Variable Selection for Sparse High-Dimensional Nonlinear Regression Models by Combining Nonnegative Garrote and Sure Independence Screening

Shuang Wu[1†*], Hongqi Xue[1†], Yichao Wu[2] and Hulin Wu[1]

[1] *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY;* [2] *Department of Statistics, North Carolina State University, Raleigh, NC*

Supplementary Material

S1 Proofs

Proof of Theorem 1. Define two $n \times p$ matrices $\boldsymbol{W} = (w_{ij})$ and $\widehat{\boldsymbol{W}} = (\hat{w}_{ij})$ with $w_{ij} = \beta_{0j} f_j(x_{ij}, \boldsymbol{\alpha}_{0j})$ and $\hat{w}_{ij} = \hat{\beta}_j f_j(x_{ij}, \hat{\boldsymbol{\alpha}}_j)$, where $\hat{\beta}_j$ and $\hat{\boldsymbol{\alpha}}_j$ are initial estimates defined in Section 2, $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$.

The proof of Theorem 1(i) follows straightly from the proof of Theorem 1 of Yuan and Lin (2007) by noting that the key steps are to establish their Equations (21), (22), (28), and (30). In our setting, these correspond to showing that $n^{-1}(\widehat{\boldsymbol{W}}^T \widehat{\boldsymbol{W}}) = n^{-1}(\boldsymbol{W}^T \boldsymbol{W}) + O_p(\delta_n)$ and every element of $n^{-1}(\hat{\boldsymbol{W}} - \boldsymbol{W})^T \boldsymbol{y}$ is of order $O_p(\delta_n)$, where $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^T$. These follow directly from Conditions (A1)-(A3) and the assumption that $\max_{1 \le j \le p}(|\hat{\beta}_j - \beta_{0j}| + \|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_{0j}\|) = O_p(\delta_n)$ and thus Theorem 1(i) holds. The proof of Theorem 1(ii) follows directly by combining the results of Theorem 1(i) and the root n consistency of the NLS estimator $\hat{\boldsymbol{\gamma}}$. $\square$

In order to prove Theorem 2, we first give a lemma. For any $N \le \delta$, let

$$\mathcal{G}_1(N) = \sup_{\boldsymbol{\gamma} \in \mathcal{A}(N)} |(P_n - P)\{\ell(\boldsymbol{\gamma}, \boldsymbol{x}, \boldsymbol{y}) - \ell(\boldsymbol{\gamma}_0, \boldsymbol{x}, \boldsymbol{y})\} I_n(\boldsymbol{x}, \boldsymbol{y})|,$$

where $\mathcal{A}(N)$ is defined in Condition (A5). The next lemma is about the upper bound of the tail probability of $\mathcal{G}_1(N)$ in the neighborhood of $\mathcal{A}(N)$.

Lemma 1. *Under Conditions (A1), (A3) and (A7), for any* $t > 0$,

$$P\Big(\mathcal{G}_1(N) \ge 4k_n^* N \max\{k_1, A_1 k_2\}(p/n)^{1/2}(1+t)\Big) \le \exp(-2t^2).$$

Proof. By Conditions (A3) and (A7), the first-order Taylor expansion, the triangular inequality and the Cauchy-Schwarz inequality, we have that on the set Ω_n,

