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.

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

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

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

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

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

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

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

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

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

Thus, we have

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

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

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

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

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

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

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

where $G = 2K(\sqrt{\overline{\gamma}}\sqrt{\gamma} + \overline{\gamma})$ and $\overline{\gamma} = \max_{1\le k\le K} \gamma_k$. 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 $\Big\| \boldsymbol{\beta}^{r+1} - \boldsymbol{B}_{k+1}^{r+1} \Big\| \le \|\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^r\|$ 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 $\xi \ge \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$, there exist $\kappa, \varepsilon > 0$ such that

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

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

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

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

For any $r \ge 1$ and any $1 \le k \le K$, 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 $\overline{\gamma} = \max_{1\le k\le K} \gamma_k$, $\underline{\gamma} = \min_{1\le k\le K} \gamma_k$, $\hat{\gamma}_k = \max(1, \gamma_k)$ 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))\| \le \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))\| \\
&\le \hat{\gamma}_k \big[\|\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\| \big] \\
&\le \hat{\gamma}_k \big[|\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})) \\
&\qquad - \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\| \big] \\
&\le 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)\| \\
&\le 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)\| \\
&\le 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)\| \\
&\le 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)\| \\
&\le 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\| \\
&\le (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

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

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

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

Thus, we have

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

– 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,

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

– 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

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

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

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

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

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

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

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

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

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

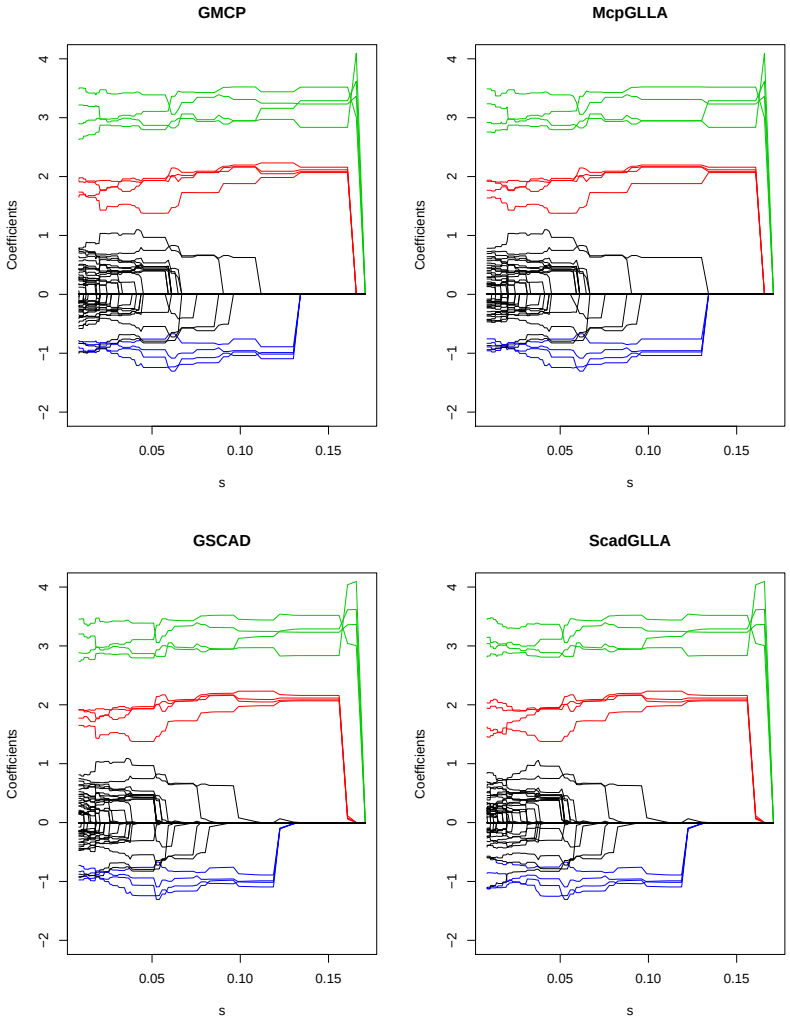

Fig. S.1 In the left, the coefficient paths of the penalized quantile regression with the group penalties (GMCP and GSCAD), and in the right, their GLLA approximations.

