Supplementary Materials for "The Mnet Method for Variable Selection"

Jian Huang[1,2], Patrick Breheny[2], Sangin Lee[2], Shuangge Ma[3], and Cun-Hui Zhang[4]

[1]*Department of Statistics & Actuarial Science, University of Iowa*
[2]*Department of Biostatistics, University of Iowa*
[3]*School of Public Health, Yale University*
[4]*Department of Statistics and Biostatistics, Rutgers University*

This supplement contains additional results concerning variable selection accuracy from the simulation studies of Section 5 as well as proofs of Proposition 1 and Theorems 1 and 2.

Variable selection accuracy: Simulation results

Section 5 focused on estimation accuracy, while this supplement contains results concerning variable selection accuracy, as measured in three ways. Let S_1, S_0, and S denote the number of true, false, and total selections by a given estimator:

$$\begin{aligned} S_1 &= \#\{j : \hat{\beta}_j \neq 0 \text{ and } \beta_j \neq 0\} \\ S_0 &= \#\{j : \hat{\beta}_j \neq 0 \text{ and } \beta_j = 0\} \\ S &= S_1 + S_0. \end{aligned}$$

For the simulations described in Sections 5.1 and 5.2, we report here the power (probability of correctly selecting a nonzero coefficient), false discovery rate (FDR, the probability of incorrectly selecting a zero coefficient), and misclassification error (MC, the number of incorrect selections), which we define as follows:

$$\text{Power} = \frac{S_1}{d^o}$$
$$\text{FDR} = \frac{S_0}{S}$$
$$\text{MC} = S_0 + 3(d^o - S_1),$$

where, as in the manuscript, d^o denotes the number of nonzero coefficients. Note that our definition of misclassification error gives three times more weight to failing to select a truly important variable than incorrectly selecting an unimportant variable. This is entirely subjective. Our main reason for choosing this weighting is that it seemed to provide a reasonable balance between the two types of incorrect selections in the settings considered here. All methods incorrectly selected unimportant variables much more often than they failed to select important variables (as would be expected by the use of prediction accuracy to choose the tuning parameters), so a 1:1 weighting of the two types of incorrect selection closely resembles FDR as an outcome.

Fixed α and γ

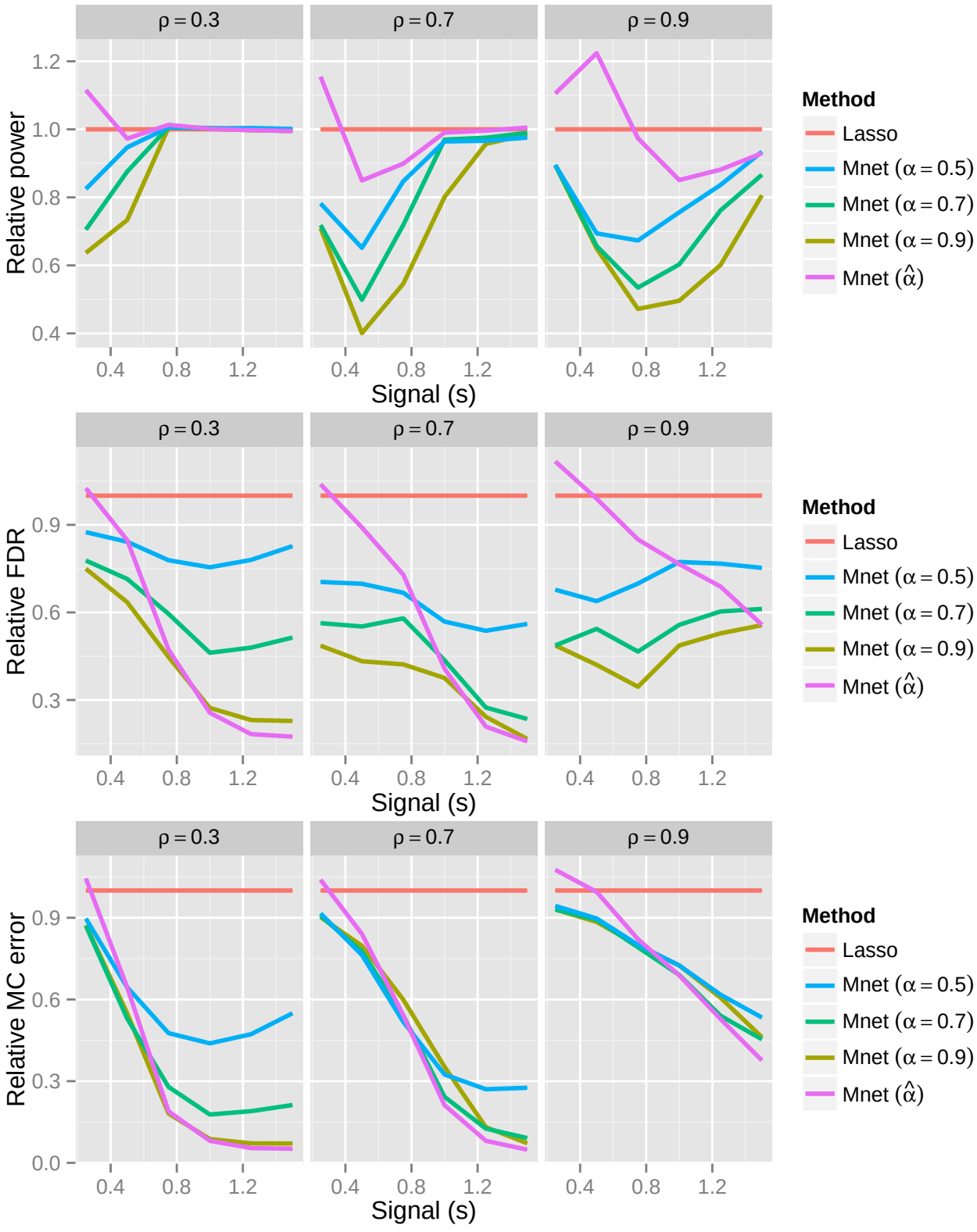
Supplemental Figure 1 displays the power, FDR, and misclassification error for the fixed-α Mnet estimator, as compared with the LASSO and variable-α Mnet estimator.