$$
\begin{aligned}
&|\ell(\boldsymbol{\gamma}, \boldsymbol{X}, Y) - \ell(\boldsymbol{\gamma}_0, \boldsymbol{X}, Y)| \leq k_n^* |\sum_{j=1}^p \beta_j f_j(X_j, \boldsymbol{\alpha}_j) - \sum_{j=1}^p \beta_{0j} f_j(X_j, \boldsymbol{\alpha}_{0j})| \\
\leq\ & k_n^* |\sum_{j=1}^p \beta_j f_j(X_j, \boldsymbol{\alpha}_j) - \sum_{j=1}^p \beta_{0j} f_j(X_j, \boldsymbol{\alpha}_j)| + k_n^* |\sum_{j=1}^p \beta_{0j} f_j(X_j, \boldsymbol{\alpha}_j) - \sum_{j=1}^p \beta_{0j} f_j(X_j, \boldsymbol{\alpha}_{0j})| \\
\leq\ & k_n^* [\sum_{j=1}^p (\beta_j - \beta_{0j})^2]^{1/2} [\sum_{j=1}^p f_j^2(X_j, \boldsymbol{\alpha}_j)]^{1/2} + k_n^* [\sum_{j=1}^p \beta_{0j}^2]^{1/2} \{\sum_{j=1}^p [f_j(X_j, \boldsymbol{\alpha}_j) - f_j(X_j, \boldsymbol{\alpha}_{0j})]^2\}^{1/2} \\
\leq\ & k_n^* \sqrt{p} k_1 [\sum_{j=1}^p (\beta_j - \beta_{0j})^2]^{1/2} + k_n^* \sqrt{p} A_1 k_2 (\sum_{j=1}^p \|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_{0j}\|^2)^{1/2} \\
\leq\ & k_n^* \sqrt{p} \max\{k_1, A_1 k_2\} \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|, \qquad \text{(S1.1)}
\end{aligned}
$$

by using $[\sum_{j=1}^p f_j^2(X_j, \boldsymbol{\alpha}_j)]^{1/2} \leq \sqrt{p} \|f_j(X_j, \boldsymbol{\alpha}_j)\|_\infty \leq \sqrt{p} k_1$ and $[\sum_{j=1}^p \beta_{0j}^2]^{1/2} \leq \sqrt{p} A_1$ with A_1 defined in Condition (A1) and $k_n^* = 2(k_1 + K_n^*)$. On the set Ω_n, by the definition of $\mathcal{A}(N)$, the above random variable is further bounded by $k_n^* \sqrt{p} \max\{k_1, A_1 k_2\} N$.

Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ be a Rademacher sequence and apply the symmetrization theorem (Lemma 2 in Fan and Song (2010)) to yield that

$$
\mathrm{E}\mathcal{G}_1(N) \leq 2\mathrm{E}\left[\sup_{\boldsymbol{\gamma} \in \mathcal{A}(N)} |P_n \boldsymbol{\varepsilon} \{\ell(\boldsymbol{\gamma}, \boldsymbol{x}, \boldsymbol{y}) - \ell(\boldsymbol{\gamma}_0, \boldsymbol{x}, \boldsymbol{y})\} I_n(\boldsymbol{x}, \boldsymbol{y})|\right]. \qquad \text{(S1.2)}
$$

By the contraction theorem (Lemma 3 in Fan and Song (2010)) and Condition (A3), we can bound the right-hand side of (S1.2) further by

$$
4k_n^* \mathrm{E}\{\sup_{\boldsymbol{\gamma} \in \mathcal{A}(N)} |P_n\, \boldsymbol{\varepsilon} [\sum_{j=1}^p \beta_j f_j(\boldsymbol{x}_j, \boldsymbol{\alpha}_j) - \sum_{j=1}^p \beta_{0j} f_j(\boldsymbol{x}_j, \boldsymbol{\alpha}_{0j})] I_n(\boldsymbol{x}, \boldsymbol{y})|\}. \qquad \text{(S1.3)}
$$

By similar arguments to (S1.1), the expectation in (S1.3) is bounded by

$$\begin{aligned}
& \mathrm{E}\|P_n\,\boldsymbol{\varepsilon}[\sum_{j=1}^p f_j^2(\boldsymbol{x}_j,\boldsymbol{\alpha}_j)]^{1/2} I_n(\boldsymbol{x},\boldsymbol{y})\| \sup_{\boldsymbol{\gamma}\in\mathcal{A}(N)} [\sum_{j=1}^p (\beta_j-\beta_{0j})^2]^{1/2} \\
+ \;& [\sum_{j=1}^p \beta_{0j}^2]^{1/2} \mathrm{E}\Big(\sup_{\boldsymbol{\gamma}\in\mathcal{A}(N)} |P_n\,\boldsymbol{\varepsilon}\{\sum_{j=1}^p [f_j(\boldsymbol{x}_j,\boldsymbol{\alpha}_j)-f_j(\boldsymbol{x}_j,\boldsymbol{\alpha}_{0j})]^2\}^{1/2} I_n(\boldsymbol{x},\boldsymbol{y})|\Big) \\
\leq \;& N\Big(\mathrm{E}\|\sum_{j=1}^p f_j^2(X_j,\boldsymbol{\alpha}_j)\|\frac{I_n(\boldsymbol{X},Y)}{n}\Big)^{1/2} + \sqrt{p}A_1\Big(\mathrm{E}\|\sum_{j=1}^p [f_j(X_j,\boldsymbol{\alpha}_j)-f_j(X_j,\boldsymbol{\alpha}_{0j})]^2\|\frac{I_n(\boldsymbol{X},Y)}{n}\Big)^{1/2} \\
\leq \;& N\max\{k_1, A_1k_2\}(p/n)^{1/2}.
\end{aligned}$$

