Variable Selection for Regression Models with Missing Data

Ramon I. Garcia, Joseph G. Ibrahim and Hongtu Zhu

Department of Biostatistics, University of North Carolina at Chapel Hill, USA

Supplementary Material

S1. Assumptions for proofs of Theorems 1 - 2

Even though the model $\mathbf{l}(\boldsymbol{\eta}) = \sum_{i=1}^{n} \mathbf{l}_i(\boldsymbol{\eta}) = \sum_{i=1}^{n} \log f(\mathbf{D}_{o,i}|\boldsymbol{\eta})$ may be misspecified, White (1994) has shown that the unpenalized ML estimate converges to the value of $\boldsymbol{\eta}$ which minimizes $E\left[\sum_{i=1}^{n} l_i(\boldsymbol{\eta})\right] = \sum_{i=1}^{n} \int l_i(\boldsymbol{\eta}) g(\mathbf{D}_{o,i}) d\mathbf{D}_{o,i}$ where $g(\cdot)$ is the true density. We denote the true value by $\boldsymbol{\eta}_n^* = \arg\sup_{\boldsymbol{\eta}} E[\mathbf{l}(\boldsymbol{\eta})]$. For simplicity, we further assume that $E[\partial\boldsymbol{\eta} l_i(\boldsymbol{\eta})] = 0$ for all i and $\boldsymbol{\eta}^* = \boldsymbol{\eta}_n^*$, for all n. Similarly, we define $\boldsymbol{\eta}_{\mathcal{S}n}^* = \text{argsup}_{\boldsymbol{\eta}\,:\,\beta_j \neq 0, j \in \mathcal{S}} E[Q(\boldsymbol{\eta}|\boldsymbol{\eta}^*)]$ and let $\boldsymbol{\eta}_{\mathcal{S}n}^* = \boldsymbol{\eta}_{\mathcal{S}}^*$, for all n.

The following assumptions are needed to facilitate development of our methods, although they may not be the weakest possible conditions.

(C1) $\boldsymbol{\eta}^*$ is unique and an interior point of the parameter space $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is compact.

(C2) $\widehat{\boldsymbol{\eta}}_0 \to \boldsymbol{\eta}^*$ in probability.

(C3) For all i, $l_i(\boldsymbol{\eta})$ is three-times continuously differentiable on $\boldsymbol{\Theta}$ and $l_i(\boldsymbol{\eta})$, $|\partial_j l_i(\boldsymbol{\eta})|^2$ and $|\partial_j \partial_k \partial_l l_i(\boldsymbol{\eta})|$ are dominated by $B_i(\mathbf{D}_{o,i})$ for all $j, k, l = 1, \cdots, d$ where $\partial_j = \partial/\partial\eta_j$. We also require that the same smoothness condition also holds for $h(\mathbf{D}_{o,i};\boldsymbol{\eta}) = E\left[\log f(\mathbf{z}_{m,i}|\mathbf{D}_{o,i};\boldsymbol{\eta})|\mathbf{D}_{o,i};\boldsymbol{\eta}\right]$.

(C4) For each $\epsilon > 0$, there exists a finite K such that

$$\sup_{n\geq 1} \frac{1}{n} \sum_{i=1}^{n} E\left[B_i(\mathbf{D}_{o,i}) 1_{[B_i(\mathbf{D}_{o,i})>K]}\right] < \epsilon$$

for all n.

(C5)

$$\lim_{n\to\infty} -\frac{1}{n}\sum_{i=1}^{n} \partial^2_{\boldsymbol{\eta}} l_i(\boldsymbol{\eta}^*) = \mathbf{A}(\boldsymbol{\eta}^*),$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \partial_{\boldsymbol{\eta}} l_i(\boldsymbol{\eta}^*)\partial_{\boldsymbol{\eta}} l_i(\boldsymbol{\eta}^*)^T = \mathbf{B}(\boldsymbol{\eta}^*),$$

$$\lim_{n\to\infty} -\frac{1}{n}\sum_{i=1}^{n} D^{20}Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*) = \mathbf{C}(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*),$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} D^{10}Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)D^{10}Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)^T = \mathbf{D}(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*),$$

where $\mathbf{A}(\boldsymbol{\eta}^*)$ and $\mathbf{C}(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)$ are positive definite and D^{ij} denotes the i-th and j-th derivatives of the first and second component of the Q function respectively.

(C6) Define $a_n = \max_j \left\{p'_{\lambda_{jn}}(|\beta^*_j|) : \beta^*_j \neq 0\right\}$, and $b_n = \max_j \left\{p''_{\lambda_{jn}}(|\beta^*_j|) : \beta^*_j \neq 0\right\}$.

1. $\max_j\{\lambda_{jn} : \beta^*_j \neq 0\} = o_p(1)$.
2. $a_n = O_p(n^{-1/2})$.
3. $b_n = o_p(1)$.

(C7) Define $d_n = \min_j\{\lambda_{jn} : \beta^*_j = 0\}$.

1. For all j such that $\beta^*_j = 0$, $\lim_{n\to\infty} \lambda^{-1}_{jn} \liminf_{\beta\to 0+} p'_{\lambda_{jn}}(\beta) > 0$ in probability.
2. $n^{1/2}d_n \xrightarrow{p} \infty$.