Because the methods select different numbers of variables, focusing exclusively on power or FDR can be misleading: models that select many variables will have greater power, while highly parsimonious models will have lower FDR. Thus, the LASSO appears more "powerful" than the fixed-α Mnet methods, but this is only because it selects a larger number of variables (and consequently has a high FDR). For this reason, our discussion concentrates on MC error, which does not inherently favor larger or smaller models.

For the most part, the results in terms of variable selection accuracy are broadly similar to the estimation accuracy results from Section 5.1: (1) the Mnet methods are much more accurate than the LASSO when the signal is reasonable strong; (2) the methods are all relatively similar when the signal is weak; and (3) there seems to be a benefit to selecting α, in that the variable-α Mnet method generally outperforms any individual fixed-α Mnet method, although this is not as dramatic in terms of variable selection as it was when we considered estimation accuracy.

Select α, fixed γ

Supplemental Figure 2 displays the power, FDR, and misclassification error for the LASSO, MCP, elastic net (Enet), and Mnet estimators. As in Section 5.2, external validation was



Supplemental Figure 1: Relative (to the LASSO) selection performance for the variable-α Mnet (with α selected by external validation) and various fixed-α Mnet estimators. Power, FDR, and misclassification (MC) error are averaged over 100 independently generated data sets.

used to select the α parameter for Enet and Mnet.

Overall, the main conclusion we would draw from Supplemental Figure 2 is that both LASSO and the elastic net select far more variables than MCP and Mnet. This increases power somewhat, but at the cost of substantially increasing FDR. In terms of MC error, Mnet and MCP are typically substantially more accurate than Enet and LASSO, although this advantage is diminished in low-signal and high-correlation settings. Comparing Mnet and MCP, the two methods perform similarly, although there seems to be a modest advantage to using Mnet when the predictors are moderately to strongly correlated.

Proofs

In the Appendix, we prove Proposition 1 and Theorems 1 and 2.

Proof of Proposition 1 The jth estimated coefficient $\hat{\beta}_j$ must satisfy the KKT conditions,

$$\begin{cases} -\frac{1}{n}x_j'(y - X\hat{\beta}) + \lambda_1(1 - |\hat{\beta}_j|/(\gamma\lambda_1))_+\text{sgn}(\hat{\beta}_j) + \lambda_2\hat{\beta}_j = 0, & \hat{\beta}_j \neq 0 \\ |x_j'(y - X\hat{\beta})| \leq \lambda_1, & \hat{\beta}_j = 0. \end{cases}$$

