VARIABLE SELECTION IN QUANTILE REGRESSION

Yichao Wu and Yufeng Liu

North Carolina State University and University of North Carolina

Supplementary Material

This note contains Lemmas 2 and 3, and technical proofs.

Lemma 2 (Convexity Lemma)**.** *Let $\{h_n(\boldsymbol{u}) : \boldsymbol{u} \in \boldsymbol{U}\}$ be a sequence of random convex functions defined on a convex, open subset $\boldsymbol{U}$ of $\mathbb{R}^d$. Suppose $h(\boldsymbol{u})$ is a real-valued function on $\boldsymbol{U}$ for which $h_n(\boldsymbol{u}) \to h(\boldsymbol{u})$ in probability, for each $\boldsymbol{u} \in \boldsymbol{U}$. Then for each compact subset K of $\boldsymbol{U}$,*

$$\sup_{\boldsymbol{u}\in K} \mid h_n(\boldsymbol{u}) - h(\boldsymbol{u}) \mid \to 0 \text{ in probability.}$$

The function $h(\cdot)$ is necessarily convex on $\boldsymbol{U}$.

Proof of Lemma 2. There are many versions of the proof for this well known Convexity Lemma. To save space, we skip its proof. Interested readers are referred to Pollard (1991). □

Denote a linear approximation to $\rho_\tau(\epsilon_i - t)$ by $D_i = (1-\tau)\{\epsilon_i < 0\} - \tau\{\epsilon_i \geq 0\}$. One intuitive interpretation of D_i is that D_i can be thought of as the first derivative of $\rho_\tau(\epsilon_i - t)$ at $t = 0$ (cf Pollard, 1991). Moreover, the condition that ϵ_i has the τ-th quantile zero implies $E(D_i) = 0$. Define $R_{i,n}(\boldsymbol{u}) = \rho_\tau(\epsilon_i - \boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n}) - \rho_\tau(\epsilon_i) - D_i\boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n}$, $W_n = \sum_{i=1}^n D_i\boldsymbol{x}_i/\sqrt{n}$, and $W_{n,11} = \sum_{i=1}^n D_i\boldsymbol{x}_{i1}/\sqrt{n}$. Then $W_n \xrightarrow{\mathcal{L}} N(\boldsymbol{0}, \tau(1-\tau)\Sigma)$ and $W_{n,11} \xrightarrow{\mathcal{L}} N(\boldsymbol{0}, \tau(1-\tau)\Sigma_{11})$.

Lemma 3. *For model (3.1) with true parameter $\boldsymbol{\beta}_0$, denote $G_n(\boldsymbol{u}) = \sum_{i=1}^n [\rho_\tau(\epsilon_i - \boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n}) - \rho_\tau(\epsilon_i)]$, where $\epsilon_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0$. Under Conditions (i) and (ii), we have, for any fixed $\boldsymbol{u}$,*

$$G_n(\boldsymbol{u}) = \frac{f(0)}{2}\boldsymbol{u}^T\frac{\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\boldsymbol{u} + W_n^T\boldsymbol{u} + o_p(1). \tag{1}$$

Proof of Lemma 3. Note first that Condition (i) ensures that the function $M(t) = E(\rho_\tau(\epsilon_i - t) - \rho_\tau(\epsilon_i))$ has a unique minimizer at zero, and its Taylor expansion at origin has the following form $M(t) = \frac{f(0)}{2}t^2 + o(t^2)$. Hence, for large n, we have

$$\begin{aligned} E(G_n(\boldsymbol{u})) &= \sum_{i=1}^n M(\frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}}) = \sum_{i=1}^n [\frac{f(0)}{2}(\frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}})^2 + o((\frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}})^2)] \\ &= \frac{f(0)}{2n}\boldsymbol{u}^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{u} + o(\frac{1}{2n}\boldsymbol{u}^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{u}). \end{aligned}$$

So, under Condition (ii), we have $E(G_n(\boldsymbol{u})) = \frac{f(0)}{2n}\boldsymbol{u}^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{u} + o(1)$.

