SUPPLEMENTARY MATERIAL FOR: ADAPTIVE ROBUST VARIABLE SELECTION

By Jianqing Fan, Yingying Fan and Emre Barut

Princeton University, University of Southern California and IBM T.J. Watson Research Center

APPENDIX A: ADDITIONAL PROOFS

A.1. Proof of Theorem 3. This proof is motivated by Pollard (1990). In the proof we use $C > 0$ to denote a generic positive constant.

Let $\boldsymbol{\theta} = \mathbf{V}_n^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)$, then $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta}$. Letting $G_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\beta}_1, \mathbf{0}) - L_n(\boldsymbol{\beta}_1^*, \mathbf{0})$, we have that

$$G_n(\boldsymbol{\theta}) = \|\rho_\tau(\boldsymbol{\varepsilon} - \mathbf{Z}_n\boldsymbol{\theta})\|_1 - \|\rho_\tau(\boldsymbol{\varepsilon})\|_1 + n\lambda_n\big(\|\mathbf{d}_0 \circ \big(\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta}\big)\|_1 - \|\mathbf{d}_0 \circ \boldsymbol{\beta}_1^*\|_1\big), \tag{A.1}$$

where we have used the shorthand notation that $\|\rho_\tau(\mathbf{u})\|_1 = \sum_{i=1}^n \rho_\tau(u_i)$ for any vector $\mathbf{u} = (u_1, \cdots, u_n)^T$. Since $L_n(\boldsymbol{\beta}_1, \mathbf{0})$ is minimized at $\boldsymbol{\beta}_1 = \widehat{\boldsymbol{\beta}}_1^o$, it follows that $G_n(\boldsymbol{\theta})$ is minimized at $\widehat{\boldsymbol{\theta}}_n = \mathbf{V}_n^{-1}(\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*)$. We consider $\boldsymbol{\theta}$ over the convex open set

$$B_0(n) = \{\boldsymbol{\theta} \in \mathbf{R}^s : \|\boldsymbol{\theta}\|_2 < c_6\sqrt{s}\},$$

with some constant $c_6 > 0$ independent of s.

The idea of the proof is to approximate the stochastic function $G_n(\boldsymbol{\theta})$ by a quadratic function, whose minimizer is shown to possess the asymptotic normality. Since $G_n(\boldsymbol{\theta})$ and the quadratic approximation are close, the minimizer of $G_n(\boldsymbol{\theta})$ enjoys the same asymptotic normality. Now, we proceed to prove Theorem 3.

Decompose $G_n(\boldsymbol{\theta})$ into its mean and centralized stochastic component:

$$G_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}) + T_n(\boldsymbol{\theta}), \tag{A.2}$$

where $Q_n(\boldsymbol{\theta}) = E[G_n(\boldsymbol{\theta})]$ and

$$T_n(\boldsymbol{\theta}) = \|\rho_\tau(\boldsymbol{\varepsilon} - \mathbf{Z}_n\boldsymbol{\theta})\|_1 - \|\rho_\tau(\boldsymbol{\varepsilon})\|_1 - E\big[\|\rho_\tau(\boldsymbol{\varepsilon} - \mathbf{Z}_n\boldsymbol{\theta})\|_1 - \|\rho_\tau(\boldsymbol{\varepsilon})\|_1\big]. \tag{A.3}$$

We first deal with the mean component $Q_n(\boldsymbol{\theta})$. Since $\|\boldsymbol{\theta}\|_2 < c_6\sqrt{s}$ over the set $B_0(n)$, it follows from the Cauchy-Schwarz inequality and the assumption of the theorem that

$$\|\mathbf{H}^{1/2}\mathbf{Z}_n\boldsymbol{\theta}\|_\infty \le \|\boldsymbol{\theta}\|_2 \max_i \|\mathbf{H}^{1/2}\mathbf{Z}_{ni}\|_2 = o\big(s^{-3}(\log s)^{-1}\big).$$

Then, since $f''(u)$ is bounded in a small neighborhood of 0, by using a similar argument as equation (7.3) of the main paper and noting that $\sum_{i=1}^n f_i(0)|\mathbf{Z}_{ni}^T\boldsymbol{\theta}|^2 = \boldsymbol{\theta}^T\big(\mathbf{Z}_n^T\mathbf{H}\mathbf{Z}_n\big)\boldsymbol{\theta} = \|\boldsymbol{\theta}\|_2^2$ we can show that

$$\begin{aligned} &E\big[\|\rho_\tau(\boldsymbol{\varepsilon} - \mathbf{Z}_n\boldsymbol{\theta})\|_1 - \|\rho_\tau(\boldsymbol{\varepsilon})\|_1\big] \\ &\quad = \|\boldsymbol{\theta}\|_2^2 + \frac{1}{2}\sum_{i=1}^n f_i'(0)|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^3 + o\Big(\sum_{i=1}^n f_i'(0)|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^3\Big). \end{aligned} \tag{A.4}$$

Furthermore, since $\sum_{i=1}^n f_i(0)|\mathbf{Z}_{ni}^T\boldsymbol{\theta}|^2 = \|\boldsymbol{\theta}\|_2^2 < c_6^2 s$, it follows that

(A.5)
$$\sum_{i=1}^n f_i'(0)|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^3 \le C\|\mathbf{H}^{1/2}\mathbf{Z}_n\boldsymbol{\theta}\|_\infty \sum_{i=1}^n f_i(0)|\mathbf{Z}_{ni}^T\boldsymbol{\theta}|^2 = o\big(s^{-2}(\log s)^{-1}\big).$$

Next, we deal with the penalty term in the expected value $Q_n(\boldsymbol{\theta})$. Since $\frac{1}{n}\mathbf{S}^T\mathbf{H}\mathbf{S}$ has bounded eigenvalues by Condition 2, it follows from the assumption of the theorem that, for any $\boldsymbol{\theta} \in B_0(n)$,

$$\|\mathbf{V}_n\boldsymbol{\theta}\|_\infty \le \|\mathbf{V}_n\boldsymbol{\theta}\|_2 \le Cn^{-1/2}\|\boldsymbol{\theta}\|_2 = o\Big(\min_{\{1\le j\le s\}}|\beta_j^*|\Big).$$

Hence, $\mathrm{sgn}(\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta}) = \mathrm{sgn}(\boldsymbol{\beta}_1^*)$ and

$$\|\mathbf{d}_0 \circ (\boldsymbol{\beta}_1^* + \mathbf{V}_n\boldsymbol{\theta})\|_1 - \|\mathbf{d}_0 \circ \boldsymbol{\beta}_1^*\|_1 = \tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta}, \tag{A.6}$$

where $\tilde{\mathbf{d}}_0$ is a s-vector with j-th component $d_j\mathrm{sgn}(\beta_j^*)$. Combining (A.4)–(A.6) yields

$$Q_n(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2 + n\lambda_n\tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta} + o\big(1\big), \tag{A.7}$$

where $o(\cdot)$ is uniformly over all $\boldsymbol{\theta} \in \mathcal{B}_0(n)$.

