Supplementary Material for "High Dimensional Variable Selection with Reciprocal L_1-Regularization"

Qifan Song and Faming Liang

This material is organized as follows. Section 1 gives the proofs of Theorem 2.1 and its extensions. Section 2 gives the proof of Theorem 3.1. Section 2 gives the proof of Equation (16) of the main text. Section 3 gives a brief review of the SAMC and SAA algorithms. Section 4 describes the proposals used by SAA for rLasso. Section 5 discusses some implementation issues of SAA for rLasso.

1 Proof of Theorem 2.1 and its extensions

Proof. We first prove the normality part. Using the same notation as in Knight and Fu (2000), we define the following class of functions $\{V_n : \mathbb{R}^p \to \mathbb{R}\}_{n=1}^{\infty}$. For any $\boldsymbol{u} \in \mathbb{R}^p$,

$$\begin{aligned} V_n(\boldsymbol{u}) &= L(\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n}) - L(\boldsymbol{\beta}^*) \\ &= \sum_i^n [(\epsilon_i - \boldsymbol{u}^T \boldsymbol{x}_i/\sqrt{n})^2 - \epsilon_i^2] + \sum_{j=1}^p \left(P_\lambda(\beta_j^* + u_j/\sqrt{n}) - P_\lambda(\beta_j^*)\right) \\ &= (I) + (II), \end{aligned}$$

where $L(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + P_\lambda(\boldsymbol{\beta})$, and ϵ_i denotes the ith element of $\boldsymbol{\epsilon}$ as defined in Equation (1) of the main text. It is easy to see that the minimum $V_n(\boldsymbol{u})$ is attained at $\hat{\boldsymbol{u}}_n = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, where $\hat{\boldsymbol{\beta}}$ is the minimizer of $L(\boldsymbol{\beta})$.

By Slutsky's and continuous mapping theorem, we have

$$(I) = \boldsymbol{u}^T \left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{u} - 2\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}}{\sqrt{n}}\boldsymbol{u} \xrightarrow{d} \boldsymbol{u}^T\Sigma\boldsymbol{u} - 2\boldsymbol{u}^T\boldsymbol{W},$$

where Σ is defined in Equation (3) of the main text, $\boldsymbol{\epsilon}^T\boldsymbol{X}/\sqrt{n} \xrightarrow{d} \boldsymbol{W}$, and $\boldsymbol{W}$ is a normal random vector with mean zero and covariance matrix $\sigma^2\Sigma$. Then it is easy to show

$$(I) \xrightarrow{e-d} \boldsymbol{u}^T\Sigma\boldsymbol{u} - 2\boldsymbol{u}^T\boldsymbol{W},$$

by Theorem 5 of Knight (1999), where $\xrightarrow{e-d}$ denotes epi-convergence in distribution for a sequence of random lower-semicontinuous functions.

For the second term of $V_n(\boldsymbol{u})$, we show that it can converge uniformly over any compact set $\boldsymbol{U} \in \mathbb{R}^p$:

$$(II) = P_\lambda(\beta_j^* + u_j/\sqrt{n}) - P_\lambda(\beta_j^*) = \begin{cases} P_\lambda(u_j/\sqrt{n}) \\ 0 \\ u_j P'_\lambda(\tilde{\beta}_j^*)/\sqrt{n} \end{cases} \rightrightarrows \begin{cases} \infty, & \text{if } \beta_j^* = 0, \quad u_j \neq 0, \\ 0, & \text{if } \beta_j^* = 0, \quad u_j = 0, \\ 0, & \text{if } \beta_j^* \neq 0, \end{cases} \tag{1}$$

where $\tilde{\beta}_j^*$ is some value between β_j^* and $\beta_j^* + u_j/\sqrt{n}$, and "$\rightrightarrows$" denotes uniform convergence as n increases. Let

$$V(\boldsymbol{u}) = \begin{cases} -2\boldsymbol{u}_{\boldsymbol{t}}^T \boldsymbol{W}_{\boldsymbol{t}} + \boldsymbol{u}_{\boldsymbol{t}}^T \Sigma_{\boldsymbol{t}} \boldsymbol{u}_{\boldsymbol{t}}, & \text{if } u_j = 0 \ \forall j \notin \boldsymbol{t}, \\ \infty, & \text{Otherwise.} \end{cases}$$

Then, by Lemma 1 of Pflug (1995), we have $V_n(\boldsymbol{u}) \xrightarrow{e-d} V(\boldsymbol{u})$, and $V(\boldsymbol{u})$ has the unique minimum $\hat{\boldsymbol{u}} = (\Sigma_{\boldsymbol{t}}^{-1} \boldsymbol{W}_{\boldsymbol{t}}, 0)^T$. Since $\boldsymbol{W}_{\boldsymbol{t}}$, the subvector of $\boldsymbol{W}$, follows $\mathrm{N}(0, \sigma^2 \Sigma_{\boldsymbol{t}})$, we have $\hat{\boldsymbol{u}}_{\boldsymbol{t}} \sim \mathrm{N}(0, \sigma^2 \Sigma_{\boldsymbol{t}}^{-1})$.

To show $\hat{\boldsymbol{u}}_n \to^d \hat{\boldsymbol{u}}$, it is sufficient to show that $\hat{\boldsymbol{u}}_n = O_p(1)$ (see Theorem 1 of Knight (1999)), where $O_p(1)$ denotes bounded in probability. Note that

$$V_n(\boldsymbol{u}) \geq \boldsymbol{u}^T \left(\frac{1}{n} \boldsymbol{X}^T \boldsymbol{X} \right) \boldsymbol{u} - 2 \frac{\boldsymbol{\epsilon}^T \boldsymbol{X}}{\sqrt{n}} \boldsymbol{u} - \sum_{j=1}^{p} P_\lambda(\beta_j^*) \triangleq \tilde{V}_n(\boldsymbol{u}).$$

Since $0 = V_n(0) \geq V_n(\hat{\boldsymbol{u}}_n) \geq \tilde{V}_n(\hat{\boldsymbol{u}}_n)$, $\tilde{V}_n(\boldsymbol{u})$ is convex, $\arg\min(\tilde{V}_n(\boldsymbol{u})) = O_p(1)$, and the eigenvalues of $\boldsymbol{X}^T \boldsymbol{X}/n$ is $O_p(1)$, it follows that $\hat{\boldsymbol{u}}_n = O_p(1)$. For more details of epi-convergence in distribution and limiting distribution of argmin estimators, see Pflug (1994, 1995), Geyer (1994, 1996) and Knight (1999,2001).