So we conclude that $\mathrm{E}\mathcal{G}_1(N) \leq 4k_n^* N \max\{k_1, A_1k_2\}(p/n)^{1/2}$. Hence, from the concentration theorem (Lemma 4 in Fan and Song (2010)), we have that

$$\begin{aligned}
& \mathrm{P}\left(\mathcal{G}_1(N) \geq 4k_n^* N\max\{k_1,A_1k_2\}(p/n)^{1/2}(1+t)\right) \\
\leq \;& \exp\left(-\frac{n[4k_n^*N\max\{k_1,A_1k_2\}(p/n)^{1/2}t]^2}{8(k_n^*)^2 q\max\{k_1^2,A_1^2k_2^2\}N^2}\right) = \exp(-2t^2).
\end{aligned}$$

This proves the lemma. □

Proof of Theorem 2. Similar to the proof of Theorem 1 in Fan and Song (2010), we define a convex combination $\boldsymbol{\gamma}_s = s\hat{\boldsymbol{\gamma}} + (1-s)\boldsymbol{\gamma}_0$ with $s = (1+\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|/N)^{-1}$. Clearly we have $0 < s < 1$ and $\|\boldsymbol{\gamma}_s - \boldsymbol{\gamma}_0\| = s\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\| = (1-s)N \leq N$, so $\boldsymbol{\gamma}_s \in \mathcal{A}(N)$. Due to the local convexity (Condition (A5)), we have

$$P_n\ell(\boldsymbol{\gamma}_s,\boldsymbol{x},\boldsymbol{y}) \leq sP_n\ell(\hat{\boldsymbol{\gamma}},\boldsymbol{x},\boldsymbol{y}) + (1-s)P_n\ell(\boldsymbol{\gamma}_0,\boldsymbol{x},\boldsymbol{y}) \leq P_n\ell(\boldsymbol{\gamma}_0,\boldsymbol{x},\boldsymbol{y}). \tag{S1.4}$$

Since $\boldsymbol{\gamma}_0$ is the minimizer of $\mathrm{E}\ell(\boldsymbol{\gamma},\boldsymbol{X},Y)$, we have from (S1.4) that

$$0 \leq \mathrm{E}[\ell(\boldsymbol{\gamma}_s,\boldsymbol{X},Y)-\ell(\boldsymbol{\gamma}_0,\boldsymbol{X},Y)] \leq (P-P_n)[\ell(\boldsymbol{\gamma}_s,\boldsymbol{X},Y)-\ell(\boldsymbol{\gamma}_0,\boldsymbol{X},Y)] \leq \mathcal{G}(N), \tag{S1.5}$$

where $\mathcal{G}(N) = \sup_{\boldsymbol{\gamma}\in\mathcal{A}(N)} |(P_n-P)[\ell(\boldsymbol{\gamma},\boldsymbol{X},Y)-\ell(\boldsymbol{\gamma}_0,\boldsymbol{X},Y)]|$.

