0.4pt=0pt

Supervised learning via the "hubNet" procedure

Leying Guan[1], Zhou Fan[1], Robert Tibshirani[1,2]

Departments of Statistics[1] and Biomedical Data Sciences[2], Stanford University

Supplementary Material

This supplementary material contains: (*i*) the optimization algorithm for the edge-out model (section S1); (*ii*) proofs for Theorem 4.5 and 4.9 in the main manuscript (section S2); (*iii*) comparisons between hubNet and other popular methods (section S3); (*iv*) comparisons between the edge-out model and hglasso (section S4).

S1 Optimization for the edge-out model

We consider the objective function (2.7). The diagonal elements of $\mathbf{B}$ are fixed at zero. Let $\mathbf{X}_{.,i}$ and $\mathbf{X}_{.,-i}$ denote the ith column of $\mathbf{X}$ and $\mathbf{X}$ with ith column removed, and let $\mathbf{B}_{-i,-i}$ denote $\mathbf{B}$ with ith row and column both removed. Let $S(x,t) = \mathrm{sign}(x)(|x| - t)_+$ be the soft-thresholding operator.

We use the following blockwise coordinate descent algorithm similar to that of Peng et al. (2010):

1. Initialize $\mathbf{B} = 0$.

2. Iterate over $i \in \{1, 2, \ldots, p\}$ until convergence:

   (a) Compute the $1 \times (p-1)$ vector $\mathbf{r}_{i,-i} = \mathbf{X}_{\cdot,i}^T(\mathbf{X}_{\cdot,-i} - \mathbf{X}_{\cdot,-i}\mathbf{B}_{-i,-i})$.

   (b) Compute the elementwise soft-thresholded vector $\beta_{i,-i} = S(\mathbf{r}_{i,-i}, \theta\gamma)$.

   (c) Update the ith row of $\mathbf{B}$:

$$\mathbf{B}_{i,-i} = \begin{cases} 0 & \|\beta_{i,-i}\|_2 \|\mathbf{X}_{\cdot,i}\|_2^2 \leq \theta(1-\gamma)\sqrt{p-1} \\ (1 - \frac{\theta(1-\gamma)\sqrt{p-1}}{\|\beta_{i.-i}\|_2 \|\mathbf{X}_{\cdot,i}\|_2^2})\beta_{i,-i} & \|\beta_{i,-i}\|_2 \|\mathbf{X}_{\cdot,i}\|_2^2 > \theta(1-\gamma)\sqrt{p-1} \end{cases}$$

It can be shown that, fixing all entries of $\mathbf{B}$ not in row i, the above update expression exactly minimizes the objective over $\mathbf{B}_{i,-i}$. Then this procedure is a blockwise coordinate descent algorithm, applied to an objective whose non-differentiable component is separable across blocks, and hence converges to the solution.

S2 Proof of Theorems 4.5 and Theorems 4.9

Denote by $\mathbf{X}_S$ and $\mathbf{X}_{S^C}$ the submatrices of $\mathbf{X}$ consisting of predictors in S and S^C, and define

$$\hat{\mathbf{\Sigma}}_{SS} := \frac{1}{n}\mathbf{X}_S^T\mathbf{X}_S, \quad \hat{\mathbf{\Sigma}}_{S^CS} := \frac{1}{n}\mathbf{X}_{S^C}^T\mathbf{X}_S, \quad \mathbf{W} := \mathbf{X}_{S^C} - \mathbf{X}_S\mathbf{\Gamma}.$$

Note that by (2.6), $\mathbf{W}$ is independent of $\mathbf{X}_S$ with independent Gaussian entries of variance at most 1. The following lemma collects probabilistic statements involving $\mathbf{X}_S$ and $\mathbf{W}$; its proof is deferred to Section S2.1.

Lemma 1. *Suppose $n, p \to \infty$, $1 \le s \le p$, and $s \ll n$. If $\lambda_{\min}(\mathbf{\Sigma}_{SS}) \ge C_{\min}$ for a constant $C_{\min} > 0$, then each of the following statements holds with probability approaching 1:*

$$\max_{j=1}^{p} \|\mathbf{X}_{\cdot,j}\|^2 \le 2n + 6\log p \tag{S2.1}$$

$$\max_{j=1}^{s} \|\mathbf{X}_{\cdot,j}\|^2 \le 2n \tag{S2.2}$$

$$\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_2 \le 2C_{\min}^{-1} \tag{S2.3}$$

$$\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_\infty \le \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty + 3(s + \sqrt{s}\log n)/(C_{\min}\sqrt{n}) \tag{S2.4}$$

$$\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{X}_S^T\mathbf{W}\|_{\infty,2} \le \sqrt{4np/C_{\min}} \tag{S2.5}$$

$$\|\mathbf{W}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_{\infty,2} \le \sqrt{4n(s + 3\log p)/C_{\min}} \tag{S2.6}$$

$$\|\mathbf{W}^T(\mathbf{Id}_{s\times s} - \tfrac{1}{n}\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{X}_S^T)\mathbf{W}\|_{\infty,2} \le 2n + \sqrt{3np} + \sqrt{6p\log p}. \tag{S2.7}$$

Proof of Theorem 4.5

Our proof draws upon a similar analysis of support recovery in the multivariate regression setting by Obozinski et al. (2011). Let us introduce $\theta_n = \theta\sqrt{p-1}/n$ and write the edge-out estimate (in the case $\gamma = 0$) as

$$\hat{\mathbf{B}}_{eo} = \underset{\mathbf{B}\in\mathbb{R}^{p\times p}\colon \mathbf{B}_{ii}=0\,\forall i}{\arg\min} \frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{B}\|_F^2 + \theta_n \sum_{i=1}^{p} \|\mathbf{B}_{i,\cdot}\|_2. \tag{S2.8}$$

Consider the restricted problem over $\mathbf{B} \in \mathbb{R}^{s\times p}$ where each predictor is regressed only on $\mathbf{X}_S$:

$$\hat{\mathbf{B}}_{\text{restricted}} = \underset{\mathbf{B}\in\mathbb{R}^{s\times p}\colon \mathbf{B}_{ii}=0\,\forall i}{\arg\min} \frac{1}{2n}\|\mathbf{X} - \mathbf{X}_S\mathbf{B}\|_F^2 + \theta_n \sum_{i\in S} \|\mathbf{B}_{i,\cdot}\|_2. \tag{S2.9}$$

The subgradient conditions for optimality of $\hat{\mathbf{B}}_{eo}$ and $\hat{\mathbf{B}}_{\text{restricted}}$ imply the following sufficient condition for recovery of S, whose proof we defer to Section S2.1:

Lemma 2. *If $\mathbf{X}_S^T\mathbf{X}_S$ is invertible, then the solution $\hat{\mathbf{B}} := \hat{\mathbf{B}}_{restricted}$ to (S2.9) is unique. If furthermore this solution satisfies*

$$\max_{j\in S^c} \frac{1}{n}\|\mathbf{X}_{\cdot,j}^T(\mathbf{X} - \mathbf{X}_S\hat{\mathbf{B}})\|_2 < \theta_n, \tag{S2.10}$$

$$\min_{i\in S} \|\hat{\mathbf{B}}_{i,\cdot}\|_2 > 0, \tag{S2.11}$$

then the solution $\hat{\mathbf{B}}_{eo}$ to (S2.8) is unique, with the first s rows non-zero and equal to $\hat{\mathbf{B}}$ and remaining rows equal to 0.

Through the remainder of this section, let $\hat{\mathbf{B}} := \hat{\mathbf{B}}_{\text{restricted}} \in \mathbb{R}^{s\times p}$ be the solution to the restricted problem (S2.9). As $s \ll n$ and $\mathbf{\Sigma}_{SS}$ is non-singular, $\mathbf{X}_S^T\mathbf{X}_S$ is invertible with probability 1. Hence, to prove Theorem 4.5, it suffices to show that (S2.10) and (S2.11) hold with high probability. Define

$$
\begin{aligned}
&\mathbf{U} := \begin{pmatrix} \mathbf{Id}_{s\times s} & \frac{1}{n}\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{X}_S^T\mathbf{W} \end{pmatrix} \in \mathbb{R}^{s\times p},\\
&\mathbf{B}^* := \begin{pmatrix} \mathbf{0}_{s\times s} & \mathbf{\Gamma} \end{pmatrix} \in \mathbb{R}^{s\times p},\\
&\hat{\mathbf{D}} := \operatorname{diag}\left(\|\hat{\mathbf{B}}_{1,.}\|_2^{-1}, ..., \|\hat{\mathbf{B}}_{s,.}\|_2^{-1}\right) \in \mathbb{R}^{s\times s},\\
&\mathbf{\Delta} \in \mathbb{R}^{s\times p}, \quad \mathbf{\Delta}_{ij} := \begin{cases} \mathbf{X}_{.,j}^T(\mathbf{X}_{.,j} - \mathbf{X}_S\hat{\mathbf{B}}_{.,j}) & i = j \\ 0 & \text{otherwise}, \end{cases}\\
&\mathcal{Z} := \left\{ \mathbf{Z} \in [-1,1]^{s\times p} : \begin{array}{ll} \mathbf{Z}_{i,.} = \hat{\mathbf{D}}_{i,i}\hat{\mathbf{B}}_{i,.} & \text{if } \|\hat{\mathbf{B}}_{i,.}\|_2 > 0 \\ \mathbf{Z}_{i,i} = 0 \text{ and } \|\mathbf{Z}_{i,.}\|_2 \le 1 & \text{if } \|\hat{\mathbf{B}}_{i,.}\|_2 = 0 \end{array} \right\}
\end{aligned}
$$

The subgradient condition for optimality of $\hat{\mathbf{B}}$ for (S2.9) implies the following, whose proof we also defer to Section S2.1.