We now prove the model consistency part. For any $j \in \boldsymbol{t}$, the asymptotic normality result implies that $\hat{\beta}_j \xrightarrow{p} \beta_j^*$; which further implies

$$P(j \in \boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) | j \in \boldsymbol{t}) \to 1, \quad \text{as } n \to \infty. \tag{2}$$

For any $j \notin \boldsymbol{t}$, the asymptotic normality result implies $P(|(\hat{\boldsymbol{u}}_n)_j| < \delta) \to 1$ for any sufficiently small $\delta > 0$. In addition, we have

$$\begin{aligned} & P\{(\hat{\boldsymbol{u}}_n)_j \neq 0, |(\hat{\boldsymbol{u}}_n)_j| < \delta\} \leq P\{\inf_{(\hat{\boldsymbol{u}}_n)_j \neq 0} V_n(\boldsymbol{u}) \leq V_n(0), |(\hat{\boldsymbol{u}}_n)_j| < \delta\} \\ < & P\left\{ -\left(\frac{\boldsymbol{\epsilon}^T \boldsymbol{X}}{\sqrt{n}} \right)^T \left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \right)^{-1} \left(\frac{\boldsymbol{\epsilon}^T \boldsymbol{X}}{\sqrt{n}} \right) - \sum_{i=1}^{p} P_\lambda(\beta_i^*) + P_\lambda(\delta/\sqrt{n}) \leq 0 \right\} \\ \to & P\left\{ P_\lambda(\delta/\sqrt{n}) - \sum_{i=1}^{p} P_\lambda(|\beta_i^*|) \leq \frac{1}{\sigma^2} \chi_p^2 \right\} \to 0, \quad \text{as } n \to \infty, \end{aligned} \tag{3}$$

where the last row follows from the asymptotics $\left(\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}}{\sqrt{n}}\right)^T\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}}{\sqrt{n}}\right) \xrightarrow{d} \chi_p^2/\sigma^2$ (by Slutsky's theorem and continuous mapping theorem). Note that in this case, we have $\beta_j^* = 0$ and $P_\lambda(\delta/\sqrt{n}) \to \infty$. Therefore,

$$P\{(\hat{\boldsymbol{u}}_n)_j = 0\} \geq P\{(\hat{\boldsymbol{u}}_n)_j = 0, |(\hat{\boldsymbol{u}}_n)_j| < \delta\} = P(|(\hat{\boldsymbol{u}}_n)_j| < \delta) - P\{(\hat{\boldsymbol{u}}_n)_j \neq 0, |(\hat{\boldsymbol{u}}_n)_j| < \delta\} \to 1,$$

which implies

$$P(j \notin \boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) | j \notin \boldsymbol{t}) \to 1, \quad \text{as } n \to \infty. \tag{4}$$

The consistency of the model selection can then be concluded by combining (2) and (4). □

Extension of Theorem 2.1

Corollary 1. *Assume conditions A_1 ($P_{\lambda_n}(0) = 0$), A_4 and*

(A_2') $P_{\lambda_n}(\cdot)$ is symmetric; $P_{\lambda_n}(\cdot/\sqrt{n})$ uniformly converges to ∞ on $(0, U]$ for any $U \in \mathbb{R}$; $\lim_n P_{\lambda_n}(\delta/\sqrt{n})/P_{\lambda_n}(\beta) = \infty$ and $\lim_n P_{\lambda_n}(\beta)/n = 0$ for any fixed δ and β.

(A_3') $P_{\lambda_n}(\cdot)$ is continuously differentiable on $\mathbb{R}\backslash\{0\}$ and $\lim_n P'_{\lambda_n}(\beta)/\sqrt{n} = 0$ for any $\beta \neq 0$.

are satisfied, then the results of Theorem 2.1 still hold.

Proof. The proof follows that of Theorem 2.1 by replacing λ by λ_n. The equation (1) holds because of condition (A_2') and (A_3'); By condition (A_2'), $P_\lambda(\delta/\sqrt{n}) - |\boldsymbol{t}|P_{\lambda_n}(\min_{i\in\boldsymbol{t}}|\beta_i^*|) \to \infty$, hence (3) holds. We only need to show that $\hat{\boldsymbol{u}}_n = O_p(1)$.

In order to show that $\arg\min V_n(\boldsymbol{u}) = O_p(1)$, it is sufficient to show that for any ϵ, there exists a compact set $M_\epsilon \in \mathbb{R}^p$ such that $P[\min_{\boldsymbol{u}\notin M_\epsilon} V_n(\boldsymbol{u}) > V_n(0) = 0] > 1 - \epsilon$ for sufficiently large n. Since $E(\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{u}/\sqrt{n}) = 0$, $\mathrm{Var}(\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{u}/\sqrt{n}) = \boldsymbol{u}^T(\boldsymbol{X}^T\boldsymbol{X}/n)\boldsymbol{u}$ and Equation (3) of the main text holds, when $|u_i|$'s are sufficiently large we have

$$\boldsymbol{u}^T\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{u} - 2\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}}{\sqrt{n}}\boldsymbol{u} \geq \sum_{j=1}^p \kappa u_j^2/2,$$

with probability greater than $1 - \epsilon$, where κ is the smallest eigenvalue of Σ. Consider the terms in the following summation

$$\sum_{j=1}^p [\kappa u_j^2/2 + P_{\lambda_n}(\beta_j^* + u_j/\sqrt{n}) - P_{\lambda_n}(\beta_j^*)] = \sum_j g_n(u_j, \beta_j^*).$$

If $\beta_j^* = 0$, $g_n(u_j, \beta_j^*) > 0$ for any large u_j. If $\beta_j^* \neq 0$, without losing generality, we assume it positive. For $u_j \in (-2\sqrt{n}\beta_j^*, 0)\backslash\{-\sqrt{n}\beta_j^*\}$, $P_{\lambda_n}(\beta_j^* + u_j/\sqrt{n}) > P_{\lambda_n}(\beta_j^*)$ and $g_n(u_j, \beta_j^*) > 0$; for $u_j > 0$, $g_n(u_j, \beta_j^*) > \kappa u_j^2/2 - u_j|P'_{\lambda_n}(\beta_j^*)|/\sqrt{n}$, by A_3' , it is positive for large n and large u_j; for $u_j < -2\sqrt{n}\beta_j^*$, $g_n(u_j, \beta_j^*) > g_n(-2u_j - 2\sqrt{n}\beta_j^*, \beta_j^*) > 0$ for large n by previous case; for $u_j = -\sqrt{n}\beta_j^*$, $g_n(u_j, \beta_j^*) = n\kappa\beta_j^{*2}/2 - P_{\lambda_n}(\beta_j^*) > 0$ for large n.

Therefore, we have shown that $P[\min_{\boldsymbol{u}\notin M_\epsilon} V_n(\boldsymbol{u}) > V_n(0) = 0] > 1 - \epsilon$. □

2 Proof of Theorem 3.1

To prove Theorem 3.1, we first prove the following lemma.