Let $\hat{r} = y - X\hat{\beta}$ and $\hat{z}_j = n^{-1}x_j'\hat{r}$. After some calculation, we have, if $\gamma\lambda_2 > 1$,

$$\hat{\beta}_j = \begin{cases} 0, & \text{if } |\hat{z}_j| \leq \lambda_1, \\ \text{sgn}(\hat{z}_j)\left|\frac{\gamma(|\hat{z}_j| - \lambda_1)}{\gamma\lambda_2 - 1}\right|, & \text{if } \lambda_1 < |\hat{z}_j| < \gamma\lambda_1\lambda_2, \\ \lambda_2^{-1}\hat{z}_j, & \text{if } |\hat{z}_j| \geq \gamma\lambda_1\lambda_2; \end{cases}$$

and if $\gamma\lambda_2 \leq 1$,

$$\hat{\beta}_j = \begin{cases} 0 & \text{if } |\hat{z}_j| \leq \lambda_1, \\ \lambda_2^{-1}\hat{z}_j & \text{if } |\hat{z}_j| > \lambda_1. \end{cases}$$

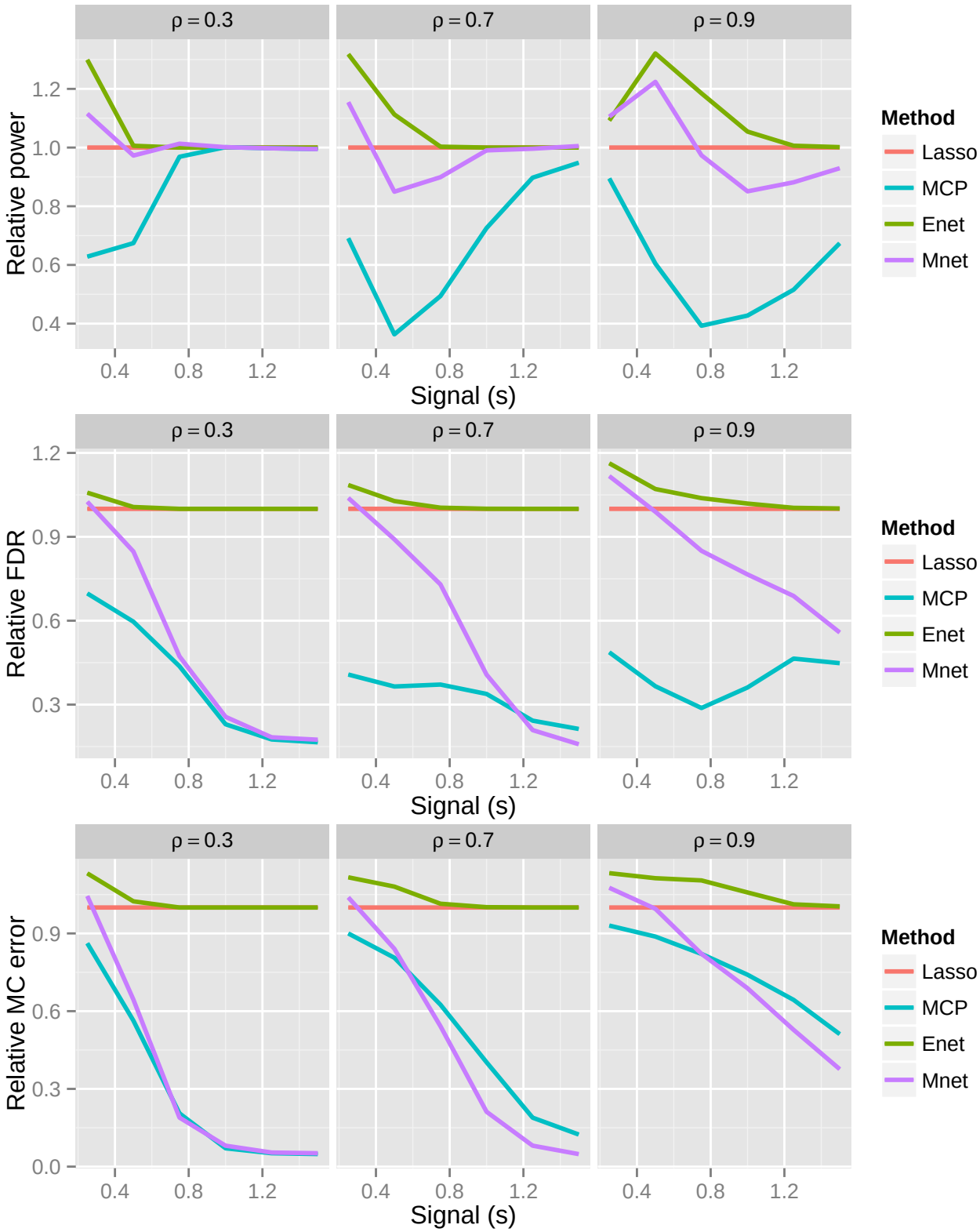
First, suppose that x_j and x_k are positively correlated. Based on the above expressions, we can show that