Proof of Theorem 1a.

Given assumptions (C1) - (C6), then it follows from White (1994) that

$$n^{-1/2}\sum_{i=1}^{n} \partial_{\boldsymbol{\eta}} l_i(\boldsymbol{\eta}^*) \xrightarrow{D} N\left(\mathbf{0}, \mathbf{B}(\boldsymbol{\eta}^*)\right) \tag{1.2}$$

and

$$n^{1/2}(\widehat{\boldsymbol{\eta}}_0 - \boldsymbol{\eta}^*) \xrightarrow{D} N\left(\mathbf{0}, \mathbf{A}(\boldsymbol{\eta}^*)^{-1}\mathbf{B}(\boldsymbol{\eta}^*)\mathbf{A}(\boldsymbol{\eta}^*)^{-1}\right). \tag{1.3}$$

To show $\widehat{\boldsymbol{\eta}}_\lambda$ is a $\sqrt{n}$-consistent maximizer of $\boldsymbol{\eta}^*$, it is enough to show that

$$P\left(\sup_{||\mathbf{u}||=C}\left\{\mathbf{l}(\boldsymbol{\eta}^*+n^{-1/2}\mathbf{u})-n\sum_{j=1}^{p}p_{\lambda_{jn}}(|\boldsymbol{\beta}_j^*+n^{-1/2}u_j|)\right\}\right.$$
$$\left.-\mathbf{l}(\boldsymbol{\eta}^*)+n\sum_{j=1}^{p}p_{\lambda_{jn}}(|\boldsymbol{\beta}_j^*|)<0\right)$$

converges to 1 for large C, since this implies there exists a local maximizer in the ball $\{\boldsymbol{\eta}^*+n^{-1/2}\mathbf{u}; ||\mathbf{u}||\le C\}$ and thus $||\widehat{\boldsymbol{\eta}}_\lambda-\boldsymbol{\eta}^*||=O_p(n^{-1/2})$. Taking a Taylor's series expansion of the penalized likelihood function, we have

$$\begin{aligned}
&\mathbf{l}(\boldsymbol{\eta}^*+n^{-1/2}\mathbf{u})-\mathbf{l}(\boldsymbol{\eta}^*)-n\sum_{j=1}^{p}p_{\lambda_{jn}}(|\boldsymbol{\beta}_j^*+n^{-1/2}u_j|)+n\sum_{j=1}^{p}p_{\lambda_{jn}}(|\boldsymbol{\beta}_j^*|)\\
\le\;& \mathbf{l}(\boldsymbol{\eta}^*+n^{-1/2}\mathbf{u})-\mathbf{l}(\boldsymbol{\eta}^*)-n\sum_{j=1}^{p_1}p_{\lambda_{jn}}(|\boldsymbol{\beta}_j^*+n^{-1/2}u_j|)+n\sum_{j=1}^{p_1}p_{\lambda_{jn}}(|\boldsymbol{\beta}_j^*|)\\
=\;& n^{-1/2}\mathbf{u}^T\partial_{\boldsymbol{\eta}}\mathbf{l}(\boldsymbol{\eta}^*)-\frac{1}{2}\mathbf{u}^T\left[-\frac{1}{n}\partial^2_{\boldsymbol{\eta}}\mathbf{l}(\boldsymbol{\eta}^*)\right]\mathbf{u}-n^{1/2}\sum_{j=1}^{p_1}\left[p'_{\lambda_{jn}}(|\boldsymbol{\beta}_j^*|)\mathrm{sgn}(\boldsymbol{\beta}_j^*)u_j\right]\\
&-\frac{1}{2}\sum_{j=1}^{p_1}\left[p''_{\lambda_{jn}}(|\boldsymbol{\beta}_j^*|)u_j^2\right]+o_p(1)\\
\le\;& n^{-1/2}\mathbf{u}^T\partial_{\boldsymbol{\eta}}\mathbf{l}(\boldsymbol{\eta}^*)-\frac{1}{2}\mathbf{u}^T\mathbf{A}(\boldsymbol{\eta}^*)\mathbf{u}+\sqrt{p_1}n^{1/2}a_n||\mathbf{u}_1||-\frac{1}{2}|b_n|||\mathbf{u}_1||^2+o_p(1)\\
\le\;& n^{-1/2}\mathbf{u}^T\partial_{\boldsymbol{\eta}}\mathbf{l}(\boldsymbol{\eta}^*)-\frac{1}{2}\mathbf{u}^T\mathbf{A}(\boldsymbol{\eta}^*)\mathbf{u}+\sqrt{p_1}n^{1/2}a_n||\mathbf{u}_1||+o_p(1), \qquad (1.4)
\end{aligned}$$

where $\mathbf{u}=(\mathbf{u}_1^T,\mathbf{u}_2^T)^T$ and $\mathbf{u}_1$ is a $p_1\times 1$ vector. The second inequality in (1.4) follows because $p_{\lambda_{jn}}(0)=0$ and $p_{\lambda_{jn}}\ge 0$. The third inequality follows from condition (C5) and the fact that $\sum_{i=1}^{p_1}|u_i|\le\sqrt{p_1}(\sum_{i=1}^{p_1}u_i^2)^{1/2}$. The last inequality follows from (C6). Since the first and third terms in (1.4) are $O_p(1)$ by (1.2) and condition (C6) - 2, and $\mathbf{u}^T\mathbf{A}(\boldsymbol{\eta}^*)\mathbf{u}$ is bounded below by $||\mathbf{u}||^2\times$ the smallest eigenvalue of $\mathbf{A}(\boldsymbol{\eta}^*)$, then the second term in (1.4) dominates the rest and all the terms can be made negative for large enough C.

