Supplementary Materials

Differential Analysis of Directed Networks

There are five parts. Firstly, we collect in Section 1 all notations used in our paper and here. We then describe the four conditions which help define the positive pair $\tilde{\tau}$ and $\tilde{\kappa}$ for Theorem 1, and further prove Theorem 1 in Section 2. In Section 3, we prove Theorem 2 which provides bounds for both estimation and prediction losses at the calibration stage. In Section 4, we prove Theorem 3 which provides bounds for both estimation and prediction losses at the construction stage. In Section 5, we prove the variable selection consistency in Theorem 4.

1 Notations

Unless otherwise claimed, we will follow the notations defined here throughout the paper and supplementary materials.

For a vector, $||\cdot||_2$ and $||\cdot||_1$ denote the ℓ_2 and ℓ_1 norms, respectively; $\|\cdot\|_\infty$ and $\|\cdot\|_{-\infty}$ are defined to be the maximum and minimum absolute values of its components, respectively; $|\cdot|_1$ implies taking element-wise absolute values of the vector so is itself a vector. For a matrix $A = (a_{ij})_{m\times n}$, $\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^m |a_{ij}|$, i.e., the maximum column sum of absolute values of its components, and $\|A\|_\infty = \max_{1\le i\le m}\sum_{j=1}^n |a_{ij}|$, i.e., the maximum row sum of absolute values of its components.

For a vector a and index set $\mathcal{S}$, a_i , a_{-i}, and $a_{\mathcal{S}}$ denote the i-th entry, the subvector excluding the i-th entry in a, and the subvector of a indexed by $\mathcal{S}$, respectively. For a matrix A, A_i and A_{-i} denote its i-th column and the submatrix of A excluding its i-th column, respectively. For a vector a_i and an index set $\mathcal{S}_i$ both sharing the same subscript, the subvector of a_i indexed by $\mathcal{S}_i$ is denoted by $a_{\mathcal{S}_i}$ for simplicity. Similarly, the submatrix of a matrix A_i including columns indexed by the set $\mathcal{S}_i$ is denoted by $A_{\mathcal{S}_i}$ for simplicity.

$a\vee b$ and $a\wedge b$ denote the maximum and minimum of a and b, respectively. $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of the corresponding matrix, respectively. $\mathbb{E}\,(\cdot)$ denotes the expectation, and $\mathbb{P}(\cdot)$ denotes the probability of an event. Symbol $\asymp$ denotes two terms at the same order. $\mathrm{tr}(\cdot)$ denotes the trace of the corresponding matrix. For a set S, $|S|$ denotes the number of its elements. For positive integers j and p, $j|p$ denotes the remainder of j when divided by p.

Throughout the paper and here, $C_1, C_2, \ldots$, $c_1, c_2, \ldots$, $\tilde{c}_1, \tilde{c}_2, \cdots$, $t_1, t_2, \ldots$ are some positive constant numbers.

2 The Conditions and Proof of Theorem 1

For each $k\in\{1,2\}$, the reduced model (3) includes p regression models, i.e., for $i=1,2,\cdots,p$,

$$\mathbf{Y}_i^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\pi}_i^{(k)} + \boldsymbol{\xi}_i^{(k)}.$$

Here we first state the four conditions in Fan and Lv [2008] which restrict the positive pairs $\tau^{(k)}$ and $\kappa^{(k)}$ so as to define $\tilde{\tau} = \max\{\tau^{(1)},\tau^{(2)}\}$ and $\tilde{\kappa} = \max\{\kappa^{(1)},\kappa^{(2)}\}$ for Theorem 1, and then prove that we can successfully screen variables for each of the above linear regression model.

Denote $Y_{ji}^{(k)}$, $\xi_{ji}^{(k)}$, and $\pi_{ji}^{(k)}$ as the j-th row of $Y_i^{(k)}$, $\boldsymbol{\xi}_i^{(k)}$, and $\boldsymbol{\pi}_i^{(k)}$, respectively. Further denote $\Sigma^{(k)}$ the variance-covariance matrix of the q random variables in observing $\mathbf{X}^{(k)}$. For any $\mathcal{M}\subset\{1,2,\cdots,q\}$, denote $\Sigma_{\mathcal{M}}^{(k)}$ the variance-covariance matrix of the random variables in observing $\mathbf{X}_{\mathcal{M}}^{(k)}$.