Hence $G_n(\boldsymbol{u}) = E(G_n(\boldsymbol{u})) + W_n^T\boldsymbol{u} + \sum_{i=1}^n(R_{i,n}(\boldsymbol{u}) - ER_{i,n}(\boldsymbol{u}))$. By routine calculation, we get $\mid R_{i,n}(\boldsymbol{u}) \mid \leq \mid \boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n} \mid \{\mid \epsilon_i \mid \leq \mid \boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n} \mid\}$. For fixed $\boldsymbol{u}$, due to the cancelation of cross-product terms, we get

$$\begin{aligned} E(\sum_{i=1}^n [R_{i,n}(\boldsymbol{u}) - ER_{i,n}(\boldsymbol{u})])^2 &= \sum_{i=1}^n E(R_{i,n}(\boldsymbol{u}) - ER_{i,n}(\boldsymbol{u}))^2 \\ &\leq \sum_{i=1}^n E(R_{i,n}(\boldsymbol{u}))^2 \\ &\leq \sum_{i=1}^n \left[\mid \frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}} \mid^2 E\{\mid \epsilon_i \mid \leq \mid \frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}} \mid\}\right] \\ &\leq (\sum_{i=1}^n \mid \frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}} \mid^2) E\{\mid \epsilon_i \mid \leq \frac{\parallel \boldsymbol{u} \parallel}{\sqrt{n}} \max_{j=1,2,\cdots,n} \parallel \boldsymbol{x}_j \parallel\} \\ &\to 0 \end{aligned} \tag{2}$$

as in Pollard (1991), where $\parallel \cdot \parallel$ denotes the Euclidean norm operator. Here the last step converging to zero holds because

$$\begin{aligned} \sum_{i=1}^n \mid \frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}} \mid^2 &= \boldsymbol{u}^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T/n)\boldsymbol{u} \to \boldsymbol{u}^T\Sigma\boldsymbol{u} \\ \max_{j=1,2,\cdots,n} \parallel \boldsymbol{x}_j \parallel /\sqrt{n} &\to 0 \text{ due to } \frac{\sum_{i=1}^n \parallel \boldsymbol{x}_i \parallel^2}{n} \to \text{trace}(\Sigma). \end{aligned}$$

Equation (2) implies that $\sum_{i=1}^n(R_{i,n}(\boldsymbol{u}) - ER_{i,n}(\boldsymbol{u})) = o_p(1)$. This completes the proof. □

Before we start the proof of Theorem 1, we want to point that $W_n^T\boldsymbol{u} = E(W_n^T\boldsymbol{u}) + O_p(\sqrt{\text{Var}(W_n^T\boldsymbol{u})})$, together with $\text{Var}(W_n^T\boldsymbol{u}) = \sum_{i=1}^n E(D_i\boldsymbol{x}_i^T\boldsymbol{u}/\sqrt{n})^2 = \tau(1-\tau)\boldsymbol{u}^T\frac{\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\boldsymbol{u}$, implies that $W_n^T\boldsymbol{u} = O_p\left(\sqrt{\tau(1-\tau)\boldsymbol{u}^T\frac{\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\boldsymbol{u}}\right)$.

Proof of Theorem 1. We use the same strategy as in Fan and Li (2001). To prove Theorem 1, it is enough to show that for any given $\delta > 0$, there exists a large constant C such that

$$P\left\{\inf_{\parallel \boldsymbol{u} \parallel = C} Q(\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n}) > Q(\boldsymbol{\beta}_0)\right\} \geq 1 - \delta \tag{3}$$

which implies that with probability at least $1 - \delta$ there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n} : \parallel \boldsymbol{u} \parallel \leq C\}$. This in turn implies that there exists a local minimizer such that