Proof of Theorem 1b.

Suppose that the conditions of Theorem 1a hold, and there exists an, $\widehat{\boldsymbol{\eta}}_\lambda$, which is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\eta}^*$. It suffices to show that for large n, the gradient of the penalized log likelihood function evaluated at $\widehat{\boldsymbol{\eta}}_\lambda$, such that

$||\widehat{\boldsymbol{\eta}}_\lambda - \boldsymbol{\eta}^*|| = O_p(n^{-1/2})$ and $||\hat{\boldsymbol{\beta}}_{(2)\lambda}|| = O_p(n^{-1/2}) = o_p(1)$, is zero. Taking a Taylor's series expansion of the penalized log likelihood function about $\boldsymbol{\eta}^*$, we have

$$\begin{aligned}
\mathbf{0} = n^{-1/2}\left[\partial_{\boldsymbol{\eta}}\mathbf{l}(\widehat{\boldsymbol{\eta}}_\lambda) - n\,\partial_{\boldsymbol{\eta}}\left\{\sum_{j=1}^{p} p_{\lambda_{jn}}(|\beta_j|)\right\}\Bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_\lambda}\right] \\
&= n^{-1/2}\partial_{\boldsymbol{\eta}}\mathbf{l}(\boldsymbol{\eta}^*) - n^{1/2}(\widehat{\boldsymbol{\eta}}_\lambda - \boldsymbol{\eta}^*)^T\left[-\frac{1}{n}\partial^2_{\boldsymbol{\eta}}\mathbf{l}(\boldsymbol{\eta}^*)\right] + O_p(n^{-1}) \\
&\quad -n^{1/2}\,\partial_{\boldsymbol{\eta}}\left\{\sum_{j=1}^{p} p_{\lambda_{jn}}(|\beta_j|)\right\}\Bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_\lambda} \\
&= O_p(1) - n^{1/2}\,\partial_{\boldsymbol{\eta}}\left\{\sum_{j=1}^{p} p_{\lambda_{jn}}(|\beta_{j\lambda}|)\right\}\Bigg|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}_\lambda} \qquad (1.5)
\end{aligned}$$

where the last equality follows from $n^{-1/2}\partial_{\boldsymbol{\eta}}\mathbf{l}(\boldsymbol{\eta}^*) = n^{1/2}(\widehat{\boldsymbol{\eta}}_\lambda - \boldsymbol{\eta}^*)^T\left[-\partial^2_{\boldsymbol{\eta}}\mathbf{l}(\boldsymbol{\eta}^*)/n\right] = O_p(1)$. Therefore, for $j = p_1+1, \ldots p$, the gradient with respect to β_j of the second term of (1.5), is $-\mathrm{sgn}(\hat{\beta}_j)n^{1/2}\lambda_{jn}[\lambda_{jn}^{-1}p'_{\lambda_{jn}}(|\hat{\beta}_j|)]$. Since $||\hat{\boldsymbol{\beta}}_{(2)\lambda}|| = o_p(1)$, $\lambda_{jn}^{-1}p'_{\lambda_{jn}}(|\hat{\beta}_j|)$ is greater than zero for large n, it follows that (1.5) is dominated by the term $-\mathrm{sgn}(\hat{\beta}_j)n^{1/2}d_n$. Since $n^{1/2}d_n \xrightarrow{p} \infty$, it must be the case that $\hat{\beta}_{j\lambda} = 0$ for $j = p_1+1, \ldots, p$, otherwise the gradient could be made large in absolute value and could not possibly be equal to zero.

Proof of Theorem 1c.