We now deal with the stochastic part $T_n(\boldsymbol{\theta})$. Define $\mathbf{D} = -\big(\rho_\tau'(\varepsilon_1), \cdots, \rho_\tau'(\varepsilon_n)\big)^T$, $\mathbf{W}_n = \mathbf{Z}_n^T\mathbf{D}$, and

$$R_n(\boldsymbol{\theta}) = \|\rho_\tau(\boldsymbol{\varepsilon} - \mathbf{Z}_n\boldsymbol{\theta})\|_1 - \|\rho_\tau(\boldsymbol{\varepsilon})\|_1 - \mathbf{W}_n^T\boldsymbol{\theta}.$$

Then $E[\mathbf{W}_n^T\boldsymbol{\theta}] = 0$ and

$$T_n(\boldsymbol{\theta}) = \mathbf{W}_n^T\boldsymbol{\theta} + r_n(\boldsymbol{\theta}), \tag{A.8}$$

where $r_n(\boldsymbol{\theta}) = R_n(\boldsymbol{\theta}) - E[R_n(\boldsymbol{\theta})]$. Here, $\mathbf{W}_n^T\boldsymbol{\theta}$ can be regarded as the first order approximation of $\|\rho_\tau(\boldsymbol{\varepsilon} - \mathbf{Z}_n\boldsymbol{\theta})\|_1 - \|\rho_\tau(\boldsymbol{\varepsilon})\|_1$. We next show $r_n(\boldsymbol{\theta})$ is uniformly small. By Lemma 3, there exists a sequence $b_n \to \infty$ such that for any $\epsilon > 0$,

$$P(|r_n(\boldsymbol{\theta})| \geq \epsilon) \leq \exp\big(-C\epsilon b_n s(\log s)\big). \tag{A.9}$$

Define $\lambda_n(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta}) - n\lambda_n\tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta} - \mathbf{W}_n^T\boldsymbol{\theta}$ and $\lambda(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$. Then by definition (A.1), $\lambda_n(\boldsymbol{\theta})$ and $\lambda(\boldsymbol{\theta})$ are both convex functions on the set $B_0(n)$. Furthermore, by definition, we can write $r_n(\boldsymbol{\theta})$ as

$$r_n(\boldsymbol{\theta}) = \lambda_n(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta}) - o(1).$$

Since $\|\boldsymbol{\theta}\|_2 < c_6\sqrt{s}$ for all $\boldsymbol{\theta} \in B_0(n)$, by Condition 1 we have that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in B_0(n)$,

$$\begin{aligned} \big|\lambda(\boldsymbol{\theta}_1) - \lambda(\boldsymbol{\theta}_2)\big| &= \big|(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\big| \\ &\leq \|(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)\|_2\|(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_2 \leq Cs\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty. \end{aligned}$$

Thus, the above result and (A.9) indicate that conditions in Lemma 4 are satisfied. Then, for any compact set $K_s = \{\|\boldsymbol{\theta}\|_2 \leq c_4\sqrt{s}\} \subset B_0(n)$ with $0 < c_4 < c_6$ some constant,

$$\sup_{\boldsymbol{\theta}\in K_s} |r_n(\boldsymbol{\theta})| = o_p(1). \tag{A.10}$$

Combining (A.2), (A.7) and (A.8) we can write that

$$\begin{aligned} G_n(\boldsymbol{\theta}) &= \|\boldsymbol{\theta}\|_2^2 + n\lambda_n\tilde{\mathbf{d}}_0^T\mathbf{V}_n\boldsymbol{\theta} + \mathbf{W}_n^T\boldsymbol{\theta} + r_n(\boldsymbol{\theta}) + o\big(1\big) & \text{(A.11)} \\ &= \|\boldsymbol{\theta} - \boldsymbol{\eta}_n\|_2^2 - \|\boldsymbol{\eta}_n\|_2^2 + r_n(\boldsymbol{\theta}) + o\big(1\big), & \text{(A.12)} \end{aligned}$$

where

$$\boldsymbol{\eta}_n = -\frac{1}{2}\big(n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0 + \mathbf{W}_n\big).$$

By a classic weak convergence result, it is easy to see that

$$\mathbf{c}^T(\mathbf{Z}_n^T\mathbf{Z}_n)^{-1/2}\mathbf{W}_n \xrightarrow{\mathscr{D}} N(0, \tau(1-\tau)),$$

for any $\mathbf{c} \in \mathbf{R}^s$ satisfying $\mathbf{c}^T\mathbf{c} = 1$. It follows immediately that

$$\mathbf{c}^T(\mathbf{Z}_n^T\mathbf{Z}_n)^{-1/2}\Big(\boldsymbol{\eta}_n + \frac{1}{2}n\lambda_n\mathbf{V}_n\tilde{\mathbf{d}}_0\Big) \xrightarrow{\mathscr{D}} N\big(0, \tau(1-\tau)\big). \tag{A.13}$$

We only need to show that the minimizer $\widehat{\boldsymbol{\theta}}$ of $G_n(\boldsymbol{\theta})$ is close to $\boldsymbol{\eta}_n$, i.e., for any $\epsilon > 0$,

$$P\Big(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\eta}_n\|_2 \geq \epsilon\Big) \to 0. \tag{A.14}$$

Hence Theorem 3 will follow from (A.13) and Slutsky's lemma.