$\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \| = O_p(1/\sqrt{n})$, which is exactly what we want to show. Note that

$$\begin{aligned} & Q(\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n}) - Q(\boldsymbol{\beta}_0) \\ = & \sum_{i=1}^{n} [\rho_\tau(y_i - \boldsymbol{x}_i^T(\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n})) - \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0)] + n\sum_{j=1}^{d}[p_{\lambda_n}(\mid \beta_{j0} + u_j/\sqrt{n} \mid) - p_{\lambda_n}(\mid \beta_{j0} \mid)] \\ \geq & \sum_{i=1}^{n} [\rho_\tau(y_i - \boldsymbol{x}_i^T(\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n})) - \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0)] + n\sum_{j=1}^{s}[p_{\lambda_n}(\mid \beta_{j0} + u_j/\sqrt{n} \mid) - p_{\lambda_n}(\mid \beta_{j0} \mid)], \end{aligned}$$

where s is the number of components in $\boldsymbol{\beta}_{10}$ and β_{j0} denotes the j-th component of $\boldsymbol{\beta}_{10}$. Due to Lemma 3, the first term on the right hand side is exactly $G_n(\boldsymbol{u}) = \frac{f(0)}{2}\boldsymbol{u}^T\frac{\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\boldsymbol{u} + W_n^T\boldsymbol{u} + o_p(1)$ for any fixed $\boldsymbol{u}$.

By applying the Convexity Lemma (Lemma 2) to $h_n(\boldsymbol{u}) = G_n(\boldsymbol{u}) - W_n^T\boldsymbol{u}$, we can strengthens this pointwise convergence to uniform convergence on any compact subsect of $\mathbb{R}^d$.

Note that, for large n,

$$n\sum_{j=1}^{s}[p_{\lambda_n}(\mid \beta_{j0} + u_j/\sqrt{n} \mid) - p_{\lambda_n}(\mid \beta_{j0} \mid)] = 0 \tag{4}$$

uniformly in any compact set of $\mathbb{R}^d$ due to the facts that $\beta_{j0} > 0$ for $j = 1, 2, \cdots, s$, SCAD penalty is flat for coefficient of magnitude larger than $a\lambda_n$, and $\lambda_n \to 0$.

Based on all the above, $Q(\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n}) - Q(\boldsymbol{\beta}_0)$ is dominated by the quadratic term $f(0)\boldsymbol{u}^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{u}/(2n)$ for $\| \boldsymbol{u} \|$ equal to sufficiently large C. Hence Condition (ii) implies that (3) holds as we have desired and this completes the proof. □

Proof of Lemma 1. For any $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$, $0 < \| \boldsymbol{\beta}_2 \| \leq Cn^{-1/2}$,

$$\begin{aligned} & Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) - Q((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T) \\ = & [Q((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) - Q((\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T)] - [Q((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T) - Q((\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T)] \\ = & G_n(\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)^T) - G_n(\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \boldsymbol{\beta}_2^T)^T) - n\sum_{j=s+1}^{d} p_{\lambda_n}(\mid \beta_j \mid) \qquad (5) \\ = & \frac{f(0)}{2}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)\frac{\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)^T + \sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)W_n \\ & -\frac{f(0)}{2}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \boldsymbol{\beta}_2^T)\frac{\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \boldsymbol{\beta}_2^T)^T - \sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \boldsymbol{\beta}_2^T)W_n \\ & +o(1) + o_p(1) - n\sum_{j=s+1}^{d} p_{\lambda_n}(\mid \beta_j \mid) \end{aligned}$$

The conditions $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$ and $0 < \| \boldsymbol{\beta}_2 \| \leq Cn^{-1/2}$ imply that

$$\frac{f(0)}{2}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)\frac{\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)^T = O_p(1)$$

$$\frac{f(0)}{2}\sqrt{n}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^T,\boldsymbol{\beta}_2^T)\frac{\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\sqrt{n}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^T,\boldsymbol{\beta}_2^T)^T=O_p(1)$$

and

$$\begin{aligned}&\sqrt{n}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^T,\mathbf{0}^T)W_n-\sqrt{n}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^T,\boldsymbol{\beta}_2^T)W_n=-\sqrt{n}(\mathbf{0}^T,\boldsymbol{\beta}_2^T)W_n\\ =\;&\sqrt{n}\sqrt{\tau(1-\tau)\boldsymbol{\beta}_2^T\Sigma_{22}\boldsymbol{\beta}_2}(1+o_p(1)).\end{aligned}$$

Note that

$$\begin{aligned}n\sum_{j=s+1}^{d}p_{\lambda_n}(\mid\beta_j\mid)&\geq n\lambda_n\left(\liminf_{\lambda\to0}\liminf_{\theta\to0+}\frac{p'_\lambda(\theta)}{\lambda}\right)\left(\sum_{j=s+1}^{d}|\beta_j|\right)(1+o(1))\\&=n\lambda_n\left(\sum_{j=s+1}^{d}|\beta_j|\right)(1+o(1)),\end{aligned}$$

where the last step follows based on the fact that $\liminf_{\lambda\to0}\liminf_{\theta\to0+}\frac{p'_\lambda(\theta)}{\lambda}=1$.