Lemma 3. *There exists $\mathbf{Z} \in \mathcal{Z}$ such that*

$$\hat{\mathbf{B}} - \mathbf{B}^* = \mathbf{U} - \theta_n\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{Z} - \tfrac{1}{n}\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{\Delta}.$$

Using these lemmas, we now verify conditions (S2.10) and (S2.11):

Lemma 4. *Suppose Assumptions 4.1, 4.3, and 4.4 hold, and θ_n satisfies (4.10). Then with probability approaching 1, (S2.11) holds and*

$$\|\hat{\mathbf{B}} - \mathbf{B}^*\|_{\infty,2} \le 2\theta_n \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty.$$

Proof:

By Lemma 3, for some $\mathbf{Z} \in \mathcal{Z}$,

$$\|\hat{\mathbf{B}} - \mathbf{B}^*\|_{\infty,2} \le \|\mathbf{U}\|_{\infty,2} + \theta_n \|\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{Z}\|_{\infty,2} + \tfrac{1}{n}\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{\Delta}\|_{\infty,2}.$$

For the first term, (S2.5) and the definition of $\mathbf{U}$ imply, with probability approaching 1,

$$\|\mathbf{U}\|_{\infty,2} \le 1 + \sqrt{4p/(C_{\min} n)}.$$

For the second term, (S2.4) and the observation $\|\mathbf{Z}\|_{\infty,2} \le 1$ imply, with probability approaching 1,

$$\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{Z}\|_{\infty,2} \le \|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_\infty \|\mathbf{Z}\|_{\infty,2} \le \|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_\infty \le \|\mathbf{\Sigma}_{SS}^{-1}\| + 3(s + \sqrt{s}\log n)/(C_{\min}\sqrt{n}).$$

For the third term, note that for all $j = 1, \ldots, p$,

$$|\mathbf{\Delta}_{jj}| \le \|\mathbf{X}_{\cdot,j}\|^2, \tag{S2.12}$$

for otherwise

$$\|\mathbf{X}_{\cdot,j} - \mathbf{X}_S\hat{\mathbf{B}}_{\cdot,j}\|_2^2 - \|\mathbf{X}_{\cdot,j}\|^2 = (2\mathbf{X}_{\cdot,j} - \mathbf{X}_S\hat{\mathbf{B}}_{\cdot,j})^T(-\mathbf{X}_S\hat{\mathbf{B}}_{\cdot,j}) > 0,$$

implying that the objective (S2.9) would decrease upon setting $\hat{\mathbf{B}}_{\cdot,j} = 0$ and contradicting optimality of $\hat{\mathbf{B}}$. Then, as $\mathbf{\Delta}$ is diagonal, (S2.2) and (S2.3) imply, with probability approaching 1,

$$\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{\Delta}\|_{\infty,2} \leq \|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_{\infty,2} \max_{j=1}^{s} |\mathbf{\Delta}_{jj}| \leq \|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_2 \max_{j=1}^{s} \|\mathbf{X}_{\cdot,j}\|_2^2 \leq 4n/C_{\min}.$$

Noting that $\|\mathbf{\Sigma}_{SS}^{-1}\|_\infty \geq \|\mathbf{\Sigma}_{SS}^{-1}\|_2 = 1/\lambda_{\min}(\mathbf{\Sigma}_{SS}) \geq 1$ by our normalization $\mathbf{\Sigma}_{jj} = 1$ for all j, we have under the given assumptions

$$\max(1, \sqrt{p/n}, \theta_n s/\sqrt{n}, \theta_n \sqrt{s/n} \log n, \ll \theta_n \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty \ll \Gamma_{\min}.$$

Then with probability approaching 1, $\|\hat{\mathbf{B}} - \mathbf{B}^*\|_{\infty,2} \leq 2\theta_n \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty$ and

$$\min_i \|\hat{\mathbf{B}}_{i,\cdot}\|_2 \geq \min_i \|\mathbf{B}^*_{i,\cdot}\|_2 - 2\theta_n \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty = \Gamma_{\min} - 2\theta_n \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty > 0.$$

■

Lemma 5. *Suppose Assumptions 4.1, 4.2, 4.3, and 4.4 hold, and θ_n satisfies (4.10). Then (S2.10) holds with probability approaching 1.*

Proof: By Lemma 4, it suffices to consider the event where $\|\hat{\mathbf{B}}_{i,\cdot}\|_2 > 0$ for all $i \in S$, and hence $\mathbf{Z} = \hat{\mathbf{D}}\hat{\mathbf{B}}$ in Lemma 3. On this event, writing $\mathbf{X} = (\mathbf{X}_S, \mathbf{X}_S\mathbf{\Gamma} + \mathbf{W}) = (\mathbf{X}_S, \mathbf{W}) + \mathbf{X}_S\mathbf{B}^*$ and applying Lemma 3,

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}_{S^C}^T(\mathbf{X} - \mathbf{X}_S\hat{\mathbf{B}})\|_{\infty,2} &= \frac{1}{n}\|\mathbf{X}_{S^C}^T(\mathbf{X}_S, \mathbf{W}) + \mathbf{X}_{S^C}^T\mathbf{X}_S(\mathbf{B}^* - \hat{\mathbf{B}})\|_{\infty,2} \\ &\leq \frac{1}{n}\|\mathbf{X}_{S^C}^T(\mathbf{X}_S, \mathbf{W}) - \mathbf{X}_{S^C}^T\mathbf{X}_S\mathbf{U}\|_{\infty,2} + \theta_n\|\hat{\mathbf{\Sigma}}_{S^C S}\hat{\mathbf{\Sigma}}_{SS}^{-1}\hat{\mathbf{D}}\hat{\mathbf{B}}\|_{\infty,2} + \frac{1}{n}\|\hat{\mathbf{\Sigma}}_{S^C S}\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{\Delta}\|_{\infty,2}.\end{aligned} \tag{S2.13}$$

For the first term of (S2.13), recalling the definition of $\mathbf{U}$, noting that $\mathbf{X}_S^T(\mathbf{Id} - \frac{1}{n}\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{X}_S^T) = 0$, and applying (S2.7), with probability approaching 1,

$$\|\mathbf{X}_{S^C}^T(\mathbf{X}_S,\mathbf{W})-\mathbf{X}_{S^C}^T\mathbf{X}_S\mathbf{U}\|_{\infty,2}=\|\mathbf{X}_{S^C}^T(\mathbf{Id}-\tfrac{1}{n}\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{X}_S^T)\mathbf{W}\|_{\infty,2}$$
$$=\|\mathbf{W}^T(\mathbf{Id}-\tfrac{1}{n}\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{X}_S^T)\mathbf{W}\|_{\infty,2}\le 2n+\sqrt{3np}+\sqrt{6p\log p}\ll n\theta_n.$$

For the third term of (S2.13), applying (S2.12), (4.11), (S2.2), and (S2.6), with probability approaching 1,

$$\|\hat{\mathbf{\Sigma}}_{S^CS}\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{\Delta}\|_{\infty,2}\le\|\hat{\mathbf{\Sigma}}_{S^CS}\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_{\infty,2}\max_{j=1}^{s}|\mathbf{\Delta}_{jj}|=\frac{1}{n}\|(\mathbf{X}_S\mathbf{\Gamma}+\mathbf{W})^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_{\infty,2}\max_{j=1}^{s}|\mathbf{\Delta}_{jj}|$$
$$\le\left(\|\mathbf{\Gamma}^T\|_{\infty,2}+\frac{1}{n}\|\mathbf{W}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_{\infty,2}\right)\max_{j=1}^{s}\|\mathbf{X}_{.,j}\|_2^2\le\frac{2n}{\sqrt{C_{\min}}}+\sqrt{\frac{16n(s+3\log p)}{C_{\min}}}\ll n\theta_n.$$