Given conditions (C1) - (C7), Theorems 1a and 1b apply. Thus, there exists a $\hat{\boldsymbol{\beta}}_\lambda = \left(\hat{\boldsymbol{\beta}}^T_{(1)\lambda}, \mathbf{0}^T\right)^T$, and $\hat{\boldsymbol{\eta}}_\lambda = \left(\hat{\boldsymbol{\beta}}^T_\lambda, \hat{\boldsymbol{\tau}}^T_\lambda, \hat{\boldsymbol{\alpha}}^T_\lambda, \hat{\boldsymbol{\xi}}^T_\lambda\right)^T$ which is a $\sqrt{n}$ local maximizer of (6). Let $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}^{*T}_{(1)}, \mathbf{0}^T\right)^T$, $\boldsymbol{\gamma}^* = \left(\boldsymbol{\beta}^{*T}_{(1)}, \boldsymbol{\tau}^{*T}, \boldsymbol{\alpha}^{*T}, \boldsymbol{\xi}^{*T}\right)^T$, $\boldsymbol{\gamma} = \left(\boldsymbol{\beta}^T_{(1)}, \boldsymbol{\tau}^T, \boldsymbol{\alpha}^T, \boldsymbol{\xi}^T\right)^T$, $\hat{\boldsymbol{\gamma}}_\lambda = \left(\hat{\boldsymbol{\beta}}^T_{(1)\lambda}, \hat{\boldsymbol{\tau}}^T_\lambda, \hat{\boldsymbol{\alpha}}^T_\lambda, \hat{\boldsymbol{\xi}}^T_\lambda\right)^T$, and $\tilde{l}(\boldsymbol{\gamma}) = l((\boldsymbol{\beta}^T_{(1)}, \mathbf{0}, \boldsymbol{\tau}^T, \boldsymbol{\alpha}^T, \boldsymbol{\xi}^T))$. Let $\tilde{\mathbf{A}}(\boldsymbol{\gamma})$ be the resulting matrix from removing the p_1+1 to p rows and columns from the matrix $\mathbf{A}((\boldsymbol{\beta}_{(1)}^T, \mathbf{0}, \boldsymbol{\tau}^T, \boldsymbol{\alpha}^T, \boldsymbol{\xi}^T))$ and similiarly define $\tilde{\mathbf{B}}$. Let,

$$\begin{aligned}
\mathbf{h}_1\left(\boldsymbol{\beta}_{(1)}\right) &= (p'_{\lambda_1}(|\beta_1|)\mathrm{sgn}(|\beta_1|), \ldots, p'_{\lambda_{p_1}}(|\beta_{p_1}|)\mathrm{sgn}(|\beta_{p_1}|))^T, \\
\mathbf{G}_1\left(\boldsymbol{\beta}_{(1)}\right) &= \mathrm{diag}\left(p''_{\lambda_1}(|\beta_1|), \ldots, p''_{\lambda_{p_1}}(|\beta_{p_1}|)\right), \\
\mathbf{h}(\boldsymbol{\gamma}^*) &= \begin{pmatrix} \mathbf{h}_1\left(\boldsymbol{\beta}_{(1)}^*\right) \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{G}(\boldsymbol{\gamma}^*) = \begin{pmatrix} \mathbf{G}_1\left(\boldsymbol{\beta}_{(1)}^*\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \text{and} \\
\boldsymbol{\Sigma}(\boldsymbol{\gamma}^*) &= \left[\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{G}(\boldsymbol{\gamma}^*)\right]^{-1} \tilde{\mathbf{B}}(\boldsymbol{\gamma}^*) \left[\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{G}(\boldsymbol{\gamma}^*)\right]^{-1}.
\end{aligned}$$

Then, using a Taylor's series expansion, we have

$$\begin{aligned}
0 &= \partial_{\boldsymbol{\gamma}} \tilde{l}(\widehat{\boldsymbol{\gamma}}_\lambda) - n\partial_{\boldsymbol{\gamma}} \left[\sum_{j=1}^{p} p_{\lambda_j}(|\beta_{\lambda_j}|)\right]\Bigg|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}_\lambda} \\
&= \partial_{\boldsymbol{\gamma}} \tilde{l}(\boldsymbol{\gamma}^*) - n\mathbf{h}(\boldsymbol{\gamma}^*) - n(\widehat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^*)^T \left[-\frac{1}{n}\partial^2_{\boldsymbol{\gamma}} \tilde{l}(\boldsymbol{\gamma}^*) + \mathbf{G}(\boldsymbol{\gamma}^*)\right] + o_p(1) \\
&= n^{-1/2}\partial_{\boldsymbol{\gamma}} \tilde{l}(\boldsymbol{\gamma}^*) - n^{1/2}\mathbf{h}(\boldsymbol{\gamma}^*) - n^{1/2}(\widehat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^*)^T \left[\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{G}(\boldsymbol{\gamma}^*)\right] + o_p(1),
\end{aligned}$$

which indicates

$$n^{1/2}\left\{\widehat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^* + \left[\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{G}(\boldsymbol{\gamma}^*)\right]^{-1} \mathbf{h}(\boldsymbol{\gamma}^*)\right\} \stackrel{D}{=} n^{-1/2}\left[\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{G}(\boldsymbol{\gamma}^*)\right]^{-1} \partial_{\boldsymbol{\gamma}} l(\boldsymbol{\gamma}^*),$$

and therefore

$$n^{1/2}\left\{\widehat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}^* + \left[\tilde{\mathbf{A}}(\boldsymbol{\gamma}^*) + \mathbf{G}(\boldsymbol{\gamma}^*)\right]^{-1} \mathbf{h}(\boldsymbol{\gamma}^*)\right\} \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\gamma}^*)\right).$$

For the SCAD penalty with $\lambda_{jn} = \lambda_n$, if $\lambda_n = o_p(1)$, $n^{1/2}\lambda_n \xrightarrow{p} \infty$ and conditions (C1) - (C5) are satisfied, then the oracle properties of Theorem 1 hold. For the ALASSO penalty, with $\lambda_{jn} = \lambda_n|\hat{\beta}_j|^{-1}$ where $\hat{\beta}_j$ is the unpenalized ML estimate, $\lambda_n = O_p(n^{-1/2})$, $n\lambda_n \xrightarrow{p} \infty$ and conditions (C1) - (C5) imply Theorem 1. Therefore, depending on the penalty function and specification of λ_{jn}, the rates of λ_{jn} which characterize the oracle properties, may be different.