Since $\boldsymbol{\gamma}_0$ is the unique minimizer of $\mathrm{E}\ell(\boldsymbol{\gamma},\boldsymbol{X},Y)$ (Condition (A2)), it follows that at $\boldsymbol{\gamma}_0$, the first-order derivative $\frac{\partial \mathrm{E}\ell(\boldsymbol{\gamma},\boldsymbol{X},Y)}{\partial\boldsymbol{\gamma}}$ equals to zero and the second-order derivative $\frac{\partial^2 \mathrm{E}\ell(\boldsymbol{\gamma},\boldsymbol{X},Y)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}$ is positive definite. By Condition (A4), $\frac{\partial^2 \mathrm{E}\ell(\boldsymbol{\gamma},\boldsymbol{X},Y)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}$ in a small neighborhood of $\boldsymbol{\gamma}_0$ is bounded away from 0 and $+\infty$. Without loss of generality, we take this small neighborhood as $\mathcal{A}(N)$. Therefore, there exists a constant V_1 such that for any $\boldsymbol{\gamma} \in \mathcal{A}(N)$,

$$\mathrm{E}[\ell(\boldsymbol{\gamma},\boldsymbol{X},Y)-\ell(\boldsymbol{\gamma}_0,\boldsymbol{X},Y)] \geq V_1\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2. \tag{S1.6}$$

Combining (S1.5) and (S1.6), it follows that

$$\|\boldsymbol{\gamma}_s - \boldsymbol{\gamma}_0\| \le [\mathcal{G}(N)/V_1]^{1/2}. \tag{S1.7}$$

Next we use (S1.7) to conclude the result. Note that for any u, $\mathrm{P}(\|\boldsymbol{\gamma}_s - \boldsymbol{\gamma}_0\| \ge u) \le \mathrm{P}(\mathcal{G}(N) \ge V_1 u^2)$. Setting $u = N/2$, we have

$$\mathrm{P}(\|\boldsymbol{\gamma}_s - \boldsymbol{\gamma}_0\| \ge N/2) \le \mathrm{P}(\mathcal{G}(N) \ge V_1 N^2/4). \tag{S1.8}$$

Using the definition of $\boldsymbol{\gamma}_s$ with $s = 1/2$, the left-hand side of (S1.8) is the same as $\mathrm{P}(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \ge N)$. Now, by taking $N = \min\{4a_n(1+t)/V_1, \delta\}$ with $a_n = 4k_n^* \max\{k_1, A_1 k_2\}\sqrt{p/n}$, we have

$$\begin{aligned} \mathrm{P}\big(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| \ge N\big) &\le \mathrm{P}\big(\mathcal{G}(N) \ge V_1 N^2/4\big) = \mathrm{P}\big(\mathcal{G}(N) \ge N a_n(1+t)\big) \\ &\le \mathrm{P}\big(\mathcal{G}(N) \ge N a_n(1+t), \Omega_{n,*}\big) + \mathrm{P}\big(\Omega_{n,*}^c\big), \end{aligned} \tag{S1.9}$$

where $\Omega_{n,*} = \{\|X_i\| \le K_n, |Y_i| \le K_n^*\}$.

On the set $\Omega_{n,*}$, since $\sup_{\boldsymbol{\gamma}\in\mathcal{A}(N)} P_n|\ell(\boldsymbol{\gamma}, X, Y) - \ell(\boldsymbol{\gamma}_0, X, Y)|(1 - I_n(X, Y)) = 0$, by the triangular inequality, we have $\mathcal{G}(N) \le \mathcal{G}_1(N) + \sup_{\boldsymbol{\gamma}\in\mathcal{A}(N)} |E[\ell(\boldsymbol{\gamma}, X, Y) - \ell(\boldsymbol{\gamma}_0, X, Y)](1 - I_n(X, Y))|$. It follows from Condition (A7) that (S1.9) is bounded by $\mathrm{P}(\mathcal{G}_1(N) \ge N a_n(1+t) + o(q/n)) + n\mathrm{P}((X, Y) \in \Omega_n^c)$. The conclusion follows from Lemma 1. □

Lemma 2. *For $j = 1, \cdots, p_n$, the marginal regression parameters $\beta_j^M = 0$ if and only if cov$[Y, f_j(X_j, \boldsymbol{\alpha}_j^M)] = 0$.*