It remains to bound the second term of (S2.13). Let $\mathbf{D}$ be as in Assumption 4.2 and write

$$\hat{\mathbf{\Sigma}}_{S^CS}\hat{\mathbf{\Sigma}}_{SS}^{-1}\hat{\mathbf{D}}\hat{\mathbf{B}}=\mathbf{\Gamma}^T\mathbf{D}\mathbf{B}^*+\mathbf{\Gamma}^T\mathbf{D}(\hat{\mathbf{B}}-\mathbf{B}^*)+\mathbf{\Gamma}^T(\hat{\mathbf{D}}-\mathbf{D})\hat{\mathbf{B}}+(\hat{\mathbf{\Sigma}}_{S^CS}\hat{\mathbf{\Sigma}}_{SS}^{-1}-\mathbf{\Gamma}^T)\hat{\mathbf{D}}\hat{\mathbf{B}}$$
$$=:\mathbf{I}+\mathbf{II}+\mathbf{III}+\mathbf{IV}.$$

By Assumption 4.2 and the definition of $\mathbf{B}^*$,

$$\|\mathbf{I}\|_{\infty,2}=\|\mathbf{\Gamma}^T\mathbf{D}\mathbf{\Gamma}\|_{\infty,2}\le 1-\delta.$$

By Lemma 4, with probability approaching 1,

$$\|\mathbf{II}\|_{\infty,2}\le\|\mathbf{\Gamma}^T\|_\infty\|\mathbf{D}(\hat{\mathbf{B}}-\mathbf{B}^*)\|_{\infty,2}\le\|\mathbf{\Gamma}^T\|_\infty\Gamma_{\min}^{-1}\|\hat{\mathbf{B}}-\mathbf{B}^*\|_{\infty,2}\le 2\|\mathbf{\Gamma}^T\|_\infty\Gamma_{\min}^{-1}\theta_n\|\mathbf{\Sigma}_{SS}^{-1}\|_\infty\ll 1.$$

$\mathbf{III}$ satisfies the same bound, as

$$\|\mathbf{III}\|_{\infty,2}\le\|\mathbf{\Gamma}^T\|_\infty\|(\hat{\mathbf{D}}-\mathbf{D})\hat{\mathbf{B}}\|_{\infty,2}=\|\mathbf{\Gamma}^T\|_\infty\max_{i\in S}\frac{|\|\mathbf{B}_{i,.}^*\|_2-\|\hat{\mathbf{B}}_{i,.}\|_2|}{\|\mathbf{B}_{i,.}^*\|_2}\le\|\mathbf{\Gamma}^T\|_\infty\|\mathbf{D}(\hat{\mathbf{B}}-\mathbf{B}^*)\|_{\infty,2}.$$

Finally, using $\mathbf{X}_{S^C} = \mathbf{X}_S\mathbf{\Gamma} + \mathbf{W}$ and applying (S2.6), with probability approaching 1,

$$\begin{aligned}\|\mathbf{IV}\|_{\infty,2} &= \|(\tfrac{1}{n}\mathbf{X}_{S^C}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1} - \mathbf{\Gamma}^T)\hat{\mathbf{D}}\hat{\mathbf{B}}\|_{\infty,2} = \frac{1}{n}\|\mathbf{W}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\hat{\mathbf{D}}\hat{\mathbf{B}}\|_{\infty,2} \\ &\leq \frac{1}{n}\|\mathbf{W}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_\infty\|\hat{\mathbf{D}}\hat{\mathbf{B}}\|_{\infty,2} \leq \frac{\sqrt{s}}{n}\|\mathbf{W}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_{\infty,2} \leq \sqrt{\frac{4s(s+3\log p)}{C_{\min}n}} \ll 1.\end{aligned}$$

Combining the above yields $\|\hat{\mathbf{\Sigma}}_{S^CS}\hat{\mathbf{\Sigma}}_{SS}^{-1}\hat{\mathbf{D}}\hat{\mathbf{B}}\|_{\infty,2} \leq 1 - \delta/2$ with probability approaching 1, which together with (S2.13) implies (S2.10). ■

Theorem 4.5 follows from Lemmas 2, 4, and 5.

Proof of Theorem 4.9

We verify the conditions of Lemma 8.2 of Zhou et al. (2009) under the given assumptions and in our asymptotic setting with random design. By (S2.1) and (S2.3), with probability approaching 1,

$$\max_{j\in S^C}\frac{\|\mathbf{X}_{\cdot,j}\|_2}{\sqrt{n}} \leq \sqrt{2 + \frac{6\log p}{n}}, \quad \lambda_{\min}(\hat{\mathbf{\Sigma}}_{SS}) \geq \frac{C_{\min}}{2}. \tag{S2.14}$$

It remains to verify the weighted incoherency condition (8.4a) of Zhou et al. (2009). Define $\mathbf{D}_{w,S} = \operatorname{diag}(w_1, \ldots, w_s) \in \mathbb{R}^{s\times s}$ and $\mathbf{D}_{w,S^C}^{-1} = \operatorname{diag}(w_{s+1}^{-1}, \ldots, w_p^{-1}) \in \mathbb{R}^{(s-p)\times(s-p)}$ where $w_k^{-1} = 0$ if $w_k = \infty$. Then

$$\|\mathbf{D}_{w,S^C}^{-1}\mathbf{X}_{S^C}^T\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{D}_{w,S}\|_\infty \leq \frac{w_{\max}(S)}{nw_{\min}(S^C)}\|\mathbf{X}_{S^C}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_\infty \leq \frac{\rho}{n}\|\mathbf{X}_{S^C}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_\infty.$$

Writing $\mathbf{X}_{S^C} = \mathbf{X}_S\mathbf{\Gamma} + \mathbf{W}$ and applying (4.11) and (S2.6), with probability approaching 1,

$$\frac{1}{n}\|\mathbf{X}_{S^C}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_\infty \leq \frac{\sqrt{s}}{n}\|\mathbf{X}_{S^C}^T\mathbf{X}_S\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_{\infty,2} \leq \sqrt{s}\|\mathbf{\Gamma}^T\|_{\infty,2} + \frac{\sqrt{s}}{n}\|\mathbf{W}^T\mathbf{X}_S\mathbf{\Sigma}_{SS}^{-1}\|_{\infty,2}$$
$$\leq \sqrt{\frac{s}{C_{\min}}} + \sqrt{\frac{4s(s+3\log p)}{nC_{\min}}} \leq \sqrt{\frac{s}{C_{\min}}}\left(1 + \sqrt{\frac{12\log p}{n}} + o(1)\right).$$

Hence under Assumption 4.7, with probability approaching 1,

$$\|\mathbf{D}_{w,S^C}^{-1}\mathbf{X}_{S^C}^T\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{D}_{w,S}\|_\infty \leq 1 - \eta - o(1) \leq 1 - \eta/2. \quad \text{(S2.15)}$$

Conditional on $\mathbf{X}$, on the event where (S2.14) and (S2.15) hold, our conclusion follows from Lemma 8.2 of Zhou et al. (2009). Then the conclusion also follows unconditionally.

S2.1 Proofs of supporting lemmas

In this section, we prove Lemmas 1, 2, and 3.