$$|\hat{\beta}_j - \hat{\beta}_k| \leq \xi|\hat{z}_j - \hat{z}_k|,$$

where ξ is given in (2.8). By the Cauchy-Schwarz inequality, $|\hat{z}_j - \hat{z}_k| = n^{-1}|(x_j - x_k)'\hat{r}| \leq n^{-1}\|x_j - x_k\|\|\hat{r}\| = n^{-1/2}\sqrt{2(1 - \rho_{jk})}\|\hat{r}\|$. Since $M(\hat{\beta}; \lambda) \leq M(0; \lambda)$ by the definition of $\hat{\beta}$, we have $\|\hat{r}\| \leq \|y\|$. Therefore

$$|\hat{\beta}_j - \hat{\beta}_k| \leq \xi|\hat{z}_j - \hat{z}_k| \leq \xi n^{-1/2}\sqrt{2(1 - \rho_{jk})}\|y\|.$$

For negative ρ_{jk}, we only need to change the sign of z_k and use the same argument. □



Supplemental Figure 2: Relative (to the LASSO) selection performance for for the MCP, elastic net (Enet) and Mnet estimator. Power, FDR, and misclassification (MC) error are averaged over 100 independently generated data sets.

To prove Theorems 1 and 2, we first need the lemma below. Let $\psi_\alpha(x) = \exp(x^\alpha) - 1$ for $\alpha \geq 1$. For any random variable X its ψ_α-Orlicz norm $\|X\|_{\psi_\alpha}$ is defined as $\|X\|_{\psi_\alpha} = \inf\{C > 0 : E\psi_\alpha(|X|/C) \leq 1\}$.

Lemma 1. *Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random variables with $\mathrm{E}\varepsilon_i = 0$ and $Var(\varepsilon_i) = 1$. Furthermore, suppose that $P(|\varepsilon_i| > x) \leq K\exp(-Cx^\alpha), i = 1, \ldots, n$ for constants C and K, and $1 \leq \alpha \leq 2$. Let $c_1, \ldots, c_n$ be constants satisfying $\sum_{i=1}^n c_i^2 = 1$. Let $X = \sum_{i=1}^n c_i\varepsilon_i$.*

(i) $\|X\|_{\psi_\alpha} \leq K_\alpha\left\{1 + (1+K)^{1/\alpha}C^{-1/\alpha}\alpha_n\right\}$, where K_α is a constant only depending on α, C and K.

(ii) Let $X_1, \ldots, X_m$ be any random variables whose Orlicz norms satisfy the inequality in (i). For any $b_n > 0$,

$$\mathrm{P}\left(\max_{1\leq j\leq m}|X_j| \geq b_n\right) \leq \frac{K_1\alpha_n(\log(m+1))^{1/\alpha}}{b_n}$$

for a positive constant K_1 only depending on α, C and K.

This lemma follows from Lemma 2.2.1 and Proposition A.1.6 of Van der Vaart and Wellner (1996). We omit the proof.

Proof of Theorem 1. Since $\hat\beta^o$ is the oracle ridge regression estimator, we have $\hat\beta^o_j = 0$ for $j \notin \mathcal{O}$ and

$$-\frac{1}{n}x_j'(y - X\hat\beta^o) + \lambda_2\hat\beta^o_j = 0, \quad \forall j \in \mathcal{O}. \tag{1}$$

If $|\hat\beta^o_j| \geq \gamma\lambda_1$, then $\rho'(|\hat\beta^o_j|;\lambda_1) = 0$. Since $c_{\min} + \lambda_2 > 1/\gamma$, the criterion (2.4) is strictly convex. By the KKT conditions, $\hat\beta = \hat\beta^o$ and $\mathrm{sgn}(\hat\beta) = \mathrm{sgn}(\beta^o)$ in the intersection of the events

$$\Omega_1(\lambda) = \left\{\max_{j\notin\mathcal{O}}\left|n^{-1}x_j'(y - X\hat\beta^o)\right| < \lambda_1\right\} \text{ and } \Omega_2(\lambda) = \left\{\min_{j\in\mathcal{O}}\mathrm{sgn}(\beta^o_j)\hat\beta^o_j \geq \gamma\lambda_1\right\}. \tag{2}$$