We now proceed to prove (A.14). First, let $B_1(n)$ be a ball with center $\boldsymbol{\eta}_n$ and radius ϵ. Since $\mathbf{c}^T\mathbf{W}_n$ has asymptotic normal distribution $N(0,1)$ for any $\mathbf{c} \in \mathbf{R}^s$ with $\mathbf{c}^T\mathbf{c} = 1$, and $n\mathbf{V}_n\mathbf{V}_n^T = (\frac{1}{n}\mathbf{S}^T\mathbf{H}\mathbf{S})^{-1}$ has bounded eigenvalues, by definition, we can bound $\boldsymbol{\eta}_n$ as

$$\begin{aligned}\|\boldsymbol{\eta}_n\|_2 &\leq \frac{1}{2}\big(\|\mathbf{W}_n\|_2 + n\lambda_n\|\mathbf{V}_n^T\tilde{\mathbf{d}}_0\|_2\big)\\ &\leq \frac{1}{2}\big(O_p(\sqrt{s}) + C\lambda_n\sqrt{n}\|\mathbf{d}_0\|_2\big) = \frac{C\sqrt{s}}{2}(1 + O_p(1)),\end{aligned} \tag{A.15}$$

where the last step is by the assumption $\lambda_n\sqrt{n}\|\mathbf{d}_0\|_2 = O(\sqrt{s})$ of the theorem. Since c_6 in the definition of $B_0(n)$ can be chosen to be much larger than $C/2$, it follows that for each fixed s, the compact set $K_s = \{\|\boldsymbol{\theta}\|_2 \leq c_4\sqrt{s}\} \subset B_0(n)$ with c_4 large enough can cover the ball $B_1(n)$ with probability arbitrarily close to 1. Therefore, by (A.10)

$$\Delta_n \equiv \sup_{\boldsymbol{\theta}\in B_1(n)} |r_n(\boldsymbol{\theta})| \leq \sup_{\boldsymbol{\theta}\in K_s} |r_n(\boldsymbol{\theta})| = o_p(1). \tag{A.16}$$

Now, we are ready to prove (A.14). Consider the behavior of $G_n(\boldsymbol{\theta})$ outside of the ball $B_1(n)$. Let $\boldsymbol{\theta} = \boldsymbol{\eta}_n + \kappa\mathbf{u} \in \mathbf{R}^s$ be a vector outside the ball $B_1(n)$, where $\mathbf{u} \in \mathbf{R}^s$ is a unit vector and κ is a constant satisfying $\kappa > \epsilon$, with ϵ the radius of $B_1(n)$. Define $\boldsymbol{\theta}^*$ as the boundary point of $B_1(n)$ that lies on the line segment connecting $\boldsymbol{\eta}_n$ and $\boldsymbol{\theta}$. Then we can write $\boldsymbol{\theta}^* = \boldsymbol{\eta}_n + \epsilon\mathbf{u} = (1 - \epsilon/\kappa)\boldsymbol{\eta}_n + \epsilon\boldsymbol{\theta}/\kappa$. By the convexity of G_n, (A.12) and (A.16),

$$\frac{\epsilon}{\kappa}G_n(\boldsymbol{\theta}) + \big(1 - \frac{\epsilon}{\kappa}\big)G_n(\boldsymbol{\eta}_n) \geq G_n(\boldsymbol{\theta}^*) \geq \epsilon^2 - \|\boldsymbol{\eta}_n\|_2^2 - \Delta_n \geq \epsilon^2 + G_n(\boldsymbol{\eta}_n) - 2\Delta_n.$$

Since $\epsilon < \kappa$, it follows that for large enough n,

$$\inf_{\{\|\boldsymbol{\theta}-\boldsymbol{\eta}_n\|>\epsilon\}} G_n(\boldsymbol{\theta}) \geq G_n(\boldsymbol{\eta}_n) + \frac{\kappa}{\epsilon}[\epsilon^2 - o_p(1)] > G_n(\boldsymbol{\eta}_n). \tag{A.17}$$

This establishes (A.14) and proves Theorem 3.

A.2. Proof of Theorem 5. The proof of Theorem 5 follows from those of Theorems 3 and 4. We use C to denote a generic constant in the proof. By Theorem 4, with asymptotic probability one, there exists a global minimizer $\widehat{\boldsymbol\beta} = (\widehat{\boldsymbol\beta}_1^T, \mathbf{0}^T)^T$ of $\widehat{L}_n(\boldsymbol\beta)$ and $\|\widehat{\boldsymbol\beta}_1 - \boldsymbol\beta_1^*\|_2 \le a_n$.

Next we study the asymptotic normality of $\widehat{\boldsymbol\beta}_1$. Following (A.1) in the proof of Theorem 3, define

$$\tilde{G}_n(\boldsymbol\theta) = \widehat{L}_n(\mathbf{V}_n\boldsymbol\theta + \boldsymbol\beta_1^*, \mathbf{0}) - \widehat{L}_n(\boldsymbol\beta_1^*, \mathbf{0}),$$

where $\mathbf{V}_n$ and $\boldsymbol\theta$ are the same as in the proof of Theorem 3. Then $\widehat{\boldsymbol\theta}_n = \mathbf{V}_n^{-1}(\widehat{\boldsymbol\beta}_1 - \boldsymbol\beta_1^*)$ is a global minimizer of $\tilde{G}_n(\boldsymbol\theta)$. The idea of the proof is to show that $G_n(\boldsymbol\theta)$ in (A.1) and $\tilde{G}_n(\boldsymbol\theta)$ are uniformly close to each other. Since $G_n(\boldsymbol\theta)$ can be well approximated by a sequence of quadratic functions, $\tilde{G}_n(\boldsymbol\theta)$ can be well approximated by the same sequence of quadratic functions. Thus, the minimizer of $\tilde{G}_n(\boldsymbol\theta)$ enjoys the same asymptotic properties as that of $G_n(\boldsymbol\theta)$.

We now proceed to prove that $G_n(\boldsymbol\theta)$ and $\tilde{G}_n(\boldsymbol\theta)$ are uniformly close. To this end, first note that for any $\boldsymbol\beta_1$ with $\|\boldsymbol\beta_1\|_2 < C\sqrt{s}$,

$$|L_n(\boldsymbol\beta_1, \mathbf{0}) - \widehat{L}_n(\boldsymbol\beta_1, \mathbf{0})| \le n\lambda_n \|\boldsymbol\beta_1\|_2 \|\mathbf{d}_0^* - \widehat{\mathbf{d}}_0\|_2 \le Cn\lambda_n\sqrt{s}\|\mathbf{d}_0^* - \widehat{\mathbf{d}}_0\|_2. \tag{A.18}$$

For $1 \le j \le s$, by the mean-value theorem,

$$|d_j^* - \hat{d}_j| = |p'_{\lambda_n}(|\beta_j^*|) - p'_{\lambda_n}(|\widehat{\beta}_j^{ini}|)| = |p''_{\lambda_n}(|\widetilde{\beta}_j|)(\beta_j^* - \widehat{\beta}_j^{ini})|, \tag{A.19}$$

where $\widetilde{\beta}_j$ lies on the segment connecting β_j^* and $\widehat{\beta}_j^{ini}$. By Condition 5 and the triangle inequality, with asymptotic probability one

$$|\widetilde{\beta}_j| \ge |\beta_j^*| - |\widehat{\beta}_j - \beta_j^*| > |\beta_j^*| - C_2\sqrt{s(\log p)/n} > 2^{-1}\min_{j<s}|\beta_j^*|.$$