Condition 1. Each $\xi_{ji}^{(k)}$ is normally distributed with mean zero. $(\Sigma^{(k)})^{-1/2}\mathbf{X}^{(k)T}$ is observed from a spherically symmetric distribution, and has the concentration property: there exist some constants $\tilde{c}_1^{(k)}, \tilde{c}_2^{(k)} > 1$ and $\tilde{c}_3^{(k)} > 0$ such that, for any $\mathcal{M}\subset\{1,2,\cdots,q\}$ with $|\mathcal{M}|\ge \tilde{c}_1^{(k)}n^{(k)}$, the eigenvalues of $|\mathcal{M}|^{-1}\mathbf{X}_{\mathcal{M}}^{(k)}(\Sigma_{\mathcal{M}}^{(k)})^{-1/2}(\Sigma_{\mathcal{M}}^{(k)T})^{-1/2}\mathbf{X}_{\mathcal{M}}^{(k)T}$ are bounded either from above by $\tilde{c}_2^{(k)}$ or from below by $1/\tilde{c}_2^{(k)}$ with probability at least $1-\exp(-\tilde{c}_3^{(k)}n^{(k)})$.

Condition 2. $\mathrm{var}(Y_{ji}^{(k)}) = O(1)$. For some $\kappa^{(k)} \geq 0$, $\tilde{c}_4^{(k)} > 0$, and $\tilde{c}_5^{(k)} > 0$,

$$\min_{j \in \mathcal{M}_{i0}^{(k)}} \left| \boldsymbol{\pi}_{ji}^{(k)} \right| \geq \frac{\tilde{c}_4^{(k)}}{(n^{(k)})^{\kappa^{(k)}}} \quad \text{and} \quad \min_{j \in \mathcal{M}_{i0}^{(k)}} \left| \mathrm{cov}\left((\boldsymbol{\pi}_{ji}^{(k)})^{-1} Y_i^{(k)}, X_j^{(k)} \right) \right| \geq \tilde{c}_5^{(k)}.$$

Condition 3. $\log(q) = O((n^{(k)})^{\tilde{c}})$ for some $\tilde{c} \in (0, 1 - 2\kappa^{(k)})$.

Condition 4. There are some $\tau^{(k)} \geq 0$ and $\tilde{c}_6^{(k)} > 0$ such that $\lambda_{\max}(\Sigma^{(k)}) \leq \tilde{c}_6^{(k)} (n^{(k)})^{\tau^{(k)}}$.

Proof of Theorem 1. Following the *Sure Independence Screening Property* by Fan and Lv [2008], there exists some $\theta^{(k)} \in (0, 1 - 2\kappa^{(k)} - \tau^{(k)})$ such that, when $d^{(k)} = |\mathcal{M}_i^{(k)}| = O((n^{(k)})^{1-\theta^{(k)}})$, we have, for some constant $C > 0$,

$$\mathbb{P}(\mathcal{M}_{i0}^{(k)} \subseteq \mathcal{M}_i^{(k)}) = 1 - \mathcal{O}\left(\exp\left\{ -\frac{C (n^{(k)})^{1-2\kappa^{(k)}}}{\log(n^{(k)})} \right\} \right).$$

Let $\theta = \min(\theta^{(1)}, \theta^{(2)})$, then for $d^{(k)} = |\mathcal{M}_i^{(k)}| \equiv d = O(n_{\min}^{1-\theta})$, we have

$$\mathbb{P}(\mathcal{M}_{i0}^{(k)} \subseteq \mathcal{M}_i^{(k)}) = 1 - \mathcal{O}\left(\exp\left\{ -\frac{C (n^{(k)})^{1-2\tilde{\kappa}}}{\log(n^{(k)})} \right\} \right).$$