We first bound $1 - \mathrm{P}(\Omega_1(\lambda))$. Let $\hat\beta_\mathcal{O} = (\hat\beta_j, j \in \mathcal{O})'$ and $Z = n^{-1/2}X$. Let $\Sigma_\mathcal{O}(\lambda_2) = \Sigma_\mathcal{O} + \lambda_2 I_\mathcal{O}$. By (1) and using $y = X_\mathcal{O}\beta^o_\mathcal{O} + \varepsilon$,

$$\hat\beta^o_\mathcal{O} = \frac{1}{n}\Sigma^{-1}_\mathcal{O}(\lambda_2)X_\mathcal{O}'y = \Sigma^{-1}_\mathcal{O}(\lambda_2)\Sigma_\mathcal{O}\beta^o_\mathcal{O} + \frac{1}{\sqrt{n}}\Sigma^{-1}_\mathcal{O}(\lambda_2)Z_\mathcal{O}'\varepsilon. \tag{3}$$

Thus

$$\hat\beta^o_\mathcal{O} - \beta^o_\mathcal{O} = \frac{1}{\sqrt{n}}\Sigma^{-1}_\mathcal{O}(\lambda_2)Z_\mathcal{O}'\varepsilon + \{\Sigma^{-1}_\mathcal{O}(\lambda_2)\Sigma_\mathcal{O} - I_\mathcal{O}\}\beta^o_\mathcal{O}. \tag{4}$$

It follows that

$$\frac{1}{n}x_j'(y - X\hat\beta^o) = \frac{1}{n}x_j'\{I_n - Z_\mathcal{O}\Sigma^{-1}_\mathcal{O}(\lambda_2)Z_\mathcal{O}'\}\varepsilon - \frac{1}{\sqrt{n}}x_j'Z_\mathcal{O}\{\Sigma^{-1}_\mathcal{O}(\lambda_2)\Sigma_\mathcal{O} - I_\mathcal{O}\}\beta^o_\mathcal{O}.$$

Denote

$$T_{j1} = \frac{1}{n} x_j' \{I_n - Z_{\mathcal{O}} \Sigma_{\mathcal{O}}^{-1}(\lambda_2) Z_{\mathcal{O}}'\} \varepsilon, \quad T_{j2} = -\frac{1}{\sqrt{n}} x_j' Z_{\mathcal{O}} \{\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\Sigma_{\mathcal{O}} - I_{\mathcal{O}}\} \beta_{\mathcal{O}}^o.$$

First consider T_{j1}. Write $T_{j1} = n^{-1/2}\sigma \|a_j\| (a_j/\|a_j\|)'(\varepsilon/\sigma)$, where $a_j = n^{-1/2}\{I_n - Z_{\mathcal{O}}\Sigma_{\mathcal{O}}^{-1}(\lambda_2) Z_{\mathcal{O}}'\} x_j$. Since $n^{-1/2}\|x_j\| = 1$, we have $\|a_j\| \le 1$. By Lemma 1,

$$\begin{aligned} \mathrm{P}(\max_{j \notin \mathcal{O}} |T_{j1}| \ge \lambda_1/2) &\le \mathrm{P}(n^{-1/2}\sigma \max_{j \notin \mathcal{O}} |(a_j/\|a_j\|)'(\varepsilon/\sigma)| \ge \lambda_1/2) \\ &\le 2K_1 \alpha_n \frac{\sigma \log^{1/\alpha}(p - d^o + 1)}{\sqrt{n}\lambda_1}, \end{aligned} \tag{5}$$

where α_n is given in (4.2).

For T_{j2}, we have $T_{j2} = n^{-1/2}\lambda_2 x_j' Z_{\mathcal{O}} \Sigma_{\mathcal{O}}^{-1}(\lambda_2)\beta_{\mathcal{O}}^o$. Since

$$n^{-1/2}\lambda_2 |x_j' Z_{\mathcal{O}} \Sigma_{\mathcal{O}}^{-1}(\lambda_2)\beta_{\mathcal{O}}^o| \le \lambda_2 (c_1 + \lambda_2)^{-1} \sqrt{c_2} \|\beta^o\|,$$

we have $|T_{j2}| < \lambda_1/2$ for every j if

$$\lambda_1/2 > \lambda_2 (c_1 + \lambda_2)^{-1} \sqrt{c_2} \|\beta^o\|. \tag{6}$$

Thus by (5), when (6) holds, $1 - \mathrm{P}(\Omega_1(\lambda)) \le \pi_1$.