This together with Condition 6 ensures that $p''_{\lambda_n}(|\widetilde{\beta}_j|) = o_p(s^{-1}\lambda_n^{-1}(n\log p)^{-1/2})$. Thus, in view of (A.19),

$$\|\widehat{\mathbf{d}}_0 - \mathbf{d}_0^*\|_2 \le o_p(s^{-1}\lambda_n^{-1}(n\log p)^{-1/2})\|\boldsymbol\beta_1^* - \widehat{\boldsymbol\beta}_1^{ini}\|_2 \le o_p(s^{-1/2}\lambda_n^{-1}n^{-1}).$$

Since $\|\boldsymbol\theta\|_2 \le C\sqrt{s}$ ensures $\|\boldsymbol\beta_1\|_2 \le C\sqrt{s}$, the above inequality combined with (A.18) entails that

$$\sup_{\boldsymbol\theta \in B_0(n)} |G_n(\boldsymbol\theta) - \tilde{G}_n(\boldsymbol\theta)| = \sup_{\|\boldsymbol\beta_1\|_2 \le C\sqrt{s}} |L_n(\boldsymbol\beta_1, \mathbf{0}) - \widehat{L}_n(\boldsymbol\beta_1, \mathbf{0})| = o_p(1),$$

where $B_0(n)$ is defined in Theorem 3. Therefore, by the above result and (A.17), for any $\boldsymbol{\theta} \in B_0(n)$ and $\|\boldsymbol{\theta} - \boldsymbol{\eta}_n\|_2 > \epsilon$ with $\epsilon > 0$ arbitrarily small,

$$\begin{aligned}\inf_{\{\|\boldsymbol{\theta}-\boldsymbol{\eta}_n\|_2>\epsilon\}} \tilde{G}_n(\boldsymbol{\theta}) &\geq \inf_{\|\boldsymbol{\theta}-\boldsymbol{\eta}_n\|_2>\epsilon} G_n(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta}\in B_0(n)} |G_n(\boldsymbol{\theta}) - \tilde{G}_n(\boldsymbol{\theta})| \\ &\geq G_n(\boldsymbol{\eta}_n) + \frac{\kappa}{\epsilon}[\epsilon^2 - o_p(1)] - o_p(1) \\ &\geq \tilde{G}_n(\boldsymbol{\eta}_n) + \frac{\kappa}{\epsilon}[\epsilon^2 - o_p(1)] - o_p(1).\end{aligned}$$

Then, it follows immediately that the minimizer $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\eta}_n\|_2 \leq \epsilon$ with asymptotic probability one. Thus $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\eta}_n = o_p(1)$. The proof of Theorem 5 is completed.

A.3. Lemmas.

LEMMA 3. *Assume conditions of Theorem 3 hold. Let* $R_{n,i}(\boldsymbol{\theta}) = \rho_\tau(\varepsilon_i - \mathbf{Z}_{n,i}^T\boldsymbol{\theta}) - \rho_\tau(\varepsilon_i) + \rho_\tau'(\varepsilon_i)\mathbf{Z}_{n,i}^T\boldsymbol{\theta}$ *and* $R_n(\boldsymbol{\theta}) = \sum_{i=1}^n R_{ni}(\boldsymbol{\theta})$. *Then for any* $\epsilon > 0$,

$$P(|R_n(\boldsymbol{\theta}) - E[R_n(\boldsymbol{\theta})]| \geq \epsilon) \leq \exp\big(-C\epsilon b_n s^2(\log s)\big),$$

where b_n *is some diverging sequence such that* $b_n s^{7/2}(\log s)\max_i \|\mathbf{Z}_{ni}\|_2 \to 0$, *and* $C > 0$ *is some constant.*

PROOF. Let $\xi_i = R_{n,i}(\boldsymbol{\theta}) - E[R_{n,i}(\boldsymbol{\theta})]$. Then $R_n(\boldsymbol{\theta}) - E[R_n(\boldsymbol{\theta})] = \sum_{i=1}^n \xi_i$. Since $R_{n,i}(\boldsymbol{\theta})$'s are independent, by Markov's inequality we obtain that for any $\epsilon > 0$ and $t > 0$,

$$\begin{aligned}P\Big(R_n(\boldsymbol{\theta}) - E[R_n(\boldsymbol{\theta})] \geq \epsilon\Big) &\leq e^{-t\epsilon}E\Big[\exp\Big(t\sum_{i=1}^n \xi_i\Big)\Big] \\ &= \exp\Big(-t\epsilon - t\sum_{i=1}^n E[R_{n,i}(\boldsymbol{\theta})]\Big)\prod_{i=1}^n E[\exp(tR_{n,i}(\boldsymbol{\theta}))]. \qquad \text{(A.20)}\end{aligned}$$

We next study $E[R_{ni}(\boldsymbol{\theta})]$ and $E\big[\exp\big(tR_{ni}(\boldsymbol{\theta})\big)\big]$ in (A.20). Using a similar argument to that for (A.4) we can prove that

$$E[R_{n,i}(\boldsymbol{\theta})] = E[\rho_\tau(\varepsilon_i - \mathbf{Z}_{n,i}^T\boldsymbol{\theta}) - \rho_\tau(\varepsilon_i)] = f_i(0)(\mathbf{Z}_{n,i}^T\boldsymbol{\theta})^2 + O\big((\mathbf{Z}_{n,i}^T\boldsymbol{\theta})^3\big),$$

where $O(\cdot)$ is uniformly over all i. Thus, it follows from the definition of $\mathbf{Z}_{n,i}$ that

$$t\sum_{i=1}^n E[R_{n,i}(\boldsymbol{\theta})] = t\|\boldsymbol{\theta}\|_2^2 + O\big(t\sum_{i=1}^n (\mathbf{Z}_{n,i}^T\boldsymbol{\theta})^3\big). \qquad \text{(A.21)}$$

