Statistical Methods & Applications
https://doi.org/10.1007/s10260-021-00580-8

REVIEW PAPER

Group penalized quantile regression

Mohamed Ouhourane[1] · Yi Yang[2] · Andréa L. Benedet[3] · Karim Oualkacha[1]

Accepted: 4 July 2021
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Quantile regression models have become a widely used statistical tool in genetics and in the omics fields because they can provide a rich description of the predictors' effects on an outcome without imposing stringent parametric assumptions on the outcome-predictors relationship. This work considers the problem of selecting grouped variables in high-dimensional linear quantile regression models. We introduce a group penalized pseudo quantile regression (GPQR) framework with both group-lasso and group non-convex penalties. We approximate the quantile regression check function using a pseudo-quantile check function. Then, using the majorization–minimization principle, we derive a simple and computationally efficient group-wise descent algorithm to solve group penalized quantile regression. We establish the convergence rate property of our algorithm with the group-Lasso penalty and illustrate the GPQR approach performance using simulations in high-dimensional settings. Furthermore, we demonstrate the use of the GPQR method in a gene-based association analysis of data from the Alzheimer's Disease Neuroimaging Initiative study and in an epigenetic analysis of DNA methylation data.

Keywords Coordinate descent algorithm · Group penalized regression · Heterogeneous · Pseudo-quantile · Variable selection · Quantile regression

For the Alzheimer's Disease Neuroimaging Initiative.

✉ Mohamed Ouhourane
Mohamed.ouhourane@gmail.com

Yi Yang
yi.yang6@mcgill.ca

Andréa L. Benedet
andrea.benedet@mail.mcgill.ca

Karim Oualkacha
oualkacha.karim@uqam.ca

[1] Department of Mathematics, Université du Québec à Montréal, Montreal, Canada

[2] Department of Mathematics and Statistics, McGill University, Montreal, Canada

[3] Translational Neuroimaging Laboratory, McGill University Research Centre for Studies in Aging, Montreal, Canada

Published online: 19 August 2021

1 Introduction

Given the high-dimensional nature of omics experiments (omics refers to genomics, metabolomics, proteomics and transcriptomics), data regularization is becoming a standard approach to better extract relevant predictors for an outcome because there is typically a wild excess of predictors over participants. These top-ranked or selected predictors can be meaningful with respect to having a functional relationship to the trait or outcome. The lasso regularized regression (Tibshirani 1996) and its generalizations are attractive data-regularization tools for analyzing high-dimensional data.

In many situations, it is reasonable to group predictors so that the predictors belonging to the same group are included or excluded from a model simultaneously. For instance, in genome-wide association studies (GWAS) (Zhou et al. 2011; Lange et al. 2014), to understand the underlying biological structure of a complex disease better (e.g. Alzheimer's disease), one might want to group single-nucleotide-polymorphisms (SNPs) within a gene or genes within a biochemical pathway and then exploit group structure effects on a disease. In epigenetics studies, considering the correlations between methylation levels (features) in nearby positions along the genome can lead to better identifying differentially methylated genomic regions between two groups (outcome) (Lakhal-Chaieb et al. 2017). Another attractive motivation of the group-variable selection models is the additive model with polynomial or non-parametric components, whereby each component/group may be expressed as a linear combination of basis functions of the original variables. In this context, the selection of important variables corresponds to the selection of groups of basis functions.

One can achieve group-variable selection by adding group penalties to the regularization-based regression approaches. The group lasso penalty (Yuan and Lin 2006; Meier et al. 2008), also denoted as an L_1/L_2 penalty, is an extension of the lasso for performing variable selection on (predefined) groups of variables in generalized linear regression models. If each group of predictors reduces to a single predictor, then the group lasso penalty reduces to a standard lasso penalty. Because both Yuan and Lin (2006)'s and Meier et al. (2008)'s approaches require orthonormality of the groups, recently, Yang and Zou (2015) used a quadratic majorization trick within block descent algorithms to relax the predictors' groupwise orthonormality assumption. Moreover, Yang and Zou (2015) developed efficient algorithms to solve group-lasso penalized regression for a class of loss functions, including ordinary least squares, logistic, and several large margin classifier loss functions.

Like the regular lasso, the group-lasso lacks the selection consistency property because it tends to overly shrink the relevant group of variables. To overcome the over-shrinkage problem and gain selection consistency, Wei and Zhu (2012) have extended the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001) and the minimax concave penalty (MCP) (Zhang 2010) for group variable selection, however both approaches require the groups to be orthonormal. Breheny and Huang (2015) suggested group-SCAD and group-MCP penalized approaches in the case of ordinary least squares (OLS) regression and a general design matrix (i.e. non-orthonormal groups).

Omics data, however, are heterogeneity prone, and covariates may have different effects on different segments of the conditional distribution of the response. These types of heterogeneity are of interest to many researchers; however, they tend to be overlooked by using (group) penalized OLS methods that only capture the effects of the covariates on the mean of the conditional distribution.

Quantile regression (QR) assesses how conditional quantiles of the response variable vary with respect to measured covariates (Koenker and Bassett 1978; Koenker and Hallock 2001). By allowing estimation of the predictor effects in different quantiles, QR provides a more complete picture of the conditional distribution of the response variable than the single estimate of the conditional mean that can be obtained via OLS regression. QR is widely used in genetics and in the omics fields, and Briollais and Durrieu (2014) provide a good review of its application in these fields.

Many recent studies have focused on penalized QR in high dimensional settings. Earlier in this decade, several authors investigated the theoretical properties of penalized QR, including Belloni and Chernozhukov (2011a), Wang (2013), and Fan et al. (2014a) for the lasso penalty; Wang et al. (2012), Fan et al. (2014b), and references therein for the non-convex penalties. More recently, several studies have focused on the computational aspect for solving the penalized QR framework, including Wu and Lange (2008) and Li and Zhu (2008) for the lasso penalty; Peng and Wang (2015) for non-convex penalties; Yi and Huang (2017) for the elastic net penalty; Juban et al. (2016) and Mkhadri et al. (2017) for the lasso, SCAD and MCP penalties.

Several authors have introduced penalized QR in the context of both semi- and non-parametric frameworks (Oh et al. 2011; Zhao et al. 2005). Waldmann et al. (2013) proposed a Bayesian semi-parametric QR additive model, where penalized splines are employed for non-parametric components, and the lasso penalty is employed for the parametric components. However, model selection is restricted to the (linear) parametric part of the model, and there is no selection in the non-parametric part of the model. Fenske et al. (2011) extended the boosting algorithm to a semi-parametric QR, which allows for selection of both linear and nonlinear effects.

Although the theoretical aspect of group penalized QR has recently been addressed by a few authors, computationally efficient algorithms for solving groupwise penalized QR have received less attention in the literature. Kato (2011) developed theoretical results for the convergence rate and the oracle property of the group-lasso QR estimator. To estimate the model parameters, the author transformed the group-lasso QR problem to a second order cone programming (SOCP) problem and then used an interior point algorithm to solve it. Anterior point algorithms, however, can be computationally challenging in the presence of high dimensional data (Efron et al. 2007). Asymptotic normality of the adaptive group-lasso QR estimator was addressed in Ciuperca (2019), for a fixed and divergent number of the groups. Hashem et al. (2016) proposed a group-lasso penalized QR for the binary response. In Hashem et al. (2016), a continuous latent variable is considered to govern the binary response, and techniques similar to those used in Bayesian lasso (binary) QR frameworks (Ji et al. 2012; Kozumi and Kobayashi 2011) are employed to develop a Bayesian Gibbs sampling procedure to estimate the model parameters. Because continuous priors are imposed on the regression parameters, sparsity cannot be achieved (i.e. draws from the posterior distributions are never exactly zero), and variable selection needs further manipulation. Finally, although Peng and Wang (2015) claimed that their R package software, `rqPen`, performs groupwise penalized QR, the method as described in their manuscript only handles single-variable-selection non-convex penalized QR and no procedure within the `rqPen R` package achieves group selection. In summary, computationally-efficient methods are lacking for group variable selection in QR.

In this work, we develop a unified computationally-efficient framework for solving penalized quantile regression with group-lasso, group-SCAD, and group-MCP penalties. Because one of the biggest challenges in solving QR lies in the non-differentiability of the loss/check function (Koenker and Hallock 2001; Hunter and Lange 2000), we rely on the pseudo-quantile check functions proposed in Aravkin et al. (2014) and Oh et al. (2011), and we use the majorization-minimization principle within block coordinate descent algorithms to solve the groupwise regularized QR problem. We also develop two additional alternative algorithms to solve the group-SCAD and group-MCP penalized QR based on the local linear approximation trick (Zou and Li 2008). Our framework, termed group penalized pseudo quantile regression (GPQR), allows for general design matrices. That is, it does not require the predictors to be groupwise orthonormal. The framework is implemented in an R software package, `GPQR`, which is publicly available in `GitHub` (https://github.com/KarimOualkach/GPQR). Moreover, we study the rate of convergence of our framework for the group-lasso penalty.

The remainder of this article proceeds as follows. In Sect. 2 we formulate our GPQR framework, we provide the convergence rate analysis of our algorithm for the group lasso penalty, and we give details about the algorithm's implementation. Evaluation of the performance of our methods through exhaustive simulation studies is considered in Sect. 3. In Sect. 4, the use of the proposed methodology is illustrated in gene-based analyses of two interesting real genetic datasets. We conclude with a discussion section.

2 Pseudo quantile regression and group penalizations

Let $\{(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)\}$ be observed data, where y_i is the observed response and $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ is a $(p+1)$-dimensional observed vector of predictors for subject $i = 1, \ldots, n$. We denote by $\boldsymbol{X}$ the design matrix with n rows and $p+1$ columns. We assume that the predictors $1, X_1, \ldots, X_p$ are put into K groups $(1, 2, 3, \ldots, p+1) = \bigcup_{k=1}^{K} I_k$, where the size of each group is p_k (the cardinality of index set I_k is p_k) and the groups are non-overlapping ($I_k \cap I_{k'} = \emptyset$ for $k \neq k'$). Because the intercept is included, we assume $I_1 = \{1\}$.

The group penalized QR problem can be formulated as

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}}\Big(R(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}) + \sum_{k=1}^{K}P_{\lambda,w_k}(\|\boldsymbol{\beta}_k\|_2)\Big), \tag{1}$$

where $\rho_\tau(u) = |\tau - I(u \le 0)| \cdot |u|$, is the so-called check/hinge function (Koenker and Hallock 2001), and $(\hat{\boldsymbol{\beta}}_\tau)_k$ is the vector of the effects of the predictors belonging to group k on the τth conditional quantile of the response. Hereafter, for ease of notation, we drop the subscript for the vector $\boldsymbol{\beta}_\tau$ when no confusion arises. $P_{\lambda,w_k}(\cdot)$ is the penalty function with regularization parameter λ and penalty weight, w_k, for the group k. The weight's default value is $w_k = \sqrt{p_k}$. Because the intercept is not penalized, $w_1 = 0$. In this work, we consider the group lasso (Glasso), group MCP (GMCP), and group SCAD (GSCAD) penalties which are defined respectively by the penalty function, $P_{\lambda,w_k}(\|\boldsymbol{\beta}_k\|_2)$, as follows

$$\lambda w_k\|\boldsymbol{\beta}_k\|_2, \tag{2}$$

$$\Big\{\Big\{ w_k(\lambda\|\boldsymbol{\beta}_k\|_2 - \frac{\|\boldsymbol{\beta}_k\|_2^2}{2\theta})\mathbb{1}_{(\|\boldsymbol{\beta}_k\|_2 \le \theta\lambda)}, w_k\frac{1}{2}\lambda^2\theta\mathbb{1}_{(\|\boldsymbol{\beta}_k\|_2 > \theta\lambda)},. \tag{3}$$

$$\Big\{\Big\{ \lambda w_k\|\boldsymbol{\beta}_k\|_2\mathbb{1}_{(\|\boldsymbol{\beta}_k\|_2 \le \lambda)}, w_k\frac{\theta\lambda\|\boldsymbol{\beta}_k\|_2 - (\|\boldsymbol{\beta}_k\|_2^2 + \lambda^2)/2}{\theta - 1}\mathbb{1}_{(\lambda < \|\boldsymbol{\beta}_k\|_2 \le \theta\lambda)}, w_k\frac{\lambda^2(\theta^2 - 1)}{2(\theta - 1)}\mathbb{1}_{(\|\boldsymbol{\beta}_k\|_2 > \theta\lambda)},. \tag{4}$$

where θ is a second tuning parameter of the GMCP and GSCAD penalties, with $\theta > 1$ for GMCP and $\theta > 2$ for GSCAD. Investigation of optimal values of θ has been discussed in the literature and fixed values, such as $\theta = 4$ for GSCAD and $\theta = 3$ for GMCP, have been suggested as suitable for many problems; however, the performance does not improve significantly with θ selected by data driven approaches, (Fan and Li 2001; Ogutu and Piepho 2014). We therefore set θ equal to the recommended values in all our simulations and real data analyses.

Solving (1) can be very computationally challenging, especially in high-dimensional settings, owing to the non-differentiability of $\rho_\tau(\cdot)$. To overcome this issue, we suggest replacing $\rho_\tau(\cdot)$ in Eq. (1) with one of the following two pseudo-quantile approximation loss functions (Mkhadri et al. 2017):

$$\Psi^{(1)}_{\tau,\delta}(u) = \begin{cases} (\tau - 1)(u + \frac{\delta}{2}) & \text{if } u < -\delta \\ \frac{(1-\tau)u^2}{2\delta} & \text{if } -\delta \le u < 0 \\ 0.5\tau u^2/\delta & \text{if } 0 \le u < \delta \\ \tau(u - 0.5\delta) & \text{if } \delta \le u \end{cases} \tag{5}$$

$$\Psi^{(2)}_{\tau,\delta}(u) = \begin{cases} (\tau-1)u - \dfrac{\delta(1-\tau)^2}{2} & \text{if } u < \dfrac{\tau-1}{\delta^{-1}} \\ \dfrac{1}{2\delta}u^2 & \text{if } \dfrac{\tau-1}{\delta^{-1}} \le u \le \tau\delta \\ \tau u - \dfrac{\delta\tau^2}{2} & \text{if } u > \tau\delta. \end{cases} \tag{6}$$

Hence, the GPQR problem, in its general form, is given by

$$\hat{\boldsymbol{\beta}}(\delta) = \arg\min_{\boldsymbol{\beta}}\Big(R_\delta(\boldsymbol{\beta}) := L(\boldsymbol{\beta}) + \sum_{k=1}^{K} P_{\lambda,w_k}(\|\boldsymbol{\beta}_k\|_2)\Big), \tag{7}$$

where $L(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\Psi_\tau(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})$ and $\Psi_\tau(\cdot) = \Psi^{(1)}_{\tau,\delta}(\cdot)$ or $\Psi^{(2)}_{\tau,\delta}(\cdot)$ is one of the two pseudo functions (5) or (6). Figure 1 (left panel) illustrates the QR check function $\rho_\tau(\cdot)$ and the pseudo loss function, $\Psi^{(1)}_{\tau,\delta}(\cdot)$, for $\tau = 0.25$ and $\delta = \{1, 2\}$. The right panel contrasts the function $\Psi^{(2)}_{\tau,\delta}(\cdot)$ and $\rho_\tau(\cdot)$ for $\delta = \{2, 4\}$ and $\tau = 0.75$. Actually, when δ becomes small, the two pseudo loss functions become close in shape to the QR check function; however, both functions are differentiable everywhere and have continuous derivatives.

The pseudo approximation (5) is proposed by Jennings et al. (1996). It has also been used by Oh et al. (2011) and Zhao et al. (2005) in the context of nonparametric QR. The pseudo approximation (6) is introduced by Aravkin et al. (2014). The first pseudo approximation is given by four intervals; however, the

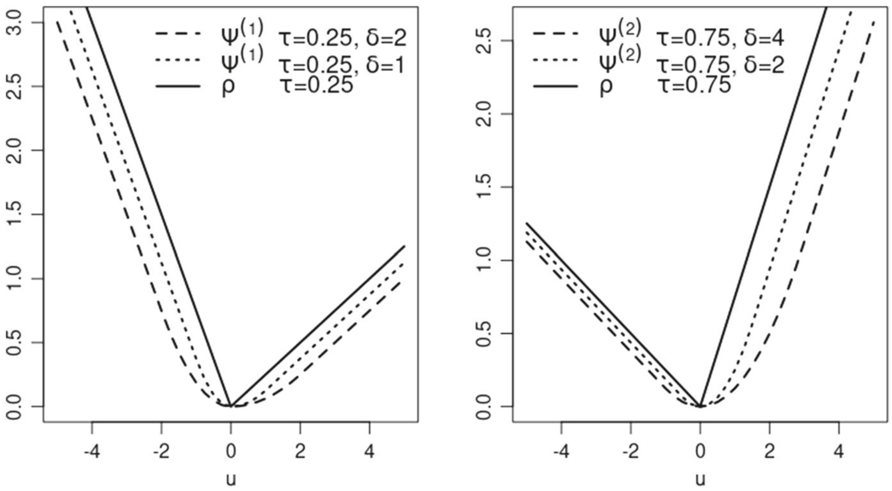

Fig. 1 Left panel: the standard quantile function $\rho_{0.25}(.)$ is shown by a solid line and the pseudo quantile function $\Psi^{(1)}_{\tau,\delta}(.)$ for $\tau = 0.25$ and $\delta = \{1, 2\}$ are shown by the dotted and dashed lines, respectively. Right panel: the standard quantile function $\rho_{0.75}(.)$ is shown by a solid line and the pseudo quantile function $\Psi^{(2)}_{\tau,\delta}$ for $\tau = 0.75$ and $\delta = \{2, 4\}$ are shown by the dotted and dashed lines, respectively

second pseudo approximation is defined only on three intervals, which leads to a difference in computation times when calculating the gradient in favor of (6) (Mkhadri et al. 2017). The next proposition provides the theoretical justifications for the success of these two approximations in providing a good solution for the initial problem (1).

Proposition 1 *For any fixed value of δ, let $\hat{\boldsymbol{\beta}}(\delta)$ be the unique minimizer of $R_\delta(\boldsymbol{\beta})$ in (7). Then we have*

$$\inf_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) \leq R(\hat{\boldsymbol{\beta}}(\delta)) \leq \inf_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + 2\kappa\delta,$$

where $R(\boldsymbol{\beta})$ is the exact group penalized quantile regression loss function defined in (1) and $\kappa = \max(\tau, 1-\tau)/2 \;\text{ or }\; \max(\tau^2, (1-\tau)^2)/2$.

The proof of Proposition 1 is detailed in Sect. 1 of the Supplementary material. The above two inequalities are true for all possible values of the tuning parameters (λ for GLasso or (λ, θ) for GMCP/GSCAD). Thus, we can compute the solution of (1) for the three group penalties by solving (7) with a small value of δ. In fact, as $\delta \to 0$, the QR with original check function $\rho_\tau(.)$ and its pseudo approximations $\Psi_\delta(.)$ are very similar. Mkhadri et al. (2017) showed that the convergence speed of pseudo-QR with the lasso penalty is greatly decreased for small values of δ, and therefore, it can be used to control the trade-off between speed and accuracy. For the GPQR framework, we followed Mkhadri et al. (2017) and set $\delta = 1$ in all analyses, which is a suitable value of δ to balance between algorithm computational efficiency and model accuracy. This is also the default value of this parameter in their SQR R package.

To solve problem (7), we propose a groupwise descent algorithm; the details are as follows. Let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_{k-1}, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{k+1}, \ldots, \tilde{\boldsymbol{\beta}}_K)$ be the current iteration and $\tilde{\boldsymbol{\beta}}_{-k} = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_{k-1}, \tilde{\boldsymbol{\beta}}_{k+1}, \ldots, \tilde{\boldsymbol{\beta}}_K)$ be the current iteration with the k-th group excluded. Suppose we are updating the k-th group of $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta}_k = (\beta_1, \ldots, \beta_{p_k})^\top$ for some $k \in \{1, \ldots, K\}$. Furthermore, consider the objective function $R_\delta(\boldsymbol{\beta})$ in (7) as a function of the k-th group $\boldsymbol{\beta}_k$, while keeping all the other groups fixed at $\tilde{\boldsymbol{\beta}}_{-k}$, that is, $R_\delta(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) := R_\delta(\boldsymbol{\beta})_{\boldsymbol{\beta}_{k'} = \tilde{\boldsymbol{\beta}}_{k'}, 1 \leq k' \leq K, k' \neq k}$ and $L(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) := L(\boldsymbol{\beta})_{\boldsymbol{\beta}_{k'} = \tilde{\boldsymbol{\beta}}_{k'}, 1 \leq k' \leq K, k' \neq k}$. Thus, at each iteration, we optimize the objective function $R_\delta(\boldsymbol{\beta})$ only in terms of the k-th group variables $\boldsymbol{\beta}_k$, while keeping all the other groups fixed at $\tilde{\boldsymbol{\beta}}_{-k}$ (i.e, $\tilde{\boldsymbol{\beta}}_k^{\text{new}} \leftarrow \arg\min_{\boldsymbol{\beta}_k} R_\delta(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$). To solve this problem efficiently for group k, we derive an upper-bound quadratic-form approximation for $R_\delta(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$ based on the quadratic majorization property of $L(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$ given in the next proposition. Then we minimize the surrogate majorizing quadratic form rather than the actual $R_\delta(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$.