Lemma 1. *Considering the linear regression (1) of the main text and the following model selection criterion*

$$\hat{\boldsymbol{\beta}} = \arg\min_{|\boldsymbol{\xi}(\boldsymbol{\beta})|\leq r}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|)\}, \tag{5}$$

where $\boldsymbol{\xi}(\boldsymbol{\beta}) = \{i : \beta_i \neq 0\}$ denotes the model corresponding to the vector $\boldsymbol{\beta}$, $|\boldsymbol{\xi}(\boldsymbol{\beta})|$ denote the size of the model $\boldsymbol{\xi}(\boldsymbol{\beta})$, and each column of $\boldsymbol{X}$ has been standardized such that $\|\boldsymbol{x}_i\| = \sqrt{n}$ for $i = 1, \ldots, p$. Suppose that the following conditions are satisfied

(a) $|\boldsymbol{t}| \leq r < n - |\boldsymbol{t}|$;

(b) for any subset model $\boldsymbol{\zeta}$,

$$nl_* \leq \min_{|\boldsymbol{\zeta}|\leq|\boldsymbol{t}|+r} ch_1(\boldsymbol{X}_{\boldsymbol{\zeta}}^T\boldsymbol{X}_{\boldsymbol{\zeta}}) \leq \max_{|\boldsymbol{\zeta}|\leq|\boldsymbol{t}|+r} ch_1'(\boldsymbol{X}_{\boldsymbol{\zeta}}^T\boldsymbol{X}_{\boldsymbol{\zeta}}) \leq nl^*;$$

(c) $P_\lambda\left(2\sqrt{\frac{2\sigma^2(|\boldsymbol{t}|+1)\log(p/e_1)}{nl_*^2} + \frac{|\boldsymbol{t}|a_\lambda}{nl_*}}\right) \geq \sigma^2(2\log(p/e_2) + 1 + 2\sqrt{\log(p/e_2)}) + |\boldsymbol{t}|a_\lambda$;

(d) $\sqrt{nl_*\underline{\beta}^2} - \sigma\sqrt{2\log(rp^r/e_3)} \geq \sqrt{\sigma^2(2r\log\frac{p}{e_2} + r + 2r\sqrt{\log(p/e_2)}) + |\boldsymbol{t}|(c_\lambda + a_\lambda)}$;

(e) $b_\lambda \leq \underline{\beta} - \sigma\frac{\sqrt{2\log(1/e_4)}}{nl_*}$;

where $\boldsymbol{t}$ denotes the true model, $|\boldsymbol{t}|$ denotes the size of $\boldsymbol{t}$, e_1, e_2, e_3 and e_4 are sufficiently small numbers, $\underline{\beta} = \min_{i\in\boldsymbol{t}}\beta_i^*$, and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)$ denotes the true regression coefficient vector. Then

$$Pr\left(\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}) = \boldsymbol{t}, \|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^o\| \leq \sqrt{|\boldsymbol{t}|a_\lambda/nl_*}\right) > 1 - 2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4, \tag{6}$$

where $\hat{\boldsymbol{\beta}}^o$ *is equal to* $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o$ *for the components corresponding to the model* $\boldsymbol{t}$ *and 0 otherwise, and* $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o$ *is the OLS estimator of* $\boldsymbol{\beta}_{\boldsymbol{t}}$*. Furthermore, we have the following upper bound for the mean estimation error,*

$$E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2) \leq \frac{2|\boldsymbol{t}|a_\lambda}{nl_*} + \frac{2|\boldsymbol{t}|\sigma^2}{nl_*} + (2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4)* \left(3\|\boldsymbol{\beta}^*\|^2 + 3\frac{n\sigma^2 + \sum P_\lambda(\beta_j^*)}{nl_*} + \frac{6rl^*\|\boldsymbol{\beta}^*\|^2}{l_*^2} + \frac{6rn\sigma^2}{n^2 l_*^2}\right).$$

Proof. Define

$$L_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|).$$

Let $\boldsymbol{\xi}(\boldsymbol{\beta}) = \{i : \beta_i \neq 0\}$ be the model extractor of $\boldsymbol{\beta}$, and let

$$R_{\boldsymbol{\xi}} = \boldsymbol{y}^T(I - X_{\boldsymbol{\xi}}(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}})\boldsymbol{y}$$

denote the residual sum of squares of the OLS estimator of the model $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\beta})$, where $\boldsymbol{X}_{\boldsymbol{\xi}}$ denotes the submatrix of $\boldsymbol{X}$ with columns corresponding to the predictors selected by $\boldsymbol{\xi}$. Therefore,

$$L_\lambda(\hat{\boldsymbol{\beta}}^o) = R_{\boldsymbol{t}} + \sum_{j=1}^{p} P_\lambda(|\hat{\beta}_j^o|),$$

where $\hat{\beta}_j^o$ denotes the jth element of $\hat{\boldsymbol{\beta}}^o$. Since $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o \sim N(\beta_{\boldsymbol{t}}^*, \sigma^2(\boldsymbol{X}_{\boldsymbol{t}}^T\boldsymbol{X}_{\boldsymbol{t}})^{-1})$, by Theorem 2.1 of Inglot (2010), condition (b) and (e), we have

$$P\left\{L_\lambda(\hat{\boldsymbol{\beta}}^o) < R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \geq 1 - 2|\boldsymbol{t}|e_4. \tag{7}$$

Next, we show that for all $\boldsymbol{\beta}$ with $\boldsymbol{\xi}(\boldsymbol{\beta})$ strictly including the true model $\boldsymbol{t}$,

$$P\left\{\min_{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})\supset\boldsymbol{t},|\boldsymbol{t}|<|\boldsymbol{\xi}(\boldsymbol{\beta})|\leq r} L_\lambda(\boldsymbol{\beta}) > R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \geq 1 - 2e_1 - 2e_2. \tag{8}$$

Since $\boldsymbol{X}$ has been standardized such that each column has a norm of $\sqrt{n}$, $\boldsymbol{X}^T\boldsymbol{\epsilon}$ is a p-vector with each entry following the Gaussian distribution N(0, $n\sigma^2$). Then, by Theorem 2.1 of Inglot (2010),

$$P\left\{|(\boldsymbol{X}^T\boldsymbol{\epsilon})_j| \leq \sqrt{n}\sigma\sqrt{2\log(p/e_1)}, \quad \text{for all } j = 1, \ldots, p\right\} \geq 1 - 2e_1,$$

where $(\boldsymbol{z})_j$ denotes the jth element of the vector $\boldsymbol{z}$.

If $\boldsymbol{\xi} \supset \boldsymbol{t}$, then

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 &= \|\boldsymbol{X}_{\boldsymbol{\xi}}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} - \boldsymbol{X}_{\boldsymbol{\xi}}\beta_{\boldsymbol{\xi}}\|^2 \\ &= \boldsymbol{\epsilon}^T(I - P_{\boldsymbol{\xi}})\boldsymbol{\epsilon} + (\boldsymbol{u}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon})^T\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}}(\boldsymbol{u}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}),\end{aligned} \tag{9}$$

where $\boldsymbol{\beta}^*_{\boldsymbol{\xi}}$ denotes the subvector of $\boldsymbol{\beta}^*$ corresponding to the model $\boldsymbol{\xi}$, $\boldsymbol{u}_{\boldsymbol{\xi}} = \boldsymbol{\beta}_{\boldsymbol{\xi}} - \boldsymbol{\beta}^*_{\boldsymbol{\xi}}$, and $P_{\boldsymbol{\xi}} = \boldsymbol{X}_{\boldsymbol{\xi}}(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T$ is the projection matrix. If $\boldsymbol{\beta}$ is outside the ellipse $\{\boldsymbol{\beta} : \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|^2 + |\boldsymbol{t}|a_\lambda = \|\boldsymbol{\epsilon}\|^2 + |\boldsymbol{t}|a_\lambda\}$, then

$$L_\lambda(\boldsymbol{\beta}) > \|\boldsymbol{\epsilon}\|^2 + |\boldsymbol{t}|a_\lambda + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \geq R_{\boldsymbol{t}} + |\boldsymbol{t}|(a_\lambda + c_\lambda),$$

by the property of the OLS estimator and the conditions (A_1) and (A_5).