Now, we consider $E\big[\exp\big(tR_{ni}(\boldsymbol{\theta})\big)\big]$. If $\mathbf{Z}_{n,i}^T\boldsymbol{\theta} > 0$, then by definition $R_{n,i}(\boldsymbol{\theta}) = (\mathbf{Z}_{n,i}^T\boldsymbol{\theta} - \varepsilon_i)1\{0 \le \varepsilon_i \le \mathbf{Z}_{n,i}^T\boldsymbol{\theta}\}$. By Condition 1 and Taylor expansion it follows that

$$
\begin{aligned}
E\big[\exp\big(tR_{n,i}(\boldsymbol{\theta})\big)\big] &\le 1 + \big(\exp(t\mathbf{Z}_{n,i}^T\boldsymbol{\theta}) - 1\big)P(0 \le \varepsilon_i \le \mathbf{Z}_{n,i}^T\boldsymbol{\theta}) \\
&\le 1 + f_i(0)t|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^2 + O(t^2|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^3).
\end{aligned}
$$

When $\mathbf{Z}_{n,i}^T\boldsymbol{\theta} < 0$, we can get the same result using a similar argument. Since $\prod_{i=1}^n(1+x_i) \le \exp(\sum_{i=1}^n x_i)$ for $x_i > 0$, in view of (A.5) and the above inequality we obtain that

$$
\begin{aligned}
\prod_{i=1}^n E\big[\exp\big(tR_{n,i}(\boldsymbol{\theta})\big)\big] &\le \exp\Big(\sum_{i=1}^n E\big[\exp\big(tR_{n,i}(\boldsymbol{\theta})\big) - 1\big]\Big) \\
&\le \exp\Big(t\|\boldsymbol{\theta}\|_2^2 + O\big(t^2\sum_{i=1}^n|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^3\big)\Big). \qquad \text{(A.22)}
\end{aligned}
$$

Substituting (A.21) and (A.22) into (A.20) gives

$$
P\Big(R_n(\boldsymbol{\theta}) - E[R_n(\boldsymbol{\theta})] \ge \epsilon\Big) \le \exp\Big(-t\epsilon + O\Big(t^2\sum_{i=1}^n|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^3\Big)\Big). \qquad \text{(A.23)}
$$

Choosing $t = 2s^2(\log s)b_n$ with $b_n \to \infty$ such that $b_n s^{7/2}(\log s)\max_i\|\mathbf{Z}_{ni}\|_2 \to 0$, and using similar idea to that for (A.5) we obtain that

$$
t\sum_{i=1}^n|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^3 \le Ct\max_i|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|\sum_{i=1}^n f_i(0)|\mathbf{Z}_{n,i}^T\boldsymbol{\theta}|^2 \le Cts^{3/2}\max_i\|\mathbf{Z}_{n,i}\|_2 \to 0.
$$

Plugging this into (A.23) yields that

$$
P\Big(R_n(\boldsymbol{\theta}) - E[R_n(\boldsymbol{\theta})] \ge \epsilon\Big) \le \exp\big(-C\epsilon b_n s^2(\log s)\big).
$$

Repeating the same argument for $P\Big(R_n(\boldsymbol{\theta}) - E[R_n(\boldsymbol{\theta})] \le -\epsilon\Big)$ completes the proof. $\square$

LEMMA 4. *Let $\lambda(\boldsymbol{\theta})$ be a positive function defined on a convex, open subset $\Theta_s = \{\boldsymbol{\theta} \in \mathbf{R}^s : \|\boldsymbol{\theta}\|_2 < c_6\sqrt{s}\}$ of $\mathbf{R}^s$, and $\{\lambda_n(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_s\}$ be a sequence of random convex functions defined on Θ_s, where $c_6 > 0$ is some constant. Suppose that there exists some $b_n \to \infty$ such that for every $\boldsymbol{\theta} \in \Theta_s$, the following holds for all $\epsilon > 0$*

$$
P\big(|\lambda_n(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})| \ge \epsilon\big) \le c_9\exp\big(-c_7 s^2(\log s)b_n\epsilon\big),
$$

where c_7, c_9 are two positive constants. Let K_s be a compact set in $\mathbf{R}^s$ such that $K_s = \{\|\boldsymbol{\theta}\|_2 \leq c_4\sqrt{s}\} \subset \Theta_s$, where $c_4 < c_6$ is some positive constant. If, for some constant $c_8 > 0$, $|\lambda(\boldsymbol{\theta}_1) - \lambda_2(\boldsymbol{\theta}_2)| \leq c_8 s\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty$ for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_s$, then

$$\sup_{\boldsymbol{\theta}\in K_s} |\lambda_n(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})| = o_p(1).$$

PROOF. The proof is an extension of the convexity lemma in Pollard (1990). The basic idea is to prove that K_s can be covered by a number of cubes, and $\lambda_n(\boldsymbol{\theta})$ and $\lambda(\boldsymbol{\theta})$ are uniformly close over the set of vertices of these cubes. Within each cube, values of both $\lambda_n(\boldsymbol{\theta})$ and $\lambda(\boldsymbol{\theta})$ do not change significantly. Thus $\lambda_n(\boldsymbol{\theta})$ and $\lambda(\boldsymbol{\theta})$ are uniformly close over K_s. In this proof we use C to denote some generic positive constant.

We proceed to prove the lemma. Since $|\lambda(\boldsymbol{\theta}_1) - \lambda(\boldsymbol{\theta}_2)| \leq c_8 s\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty$ for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_s$, it follows that for a fixed $\epsilon > 0$, the function $\lambda(\boldsymbol{\theta})$ varies by less than ϵ/s over each cube of side $\delta \equiv \epsilon/(s^2 c_8)$ that intersects K_s. Note that K_s can be covered by less than $(2c_4\sqrt{s}/\delta)^s = (2c_4c_8s^{5/2})^s$ such cubes. Then in total, there are less than $2^s(2c_4c_8s^{5/2})^s$ vertices. Denote by $\mathfrak{V}_s$ the set of all such vertices whose cubes intersect K_s. Since c_6 can be much larger than c_4 and the edge of each cube, $\delta = \epsilon/(c_8 s^2)$, is small and decreases with s, all vertices in $\mathfrak{V}_s$ fall in Θ_s as well. Thus by the pointwise convergence assumption in the Lemma, it is easy to derive that for any $\epsilon > 0$, as $b_n \to \infty$,

$$P\Big(\max_{\boldsymbol{\theta}\in\mathfrak{V}_s} |\lambda_n(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})| \geq \epsilon/s\Big) \leq c_6 \exp\big(Cs\log(Cs) - c_7 s(\log s) b_n \epsilon\big) \to 0.$$

Therefore,

$$M_n \equiv \max_{\boldsymbol{\theta}\in\mathfrak{V}_s} |\lambda_n(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})| = o_p(\epsilon/s). \tag{A.24}$$