Under the assumptions of Theorem 1 for the SCAD and ALASSO penalty functions, $\mathbf{h}(\boldsymbol{\eta}^*) \to 0$, therefore the asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}_\lambda$ is $n^{-1}\boldsymbol{\Sigma}(\boldsymbol{\gamma}^*)$. Using Louis's formula (Louis (1982)), an estimate of $\boldsymbol{\Sigma}(\boldsymbol{\gamma}^*)$ is,

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}_\lambda) \approx n^{-1}[\hat{\mathbf{A}}(\hat{\boldsymbol{\gamma}}_\lambda) + \mathbf{G}(\hat{\boldsymbol{\gamma}}_\lambda)]^{-1}\hat{\mathbf{B}}(\hat{\boldsymbol{\gamma}}_\lambda)\left[\mathbf{A}(\hat{\boldsymbol{\gamma}}_\lambda) + \mathbf{G}(\hat{\boldsymbol{\gamma}}_\lambda)\right]^{-1}, \qquad (1.6)$$

where

$$
\begin{aligned}
\dot{Q}_i(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^*) &= \partial_{\boldsymbol{\gamma}}\left[\int \log f(D_{c,i};\boldsymbol{\gamma},\boldsymbol{\beta}_{(2)}=\mathbf{0}) f(\mathbf{z}_{m,i}|\mathbf{D}_{o,i};\boldsymbol{\gamma}^*,\boldsymbol{\beta}_{(2)}=\mathbf{0})d\mathbf{z}_{m,i}\right]\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}, \\
\hat{\mathbf{B}}(\boldsymbol{\gamma}^*) &= n^{-1}\sum_{i=1}^{n}\dot{Q}_i(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^*)\dot{Q}_i(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^*)^T, \\
\dot{Q}(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^*) &= \partial_{\boldsymbol{\gamma}}Q((\boldsymbol{\beta}_{(1)},\mathbf{0},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\xi})|\boldsymbol{\gamma}^*,\boldsymbol{\beta}_{(2)}=\mathbf{0})\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} \\
\ddot{Q}(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^*) &= \partial^2_{\boldsymbol{\gamma}}Q((\boldsymbol{\beta}_{(1)},\mathbf{0},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\xi})|\boldsymbol{\gamma}^*,\boldsymbol{\beta}_{(2)}=\mathbf{0})\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}, \text{and} \\
\hat{\mathbf{A}}(\boldsymbol{\gamma}^*) &= -n^{-1}\ddot{Q}(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^*)+n^{-1}\dot{Q}(\boldsymbol{\gamma}^*|\boldsymbol{\eta}^*)\dot{Q}(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^*)^T \\
&\quad -n^{-1}\,E[(\partial_{\boldsymbol{\gamma}}\log f(\mathbf{D}_c;\boldsymbol{\gamma},\boldsymbol{\beta}_{(2)}=\mathbf{0}))^{\oplus}|\mathbf{D}_o;\boldsymbol{\gamma}^*,\boldsymbol{\beta}_{(2)}=\mathbf{0}]\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}
\end{aligned}
$$

where $\mathbf{v}^{\oplus}=\mathbf{v}\mathbf{v}^T$.

Proof of Theorem 2a.

To prove Theorem 2, we first show that for $\boldsymbol{\eta}_{tn}\xrightarrow{p}\boldsymbol{\eta}_t$, $t=1,2$,

$$
\begin{aligned}
Q(\boldsymbol{\eta}_{1n}|\boldsymbol{\eta}_{2n})-Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2) &= o_p(n) \\
E[Q(\boldsymbol{\eta}_{1n}|\boldsymbol{\eta}_{2n})]-E[Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2)] &= o_p(n) \\
Q(\boldsymbol{\eta}_{1n}|\boldsymbol{\eta}_{2n})-E[Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2)] &= o_p(n). \qquad (1.7)
\end{aligned}
$$

First we note that conditions (C3) and (C4) imply $[Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2)-E[Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2)]/n$ converges in probability to 0 for all $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2\in\boldsymbol{\Theta}$. Furthermore, because conditions (C3) and (C4) satisfy the W-LIP assumption of Lemma 2 of Andrews (1992), we obtain the uniform continuity and stochastic continuity of $E[Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2)]$ and $[Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2)-E(Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2))]/n$ respectively. Because the stochastic continuity and pointwise convergence properties satisfy the assumptions of Theorem 3 of Andrews (1992), we have

$$
\sup_{(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\in\boldsymbol{\Theta}\times\boldsymbol{\Theta}}\frac{1}{n}|Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2)-E[Q(\boldsymbol{\eta}_1|\boldsymbol{\eta}_2)]|\xrightarrow{p}0, \qquad (1.8)
$$

which implies (1.7).