Proposition 2 *Let $\mathbf{X}_k$ be the sub-matrix of $\mathbf{X}$ corresponding to group k. The quadratic majorization condition is satisfied for both pseudo loss approximations. That is, for all $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ we have*

$$L(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) \leq L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top \nabla_k L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + \frac{1}{2}(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top \mathbf{H}_k (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k), \tag{8}$$

where $\mathbf{H}_k = \frac{2\boldsymbol{X_k}^\top \boldsymbol{X_k}/n}{\delta/max(\tau,1-\tau)}$ for $\Psi^{(1)}_{\tau,\delta}(\cdot)$, and $\mathbf{H}_k = \frac{2\boldsymbol{X_k}^\top \boldsymbol{X_k}/n}{\delta}$ for $\Psi^{(2)}_{\tau,\delta}(\cdot)$.

The proof of Proposition 2 is detailed in Sect. 2 of the Supplementary material. The upper bound in (8) can be further relaxed to get the following upper bound approximation of $R_\delta(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$

$$\begin{aligned} R_\delta(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) &\leq Q(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) \\ &:= L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top \nabla_k L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) \\ &\quad + \frac{\gamma_k}{2}(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) + P_{\lambda, w_k}(\|\boldsymbol{\beta}_k\|_2), \end{aligned} \tag{9}$$

where γ_k is the largest eigenvalue of the matrix $\mathbf{H}_k$.

Thus, we minimize the quadratic form $Q(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$ groupwise, while cycling through groups. The update solution using $Q(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$ of (9) has a closed form for the three group penalties.

Note, after updating all the groups in a cycle, one can verify that the objective function (7) is decreased (i.e., it satisfies the descent property) using the majorization-minimization principle Hunter and Lange (2000, 2004). This assures the convergence of the GPQR algorithms.

Validation of GPQR convergence is carried out through simulation scenarios in Sect. 3.3 to demonstrate that the algorithm solution satisfies the Karush–Kuhn–Tucker (KKT) conditions. The derivation of both the theoretical and numerical KKT conditions of the GPQR algorithm are outlined in Sects. 6 and 7 of the Supplementary material, respectively. Our KKT conditions are calculated based on the pseudo QR objective function $R_\delta(\boldsymbol{\beta})$ given in (7).

Next, we present the GPQR framework in detail for each group-penalty.

2.1 Pseudo QR with group-Lasso penalty

This section gives details of the GPQR algorithm with the Glasso penalty and its convergence rate properties.

In this case, the penalty term $P_{\lambda, w_k}(\|\boldsymbol{\beta}_k\|_2)$ in (9) is replaced by the Glasso penalty given in (2). By employing the proximal gradient algorithm for $Q(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$ to update $\boldsymbol{\beta}_k$, one can write

$$\begin{aligned} \tilde{\beta}_k^{\text{new}} &= \arg\min_{\beta_k} Q\left(\beta_k, \tilde{\beta}_{-k}\right) \\ &= \arg\min_{\beta_k} \frac{1}{2}\left\|\beta_k - \left(\tilde{\beta}_k - \gamma_k^1 \nabla_k L\left(\tilde{\beta}_k, \tilde{\beta}_{-k}\right)\right)\right\|_2^2 + \lambda\omega_k\gamma_k^{-1}\|\beta_k\|_2 \\ &= \text{prox}_{\lambda\omega_k\gamma_k^{-1}h}\left(\tilde{\beta}_k - \gamma_k^{-1}\nabla_k L\left(\tilde{\beta}_k, \tilde{\beta}_{-k}\right)\right) \end{aligned} \tag{10}$$

where the proximal mapping of the function $h(.) = \|.\|_2$ is given by

$$\text{prox}_{\lambda h}(\mathbf{u}) = \arg\min_{\mathbf{v}} \lambda h(\mathbf{v}) + \frac{1}{2}\|\mathbf{v} - \mathbf{u}\|_2^2.$$

The following algorithm gives details of the GPQR with the Glasso penalty.

Algorithm 1 The GPQR algorithm for the Glasso penalty

```
Calculate $\gamma_k$, the maximum eigenvalue of $\mathbf{H}_k$ for $k = 1, \ldots, K$, and initialize $\tilde{\boldsymbol{\beta}}$;
repeat
    for $k = 1, 2, \ldots, K$ do
        $\tilde{\boldsymbol{\beta}}_k^{\text{new}} \leftarrow \text{prox}_{\lambda w_k \gamma_k^{-1} h}(\tilde{\boldsymbol{\beta}}_k - \gamma_k^{-1} \nabla_k L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}))$
    end
until Convergence of $\tilde{\boldsymbol{\beta}}$;
Return $\tilde{\boldsymbol{\beta}}$
```

The next theorem provides the convergence rate analysis of the GPQR algorithm with the Glasso penalty.

Theorem 1 *The GPQR algorithm with Glasso penalty* (*Algorithm* 1) converges at least linearly to the global solution $\boldsymbol{\beta}^*$.

The proof of Theorem 1 is relegated to Sect. 3 of the Supplementary material.

2.2 Pseudo QR with GSCAD and GMCP penalties

The nonconvex group penalties, GSCAD and GMCP, are used both to perform group variable-selection and to reduce the bias towards zero introduced by the Glasso. For instance, to understand the effect of the GSCAD penalty (4) compared with the Glasso (2), let us consider its derivative function, which relies directly on the shrinkage amount of the parameters. For small values of $\|\boldsymbol{\beta}_k\|_2$ (i.e. $\|\boldsymbol{\beta}_k\|_2 \leq \lambda$), GSCAD exercises the same shrinkage on the parameters' effects, as the Glasso does (i.e. $P'_{\lambda,w_k}(\|\boldsymbol{\beta}_k\|_2) = \lambda$). However, the GSCAD penalty continuously reduces the shrinkage for $\lambda \leq \|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda$, and the shrinkage becomes zero when $\|\boldsymbol{\beta}_k\|_2 \geq \lambda\theta$ (i.e. $P'_{\lambda,w_k}(\|\boldsymbol{\beta}_k\|_2) = 0$). A similar reasoning can explain the GMCP penalty effect (Breheny and Huang 2011).

The following proposition gives closed form solutions to the update, $\tilde{\boldsymbol{\beta}}_k^{\text{new}}$, in (9) when $P_{\lambda,w_k}(\|\boldsymbol{\beta}_k\|_2)$ is given by (3) for GMCP, and by (4) for the GSCAD.

Proposition 3 *Let $Q(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}})$ be the surrogate function given by (9) and let $P_{\lambda,w_k}(\|\boldsymbol{\beta}_k\|_2)$ be one of the two penalties given in (3) and (4). The closed form solutions to (9) of $\tilde{\boldsymbol{\beta}}_k^{\text{new}}$ for the GPQR algorithm with the GMCP and GSCAD penalties are, respectively, given by*

$$\tilde{\beta}_k^{\text{new}} \longleftarrow F(\mathbf{Z}_k) = \begin{cases} \dfrac{1}{\gamma_k - w_k/\theta} \dfrac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} S(\|\mathbf{Z}_k\|_2, \lambda w_k), & \text{if } \|\mathbf{Z}_k\|_2 \leq \gamma_k\theta\lambda \\ \dfrac{1}{\gamma_k}\mathbf{Z}_k, & \text{if } \|\mathbf{Z}_k\|_2 > \gamma_k\theta\lambda, \end{cases} \tag{11}$$

$$\tilde{\boldsymbol{\beta}}_k^{\text{new}} \longleftarrow F(\mathbf{Z}_k) = \begin{cases} \frac{1}{\gamma_k}\frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2}S(\|\mathbf{Z}_k\|_2, \lambda w_k), & \text{if} \quad \|\mathbf{Z}_k\|_2 \leq (w_k+\gamma_k)\lambda \\ \frac{1}{\gamma_k - \frac{w_k}{\theta-1}}\frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2}(\|\mathbf{Z}_k\|_2 - \frac{\lambda w_k \theta}{\theta - 1}), & \text{if} \quad (w_k+\gamma_k)\lambda < \|\mathbf{Z}_k\|_2 \leq \gamma_k\theta\lambda \\ \frac{1}{\gamma_k}\mathbf{Z}_k, & \text{if} \quad \|\mathbf{Z}_k\|_2 > \gamma_k\theta\lambda \end{cases}, \tag{12}$$

where $\mathbf{Z}_k = \gamma_k\tilde{\boldsymbol{\beta}}_k - \nabla_k L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k})$, and S(.) is the soft-threshold operator, defined as

$$S(\|\mathbf{z}\|_2, \lambda) := \begin{cases} 0, & \text{if } \|\mathbf{z}\|_2 \leq \lambda \\ \|\mathbf{z}\|_2 - \lambda, & \text{if } \|\mathbf{z}\|_2 > \lambda. \end{cases}$$

The proof of Proposition 3 is detailed in Sect. 4 of the Supplementary material.

The following algorithm summarizes the steps of the GPQR framework with the GMCP or GSCAD penalty:

```
Algorithm 2 The GPQR algorithm with the GMCP or GSCAD penalty
Calculate $\gamma_k$, the maximum eigenvalue of $\mathbf{H}_k$ for $k = 1, \ldots, K$, and initialize $\tilde{\boldsymbol{\beta}}$;
repeat
    for $k = 1, 2, \ldots, K$ do
        $\tilde{\boldsymbol{\beta}}_k^{\text{new}} \leftarrow F(-\nabla_k L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + \gamma_k\tilde{\boldsymbol{\beta}}_k)$
    end
    where $F(\cdot)$ is given by (11) and (12) for GMCP and GSCAD penalties, respectively;
until Convergence of $\tilde{\boldsymbol{\beta}}$;
Return $\tilde{\boldsymbol{\beta}}$.
```

The convexity of $P_{\lambda,w_k}(t)$ for the Glasso is a crucial property for proving the convergence, at least linearly, of the GPQR in Theorem 1. However, this property is not available for the non-convex GMCP and GSCAD penalties.

2.3 Pseudo QR with group local linear approximation penalty

In this section, we propose to extend the local linear approximation (LLA) trick to solve the GPQR with the GMCP and GSCAD penalties to remedy the possible computational weakness of the two nonconvex penalties.

The LLA approximation is based on the first order Taylor expansion of the MCP or SCAD penalty functions around $\|\tilde{\boldsymbol{\beta}}_k\|_2$. Thus, one can write

$$P_{\lambda,w_k}(\|\boldsymbol{\beta}_k\|_2) \approx P_{\lambda,w_k}(\|\tilde{\boldsymbol{\beta}}_k\|_2) + P'_{\lambda,w_k}(\|\tilde{\boldsymbol{\beta}}_k\|_2)(\|\boldsymbol{\beta}_k\|_2 - \|\tilde{\boldsymbol{\beta}}_k\|_2), \tag{13}$$

where $P_{\lambda,w_k}(.)$ is one of the two penalties given in (3) and (4).

Substituting (13) into (9) leads to the following update for the GPQR with the group local linear approximation (GLLA) penalty

$$\tilde{\boldsymbol{\beta}}_k^{\text{new}} = \arg\min_{\boldsymbol{\beta}_k} Q(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) + \lambda w'_k \|\boldsymbol{\beta}_k\|_2, \tag{14}$$

where $w'_1 = 0$ and $w'_k = \dfrac{w_k P'_{\lambda,w_k}(\|\tilde{\boldsymbol{\beta}}_k\|_2)}{\lambda}$ for $k = 2, \ldots, K$. The weight w'_k depends on the penalty function through the first derivative, $P'_{\lambda,w_k}(\|\tilde{\boldsymbol{\beta}}_k\|_2)$, which is given for the GMCP and GSCAD, respectively, as follows:

$$\begin{cases} \lambda - \dfrac{\|\tilde{\beta}_k\|_2}{\theta}, & \text{if } \|\tilde{\beta}_k\|_2 \leq \theta\lambda \\ 0, & \text{if } \|\tilde{\beta}_k\|_2 > \theta\lambda, \end{cases}$$

$$\begin{cases} \lambda, & \text{if } \|\beta_k\|_2 \leq \lambda \\ \dfrac{\theta\lambda}{\theta - 1} - \dfrac{\|\tilde{\beta}_k\|_2}{\theta - 1}, & \text{if } \lambda < \|\beta_k\|_2 \leq \theta\lambda \\ 0, & \text{if } \|\beta_k\|_2 > \theta\lambda. \end{cases}$$

The problem (14) can be solved using a Glasso-type update similar to Algorithm 1 described in Sect. 2.1. Thus, we use the proximal gradient algorithm in (10) to solve it.

The details of the GPQR approach with the GLLA penalty is described in Algorithm 3.

Algorithm 3 The GPQR algorithm with the GLLA penalty

```
Initialize $\tilde{\boldsymbol{\beta}}$ and set $(w'_1, \gamma_1) = (0, 0)$;
Calculate $\gamma_k$ as the maximum eigenvalue of $\mathbf{H}_k$ and $w'_k = \lambda^{-1} P'_{\lambda,w_k}(\|\tilde{\boldsymbol{\beta}}_k\|_2)$ for $k = 2, \ldots, K$;
repeat
    for $k = 1, 2, \ldots, K$ do
        $\tilde{\boldsymbol{\beta}}_k^{\text{new}} \leftarrow \text{prox}_{\lambda w'_k \gamma_k^{-1} h}(\tilde{\boldsymbol{\beta}}_k - \gamma_k^{-1} \nabla_k L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}))$,
    end
    for $k = 2, 3, \ldots, K$ do
        $w_k'^{new} \leftarrow \dfrac{w_k P'_{\lambda,w_k}(\|\tilde{\boldsymbol{\beta}}_k^{\text{new}}\|_2)}{\lambda}$,
    end
until Convergence of $\tilde{\boldsymbol{\beta}}$;
Return $\tilde{\boldsymbol{\beta}}$.
```

Note that the GLLA penalty is a convex majorant of the GMCP (or GSCAD) penalty. Thus, for each fixed value of λ, the GLLA allows a search of the solution in a locally convex region, and consequently it may lead to stable and smooth path solutions.

A comparison of the GLLA approximation and the exact GMCP and GSCAD penalties is illustrated in Sect. 5 of the Supplementary material. Figure S.1 shows that the exact and approximate path solutions of the GPQR algorithm with nonconvex penalties are nearly identical for all values of the tuning parameter λ. This proves the efficiency of the GLLA approximation.

2.4 Implementation

In this section we give details about the implementation of the proposed GPQR algorithms.

The intercept term is always included in all our models. Each GPQR model is solved by using a fine grid of λ. We proceeded by choosing $\lambda_{\max}$ which is the smallest λ that allows all groups, $\boldsymbol{\beta}_k, (2 \leqslant k \leqslant K)$, to be zero except the intercept. To obtain $\lambda_{\max}$, we first calculated the estimates, $\hat{\beta}_0$, for the null model with only the intercept:

$$\hat{\beta}_0 = \arg\min_{\beta_0} \frac{1}{n} \sum_{i=1}^{n} \Psi_\tau(y_i - \beta_0). \tag{15}$$

According to the KKT conditions of (15), we derived the following formula:

$$\lambda_{\max} = \max_{k=2,\ldots,K} \|\nabla_k L(\hat{\beta}_0, \mathbf{0})\|_2 / \omega_k.$$

Let $\lambda_{\min} = \eta \lambda_{\max}$, where $0 < \eta < 1$ is a small number. We generated a sequence of λs by placing 98 evenly spaced points, $\{\lambda^{[l]}\}_{l=2}^{99}$, between $\lambda_{\max}$ and $\lambda_{\min}$ in log-scale and let $\lambda^{[1]} = \lambda_{\max}$ and $\lambda^{[100]} = \lambda_{\min}$.

We also used the warm-start trick in solving the solution paths: the solution of $\widehat{\boldsymbol{\beta}}$ at $\lambda^{[l-1]}$ is taken as the initial value for solving the solution of $\widehat{\boldsymbol{\beta}}$ at $\lambda^{[l]}$.

For computing efficiency at each λ, we used the "strong rule statement" proposed by Tibshirani et al. (2012), which screens out group predictors. Let $\hat{\boldsymbol{\beta}}^{[l]}$ be the solution at $\lambda^{[l]}$. For finding the solution $\boldsymbol{\beta}^{[l+1]}$ at $\lambda^{[l+1]}$, we introduced a supplementary screening step to check whether a group k satisfies the following condition:

$$\|\nabla_k L(\widehat{\boldsymbol{\beta}}^{[l]})\|_2 \geq \omega_k (2\lambda^{[l+1]} - \lambda^{[l]}). \tag{16}$$

Let $\mathbf{S}$ be the subset of the predictors' groups that are not discarded by condition (16) and $\mathbf{S}^c$, its complement. According to the strong rule, at $\lambda^{[l+1]}$, the coefficients of the groups in the set $\mathbf{S}$ are very likely to be active and those of the groups in the complement set $\mathbf{S}^c$ are very likely to be inactive. If this statement is correct, then solving the proposed GPQR models will only require a reduced data set, $(\mathbf{y}, \boldsymbol{X}_{\mathbf{S}})$, where $\boldsymbol{X}_{\mathbf{S}}$ is the restricted matrix where the columns are the groups belonging to $\mathbf{S}$. Denote this solution as $\hat{\boldsymbol{\beta}}_{\mathbf{S}}$. Then, one must verify if the strong rule statement is well confirmed at $\lambda^{[l+1]}$ by verifying if $\tilde{\boldsymbol{\beta}}^{[l+1]} = (\hat{\boldsymbol{\beta}}_{\mathbf{S}}, \mathbf{0})$ satisfies the KKT conditions. Following the calculation details in Sects. 6 and 7 of the Supplementary material, this means that for the GLasso, GMCP, and GSCAD, any group k from the inactive set, $\mathbf{S}^c$, needs to satisfy the following inequality

$$\|\nabla_k L(\tilde{\boldsymbol{\beta}}^{[l+1]})\|_2 \leq \omega_k \lambda^{[l+1]}.$$

For GLLA, the inactive group, k, needs to verify

$$\|\nabla_k L(\tilde{\boldsymbol{\beta}}^{[l+1]})\|_2 \le \omega'_k \lambda^{[l+1]},$$

where ω'_k is the weight given in (14).

If there are no violations of the strong rule statement, then the solution at $\lambda = \lambda^{[l+1]}$ is $\tilde{\boldsymbol{\beta}}^{[l+1]} = (\hat{\boldsymbol{\beta}}_{\mathbf{S}}, \mathbf{0})$, otherwise we add the subset of the violator groups, denoted as $\mathbf{V}$, into the active set, $\mathbf{S} = \mathbf{S} \cup \mathbf{V}$, and repeat the whole procedure with the reduced data set $(\mathbf{y}, \boldsymbol{X}_{\mathbf{S}})$.

3 Numerical experiments

We conducted simulation studies with four scenarios to illustrate the methodology presented in this work. In the first scenario, we aimed both (i) to graphically illustrate key advantages of using group penalized quantile regression approaches to detect heterogeneous effects of predictors, as alternatives to group penalized Least-Square (LS) regression methods, and (ii) to compare the proposed approaches with existing penalized QR methods, namely, the regularized Bayesian QR(BQR) method (Alhamzawi et al. 2012), the standard quantile regression with the lasso, SCAD, and MCP penalties (Mkhadri et al. 2017), and Boosting Additive QR (BAQR) (Fenske et al. 2011). The LS methods are implemented in the *grpreg* R package (Breheny 2015), with Glasso, GSCAD, and GMCP penalties. The BQR approach is implemented in the *Brq* R package (Alhamzawi et al. 2012). The BAQR is implemented in the *mboost* R package (Hofner et al. 2014). The standard penalized quantile regression of (Mkhadri et al. 2017) uses the pseudo quantile approximation functions [(5) and (6)] to fit the QR, and is implemented in the *SQR* R package.

The second and the third scenarios targeted evaluation of the proposed approach's performance in terms of computational efficiency and prediction accuracy. The fourth scenario aimed to evaluate the GPQR algorithms convergence based on the numerical KKT conditions derived in Sect. 7 of the Supplementary material.

3.1 Simulation setting of scenarios 1, 2 and 3

3.1.1 Setting of scenario 1

To illustrate key advantages of the proposed method compared with single-variable selection QR methods, we focused on a setting in which the predictors are highly correlated in this scenario. The model is based on an illustration example in Mkhadri and Ouhourane (2013). We set the sample size to $n = 100$ observations and $p = 20$ predictors. The predictors $X_j, j = 1, \ldots, 20$, were generated as follows:

- We generated $Z_j, j = 1, \ldots, 11$, following the standard normal distribution;
- We set $X_j = Z_1 + \epsilon^x_j, j = 1, \ldots, 4, \epsilon^x_j \sim N(0, 0.1)$;
- $X_j = Z_2 + \epsilon^x_j, j = 5, \ldots, 8, \epsilon^x_j \sim N(0, 0.1)$;
- $X_j = Z_3 + \epsilon^x_j, j = 9, \ldots, 12, \epsilon^x_j \sim N(0, 0.1)$;

- $X_j = Z_{j-9}, j = 13, \ldots, 20.$

Thus, we set the predictors' effects to be

$$\boldsymbol{\beta} = (\underbrace{3,3,3,3}_{G_1},\underbrace{2,2,2,2}_{G_2},\underbrace{-1,-1,-1,-1}_{G_3},\underbrace{0,\ldots,0}_{G_4-G_{11}})^\top$$

and $\sigma = 3$. The response Y is generated from the following location-scale linear regression model

$$Y = \sum_{j=1}^{20} \boldsymbol{\beta}_j X_j + \Phi(X_{20})\epsilon, \quad \epsilon \sim N(0,3),$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. Many authors consider that using $\Phi(.)$ in variance simulation generates a model with heteroscedasticity (Wang et al. 2012; Gu and Zou 2016). The predictors X_1, X_2, X_3, and X_4 form group G_1, for which the underlying common factor is Z_1; the predictors X_5, X_6, X_7, and X_8 form the second group G_2, for which the underlying common factor is Z_2; finally, X_9, X_{10}, X_{11}, and X_{12} form the third group G_3, for which the underlying common factor is Z_3. The within-group correlations are high. An oracle estimator would identify the groups G_1, G_2, and G_3 as the important variables, and variable X_{20} (i.e. $G_{11} = \Phi(X_{20})$) when $\tau \neq 0.5$.