Proof of Lemma 1

Our normalization $\mathbf{\Sigma}_{jj} = 1$ implies $\|\mathbf{X}_{\cdot,j}\|_2^2 \sim \chi_n^2$ for each $j = 1, \ldots, p$. We use the chi-squared tail bound

$$P[\chi_n^2 > n + 2\sqrt{nt} + 2t] \leq \exp(-t) \quad \text{(S2.16)}$$

for all $t > 0$, from Lemma 1 of Laurent and Massart (2000). Then

$$P[\|\mathbf{X}_{\cdot,j}\|_2^2 > 2n+6\log p] \leq P[\|\mathbf{X}_{\cdot,j}\|_2^2 > n+2\sqrt{2n\log p}+4\log p] \leq \exp(-2\log p),$$

and a union bound over $j = 1, \ldots, p$ yields (S2.1). Also, $P[\|\mathbf{X}_{\cdot,j}\|_2^2 > 2n] \le \exp(-n/8)$, and as $s \ll n$, a union bound over $j = 1, \ldots, s$ yields (S2.2). For (S2.3) and (S2.4),

$$\|\hat{\mathbf{\Sigma}}_{SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\|_2 \le \|\mathbf{\Sigma}_{SS}^{-1/2}\|_2 \|\mathbf{\Sigma}_{SS}^{1/2}\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{\Sigma}_{SS}^{1/2} - \mathbf{Id}\|_2 \|\mathbf{\Sigma}_{SS}^{-1/2}\|_2 \le C_{\min}^{-1}\|\tilde{\mathbf{\Sigma}}_{SS}^{-1} - \mathbf{Id}\|_2$$

where $\tilde{\mathbf{\Sigma}}_{SS} \stackrel{L}{=} n^{-1}\mathbf{Z}^T\mathbf{Z}$ for $\mathbf{Z} \in \mathbb{R}^{n \times s}$ having i.i.d. standard Gaussian entries. Corollary 5.35 of Vershynin (2012) implies

$$\left(1 - \frac{\sqrt{s} + \log n}{\sqrt{n}}\right)^2 \le \lambda_{\min}(\tilde{\mathbf{\Sigma}}_{SS}) \le \lambda_{\max}(\tilde{\mathbf{\Sigma}}_{SS}) \le \left(1 + \frac{\sqrt{s} + \log n}{\sqrt{n}}\right)^2$$

with probability approaching 1. As $s \ll n$, this implies for any $\delta > 0$, with probability approaching 1

$$\|\tilde{\mathbf{\Sigma}}_{SS}^{-1} - \mathbf{Id}\|_2 \le (2 + \delta)\left(\frac{\sqrt{s} + \log n}{\sqrt{n}}\right).$$

Then (S2.3) follows from $\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_2 \le \|\hat{\mathbf{\Sigma}}_{SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\|_2 + \|\mathbf{\Sigma}_{SS}^{-1}\|_2 \le 2C_{\min}^{-1}$, and (S2.4) from

$$\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_\infty \le \|\hat{\mathbf{\Sigma}}_{SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\|_\infty + \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty \le \sqrt{s}\|\hat{\mathbf{\Sigma}}_{SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\|_2 + \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty \le \frac{3(s + \sqrt{s}\log n)}{C_{\min}\sqrt{n}} + \|\mathbf{\Sigma}_{SS}^{-1}\|_\infty.$$

For the remaining three statements, denote $\mathbf{S} = \operatorname{diag}(\sigma_{j+1}, \ldots, \sigma_p) \in \mathbb{R}^{(p-s)\times(p-s)}$, so $\mathbf{W} = \mathbf{ZS}$ where $\mathbf{Z} \in \mathbb{R}^{n \times (p-s)}$ is independent of $\mathbf{X}_S$ with i.i.d. standard Gaussian entries. Denote $\mathbf{P} = \frac{1}{\sqrt{n}}\hat{\mathbf{\Sigma}}_{SS}^{-1/2}\mathbf{X}_S^T$, so that $\mathbf{P}^T\mathbf{P}$ is the projection in $\mathbb{R}^n$ onto the column span of $\mathbf{X}_S$. With probability 1, this column span is of rank s, so $\mathbf{P}$ is an orthogonal projection from $\mathbb{R}^n$ to $\mathbb{R}^s$.

Applying $\sigma_j \leq 1$ for each j,

$$\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\mathbf{X}_S^T\mathbf{W}\|_{\infty,2} = \sqrt{n}\|\hat{\mathbf{\Sigma}}_{SS}^{-1/2}\mathbf{PZS}\|_{\infty,2} \leq \sqrt{n}\|\hat{\mathbf{\Sigma}}_{SS}^{-1/2}\mathbf{PZ}\|_{\infty,2}.$$

Conditional on $\mathbf{X}_S$, the columns of $\hat{\mathbf{\Sigma}}_{SS}^{-1/2}\mathbf{PZ}$ are independent and distributed as $N(0, \hat{\mathbf{\Sigma}}_{SS}^{-1})$, so each ith row of $\hat{\mathbf{\Sigma}}_{SS}^{-1/2}\mathbf{PZ}$ consists of independent Gaussian entries with variance $(\hat{\mathbf{\Sigma}}_{SS}^{-1})_{ii} \leq \|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_2$. Then by (S2.16),

$$P[\|(\hat{\mathbf{\Sigma}}_{SS}^{-1/2}\mathbf{PZ})_{i,\cdot}\|_2^2 > 2p\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_2 \mid \mathbf{X}_S] \leq \exp(-p/8),$$

and (S2.5) follows by taking a union bound over $i = 1, \ldots, s$, recalling $s \leq p$, and applying (S2.3). Similarly, $\|\mathbf{W}^T\mathbf{X}_S\mathbf{\Sigma}_{SS}^{-1}\|_{\infty,2} \leq \sqrt{n}\|\mathbf{Z}^T\mathbf{P}^T\mathbf{\Sigma}_{SS}^{-1/2}\|_{\infty,2}$, and conditional on $\mathbf{X}_S$ each row of $\mathbf{Z}^T\mathbf{P}^T\hat{\mathbf{\Sigma}}_{SS}^{-1}$ is distributed as $N(0, \hat{\mathbf{\Sigma}}_{SS}^{-1})$. Then (S2.16) implies

$$P[\|(\mathbf{Z}^T\mathbf{P}^T\hat{\mathbf{\Sigma}}_{SS}^{-1/2})_{j,\cdot}\|_2^2 > (2s + 6\log p)\|\hat{\mathbf{\Sigma}}_{SS}^{-1}\|_2 \mid \mathbf{X}_S] \leq \exp(-2\log p),$$

and (S2.3) and a union bound over $j = s+1, \ldots, p$ yields (S2.6). Finally,

$$\|\mathbf{W}^T(\mathbf{Id} - \tfrac{1}{n}\mathbf{X}_S\mathbf{\Sigma}_{SS}^{-1}\mathbf{X}_S^T)\mathbf{W}\|_{\infty,2} \leq \|\mathbf{Z}^T(\mathbf{Id} - \mathbf{P}^T\mathbf{P})\mathbf{Z}\|_{\infty,2},$$

and conditional on $\mathbf{X}_S$, $\mathbf{Z}^T(\mathbf{Id} - \mathbf{P}^T\mathbf{P})\mathbf{Z}$ is equal in law to $\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}$ where $\tilde{\mathbf{Z}} \in \mathbb{R}^{(n-s)\times(p-s)}$ has i.i.d. standard Gaussian entries. Writing $\|\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}\|_{\infty,2} \leq \|\tilde{\mathbf{Z}}^T\|_{\infty,2}\|\tilde{\mathbf{Z}}\|_2$, Corollary 5.35 of Vershynin (2012) implies $\|\tilde{\mathbf{Z}}\|_2 \leq \sqrt{2n} + \sqrt{p}$ with probability approaching 1, while (S2.16) implies $\|\tilde{\mathbf{Z}}\|_{\infty,2}^2 \leq 2n + 6\log p$ with probability approaching 1. Then (S2.7) follows from combining these bounds and observing $n\log p \ll np$.

Proof of Lemma 2

Denote by $J_{eo}(\mathbf{B})$ the objective function in (S2.8) and by $J_{\text{restricted}}(\mathbf{B})$ the objective function in (S2.9). (The former is a function of $\mathbf{B} \in \mathbb{R}^{p\times p}$: $\mathbf{B}_{ii} = 0$ and the latter of $\mathbf{B} \in \mathbb{R}^{s\times p}$: $\mathbf{B}_{ii} = 0$.) If $\mathbf{X}_S^T\mathbf{X}_S$ is invertible, then $J_{\text{restricted}}$ is strictly convex and $|J_{\text{restricted}}(\mathbf{B})| \to \infty$ as $\|\mathbf{B}\|_F \to \infty$, hence there is a unique solution $\hat{\mathbf{B}}_{\text{restricted}}$ to (S2.9). Denote by ∂J_{eo} and $\partial J_{\text{restricted}}$ the subdifferentials of J_{eo} and $J_{\text{restricted}}$. Note that $\|\mathbf{X} - \mathbf{X}\mathbf{B}\|_F^2$ is differentiable in $\mathbf{B}$ and the penalty decomposes across rows of $\mathbf{B}$, hence $\partial J_{eo}(\mathbf{B}) = \mathcal{D}_1(\mathbf{B}) \times \cdots \times \mathcal{D}_p(\mathbf{B})$, where $\mathcal{D}_i(\mathbf{B})$ is the set of vectors of the form

$$-\frac{1}{n}\mathbf{X}_{\cdot,i}^T(\mathbf{X}_{\cdot,-i} - \mathbf{X}\mathbf{B}_{\cdot,-i}) + \theta_n \begin{cases} \mathbf{B}_{i,-i}/\|\mathbf{B}_{i,-i}\|_2 & \mathbf{B}_{i,-i} \neq 0 \\ \{\mathbf{Z}_{i,-i} : \|\mathbf{Z}_{i,-i}\|_2 \leq 1\} & \mathbf{B}_{i,-i} = 0 \end{cases}$$

where $\mathbf{X}_{\cdot,-i}$ and $\mathbf{B}_{\cdot,-i}$ denote $\mathbf{X}$ and $\mathbf{B}$ with ith columns removed. Similarly, $\partial J_{\text{restricted}}(\mathbf{B}) = \mathcal{D}_1(\mathbf{B})' \times \cdots \times \mathcal{D}_s(\mathbf{B})'$ where $\mathcal{D}_i(\mathbf{B})'$ is the set of vectors of the form

$$-\frac{1}{n}\mathbf{X}_{\cdot,i}^T(\mathbf{X}_{\cdot,-i} - \mathbf{X}_S\mathbf{B}_{\cdot,-i}) + \theta_n \begin{cases} \mathbf{B}_{i,-i}/\|\mathbf{B}_{i,-i}\|_2 & \mathbf{B}_{i,-i} \neq 0 \\ \{\mathbf{Z}_{i,-i} : \|\mathbf{Z}_{i,-i}\|_2 \leq 1\} & \mathbf{B}_{i,-i} = 0. \end{cases}$$