We also need to show that the hypothetical estimator

$$
\bar{\boldsymbol{\eta}}_{\mathcal{S}}=\text{argsup}_{\boldsymbol{\eta}\,:\,\beta_j\neq 0,j\in\mathcal{S}}Q(\boldsymbol{\eta}|\boldsymbol{\eta}^*)
$$

is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\eta}^*_{\mathcal{S}}$. To prove this, it is enough to show that

$$
P\left[\sup_{||\mathbf{u}||=C}Q(\boldsymbol{\eta}^*_{\mathcal{S}}+n^{-1/2}\mathbf{u}|\boldsymbol{\eta}^*)\le Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)\right]\ge 1-\epsilon
$$

for large C, since this implies there exists a local maximizer in the ball $\{\boldsymbol{\eta} + n^{-1/2}\mathbf{u}; ||\mathbf{u}|| \le C\}$ and thus $||\bar{\boldsymbol{\eta}}_{\mathcal{S}} - \boldsymbol{\eta}^*_{\mathcal{S}}|| = O_p(n^{-1/2})$. Taking a Taylor's series expansion of the first component of the Q function, we have

$$\begin{aligned} & Q(\boldsymbol{\eta}^*_{\mathcal{S}} + n^{-1/2}\mathbf{u}|\boldsymbol{\eta}^*) - Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*) \\ = \; & n^{-1/2}\mathbf{u}^T D^{10} Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*) - \frac{1}{2}\mathbf{u}^T \left[-\frac{1}{n} D^{20} Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*) \right] \mathbf{u} + o_p(1) \\ = \; & n^{-1/2}\mathbf{u}^T D^{10} Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*) - \frac{1}{2}\mathbf{u}^T \mathbf{C}(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)\mathbf{u} + o_p(1). \end{aligned} \tag{1.9}$$

Conditions (C3) and (C5) ensure that $n^{-1/2} D^{10} Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*) \xrightarrow{D} N(0, \mathbf{D}(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)) = O_p(1)$ and $\mathbf{C}(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)$ is positive definite. Therefore, the second term dominates the rest and (1.9) can be made negative for large enough C.

Let $\widetilde{\boldsymbol{\eta}}_{\mathcal{S}_\lambda} = \underset{\boldsymbol{\eta}:\, \beta_j = 0, j \in \mathcal{S}_\lambda}{\operatorname{argsup}} Q(\boldsymbol{\eta}|\widehat{\boldsymbol{\eta}}_0)$. Since $\widehat{\boldsymbol{\eta}}_0 \xrightarrow{p} \boldsymbol{\eta}^*$ and $\bar{\boldsymbol{\eta}}_{\mathcal{S}_\lambda} \xrightarrow{p} \boldsymbol{\eta}^*_{\mathcal{S}_\lambda}$, we have

$$\begin{aligned} & \frac{1}{n}\mathrm{dIC}_Q(\boldsymbol{\lambda}, 0) = \frac{1}{n}(\mathrm{IC}_Q(\boldsymbol{\lambda}) - \mathrm{IC}_Q(0)) \\ = \; & \frac{1}{n}\left[2Q(\widehat{\boldsymbol{\eta}}_0|\widehat{\boldsymbol{\eta}}_0) - 2Q(\widehat{\boldsymbol{\eta}}_\lambda|\widehat{\boldsymbol{\eta}}_0) + \hat{c}_n(\widehat{\boldsymbol{\eta}}_\lambda) - \hat{c}_n(\widehat{\boldsymbol{\eta}}_0)\right] \\ \ge \; & \frac{2}{n}\left[Q(\widehat{\boldsymbol{\eta}}_0|\widehat{\boldsymbol{\eta}}_0) - Q(\widetilde{\boldsymbol{\eta}}_{\mathcal{S}_\lambda}|\widehat{\boldsymbol{\eta}}_0)\right] + o_p(1) \\ = \; & \frac{2}{n}\left[Q(\widehat{\boldsymbol{\eta}}_0|\widehat{\boldsymbol{\eta}}_0) - Q(\widetilde{\boldsymbol{\eta}}_{\mathcal{S}_\lambda}|\boldsymbol{\eta}^*)\right] + o_p(1) \\ \ge \; & \frac{2}{n}\left[Q(\widehat{\boldsymbol{\eta}}_0|\widehat{\boldsymbol{\eta}}_0) - Q(\bar{\boldsymbol{\eta}}_{\mathcal{S}_\lambda}|\boldsymbol{\eta}^*)\right] + o_p(1) \\ = \; & \frac{2}{n}E\left[Q(\boldsymbol{\eta}^*|\boldsymbol{\eta}^*)\right] - E\left[Q\left(\boldsymbol{\eta}^*_{\mathcal{S}_\lambda}|\boldsymbol{\eta}^*\right)\right] + o_p(1) \\ \ge \; & \frac{2}{n}\min_{\mathcal{S} \not\supset \mathcal{S}_T}\left\{E\left[Q(\boldsymbol{\eta}^*|\boldsymbol{\eta}^*)\right] - E\left[Q(\boldsymbol{\eta}^*_{\mathcal{S}}|\boldsymbol{\eta}^*)\right]\right\} + o_p(1), \end{aligned}$$

where the second and fourth inequalities follow because $Q(\widehat{\boldsymbol{\eta}}_\lambda|\widehat{\boldsymbol{\eta}}_0) \le Q(\widetilde{\boldsymbol{\eta}}_{\mathcal{S}_\lambda}|\widehat{\boldsymbol{\eta}}_0)$ and $Q(\widetilde{\boldsymbol{\eta}}_{\mathcal{S}_\lambda}|\boldsymbol{\eta}^*) \le Q(\bar{\boldsymbol{\eta}}_{\mathcal{S}_\lambda}|\boldsymbol{\eta}^*)$ for all $\boldsymbol{\lambda}$ and the third and fifth equalities follow from (1.7). Therefore, we have

$$Pr\left(\inf_{\boldsymbol{\lambda} \in R^p_u} \mathrm{IC}_Q(\boldsymbol{\lambda}) > \mathrm{IC}_Q(0)\right) \to 1,$$

which yields Theorem 2a.