□

3 Proof of Theorem 2

Note that $\boldsymbol{\xi}^{(k)} = \mathcal{E}^{(k)}(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})^{-1}$ for $k \in \{1, 2\}$. Suppose that the singular values of both $(\mathbf{I} - \boldsymbol{\Gamma}^{(k)})$ are positively bounded from below by a constant c. Denote $\sigma_i^{(k)2} = \mathrm{var}(\epsilon_{ji}^{(k)})$ and $\tilde{\sigma}_i^{(k)2} = \mathrm{var}(\xi_{ji}^{(k)})$. Then $\tilde{\sigma}_i^{(k)} \leq \sigma_{p\max}/c = \max_{1 \leq i \leq p} (\sigma_i^{(1)} \vee \sigma_i^{(2)})$/c.

Lemma 1. *Under Assumptions 1-3, for each network $k \in \{1, 2\}$ in the calibration step, there exist positive constants $C_1^{(k)}$ and $C_2^{(k)}$ such that, with probability at least $1 - e^{-f^{(k)}}$,*

1. *(Estimation Loss)* $||\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)}||_2^2 \leq C_1^{(k)} \left(r_i^{(k)} \vee d \vee f^{(k)} \right) / n^{(k)}$;
2. *(Prediction Loss)* $||\mathbf{X}^{(k)}(\hat{\boldsymbol{\pi}}_i^{(k)} - \boldsymbol{\pi}_i^{(k)})||_2^2 / n^{(k)} \leq C_2^{(k)} \left(r_i^{(k)} \vee d \vee f^{(k)} \right) / n^{(k)}$.

Proof of Lemma 1. We have the closed form ridge estimator $\hat{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)}$ for the linear model $\mathbf{Y}_i^{(k)} = \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} \boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)} + \boldsymbol{\xi}_i^{(k)}$.

$$\hat{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)} = \left(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} + \lambda_i^{(k)} I_d \right)^{-1} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \mathbf{Y}_i^{(k)},$$

where $\lambda_i^{(k)}$ is the ridge tuning parameter. Plugging in the equation $\mathbf{Y}_i^{(k)} = \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} \boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)} + \boldsymbol{\xi}_i^{(k)}$, we have

$$\begin{aligned} \hat{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)} = & \left\{ \left(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} + \lambda_i^{(k)} I_d \right)^{-1} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} \boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)} \right\} \\ & + \left\{ \left(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} + \lambda_i^{(k)} I_d \right)^{-1} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \boldsymbol{\xi}_i^{(k)} \right\}. \end{aligned}$$

The difference between the ridge estimator $\hat{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)}$ and the true $\boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)}$ can be written as

$$\hat{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)} - \boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)} = -\lambda_i^{(k)} \left(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} + \lambda_i^{(k)} I_d \right)^{-1} \boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)} + \left(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)} + \lambda_i^{(k)} I_d \right)^{-1} \mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T} \boldsymbol{\xi}_i^{(k)}.$$

For simplicity, we denote the composite forms of $\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$ and $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ as follows,

$$\begin{aligned}\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} &= -\lambda_i^{(k)}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)} I_d\big)^{-1}\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}};\\ \tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}} &= \mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)} I_d\big)^{-1}.\end{aligned}$$

Then we have the following simplified form of the difference,

$$\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}} = \tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} + \tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}.$$

To obtain the ℓ_2 norm losses of estimation and prediction, we write

$$\begin{aligned}&||\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2^2\\ &= \underbrace{\tilde{\boldsymbol{\pi}}^{(k)T}_{\mathcal{M}_i^{(k)}}\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}}_{T_{21}} + \underbrace{2\tilde{\boldsymbol{\pi}}^{(k)T}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}}_{T_{22}} + \underbrace{\boldsymbol{\xi}_i^{(k)T}\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}}_{T_{23}},\\ &||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}\big)||_2^2\\ &= \underbrace{\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big)\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}}_{T_{24}} + \underbrace{2\tilde{\boldsymbol{\pi}}^{(k)T}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big)\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}}_{T_{25}}\\ &\quad + \underbrace{\boldsymbol{\xi}_i^{(k)T}\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big)\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}\boldsymbol{\xi}_i^{(k)}}_{T_{26}}.\end{aligned}$$