If $\boldsymbol{\beta}$ is inside the ellipse, it follows from (9) that

$$\boldsymbol{\epsilon}^T(I - P_{\boldsymbol{\xi}})\boldsymbol{\epsilon} + (\boldsymbol{u}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon})^T\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}}(\boldsymbol{u}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}) \leq \|\boldsymbol{\epsilon}\|^2 + |\boldsymbol{t}|a_\lambda,$$

which implies by condition (b) that

$$\|\boldsymbol{u}_{\boldsymbol{\xi}}\| \leq \|(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}\| + \frac{1}{\sqrt{nl_*}}\sqrt{\boldsymbol{\epsilon}^T P_{\boldsymbol{\xi}}\boldsymbol{\epsilon} + |\boldsymbol{t}|a_\lambda}. \tag{10}$$

When all entries of $\boldsymbol{X}^T\boldsymbol{\epsilon}$ are bounded by $\sqrt{n}\sigma\sqrt{2\log(p/e_1)}$, we have

$$\|\boldsymbol{u}_{\boldsymbol{\xi}}\| \leq 2\sqrt{\frac{2\sigma^2|\boldsymbol{\xi}|\log(p/e_1)}{nl_*^2} + \frac{|\boldsymbol{t}|a_\lambda}{nl_*}}. \tag{11}$$

It is easy to show that $P_\lambda(\sqrt{\cdot})$ is convex and thus

$$\begin{aligned}
\sum_{j=1}^{|\boldsymbol{\xi}|} P_\lambda(|\beta_{\boldsymbol{\xi},j}|) &\geq \sum_{j=1}^{|\boldsymbol{\xi}|} P_\lambda(|\beta^*_{\boldsymbol{\xi},j}| + |u_{\boldsymbol{\xi},j}|) \geq |\boldsymbol{t}|c_\lambda + \sum_{\{j:\beta^*_{\boldsymbol{\xi},j}=0\}} P_\lambda(\sqrt{|u_{\boldsymbol{\xi},j}|^2}) \\
&\geq |\boldsymbol{t}|c_\lambda + (|\boldsymbol{\xi}| - |\boldsymbol{t}|)P_\lambda\left(2\sqrt{\frac{2\sigma^2|\boldsymbol{\xi}|\log(p/e_1)}{(|\boldsymbol{\xi}| - |\boldsymbol{t}|)nl_*^2} + \frac{|\boldsymbol{t}|a_\lambda}{(|\boldsymbol{\xi}| - |\boldsymbol{t}|)nl_*}}\right) \\
&\geq |\boldsymbol{t}|c_\lambda + (|\boldsymbol{\xi}| - |\boldsymbol{t}|)P_\lambda\left(2\sqrt{\frac{2\sigma^2(|\boldsymbol{t}| + 1)\log(p/e_1)}{nl_*^2} + \frac{|\boldsymbol{t}|a_\lambda}{nl_*}}\right),
\end{aligned} \tag{12}$$

where $\beta_{\boldsymbol{\xi},j}$, $\beta^*_{\boldsymbol{\xi},j}$ and $u_{\boldsymbol{\xi},j}$ denote the jth elements of $\boldsymbol{\beta}_{\boldsymbol{\xi}}$, $\boldsymbol{\beta}^*_{\boldsymbol{\xi}}$ and $\boldsymbol{u}_{\boldsymbol{\xi}}$, respectively; the third inequality follows from (11) and the convexity of $P_\lambda(\sqrt{\cdot})$; and the last inequality follows from the facts that both $|\boldsymbol{\xi}|/(|\boldsymbol{\xi}| - |\boldsymbol{t}|)$ and $|\boldsymbol{t}|/(|\boldsymbol{\xi}| - |\boldsymbol{t}|)$ are decreasing functions of $|\boldsymbol{\xi}|$.

In addition, we have

$$\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\xi}}\boldsymbol{\beta}_{\boldsymbol{\xi}}\|^2 \geq R_{\boldsymbol{t}} - (R_{\boldsymbol{t}} - R_{\boldsymbol{\xi}}) = R_{\boldsymbol{t}} - \sigma^2 Z^2_{|\boldsymbol{\xi}|-|\boldsymbol{t}|}(\boldsymbol{\xi}), \tag{13}$$

where $Z^2_{|\boldsymbol{\xi}|-|\boldsymbol{t}|}(\boldsymbol{\xi})$ follows a χ^2-distribution of degree of freedom $|\boldsymbol{\xi}| - |\boldsymbol{t}|$. By Theorem 4.1 of Inglot (2010) and Bonferroni inequality, with probability greater than $1 - \sum_{i=1}^{r-|\boldsymbol{t}|} e_2^i$ (which is greater than $1 - 2e_2$), for all $\boldsymbol{\xi}$ with $\boldsymbol{\xi} \supset \boldsymbol{t}$,

$$Z^2_{|\boldsymbol{\xi}|-|\boldsymbol{t}|}(\boldsymbol{\xi}) \leq 2(|\boldsymbol{\xi}| - |\boldsymbol{t}|)\log(p/e_2) + |\boldsymbol{\xi}| - |\boldsymbol{t}| + 2(|\boldsymbol{\xi}| - |\boldsymbol{t}|)\sqrt{\log(p/e_2)}. \tag{14}$$

Combining (12), (13), (14) and condition (c), one can show (8) by Bonferroni inequality.

Third, we show that for all $\boldsymbol{\beta}$ with $\boldsymbol{\xi}(\boldsymbol{\beta}) \not\supseteq \boldsymbol{t}$,

$$P\left\{\min_{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})\not\supseteq\boldsymbol{t},|\boldsymbol{\xi}(\boldsymbol{\beta})|\le r} L_\lambda(\boldsymbol{\beta}) > R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \ge 1 - 2e_2 - 2e_3. \tag{15}$$

If $\boldsymbol{\xi} \not\supseteq \boldsymbol{t}$, let $\boldsymbol{\zeta} = \boldsymbol{t} \cup \boldsymbol{\xi}$, then

$$L_\lambda(\boldsymbol{\beta}) > R_{\boldsymbol{\xi}} = (R_{\boldsymbol{\xi}} - R_{\boldsymbol{\zeta}}) - (R_{\boldsymbol{t}} - R_{\boldsymbol{\zeta}}) + R_{\boldsymbol{t}}, \tag{16}$$

where $(R_{\boldsymbol{\xi}} - R_{\boldsymbol{\zeta}})/\sigma^2$ is noncentral $\chi^2_{|\boldsymbol{\zeta}|-|\boldsymbol{\xi}|}(C)$ distribution with noncentral parameter

$$C = \boldsymbol{\beta}_{\boldsymbol{t}}^* \boldsymbol{X}_{\boldsymbol{t}}^T (P_{\boldsymbol{\zeta}} - P_{\boldsymbol{\xi}}) \boldsymbol{X}_{\boldsymbol{t}} \boldsymbol{\beta}_{\boldsymbol{t}}^* / \sigma^2 \ge n l_* \underline{\beta}^2 / \sigma^2.$$