For any $\boldsymbol{\theta} \in K_s$, it will fall into a cube and thus can be written as the convex combination of this cube's vertices $\{\boldsymbol{\theta}_i\}$ in $\mathfrak{V}_s$, that is, $\boldsymbol{\theta} = \sum_i \alpha_i \boldsymbol{\theta}_i$ with $\alpha_i \in [0, 1)$. Then by the convexity of $\lambda_n(\boldsymbol{\theta})$ and (A.24),

$$\begin{aligned}\lambda_n(\boldsymbol{\theta}) &\leq \sum \alpha_i \lambda_n(\boldsymbol{\theta}_i) \leq \sum_i \alpha_i\{|\lambda_n(\boldsymbol{\theta}_i) - \lambda(\boldsymbol{\theta}_i)| + |\lambda(\boldsymbol{\theta}_i) - \lambda(\boldsymbol{\theta})| + \lambda(\boldsymbol{\theta})\} \\ &\leq \max_i |\lambda_n(\boldsymbol{\theta}_i) - \lambda(\boldsymbol{\theta}_i)| + \epsilon/s + \lambda(\boldsymbol{\theta}) \leq M_n + \epsilon/s + \lambda(\boldsymbol{\theta}).\end{aligned}$$

Thus,

$$P\Big(\sup_{\boldsymbol{\theta}\in K_s} (\lambda_n(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})) > 2\epsilon/s\Big) \to 0. \tag{A.25}$$

Next we prove the lower bound. Each $\boldsymbol{\theta}$ in K_s lies within a δ-cube with a vertex $\boldsymbol{\theta}_0$ in $\mathfrak{D}_s$, thus

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \sum_{i=1}^{s} \delta_i \mathbf{e}_i \text{ with } |\delta_i| < \delta,$$

where $\mathbf{e}_1, \cdots, \mathbf{e}_s$ denote s coordinate directions. Without loss of generality suppose $0 \leq \delta_i < \delta$ for each i. Define $\boldsymbol{\theta}_i$ to be the vertex $\boldsymbol{\theta}_0 - \delta \mathbf{e}_i$ in $\mathfrak{D}_s$. Then $\boldsymbol{\theta}_0$ can be written as a convex combination of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_i$:

$$\boldsymbol{\theta}_0 = \frac{\delta}{\delta + \sum_j \delta_j} \boldsymbol{\theta} + \sum_i \frac{\delta_i}{\delta + \sum_j \delta_j} \boldsymbol{\theta}_i.$$

Denote by $a = \frac{\delta}{\delta + \sum_j \delta_j}$ and $a_i = \frac{\delta_i}{\delta + \sum_j \delta_j}$. Since $0 \leq \delta_j < \delta$, it follows that the coefficient for $\boldsymbol{\theta}$ can be bounded as $a \geq \frac{1}{1+s}$. Since $\lambda_n(\cdot)$ is convex, by (A.24) we obtain that

$$\begin{aligned} a\lambda_n(\boldsymbol{\theta}) &\geq \lambda_n(\boldsymbol{\theta}_0) - \sum_i a_i \lambda_n(\boldsymbol{\theta}_i) \geq \lambda(\boldsymbol{\theta}_0) - \sum_i a_i \lambda(\boldsymbol{\theta}_i) - 2M_n \\ &\geq \lambda(\boldsymbol{\theta}) - \frac{\epsilon}{s} - \sum_i a_i (\lambda(\boldsymbol{\theta}) + \frac{\epsilon}{s}) - 2M_n \geq a\lambda(\boldsymbol{\theta}) - \frac{2\epsilon}{s} - 2M_n. \end{aligned}$$

Since $a > (1+s)^{-1}$, it follows from the above inequality and (A.24) that

$$\lambda_n(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta}) \geq (-\frac{2\epsilon}{s} - 2M_n)(s+1) \geq -\frac{2\epsilon(s+1)}{s} - o(1).$$

Therefore,

$$P\Big(\inf_{\boldsymbol{\theta} \in K_s} (\lambda_n(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})) < -3(1+s)\frac{\epsilon}{s}\Big) \to 0. \tag{A.26}$$

Since we can choose ϵ arbitrarily small, the uniform convergence result follows easily by combining (A.25) with (A.26). $\square$

A.4. Proof of Proposition 1. Since s is finite, the summation of the probability below is of the same order as the maximum of the probability below. The distribution result (equation (3.1) in the main paper) entails that

$$\begin{aligned} P(\|\frac{1}{n}\mathbf{S}^T \boldsymbol{\varepsilon}\|_\infty > z) = P(\max_{1 \leq j \leq s} |\tilde{\mathbf{x}}_j^T \boldsymbol{\varepsilon}| > nz) &\sim \sum_{j=1}^{s} P(|\tilde{\mathbf{x}}_j^T \boldsymbol{\varepsilon}| > Cnz) \\ &\sim \sum_{j=1}^{s} \|\tilde{\mathbf{x}}_j\|_\alpha^\alpha c_\alpha (Cnz)^{-\alpha}, \end{aligned}$$

where $C > 0$ is some generic constant. For any sequence $\tilde{b}_n$ such that $\tilde{b}_n \to \infty$, by letting $z = n^{-1} b_n \tilde{b}_n$ with $b_n = (\sum_{j=1}^s \|\tilde{\mathbf{x}}_j\|_\alpha^\alpha)^{1/\alpha}$, we have that

$$P(\|n^{-1}\mathbf{S}^T\boldsymbol{\varepsilon}\|_\infty > n^{-1}b_n\tilde{b}_n) \to 0.$$

In other words, $\|n^{-1}\mathbf{S}^T\boldsymbol{\varepsilon}\|_\infty = O_p(n^{-1}b_n)$. Hence, for the following condition, $\widetilde{\boldsymbol{\beta}}_{\mathcal{M}^*} + \lambda_n \mathrm{sgn}(\widetilde{\boldsymbol{\beta}}_{\mathcal{M}^*}) = \boldsymbol{\beta}^*_{\mathcal{M}^*} + n^{-1}\mathbf{S}^T\boldsymbol{\varepsilon}$, to have a solution $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \cdots, \widetilde{\beta}_s)^T$ with $\widetilde{\beta}_j > 0$ for all $j = 1, \cdots, s$, the necessary conditions are $\beta_0 n b_n^{-1} \to \infty$ and $\lambda_n < \beta_0$. Combining these conditions we have

$$\lambda_n < \beta_0 \equiv n^{-1} b_n \tilde{b}_n, \tag{A.27}$$

with some diverging sequence $\tilde{b}_n$.