Firstly, we will derive the bound for T_{24}, T_{25} and T_{26} terms, then we can obtain similar results for term T_{21}, T_{22} and T_{23} by simply removing the matrix $\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$. Denote the singular value decomposition $\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} = U_i^{(k)T}V_i^{(k)}U_i^{(k)}$, where $U_i^{(k)}$ is a unitary matrix, $V_i^{(k)}$ is a diagonal matrix with eigenvalues v_i. Therefore, the shared component of $\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}}$ and $\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}$ can be rewritten as

$$\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)} I_d\big)^{-1} = U_i^{(k)T}\big(V_i^{(k)} + \lambda_i^{(k)} I_d\big)^{-1}U_i^{(k)}.$$

By Assumption 3, there are some constants c_1, c_2 such that $\max_{||\delta||_2=1}(n^{(k)})^{-1/2}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\delta||_2 \le c_1$ and $\min_{||\delta||_2=1}(n^{(k)})^{-1/2}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\delta||_2 \ge c_2$. Thus, $\lambda_{\max}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big) < c_1^2 n^{(k)}$ and $\lambda_{\min}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big) > c_2^2 n^{(k)}$. That is, $v_j \asymp n^{(k)}$ for each eigenvalue. Let $b = U_i^{(k)}\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$, then $||b||_2 = ||\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2$. Noting that $\lambda_i^{(k)} = o(n^{(k)})$ in Assumption 3, we can bound the term T_{24} as follows,

$$\begin{aligned}T_{24} &= \tilde{\boldsymbol{\pi}}^{(k)T}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big)\tilde{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} = \lambda_i^{(k)2} b^T V_i^{(k)}\big(V_i^{(k)} + \lambda_i^{(k)} I_d\big)^{-1}V_i^{(k)}\big(V_i^{(k)} + \lambda_i^{(k)} I_d\big)^{-1} b\\ &= \lambda_i^{(k)2}\sum_{j=1}^{d}\frac{v_j b_{ij}^2}{\big(v_j + \lambda_i^{(k)}\big)^2} = \mathcal{O}\left(\lambda_i^{(k)2}||\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2^2/n^{(k)}\right) = \mathcal{O}\left(r_i^{(k)}\right).\end{aligned} \tag{1}$$

Similarly, removing the term $\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$, we have

$$T_{21} = \mathcal{O}\left(\lambda_i^{(k)2}||\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2^2/n^{(k)}\right) = \mathcal{O}\left(r_i^{(k)}/n^{(k)}\right). \tag{2}$$

Noting that T_{25} follows a Gaussian distribution, we can write the probability of deviation of T_{25} with the classical Gaussian tail inequality, for any positive number t,

$$\mathbb{P}\left(T_{25} \le t\right) \ge 1 - \exp\left(-\frac{1}{2}t^2/\mathrm{var}(T_{25})\right).$$

Furthermore,

$$\begin{aligned}
\mathrm{var}(T_{25}) &= 4\tilde{\sigma}_i^{(k)2}\tilde{\boldsymbol{\pi}}_{\boldsymbol{(i)}}^{\boldsymbol{(k)T}}\big(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\big)\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\big(\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\big)\tilde{\boldsymbol{\pi}}_{\mathcal{M}_i^{(k)}}^{(k)} \\
&= 4\tilde{\sigma}_i^{(k)2}\lambda_i^{(k)2}b^T(V+\lambda_i^{(k)}I_d)^{-1}V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1} \\
&\quad \times V_i^{(k)}\big(V_i^{(k)}+\lambda_i^{(k)}I_d\big)^{-1}V_i^{(k)}\big(V_i^{(k)}+\lambda_i^{(k)}I_d\big)^{-1}b \\
&= 4\tilde{\sigma}_i^{(k)2}\lambda_i^{(k)2}\sum_{j=1}^{d}\frac{v_j^3 b_{ij}^2}{(v_j+\lambda_i^{(k)})^4} = \mathcal{O}\left(\tilde{\sigma}_i^{(k)2}\lambda_i^{(k)2}||\boldsymbol{\pi}_{\mathcal{M}_i^{(k)}}^{(k)}||_2^2/n^{(k)}\right) = \mathcal{O}\left(\tilde{\sigma}_i^{(k)2}r_i^{(k)}\right).
\end{aligned}$$