If $\sqrt{n l_* \underline{\beta}^2/\sigma^2} > \sqrt{2\log(rp^r/e_3)}$, then by Theorem 2.1 of Inglot (2010), with probability greater then $1 - 2e_3$, for all possible $\boldsymbol{\xi}$ with $\boldsymbol{t} \not\subseteq \boldsymbol{\xi}$,

$$\begin{aligned} R_{\boldsymbol{\xi}} - R_{\boldsymbol{\zeta}} &> \left\{\sqrt{n l_* \underline{\beta}^2} - \sigma\sqrt{2\log(rp^r/e_3)}\right\}^2 \\ &\ge \sigma^2(2r\log(p/e_2) + r + 2r\sqrt{\log(p/e_2)}) + |\boldsymbol{t}|(c_\lambda + a_\lambda). \end{aligned} \tag{17}$$

Combining (16), (17) and(14), one can show (15) by Bonferroni inequality.

Finally, we combine (7), (8) and (15), and conclude that

$$\begin{aligned} P\left\{\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}) = \boldsymbol{t}\right\} &\ge P\left\{L_\lambda(\hat{\boldsymbol{\beta}}^o) < R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \\ &\quad + P\left\{\min_{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})\supset\boldsymbol{t},|\boldsymbol{t}|<|\boldsymbol{\xi}(\boldsymbol{\beta})|\le r} L_\lambda(\boldsymbol{\beta}) > R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \\ &\quad + P\left\{\min_{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})\not\supseteq\boldsymbol{t},|\boldsymbol{\xi}(\boldsymbol{\beta})|\le r} L_\lambda(\beta) > R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} - 2 \\ &\ge 1 - 2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4, \end{aligned}$$

by Bonferroni inequality.

Suppose that $\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}) = \boldsymbol{t}$. Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} = \min_{\{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})=\boldsymbol{t}\}} L_\lambda(\boldsymbol{\beta})$. Then,

$$\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}\|^2 + |\boldsymbol{t}|c_\lambda < R_{\boldsymbol{t}} + |\boldsymbol{t}|(a_\lambda + c_\lambda).$$

It follows from the decomposition $\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}\|^2 = R_{\boldsymbol{t}} + \|\boldsymbol{X}_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o - \boldsymbol{X}_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}\|^2$ that

$$(\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o)^T \boldsymbol{X}_{\boldsymbol{t}}^T \boldsymbol{X}_{\boldsymbol{t}} (\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o) \le |\boldsymbol{t}|a_\lambda,$$

which, by condition (b), implies

$$\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^o\| = \|\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o\| \le \sqrt{|\boldsymbol{t}|a_\lambda/n l_*}.$$

This concludes (6).

Let $\boldsymbol{\xi} = \boldsymbol{\xi}(\hat{\boldsymbol{\beta}})$ for some $\|\boldsymbol{\xi}\| \leq r$, and let $\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} = \min_{\{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})=\boldsymbol{\xi}\}} L_\lambda(\boldsymbol{\beta})$. Consider the case that $\boldsymbol{\xi} \neq \boldsymbol{t}$, then

$$\begin{aligned}&(\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y})^T(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})(\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y})\\ &\leq \|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\xi}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}}\|^2 < L_\lambda(\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}}) \leq L_\lambda(\boldsymbol{\beta}^*) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \sum P_\lambda(\beta_j^*),\end{aligned} \tag{18}$$

where the first inequality follows from the decomposition

$$\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\xi}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}}\|^2 = R_{\boldsymbol{\xi}} + \|\boldsymbol{X}_{\boldsymbol{\xi}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - \boldsymbol{X}_{\boldsymbol{\xi}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}}^o\|^2,$$

and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}}^o = (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y}$ denotes the OLS estimator of $\boldsymbol{\beta}_{\boldsymbol{\xi}}$. Thus, $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y}\|^2 \leq (\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \sum P_\lambda(\beta_j^*))/nl_*$ and $\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y} = \boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}$, where $\|\boldsymbol{X}\boldsymbol{\beta}^*\|^2 \leq nl^*\|\boldsymbol{\beta}^*\|^2$ and each row of $\boldsymbol{X}_{\boldsymbol{\xi}}^T$ has been standardized to have a norm of $\sqrt{n}$. It follows that $(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}\boldsymbol{\beta}^*)_j \leq n\sqrt{l^*}\|\boldsymbol{\beta}^*\|$ for $j = 1, \ldots, |\boldsymbol{\xi}|$. Furthermore,

$$\begin{aligned}\|(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y}\|^2 &\leq \frac{1}{n^2l_*^2}\|\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y}\|^2 \leq \frac{2}{n^2l_*^2}(\|\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}\boldsymbol{\beta}^*\|^2 + \|\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}\|^2)\\ &\leq \frac{2rn^2l^*\|\boldsymbol{\beta}^*\|^2}{n^2l_*^2} + \frac{2\boldsymbol{\epsilon}^T\boldsymbol{X}_{\boldsymbol{\xi}}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}}{n^2l_*^2}.\end{aligned} \tag{19}$$

Following from (19),

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - \boldsymbol{\beta}^*\|^2 &\leq 3\|\boldsymbol{\beta}^*\|^2 + 3\|\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y}\|^2 + 3\|(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{y}\|^2\\ &\leq 3\|\boldsymbol{\beta}^*\|^2 + 3\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \sum P_\lambda(\beta_j^*)}{nl_*} + \frac{6rn^2l^*\|\boldsymbol{\beta}^*\|^2}{n^2l_*^2} + \frac{6\boldsymbol{\epsilon}^T\boldsymbol{X}_{\boldsymbol{\xi}}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}}{n^2l_*^2}.\end{aligned} \tag{20}$$

Combining (6) and (20), we have

$$\begin{aligned}E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2) &\leq \frac{2|\boldsymbol{t}|a_\lambda}{nl_*} + 2E(\|\hat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*\|^2) + (2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4)\\ &\quad\times E\left(3\|\boldsymbol{\beta}^*\|^2 + 3\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \sum P_\lambda(\beta_j^*)}{nl_*} + \frac{6rn^2l^*\|\boldsymbol{\beta}^*\|^2}{n^2l_*^2} + \frac{6\boldsymbol{\epsilon}^T\boldsymbol{X}_{\boldsymbol{\xi}}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}}{n^2l_*^2}\right)\\ &\leq \frac{2|\boldsymbol{t}|a_\lambda}{nl_*} + \frac{2|\boldsymbol{t}|\sigma^2}{nl_*} + (2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4)\\ &\quad\times\left(3\|\boldsymbol{\beta}^*\|^2 + 3\frac{n\sigma^2 + \sum P_\lambda(\beta_j^*)}{nl_*} + \frac{6rl^*\|\boldsymbol{\beta}^*\|^2}{l_*^2} + \frac{6rn\sigma^2}{n^2l_*^2}\right).\end{aligned}$$

This concludes the proof of the lemma. □

Remark:

1. The conditions (c) and (d) look very technical, but can be interpreted intuitively. In order to bring sparsity into the model, the shape of the penalty function around zero is crucial. Traditional penalty functions, such as those used in Lasso, SCAD or MCP, are singular at zero and have the largest derivative at zero, such that the coefficients of false predictors can shrink faster than those of true predictors. rLasso brings sparsity into the model in a different way: By giving a very large penalty around zero (i.e. condition (c)) such that the model cannot afford a small coefficient for the false predictor. Condition (d) restricts the dimensionality and eigen-structure of the design matrix. An arbitrarily large p or an arbitrarily small l_* increases the probability that the linear effect of a true predictor can be almost totally replaced by some combination of false predictors.