Proof of Theorem 2b.

Under the assumptions of Theorem 2b, we have

$$\begin{aligned}
n^{-1/2}\delta_Q(\boldsymbol{\lambda}_2,\boldsymbol{\lambda}_1) &= n^{-1/2}(\mathrm{IC}_Q(\boldsymbol{\lambda}_2)-\mathrm{IC}_Q(\boldsymbol{\lambda}_1))\\
&= 2n^{-1/2}\left(Q(\widehat{\boldsymbol{\eta}}_{\lambda_1}|\widehat{\boldsymbol{\eta}}_0)-2Q(\widehat{\boldsymbol{\eta}}_{\lambda_2}|\widehat{\boldsymbol{\eta}}_0)\right)+n^{-1/2}(\hat{c}(\widehat{\boldsymbol{\eta}}_{\lambda_2})-\hat{c}(\widehat{\boldsymbol{\eta}}_{\lambda_1}))\\
&= 2n^{-1/2}\Big(Q(\widehat{\boldsymbol{\eta}}_{\lambda_1}|\widehat{\boldsymbol{\eta}}_0)-E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_1}}|\widehat{\boldsymbol{\eta}}_0)]\Big)-2n^{-1/2}\Big(Q(\widehat{\boldsymbol{\eta}}_{\lambda_2}|\widehat{\boldsymbol{\eta}}_0)-E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_2}}|\widehat{\boldsymbol{\eta}}_0)]\Big)\\
&\quad +2n^{-1/2}\Big(E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_2}}|\widehat{\boldsymbol{\eta}}_0)]-E[Q(\boldsymbol{\eta}^*_{\mathcal{S}_{\lambda_1}}|\widehat{\boldsymbol{\eta}}_0)]\Big)+n^{-1/2}\delta_c(\boldsymbol{\lambda}_2,\boldsymbol{\lambda}_1)\\
&= O_p(1)+n^{-1/2}\delta_{c21}\xrightarrow{p}\infty.
\end{aligned}$$

Thus $\mathrm{IC}_Q(\boldsymbol{\lambda}_2) > \mathrm{IC}_Q(\boldsymbol{\lambda}_1)$ in probability, which yields Theorem 2b. Proof of Theorem 2c is similar to that of Theorem 2b.

S2. Statistical model for application of SIAS method to linear regression simulations.

To implement SIAS, we assume the response model is $y_i \sim N(\mathbf{u}_i^T\boldsymbol{\beta},\sigma^2)$, the covariate distribution is $\mathbf{u}_i \sim N(\boldsymbol{\mu}_u,\boldsymbol{\Sigma}_u)$ for $i=1,\ldots,n$ and the missing covariates are MAR. For the prior distribution of all the parameters we assume

$$\pi(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2,\boldsymbol{\mu}_u,\boldsymbol{\Sigma}_u)=\prod_{j=1}^{p}\{\pi(\beta_j|\gamma_j)\pi(\gamma_j)\}\pi(\sigma^2)\pi(\boldsymbol{\mu}_u|\boldsymbol{\Sigma}_u)\pi(\boldsymbol{\Sigma}_u)$$

where $\boldsymbol{\mu}_u|\boldsymbol{\Sigma}_u \sim N_8(\mathbf{0},\delta^{-1}\boldsymbol{\Sigma}_u)$, $\boldsymbol{\Sigma}_u^{-1}\sim \mathrm{Wishart}(r,\mathbf{I}_8)$, $\sigma^{-2}\sim \mathrm{Gamma}(\nu/2,\nu\omega/2)$, $\beta_j \sim (1-\gamma_j)N(0,t_j^2)+\gamma_j N(0,c_j^2t_j^2)$ and $\gamma_j \sim \mathrm{Bernoulli}(1/2)$. The hyperparameters were selected to reflect a lack of prior information on the parameters, i.e. $\delta=\nu=\omega=.001$, $r=8$. For the values of t_j and c_j, we use those suggested by George and McCulloch (1993) where $(\sigma^2_{\beta_j}/t_j^2,c_j^2)=(1,5),(1,10),(10,100),(10,300)$ and $\sigma^2_{\beta_j}$ was estimated using preliminary simulations.

We performed 5,000 simulations after a burn-in period of 5000 iterations. The posterior probability of $\boldsymbol{\gamma}$ was calculated from the posterior simulations and the model with the highest probability was selected as the 'best' model. The results of $(\sigma^2_{\beta_j}/t_j^2,c_j^2)=(1,10)$ are presented since it gives the best model with the highest posterior probability.