We next check if the following condition is satisfied, $\|\mathbf{Q}^T\boldsymbol{\varepsilon}\|_\infty \le n\lambda_n$. Combining (3.1) and (A.27) ensures that for $j > s$,

$$P(|\tilde{\mathbf{x}}_j^T\boldsymbol{\varepsilon}| > n\lambda_n) \ge P(|\tilde{\mathbf{x}}_j^T\boldsymbol{\varepsilon}| > b_n\tilde{b}_n) \sim \|\tilde{\mathbf{x}}_j\|_\alpha^\alpha c_\alpha (b_n\tilde{b}_n)^{-\alpha}.$$

Since we assumed $\mathrm{supp}(\tilde{\mathbf{x}}_j) \cap \mathrm{supp}(\tilde{\mathbf{x}}_i) = \varnothing$ for any $i, j \in \{s+1, \cdots, p\}$, it follows that $\mathbf{Q}^T\boldsymbol{\varepsilon}$ is a vector of independent random variables with components $\tilde{\mathbf{x}}_j^T\boldsymbol{\varepsilon}$. Then,

$$P\Big(\|\mathbf{Q}^T\boldsymbol{\varepsilon}\|_\infty > n\lambda_n\Big) = 1 - P\Big(\|\mathbf{Q}^T\boldsymbol{\varepsilon}\|_\infty \le n\lambda_n\Big) = 1 - \prod_{j>s}\big(1 - P(|\tilde{\mathbf{x}}_j^T\boldsymbol{\varepsilon}| > n\lambda_n)\big)$$
$$\ge 1 - \prod_{j>s}\big(1 - C\|\tilde{\mathbf{x}}_j\|_\alpha^\alpha (b_n\tilde{b}_n)^{-\alpha}\big) = 1 - \exp\Big(\sum_{j>s}\log\big(1 - C\|\tilde{\mathbf{x}}_j\|_\alpha^\alpha (b_n\tilde{b}_n)^{-\alpha}\big)\Big).$$

Since $\log(1+x) \le x$ for all $x > -1$, we have that

$$\sum_{j>s}\log\big(1 - C\|\tilde{\mathbf{x}}_j\|_\alpha^\alpha (b_n\tilde{b}_n)^{-\alpha}\big) \le -C\sum_{j>s}\|\tilde{\mathbf{x}}_j\|_\alpha^\alpha (b_n\tilde{b}_n)^{-\alpha}.$$

Combining the above two inequalities, if $\sum_{j>s}\|\tilde{\mathbf{x}}_j\|_\alpha^\alpha (b_n\tilde{b}_n)^{-\alpha} \to c_0/C \in (0, \infty]$, then

$$P\Big(\|\mathbf{Q}^T\boldsymbol{\varepsilon}\|_\infty > n\lambda_n\Big) \ge 1 - e^{-c_0}.$$

That is, with probability at least $1 - e^{-c_0}$,

$$\|\mathbf{Q}^T\boldsymbol{\varepsilon}\|_\infty \le n\lambda_n,$$

fails to hold and Lasso does not have the model selection oracle property. In fact, if $\tilde{b}_n \le O(n^{-1}(\sum_{j>s}\|\tilde{\mathbf{x}}_j\|_\alpha^\alpha)^{1/\alpha} b_n^{-1})$, or equivalently by (A.27),

$$\beta_0 \le O(n^{-1}\big(\sum_{j>s}\|\tilde{\mathbf{x}}_j\|_\alpha^\alpha)^{1/\alpha}\big),$$

then $c_0/C \in (0, \infty]$. In other words, unless we have

$$n\beta_0\{\sum_{j>s} \|\tilde{\mathbf{x}}_j\|_\alpha^\alpha\}^{-1/\alpha} \to \infty, \tag{A.28}$$

Lasso is not able to recover the true model and the correct sign.

Finally, since $\|\tilde{\mathbf{x}}_j\|_2 = \sqrt{n}$, $|\text{supp}(\tilde{\mathbf{x}}_j)| = O(n^{1/2})$ and $\max_{ij} |x_{ij}| = O(n^{1/4})$, the nonzero components of $\mathbf{Q}$ are all of the same order $O(n^{1/4})$. Consequently, $\|\tilde{\mathbf{x}}_j\|_\alpha^\alpha$ must be all of the same order $O(n^{(2+\alpha)/(4)})$.

Then, $n\{\sum_{j>s} \|\tilde{\mathbf{x}}_j\|_\alpha^\alpha\}^{-1/\alpha} = O\left(n^{\frac{3}{4}-\frac{1}{\alpha}}\right)$ and the condition (A.28) above becomes $n^{\frac{3}{4}-\frac{1}{\alpha}}\beta_0 \to \infty$.

APPENDIX B: REAL DATA EXAMPLE

In this section, we use expression quantitative trait locus (eQTL) mapping to illustrate the performance of R-Lasso and AR-Lasso. eQTL studies aim at finding the variations of genotype in a certain part of a chromosome that are associated with the gene expression levels.

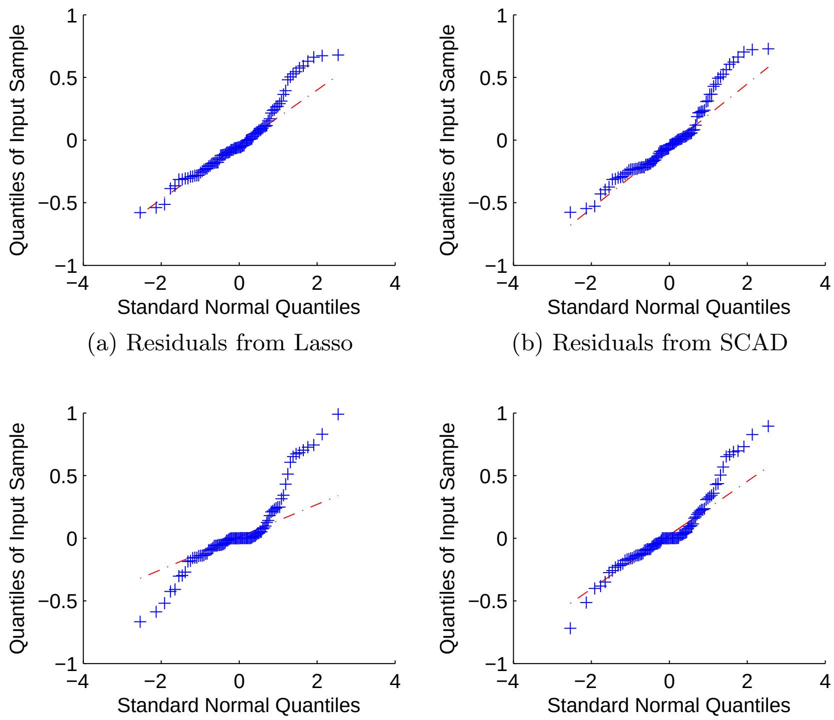
In this study, we conducted a cis-eQTL mapping for the gene *CHRNA6*, cholinergic receptor, nicotinic, alpha 6. *CHRNA6* is located on the 8th chromosome, in the cytogenetic location 8p11. *CHRNA6* is thought to be related to activation of dopamine releasing neurons with nicotine (Thorgeirsson et al., 2010). Therefore, *CHRNA6* has been the subject of many nicotine addiction studies on people with western European heritage (Saccone et al., 2009; Thorgeirsson et al., 2010).