2. If, furthermore, there exists a sufficient small number e_5 and the following condition holds

   (f) $(r - |\boldsymbol{t}|)c_\lambda > |\boldsymbol{t}|a_\lambda + \sigma^2(n - |\boldsymbol{t}| + 2\log(1/e_5) + \sqrt{(n - |\boldsymbol{t}|)\log(1/e_5)})$,

   then in probability greater than $1 - e_5$, the following inequality holds

   $$(r + 1)c_\lambda > |\boldsymbol{t}|(c_\lambda + a_\lambda) + R_{\boldsymbol{t}},$$

   which implies that for any $\boldsymbol{\beta}$ with $|\boldsymbol{\xi}(\boldsymbol{\beta})| > r$, $L_\lambda(\boldsymbol{\beta}) > L_\lambda(\hat{\boldsymbol{\beta}}^o)$ holds. Hence the constraint $|\boldsymbol{\xi}(\boldsymbol{\beta})| \leq r$ is automatically satisfied in minimization of $L_\lambda(\boldsymbol{\beta})$. In this case, (5) is equivalent to

   $$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(\beta_j)\}$$

   without the model size constraint.

To prove Theorem 3.1, we let $e_1 = e_2 = e_3 = e_4 = \exp(-K_n)$. Thus, the conditions of Lemma 1 are satisfied when n is sufficiently large, and this concludes the consistency of rLasso for variable selection and parameter estimation.

3 Proof of Equation (16) of the Main Text

If λ is sufficiently large, then $\hat{\boldsymbol{\beta}}_n(\lambda) = 0$. As λ decreases to some threshold value λ_m, $\hat{\boldsymbol{\beta}}_n$ will jump away from zero. Consider the case that only the predictor $\boldsymbol{x}_k$ has the largest absolute marginal correlation with the response variable $\boldsymbol{y}$, i.e., $|\text{cor}(\boldsymbol{y}, \boldsymbol{x}_k)| > max_{j\neq k}|\text{cor}(\boldsymbol{y}, \boldsymbol{x}_j)|$. Then, only $\boldsymbol{x}_k$ will be included into the model, i.e. $\hat{\boldsymbol{\beta}}_n(\lambda_m) = (0, \ldots, 0, \hat{\beta}_k, 0, \ldots, 0)^T$. At this critical

jump point, we have

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (y_i - x_{ik}\hat{\beta}_k)^2 + \lambda_m/|\hat{\beta}_k|. \tag{21}$$

Since the rLasso objective function is convex given the signs of $\boldsymbol{\beta}$, hence we have

$$-2\sum_{i=1}^{n} (y_i - x_{ik}\hat{\beta}_k)x_{ik} - \text{sign}(\hat{\beta}_k)\frac{\lambda_m}{|\hat{\beta}_k|^2} = 0. \tag{22}$$

Combined (21) and (22) with the fact $\text{sign}(\hat{\beta}_k) = \text{sign}(\text{cor}(\boldsymbol{y}, \boldsymbol{x}_k))$, it is easy to derive that

$$\hat{\beta}_k = \frac{4\sum x_{ik} y_i}{3\sum x_{ik}^2}, \quad \lambda_m = \frac{\sum x_{ik}^2 |\hat{\beta}_k|^3}{2}, \tag{23}$$

which concludes the proof.

4 A Brief Review of the SAMC and SAA Algorithm

Suppose that we want to draw samples from the distribution

$$f(x, \tau) = \frac{1}{Z}\exp(-U(x)/\tau), \quad x \in \mathcal{X}, \tag{24}$$

where Z is the normalizing constant, τ is called the temperature, $U(x)$ is called the energy function, and $\mathcal{X}$ is the sample space. Furthermore, suppose that the sample space $\mathcal{X}$ has been partitioned according to the energy function into m disjoint subregions denoted by $E_1 = \{x : U(x) \leq u_1\}$, $E_2 = \{x : u_1 < U(x) \leq u_2\}$, ..., $E_{m-1} = \{x : u_{m-2} < U(x) \leq u_{m-1}\}$ and $E_m = \{x : U(x) > u_{m-1}\}$, where $u_1 < u_2 < \ldots < u_{m-1}$ are pre-specified numbers.

Let $\psi(x, \tau) = \exp(-U(x)/\tau)$ and let $\omega_i(\tau) = \int_{E_i} \psi(x, \tau)dx$ for $i = 1, \ldots, m$. Without loss of generality, we assume $\omega_i > 0$ for all $i = 1, \ldots, m$. As shown in (30), the case for $\omega_i = 0$ is trivial with the estimates of $\log(w_i)$ simply going to $-\infty$. SAMC seeks to sample from the trial distribution

$$f_\omega(x) \propto \sum_{i=1}^{m} \frac{\pi_i \psi(x, \tau)}{\omega_i(\tau)} I(x \in E_i), \tag{25}$$

where π_i's are pre-specified frequency values such that $\pi_i > 0$ for all i and $\sum_{i=1}^{m} \pi_i = 1$. In Liang *et al.* (2007), $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^T$ is called the desired sampling distribution of the subregions. It is easy to see that if $\omega_1(\tau), \ldots, \omega_m(\tau)$ can be well estimated, sampling from $f_\omega(x)$ will result in a "random walk" in the space of subregions (by regarding each subregion as a "point") with each subregion being sampled with a frequency proportional to π_i. Hence, the local-trap problem can be avoided essentially, provided that the sample space has been partitioned appropriately.

SAMC provides a systematic means, as described below, to estimate $\omega_1(\tau), \ldots, \omega_m(\tau)$ under the framework of stochastic approximation (Robbins and Monro, 1951; Benveniste *et al.*, 1990).