Proof. As defined in (7), $\boldsymbol{\gamma}_j^M$ is the minimizer of $\mathrm{E}\ell(\boldsymbol{\gamma}_j, X_j, Y)$, it follows that the first-order derivative $\frac{\partial \mathrm{E}\ell(\boldsymbol{\gamma}_j, X_j, Y)}{\partial \boldsymbol{\gamma}_j}$ at $\boldsymbol{\gamma}_j^M$ equals to zero. By $\mathrm{E}f_j(X_j, \boldsymbol{\alpha}_j^M) = 0$ for $j = 1, \cdots, p_n$, the score equation of the marginal regression at β_j^M takes the following form:

$$\begin{aligned} &\frac{\partial \mathrm{E}[Y - \beta_j^M f_j(X_j, \alpha_j^M)]^2}{\partial \beta_j^M} = -2\mathrm{E}[Y - \beta_j^M f_j(X_j, \alpha_j^M)] f_j(X_j, \alpha_j^M) \\ = &\ -2\mathrm{E}[Y f_j(X_j, \alpha_j^M)] + 2\beta_j^M \mathrm{E}f_j^2(X_j, \boldsymbol{\alpha}_j^M) \\ = &\ -2\sum_{k\in\mathcal{M}_*} \beta_{0k}\mathrm{E}[f_k(X_k, \alpha_{0k}) f_j(X_j, \alpha_j^M)] + 2\beta_j^M \mathrm{E}f_j^2(X_j, \boldsymbol{\alpha}_j^M). \end{aligned}$$

Moreover, we have

$$\mathrm{cov}[Y, f_j(X_j, \alpha_j^M)] = \mathrm{cov}\{[\sum_{k\in\mathcal{M}_*} \beta_{0k} f_k(X_k, \alpha_{0k})], f_j(X_j, \alpha_j^M)\} = \sum_{k\in\mathcal{M}_*} \beta_{0k} E[f_k(X_k, \alpha_{0k}) f_j(X_j, \alpha_j^M)].$$

Then

$$0 = \frac{\partial \mathrm{E}[Y - \beta_j^M f_j(X_j, \alpha_j^M)]^2}{\partial \beta_j^M} = -2\mathrm{cov}[Y, f_j(X_j, \alpha_j^M)] + 2\beta_j^M \mathrm{E} f_j^2(X_j, \boldsymbol{\alpha}_j^M).$$

So we have $\beta_j^M = \mathrm{cov}[Y, f_j(X_j, \alpha_j^M)]/\mathrm{E} f_j^2(X_j, \boldsymbol{\alpha}_j^M)$. Thus this lemma holds. □

Corollary 1. If the partial orthogonality condition holds, that is, $\{f_j(X_j, \boldsymbol{\alpha}_j^M), j \notin \mathcal{M}_*\}$ is independent of $\{f_j(X_j, \boldsymbol{\alpha}_j^M), j \in \mathcal{M}_*\}$, then $\beta_j^M = 0$, for $j \notin \mathcal{M}_*$.

Lemma 3. *Under Condition (B1), if cov*$[Y, f_j(X_j, \boldsymbol{\alpha}_j^M)] \geq c_1 n^{-\kappa}$ *for* $j \in \mathcal{M}_*$ *with two constants* $c_1 > 0$ *and* $0 < \kappa < 1/2$*, then there exists a positive constant* c_2 *such that*

$$\min_{j \in \mathcal{M}_*} |\beta_j^M| \geq c_2 n^{-\kappa}.$$

Proof. From the proof of Lemma 2, we know that $\beta_j^M = \mathrm{cov}[Y, f_j(X_j, \alpha_j^M)]/\mathrm{E} f_j^2(X_j, \boldsymbol{\alpha}_j^M)$, then under Condition (B1), Lemma 3 holds by taking $c_2 = c_1/k_4$. □

Lemma 4. *If Condition (B4) holds,*

$$P(|Y| \geq y) \leq s_2 \exp[-m_2 y^{\frac{a}{a-1}}]$$

with $s_2 = s_0^{\nu_n} s_1$ *and* $m_2 = 2^{-\frac{a}{a-1}} (m_0 A_1 \nu_n + m_1)^{-\frac{1}{a-1}}$.