Then $\sqrt{n}\lambda_n\to\infty$ implies that $n\lambda_n=\sqrt{n}(\sqrt{n}\lambda_n)$ is of higher order than $\sqrt{n}$. This implies that, in (5), the last term dominates in magnitude and, as a result, $Q((\boldsymbol{\beta}_1^T,\mathbf{0}^T)^T)-Q((\boldsymbol{\beta}_1^T,\boldsymbol{\beta}_2^T)^T)<0$ for large n. The completes the proof. □

Proof of Theorem 2. Similarly as in Fan and Li (2001), Part (a) holds simply due to Lemma 1. Next we prove part (b). By Theorem 1, we can show that there exists a root-n consistent minimizer $\hat{\boldsymbol{\beta}}_1$ of $Q((\boldsymbol{\beta}_1^T,\mathbf{0}^T)^T)$ as a function of $\boldsymbol{\beta}_1$.

From the proof of Theorem 1, we see that $\sqrt{n}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})$ minimizes $G_n((\boldsymbol{\theta}^T,\mathbf{0}^T)^T)+n\sum_{j=1}^s p_{\lambda_n}(\mid\beta_{j0}+\frac{\theta_j}{\sqrt{n}}\mid)$, where $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_s)^T\in\mathbb{R}^s$. Notice that, as in the proof of Theorem 1, Lemma 3 and the convexity lemma imply that

$$\begin{aligned}G_n((\boldsymbol{\theta}^T,\mathbf{0}^T)^T)&=\frac{f(0)}{2}(\boldsymbol{\theta}^T,\mathbf{0}^T)\frac{\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^T}{n}(\boldsymbol{\theta}^T,\mathbf{0}^T)^T+(\boldsymbol{\theta}^T,\mathbf{0}^T)W_n+o_p(1)\\&=\frac{f(0)}{2}\boldsymbol{\theta}^T\frac{\sum_{i=1}^n\boldsymbol{x}_{i1}\boldsymbol{x}_{i1}^T}{n}\boldsymbol{\theta}+\boldsymbol{\theta}^T\sum_{i=1}^n D_i\boldsymbol{x}_{i1}/\sqrt{n}+o_p(1)\end{aligned}$$

uniformly in any compact subset of $\mathbb{R}^s$. Notice that, for large n, $n\sum_{j=1}^s p_{\lambda_n}(\mid\beta_{j0}+\theta_j/\sqrt{n}\mid)=n\sum_{j=1}^s p_{\lambda_n}(\mid\beta_{j0}\mid)$ uniformly in any compact set of $\mathbb{R}^s$, due to (4). Hence we have

$$\begin{aligned}&G_n((\boldsymbol{\theta}^T,\mathbf{0}^T)^T)+n\sum_{j=1}^s p_{\lambda_n}(\mid\beta_{j0}+\frac{\theta_j}{\sqrt{n}}\mid)\\ =\;&\frac{1}{2}\boldsymbol{\theta}^T(f(0)\frac{\sum_{i=1}^n\boldsymbol{x}_{i1}\boldsymbol{x}_{i1}^T}{n})\boldsymbol{\theta}+(\sum_{i=1}^n D_i\boldsymbol{x}_{i1}/\sqrt{n})^T\boldsymbol{\theta}+n\sum_{j=1}^s p_{\lambda_n}(\mid\beta_{j0}\mid)+r_n(\boldsymbol{\theta})\\ =\;&\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\zeta}_n)^T(f(0)\frac{\sum_{i=1}^n\boldsymbol{x}_{i1}\boldsymbol{x}_{i1}^T}{n})(\boldsymbol{\theta}-\boldsymbol{\zeta}_n)-\frac{1}{2}\boldsymbol{\zeta}_n^T(f(0)\frac{\sum_{i=1}^n\boldsymbol{x}_{i1}\boldsymbol{x}_{i1}^T}{n})\boldsymbol{\zeta}_n+n\sum_{j=1}^s p_{\lambda_n}(\mid\beta_{j0}\mid)+r_n(\boldsymbol{\theta})\end{aligned}$$