3.1.2 Setting of scenario 2

This scenario considers an additive model involving both continuous and categorical factors (i.e. groups of predictors) to relate y to the predictors. The model in this scenario is based in part on simulation studies conducted in Yuan and Lin (2006). We generated 21 independent random variables $Z_1, \ldots, Z_{20}$ and W from N(0, 1). We set the predictors to be defined as $X_1 = Z_1$ and $X_j = (Z_j + W)/\sqrt{2}$, for $j = 2, \ldots, 20$. Furthermore, each predictor $X_j, j = 11, \ldots, 20$, was trichotomized as $\tilde{X_j} = 0$ if X_j is smaller than $\Phi^{-1}(1/3)$, $\tilde{X_j} = 1$ if X_j is larger than $\Phi^{-1}(2/3)$, and $\tilde{X_j} = 2$ if X_j is between $\Phi^{-1}(1/3)$ and $\Phi^{-1}(2/3)$. The response was then simulated from the heterogeneous additive model

$$Y = \underbrace{3X_3^3 + X_3^2 + X_3}_{G3} + \underbrace{\frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6}_{G6} + \underbrace{2I(\tilde{X}_{11} = 0) + I(\tilde{X}_{11} = 1)}_{G11} + \Phi(X_1)\epsilon;$$

where $\epsilon \sim N(0,1)$, and $I(\cdot)$ is the indicator function. In this scenario, each continuous factor was represented by a polynomial of degree 3 and each categorical factor was represented by two levels of its corresponding trichotomized variable. Thus, by construction, we have a total $p = 50$ (i.e. 30 continuous and 20 categorical variables), and we set the sample size to $n = 50$.

3.1.3 Setting of scenario 3

In this scenario, we considered an additive model involving continuous factors represented by polynomials of degree 3 to link y to the predictors. The data generation is motivated in part by a simulation study carried out in Peng and Wang (2015). First, we simulated $(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_p)^\top$ from a multivariate normal distribution $N_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{jk})_{p\times p}$ and $\sigma_{jk} = 0.5^{|j-k|}$. Second, we set $X_1 = \Phi(\tilde{X}_1)$ and $X_j = \tilde{X}_j$ for $j = 2, \ldots, p$. Third, each variable from {6,12,15,20} was represented through a third-order polynomial. Then, we simulated the response variable from the following regression model:

$$Y = \underbrace{X_6 + X_6^2 + X_6^3} + \underbrace{\frac{1}{3}X_{12} - X_{12}^2 + \frac{2}{3}X_{12}^3} + \underbrace{\frac{1}{2}X_{15} - X_{15}^2 + \frac{1}{2}X_{15}^3} + \underbrace{X_{20} + X_{20}^2 + X_{20}^3} + X_1\epsilon,$$

where $\epsilon \sim N(0, 1)$. We considered $n = 300$ and $p = 1000$. In this scenario, we considered $\{X_j, X_j^2, X_j^3\}$ as a group when fitting penalized LS and all the proposed models. Thus, the final design matrix consists of $q = 3p = 3000$ variables.

Note, in both Scenarios 2 and 3, X_1 plays the role of the heteroskedastic predictor and does not influence the center of the response conditional distribution. Thus, one of the aims of these two settings is to test the GPQR ability to select X_1 when considering lower and/or upper conditional quantiles (i.e. $\tau \neq 0.5$).

We implemented 100 Monte Carlo replications in each of the three scenarios. Each replication consists of a training dataset of 300 observations, and a test dataset of 300 observations. The training dataset is used to fit the proposed models and their competitors (at a desired τth quantile) to determine the optimal λ using five-fold cross-validation (CV). For our models, the optimal λ corresponds to the value of λ that gives a small value for the quantile-based prediction errors (QPE_τ) defined as

$$QPE_\tau = \frac{1}{n}\sum_{i\in validation} \rho_\tau(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}).$$

The performance evaluation of the methods, including the LS methods, is computed on the test data sets, and is based on the following statistics:

- False Positive FP: the number of the groups of variables with zero coefficients incorrectly included in the final model;
- P1: the proportion of the true active/non-null groups, $\boldsymbol{\beta}_k \neq 0$, that are selected;
- P2: the proportion of simulation runs X_1 (or X_{20} in Scenario 1) is selected;
- AE: the absolute estimate error defined by $\sum_{j=0}^{p} |\hat{\beta}_j - \beta_j|$;
- The quantile-based prediction error (QPE_τ) defined as

$$QPE_\tau = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}});$$

- The root mean square error (RMSE) defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Quantile_{Y_i}(\tau|\boldsymbol{x}_i) - \widehat{Quantile}_{Y_i}(\tau|\boldsymbol{x}_i))^2},$$

where $\widehat{Quantile}_{Y_i}(\tau|\boldsymbol{x}_i))$ is the estimated value of the true quantile, $Quantile_{Y_i}(\tau|\boldsymbol{x}_i))$, of the Y_i conditional on $\boldsymbol{x}_i$. The QPE_τ and $RMSE$ statistics have been used recently in Xu et al. (2020) for model-prediction evaluation in the expectile regression framework. Note, for the LS methods, QPE is defined as the absolute deviation/prediction error (i.e. $QPE_{0.5}$), and $RMSE = [\sum_{i=1}^{n}(Quantile_{Y_i}(0.5|\boldsymbol{x}_i) - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{LS})^2/n]^{1/2}$.

In Scenario 1, two locations were investigated with the quantile-based models, $\tau = 0.5$ and 0.95; in Scenarios 2 and 3, the proposed methods were fitted for three locations/quantiles, $\tau = 0.5, 0.75$ and 0.95.

3.2 Simulation results of scenarios 1, 2, and 3

In this section we outline and discuss the results of the first three scenarios.

3.2.1 Results of scenario 1

Graphical illustration results (based on one replication) Figure 2 shows the path solutions for the grid on $[\lambda_{\min}, \lambda_{\max}]$ of λ, for the GPQR and LS methods with the Glasso, GSCAD, and GMCP. The GPQR is fitted for two locations, $\tau = \{0.5, 0.95\}$. Figure 2 shows that the coefficients' profiles of the GPQR with $\tau \in \{0.50, 0.95\}$ tend to be smooth, however, the LS paths fluctuate widely, and some coefficients are in opposite directions/signs to their true values. This poor behavior of the LS methods is remarkably confirmed by the AE statistic in Table 1, which shows substantial bias of the LS parameters' estimators. Furthermore, the heteroskedastic variable $\Phi(X_{20})$ in the scale component, represented by G_{11}, is often recovered when fitting the GPQR model for the 0.95th conditional quantile (pink group); this is not the case for the GPQR model with $\tau = 0.5$ and for the LS methods. This shows that the GPQR framework can be useful for detecting heteroskedastic groups of variables.

Numerical results Table 1 outlines the results for averages, over 100 replications, of the six statistics defined earlier. Notice that, the BQR, standard penalized QR, and BAQR methods are designed for individual variable selection, and therefore, they do not enforce the selection of a whole group of variables. Thus, for fair comparison with these three methods in this scenario, the false positive (FP) and P1 statistics were calculated for all methods as the number of predictors with zero coefficients incorrectly included in the final model and the proportion of the true active variables, respectively.

Table 1 shows that the GPQR outperforms all the other methods for almost all the statistics, except for the FP statistic, for which the BQR and BAQR have the smallest values. By contrast, because in this scenario the predictors are highly correlated, the single-variable selection methods suffer from *unstable* selection of

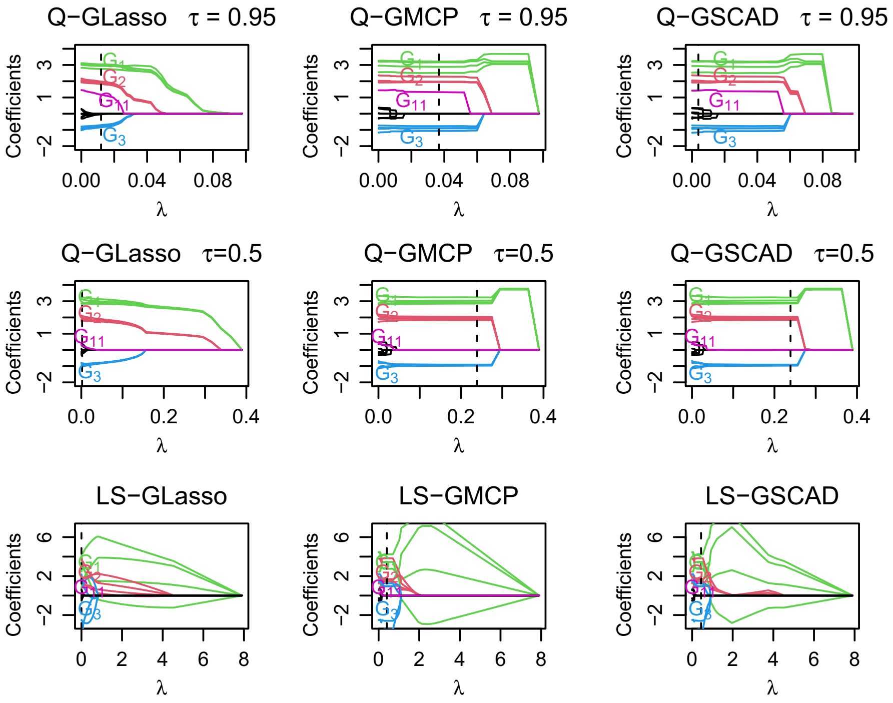

Fig. 2 At the top from left to right, the coefficient paths of the penalized quantile regression with the three group penalties (Q-Glasso, Q-GMCP, and Q-GSCAD) and $\tau = 0.95$ are shown as a function of the tuning parameter λ; the vertical dashed line reports λ selected by five-fold CV. The middle from left to right shows the coefficient paths of the penalized quantile regression with $\tau = 0.5$. The bottom row from left to right shows the coefficient paths corresponding to the grpreg package with the same three penalties. The group coefficients G_1, G_2, G_3, and G_{11} are plotted in green, red, blue and pink, respectively. The black line corresponds to the noisy groups of predictors

correlated predictors (Wang et al. 2019). This is well illustrated in the results of the P1 (especially for the BQR) and AE statistics in Table 1. The LS methods perform well in general, in this scenario, and surprisingly detect the heteroskedastic predictor, X_{20}, especially with the Glasso penalty, which has a high value of the P2 statistic (70%). When it is fitted for $\tau = 0.95$, the GPQR approach outperforms the LS for the P2 statistic, which reaches 94% for P2 using the GPQR and Glasso penalty. Yet, the LS results of the AE statistic reveal substantial bias for the model-parameter estimators. We also reported the Time statistic (in seconds) for computational efficiency comparison. The results of this statistic in Table 1 show that all the methods have comparable run-times, except for the BQR. This is not surprising because the BQR uses a Markov chain Monte Carlo (MCMC) algorithm to estimate the solution. Moreover, the BQR does not enforce sparsity, and so, it provides non-exact zero coefficient estimates and builds on the parameters' posterior distribution to provide credible intervals for variable selection. The performance of Boosting algorithms is influenced by two principal tuning parameters, including *mstop*, which is the maximum number of iterations the

Table 1 Simulation results of FP, P1, P2, AE, RMSE, QPE_τ, and running time (Time) for scenario 1, based on 100 replications

Stats	$\tau = 0.50$										
	LS-GLasso	Q-Lasso	Q-GLasso	LS-GMCP	Q-MCP	Q-GMCP	LS-GSCAD	Q-SCAD	Q-GSCAD	BQR	BAQR
FP	5.14	4.68	4.18	1.18	2.1	1.12	1.21	4.8	1.28	**0.40**	0.85
P1	**100%**	98%	**100%**	**100%**	81%	**100%**	**100%**	73%	**100%**	37%	64%
P2	70%	56%	55%	**4%**	28%	12%	15%	66%	14%	12%	12%
AE	16.4	10.0	**1.61**	15.6	10.1	3.84	15.9	23.1	3.8	16.9	14.11
RMSE	0.44	0.76	0.31	0.36	0.51	**0.24**	1.04	0.55	**0.22**	35.26	1.62
QPE_τ	0.68	0.72	0.65	0.66	0.69	**0.64**	0.76	0.69	**0.64**	2.41	1.04
Time	0.09	0.09	0.10	**0.08**	0.09	**0.08**	0.09	0.09	**0.08**	1.09	0.60

Stats	$\tau = 0.95$							
	Q-Lasso	Q-GLasso	Q-MCP	Q-GMCP	Q-SCAD	Q-GSCAD	BQR	BAQR
FP	5.35	5.05	1.8	**1.67**	2.57	1.94	2.25	2.34
P1	99%	**100%**	63%	96%	52%	98%	39%	75%
P2	91%	**94%**	60%	67%	66%	79%	90%	53%
AE	14.50	**2.82**	16.76	3.41	24.31	2.89	21.87	18.12
RMSE	7.02	**6.64**	55.45	30.57	61.54	18.29	30.35	30.46
QPE_τ	**0.28**	**0.28**	0.37	0.33	0.39	**0.29**	1.17	0.42
Time	0.22	0.22	**0.10**	**0.10**	0.11	0.11	1.12	6.73

The results are reported for our GPQR (Q-GLasso, Q-GSCAD, Q-MCP), the group-variable least squares method (LS) (LS-GLasso, LS-GSCAD, LS-GMCP), and three single-variable methods: the Bayesian quantile regression (BQR); the standard quantile regression with lasso (Q-lasso), MCP (Q-MCP), and SCAD (Q-SCAD) penalties; and boosting quantile regression (BAQR)

The best results for each statistic are highlighted in bold font

boosting algorithm will run for. Large *mstop* values lead to including more components. Oppositely, smaller *mstop* values lead to excluding more components (Mayr et al. 2014). Consequently, the FP and P1 statistics are sensitive to *mstop*. In Scenario 1, the optimum value of this parameter is selected via cross-validation.

Note, we have only reported the results of the GPQR with the check function approximation (5) in this scenario. The unreported results of the GPQR with check function approximation (6) are similar to those presented in Table 1.

3.2.2 Results of scenarios 2 and 3

The simulation results of the average, over 100 replications, of the FP, P1, P2, AE, RMSE, and QPE_τ statistics are outlined in Tables 2 and 3 for Scenarios 2 and 3, respectively. The six statistics are calculated for the GPQR approach with all suggested group penalties. In these two scenarios, we reported results for both pseudo check function approximations, (5) and (6).

Tables 2 and 3 show that all models select the true active groups, with the P_1 statistic always around 100%. By contrast, the FP statistic reveals that the GPQR with the GMCP and GSCAD penalties tends to provide less false positives than the Glasso.

The P2 statistic shows how many times the heterogeneous variable, X_1, is selected in each model fit. For $\tau = 0.5$, it is expected that X_1 will not be selected because it has no effect on the center of y. However, as τ increases, the proportion of selecting X_1 increases for all approaches. For $\tau = 0.75$, P2 ranges between $(17\%, 73\%)$, and when $\tau = 0.95$, P_2 is approximately around 100%.

3.3 Checking the KKT conditions

In this section we test the accuracy of the proposed algorithms' solutions by checking their numerical KKT conditions, defined in Sect. 7 of the Supplementary material. More precisely, because we are using the majorization-minimization principle to solve (7), the aim of this scenario is to evaluate if the minimizer $\widehat{\boldsymbol{\beta}}$, obtained by solving (9), satisfies the first-order optimality conditions (i.e. the KKT conditions) for the objective function (7). This ensures that the GPQR algorithms converge to the desired solution. Derivation of the KKT conditions is given in more detail in Sects. 6 and 7 of the Supplementary material.

3.3.1 Setting of scenario 4

We designed this simulation scenario following a numerical example suggested in Yang and Zou (2015). First, we simulated q initial predictors, $X_1, X_2, \ldots, X_q$, from a centered multivariate normal distribution with a compound symmetry correlation matrix, $\boldsymbol{\Gamma}$, with $\boldsymbol{\Gamma}_{jj'} = \rho$, for all $j \neq j'$. We then generated the response following the regression model

Table 2 Simulation results of FP, P1, P2, AE, RMSE, and QPE_τ for Scenario 2 based on 100 replications (5) and (6), respectively

Stats	$GLasso_{1^*}$	$GLasso_2$	$GMCP_1$	$GMCP_2$	$GSCAD_1$	$GSCAD_2$	$GLLA$ MCP_1	$GLLA$ MCP_2	$GLLA$ $SCAD_1$	$GLLA$ $SCAD_2$
$\tau = 0.50$										
FP	3.39	2.96	1.21	0.93	**0.74**	0.81	1.06	1.22	0.92	1.16
P1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
P2	45%	39%	5%	**0%**	**0%**	**0%**	5%	**0%**	**0%**	**0%**
AE	15.73	**15.16**	16.00	15.24	15.93	15.23	15.25	15.24	15.25	15.24
RMSE	0.030	0.031	0.042	0.040	**0.027**	0.028	0.041	0.041	0.028	0.028
QPE_τ	0.29	0.30	0.22	**0.21**	**0.21**	**0.21**	0.22	**0.21**	**0.21**	0.22
$\tau = 0.75$										
FP	2.60	2.64	**1.61**	2.00	1.71	1.70	2.08	2.01	2.02	1.83
P1	100%	100%	98%	98%	100%	100%	100%	100%	100%	100%
P2	76%	81%	43%	31%	42%	39%	50%	47%	38%	46%
AE	**15.24**	15.48	15.90	15.76	15.33	15.52	15.55	15.81	**15.23**	15.44
RMSE	**0.099**	**0.100**	0.114	0.113	0.111	0.109	0.115	0.114	0.113	0.111
QPE_τ	0.184	**0.181**	0.185	0.186	0.185	0.186	0.187	0.185	0.184	0.183
$\tau = 0.95$										
FP	2.42	2.22	1.23	1.12	**1.10**	1.28	1.13	1.20	1.27	1.14
P1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
P2	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
AE	15.30	**15.11**	15.64	15.53	15.63	15.52	15.65	15.53	15.64	15.51
RMSE	**0.564**	0.566	0.641	0.638	0.620	0.621	0.644	0.643	0.617	0.623
QPE_τ	**0.082**	0.085	0.092	0.093	0.089	0.090	0.088	0.090	0.091	0.088

The six statistics are calculated for the GPQR approach with all suggested group penalties. *Subscripts 1 and 2 indicate that the GPQR is fitted using the pseudo check function

The best results for each statistic are highlighted in bold font

Table 3 Simulation results of FP, P1, P2, AE, RMSE and QPE_τ for scenario 3, based on 100 replications(5) and (6), respectively

Stats	$GLasso_{1^*}$	$GLasso_2$	$GMCP_1$	$GMCP_2$	$GSCAD_1$	$GSCAD_2$	$GLLA$ MCP_1	$GLLA$ MCP_2	$GLLA$ $SCAD_1$	$GLLA$ $SCAD_2$
$\tau = 0.50$										
FP	4.64	3.92	0.79	0.74	0.88	0.63	0.72	**0.69**	0.75	0.71
P1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
P2	58%	67%	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**
AE	14.96	14.77	15.22	14.33	15.21	**14.30**	14.35	14.33	14.32	14.32
RMSE	**0.404**	0.410	0.430	0.428	0.535	0.541	0.439	0.435	0.529	0.530
QPE_τ	0.604	0.625	**0.542**	0.549	0.562	0.566	0.550	0.545	0.571	0.569
$\tau = 0.75$										
FP	14.6	16.1	2.05	2.12	1.41	**1.37**	1.99	1.87	1.44	1.67
P1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
P2	70%	73%	17%	38%	30%	57%	20%	53%	27%	63%
AE	**22.9**	23.1	24.1	24.0	24.1	24.1	24.1	24.0	24.0	24.1
RMSE	1.121	1.117	0.723	0.727	**0.681**	0.687	0.724	0.732	0.684	0.678
QPE_τ	0.509	0.512	0.454	**0.451**	0.460	0.465	0.517	0.522	0.466	0.461
$\tau = 0.95$										
FP	12.23	11.65	2.27	2.32	1.43	1.51	2.36	2.52	**1.37**	1.41
P1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
P2	93%	90%	91%	**97%**	93%	94%	83%	**97%**	94%	93%
AE	22.3	**22.0**	24.2	24.3	24.3	24.3	24.2	24.3	24.3	24.2
RMSE	10.95	11.17	**4.342**	4.523	4.721	4.635	4.882	4.404	5.14	5.22
QPE_τ	0.247	0.251	0.199	0.209	0.202	0.207	**0.197**	0.200	0.203	0.201

The six statistics are calculated for the GPQR approach with all suggested group penalties. * Subscripts 1 and 2 indicate that the GPQR is fitted using the pseudo check functions

The best results for each statistic are highlighted in bold font

$$Y = \sum_{j=1}^{q} \left(\frac{2}{3} X_j - X_j^2 + \frac{1}{3} X_j^3 \right) \beta_j + \epsilon,$$

where $\beta_j = (-1)^j \exp\{-(2j-1)/20\}$, the error term ϵ is generated from $N(0, \sigma^2)$, and σ^2 is chosen so that the signal-to-noise ratio (SNR) is 3 (i.e. $SNR = \|\boldsymbol{X\beta}\|_2 / \sqrt{n}\sigma$). We considered $\{X_j, X_j^2, X_j^3\}$ as a group when fitting all the proposed models. Thus, the final design matrix of the predictors has $p = 3q$ columns. In this scenario, we set two values of $q = 1000, 3000$, and we fixed $n = 100$. For all group penalties, we fitted three conditional quantile regression models, with $\tau = 0.50, 0.75, 0.95$.