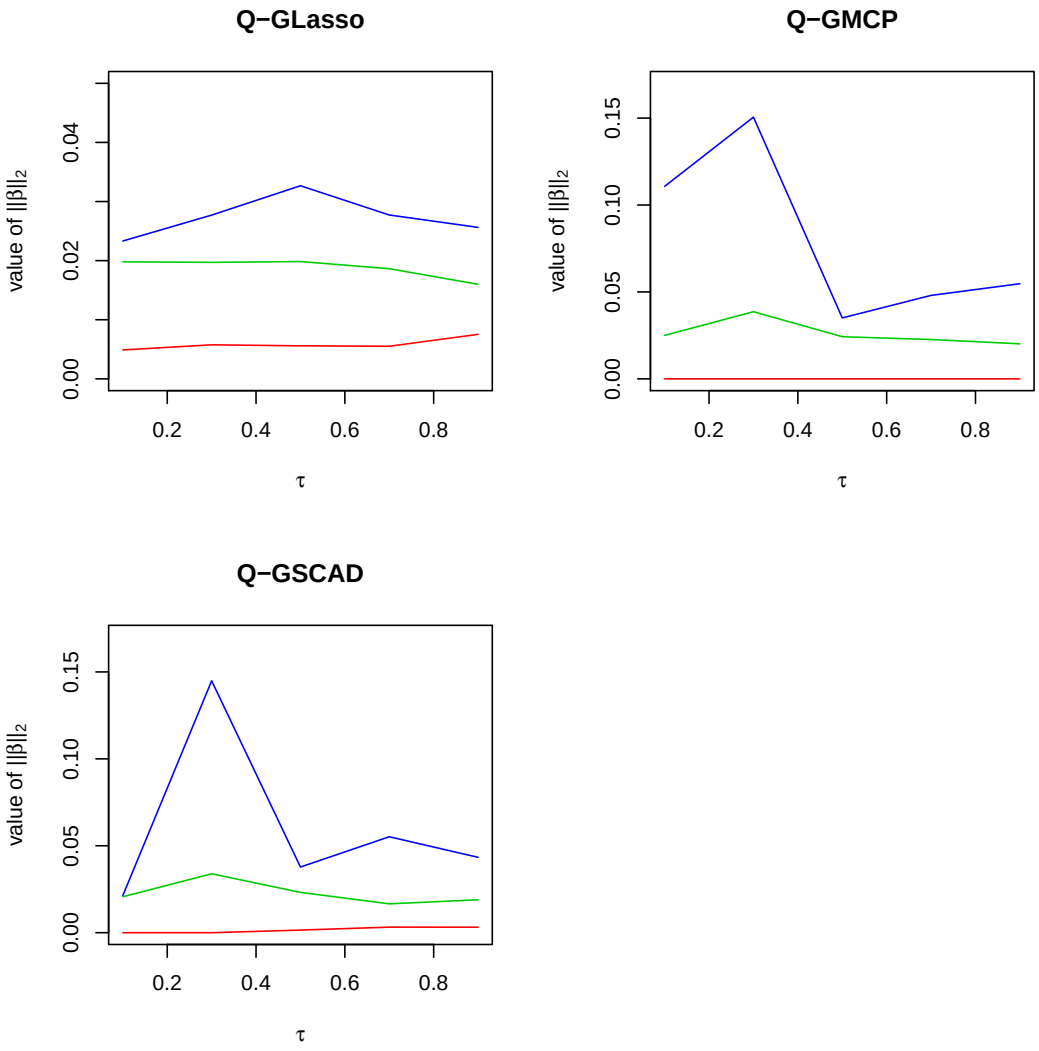

Fig. S.2 L2-norm of the optimal solution coefficients correspond to three important genes are shown as a function of the τ conditional quantile parameter. The genes APOE, TOMM40 APOC1 are plotted in blue, green and red, respectively.