S3. Simulation results evaluating performance of standard errors of penalized estimates for linear regression simulations

Table 1.1: Standard errors of penalized estimates of linear regression model with covariates missing at random

Method	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_5$		
	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
SCAD-RE	.138	.164	.042	.170	.187	.039	.160	.180	.039
SCAD-IC_Q	.141	.161	.039	.178	.180	.048	.163	.175	.038
ALASSO-RE	.157	.161	.031	.183	.180	.035	.165	.173	.036
ALASSO-IC_Q	.139	.164	.039	.198	.185	.037	.166	.176	.038
Oracle	.138	.155	.036	.179	.157	.040	.147	.139	.028

In order to test the accuracy of the asymptotic error formula (1.6), we estimated the standard errors of the significant coefficients, β_1, β_3, and β_5 for the linear regression model using $n = 60, \sigma = 1$ with the covariates missing at random. The median of the absolute deviations $|\hat{\beta}_{j\lambda} - \beta_j^*|$ divided by .6745, denoted by SD, of the 100 penalized estimates can be regarded as the true standard error. The median of the estimated standard errors is denoted as SD_m. The median absolute deviation error divided by .6745, denoted SD_{mad}, measures the overall performance of the standard error formula. The results, which are presented in Table 1.1, indicate that the standard error estimate does a good job of estimating the true standard error. All of the SD_{mad} values were less than .05.

S4. Statistical model for application of SIAS method to Melanoma data

Using the definition of y_i, $\mathbf{z}_i$ and $\mathbf{x}_i$ in the main document, we assume a logistic regression model on $y_i|\mathbf{x}_i, \boldsymbol{\beta}$ with $E(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\gamma_i)/(1 + \exp(\gamma_i))$, where $\gamma_i = (1,\ \mathbf{z}_i,\ \mathbf{x}_i)^T\boldsymbol{\beta}$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_6)^T$. We assume the covariates are MAR with the following covariate distribution

$$f(\mathbf{z}_i|\mathbf{x}_i; \boldsymbol{\alpha}) = f(z_{i3}|z_{i1}, z_{i2}, \mathbf{x}_i; \boldsymbol{\alpha}_3) f(z_{i1}, z_{i2}|\mathbf{x}_i; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$$

for $i = 1, \ldots, n$. Since $\mathbf{x}_i$ are completely observed, they are conditioned on throughout. We take a $(z_{i1}, z_{i2}|\mathbf{x}_i) \sim \boldsymbol{N}_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2})$ and $\mu_{is} = \alpha_{s0} + \sum_{j=1}^{3} \alpha_{sj} x_{ij}$ for $s = 1, 2,\ i = 1, \ldots, n$ and $\boldsymbol{\Sigma}$ is an unstructured 2×2 covariance matrix. We also assume a logistic regression model for x_{i3} conditional on $(z_{i1}, z_{i2}, \mathbf{x}_i)$ with with $E(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\psi_i)/(1 + \exp(\psi_i))$, where

$\psi_i = (1, z_{i1}, z_{i2}, \mathbf{x}_i)^T \boldsymbol{\varphi}$, and $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \ldots, \varphi_5)^T$. Let $\boldsymbol{\nu}_j = (\alpha_{1j}, \alpha_{2j})^T$ for $j = 0, \ldots, 3$. For the prior distribution, we assume

$$\pi(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\nu}_0, \ldots, \boldsymbol{\nu}_3, \boldsymbol{\Sigma}) = \prod_{j=1}^{p} \{\pi(\beta_j|\gamma_j)\pi(\gamma_j)\} \prod_{l=0}^{5} \pi(\varphi_l) \prod_{k=0}^{3} \pi(\boldsymbol{\nu}_k|\boldsymbol{\Sigma})\pi(\boldsymbol{\Sigma}),$$

where $\varphi_l \sim N(0, \delta^{-1})$ for $l = 0, \ldots, 5$, $\boldsymbol{\nu}_k|\boldsymbol{\Sigma} \sim N_2(\mathbf{0}, \delta^{-1}\boldsymbol{\Sigma})$ for $k = 0, \ldots, 3$, $\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(r, \mathbf{I}_2)$, $\beta_j \sim (1-\gamma_j)N(0, t_j^2) + \gamma_j N(0, c_j^2 t_j^2)$ and $\gamma_j \sim \text{Bernoulli}(1/2)$ for $j = 1, \ldots, 6$.

The hyperparameters were selected to reflect lack of prior information on the parameters, i.e. $\delta = .001$, $r = 2$. We set $(\sigma_{\beta_j}^2 / t_j^2, c_j^2) = (1, 10)$. The posterior probability of $\boldsymbol{\gamma}$ was calculated from 5000 simulated observations after 5,000 burn-in iterations and the model with the highest probability was selected as the 'best' model.

References

Andrews, D. W. K. (1992). Generic uniform convergence. *Econometric Theory* **8**, 241-57.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.

Louis, T. A. (1983). Finding the observed information matrix when using the EM algorithm. *Journal of Royal Statistical Society, Series B* **44**, 226-233.

White, H. (1994). *Estimation, Inference, and Specification Analysis.* Cambridge University Press, New York.