For all algorithms, we calculated the number of coefficients among p coefficients that violated the KKT condition check at each λ value. This number is then averaged over the 100λ values. We repeated this process 10 times on 10 independent datasets. Table 4 reports the results that are averaged over 100λ values and averaged over the 10 independent runs.

4: Table 4 shows that all exact group-penalized methods have a zero-violation count, except the GSCAD which has 1 violation. The GPQR with the GLLA penalty also has small violation counts. Thus, one can argue that all the proposed approaches are accurate algorithms that pass the KKT checks without severe violation.

Table 4 The reported numbers are the average number of coefficients among the p coefficients that violated the KKT condition check using the GPQR with the Glasso, GMCP, GSCAD, and GLLA penalties

Method	$(n,p) = (100, 3000)$			$(n,p) = (100, 9000)$		
	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
GLasso$_1$	0	0	0	0	0	0
GLasso$_2$	0	0	0	0	0	0
GMCP$_1$	0	0	0	0	0	0
GMCP$_2$	0	0	0	0	0	0
GSCAD$_1$	0	0	1	0	0	0
GSCAD$_2$	0	0	1	0	0	0
McpGLLA$_1$	8	4	2	10	5	3
McpGLLA$_2$	8	4	2	10	5	3
ScadGLLA$_1$	3	1	1	3	2	1
ScadGLLA$_2$	3	1	1	3	2	1

Subscripts 1 and 2 indicate that the GPQR is fitted using the check functions (5) and (6) respectively. Results are averaged over the λ sequence of 100 values and averaged over 10 independent runs

4 Real data

4.1 Gene-based analysis of Alzheimer's disease neuroimaging initiative (ADNI) data

The data used in the preparation of this article were obtained from the ADNI database (https://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by principal investigator Michael W. Weiner, MD. The ADNI's primary goal is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

It is known that the pathogenic relevance in AD presents a decrease of the biomarker cerebrospinal fluid amyloid-β42 (CSF $A\beta$42) levels and an increase in the biomarker cerebrospinal fluid total tau (CSF T-tau) levels (Li et al. 2017). Moreover, it is known that individuals with a family history of AD have a higher risk for AD than those without a family history. This reveals that underlying genetic factors may play a key role in AD (Hohman et al. 2014). In fact, in several GWAS, the two biomarkers CSF $A\beta$42 and CSF T-tau have been reported to be associated with several SNPs falling within or near the genes *APOE*, *TOMM40* and *APOC1*, located in Chromosome 19 (Kim et al. 2011).

To illustrate the use of our framework in GWAS, we conducted a gene-based association study using the GPQR approaches in the ADNI cohort. More precisely, we considered the CSF T-tau/$A\beta$42 ratio as an AD imaging quantitative trait (response) on 442 subjects. As predictors, we used single-nucleotide polymorphisms (SNPs) falling within a genomic region of 629 kilobase pairs located around the three genes of interest (*APOE*, *TOMM40,* and *APOC1*). This region results in $K =$ 17 genes/groups with observed genotypes of 1162 SNPs of the ADNI samples. We then assigned the SNPs to genes based on their base-pair coordinates. We used the R package `biomaRt`(Durinck et al. 2009) to extract the genes' start-end genomic coordinates.

This analysis aims to replicate/select the three genes of interest as associated with the response variable, CSF T-tau/$A\beta$42, using the GPQR framework and compare its performance with that of group penalized LS methods. In our analyses, all models were fitted with 17 penalized genes/groups, and we adjusted for sex, age, and diagnostic without penalization because such covariates are known to be potential confounding factors for AD.

We conducted two analyses for this data. First, we fitted the GPQR and group penalized LS methods for all 442 analyzed subjects with five-fold CV to obtain a better model estimation. In the second analysis, we aimed to evaluate the prediction performance of the methods. Thus, we randomly divided the data into a training sample of two-thirds of the observations and the remainder making up the test data. The model is fitted to the training data and the prediction errors are calculated on the test data. The tuning parameters were selected by five-fold CV on the training data. The whole procedure was repeated 100 times and we reported the empirical distribution of the prediction-errors and model-size statistics using box plots, for all methods. The model-size statistic is defined as the number of significant genes. Figures 3, 4, and Table 5 of the main manuscript summarize the results from the gene-based association study of the ADNI cohort in the center of the response variable (i.e. mean and median).

In Fig. 3, the LS methods tend to select the null model, whereas the median regressions (QR with $\tau = 0.5$) select a model with a moderate number of significant

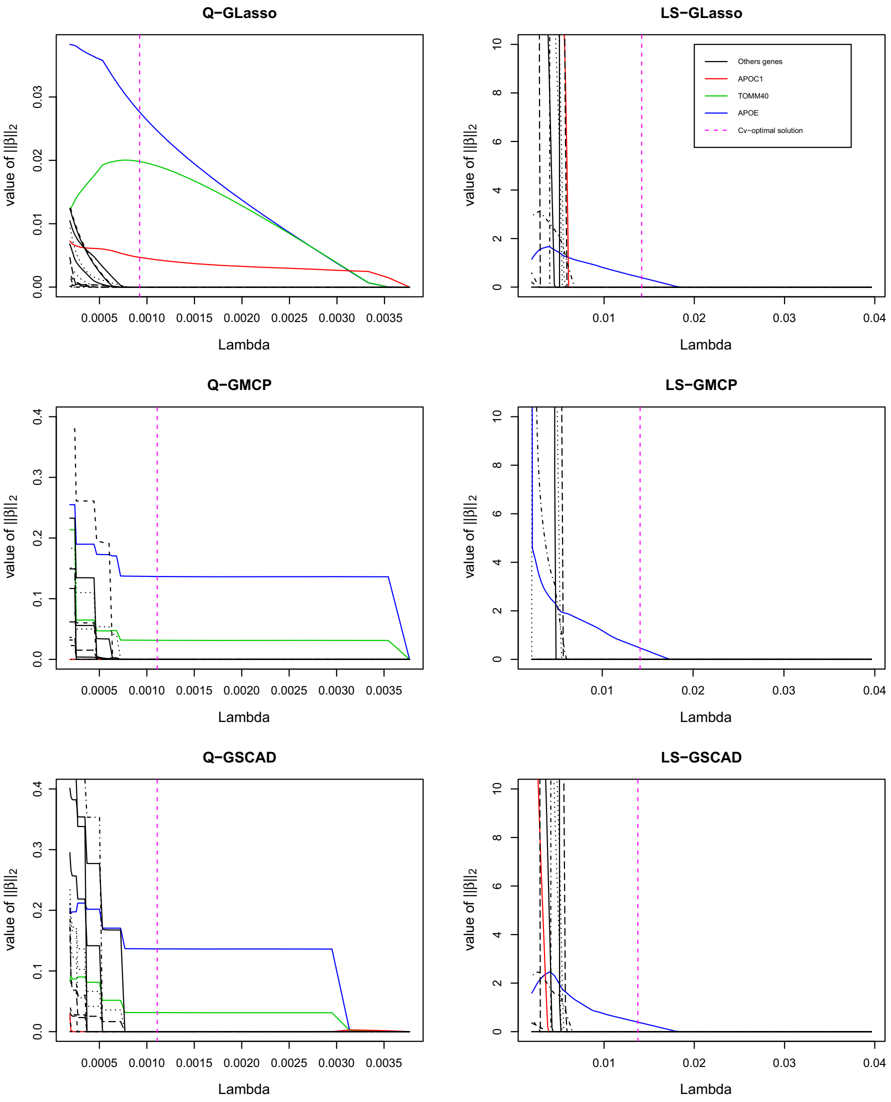

Fig. 3 On the left from top to bottom, the L2-norm of the coefficient paths of the Q-GLasso, Q-GMCP, and Q-GSCAD, respectively, with $\tau = 0.5$, are shown as a function of the tuning parameter λ. On the right from top to bottom are the coefficient paths of the same group methods with LS

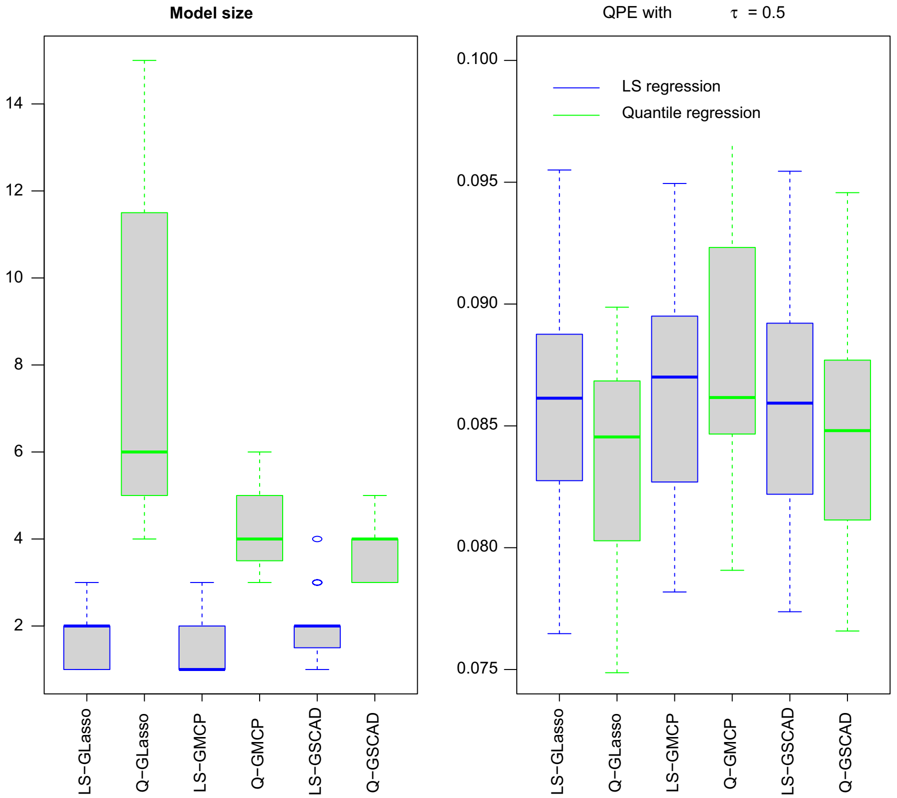

Fig. 4 Comparison of the number of selected genes (Model Size) and the prediction accuracy (QPE_τ) based on 100 replications, for the ADNI data. The group quantile methods are fitted with $\tau = 0.5$

Table 5 The number of times (in %) the genes *APOE, TOMM40,* and *APOC1* are selected based on 100 replications, for the ADNI data

Genes	LS-GLasso	Q-GLasso	LS-GMCP	Q-GMCP	LS-GSCAD	Q-GSCAD
APOC1	8.0	81.6	3.5	23.1	8.0	20.4
TOMM40	0.0	26.5	0.0	0.6	0.0	0.0
APOE	43.4	81.8	19.0	34.7	45.1	20.8

The group quantile methods are fitted with $\tau = 0.5$

genes. Interestingly, at least two of the three genes of interest are selected as active groups by the GPQR using the five-fold CV criterion (pink vertical line). This is also in agreement with the results of Fig. 4 (left panel) which shows the distribution of the model-size statistic for the 100 replications of the second analysis. In Table 5, when comparing the methods based on the selection of the genes of interest (*APOE*, *TOMM40,* and *APOC1*), we notice that the GPQR with the Glasso penalty (Q-

Glasso) is better than the OLS with the Glasso penalty (LS-Glasso). In fact, the proportions of the three genes detected by the Q-Glasso are significantly larger than the LS-Glasso.

By contrast, the right panel of Fig. 4 shows an improvement in predictive performance when using the QR approaches. This is also in accordance with the results reported in Fig. 3 and Table 5.

We implemented further analyses of the ADNI data to investigate the effects of the important three genes in the lower and upper tails of the conditional distribution of the response variable. Thus, the GPQR model was fitted for four additional locations, $\tau \in \{0.1, 0.25, 0.75, 0.9\}$.

Figure S.2 shows $\|\hat{\boldsymbol{\beta}}_k\|_2$ for the three important genes in the ordinate axis as a function of a grid of values of $\tau \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. For each value of τ, we used the five-fold CV procedure to obtain the optimal λ_τ; Fig. S.2 reports the GPQR solution with optimal λ_τ for all 442 subjects of the ADNI cohort. Although one would expect that the genes' effects could be more important for subjects with higher levels of the response variable (i.e. higher quantiles), the results in Fig. S.2 show no significant evidence of this expectation. This might be explained by the presence of both relevant and noisy SNPs within the same gene, which can add some estimation instability to the overall gene effect using the GPQR. Sparse-group selection methods, which achieve both group selection and single-variable selection within each group might be suitable in such situations (Simon et al. 2013; Friedman et al. 2010).

Figure S.3 highlights the results of the L2-norm of the coefficient paths of the Q-Glasso, Q-GMCP, and Q-GSCAD, respectively, as a function of the tuning parameter λ, for $\tau = 0.25$ and 0.75. It shows similar patterns to those in the analysis of the GPQR with the Glasso and $\tau = 0.5$. However, the GPQR with the GMCP (or GSCAD) behaves differently for $\tau = 0.25$ and $\tau = 0.75$. In fact, fitting the GPQR with group MCP/SCAD detects *APOE* and *APOC1* for $\tau = 0.75$; but, when the 0.25-th quantile model is fitted, it selects *APOE* and *TOMM40*.

For more investigation, we also analyzed the 0.25-th and 0.75-th quantiles, similar to the second analysis of the median/center regression (i.e. we conducted 0.25-th and 0.75-th quantile regression models for 100 random training/test replications of the ADNI cohort). Table S.1 of the Supplementary materials summarizes the results for this analysis.

The results are based on 100 random training/test data replications of the ADNI cohort. Each replication consists of a random split of the whole cohort dataset to training (67% observations) and test (33% observations) datasets. The model is fitted to the training data to choose the optimal solution and the tuning parameter using five-fold CV. Then, the prediction performance is evaluated in the test data. Table S.1 outlines the average, over 100 replications, of the following three statistics: (1) the number of times (in proportion) the three genes of interest (*APOE*, *TOMM40,* and *APOC1*) were selected, (2) the quantile-based error prediction (QPE_τ), and (3) the model size (Size) statistic.

Table S.1 shows that the GPQR behaves relatively differently when looking for the effects of the genes in the different locations of the response conditional-distribution, particularly for the model-size statistic. This table shows that the proportions of the three genes detected for $\tau = 0.25$ are larger than for $\tau = 0.75$. Table S.1 also shows inconsistency in the results of the three penalties for the same location, except for the QPE_τ statistic, which is stable across different specifications. This might be explained, on one hand, by the known sensitivity of the lasso-type penalized regression models to the five-fold assignment used in the CV procedure, (Roberts and Nowak 2014). On the other hand, a good tuning parameter choice depends on the unknown parameter σ^2 which is the homogeneous noise variance in linear models (Bickel et al. 2009). For the ADNI data, more knowledge about the standard deviation is necessary and this needs more data investigation. The Discussion section emphasizes this issue and provides tentative solutions.

4.2 Gene-based analysis of the DNA methylation data near the *BLK* gene

This section illustrates the GPQR approach performance for binary classification using DNA methylation around the *BLK* gene, located in chromosome 8, to detect differentially methylated regions (DMRs). DMRs refer to genomic regions with significantly different methylation levels between two groups of samples (e.g.: case-controls). The data consists of methylation levels of 5986 cytosine-guanine dinucleotides (i.e. CpG sites) within a genomic region of 2 million base pairs (i.e. 2 Mb pairs located in Chromosome 8, ranging between positions Chr8-10321522 and Chr8-12391296). The methylation levels in these CpG sites (predictors) are measured in 40 samples using bisulfite sequencing (Lakhal-Chaieb et al. 2017). Each sample corresponds to one of three cell types: B cells (8 samples), T cells (19 samples), and Monocytes (13 samples). These samples are derived from whole blood collected from a cohort of healthy individuals from Sweden. This genomic region is known to be hypomethylated near the *BLK* gene in B-cells, compared to other cell types (Hertz et al. 1999). We first coded the cell types as $y = \{0, 1\}$ variable, with $y = 1$ corresponding to B-cells and $y = 0$ corresponding to T- and Monocyte-cell types. To build groups of predictors (CpGs sites), we proceeded in a similar way to Sect. 4.1. That is, we extracted the start-end genomic positions of all genes belonging to the 2Mb region. $K = 36$ genes fall within this region. Then, we used prior information about the genomic position of each CpG site and assigned each CpG to a corresponding gene based on its base pair coordinate. More precisely, if the genomic position of a CpG site is between the start and end positions of a gene, we considered that the CpG belongs to this gene/group. The CpG assignment procedure is implemented in the *biomaRt* R package. In total, 4427 of all the 5986 CpG sites spread over the 36 genes. The size of the studied groups ranges between 1 and 756, with 398 CpG sites falling between the start-end coordinates of the *BLK* gene.

The $\{0, 1\}$ response variable is then fitted by the group penalized LS and GPQR methods with $\tau = 0.5$. Given the binary nature of the response variable, we also compared the proposed methods with support vector machine (SVM) and logistic regression with a group lasso penalty. Both methods are implemented in the *gglasso* R package (Yang and Zou 2013).

This analysis aims to test the performance of our methods in detecting the group of CpG sites belonging to the *BLK* gene as a DMR for the $0-1$ response, and to test the power of the GPQR in classification. The classification function is $I(\text{fitted value} > 0.5)$, where $I(A)$ is the indicator function which equals 1 if A is true and 0 if A is false.

In Fig. 5, the x and y axes correspond respectively to the genomic position, say t_j, of the j-th CpG site and the coefficient value $(\hat{\beta}_j)_{1 \leq j \leq 4427}$ of the optimal solution that is obtained using five-fold CV. More precisely, each blue dot point in Fig.5 represents a pair $(t_j, \hat{\beta}_j)$, for $j = 1, \ldots, 4427$. As we can see, the region around 11.3 Mb with size 150kb (i.e., the region delimited by the two vertical lines) is detected/selected by the quantile, SVM, and logistic regression methods, but not with the LS approach. This region is known to be the DMR between the DNA methylation profiles of the B-cells and T/Mono cells (Turgeon et al. 2016). These results are also in agreement with Lakhal-Chaieb et al. (2017)'s analysis.

In a second analysis of this DNA methylation data, we randomly divided the data into a training sample of 30 observations with the remainder making up the test data. The model is fitted to the training data and the misclassification error rate (MER) is calculated on test data. The MER is defined as the ratio of the number of misclassified observations to the total number of observations. The tuning parameters are selected by five-fold CV on the training data, and τ is fixed to 0.5 for all our algorithms in this analysis. The whole procedure is repeated 100 times. The results of this analysis are shown in Figs. 6 and 7.

In Fig. 6, the y axis represents the rate of selection of each gene, over 100 replications, of the methylation data analysis. Each segment represents a gene, with large segments corresponding to genes containing a high number of CpG sites (i.e. large genes), and vice versa. The DMR region is always selected by the quantile regression and SVM methods; it is often selected by logistic regression, but the region is never selected by the LS methods. Furthermore, our proposed quantile approaches and SVM outperform the LS approaches and the logistic regression in terms of classification prediction accuracy. This is illustrated by the results of the MER statistic in Fig. 7.

5 Discussion

In this work, we have proposed a unified and computationally-efficient block descent algorithm for solving the group penalized quantile regression in high-dimensional settings. The framework, called GPQR, fits quantile regression with the most appealing group penalties, namely, the group lasso penalty, the group non-convex penalties (SCAD and MCP) and their local approximations. The GPQR allows for the selection of important (heterogeneous) groups of predictors and provides estimates of their effects on the response simultaneously.