As $\mathbf{X}\hat{\mathbf{B}}_{eo} = \mathbf{X}_S\hat{\mathbf{B}}_{\text{restricted}}$, we have $\mathcal{D}_i(\hat{\mathbf{B}}_{eo}) = \mathcal{D}_i(\hat{\mathbf{B}}_{\text{restricted}})'$ for each $i \in S$. By optimality of $\hat{\mathbf{B}}_{\text{restricted}}$ for (S2.9), $0 \in \partial J_{\text{restricted}}(\hat{\mathbf{B}}_{\text{restricted}})$, hence $0 \in$

$\partial\mathcal{D}_i(\hat{\mathbf{B}}_{\text{restricted}})' = \mathcal{D}_i(\hat{\mathbf{B}}_{eo})$ for each $i \in S$. On the other hand, condition (S2.10) implies $0 \in \partial\mathcal{D}_i(\hat{\mathbf{B}}_{eo})$ for each $i \in S^C$. Then $0 \in \partial J_{eo}(\hat{\mathbf{B}}_{eo})$, so $\hat{\mathbf{B}}_{eo}$ solves (S2.8). In fact, the strict inequality in condition (S2.10) implies that 0 is in the interior of $\mathcal{D}_i(\hat{\mathbf{B}}_{eo})$ for each $i \in S^C$. If $\tilde{\mathbf{B}}$ is any solution to (S2.9), then $\operatorname{Tr}\mathbf{D}^T(\tilde{\mathbf{B}} - \hat{\mathbf{B}}_{eo}) \leq 0$ for any $\mathbf{D} \in \partial J_{eo}(\hat{\mathbf{B}}_{eo})$, which implies $(\tilde{\mathbf{B}} - \hat{\mathbf{B}}_{eo})_{i,\cdot} = \tilde{\mathbf{B}}_{i,\cdot} = 0$ for all $i \in S^C$. As $\hat{\mathbf{B}}_{\text{restricted}}$ is the unique solution to (S2.9), this implies $\tilde{\mathbf{B}} = \hat{\mathbf{B}}_{eo}$, so $\hat{\mathbf{B}}_{eo}$ is the unique solution to (S2.8).

Proof of Lemma 3

Let $\mathcal{D}_i(\hat{\mathbf{B}})'$ for $i \in S$ be as in the proof of Lemma 2 above. Optimality of $\hat{\mathbf{B}}$ implies $0 \in \mathcal{D}_i(\hat{\mathbf{B}})'$ for each $i \in S$, i.e. for some $\mathbf{Z} \in \mathcal{Z}$,

$$0 = -\frac{1}{n}\mathbf{X}_{\cdot,i}^T(\mathbf{X} - \mathbf{X}_S\hat{\mathbf{B}}) + \theta_n\mathbf{Z}_{i,\cdot} + \frac{1}{n}\mathbf{X}_{\cdot,i}^T(0, \ldots, 0, \mathbf{X}_{\cdot,i} - \mathbf{X}_S\hat{\mathbf{B}}_{\cdot,i}, 0, \ldots, 0).$$

Combining this condition across $i \in S$ and recalling $\mathbf{X} = (\mathbf{X}_S, \mathbf{X}_S\mathbf{\Gamma} + \mathbf{W}) = (\mathbf{X}_S, \mathbf{W}) + \mathbf{X}_S\mathbf{B}^*$,

$$0 = -\frac{1}{n}\mathbf{X}_S^T(\mathbf{X} - \mathbf{X}_S\hat{\mathbf{B}}) + \theta_n\mathbf{Z} + \frac{1}{n}\mathbf{\Delta} = -\frac{1}{n}\mathbf{X}_S^T(\mathbf{X}_S, \mathbf{W}) - \hat{\mathbf{\Sigma}}_{SS}(\mathbf{B}^* - \hat{\mathbf{B}}) + \theta_n\mathbf{Z} + \frac{1}{n}\mathbf{\Delta}.$$

The lemma follows by rearranging and substituting the definition of $\mathbf{U}$.

S3 Comparisons between hubNet and other methods

S3.1 Comparisons of simulation results between hubNet, lasso, elasticNet and adaptive lasso

We first compare performance under different settings between four methods: hubNet, lasso, elastic net and the adaptive lasso with weights set to the inverse absolute values of the univariate regression coefficients.

We experimented with the following four scenarios:

(a) A favorable model:

$$Y = \mathbf{X}_S\beta + \epsilon,\ \beta = \mathbf{1},\ \epsilon \sim N(0,1)$$

$$X_j = \mathbf{X}_S\Gamma_j + \epsilon_j,\ j \in T,\ \Gamma_{ij} \sim N(0,4),\ \epsilon_j \sim N(0,1)$$

$$X_j = \epsilon_j,\ j \notin T,\ \epsilon_j \sim N(0,1)$$

The set S contains the first s features, and T contains 20% of the remaining features. Hence the model (2.6) is correct but with only 20% of non-core features depending on $\mathbf{X}_S$.

(b) An adversarial model:

$$Y = \mathbf{X}_{S_1}\beta + \epsilon,\ \beta = \mathbf{1},\ \epsilon \sim N(0,1)$$

$$X_j = \mathbf{X}_{S_2}\Gamma_j + \epsilon_j,\ j \in T,\ \Gamma_{ij} \sim N(0,0.25),\ \epsilon_j \sim N(0,1)$$

$$X_j = \epsilon_j,\ j \notin S_2 \cup T$$

S_2 contains the first s features and T contains 20% of the remaining features, of which s belong to S_1. Hence a core set S_2 influences T, but Y is explained directly by certain features in T rather than $\mathbf{X}_{S_2}$.

(c) An extreme adversarial model:

$$Y = \mathbf{X}_{S_1}\beta + \epsilon,\ \beta = \mathbf{1},\ \epsilon \sim N(0,1)$$

$$X_j = \mathbf{X}_{S_2}\Gamma_j + \epsilon_j,\ j \notin S_2,\ \Gamma_{ij} \sim N(0, 0.25),\ \epsilon_j \sim N(0,1)$$

$$X_j = \epsilon_j,\ j \in S_2$$

S_2 contains the first s features and S_1 contains the next s features. This setup is the same as in (b) above, except T is now the set of all features outside S_2.

(d) A neutral model:

$$Y = \mathbf{X}_S\beta + \epsilon,\ \beta = \mathbf{1},\ \epsilon \sim N(0,1)$$

$$X \sim N(0, \mathbf{\Sigma})$$

S contains the first s features, and $\mathbf{\Sigma}$ is a random positive-definite covariance matrix (generated using the R function `genPositiveDefMat`) with the ratio of largest to smallest eigenvalue set to 10.

For each scenario, we consider $(n, p, s) = (100, 500, 10)$ and $(200, 1000, 20)$, and we also scale each feature to have variance 1 before applying each of the four methods. For hubNet, the edge-out tuning parameter θ is set by minimizing GCV, and we fix $\gamma = 1/2$. For the elastic net, we also fix $\alpha = 1/2$. The main tuning parameter λ in all four methods (corresponding to the tuning parameter for the adaptive lasso step in hubNet) is set by 10-fold cross-validation.

We evaluate performance using the proportion of falsely detected features (FP), the proportion of true features that are undetected (FN), the cross-validation mean square prediction error in the training set (cvm), mean square prediction error in the test set, and the total number of selected features. A summary of these values averaged across 100 repetitions of each scenario is presented in Tables 1 to 4, with standard deviations reported for cvm and test error.