Now consider the event Ω_2. Let e_j be the jth unit vector of length d^o. By (4),

$$\hat{\beta}_j^o - \beta_j^o = S_{j1} + S_{j2}, \ j \in \mathcal{O},$$

where $S_{j1} = n^{-1} e_j' (\Sigma_{\mathcal{O}} + \lambda_2 I)^{-1} X_{\mathcal{O}}' \varepsilon$ and $S_{j2} = -\lambda_2 e_j' (\Sigma_{\mathcal{O}} + \lambda_2 I)^{-1}\beta_{\mathcal{O}}^o$. Therefore, $\mathrm{sgn}(\beta_j^o)\hat{\beta}_j^o \ge \gamma\lambda_1$ if $|\beta_j^o| + \mathrm{sgn}(\beta_j^o)(S_{j1} + S_{j2}) \ge \gamma\lambda_1$, which in turn is implied by

$$|S_{j1} + S_{j2}| \le \beta_*^o - \gamma\lambda_1, \ \forall j.$$

It follows that $1 - \mathrm{P}(\Omega_2(\lambda)) \le \mathrm{P}(\max_{j \in \mathcal{O}}(|S_{j1} + S_{j2}| > \beta_*^o - \gamma\lambda_1)$. Since $|S_{j2}| \le \lambda_2 \|\beta^o\|/(c_1 + \lambda_2)$, we have $|S_{j2}| < (\beta_*^o - \gamma\lambda_1)/2$ if $\beta_*^o > \gamma\lambda_1 + 2\lambda_2\|\beta^o\|/(c_1 + \lambda_2)$. Similarly to (5), by Lemma 1, when $\beta_*^o > \gamma\lambda_1 + 2\lambda_2\|\beta^o\|/(c_1 + \lambda_2)$,

$$\mathrm{P}(\max_{j \in \mathcal{O}}(|S_{j1} + S_{j2}| > \beta_*^o - \gamma\lambda_1) \le 2K_1\alpha_n \frac{\sigma\sqrt{c_2}\log^{1/\alpha}(d^o + 1)}{\sqrt{n}(\beta_*^o - \gamma\lambda_1)(c_1 + \lambda_2)}. \tag{7}$$

By (7) and the restrictions on λ_1 and β_*^o, $1 - P(\Omega_2(\lambda)) \le \pi_2$. □

Proof of Theorem 2. Let

$$\tilde{y} = \begin{pmatrix} y \\ 0_p \end{pmatrix}, \ \tilde{X} = \begin{pmatrix} X \\ \sqrt{n\lambda_2}\, I_p \end{pmatrix},$$

where 0_p is a p-dimensional vector of zeros. We have

$$\hat{\beta}(\lambda) = \underset{b}{\operatorname{argmin}}\{\frac{1}{2n}\|\tilde{y} - \tilde{X}b\|_2^2 + \sum_{j=1}^{p}\rho(|b_j|, \lambda_1)\}.$$

Thus the Mnet estimator can be considered an MCP estimator based on $(\tilde{y}, \tilde{X})$.

Denote $\tilde{P}_B = \tilde{X}_B(\tilde{X}_B'\tilde{X}_B)^{-1}\tilde{X}_B'$. For $m \geq 1$ and $u \in \mathbb{R}^n$, define

$$\tilde{\zeta}(u; m, \mathcal{O}, \lambda_2) = \max\left\{\frac{\|(\tilde{P}_B - \tilde{P}_{\mathcal{O}})v\|_2}{(mn)^{1/2}} : v = (u', 0_p')', \mathcal{O} \subseteq B \subseteq \{1, \ldots p\},\ |B| = m + |\mathcal{O}|\right\}.$$

Here $\tilde{\zeta}$ depends on λ_2 through $\tilde{P}$. We make this dependence explicit in the notation. By Lemma 1 of **?**, in the event

$$\lambda_1 \geq 2\sqrt{c^*}\,\tilde{\zeta}(y; m, \mathcal{O}, \lambda_2) \tag{8}$$

for $m = d^* - d^o$, we have

$$\#\{j : \hat{\beta}_j \neq 0\} \leq (K_* + 1)d^o \equiv p^*.$$

Thus in the event (8), the original p-dimensional problem reduces to a p_*-dimensional problem. Since $p_* \leq d^*$, the conditions of Theorem 2 implies that the conditions of Theorem 1 are satisfied for $p = p_*$. So the result follows from Theorem 1.