We provided a detailed theoretical justification of the linear convergence rate property of the GPQR with group lasso penalty. Moreover, simulation studies have confirmed that the quantile regression performs better for group variable-selection

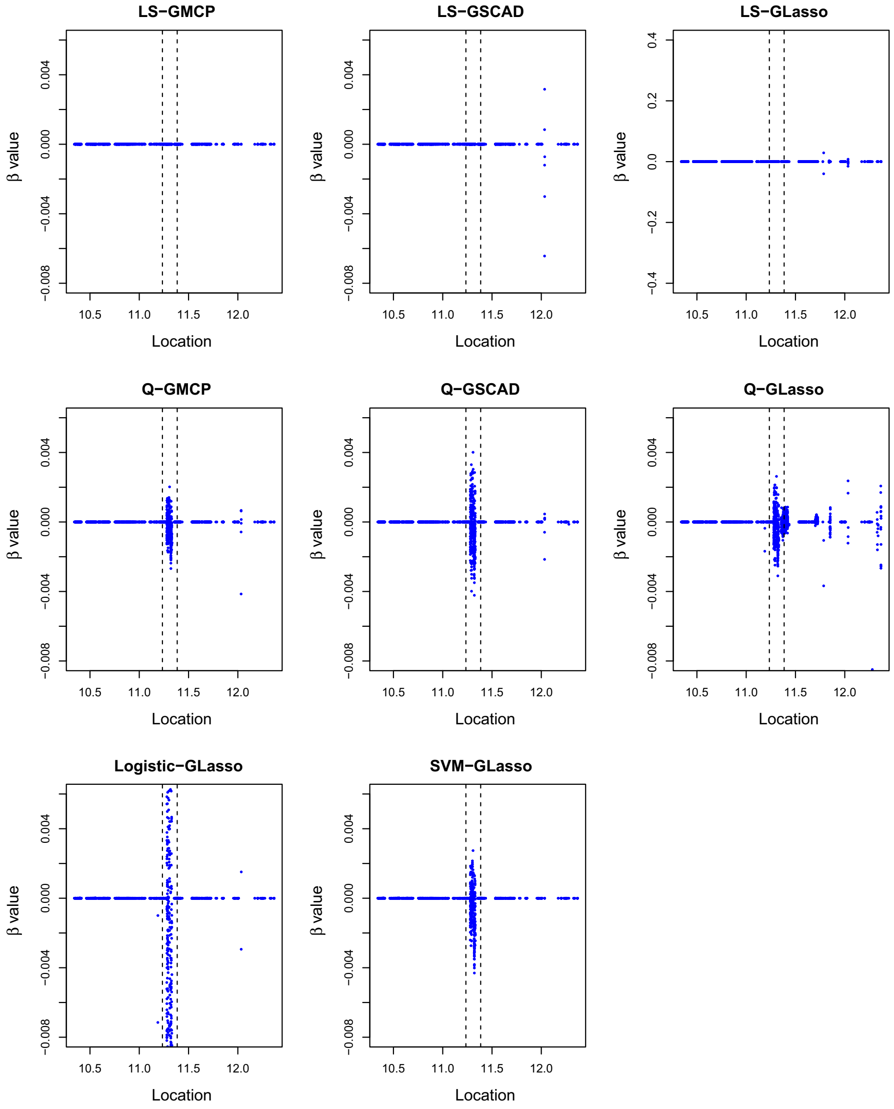

Fig. 5 At the top from left to right, the optimal value (five-fold CV) for the regression coefficients of the LS-methods with the three group penalties (GMCP, GSCAD, and Glasso) are shown as a function of the genomic position. The middle from left to right shows the coefficient values of the same group penalties for the quantile regression, with $\tau = 0.5$. The bottom row shows the coefficient value of the SVM and logistic regression with the Glasso penalty. The x and y axes correspond respectively to the genomic position, t_j, of the j-th CpG site and the coefficient value $(\hat{\beta}_j)_{1 \leq j \leq 4427}$ of the optimal solution

than single-variable quantile regression methods and group-variable selection least-squares approaches in terms of prediction accuracy, variable selection, and detection of heteroscedasticity.

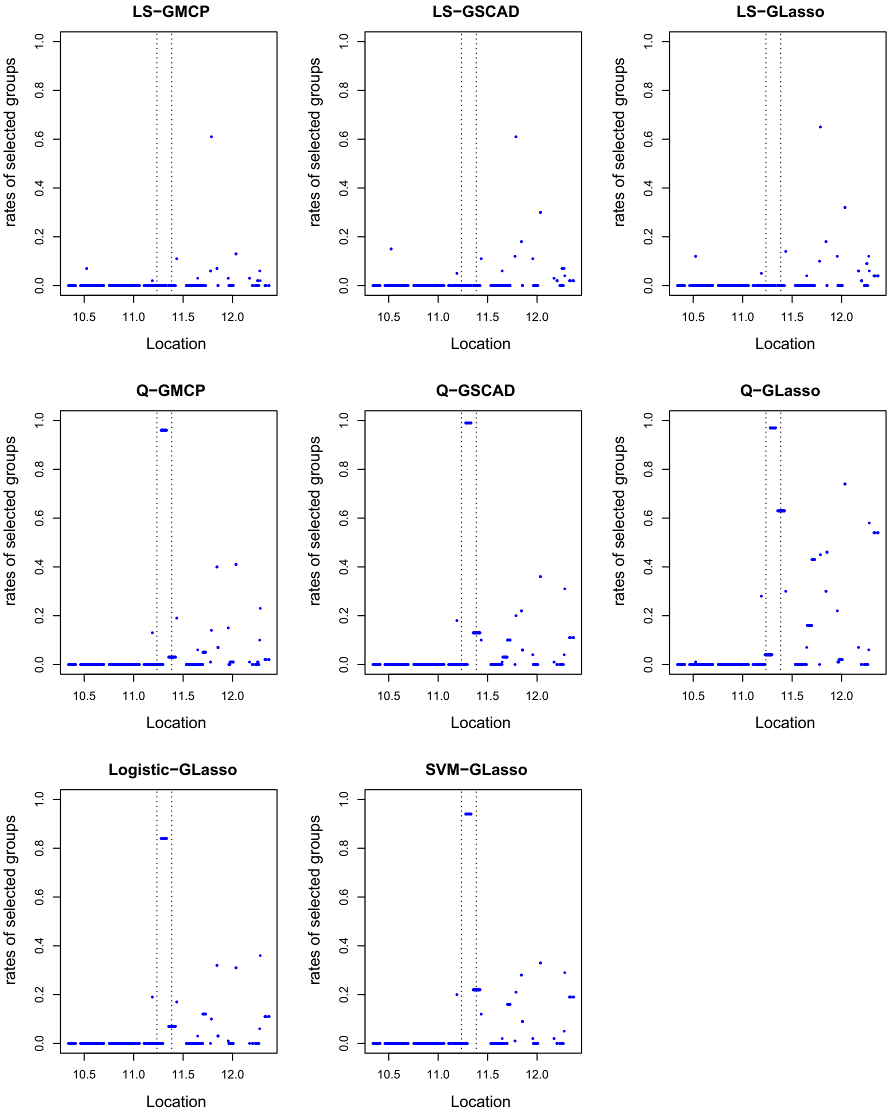

Fig. 6 Comparison of the proportion of selected genes for the DNA methylation data. At the top from left to right, the proportion of LS-GMCP, LS-GSCAD, and LS-Glasso are shown as a function of the genomic position. The middle from left to right shows the proportion of the same group penalties for the quantile regression, with $\tau = 0.5$. The bottom row shows the proportion of the SVM and logistic regression with the Glasso penalty. The x and y axes correspond respectively to the genomic position, t_j, of the j-th CpG site and the proportion of non-zero $(\hat{\beta}_j)_{1 \leq j \leq 4427}$

Although the GPQR demonstrated its utility in selecting relevant genes in the gene-based selection analysis of the ADNI cohort, the results also showed some inconsistency across the three penalties. As we stated in Sect. 4, this might be a result of the method sensitivity to the fold assignment used in the CV procedure,

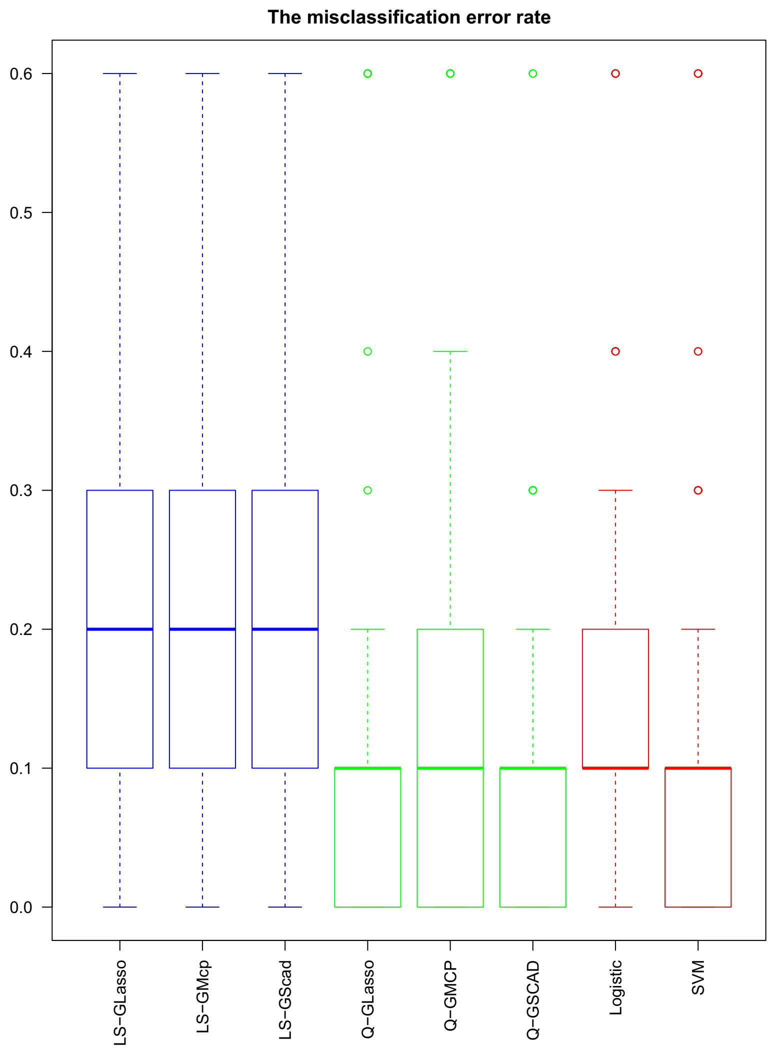

Fig. 7 Comparison of the MER for the DNA methylation data. The MER of the GPQR, LS-methods, and logistic/SVM are plotted in blue, green and red, respectively

(Roberts and Nowak 2014). By contrast, a good tuning parameter choice depends on the unknown parameter σ^2 which is the homogeneous noise variance in linear models (Bickel et al. 2009). For real data situations, knowledge of the standard deviation is lacking and needs more data investigation; yet error variance homogeneity might be a strong assumption for real data applications. In the literature, pivotal penalized methods have been developed to alleviate this problem (Belloni et al. 2011b) (i.e. pivotal in a sense that the method neither relies on knowledge of the standard deviation σ nor does it need to pre-estimate σ). Adapting the GPQR to a pivotal group-variable selection approach within our GPQR framework might be a valuable research avenue. In summary, for real data analysis, our guideline for users is to use the GPQR with the Glasso penalty if the global aim of the analysis is prediction accuracy and to use the GPQR with the GSCAD or GMCP penalties if selection consistency and sparsity are the primary goals.

Like standard QR models, the GPQR framework might be susceptible to the well-known crossing-quantile issue. In high-dimensional settings, the crossing-quantile issue might be more persistent because of the regularized estimation of the quantile curves. In fact, because GPQR curves are estimated individually, the monotonicity of the curves might be violated for some empirical data. Moreover, penalization shrinkage levels may be different for different locations, which renders the crossing even more frequent in high-dimensional settings. To circumvent the crossing-quantile issue, in low-dimensional settings, several authors suggest simultaneous constrained estimation of a sequence of quantile curves. For instance, Koenker (1984) assumed the equality-of-slopes condition (i.e. quantile planes are parallel). Liu and Wu (2009) proposed a sequential procedure for estimation of the quantile that guarantees non-crossing, by constraining the current curve not to cross the previous curve. Bondell et al. (2010) added constraints to the quantile regression optimization problem to ensure non-crossing quantile hyperplanes for any given sample. The crossing-quantiles question for penalized QR and/or high-dimensional settings has seen limited consideration in the literature. Combining high-dimensional and crossing problems might be an interesting avenue of research in QR.

Finally, high-throughput time-varying omics data are becoming available in many genetic studies. Performing predictions and modelling this data is key for understanding the complexity of human health, disease susceptibility, and causations. An interesting direction to pursue would be to study penalized QR in high-dimensional longitudinal/panel data. In low-dimensional settings (i.e., $n > p$), the penalized QR fixed-effect model proposed in Koenker (2004) has been proved very useful for capturing both unobserved individual-specific heterogeneity (i.e. within-subject variation) and effects of heterogeneous predictors in the presence of time-varying data. The extension of our framework to QR fixed-effect models in the presence of high-dimensional longitudinal data could be an interesting avenue to explore.

6 Software

Algorithms 1–3 are implemented in an `R` package, `GPQR`, which is publicly available in `GitHub` (https://github.com/KarimOualkach/GPER.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10260-021-00580-8.

Acknowledgements This work is supported by the Natural Sciences and Engineering Research Council of Canada through an individual discovery research grant to Karim Oualkacha and by the Fonds de recherche du Québec-Santé through individual Grant #31110 to Karim Oualkacha. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense Award Number W81XWH-12-2-0012). The ADNI is funded by the National Institute on Aging, the National Institute of

Biomedical Imaging and Bioengineering, and through generous contributions from: the AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institute of Health Research provides funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Alhamzawi R, Yu K, Benoit DF (2012) Bayesian adaptive lasso quantile regression. Stat Modell 12(3):279–297

Aravkin AY, Kambadur A, Lozano AC, Luss R (2014) Sparse quantile huber regression for efficient and robust estimation. arXiv preprint arXiv:1402.4624

Belloni A, Chernozhukov V et al (2011) l1-penalized quantile regression in high-dimensional sparse models. Ann Stat 39(1):82–130

Belloni A, Chernozhukov V, Wang L (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika 98(4):791–806

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

Bondell HD, Reich BJ, Wang H (2010) Noncrossing quantile regression curve estimation. Biometrika 97(4):825–838

Breheny P (2015) grpreg: regularization paths for regression models with grouped covariates. R Package Version 2:1–8

Breheny P, Huang J (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann Appl Stat 5(1):232

Breheny P, Huang J (2015) Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat Comput 25(2):173–187

Briollais L, Durrieu G (2014) Application of quantile regression to recent genetic and-omic studies. Hum Genet 133(8):951–966

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

Durinck S, Spellman PT, Birney E, Huber W (2009) Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nat Protoc 4(8):1184

Efron B, Hastie T, Tibshirani R (2007) Discussion: the dantzig selector: statistical estimation when p is much larger than n. Ann Stat 35(6):2358–2364

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

Fan J, Fan Y, Barut E (2014) Adaptive robust variable selection. Ann Stat 42(1):324

Fan J, Xue L, Zou H (2014) Strong oracle optimality of folded concave penalized estimation. Ann Stat 42(3):819

Fenske N, Kneib T, Hothorn T (2011) Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. J Am Stat Assoc 106(494):494–510

Friedman J, Hastie T, Tibshirani R (2010) A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736

Gu Y, Zou H et al (2016) High-dimensional generalizations of asymmetric least squares regression and their applications. Ann Stat 44(6):2661–2694

Hashem H, Vinciotti V, Alhamzawi R, Yu K (2016) Quantile regression with group lasso for classification. Adv Data Anal Classif 10(3):375–390

Hertz JM, Schell G, Doerfler W (1999) Factors affecting de novo methylation of foreign DNA in mouse embryonic stem cells. J Biol Chem 274(34):24232–24240

Hofner B, Mayr A, Robinzonov N, Schmid M (2014) Model-based boosting in R: a hands-on tutorial using the R package mboost. Comput Stat 29(1–2):3–35

Hohman TJ, Koran MEI, Thornton-Wells TA (2014) Genetic modification of the relationship between phosphorylated tau and neurodegeneration. Alzheimer's & dementia J Alzheimer's Assoc 10(6):637–645

Hunter DR, Lange K (2000) Quantile regression via an MM algorithm. J Comput Gr Stat 9(1):60–77

Hunter DR, Lange K (2004) A tutorial on MM algorithms. Am Stat 58(1):30–37

Jennings L, Wong K, Teo K (1996) Optimal control computation to account for eccentric movement. ANZIAM J 38(2):182–193

Ji Y, Lin N, Zhang B (2012) Model selection in binary and tobit quantile regression using the Gibbs sampler. Comput Stat Data Anal 56(4):827–839

Juban R, Ohlsson H, Maasoumy M, Poirier L, Kolter JZ (2016) A multiple quantile regression approach to the wind, solar, and price tracks of gefcom2014. Int J Forecast 32(3):1094–1102

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

Kim S, Swaminathan S, Shen L, Risacher S, Nho K, Foroud T, Shaw L, Trojanowski J, Potkin S, Huentelman M et al (2011) Genome-wide association study of CSF biomarkers aβ1-42, t-tau, and p-tau181p in the ADNI cohort. Neurology 76(1):69–79

Koenker R (1984) A note on l-estimates for linear models. Stat Prob Lett 2(6):323–325

Koenker R (2004) Quantile regression for longitudinal data. J Multivar Anal 91(1):74–89

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

Koenker R, Hallock KF (2001) Quantile regression. J Econ Perspect 15(4):143–156

Kozumi H, Kobayashi G (2011) Gibbs sampling methods for Bayesian quantile regression. J Stat Comput Simul 81(11):1565–1578

Lakhal-Chaieb L, Greenwood CM, Ouhourane M, Zhao K, Abdous B, Oualkacha K (2017) A smoothed EM-algorithm for DNA methylation profiles from sequencing-based methods in cell lines or for a single cell type. Stat Appl Genet Mol Biol 16(5–6):333–347

Lange K, Papp JC, Sinsheimer JS, Sobel EM (2014) Next-generation statistical genetics: modeling, penalization, and optimization in high-dimensional data. Annu Rev Stat Appl 1(1):279–300

Li Y, Zhu J (2008) L 1-norm quantile regression. J Comput Gr Stat 17(1):163–185

Li J, Zhang Q, Chen F, Meng X, Liu W, Chen D, Yan J, Kim S, Wang L, Feng W et al (2017) Genome-wide association and interaction studies of CSF t-tau/aβ42 ratio in ADNI cohort. Neurobiol Aging 57:247-e1

Liu Y, Wu Y (2009) Stepwise multiple quantile regression estimation using non-crossing constraints. Stat Interface 2(3):299–310

Mayr A, Binder H, Gefeller O, Schmid M (2014) The evolution of boosting algorithms-from machine learning to statistical modelling. arXiv preprint arXiv:1403.1452

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

Mkhadri A, Ouhourane M (2013) An extended variable inclusion and shrinkage algorithm for correlated variables. Comput Stat Data Anal 57(1):631–644

Mkhadri A, Ouhourane M, Oualkacha K (2017) A coordinate descent algorithm for computing penalized smooth quantile regression. Stat Comput 27(4):865–883

Ogutu JO, Piepho H-P (2014) Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD. BMC Proc 8(Suppl 5):S7

Oh H-S, Lee TC, Nychka DW (2011) Fast nonparametric quantile regression with arbitrary smoothing methods. J Comput Gr Stat 20(2):510–526

Peng B, Wang L (2015) An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. J Comput Gr Stat 24(3):676–694

Roberts S, Nowak G (2014) Stabilizing the lasso against cross-validation variability. Comput Stat Data Anal 70:198–211

Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. J Comput Gr Stat 22(2):231–245

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

Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, Tibshirani RJ (2012) Strong rules for discarding predictors in lasso-type problems. J R Stat Soc Ser B (Stat Methodol) 74(2):245–266

Turgeon M, Oualkacha K, Ciampi A, Miftah H, Dehghan G, Zanke BW, Benedet AL, Rosa-Neto P, Greenwood CM, Labbe A; Alzheimer's Disease Neuroimaging Initiative (2018) Principal component of explained variance: an efficient and optimal data dimension reduction framework for association studies. Stat Methods Med Res 27(5):1331–1350. https://doi.org/10.1177/0962280216660128

Waldmann E, Kneib T, Yue YR, Lang S, Flexeder C (2013) Bayesian semiparametric additive quantile regression. Stat Modell 13(3):223–252

Wang L (2013) The l1 penalized LAD estimator for high dimensional linear regression. J Multivar Anal 120:135–151

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

Wang H, Lengerich BJ, Aragam B, Xing EP (2019) Precision lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. Bioinformatics 35(7):1181–1187

Wei F, Zhu H (2012) Group coordinate descent algorithms for nonconvex penalized regression. Comput Stat Data Anal 56(2):316–326

Wu TT, Lange K et al (2008) Coordinate descent algorithms for lasso penalized regression. Ann Appl Stat 2(1):224–244

Xu QF, Ding XH, Jiang CX, Yu KM, Shi L (2020) An elastic-net penalized expectile regression with applications. J Appl Stat. https://doi.org/10.1080/02664763.2020.1787355

Yang Y, Zou H (2013) An efficient algorithm for computing the HHSVM and its generalizations. J Comput Gr Stat 22(2):396–415

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

Yi C, Huang J (2017) Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. J Comput Gr Stat 26(3):547–557

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Stat Methodol) 68(1):49–67

Zhang C-H et al (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38(2):894–942

Zhao G, Teo KL, Chan K (2005) Estimation of conditional quantiles by a new smoothing approximation of asymmetric loss functions. Stat Comput 15(1):5–11

Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel EM, Lange K (2011) Penalized regression for genome-wide association screening of sequence data. Pac Symp Biocomput 2011:106–117. https://doi.org/10.1142/9789814335058_0012. PMID: 21121038; PMCID: PMC5049883

Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models. Ann Stat 36(4):1509

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Noname manuscript No.
(will be inserted by the editor)

Group Penalized Smooth Quantile Regression

Received: date / Accepted: date

1 Proof of Proposition 1

From Proposition 1 of Mkhadri et al. (2017), we have

$$-\delta\kappa \leq \Psi_\tau(u) - \rho_\tau(u) \leq \delta\kappa \quad \forall u \in \mathbb{R},$$

where the constant $\kappa = sup(\tau, 1-\tau)/2$ or $sup(\tau^2, (1-\tau)^2)/2$. This yields to the following inequalities

$$-\delta\kappa + R(\boldsymbol{\beta}) \leq R_\delta(\boldsymbol{\beta}) \leq \delta\kappa + R(\boldsymbol{\beta}). \tag{A-1}$$

Let $\hat{\boldsymbol{\beta}}$ be the unique minimizer of $R(\boldsymbol{\beta})$ in (1), then we have

$$\begin{aligned}\inf_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) &\leq R(\hat{\boldsymbol{\beta}}(\delta)) \\ &\overset{(a)}{\leq} R_\delta(\hat{\boldsymbol{\beta}}(\delta)) + \delta\kappa \\ &\overset{(b)}{\leq} R_\delta(\hat{\boldsymbol{\beta}}) + \delta\kappa \\ &\overset{(c)}{\leq} R(\hat{\boldsymbol{\beta}}) + \delta\kappa + \delta\kappa \\ &\leq \inf_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) + 2\delta\kappa.\end{aligned}$$

Inequality (a) is due to the first inequality in (A-1), inequality (b) is due to $\hat{\boldsymbol{\beta}}(\delta)$ is the minimizer of $R_\delta(\boldsymbol{\beta})$ and inequality (b) is due to the second inequality in (A-1). This ends the proof of Proposition 1.