HubNet outperforms the other three methods in scenario (a) as expected. Perhaps surprisingly, it also seems to outperform the other methods under scenarios (b) and (d). In the extreme adversarial scenario (c), hubNet performs worse than the other methods, although this can be detected in cross-validation.

In Figure 1 ,we track FP and FN along the solution paths of the various methods as λ varies. The results are in line with the above.

Table 1: Comparison of hubNet with other methods in scenario (a)

$(n, p, s) = (100, 500, 10)$					
	cvm(se)	FN	FP	features	test.error(se)
lasso	1.555(0.297)	0.94	0.98	27.44	1.634(0.313)
elasticNet	1.599(0.298)	0.90	0.97	40.69	1.685(0.317)
adaptiveLasso	1.251(0.21)	0.93	0.97	24.31	1.497(0.268)
hubNet	1.199(0.201)	0.00	0.25	15.55	1.3(0.227)
$(n, p, s) = (200, 1000, 20)$					
	cvm(se)	FN	FP	features	test.error(se)
lasso	1.55(0.196)	0.94	0.98	58.33	1.631(0.242)
elasticNet	1.564(0.183)	0.91	0.97	72.76	1.638(0.239)
adaptiveLasso	1.279(0.136)	0.94	0.97	36.65	1.421(0.193)
hubNet	1.174(0.12)	0.00	0.19	25.83	1.256(0.153)

Table 2: Comparison of hubNet with other methods in scenario (b)

$(n, p, s) = (100, 500, 10)$					
	cvm(se)	FN	FP	features	test.error(se)
lasso	5.479(2.233)	0.03	0.85	66.33	4.588(2.239)
elasticNet	7.017(2.156)	0.05	0.86	72.94	6.14(2.563)
adaptiveLasso	4.878(1.773)	0.16	0.79	41.65	5.867(2.623)
hubNet	3.716(1.405)	0.01	0.77	44.37	3.247(1.394)
$(n, p, s) = (200, 1000, 20)$					
	cvm(se)	FN	FP	features	test.error(se)
lasso	15.277(4.159)	0.13	0.85	126.80	12.611(5.519)
elasticNet	17.328(3.555)	0.15	0.86	126.91	15.485(4.568)
adaptiveLasso	12.125(2.536)	0.22	0.76	67.57	13.183(3.658)
hubNet	6.685(3.369)	0.02	0.67	61.82	6.011(3.117)

Table 3: Comparison of hubNet with other methods in scenario (c)

$(n, p, s) = (100, 500, 10)$					
	cvm(se)	FN	FP	features	test.error(se)
lasso	2.619(0.821)	0.00	0.82	57.68	2.531(0.807)
elasticNet	3.53(1.183)	0.00	0.86	71.89	3.143(0.984)
adaptiveLasso	5.988(1.889)	0.19	0.79	40.86	6.258(2.086)
hubNet	4.815(1.988)	0.08	0.53	19.77	4.751(2.288)
$(n, p, s) = (200, 1000, 20)$					
	cvm(se)	FN	FP	features	test.error(se)
lasso	2.776(0.525)	0.00	0.77	86.72	2.866(0.642)
elasticNet	3.915(0.809)	0.00	0.80	99.71	3.664(0.877)
adaptiveLasso	13.466(2.344)	0.24	0.80	77.10	13.135(2.883)
hubNet	21.302(4.784)	0.78	0.85	23.26	21.209(5.111)

Table 4: Comparison of hubNet with other methods in scenario (d)

$(n, p, s) = (100, 500, 10)$					
	cvm(se)	FN	FP	features	test.error(se)
lasso	2.486(0.515)	0.00	0.80	54.21	2.683(0.779)
elasticNet	3.948(1.111)	0.00	0.85	69.60	3.649(1.323)
adaptiveLasso	2.038(1.632)	0.01	0.70	37.96	3.085(2.723)
hubNet	1.717(0.356)	0.00	0.72	39.00	2.16(0.619)
$(n, p, s) = (200, 1000, 20)$					
	cvm(se)	FN	FP	features	test.error(se)
lasso	2.38(0.364)	0.00	0.80	104.40	2.668(0.623)
elasticNet	3.374(0.694)	0.00	0.84	126.78	3.317(0.889)
adaptiveLasso	3.475(1.825)	0.02	0.49	41.74	4.615(2.687)
hubNet	1.647(0.207)	0.00	0.69	66.77	2.137(0.416)

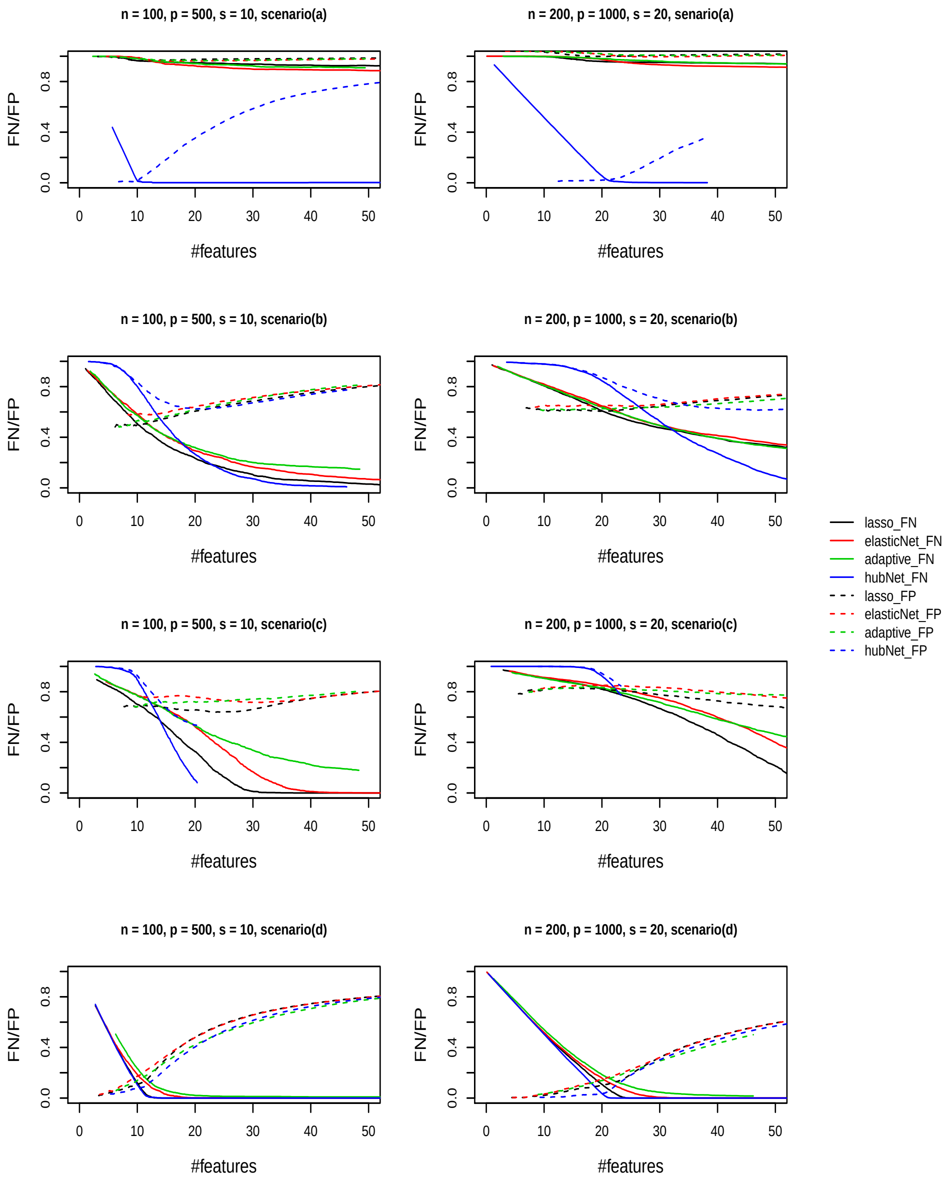

Figure 1: *False positive and false negative paths under four generating models.*

Table 5: *Comparisons between hubNet , PCR and sparse PCR on two real data sets.*

		cvm(se)	Num. features	test error
Kidney Cancer Data	hubNet	9.98(0.40)	1	0.008
$p = 14814$	PCR	11.1(0.33)	10	0.36
$n_{\text{train}} = 88, n_{\text{test}} = 89$	SPCR(10 non-zeros)	10.3(0.60)	7	0.456
	SPCR(50 non-zeros)	10.1(0.40)	1	0.564
	SPCR(100 non-zeros)	10.0(0.40)	1	0.137
		cvm(se)	Num. features	test p-value
DLBCL-patient Data	hubNet	10.9(0.36)	21	0.020
$p = 7399$	PCR	11.1(0.33)	0	–
$n_{\text{train}} = 156, n_{\text{test}} = 79$	SPCR(10 non-zeros)	11.01(0.40)	18	0.738
	SPCR(50 non-zeros)	11.06(0.35)	7	0.829
	SPCR(100 non-zeros)	11.07(0.26)	1	0.473

S3.2 Comparisons of hubNet, PC regression and sparse PC regression on real datasets

The comparisons between hubNet, PCR and sparse PCR are summarized in Table 5, and plots of test p-values versus number of non-zero features are given in Figures 2 and 3. The model trained using hubNet has much better performance on the test data set, and it uses original features which are easier to interpret and validate.