Proof. Under Condition (B4), since $Y = \sum_{j=1}^{p_n} \beta_j f_j(X_j, \boldsymbol{\alpha}_j) + \epsilon$, we have

$$\begin{aligned} \mathrm{E}\exp(tY) &= \prod_{j=1}^{p_n} \mathrm{E}[\exp(t\beta_j f_j(X_j, \boldsymbol{\alpha}_j))]\mathrm{E}[\exp(t\epsilon)] \\ &\leq [s_0 \exp(m_0 A_1 t^a)]^{\nu_n} s_1 \exp(m_1 t^a) \leq s_0^{\nu_n} s_1 \exp[(m_0 A_1 \nu_n + m_1) t^a]. \end{aligned}$$

By the exponential Chebyshev's inequality, we have

$$\mathrm{P}(Y \geq y) \leq \exp(-ty)\mathrm{E}[\exp(tY)] \leq s_0^{\nu_n} s_1 \exp(-ty) \exp[(m_0 A_1 \nu_n + m_1) t^a].$$

Take $t = [\frac{y}{2(m_0 A_1 \nu_n + m_1)}]^{\frac{1}{a-1}}$ for $a > 1$, the above inequality can be simplified as $\mathrm{P}(Y \geq y) \leq s_2 \exp[-m_2 y^{\frac{a}{a-1}}]$, with $s_2 = s_0^{\nu_n} s_1$ and $m_2 = 2^{-\frac{a}{a-1}} (m_0 A_1 \nu_n + m_1)^{-\frac{1}{a-1}}$. By similar arguments, we have

$$\begin{aligned} &\mathrm{P}(Y \leq -y) = \mathrm{P}(-Y \geq y) \leq \exp(-ty)\mathrm{E}[\exp(-tY)] \\ \leq\ & s_0^{\nu_n} s_1 \exp(-ty) \exp[(m_0 A_1 \nu_n + m_1)(-t)^a] \leq s_2 \exp[-m_2 y^{\frac{a}{a-1}}]. \end{aligned}$$

Thus the lemma holds. □

Proof of Theorem 3. Note that

$$\begin{aligned}&\mathrm{E}[\ell(\boldsymbol{\gamma}_j, X_j, Y) - \ell(\boldsymbol{\gamma}_j^M, X_j, Y)][1 - I_n(X_j, Y)]\\ \leq\ & |\mathrm{E}[\beta_j^2 f_j^2(X_j, \alpha_j)]I(|X_j) \geq K_n)| + |\mathrm{E}[(\beta_j^M)^2 f_j^2(X_j, \alpha_j^M)]I(|X_j| \geq K_n)| + B(\beta_j) + B(\beta_j^M),\end{aligned}$$

where $B(\beta_j) = |\mathrm{E}[Y\beta_j f_j(X_j, \alpha_j)][1-I_n(X_j, Y)]|$. The first two terms are of order $o(1/n)$ by Condition (B3), and the last two terms can be bounded following from Lemma 4 and the Cauchy-Schwarz inequality. By Theorem 2, for any $t > 0$, we have

$$\mathrm{P}\left(\sqrt{n}\|\hat{\boldsymbol{\gamma}}_j^M - \boldsymbol{\gamma}_j^M\| \geq 16k_n^* \max\{k_1, A_1k_2\}(1+t)/V_1\right) \leq \exp(-2t^2) + ns_2 \exp[-m_2(K_n^*)^{\frac{a}{a-1}}].$$

For any $c_3 > 0$, by taking $1 + t = c_3V_1n^{1/2-\kappa}/(16k_n^* \max\{k_1, A_1k_2\})$, it follows that $\mathrm{P}\left(\|\hat{\boldsymbol{\gamma}}_j^M - \boldsymbol{\gamma}_j^M\| \geq c_3n^{-\kappa}\right) \leq \exp[-c_4n^{1-2\kappa}/(k_n^*)^2] + ns_2 \exp[-m_2(K_n^*)^{\frac{a}{a-1}}]$ holds for some positive constant c_4. Then Theorem 3(i) follows from the union bound of probability.