2 Proof of Proposition 2

Proof. Following Mkhadri et al. (2017), we can show that the smooth quantile loss function $\Psi_\tau(.)$ has a Lipschitz continuous derivative $\Psi'_\tau(.)$, i.e.

$$\text{when} \quad \Psi_\tau = \Psi^{(1)}_{\tau,\delta}: \quad |\Psi'_\tau(u) - \Psi'_\tau(v)| \leq \frac{\max(\tau, 1-\tau)}{\delta}|u-v| \quad \forall u, v \in \mathbb{R},$$

$$\text{when} \quad \Psi_\tau = \Psi^{(2)}_{\tau,\delta}: \quad |\Psi'_\tau(u) - \Psi'_\tau(v)| \leq \frac{1}{\delta}|u-v| \quad \forall u, v \in \mathbb{R}.$$

Thus, we have

$$|\Psi'_\tau(u) - \Psi'_\tau(v)| \leq c|u-v| \quad \forall u, v \in \mathbb{R}, \tag{A-2}$$

where $c = \frac{\max(\tau, 1-\tau)}{\delta}$ for $\Psi^{(1)}_{\tau,\delta}$ and $c = \frac{1}{\delta}$ for $\Psi^{(2)}_{\tau,\delta}$.
For $\boldsymbol{\beta}_k$ and $\tilde{\boldsymbol{\beta}}_k$, let $\mathbf{V}_k = \boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k$ and define $g(t) = L(\tilde{\boldsymbol{\beta}}_k + t\mathbf{V}_k, \tilde{\boldsymbol{\beta}}_{-k})$. Thus, we have $g(0) = L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k})$, $g(1) = L(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$.
By the mean value theorem, $\exists a \in (0,1)$ such that

$$g(1) = g(0) + g'(a) = g(0) + g'(0) + (g'(a) - g'(0)). \tag{A-3}$$

Since we have

$$g'(t) = n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_{i,k}^{\top}\mathbf{V}_k \Psi'_\tau(y_i - \boldsymbol{x}_{i,-k}^{\top}\tilde{\boldsymbol{\beta}}_{-k} - \boldsymbol{x}_{i,k}^{\top}\tilde{\boldsymbol{\beta}}_k + t\boldsymbol{x}_{i,k}^{\top}\mathbf{V}_k))$$

it follows that $g'(0) = (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top}\nabla_k L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k})$, and thus, one can write

$$\begin{aligned}
\mid g'(a) - g'(0)| &= |n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_{ik}^{\top}\mathbf{V}_k[\Psi'_\tau(y_i - \boldsymbol{x}_{i,-k}^{\top}\tilde{\boldsymbol{\beta}}_{-k} - \boldsymbol{x}_{ik}^{\top}(\tilde{\boldsymbol{\beta}}_k + a\mathbf{V}_k)) - \Psi'_\tau(y_i - \boldsymbol{x}_{i,-k}^{\top}\tilde{\boldsymbol{\beta}}_{-k} - \boldsymbol{x}_{ik}^{\top}\tilde{\boldsymbol{\beta}}_k)]| \\
&\le n^{-1}\sum_{i=1}^{n} |\boldsymbol{x}_{ik}^{\top}\mathbf{V}_k||\Psi'_\tau(y_i - \boldsymbol{x}_{i,-k}^{\top}\tilde{\boldsymbol{\beta}}_{-k} - \boldsymbol{x}_{ik}^{\top}(\tilde{\boldsymbol{\beta}}_k + a\mathbf{V}_k)) - \Psi'_\tau(y_i - \boldsymbol{x}_{i,-k}^{\top}\tilde{\boldsymbol{\beta}}_{-k} - \boldsymbol{x}_{ik}^{\top}\tilde{\boldsymbol{\beta}}_k)| \\
&\overset{(a)}{\le} n^{-1}\sum_{i=1}^{n} |\boldsymbol{x}_{ik}^{\top}\mathbf{V}_k|c|a\boldsymbol{x}_{ik}^{\top}\mathbf{V}_k| \\
&\le cn^{-1}\sum_{i=1}^{n} \|\boldsymbol{x}_{ik}^{\top}\mathbf{V}_k\|^2 \\
&\le cn^{-1}\mathbf{V}_k^{\top}\boldsymbol{X}_k^{\top}\boldsymbol{X}_k\mathbf{V}_k.
\end{aligned}$$

Inequality (a) is due to equation (A-2).Using the last inequality and (A-3) leads to the following inequality

$$\begin{aligned}
L(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) \le L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top}\nabla_k L(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + \\
cn^{-1}(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top}\boldsymbol{X}_k^{\top}\boldsymbol{X}_k(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k).
\end{aligned}$$

This ends the proof of Proposition 2. □

3 The convergence analysis of Algorithm 2: proof of Theorem 1

Some properties of the smooth quantile loss function, $L(\boldsymbol{\beta}) = n^{-1}\mathbf{1}_n^{\top}\Psi_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})$, are used in the steps of the Theorem's proof; they are given first. The smooth quantile check function, Ψ_τ, can be either $\Psi_{\tau,\delta}^{(1)}$ or $\Psi_{\tau,\delta}^{(2)}$ and $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector of all ones.

Since we have

$$\nabla L(\boldsymbol{\beta}) = -n^{-1}\boldsymbol{X}^{\top}\Psi'_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

then, using (A-2), it follows that

$$\begin{aligned}
\|\nabla L(\boldsymbol{\beta}) - \nabla L(\boldsymbol{\beta}')\| &= n^{-1}\|\boldsymbol{X}^{\top}(\Psi'_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) - \Psi'_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}'))\| \\
&\le n^{-1}\|\mathbf{X}\|\|\Psi'_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) - \Psi'_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}')\| \\
&\le cn^{-1}\|\mathbf{X}\|\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}')\| \\
&\le cn^{-1}\|\mathbf{X}\|^2\|\boldsymbol{\beta} - \boldsymbol{\beta}'\| \\
&\le \gamma\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|, \qquad \forall \boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p,
\end{aligned}$$

where γ is the largest eigenvalue of $cn^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}$, and $c = \frac{\max(\tau,1-\tau)}{\delta}$ for $\Psi_{\tau,\delta}^{(1)}(u)$ and $c = \frac{1}{\delta}$ for $\Psi_{\tau,\delta}^{(2)}(u)$. This implies that the gradient of $L(\cdot)$ is uniformly Lipschitz continuous with Lipschitz constant γ. When restricted to each block, we have

$$\nabla_k L(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^{n}\Psi'_\tau(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\beta})\boldsymbol{x}_{ik} = -n^{-1}\boldsymbol{X}_k^{\top}\Psi'_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}),\ k = 1, \dots, K.$$

Thus, we have

$$\begin{aligned}
\|\nabla_k L(\mathbf{u}_k; \boldsymbol{\beta}_{-k}) - \nabla_k L(\mathbf{v}_k, \boldsymbol{\beta}_{-k})\| &\le n^{-1}c\|\mathbf{X}_k\|^2\|\mathbf{u}_k - \mathbf{v}_k\| \\
&\le \gamma_k\|\mathbf{u}_k - \mathbf{v}_k\|, \qquad \forall \mathbf{u}_k, \mathbf{v}_k \in \mathbb{R}^{p_k}, \forall k \in \{1, \dots, K\},
\end{aligned}$$

where γ_k is the largest eigenvalue of $cn^{-1}\mathbf{X}_k^{\top}\mathbf{X}_k$. This implies that the gradient of $L(\cdot)$ is block-wise uniformly Lipschitz continuous with Lipschitz constant γ_k.

Moreover, for group k, let $u_k(\cdot; \boldsymbol{\beta}_{-k})$ be the quadratic majorization function of $L(., \boldsymbol{\beta}_{-k})$, at $\boldsymbol{\beta}_k$, defined as follows

$$u_k(\mathbf{v}_k; \boldsymbol{\beta}_{-k}) = L(\boldsymbol{\beta}) + \langle \nabla_k L(\boldsymbol{\beta}),\ \mathbf{v}_k - \boldsymbol{\beta}_k\rangle + \frac{\gamma_k}{2}\|\mathbf{v}_k - \boldsymbol{\beta}_k\|^2.$$

Note that we omit the dependency u_k on $\boldsymbol{\beta}_k$ to ease exposition. The function $u_k(\mathbf{v}_k; \boldsymbol{\beta}_{-k})$satisfies the following conditions

1. $u_k(\boldsymbol{\beta}_k; \boldsymbol{\beta}_{-k}) = L(\boldsymbol{\beta})$;
2. $u_k(\mathbf{v}_k; \boldsymbol{\beta}_{-k}) \geq L(\mathbf{v}_k, \boldsymbol{\beta}_{-k})$, for $\mathbf{v}_k \neq \boldsymbol{\beta}_k$;
3. $\nabla u_k(\boldsymbol{\beta}_k; \boldsymbol{\beta}_{-k}) = \nabla_k L(\boldsymbol{\beta}_k, \boldsymbol{\beta}_{-k})$.

We can verify that $u_k(\cdot; \boldsymbol{\beta}_{-k})$ is strongly convex:

$$\begin{aligned} u_k(\mathbf{u}_k; \boldsymbol{\beta}_{-k}) \geq u_k(\mathbf{v}_k; \boldsymbol{\beta}_{-k}) + \left\langle \nabla u_k(\mathbf{v}_k; \boldsymbol{\beta}_{-k}), \mathbf{u}_k - \mathbf{v}_k \right\rangle \\ + \tfrac{\gamma_k}{2} \|\mathbf{u}_k - \mathbf{v}_k\|^2 \qquad \forall \mathbf{u}_k, \mathbf{v}_k \in \mathbb{R}^{p_k}, \forall k. \end{aligned} \tag{A-4}$$

Further, we have

$$\begin{aligned} \|\nabla u_k(\mathbf{v}_k; \boldsymbol{\beta}_{-k}) - \nabla u_k(\mathbf{v}_k; \boldsymbol{\beta}'_{-k})\| &= \|\nabla_k L(\boldsymbol{\beta}) - \nabla_k L(\boldsymbol{\beta}') + \gamma_k(\boldsymbol{\beta}_k - \boldsymbol{\beta}'_k)\| \\ &\leq \|\nabla_k L(\boldsymbol{\beta}) - \nabla_k L(\boldsymbol{\beta}')\| + \gamma_k \|\boldsymbol{\beta}_k - \boldsymbol{\beta}'_k\| \\ &\leq n^{-1} \|\boldsymbol{X}_k^\top (\Psi'_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) - \Psi'_\tau(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}'))\| + \gamma_k \|\boldsymbol{\beta}_k - \boldsymbol{\beta}'_k\| \\ &\overset{(a)}{\leq} n^{-1} c \|\boldsymbol{X}_k\| \|\boldsymbol{X}\| \|(\boldsymbol{\beta} - \boldsymbol{\beta}')\| + \gamma_k \|\boldsymbol{\beta}_k - \boldsymbol{\beta}'_k\| \\ &\leq G_k \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|, \qquad \forall \mathbf{v}_k \in \mathbb{R}^{p_k}, \forall k, \boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^{p+1}, \end{aligned} \tag{A-5}$$

where $G_k = \sqrt{\gamma_k}\sqrt{\gamma} + \gamma_k$. Inequality (a) is due to equation (A-2)
The proof of Theorem 1 relies on the iteration complexity analysis which is given next. This analysis is divided into three parts: the sufficient descent step, the cost-to-go estimate step, and the local error bound step. Similar techniques can be found in Luo and Tseng (1992), Luo and Tseng (1993), Zhang et al. (2013), Sun and Hong (2015) and Hong et al. (2017).
Iteration Complexity Analysis. For ease of exposition, let us rewrite (7) as the following unconstrained optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} Q(\boldsymbol{\beta}) := \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} L(\boldsymbol{\beta}) + \sum_{k=1}^{K} h_k(\boldsymbol{\beta}_k), \tag{A-6}$$

where $L(\boldsymbol{\beta})$ is the smooth quantile loss function which is smooth convex in $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ while $h_k(\boldsymbol{\beta}_k) = w_k \lambda \|\boldsymbol{\beta}_k\|$ is nonsmooth convex in $\boldsymbol{\beta}_k$ for each $k = 1, \ldots, K$. We have the following cyclic block-coordinate update of $\boldsymbol{\beta}_k$ by (11)

$$\boldsymbol{\beta}_k := \mathbf{prox}_{\gamma_k^{-1} h_k}(\boldsymbol{\beta}_k - \gamma_k^{-1} \nabla_k L(\boldsymbol{\beta})).$$

The following notation is convenient for this iteration complexity analysis. Let $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ be a K-block partition of the optimization variable $\boldsymbol{\beta}$ (i.e., $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top \in \mathbb{R}^{p+1}$, with $\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}$ and $\sum_{k=1}^K p_k = p+1$). Also, denote the subvector of $\boldsymbol{\beta}$ with its kth component removed by $\boldsymbol{\beta}_{-k} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_{k-1}^\top, \boldsymbol{\beta}_{k+1}^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top$ and recover $\boldsymbol{\beta}$ from $\boldsymbol{\beta}_{-k}$ by $\boldsymbol{\beta} = (\boldsymbol{\beta}_k^\top, \boldsymbol{\beta}_{-k}^\top)^\top$. Moreover, in the cyclic coordinate descent algorithm, let $\boldsymbol{\beta}^r$ be the update of $\boldsymbol{\beta}$ after the rth cycle, $r \geq 0$. When updating $\boldsymbol{\beta}_k$ in the $(r+1)$th cycle using the proximal operator (i.e. GPQR Algorithm 2), the following notations are also adopted

$$\begin{aligned} &\boldsymbol{B}_k^{r+1} = [(\boldsymbol{\beta}_1^{r+1})^\top, \ldots, (\boldsymbol{\beta}_{k-1}^{r+1})^\top, (\boldsymbol{\beta}_k^{r})^\top, (\boldsymbol{\beta}_{k+1}^{r})^\top, \ldots, (\boldsymbol{\beta}_K^{r})^\top]^\top, \ k = 2, \ldots, K, \\ &\boldsymbol{B}_{-k}^{r+1} = [(\boldsymbol{\beta}_1^{r+1})^\top, \ldots, (\boldsymbol{\beta}_{k-1}^{r+1})^\top, (\boldsymbol{\beta}_{k+1}^{r})^\top, \ldots, (\boldsymbol{\beta}_K^{r})^\top]^\top, \ k = 2, \ldots, K, \\ &\boldsymbol{\beta}_{-k} = [(\boldsymbol{\beta}_1)^\top, \ldots, (\boldsymbol{\beta}_{k-1})^\top, (\boldsymbol{\beta}_{k+1})^\top, \ldots, (\boldsymbol{\beta}_K)^\top]^\top, k = 2, \ldots, K. \end{aligned}$$

By definition we have $\boldsymbol{B}_1^{r+1} := \boldsymbol{\beta}^r$ and $\boldsymbol{B}_{K+1}^{r+1} := \boldsymbol{\beta}^{r+1}$.

*Sufficient Descent.*Consider the proximal gradient method applied to solving the following problem

$$\min_{\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}} Q(\boldsymbol{\beta}_k, \boldsymbol{B}_{-k}^{r+1}) = \min_{\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}} L(\boldsymbol{\beta}_k, \boldsymbol{B}_{-k}^{r+1}) + h_k(\boldsymbol{\beta}_k).$$

By the convexity of $h_k(\cdot)$, there exists $\zeta_k^{r+1} \in \partial h_k(\boldsymbol{\beta}_k^{r+1})$ such that

$$h_k(\boldsymbol{\beta}_k^r) - h_k(\boldsymbol{\beta}_k^{r+1}) \geq \left\langle \zeta_k^{r+1}, \boldsymbol{\beta}_k^r - \boldsymbol{\beta}_k^{r+1} \right\rangle, \ \forall \boldsymbol{\beta}_k^r, \tag{A-7}$$

where ∂h_k is is a sub-gradient of h_k.
Using (A-4) and (A-7), one has

$$\begin{aligned}
&Q(\boldsymbol{\beta}_k^r, \boldsymbol{B}_{-k}^{r+1}) - Q(\boldsymbol{\beta}_k^{r+1}, \boldsymbol{B}_{-k}^{r+1})\\
=&u_k(\boldsymbol{\beta}_k^r; \boldsymbol{B}_{-k}^{r+1}) + h_k(\boldsymbol{\beta}_k^r) - \left(u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}) + h_k(\boldsymbol{\beta}_k^{r+1})\right)\\
\geq&\left\langle \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}),\, \boldsymbol{\beta}_k^r - \boldsymbol{\beta}_k^{r+1}\right\rangle + h_k(\boldsymbol{\beta}_k^r) - h_k(\boldsymbol{\beta}_k^{r+1}) + \frac{\gamma_k}{2}\|\boldsymbol{\beta}_k^r - \boldsymbol{\beta}_k^{r+1}\|^2\\
\geq&\left\langle \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}) + \zeta_k^{r+1},\, \boldsymbol{\beta}_k^r - \boldsymbol{\beta}_k^{r+1}\right\rangle + \frac{\gamma_k}{2}\|\boldsymbol{\beta}_k^r - \boldsymbol{\beta}_k^{r+1}\|^2\\
\overset{(a)}{\geq}&\frac{\gamma_k}{2}\|\boldsymbol{\beta}_k^r - \boldsymbol{\beta}_k^{r+1}\|^2.
\end{aligned}$$

Inequality (a) is due to the optimality condition

$$\left\langle \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}) + \zeta_k^{r+1},\, \boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\right\rangle \leq 0. \tag{A-8}$$

Thus, it follows that

$$Q(\boldsymbol{\beta}^r) - Q(\boldsymbol{\beta}^{r+1}) = \sum_{k=1}^{K}\left[Q(\boldsymbol{\beta}_k^r, \boldsymbol{B}_{-k}^{r+1}) - Q(\boldsymbol{\beta}_k^{r+1}, \boldsymbol{B}_{-k}^{r+1})\right] \geq \frac{\underline{\gamma}}{2}\|\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r+1}\|^2, \tag{A-9}$$

where $\underline{\gamma} = \min_{1\leq k\leq K}\gamma_k$.