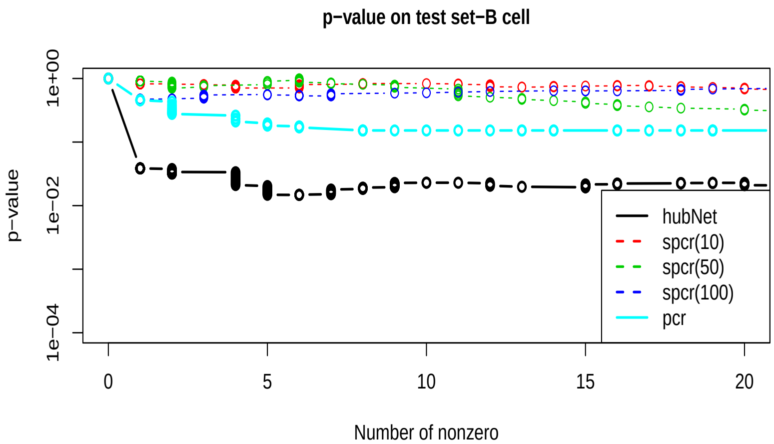

Figure 2: *p-values of LR statistics for B-cell lymphoma dataset*

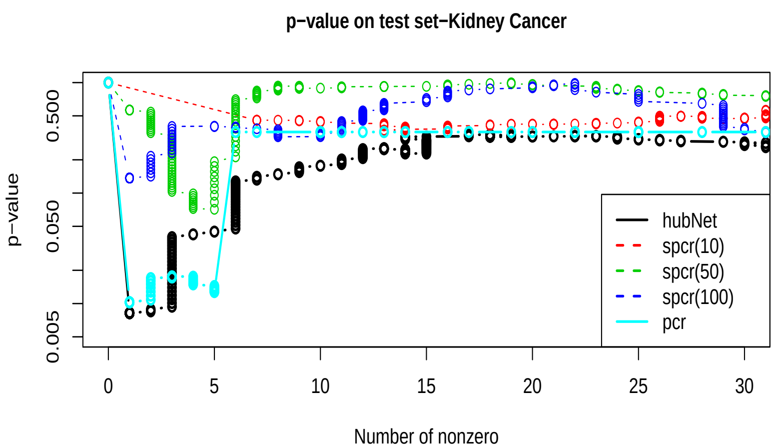

Figure 3: *p-values of LR statistics for Kidney cancer dataset*

S4 Recovery of hub nodes and speed comparisons

In this section, we compare the edge-out method with the `hglasso` method of Tan et al. (2014) in terms of computational speed and recovery of the underlying hub structure. We also compare the edge-out procedure with individual lasso regressions to show that the grouped ℓ_2 penalty can significantly improve the identification of hub predictors.

We generate $\mathbf{X}$ according to three settings:

1. For a core set S of size s, let $\mathbf{A} \in \{0,1\}^{p \times p}$ have all diagonal entries 1, all entries in row i and column i equal to 1 for all $i \in S$, and remaining entries 0. Define

$$\mathbf{E} = \begin{cases} 0 & \mathbf{A}_{ij} = 0 \\ \text{Unif}([-0.15, -0.015] \cup [0.015, 0.15]) & \text{otherwise,} \end{cases}$$

   $\bar{\mathbf{E}} = \frac{1}{2}(\mathbf{E} + \mathbf{E}^T)$, and $\boldsymbol{\Sigma}^{-1} = \bar{\mathbf{E}} + (0.2 - \lambda_{\min}(\bar{\mathbf{E}}))\mathbf{Id}$, and generate the rows of $\mathbf{X}$ from $N(0, \boldsymbol{\Sigma})$.

2. For two predictor sets S_1 and S_2 of sizes $s/2$, let

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{pmatrix}$$

with $\mathbf{A}_1, \mathbf{A}_2$ generated as above with core sets S_1, S_2. Construct $\mathbf{X}$ from $\mathbf{A}$ in the same way as above.

3. For a core set S of size s, generate $\mathbf{\Gamma} \in \mathbb{R}^{s\times(p-s)}$ with i.i.d. entries distributed as $N(0,4)$ truncated above and below at ± 2. Then generate each row $\mathbf{X}_{i,\cdot}$ of $\mathbf{X}$ such that $\mathbf{X}_{ij} \sim N(0,1)$ for $j \in S$ and $\mathbf{X}_{ij} = \mathbf{X}_{i,S}\mathbf{\Gamma}_{\cdot,j} + \epsilon_{ij}$ for $j \notin S$ and $\epsilon_{ij} \sim N(0,1)$.

In each setting, we re-standardize the predictors to have variance 1.

We set $(n,p,s) = (100, 200, 4)$ and compare edge-out and hglasso by the number of correctly identified hub nodes as well as their corresponding absolute row sums in the estimated matrix. (This matrix is $\hat{\mathbf{B}}_{eo}$ for edge-out and $\hat{\mathbf{V}}^T$ in the hglasso decomposition $\mathbf{\Sigma}^{-1} = \mathbf{Z} + \mathbf{V} + \mathbf{V}^T$ where $\mathbf{Z}$ is sparse and $\mathbf{V}^T$ has few non-zero rows.) Edge-out was applied with only the ℓ_2 penalty (eol2) or with $\gamma = 0.5$ (eol12), and hglasso with $\lambda_1 = 1000$ and $\lambda_2 = 0.2$ or 0.5. Results are shown in Figure 4: the left column of the figure tracks the number of correctly identified hubs as the main tuning parameter (θ for edge-out and λ_3 for hglasso) varies, while the right column tracks the maximum rank of any hub node when all nodes are ranked in decreasing order of their absolute row sums. (A maximum rank of 4 indicates that all four hub nodes have larger absolute row sums than all remaining nodes.) We observe that both variants of edge-out perform well in all three settings;

hglasso performs well in settings 1 and 3 for $\lambda_2 = 0.2$ but not for setting 2 under the tested tuning parameters.