Specifically, let τ_n be as in (4.2) and λ_n^* as in (4.7). Let π_2 be as in (4.4). Denote

$$\pi_1^* = K_1\lambda_1^*/\lambda_1.$$

We show that if $\lambda_1 > 2\lambda_2\sqrt{c_2}\|\beta^o\|/(c_1 + \lambda_2)$, then

$$\mathrm{P}\big(2\sqrt{c^*}\,\tilde{\zeta}(y; m, \mathcal{O}, \lambda_2) > \lambda_1\big) \leq \pi_1^* + \pi_3. \tag{9}$$

Therefore, by Theorem 1, we have

$$\mathrm{P}(\mathrm{sgn}(\hat{\beta}) \neq \mathrm{sgn}(\beta^o) \text{ or } \hat{\beta}(\lambda) \neq \hat{\beta}^o(\lambda_2)) \leq \pi_1 + \pi_1^* + \pi_2 + \pi_3. \tag{10}$$

Then Theorem 2 follows from this inequality.

We now prove (9). By the definition of $\tilde{P}$,

$$\|(\tilde{P}_B - \tilde{P}_{\mathcal{O}})\tilde{y}\|_2^2 = y'\{Z_B(\Sigma_B + \lambda_2 I_B)^{-1}Z_B' - Z_{\mathcal{O}}(\Sigma_{\mathcal{O}} + \lambda_2 I_{\mathcal{O}})^{-1}Z_{\mathcal{O}}'\}y, \tag{11}$$

where $Z_B = n^{-1/2}X_B$. Let $P_B(\lambda_2) = Z_B(\Sigma_B + \lambda_2 I_B)^{-1}Z_B'$ and write $P_B = P_B(0)$. We have

$$\|(\tilde{P}_B - \tilde{P}_{\mathcal{O}})\tilde{y}\|_2^2 = \|(P_B - P_{\mathcal{O}})y\|_2^2 + y'(P_B(\lambda_2) - P_B)y - y'(P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})y. \tag{12}$$

Let $T_{B1} = \|(P_B - P_{\mathcal{O}})y\|_2^2$ and $T_{B2} = y'(P_B(\lambda_2) - P_B)y - y'(P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})y$. Let $\eta = \lambda_1/(2\sqrt{c^*})$. Note that $(P_B - P_{\mathcal{O}})y = (P_B - P_{\mathcal{O}})\varepsilon$, since $y = X_{\mathcal{O}}\beta^o + \varepsilon$ and $\mathcal{O} \subseteq B$. Therefore, $T_{B1} = \|(P_B - P_{\mathcal{O}})\varepsilon\|^2$.

Consider T_{B2}. Since $y = X_B\beta_B^o + \varepsilon$, some algebra shows that

$$y'(P_B(\lambda_2)-P_B)y = n\beta_B^{o\prime}Z_B'(P_B(\lambda_2)-P_B)Z_B\beta_B^o+2\sqrt{n}\beta_B^{o\prime}Z_B'(P_B(\lambda_2)-P_B)\varepsilon+\varepsilon'(P_B(\lambda_2)-P_B)\varepsilon,$$

and $n\beta_B^{o\prime}Z_B'(P_B(\lambda_2) - P_B)Z_B\beta_B^o = -n\lambda_2\|\beta_B^o\|^2 + n\lambda_2^2\beta_B^{o\prime}\Sigma_B^{-1}(\lambda_2)\beta_B^o$. These two equations and the identity $\|\beta_B^o\|^2 - \|\beta_{\mathcal{O}}\|^2 = 0$ imply that $T_{B2} = S_{B1} + S_2 + S_{B3} + S_{B4}$, where

$$\begin{aligned}
S_{B1} &= 2\sqrt{n}\{\beta_B^{o\prime}Z_B'(P_B(\lambda_2) - P_B) - \beta_{\mathcal{O}}^{o\prime}Z_{\mathcal{O}}'(P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})\}\varepsilon,\\
S_2 &= \varepsilon'\{P_{\mathcal{O}} - P_{\mathcal{O}}(\lambda_2)\}\varepsilon,\\
S_{B3} &= \varepsilon'\{P_B(\lambda_2) - P_B\}\varepsilon,\\
S_{B4} &= n\lambda_2^2\{\beta_B^{o\prime}\Sigma_B^{-1}(\lambda_2)\beta_B^o - \beta_{\mathcal{O}}^{o\prime}\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\beta_{\mathcal{O}}^o\}.
\end{aligned}$$