*Cost-to-go Estimate.*Let $\mathcal{X}^* = \{\boldsymbol{\beta}^* | Q(\boldsymbol{\beta}^*) = \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})\}$ be the optimal solution set of problem (A-6). Let $\bar{\boldsymbol{\beta}}^r = (\bar{\boldsymbol{\beta}}_1^r, \ldots, \bar{\boldsymbol{\beta}}_K^r) \in \mathcal{X}^*$ be a point in $\mathcal{X}^*$ such that $\mathrm{d}_{\mathcal{X}^*}(\boldsymbol{\beta}^r) = \min_{\boldsymbol{\beta}\in\mathcal{X}^*}\|\boldsymbol{\beta} - \boldsymbol{\beta}^r\| = \|\bar{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^r\|$. We have

$$\begin{aligned}
&\left\langle \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}),\, \boldsymbol{\beta}_k^{r+1} - \bar{\boldsymbol{\beta}}_k^r\right\rangle + \left[h_k(\boldsymbol{\beta}_k^{r+1}) - h_k(\bar{\boldsymbol{\beta}}_k^r)\right]\\
\leq&\left\langle \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}) + \zeta_k^{r+1},\, \boldsymbol{\beta}_k^{r+1} - \bar{\boldsymbol{\beta}}_k^r\right\rangle \qquad\qquad (\text{A-10})\\
\leq&0,
\end{aligned}$$

where the first inequality is due to the inequality (A-7), and the last inequality, we use the optimality conditions in (A-8).
On the other hand, we also have that

$$\begin{aligned}
Q(\boldsymbol{\beta}^{r+1}) - Q(\bar{\boldsymbol{\beta}}^r) &= L(\boldsymbol{\beta}^{r+1}) - L(\bar{\boldsymbol{\beta}}^r) + \sum_{k=1}^{K} h_k(\boldsymbol{\beta}_k^{r+1}) - \sum_{k=1}^{K} h_k(\bar{\boldsymbol{\beta}}_k^r)\\
&\leq \left\langle \nabla L(\boldsymbol{\beta}^{r+1}),\, \boldsymbol{\beta}^{r+1} - \bar{\boldsymbol{\beta}}^r\right\rangle + \sum_{k=1}^{K} h_k(\boldsymbol{\beta}_k^{r+1}) - \sum_{k=1}^{K} h_k(\bar{\boldsymbol{\beta}}_k^r)\\
&= \sum_{k=1}^{K}\left\langle \nabla_k L(\boldsymbol{\beta}^{r+1}),\, \boldsymbol{\beta}_k^{r+1} - \bar{\boldsymbol{\beta}}_k^r\right\rangle + \sum_{k=1}^{K}\left[h_k(\boldsymbol{\beta}_k^{r+1}) - h_k(\bar{\boldsymbol{\beta}}_k^r)\right]\\
&= \sum_{k=1}^{K}\left\langle \nabla_k L(\boldsymbol{\beta}^{r+1}) - \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}),\, \boldsymbol{\beta}_k^{r+1} - \bar{\boldsymbol{\beta}}_k^r\right\rangle\\
&\quad + \sum_{k=1}^{K}\left\langle \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}),\, \boldsymbol{\beta}_k^{r+1} - \bar{\boldsymbol{\beta}}_k^r\right\rangle + \sum_{k=1}^{K}\left[h_k(\boldsymbol{\beta}_k^{r+1}) - h_k(\bar{\boldsymbol{\beta}}_k^r)\right].
\end{aligned} \tag{A-11}$$

Combine (A-10) and (A-11), we get

$$
\begin{aligned}
\left(Q(\boldsymbol{\beta}^{r+1}) - Q(\bar{\boldsymbol{\beta}}^r)\right)^2 &\leq \left(\sum_{k=1}^K \left\langle \nabla_k L(\boldsymbol{\beta}^{r+1}) - \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}),\, \boldsymbol{\beta}_k^{r+1} - \bar{\boldsymbol{\beta}}_k^r \right\rangle\right)^2 \\
&\overset{(a)}{\leq} \left(\sum_{k=1}^K \left\| \nabla_k L(\boldsymbol{\beta}^{r+1}) - \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}) \right\|^2\right) \left(\sum_{k=1}^K \left\| \boldsymbol{\beta}_k^{r+1} - \bar{\boldsymbol{\beta}}_k^r \right\|^2\right) \\
&= \left(\sum_{k=1}^K \left\| \nabla_k L(\boldsymbol{\beta}^{r+1}) - \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}) \right\|^2\right) \left\| \boldsymbol{\beta}^{r+1} - \bar{\boldsymbol{\beta}}^r \right\|^2 \\
&\overset{(b)}{=} \left(\sum_{k=1}^K \left\| \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{\beta}_{-k}^{r+1}) - \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k+1}^{r+1}) \right\|^2\right) \left\| \boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r + \boldsymbol{\beta}^r - \bar{\boldsymbol{\beta}}^r \right\|^2 \\
&\overset{(c)}{\leq} \left(\sum_{k=1}^K G_k^2 \left\| \boldsymbol{\beta}^{r+1} - \boldsymbol{B}_{k+1}^{r+1} \right\|^2\right) \cdot 2 \left(\left\| \boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r \right\|^2 + \left\| \boldsymbol{\beta}^r - \bar{\boldsymbol{\beta}}^r \right\|^2 \right) \\
&\overset{(d)}{\leq} (2\sum_{k=1}^K G_k^2) \left\| \boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r \right\|^2 \left(\left\| \boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r \right\|^2 + \left\| \boldsymbol{\beta}^r - \bar{\boldsymbol{\beta}}^r \right\|^2 \right) \\
&\leq G \left\| \boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r \right\|^2 \left(\left\| \boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r \right\|^2 + \mathrm{d}_{\mathcal{X}^*}^2(\boldsymbol{\beta}^r) \right), \qquad \text{(A-12)}
\end{aligned}
$$

where $G = 2K(\sqrt{\overline{\gamma}}\sqrt{\gamma} + \overline{\gamma})$ and $\overline{\gamma} = \max_{1 \leq k \leq K} \gamma_k$. Inequality (a) in (A-12) is due to the Cauchy-Schwarz inequality, equality (b) is due to that $\nabla_k L(\boldsymbol{\beta}^{r+1}) = \nabla_k L(\boldsymbol{\beta}_k^{r+1}, \boldsymbol{\beta}_{-k}^{r+1}) = \nabla u_k(\boldsymbol{\beta}_k^{r+1}, \boldsymbol{\beta}_{-k}^{r+1})$. In inequalities (c) and (d), we use the inequality (A-5) and $\left\| \boldsymbol{\beta}^{r+1} - \boldsymbol{B}_{k+1}^{r+1} \right\| \leq \|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|$ for all k, respectively.

*Local error bound.*Let $\mathbf{d}_{\mathcal{X}^*}(\boldsymbol{\beta}) \equiv \min_{\boldsymbol{\beta}^* \in \mathcal{X}^*} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|$. Note that the function $p(\mathbf{z}) = n^{-1}\mathbf{1}_n^\top \Psi_\tau(\mathbf{y} - \mathbf{z})$ is strongly convex in $\mathbf{z} \in \mathbb{R}^n$. We can see that $L(\boldsymbol{\beta}) = p(\boldsymbol{X}\boldsymbol{\beta})$. It follows from Zhang et al. (2013) that for any $\xi \geq \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$, there exist $\kappa, \varepsilon > 0$ such that

$$
\mathrm{d}_{\mathcal{X}^*}(\boldsymbol{\beta}) \leq \kappa \|\boldsymbol{\beta} - \mathbf{prox}_h(\boldsymbol{\beta} - \nabla L(\boldsymbol{\beta}))\|, \qquad \text{(A-13)}
$$

for all $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \mathbf{prox}_h(\boldsymbol{\beta} - \nabla L(\boldsymbol{\beta}))\| \leq \varepsilon$ and $Q(\boldsymbol{\beta}) \leq \xi$.

Now we are ready to prove Theorem 1.

Theorem 1 *The GPQR algorithm (Algorithm 2) converges at least linearly to a solution in $\mathcal{X}^*$.*

Proof. We first show that there exist some $\sigma > 0$ such that

$$
\|\boldsymbol{\beta}^r - \mathbf{prox}_h(\boldsymbol{\beta}^r - \nabla L(\boldsymbol{\beta}^r))\| \leq \sigma \|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|, \ \forall r \geq 1. \qquad \text{(A-14)}
$$

For any $r \geq 1$ and any $1 \leq k \leq K$, by the optimality of

$$
\boldsymbol{\beta}_k^{r+1} := \arg\min_{\boldsymbol{\beta}_k} u_k(\boldsymbol{\beta}_k; \boldsymbol{B}_{-k}^{r+1}) + h_k(\boldsymbol{\beta}_k),
$$

we have

$$
\boldsymbol{\beta}_k^{r+1} = \mathbf{prox}_{\gamma_k^{-1} h_k}(\boldsymbol{\beta}_k^r - \gamma_k^{-1} \nabla u_k(\boldsymbol{\beta}_k^r; \boldsymbol{B}_{-k}^{r+1})).
$$

Let $\overline{\gamma} = \max_{1 \leq k \leq K} \gamma_k$, $\underline{\gamma} = \min_{1 \leq k \leq K} \gamma_k$, $\hat{\gamma}_k = \max(1, \gamma_k)$ and $\tilde{\gamma}_k = \max(1, \gamma_k^{-1})$. It follows from Lemma 4.3 of Kadkhodaie et al. (2014) that

$$
\begin{aligned}
&\|\boldsymbol{\beta}_k^r - \mathbf{prox}_{h_k}(\boldsymbol{\beta}_k^r - \nabla_k L(\boldsymbol{\beta}^r))\| \leq \hat{\gamma}_k \|\boldsymbol{\beta}_k^r - \mathbf{prox}_{\gamma_k^{-1} h_k}(\boldsymbol{\beta}_k^r - \gamma_k^{-1} \nabla_k L(\boldsymbol{\beta}^r))\| \\
&\leq \hat{\gamma}_k \left[\|\boldsymbol{\beta}_k^{r+1} - \mathbf{prox}_{\gamma_k^{-1} h_k}(\boldsymbol{\beta}_k^r - \gamma_k^{-1} \nabla_k L(\boldsymbol{\beta}^r))\| + \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\| \right] \\
&\leq \hat{\gamma}_k \left[|\mathbf{prox}_{\gamma_k^{-1} h_k}(\boldsymbol{\beta}_k^{r+1} - \gamma_k^{-1} \nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1})) \right. \\
&\qquad \left. - \mathbf{prox}_{\gamma_k^{-1} h_k}(\boldsymbol{\beta}_k^r - \gamma_k^{-1} \nabla_k L(\boldsymbol{\beta}^r))| + \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\| \right] \\
&\leq 2\hat{\gamma}_k \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\| + \hat{\gamma}_k \gamma_k^{-1} \|\nabla u_k(\boldsymbol{\beta}_k^{r+1}; \boldsymbol{B}_{-k}^{r+1}) - \nabla_k L(\boldsymbol{\beta}^r)\| \\
&\leq 2\hat{\gamma}_k \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\| + \hat{\gamma}_k \gamma_k^{-1} \|\nabla_k L(\boldsymbol{B}_k^{r+1}) + \gamma_k(\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r) - \nabla_k L(\boldsymbol{\beta}^r)\| \\
&\leq 3\hat{\gamma}_k \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\| + \hat{\gamma}_k \tilde{\gamma}_k \|\nabla_k L(\boldsymbol{B}_k^{r+1}) - \nabla_k L(\boldsymbol{\beta}^r)\| \\
&\leq 3\hat{\gamma}_k \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\| + \hat{\gamma}_k \tilde{\gamma}_k \|\nabla L(\boldsymbol{B}_k^{r+1}) - \nabla L(\boldsymbol{\beta}^r)\| \\
&\leq 3\hat{\gamma}_k \|\boldsymbol{\beta}_k^{r+1} - \boldsymbol{\beta}_k^r\| + \hat{\gamma}_k \tilde{\gamma}_k \gamma \|\boldsymbol{B}_k^{r+1} - \boldsymbol{\beta}^r\| \\
&\leq (3 + \gamma\tilde{\gamma}_k)\hat{\gamma}_k \|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|.
\end{aligned}
$$

It follows that

$$\|\boldsymbol{\beta}^r - \mathbf{prox}_h(\boldsymbol{\beta}^r - \nabla L(\boldsymbol{\beta}^r))\| \leq (3+\gamma\tilde{\gamma})\hat{\gamma}K\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|,$$

where $\hat{\gamma} = \max(1, \bar{\gamma})$ and $\tilde{\gamma} = \max(1, \underline{\gamma}^{-1})$. Therefore, when we take $\sigma = (3+\gamma\tilde{\gamma})\hat{\gamma}K$, we get the desired result in (A-14). Note that the sufficient descent property (A-9) implies that $\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\| \to 0$ as $r \to \infty$. It follows from (A-14) that $\|\boldsymbol{\beta}^r - \mathbf{prox}_h(\boldsymbol{\beta}^r - \nabla L(\boldsymbol{\beta}^r))\| \to 0$ as $r \to \infty$. Thus, by (A-13) we have $\mathrm{d}_{\mathcal{X}^*}(\boldsymbol{\beta}^r) \to 0$ as $r \to \infty$. Consequently, using (A-12), we have $Q(\boldsymbol{\beta}^r) \to Q^* := \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$, which shows that the GPQR algorithm converges to the global minimum.

Now, let $c_1 = \underline{\gamma}/2$, $c_2 = \sqrt{G}$, and $\Delta^r = Q(\boldsymbol{\beta}^r) - Q^*$. By the local error bound (A-13) and the cost-to-go estimate (A-12), we obtain

$$\begin{aligned}
\Delta^{r+1} &\leq c_2\sqrt{\left\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\right\|^2\left(\left\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\right\|^2 + \mathrm{d}^2_{\mathcal{X}^*}(\boldsymbol{\beta}^r)\right)} \\
&\leq c_2\sqrt{\left\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\right\|^2\left(\left\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\right\|^2 + \kappa^2\|\boldsymbol{\beta}^r - \mathbf{prox}_h(\boldsymbol{\beta}^r - \nabla L(\boldsymbol{\beta}^r))\|^2\right)} \\
&\overset{(a)}{\leq} c_2\sqrt{\left\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\right\|^2\left(\left\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\right\|^2 + \kappa^2\sigma^2\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2\right)} \\
&\leq (c_2\sqrt{1+\kappa^2\sigma^2})\|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|^2 \\
&\overset{(b)}{\leq} (c_2\sqrt{1+\kappa^2\sigma^2})c_1^{-1}[Q(\boldsymbol{\beta}^r) - Q(\boldsymbol{\beta}^{r+1})] \\
&= (c_2\sqrt{1+\kappa^2\sigma^2})c_1^{-1}(\Delta^r - \Delta^{r+1}).
\end{aligned}$$

Inequality (a) is due to (A-14), and inequality (b) is due to (A-9). This implies that

$$\Delta^{r+1} \leq \frac{c_3}{1+c_3}\Delta^r, \tag{A-15}$$

where $c_3 = (c_2\sqrt{1+\kappa^2\sigma^2})c_1^{-1}$. We can see from (A-15) that $Q(\boldsymbol{\beta}^r)$ approaches Q^* with at least linear rate of convergence. From (A-9) again, this further implies that the sequence $\{\boldsymbol{\beta}^r\}$ converges at least linearly. □

4 Proof of Proposition 3

For Group SCAD penalty

The KKT conditions of the objective function in equation (9) of the main manuscript, with $P_{\lambda,\omega_k}(\|\boldsymbol{\beta}_k\|_2)$ is given by (4), can be written as

$$-\mathbf{Z}_k + \gamma_k\boldsymbol{\beta}_k + P'_{\lambda,\omega_k}(\|\boldsymbol{\beta}_k\|_2) = 0,$$

where $\mathbf{Z}_k = -\nabla_k L(\tilde{\boldsymbol{\beta}}) + \gamma_k\tilde{\boldsymbol{\beta}}_k$.

- If $\|\boldsymbol{\beta}_k\|_2 \leq \lambda$ then $-\mathbf{Z}_k + \gamma_k\boldsymbol{\beta}_k + \lambda w_k\mathbf{u} = 0$ where $\mathbf{u}$ is the sub-gradient and $\|\mathbf{u}\|_2 \leq 1$
  - If $\boldsymbol{\beta}_k = 0$, then

$$\begin{aligned}
&\Rightarrow -\mathbf{Z}_k + \lambda w_k\mathbf{u} = 0 \\
&\Rightarrow \|\mathbf{Z}_k\|_2 \leq \lambda w_k.
\end{aligned}$$

  - If $\boldsymbol{\beta}_k \neq 0$, then one has

$$\begin{aligned}
-\mathbf{Z}_k + \gamma_k\boldsymbol{\beta}_k + \lambda w_k\frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} = 0 &\Rightarrow \|\mathbf{Z}_k\|_2 \leq \gamma_k\|\boldsymbol{\beta}_k\|_2 + \lambda w_k \\
&\Rightarrow \|\mathbf{Z}_k\|_2 \leq \lambda(w_k + \gamma_k).
\end{aligned}$$

  Moreover, we have

$$-\mathbf{Z}_k + \gamma_k\boldsymbol{\beta}_k + \lambda w_k\frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = 0 \ (\mathit{since}\ \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}),$$

which implies

$$\boldsymbol{\beta}_k = \frac{1}{\gamma_k} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} (\|\mathbf{Z}_k\|_2 - \lambda w_k).$$

- If $\lambda \leq \|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda$, then $-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \frac{\theta\lambda w_k}{\theta-1} \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{w_k}{\theta-1}\boldsymbol{\beta}_k = 0$. It follows that

$$\mathbf{Z}_k = [\gamma_k + \frac{w_k}{\theta - 1}(\frac{\theta\lambda}{\|\boldsymbol{\beta}_k\|_2} - 1)]\boldsymbol{\beta}_k,$$

which implies that

$$\|\mathbf{Z}_k\|_2 = (\gamma_k - \frac{w_k}{\theta - 1})\|\boldsymbol{\beta}_k\|_2 + \frac{w_k \lambda \theta}{\theta - 1} \quad \textit{and} \quad \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}.$$

Thus, we have

$$\begin{aligned}
&\lambda \leq \|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda \\
&\Rightarrow (\gamma_k - \tfrac{w_k}{\theta-1})\lambda + \lambda w_k \tfrac{\theta}{\theta-1} \leq (\gamma_k - \tfrac{w_k}{\theta-1})\|\boldsymbol{\beta}_k\|_2 + \lambda w_k \tfrac{\theta}{\theta-1} \leq (\gamma_k - \tfrac{w_k}{\theta-1})\theta\lambda + \lambda w_k \tfrac{\theta}{\theta-1} \\
&\Rightarrow \lambda(\gamma_k + w_k) \leq \|\mathbf{Z}_k\|_2 \leq \gamma_k \theta\lambda \\
&\Rightarrow \boldsymbol{\beta}_k = \tfrac{1}{\gamma_k - \frac{w_k}{\theta-1}} \tfrac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2}(\|\mathbf{Z}_k\|_2 - \lambda w_k \tfrac{\theta}{\theta-1}).
\end{aligned}$$

- If $\|\boldsymbol{\beta}_k\|_2 \geq \theta\lambda$, then $-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k = 0$. This implies that

$$\|\mathbf{Z}_k\|_2 \geq \gamma_k \theta\lambda \quad \text{and} \quad \boldsymbol{\beta}_k = \frac{1}{\gamma_k}\mathbf{Z}_k.$$

To conclude, we have

$$\widehat{\boldsymbol{\beta}}_k = \begin{cases} \frac{1}{\gamma_k} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} S(\|\mathbf{Z}_k\|_2, \lambda w_k), & \text{if} \quad \|\mathbf{Z}_k\|_2 \leq \lambda(\gamma_k + w_k) \\ \frac{1}{\gamma_k - \dfrac{w_k}{\theta - 1}} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2}(\|\mathbf{Z}_k\|_2 - \frac{\lambda w_k \theta}{\theta - 1}), & \text{if} \quad \lambda(\gamma_k + w_k) < \|\mathbf{Z}_k, \|_2 \leq \gamma_k \theta\lambda \\ \frac{1}{\gamma_k}\mathbf{Z}_k, & \text{if} \quad \|\mathbf{Z}_k\|_2 > \gamma_k \theta\lambda. \end{cases}$$

For Group MCP penalty

Again, the KKT conditions of the objective function in equation (9) of the main manuscript, with $P_{\lambda,\omega_k}(\|\boldsymbol{\beta}_k\|_2)$ is given by (3), can be written as

$$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + P'_{(\lambda,\omega_k)}(\|\boldsymbol{\beta}_k\|_2) = 0,$$

where $\mathbf{Z}_k = -\nabla_k L(\tilde{\boldsymbol{\beta}}) + \gamma_k \tilde{\boldsymbol{\beta}}_k$.

- If $\|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda$, then $-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \lambda \mathbf{u} - \frac{w_k}{\theta}\boldsymbol{\beta}_k = 0$, where $\mathbf{u}$ is the sub-gradient and $\|\mathbf{u}\|_2 \leq 1$.
  - If $\boldsymbol{\beta}_k = 0$, then one has

$$-\mathbf{Z}_k + \lambda w_k \mathbf{u} = 0,$$

which implies that

$$\|\mathbf{Z}_k\|_2 \leq \lambda w_k.$$

  - If $\boldsymbol{\beta}_k \neq 0$, then

$$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \lambda w_k \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{w_k}{\theta}\boldsymbol{\beta}_k = 0,$$

which implies that

$$\|\mathbf{Z}_k\|_2 = (\gamma_k - \frac{w_k}{\theta})\|\boldsymbol{\beta}_k\|_2 + \lambda w_k.$$

Thus,

$$\begin{aligned}
&\|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda \\
&\Rightarrow (\gamma_k - \tfrac{w_k}{\theta})\|\boldsymbol{\beta}_k\|_2 + \lambda w_k \leq (\gamma_k - \tfrac{w_k}{\theta})\theta\lambda + \lambda w_k \\
&\Rightarrow \|\mathbf{Z}_k\|_2 \leq \gamma_k \theta\lambda \\
&\Rightarrow \boldsymbol{\beta}_k = \tfrac{1}{\gamma_k - \frac{w_k}{\theta}} \tfrac{\mathbf{Z}^{(k)}}{\|\mathbf{Z}_k\|_2}(\|\mathbf{Z}_k\|_2 - \lambda w_k).
\end{aligned}$$

– If $\|\boldsymbol{\beta}_k\|_2 \geq \theta\lambda$, then we have $-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k = 0$. This implies that

$$\|\mathbf{Z}_k\|_2 \geq \gamma_k \theta \lambda \quad \text{and} \quad \boldsymbol{\beta}_k = \frac{1}{\gamma_k}\mathbf{Z}_k.$$

To sum up, we have

$$\widehat{\boldsymbol{\beta}}_k = \begin{cases} \frac{1}{\gamma_k - w_k/\theta} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} S(\|\mathbf{Z}_k\|_2, \lambda w_k), & \text{if } \|\mathbf{Z}_k\|_2 \leq \gamma_k \theta \lambda \\ \frac{1}{\gamma_k}\mathbf{Z}_k, & \text{if } \|\mathbf{Z}_k\|_2 > \gamma_k \theta \lambda. \end{cases}$$

5 Solution path comparison of GLLA and GSCAD/GMCP penalties

Illustration of the GPQR approach with GLLA approximation compared to the exact GMCP and GSCAD penalties are given in Figure S.1. In this example, we used the smoothed check function $\Psi^{(1)}_{\tau,\delta}(u)$ to approximate the standard quantile check function, with $\delta = 1$. We generated n observations of p-dimensional vector $\boldsymbol{x}_i, i = 1, \ldots, n$, following a multivariate normal distribution, with $p = 200$ and $n = 100$. We divided the p variables into $K = 191$ groups, and assigned non-zero coefficients to the first three groups and set the 188 coefficients of the remaining 188 groups to be zero:

$$\boldsymbol{\beta} = (\underbrace{3, 3, 3, 3}_{G_1}, \underbrace{2, 2, 2, 2}_{G_2}, \underbrace{-1, -1, -1, -1}_{G_3}, \underbrace{0, \ldots, 0}_{G_4 - G_{191}})^\top.$$

The response $y_i, i = 1, \ldots, n$, is generated from the following linear regression model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, 1).$$

Fig. S.1

6 Checking the theoretical KKT conditions

In this section, we establish the theoretical KKT conditions of GPQR solution. When our GPQR algorithm converges to the final solution, it must satisfy those conditions, which means that the algorithm converges and finds the right answer.