To prove Theorem 3(ii), note that on the event $A_n \equiv \{\max_{j\in\mathcal{M}_*} \|\hat{\beta}_j^M - \beta_j^M\| \leq c_2n^{-\kappa}/2\}$, by Lemma 3 and the triangular inequality, we have $|\hat{\beta}_j^M| \geq c_2n^{-\kappa}/2$, for all $j \in \mathcal{M}_*$. Hence, by the choice of ζ_n, we have $\mathcal{M}_* \subset \hat{\mathcal{N}}_{\zeta_n}$. The result now follows from a simple union bound:

$$\mathrm{P}(A_n^c) \leq \nu_n\{\exp[-c_4n^{1-2\kappa}/(k_n^*)^2] + ns_2 \exp[-m_2(K_n^*)^{\frac{a}{a-1}}]\}.$$

This completes the proof. □

Proof of Theorem 4. From the proof of Lemma 2, we have that

$$\beta_j^M = \sum_{k\in\mathcal{M}_*} \beta_{0k}\mathrm{E}[f_k(X_k, \boldsymbol{\alpha}_{0k})f_j(X_j, \boldsymbol{\alpha}_j^M)]/\mathrm{E}f_j^2(X_j, \boldsymbol{\alpha}_j^M).$$

From Condition (B1), it follows that

$$\|\boldsymbol{\beta}^M\|^2 = \|\boldsymbol{\beta}_0^T \mathrm{cov}[\boldsymbol{f}(X, \boldsymbol{\alpha}_0), \boldsymbol{f}(X, \boldsymbol{\alpha}^M)]\boldsymbol{\beta}_0\|^2 \Big/ [\mathrm{E}f_j^2(X_j, \boldsymbol{\alpha}_j^M)]^2 \leq k_5^{-2}\lambda_{\max}(\Sigma_{\boldsymbol{\alpha}})\|\Sigma_{\boldsymbol{\alpha}}^{1/2}\boldsymbol{\beta}_0\|^2.$$

From Condition (B7), we have that $\|\boldsymbol{\beta}^M\|^2 \leq C^*\lambda_{\max}(\Sigma_{\boldsymbol{\alpha}})$ for some positive constant C^*. Then the number of $\{j : |\beta_j^M| > \rho n^{-\kappa}\}$ cannot exceed $O\{n^{2\kappa}\lambda_{\max}(\Sigma_{\boldsymbol{\alpha}})\}$ for any $\rho > 0$. Thus, on the set $B_n = \{\max_{1\leq j\leq p_n} |\hat{\beta}_j^M - \beta_j^M| \leq \rho n^{-\kappa}\}$, the number of $\{j : |\beta_j^M| > 2\rho n^{-\kappa}\}$ cannot exceed the number of $\{j : |\beta_j^M| > \rho n^{-\kappa}\}$, which is bounded by $O\{n^{2\kappa}\lambda_{\max}(\Sigma_{\boldsymbol{\alpha}})\}$. By taking $\rho = c_5/2$, we have

$$\mathrm{P}\big(|\hat{\mathcal{N}}_{\zeta_n}| \leq O\{n^{2\kappa}\lambda_{\max}(\Sigma_{\boldsymbol{\alpha}})\}\big) \geq \mathrm{P}(B_n).$$

The conclusion follows from Theorem 3(i). □

Lemma 5. *For $j = 1, \cdots, p_n$, the marginal residual increment $R_j^* = 0$ if and only if $cov[Y, f_j(X_j, \boldsymbol{\alpha}_j^M)] = 0$.*

Proof. From the definition of $\boldsymbol{\gamma}_0^M$ and the proof of Lemma 2, we have

$$\begin{aligned} \mathrm{R}_j^* &= 2\beta_j^M \mathrm{E}[Y f_j(X_j, \boldsymbol{\alpha}_j^M)] - (\beta_j^M)^2 \mathrm{E} f_j^2(X_j, \boldsymbol{\alpha}_j^M) \\ &= (\beta_j^M)^2 \mathrm{E} f_j^2(X_j, \boldsymbol{\alpha}_j^M) = \mathrm{cov}[Y, f_j(X_j, \boldsymbol{\alpha}_j)]. \end{aligned}$$