Using the singular value decomposition, it can be verified that $S_{B3} \le 0$. Also, since $\beta_B^o = (\beta_{\mathcal{O}}^{o\prime}, 0'_{|B|-d^o})'$ and by the formula of the block matrix inverse, it can be verified that $S_{B4} \le 0$. Therefore,

$$T_{B1} + T_{B2} \le T_{B1} + |S_{B1}| + S_2. \tag{13}$$

Note that $S_2 \ge 0$. When $\alpha = 2$, by Lemma 2 and Proposition 3 of **?**,

$$\mathrm{P}(\max_{B:|B|=m+d^o} T_{B1} > mn\lambda_1^2/(4c^*)) \le K_1\frac{2\sqrt{c^*}\sqrt{m}\{m\log(p-d^o)+1\}^{1/\alpha}}{\sqrt{m}\sqrt{n}\lambda_1}.$$

When $1 \le \alpha < 2$, since $P_B - P_{\mathcal{O}}$ is a rank m projection matrix and there are $\binom{p-d^o}{m}$ ways to choose B from $\{1,\ldots,p\}$, by Lemma 1,

$$\begin{aligned}
\mathrm{P}(\max_{B:|B|=m+d^o} T_{B1} > mn\lambda_1^2/(4c^*)) &\le K_1\frac{\alpha_n 2\sqrt{c^*}\sqrt{m}\log^{1/\alpha}\left(m\binom{p-d^o}{m}\right)}{\sqrt{m}\sqrt{n}\lambda_1}\\
&= K_1\frac{\alpha_n 2\sqrt{c^*}\log^{1/\alpha}\left(m\binom{p-d^o}{m}\right)}{\sqrt{n}\lambda_1},\\
&\le K_1\frac{\alpha_n 2\sqrt{c^*}\{m\log(p-d^o+1)\}^{1/\alpha}}{\sqrt{n}\lambda_1},
\end{aligned}$$

where K_1 is a constant that only depends on the tail probability of the error distribution in (A2b). Here we used the inequality $\log\left(\binom{p-d^o}{m}\right) \le m\log(e(p-d^o)/m)$.

Let $\mu^o = \sqrt{n}Z_{\mathcal{O}}\beta_{\mathcal{O}}^o$. Since $Z_B\beta_B^o = Z_{\mathcal{O}}\beta_{\mathcal{O}}^o = \mu^o/\sqrt{n}$, we have $S_{B1} = 2\mu^{o\prime}(P_B(\lambda_2) - P_B - (P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})\}\varepsilon$. Write $S_{B1} = 2\|a_B\|(a_B/\|a_B\|)'\varepsilon$, where

$$\|a_B\| = \|\{P_B(\lambda_2) - P_B - (P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})\}\mu^o\| \le \frac{2\lambda_2\|\mu^o\|}{c_* + \lambda_2}.$$

Therefore,

$$\begin{aligned}
\mathrm{P}(\max_{B:|B|=m+d^o} S_{B1} > mn\lambda_1^2/(8c^*)) &\leq \mathrm{P}\Big(\frac{4\lambda_2\|\mu^o\|}{c_*+\lambda_2}\max_{B:|B|=m+d^o}|(a_B/\|a_B\|)'\varepsilon| > \frac{mn\lambda_1^2}{8c^*}\Big) \\
&\leq K_1\alpha_n\frac{32c^*\|\mu^o\|\lambda_2\log^{1/\alpha}(\binom{p-d^o}{m})}{mn\lambda_1^2(c_*+\lambda_2)}, \\
&\leq K_1\alpha_n\frac{32c^*\|\mu^o\|\lambda_2 m^{1/\alpha}\{\log(p-d^o+1)\}^{1/\alpha}}{mn\lambda_1^2(c_*+\lambda_2)}.
\end{aligned}$$

By assumption, $\lambda_2\|\mu^o\| \leq \lambda_1/2(c_1+\lambda_2) \leq \lambda_1/2(c_*+\lambda_2)$, thus

$$\mathrm{P}(\max_{B:|B|=m+d^o} S_{B1} > mn\lambda_1^2/(8c^*)) \leq K_1\alpha_n\frac{16c^*m^{1/\alpha}\{\log(p-d^o+1)\}^{1/\alpha}}{mn\lambda_1(c_*+\lambda_2)^2}. \tag{14}$$

For S_2, by Lemma 1,

$$\mathrm{P}(S_2 > mn\lambda_1^2/(8c^*)) \leq K_1\alpha_n\frac{8c^*\sigma\lambda_2\sqrt{d^o}\log^{1/\alpha}(d^o+1)}{mn(c_*+\lambda_2)}. \tag{15}$$

Inequality (10) follows from (13) to (15). □