Let θ_{ti} denote the working estimate of $\log(\omega_i/\pi_i)$ obtained at iteration t, let $\theta_t = (\theta_{t1}, \ldots, \theta_{tm})^T$, and let $\{a_t\}$ denote a positive, non-increasing sequence satisfying the conditions

$$(i) \quad \sum_{t=1}^{\infty} a_t = \infty, \quad (ii) \quad \sum_{t=1}^{\infty} a_t^{\zeta} < \infty, \tag{26}$$

for some $\zeta \in (1, 2)$. Since $f_\omega(x)$ is invariant with respect to a scale change of $\boldsymbol{\omega}(\tau) = (\omega_1(\tau), \ldots, \omega_m(\tau))^T$, i.e., $f_{c\omega}(x) = f_\omega(x)$ for any number $c > 0$, the domain of θ_t can be restricted to a compact set Θ by adjusting θ_t with a constant vector. As in Liang *et al.* (2007), we set $\Theta = [-10^{100}, 10^{100}]^m$ in this paper, although, as a practical matter, this is equivalent to setting $\Theta = \mathbb{R}^m$. It follows from (30) (presented below) that

$$\lim_{t\to\infty} [\theta_{ti} - \theta_{tj}] = \log(\omega_i(\tau)) - \log(\omega_j(\tau)) - \log(\pi_i) + \log(\pi_j). \tag{27}$$

Hence, the range of Θ also represents the maximum resolution allowed for the estimates of $\omega_1, \ldots, \omega_m$, when $\boldsymbol{\pi}$ is uniform over the subregions.

The SAMC algorithm iterates between the following two steps:

(a) (Sampling) Simulate a sample x_t by a single MH update with the target distribution

$$f_{\theta_t}(x) \propto \sum_{i=1}^{m} \frac{\psi(x, \tau)}{e^{\theta_{ti}}} I(x \in E_i). \tag{28}$$

(b) (Weight updating) Set

$$\theta^* = \theta_t + a_{t+1}(\tilde{\boldsymbol{e}}_t - \boldsymbol{\pi}), \tag{29}$$

where $\tilde{\boldsymbol{e}}_t = (\tilde{e}_{t,1}, \ldots, \tilde{e}_{t,m})$ and $\tilde{e}_{t,i} = 1$ if $x_t \in E_i$ and 0 otherwise. If $\theta^* \in \Theta$, set $\theta_{t+1} = \theta^*$; otherwise, set $\theta_{t+1} = \theta^* + \boldsymbol{c}^*$, where $\boldsymbol{c}^* = (c^*, \ldots, c^*)$ can be an arbitrary constant vector satisfying the condition $\theta^* + \boldsymbol{c}^* \in \Theta$.

Under mild conditions, Liang *et al.* (2007) showed that as $t \to \infty$,

$$\theta_{ti} \to \begin{cases} \text{Const} + \log(\omega_i(\tau)) - \log(\pi_i + \bar{\pi}_0), & \text{if } E_i \neq \emptyset, \\ -\infty. & \text{if } E_i = \emptyset, \end{cases} \tag{30}$$

where $\bar{\pi}_0 = \sum_{j \in \{i: E_i = \emptyset\}} \pi_j/(m - m_0)$ and m_0 is the number of empty subregions, and Const represents an arbitrary constant. The subregion E_i is called an empty subregion if $\omega_i = 0$.

Furthermore, Liang (2009) showed that SAMC is invariant with respect to an importance weight. Mathematically, this can be explained as follows. Let $(x_1, w_1), \ldots, (x_N, w_N)$ denote a set of samples collected at the sampling step of SAMC, where

$$w_t = \sum_{i=1}^{m} e^{\theta_{ti}} I(x_t \in E_i). \tag{31}$$

Let $y_1, \ldots, y_{N'}$ denote the distinct samples among $x_1, \ldots, x_N$. If we generate a random sample Y such that

$$P(Y = y) = \frac{\sum_{t=1}^{N} w_t I(x_t = y)}{\sum_{t=1}^{N} w_t}, \quad y \in \{y_1, \ldots, y_{N'}\},$$

then Y is asymptotically distributed with respect to the density $f(y, \tau)$. This property implies that SAMC can be used as a usual importance sampling algorithm. For any integrable function $h(x)$, the expectation $E_f h(x) = \int h(x) f(x, \tau) dx$ can be estimated by

$$\widehat{E_f h(x)} = \frac{\sum_{t=1}^{N} w_t h(x_t)}{\sum_{t=1}^{N} w_t}, \tag{32}$$

and, as $n \to \infty$, $\widehat{E_f h(x)} \to E_f h(x)$ almost surely for the same reason that the usual importance sampling estimate converges.

Compared to conventional MCMC algorithms, such as the Metropolis-Hastings algorithm, SAMC has a significant advantage in sample space exploration. This is due to its self-adjusting mechanism: If a subregion is visited at iteration t, θ_t will be updated accordingly such that this subregion has a smaller probability to be revisited in the next iteration. Mathematically, if $x_t \in E_i$, then $\theta_{t+1,i} \leftarrow \theta_{t,i} + a_{t+1}(1 - \pi_i)$ and $\theta_{t+1,j} \leftarrow \theta_{t,i} - a_{t+1}\pi_j$ for $j \neq i$. This mechanism makes SAMC essentially immune to the local trap problem and particularly suitable for sampling of high dimensional space.

Asymptotically, SAMC can be used for minimizing the energy function $U(x)$, because SAMC is ergodic. However, when τ is large, $f(x, \tau)$ is quite flat on $\mathcal{X}$, the search for global energy minima will be very inefficient. To serve the purpose of optimization, Liang et al. (2014) proposed the SAA algorithm, which combines SAMC and the simulated annealing algorithm. The SAA algorithm allows τ to be decreasing in a square-root cooling schedule, i.e.,

$$\tau_t = \tau_* + C/\sqrt{t},$$

where τ_* is a tiny value such that $\int_{O(\hat{x})} f(x, \tau_*) \approx 1$, and $O(\hat{x})$ is a small open set around $\hat{x} = \arg\min U(x)$. SAA can be run in the same way as SAMC except that the temperature is

decreasing with iterations in SAA, but fixed in SAMC. Liang et al. (2014) showed that

$$\theta_{ti} \to \begin{cases} \text{Const} + \log(\omega_i(\tau_*)) - \log(\pi_i + \bar{\pi}_0), & \text{if } E_i \neq \emptyset, \\ -\infty. & \text{if } E_i = \emptyset, \end{cases} \tag{33}$$

and the SAA algorithm is able to locate the global minima if τ_* is sufficiently close to zero.

5 Proposal setup for High dimensional rLasso

To minimize the function $L(\boldsymbol{\omega})$, $\boldsymbol{\omega} \in \mathcal{W} = \{-1, 0, 1\}^p$, SAA works by sampling from a sequence of distributions given by

$$f(\boldsymbol{\omega}, \tau_t) \propto \exp(-L(\boldsymbol{\omega})/\tau_t) I(\sum |\omega_i| < r), \quad \boldsymbol{\omega} \in \mathcal{W} = \{-1, 0, 1\}^p,$$

where $\tau_t = \tau_* + C/\sqrt{t}$ for some constants $\tau_* > 0$ and $C > 0$. To sample from the sequence of distributions, we specify four types of moves, birth, death, sign-change and exchange, which are described in sequel as follows.