where $\boldsymbol{\zeta}_n = -(f(0)\frac{\sum_{i=1}^n \boldsymbol{x}_{i1}\boldsymbol{x}_{i1}^T}{n})^{-1}W_{n,11}$ and the residual $r_n(\boldsymbol{\theta}) \to 0$ in probability uniformly in any compact subset of $I\!R^s$. Notice further that the term $n\sum_{j=1}^s p_{\lambda_n}(\mid \beta_{j0} \mid)$ does not depend on $\boldsymbol{\theta}$. So this implies that, for large n, the local minimizer $\hat{\boldsymbol{\theta}}$ is very close to $\boldsymbol{\zeta}_n$ and satisfies $\hat{\boldsymbol{\theta}} - \boldsymbol{\zeta}_n = o_p(1)$.

That is, the minimizer $\hat{\theta}$ satisfies $\hat{\theta} = -(f(0)\frac{\sum_{i=1}^n \boldsymbol{x}_{i1}\boldsymbol{x}_{i1}^T}{n})^{-1}(\sum_{i=1}^n D_i\boldsymbol{x}_{i1}/\sqrt{n}) + o_p(1)$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) = -(f(0)\frac{\sum_{i=1}^n \boldsymbol{x}_{i1}\boldsymbol{x}_{i1}^T}{n})^{-1}(\sum_{i=1}^n D_i\boldsymbol{x}_{i1}/\sqrt{n}) + o_p(1)$. Applying *Slutsky*'s theorem, we get $\sqrt{n}f(0)\Sigma_{11}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \tau(1-\tau)\Sigma_{11})$. This completes the proof. □

Proof of Theorem 3. Note that

$$\begin{aligned} & Q_1(\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n}) - Q_1(\boldsymbol{\beta}_0) \\ = & \sum_{i=1}^n [\rho_\tau(y_i - \boldsymbol{x}_i^T(\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n})) - \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0)] + n\lambda_n \sum_{j=1}^d [\tilde{w}_j \mid \beta_{j0} + u_j/\sqrt{n} \mid - \tilde{w}_j \mid \beta_{j0} \mid]. \end{aligned}$$

We consider the second term first, for $j = 1, 2, \cdots, s$, we have $\beta_{j0} \neq 0$; as a result, $\tilde{w}_j \xrightarrow{P} |\beta_{j0}|^{-\gamma}$; hence $n\lambda_n[\tilde{w}_j \mid \beta_{j0} + u_j/\sqrt{n} \mid - \tilde{w}_j \mid \beta_{j0} \mid] \xrightarrow{P} 0$ as $\sqrt{n}(\mid \beta_{j0} + u_j/\sqrt{n} \mid - \mid \beta_{j0} \mid) \to u_j \operatorname{sign}(\beta_{j0})$ and $\sqrt{n}\lambda_n \to 0$. On the other hand, for $j = s+1, s+2, \cdots, d$, the true coefficient $\beta_{j0} = 0$; so $\sqrt{n}\lambda_n\tilde{w}_j = n^{(1+\gamma)/2}\lambda_n(\sqrt{n}|\tilde{\beta}_j|)^{-\gamma}$ with $\sqrt{n}\tilde{\beta}_j = O_p(1)$; so it follows that $n\lambda_n[\tilde{w}_j \mid \beta_{j0} + u_j/\sqrt{n} \mid - \tilde{w}_j \mid \beta_{j0} \mid] \xrightarrow{P} \infty$ when $u_j \neq 0$ and $= 0$ otherwise due to $\sqrt{n} \mid u_j/\sqrt{n} \mid = |u_j|$ for large n. These facts and the result of Lemma 3 imply that