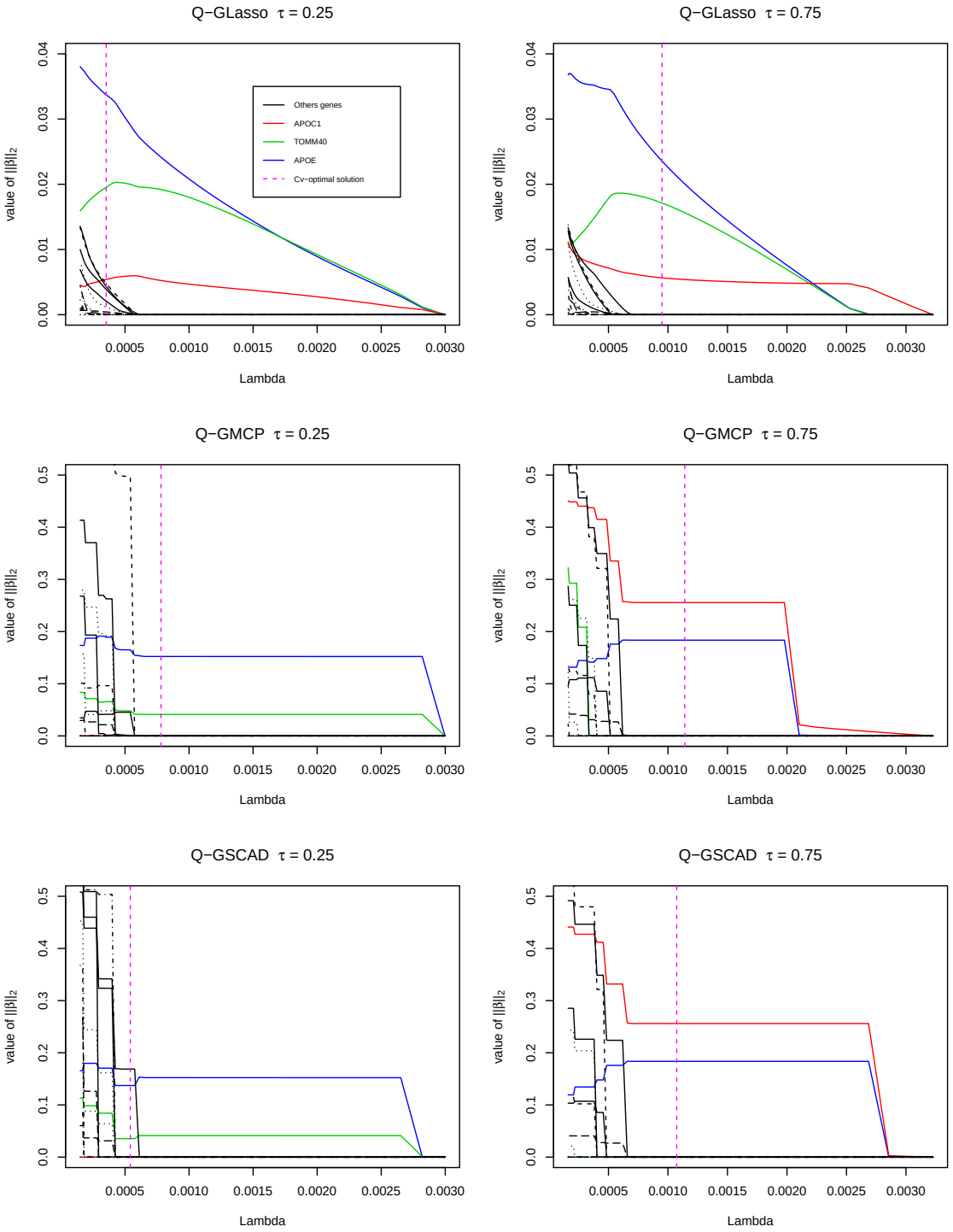

Fig. S.3 At the left and from top to bottom, L2-norm of the coefficient paths of Q-GLasso, Q-GMCP and Q-GSCAD respectively, with $\tau = 0.25$, are shown as a function of a tuning parameter λ. At the right and from top to bottom, the coefficient paths of the same group methods with $\tau = 0.75$.