Let $\boldsymbol{\omega}_t$ denote the current state at iteration t. Let $\xi_t = \{i : \omega_{t,i} \neq 0\}$ denote the set of predictors selected by $\boldsymbol{\omega}_t$, and let $\xi_t^c = \{i : i \notin \xi_t\}$ denote the set of predictors excluded from the set ξ_t. In what follows, we will also refer ξ_t as a model simulated at iteration t. Let $\boldsymbol{\omega}_* = (\omega_{*1}, \ldots, \omega_{*p})$ denote the proposed state.

In the birth step, we randomly select a predictor, say x_i, from the set ξ_t^c and assign $\omega_{*,i}$ a value 1 or -1 with equal probability, and set $\omega_{*,j} = \omega_{t,j}$ for all $j \neq i$. The corresponding proposal probability is

$$P(\omega_{*,i} = \pm 1, \omega_{*,j} = \omega_{t,j}, \text{for } i \in \xi_t^c,\, j \neq i | \text{birth}, \boldsymbol{\omega}_t) = \frac{1}{2|\xi_t^c|}. \tag{34}$$

In the death step, we randomly select a predictor, say x_i, from the set ξ_t and set $\omega_{*,i} = 0$, and set $\omega_{*,j} = \omega_{t,j}$ for all $j \neq i$. The corresponding proposal probability is

$$P(\omega_{*,i} = 0, \omega_{*,j} = \omega_{t,j}, \text{for } i \in \xi_t,\, j \neq i | \text{death}, \boldsymbol{\omega}_t) = \frac{1}{|\xi_t|}. \tag{35}$$

In the exchange step, we randomly select a predictor, say x_i, from the set ξ_t^c, and randomly select another predictor, say x_j, from the set ξ_t, and then form $\boldsymbol{\omega}_*$ by exchanging the values of $\omega_{t,i}$ and $\omega_{t,j}$. The corresponding proposal probability is

$$P(\omega_{*,i} = \omega_{t,j}, \omega_{*,j} = \omega_{t,i}, \omega_{*,k} = \omega_{t,k}, \text{for } i \in \xi_t,\, j \in \xi_t^c,\, k \neq i,\, k \neq j | \text{exchange}, \boldsymbol{\omega}_t) = \frac{1}{|\xi_t||\xi_t^c|}. \tag{36}$$

In the sign-change step, we randomly select a predictor, say x_i, from the set ξ_t and change the sign of $\omega_{t,i}$ (from -1 to 1 or from 1 to -1), and remains other values of $\omega_{t,j}$ unchanged. The corresponding proposal probability is

$$P(\omega_{*,i} = -\omega_{t,i}, \omega_{*,j} = \omega_{t,j}, \text{for } i \in \xi_t, j \neq i | \text{sign-change}, \boldsymbol{\omega}_t) = \frac{1}{|\xi_t|}. \tag{37}$$

Since the death move cannot be performed for the minimal size model, and the birth move cannot be performed for the maximum size models, we specify the following proposal probabilities for the four operators conditioned on $|\xi_t|$, the number of predictors included in the model ξ_t:

$$\begin{cases} P(\text{birth}\big||\xi_t| = 1) = P(\text{sign-change}\big||\xi_t| = 1) = 1/2, \\ P(\text{birth}\big|1 < |\xi_t| < r) = P(\text{death}\big|1 < |\xi_t| < r) = (\text{exchange}\big|1 < |\xi_t| < r) \\ = P(\text{sign-change}\big|1 < |\xi| < r) = 1/4, \\ P(\text{death}\big||\xi_t| = r) = P(\text{sign-change}\big||\xi_t| = r) = 1/2, \end{cases} \tag{38}$$

where $r < p$ denotes the maximum model size considered by the user. Given (34), (35), (36), (37) and (38), the transition probability ratio can be compute accordingly.

6 Implementation Issues of SAA for rLasso

For an efficient implementation of SAA for optimization of rLasso, several issues need to be taken care.

- *Sample space partitioning.* The sample space is partitioned according to the energy function. Given the energy function $L(\boldsymbol{\omega})$, the sample space can be partitioned as follows: $E_1 = \{\boldsymbol{\omega} : L(\boldsymbol{\omega}) \leq u_1\}$, $E_2 = \{\boldsymbol{\omega} : u_1 < L(\boldsymbol{\omega}) \leq u_2\}$, ..., $E_{m-1} = \{\boldsymbol{\omega} : u_{m-2} < L(\boldsymbol{\omega}) \leq u_{m-1}\}$, $E_m = \{\boldsymbol{\omega} : L(\boldsymbol{\omega}) > u_{m-1}\}$, where $u_i = u_1 + (i-1)\Delta u$ for $i = 1, \ldots, m-1$, are pre-specified numbers. Generally, we recommend that u_1 to be a small number such that E_1 is empty, and set m to be a large number such that the models falling into the subregion E_m are not of interest at all. In the simulations of this paper, we set $\Delta u = 20$, $u_1 = 0$ and $m = 51$. We note that the choice of Δu is not as crucial to SAA as to SAMC.

- *Desired sampling distribution.* For the choice of π_i's, if only very small probabilities are assigned to high energy regions, it will reduce the motivation of SAA to escape from local traps. On the other hand, if very high probabilities are assigned to high energy regions, the resulting sampling will not focus on low energy regions. A balanced choice is $\pi_1 = \pi_2 = \cdots = \pi_m = 1/m$.

- *Gain factor sequence.* In this paper, we choose the gain factor sequence in the form

$$a_t = \left(\frac{t_0}{\max\{t, t_0\}} \right)^{0.75}, \quad t \geq 1,$$

where $t_0 > 0$ is a pre-specified number. The choice of t_0 can depend on the complexity of the sample space. In general, a large value of m should associate with a large value of t_0.

- *Cooling schedule.* In the simulation studies of this paper, τ_t is set to be decreasing according to a square-root cooling schedule with $\tau_* = 0.005$ and $C = 0.05$.

References

Benveniste, A., Métivier, M., and Priouret, P. (1990), *Adaptive Algorithms and Stochastic Approximations*, New York: Springer-Verlag.

Geyer, C. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics*, **22**, 1993-2010.

Geyer, C. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, **30**, 339-351.

Knight, K. (1999). Epi-convergence in distribution and stochastic equi-semicontinuity. Unpublished manuscript.

Knight, K. (2001). Limiting Distributions of Linear Programming Estimators. *Journal of Econometrics*, **95**, 347-374.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics*, **28**, 1356-1378.

Liang, F. (2009). On the use of SAMC for Monte Carlo integration. *Statistics & Probability Letters*, **79**, 581-587.

Liang, F., Cheng, Y. and Lin, G. (2014). Simulated Stochastic Approximation Annealing for Global Optimization with a Square-Root Cooling Schedule. *J. Amer. Statist. Assoc.* **109**, 847-863.

Liang, F., Liu, C., Carroll, R.J. (2007). Stochastic approximation in Monte Carlo computation. *J. Amer. Statist. Assoc.* **102**, 305-320.

Pflug, G. (1994). On an argmax-distribution connected to the Poisson process. In *Asymptotic Statistics*, Physica-Verlag, pp. 123-130.

Pflug, G. (1995), Asymptotic stochastic programs. *Mathematics of Operations Research* **20**, 769-789.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *Annals of Mathematical Statistics*, **22**, 400-407.