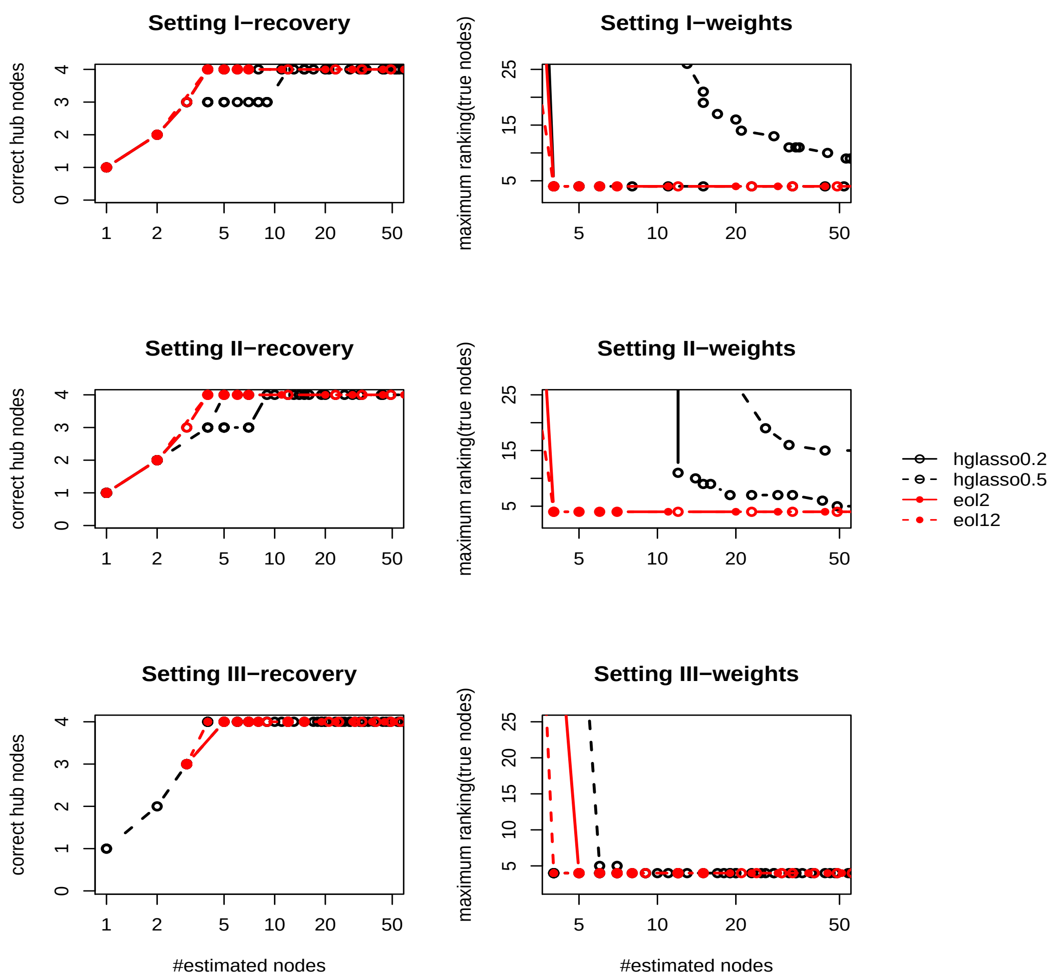

Figure 4: *Comparison of hub detection accuracy of* `edge-out` *and* `hglasso`*, by inclusion of hub predictors on the left and ranking of hub predictors on the right, as the number of total included predictors increases.*

Figure 5 compares the speed of these two methods, with one of n, p

fixed while the other grows. We see that the edge-out algorithm is much faster and appears to scale quadratically in p and linearly in n.

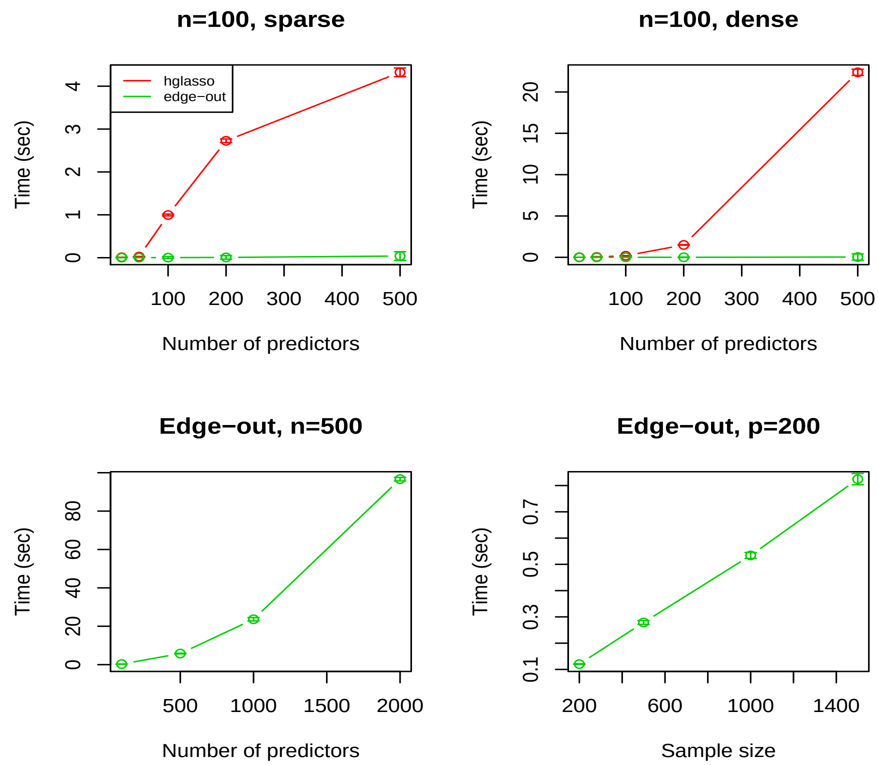

Figure 5: *Speed comparisons. In the top row we compare the computation times for the* `hglasso` *and* `edge-out` *algorithms, as the number of predictors increases, for sparse and dense problems. The bottom row examines just* `edge-out`*, with n or p fixed, for larger problems. We were not able to run* `hglasso` *in these latter settings.*

Next, we increase the number of hub predictors s to 10, and we compare edge-out with and without the ℓ_1 penalty to individual lasso regressions (corresponding to the special case of edge-out with $\gamma = 1$).

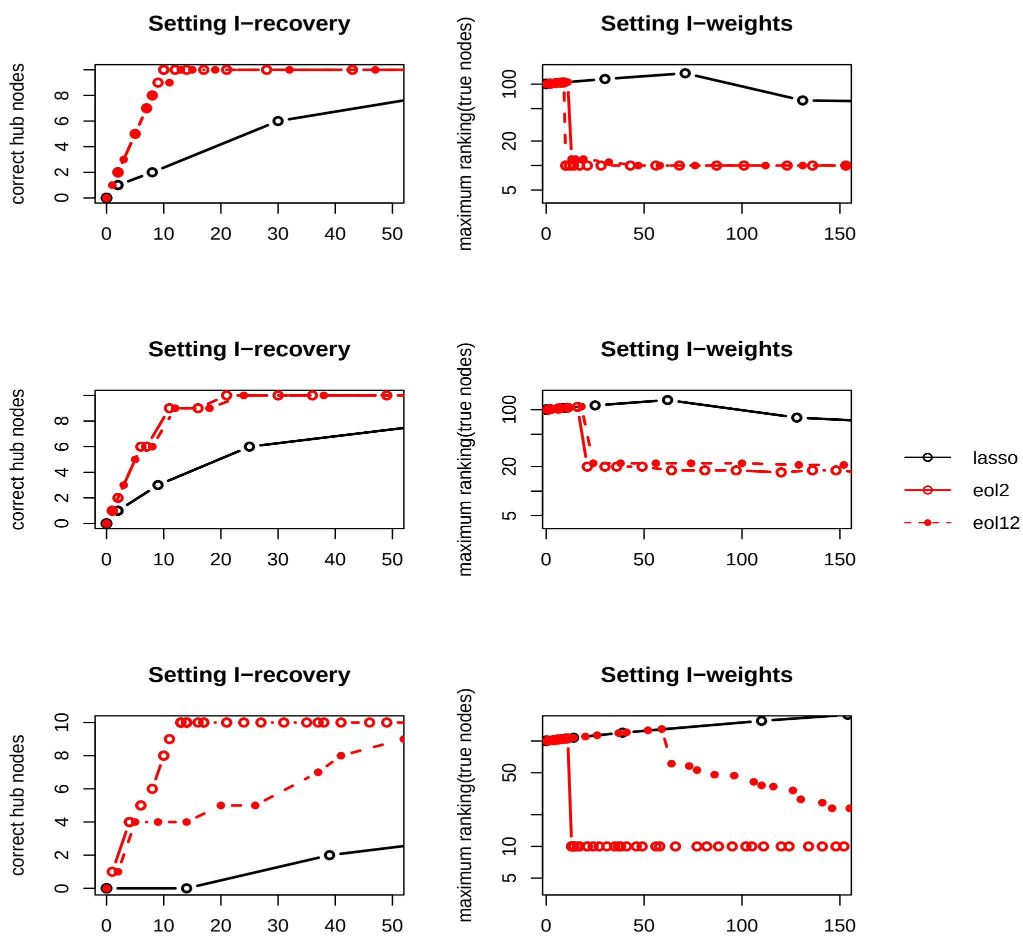

Figure 6: *Comparison of hub detection accuracy of* `edge-out` *with* $\gamma = 0$ *(only* ℓ_2 *penalty),* $\gamma = 1/2$ *(combined* ℓ_1 *and* ℓ_2 *penalty), and* $\gamma = 1$ *(individual lasso regressions), using the same metrics as in Figure 4.*

Performance is significantly better in all three examples when we include the ℓ_2 or grouped lasso penalty in edge-out, rather than using only the lasso penalty. Performance of edge-out with and without the ℓ_1 penalty is similar in the first and second examples, as the hub predictors have varying levels of influence on the other predictors, and many of these influences are small. In contrast, inclusion of the ℓ_1 penalty in the third example yields worse performance, because the hub predictors in this example have a strong influence on all of the non-hub predictors.

References

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47.

Peng, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to intregative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53–77.

Tan, K. M., London, P., Mohan, K., Lee, S.-I., Fazel, M., and Witten, D. M. (2014). Learning graphical models with hubs. *Journal of Machine Learning Research*, 15(1):3297–3331.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. C. and Kutyniok, G., editors, *Compressed Sensing*, pages 210–268. Cambridge University Press.

Zhou, S., van de Geer, S., and Bühlmann, P. (2009). Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*.