$$Q_1(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}) - Q_1(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} V(\boldsymbol{u}) = \begin{cases} \frac{f(0)}{2}\boldsymbol{u}_1\Sigma_{11}\boldsymbol{u}_1 + W_{n,11}^T\boldsymbol{u}_1 & \text{when } u_j = 0 \text{ for } j \geq s+1 \\ \infty & \text{otherwise,} \end{cases}$$

where $\boldsymbol{u}_1 = (u_1, u_2, \cdots, u_s)^T$. Noticing that $Q_1(\boldsymbol{\beta}_0 + \boldsymbol{u}/\sqrt{n}) - Q_1(\boldsymbol{\beta}_0)$ is convex in $\boldsymbol{u}$ and V has a unique minimizer, the epi-convergence results of Geyer (1994) imply that

$$\operatorname{argmin} Q_1(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}) = \sqrt{n}(\hat{\boldsymbol{\beta}}^{(AL)} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \operatorname{argmin} V(\boldsymbol{u}),$$

which establishes the asymptotic normality part. Next we show the consistency property of model selection.

For any $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$, $0 < \| \boldsymbol{\beta}_2 \| \leq Cn^{-1/2}$,

$$\begin{aligned} & Q_1((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) - Q_1((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T) \\ = & [Q_1((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) - Q_1((\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T)] - [Q((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T) - Q((\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T)] \\ = & G_n(\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \mathbf{0}^T)^T) - G_n(\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^T, \boldsymbol{\beta}_2^T)^T) - n\lambda_n \sum_{j=s+1}^d \tilde{w}_j \mid \beta_j \mid . \end{aligned}$$

Note here the first two terms are exactly the same as in (5) and hence can be bounded similarly.

However the third term goes to $-\infty$ as $n \to \infty$ due to the following

$$n\lambda_n \sum_{j=s+1}^{d} \tilde{w}_j \mid \beta_j \mid \quad = \quad (n^{(1+\gamma)/2}\lambda_n)\sqrt{n} \sum_{j=s+1}^{d} \left|(\sqrt{n}|\tilde{\beta}_j|)^{-\gamma}\right| |\beta_j| \to \infty.$$

Hence the condition that $n^{(1+\gamma)/2}\lambda_n \to \infty$ implies that $n\lambda_n \sum_{j=s+1}^{d} \tilde{w}_j \mid \beta_j \mid$ is of higher order than any other terms and dominates as a result. This in turn implies that $Q_1((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T) - Q_1((\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T) < 0$ for large n. This proves the consistency of model selection of adaptive lasso penalized quantile regression. □

Proof of Corollary 1. Notice from the proofs of Theorem 1, Lemma 1, Theorem 2, and Theorem 3, it is enough to establish an asymptotic approximation similar as (1).

Note that the check function $\rho_\tau(\cdot)$ can be rewritten as $\rho_\tau(r) = |r|/2 + (\tau - 1/2)r$. Hence,

$$\begin{aligned} G_n(\boldsymbol{u}) \quad &= \quad \sum_{i=1}^{n} [\rho_\tau(\epsilon_i - \boldsymbol{x}_i^T \boldsymbol{u}/\sqrt{n}) - \rho_\tau(\epsilon_i)] \\ &= \quad \sum_{i=1}^{n} \frac{-\boldsymbol{x}_i^T \boldsymbol{u}}{\sqrt{n}} (\frac{\text{sign}(\epsilon_i)}{2} + (\tau - \frac{1}{2})) + \sum_{i=1}^{n} \int_0^{\frac{\boldsymbol{x}_i^T \boldsymbol{u}}{\sqrt{n}}} (I(\epsilon_i \le s) - I(\epsilon_i \le 0)) ds \end{aligned}$$

as in Knight (1999). By the same argument as in Knight (1999), we get $G_n(\boldsymbol{u}) \xrightarrow{\mathcal{L}} -\boldsymbol{u}^T \boldsymbol{V} + \varsigma(\boldsymbol{u})$ for some multivariate normal random vector $\boldsymbol{V}$ with mean zero. Then the result follows from the strictly convexity of $\varsigma(\cdot)$. □