The data are from 90 individuals from the international 'HapMap' project (The International HapMap Consortium, 2005), all with western Europe ancestry. The data are available on ftp://ftp.sanger.ac.uk/pub/genevar/. The normalized expression data was generated with an Illumina Sentrix Human-6 Expression Bead Chip (Stranger et al., 2007). The SNPs under investigation are located at 1 megabase upstream and downstream of the transcription start site (TSS) of *CHRNA6*; in this range, there were 554 SNPs. The additive coding for SNPs was employed, with 0, 1 and 2 representing the major, heterozygous, and minor populations, respectively. We further screened the SNPs using a variation of the independent screening method (SIS) of Fan and Lv (2008). We kept the top 100 SNPs that had correlation with the gene expression levels. Finally, we applied Lasso, SCAD, R-Lasso and AR-Lasso to the screened variables. The quantile parameter, τ was set to 0.5 for R-Lasso and AR-Lasso, corresponding to the median regression. The tuning parameter for all methods was chosen using five-fold cross validation. The selected SNPs as well as their regression coefficients and distances from the main transcription site are given in Table 1.

Table 1
Selected SNPs for the eQTL study.

SNP	*Lasso*	*SCAD*	*R-Lasso*	*AR-Lasso*	*Distance from TSS (in kb)*
rs7823138			-0.0046		-963
rs10090395			0.0513	0.1213	-941
rs3739368			0.0684		-921
rs4737019			0.0331		-889
rs7004640	-0.0170		-0.0114		-872
rs4737023				-0.0124	-849
rs10504049	-0.0216	-0.0096	-0.0082	-0.0505	-800
rs11990460			-0.0918		-769
rs6996712			0.0853	0.0139	-694
rs4466388	0.1214	0.1676	0.1603	0.1299	-681
rs4736825	0.0504			0.1255	-653
rs7819109			0.0716		-564
rs7012976			-0.0925		-529
rs6474389	0.0155				-420
rs3136797				-0.0124	-381
rs12542076				-0.0167	-247
rs13281070			0.0615		-233
rs5024226	0.0155				-93
rs4305884	0.0155				-89
rs10958726			0.0513	0.1213	-18
rs6985527	-0.0170		-0.0114		54
rs11995681	-0.0216	-0.0096	-0.0082	-0.0505	89
rs7818669	0.0538	0.0114	0.1056	0.1013	123
rs11775022			0.0502		138
rs10092934	0.0155				468
rs7016102	0.0770		0.0856	0.0293	538
rs11776934			0.0853	0.0139	590
rs12545574	-0.0363		-0.0942	-0.0093	749
rs9298634	0.0467	0.0171		0.0753	751
rs4737107	0.0222				780
rs10098088			0.0363		809

It is seen that robust regression methods (R-Lasso and AR-Lasso) found more of the variables to be significant. R-Lasso and AR-Lasso selected 21 and 15 variables, respectively, whereas Lasso and SCAD only found 15 and 5 of the variables to be significant. Only 4 SNPs were included in all of the models, (rs10504049, rs4466388, rs7818669, rs708190). Furthermore, none of these SNPs were covered in the previous study (Saccone et al., 2009). We speculate that the difference is due to the fact that the previous studies focused on SNPs that are only 50 kb upstream and downstream. Additionally, these studies did not consider multiple regression which makes significant use of the correlation structure in the data. In addition, among all the SNPs



(a) Residuals from Lasso (b) Residuals from SCAD

(c) Residuals from R-Lasso (d) Residuals from AR-Lasso

Fig 1: QQ-plots of residuals for different methods in the eQTL study

that are chosen by these four methods, only one of them (rs10958726) appears in the paper by Saccone et al. (2009), and only R-Lasso and AR-Lasso found this SNP to be important.

As it was observed in the finite sample simulations, SCAD and AR-Lasso consistently chose a smaller set of variables than their counterparts. Furthermore, almost two thirds of the selected SNPs lie to the left of the transcription site.

We would also like to note that the residuals from the fitted regressions had very heavy right tails. This suggests that, at least for this particular eQTL study, it is a lot more reasonable to use methods based on quantile regression. The QQ-plots of the residuals from different regression methods are shown in Figure 1.

REFERENCES

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70** 849–911.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437** 1299–1320.

Nolan, J. P. (2012). Stable Distributions - Models for Heavy Tailed Data. Birkhauser. (In progress, Chapter 1 online at academic2.american.edu/∼jpnolan).

Pollard, D. (1990). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199.

Saccone, N.L., Saccone, S.F., Hinrichs, A.L., Stitzel, J.A., Duan, W., Pergadia, M.L., Agrawal, A., Breslau, N., Grucza, R.A., Hatsukami, D., Johnson, E.O., Madden, P.A.F., Swan, G.E., Wang, J.C., Goate, A.M., Rice, J.P., Bierut, L.J. (2009). Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (*CHRN*) genes. *Am. J. Med. Genet. Part B* **150** B 453-466.

Stranger, B., Forrest, S., Dunning, M., Ingle, C., Beazley, C., Thorne, N., Redon, R., Bird, C., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S., Tavar, S., Deloukas, P., Hurles, M. and Dermitzakis, E. (2007). Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science* **315 (5813)** 848–853.

Thorgeirsson, T. et al. (2010). Sequence variants at *CHRNB3-CHRNA6* and *CYP2A6* affect smoking behavior. *Nat. Genet.* **42** 448-453.

J. Fan
Department of Operations Research and Fin Eng
Princeton University
Princeton, New Jersey 08544
USA
E-mail: jqfan@princeton.edu

Y. Fan
Data Sciences and Operations Department
Marshall School of Business
University of Southern California
Los Angeles, California 90089
USA
E-mail: fanyingy@marshall.usc.edu

E. Barut
IBM T.J. Watson Research Center
Yorktown Heights, New York 10598
USA
E-mail: abarut@us.ibm.com