Thus the lemma holds. □

Lemma 6. *If $cov[Y, f_j(X_j, \boldsymbol{\alpha}_j^M)] \geq c_6 n^{-\kappa}$ for $j \in \mathcal{M}_*$ with two constants $c_6 > 0$ and $0 < \kappa < 1/2$, then*

$$\min_{j \in \mathcal{M}_*} |R_j^*| \geq c_6 n^{-\kappa}.$$

Proof of Theorem 5. By Taylor's expansion, we have

$$\begin{aligned} 2\mathrm{R}_{j,n} &= (\hat{\boldsymbol{\gamma}}_0^M - \hat{\boldsymbol{\gamma}}_j^M)^T \frac{\partial^2 P_n \ell(\boldsymbol{\gamma}, \boldsymbol{x}_j, \boldsymbol{y})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} |_{\boldsymbol{\gamma} = \boldsymbol{\gamma}_*} (\hat{\boldsymbol{\gamma}}_0^M - \hat{\boldsymbol{\gamma}}_j^M) \\ &= (\hat{\beta}_j^M)^2 \frac{\partial^2 P_n \ell(\boldsymbol{\gamma}, \boldsymbol{x}_j, \boldsymbol{y})}{\partial \boldsymbol{\beta}^2} |_{\boldsymbol{\gamma} = \boldsymbol{\gamma}_*} = \frac{2}{n} (\hat{\beta}_j^M)^2 \sum_{i=1}^n f_j^2(x_{ij}, \boldsymbol{\alpha}_{j,*}), \end{aligned} \quad \text{(S1.10)}$$

where $\boldsymbol{\gamma}_* = (\beta_{j,*}, \boldsymbol{\alpha}_{j,*}^T)^T$ lies between $\hat{\boldsymbol{\gamma}}_0^M$ and $\hat{\boldsymbol{\gamma}}_j^M$. And we have

$$\frac{2}{n} \sum_{i=1}^n f_j^2(x_{ij}, \boldsymbol{\alpha}_{j,*}) \geq \min_{\boldsymbol{\alpha}_j \in \mathcal{H}_j, f_j(X_j, \boldsymbol{\alpha}_j) \leq K} \frac{2}{n} \sum_{i=1}^n f_j^2(x_{ij}, \boldsymbol{\alpha}_j) I\{f_j(X_j, \boldsymbol{\alpha}_j) \leq K\}$$

for any given K. By the Hoeffding inequality, it follows that

$$P\{|(P_n - P) f_j^2(X_j, \boldsymbol{\alpha}_j) I\{|f_j(X_j, \boldsymbol{\alpha}_j)| \leq K\}| > \rho\} \leq \exp\{-2n\rho^2/(4K^4)\}$$

for any $\rho > 0$. By taking $\rho = n^{-\kappa/2}$, we have

$$P\{|(P_n - P) f_j^2(X_j, \boldsymbol{\alpha}_j) I\{|f_j(X_j, \boldsymbol{\alpha}_j)| \leq K\}| > n^{-\kappa/2}\} \leq \exp\{-2n^{1-\kappa}/(4K^4)\}.$$

Since $\mathrm{E} f_j^2(X_j, \boldsymbol{\alpha}_j) \geq k_5$, consequently, with probability tending to 1 exponentially fast, we have $\frac{2}{n} \sum_{i=1}^n f_j^2(x_{ij}, \boldsymbol{\alpha}_{j,*}) \geq k_5/2$. The desired result thus follows from Theorem 3. □

Proof of Theorem 6. By (S1.10) and Condition (B1), with probability tending to 1 exponentially fast, we have that

$$2\mathrm{R}_{j,n} \leq \max_{\boldsymbol{\alpha}_j \in \mathcal{H}_j} \frac{2}{n} (\hat{\boldsymbol{\beta}}_j^M)^2 \sum_{i=1}^n f_j^2(x_{ij}, \boldsymbol{\alpha}_j) \leq 4\mathrm{E} f_j^2(X_j, \boldsymbol{\alpha}_j) \leq 4k_4 (\hat{\boldsymbol{\beta}}_j^M)^2,$$

uniformly in j. Then if $\mathrm{R}_{j,n} > c_7 n^{-2\kappa}$, then $|\hat{\beta}_j^M| \geq D^* n^{-\kappa}$, with exception on a set with negligible probability, where $D^* = [c_7/(2k_4)]^{1/2}$. This implies that $|\hat{\mathcal{M}}_{\xi_n}| \leq |\hat{\mathcal{N}}_{\zeta_n}|$ with $\xi_n = D^* n^{-\kappa}$. The conclusion then follows from Theorem 4. □

References

Fan, J. and R. Song (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics* **38**, 3567–3604.