For GPQR with GLasso penalty, the KKT conditions of the objective function in equation (7) of the main manuscript with $P_{\lambda,\omega_k}(\|\boldsymbol{\beta}_k\|_2)$ is given by (2) can be written as

$$\nabla_k L(\boldsymbol{\beta}) + \lambda w_k \partial \|\boldsymbol{\beta}_k\|_2 = 0,$$

If $\boldsymbol{\beta}_k = 0$, then we have

$$\nabla_k L(\boldsymbol{\beta}) + \lambda w_k \mathbf{u} = 0,$$

where $\mathbf{u}$ is the sub-gradient of $\|\boldsymbol{\beta}_k\|_2$ and $\|\mathbf{u}\|_2 \leq 1$
which implies

$$\|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda w_k. \tag{A-16}$$

If $\boldsymbol{\beta}_k \neq 0$, then we have

$$\nabla_k L(\boldsymbol{\beta}) + \lambda w_k \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} = 0. \tag{A-17}$$

Combining (A-16) amd (A-17), we get

$$\begin{cases} \nabla_k L(\boldsymbol{\beta}) + \lambda \omega_k \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} = \mathbf{0}, & if \ \ \boldsymbol{\beta}_k \neq 0 \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda w_k, & if \ \ \boldsymbol{\beta}_k = 0. \end{cases}$$

Following the same reasoning as for GLasso and as in Proposition 3 , the exact KKT conditions of GMCP, GSCAD and GLLA are given for each solution $\boldsymbol{\beta}_k, \{k = 1, \ldots, K\}$ respectively, as

$$\begin{cases} \nabla_k L(\boldsymbol{\beta}) + \lambda \omega_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{\boldsymbol{\beta}_k}{\theta} = \mathbf{0}, & if \ \ \boldsymbol{\beta}_k \neq 0 \ \ and \ \ \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda \omega_k, & if \ \ \boldsymbol{\beta}_k = 0 \ \ and \ \ \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 = 0, & if \ \ \|\boldsymbol{\beta}_k\|_2 > \theta\lambda. \end{cases}$$

$$\begin{cases} \nabla_k L(\boldsymbol{\beta}) + \lambda\omega_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} = \mathbf{0}, & if \;\; \boldsymbol{\beta}_k \neq 0 \;\; and \;\; \|\boldsymbol{\beta}_k\|_2 \leqslant \lambda \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega_k, & if \;\; \boldsymbol{\beta}_k = 0 \;\; and \;\; \|\boldsymbol{\beta}_k\|_2 \leqslant \lambda \\ \nabla_k L(\boldsymbol{\beta}) + \frac{\theta}{\theta-1}\lambda\omega_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{\boldsymbol{\beta}_k}{(\theta-1)} = \mathbf{0}, & if \;\; \lambda < \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 = 0, & if \;\; \|\boldsymbol{\beta}_k\|_2 > \theta\lambda. \end{cases}$$

$$\begin{cases} \nabla_k L(\boldsymbol{\beta}) + \lambda\omega'_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} = \mathbf{0}, & if \;\; \boldsymbol{\beta}_k \neq 0 \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega'_k, & if \;\; \boldsymbol{\beta}_k = 0. \end{cases}$$

7 Checking the numerical KKT conditions

The theoretical solution for the GPQR algoithm always passes the KKT condition check defined in the previous section. However, a numerical solution could only approach this theoretical value within certain precision therefore may fail the KKT check. In order to adapt the exact KKT conditions to the numerical solution. Numerically, we declare $\boldsymbol{\beta}_k$ passes the KKT condition check for GLasso, GMCP, GSCAD and GLLA, respectively if

$$\begin{cases} \|\nabla_k L(\boldsymbol{\beta}) + \lambda\omega_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}\|_2 \leqslant \epsilon, & if \;\; \boldsymbol{\beta}_k \neq 0 \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega_k + \epsilon, & if \;\; \boldsymbol{\beta}_k = 0, \end{cases}$$

$$\begin{cases} \|\nabla_k L(\boldsymbol{\beta}) + \lambda\omega_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{\boldsymbol{\beta}_k}{\theta}\|_2 \leqslant \epsilon, & if \;\; \boldsymbol{\beta}_k \neq 0 \;\; and \;\; \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega_k + \epsilon, & if \;\; \boldsymbol{\beta}_k = 0 \;\; and \;\; \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \epsilon, & if \;\; \|\boldsymbol{\beta}_k\|_2 > \theta\lambda, \end{cases}$$

$$\begin{cases} \|\nabla_k L(\boldsymbol{\beta}) + \lambda\omega_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}\|_2 \leqslant \epsilon, & if \;\; \boldsymbol{\beta}_k \neq 0 \;\; and \;\; \|\boldsymbol{\beta}_k\|_2 \leqslant \lambda \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega_k + \epsilon, & if \;\; \boldsymbol{\beta}_k = 0 \;\; and \;\; \|\boldsymbol{\beta}_k\|_2 \leqslant \lambda \\ \|\nabla_k L(\boldsymbol{\beta}) + \frac{\theta}{\theta-1}\lambda\omega_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{\boldsymbol{\beta}_k}{(\theta-1)}\|_2 \leqslant \epsilon, & if \;\; \lambda < \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \epsilon, & if \;\; \|\boldsymbol{\beta}_k\|_2 > \theta\lambda, \end{cases}$$

$$\begin{cases} \|\nabla_k L(\boldsymbol{\beta}) + \lambda\omega'_k \cdot \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}\|_2 \leqslant \epsilon, & if \;\; \boldsymbol{\beta}_k \neq 0 \\ \|\nabla_k L(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega'_k + \epsilon, & if \;\; \boldsymbol{\beta}_k = 0. \end{cases}$$

for a small $\epsilon > 0$. In this paper we set $\epsilon = 10^{-4}$

8 ADNI data analysis

In this section we present additional results of the GPQR approach in the gene-based association study of the ADNI cohort. In this analysis we fitted the GPQR model for two additional locations, $\tau = 0.25, 0.75$.

Figure S.2 highlights results of the L2-norm of the coefficient paths of Q-GLasso, Q-GMCP and Q-GSCAD respectively, with $\tau = 0.25$, or 0.75, as a function of the tuning parameter λ. The results of Figure S.2 obtained by fitting GPQR for all 442 analyzed subjects of the ADNI cohort.

Fig. S.2

Fig. S.3

Table S.1

References

Hong, M., Wang, X., Razaviyayn, M., and Luo, Z.-Q. (2017). Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163(1-2):85–114.

Kadkhodaie, M., Sanjabi, M., and Luo, Z.-Q. (2014). On the linear convergence of the approximate proximal splitting method for non-smooth convex optimization. *Journal of the Operations Research Society of China*, 2(2):123–141.

Luo, Z.-Q. and Tseng, P. (1992). On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425.

Luo, Z.-Q. and Tseng, P. (1993). Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178.

Mkhadri, A., Ouhourane, M., and Oualkacha, K. (2017). A coordinate descent algorithm for computing penalized smooth quantile regression. *Statistics and Computing*, 27(4):865–883.

Sun, R. and Hong, M. (2015). Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. In *Advances in Neural Information Processing Systems*, pages 1306–1314.

Zhang, H., Jiang, J., and Luo, Z.-Q. (2013). On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. *Journal of the Operations Research Society of China*, 1(2):163–186.

Tables

	Q-GLasso	Q-GMCP	Q-GSCAD	Q-GLass	Q-GMCP	Q-GSCAD
Genes		$\tau = 0.25$			$\tau = 0.75$	
APOC1	98.8	97.9	33.7	29.8	7.7	47.9
TOMM40	85.8	29.0	0.0	0.0	0.0	0.0
APOE	94.2	69.2	38.4	14.4	4.8	5.0
	Q-GLasso	Q-GMCP	Q-GSCAD	Q-GLass	Q-GMCP	Q-GSCAD
		$\tau = 0.25$			$\tau = 0.75$	
QPE_τ	0.033	0.032	0.033	0.073	0.075	0.073
Size	13.38	6.61	4.38	3.69	3.46	4.01

Table S.1 top: comparison of the number of times (in %) the genes *APOE, TOMM40* and *APOC1* are selected, based on 100 replications, for ADNI data. bottom: average of the quantile-based error prediction (QPE_τ)

and the number of selected groups/genes (Model Size) computed on the 100 runs' test sets. The group quantile methods are fitted with $\tau = 0.25, 0.75$.

Figure Captions

Fig. S.1. In the left, the coefficient paths of the penalized quantile regression with the group penalties (GMCP and GSCAD), and in the right, their GLLA approximations.

Fig. S.2. L2-norm of the optimal solution coefficients correspond to three important genes are shown as a function of the τ conditional quantile parameter. The genes APOE, TOMM40 APOC1 are plotted in blue, green and red, respectively.

Fig. S.3. At the left and from top to bottom, L2-norm of the coefficient paths of Q-GLasso, Q-GMCP and Q-GSCAD respectively, with $\tau = 0.25$, are shown as a function of a tuning parameter λ. At the right and from top to bottom, the coefficient paths of the same group methods with $\tau = 0.75$.

Figures

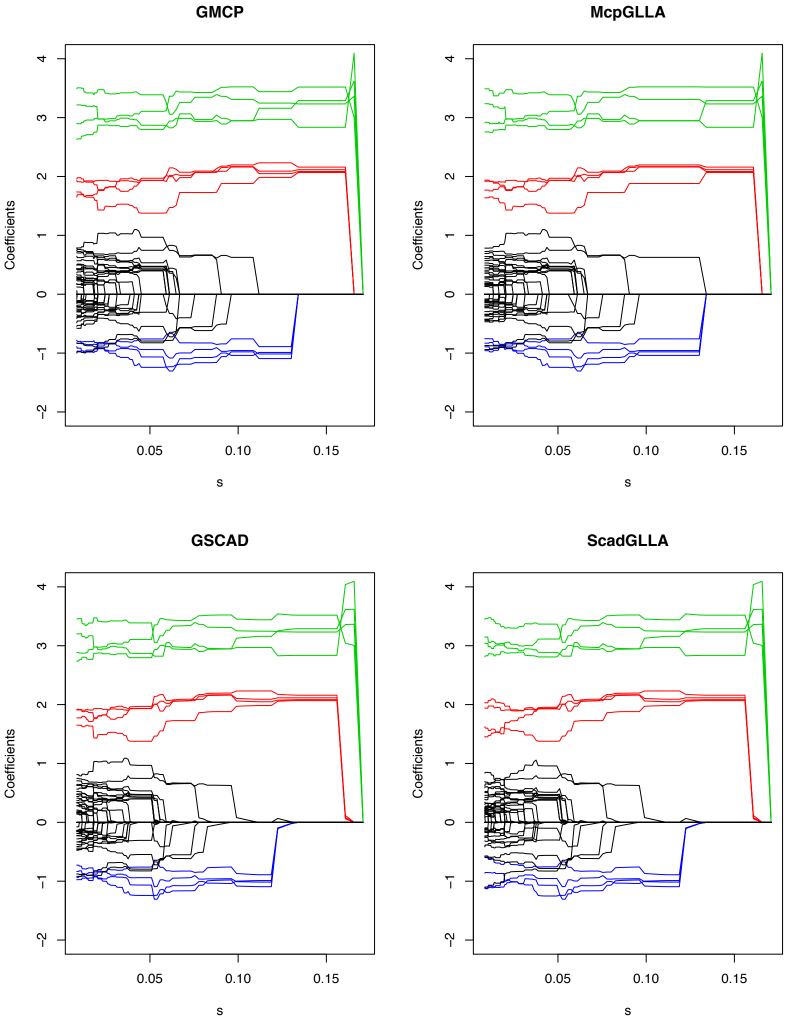

Fig. S.1 In the left, the coefficient paths of the penalized quantile regression with the group penalties (GMCP and GSCAD), and in the right, their GLLA approximations.

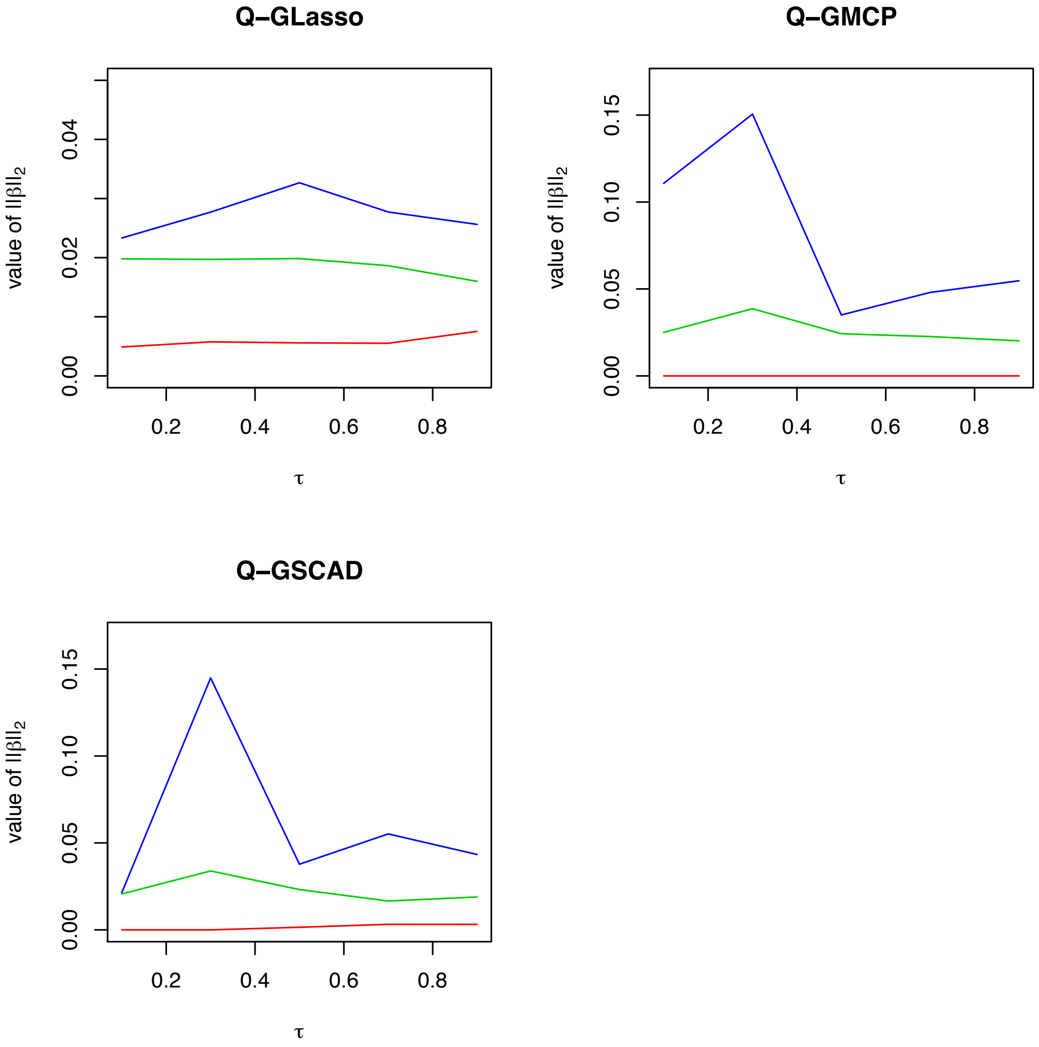

Fig. S.2 L2-norm of the optimal solution coefficients correspond to three important genes are shown as a function of the τ conditional quantile parameter. The genes APOE, TOMM40 APOC1 are plotted in blue, green and red, respectively.

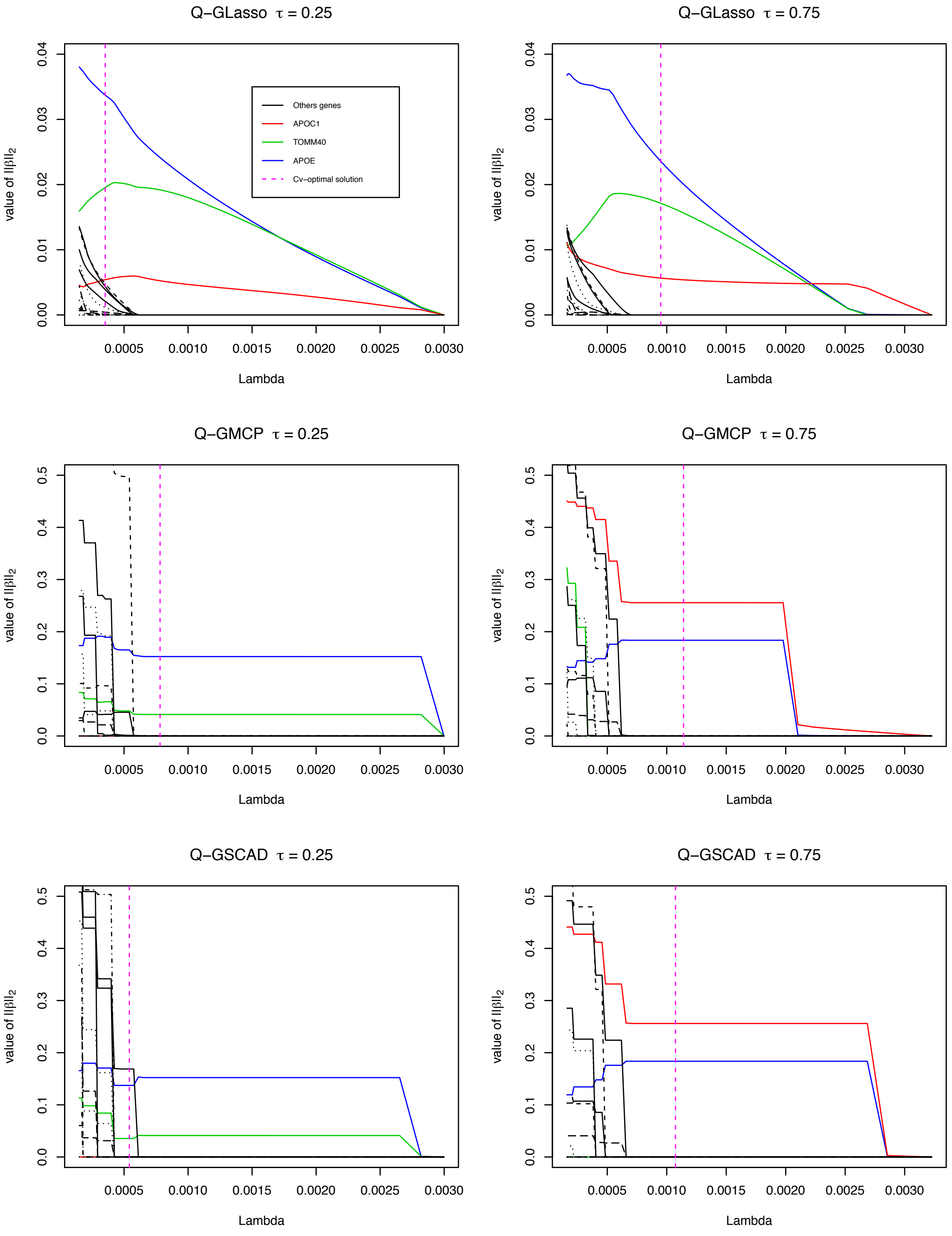

Fig. S.3 At the left and from top to bottom, L2-norm of the coefficient paths of Q-GLasso, Q-GMCP and Q-GSCAD respectively, with $\tau = 0.25$, are shown as a function of a tuning parameter λ. At the right and from top to bottom, the coefficient paths of the same group methods with $\tau = 0.75$.