Letting $t=\sqrt{2\mathrm{var}(T_{25})(f^{(k)}+\log 2)}$, we obtain that, with probability at least $1-e^{-f^{(k)}}/2$,

$$T_{25}=\mathcal{O}\left(\sqrt{r_i^{(k)}f^{(k)}}\right). \tag{3}$$

Similarly, removing $\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}$, we can obtain that, concurring with (3),

$$T_{22}=\mathcal{O}\left(\sqrt{r_i^{(k)}f^{(k)}}/n^{(k)}\right). \tag{4}$$

The term T_{26} follows a non-central χ^2 distribution. We can invoke the Hanson-Wright inequality [Rudelson et al., 2013] to bound the probability of its extreme deviation, for some constant $t_2>0$,

$$\begin{aligned}
&\mathbb{P}(T_{26}\le \mathbb{E}\,(T_{26})+t) \\
&\ge\ 1-\exp\left\{\frac{-t^2t_2}{\tilde{\sigma}_i^{(k)4}||\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}||_F^2}\wedge\frac{-tt_2}{\tilde{\sigma}_i^{(k)2}||\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}||_{op}}\right\}.
\end{aligned} \tag{5}$$

To understand this probabilistic bound, we need to calculate $\mathbb{E}\,(T_{26})$ and the two involved norms. Firstly,

$$\begin{aligned}
\mathbb{E}\,(T_{26}) &= \tilde{\sigma}_i^{(k)2}\mathrm{tr}\left(\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}\right) \\
&= \tilde{\sigma}_i^{(k)2}\mathrm{tr}\left(V_i^{(k)}\big(V_i^{(k)}+\lambda_i^{(k)}I_d\big)^{-1}V_i^{(k)}\big(V_i^{(k)}+\lambda_i^{(k)}I_d\big)^{-1}\right) \\
&= \tilde{\sigma}_i^{(k)2}\sum_{j=1}^{d}\frac{v_j^2}{(v_j+\lambda_i^{(k)})^2}=\mathcal{O}\left(d\tilde{\sigma}_i^{(k)2}\right).
\end{aligned} \tag{6}$$

The Frobenius norm can be simplified as follows,

$$\begin{aligned}
&||\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}||_F^2 \\
&= \mathrm{tr}\left(\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)T}\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)}\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)T}\right) \\
&= \mathrm{tr}\left(((\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)})^T\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)})(\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)})^T\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}((\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)})^T\mathbf{X}_{\mathcal{M}_i^{(k)}}^{(k)})(\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)})^T\tilde{\mathbf{X}}_{\mathcal{M}_i^{(k)}}^{(k)}\right) \\
&= \mathrm{tr}\left(V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}V_i^{(k)}(V_i^{(k)}+\lambda_i^{(k)}I_d)^{-1}\right) \\
&= \sum_{j=1}^{d}\frac{v_j^4}{(v_j+\lambda_i^{(k)})^4}=\mathcal{O}\,(d).
\end{aligned} \tag{7}$$

Note that $\lambda_{\max}(\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}) \asymp n^{(k)}$, then, the operator norm can be simplified as follows,

$$
\begin{aligned}
&||\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}||_{op} \\
&= \quad ||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)} I_d\big)^{-1}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\big(\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}} + \lambda_i^{(k)} I_d\big)^{-1}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}||_{op} \\
&= \quad \mathcal{O}\left(\lambda_{\max}\big(\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\big)/n^{(k)2}\right) = \mathcal{O}(1).
\end{aligned} \tag{8}
$$

Letting

$$
\begin{aligned}
t = &\sqrt{\tilde{\sigma}_i^{(k)4}||\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}||_F^2 \times (f^{(k)} + \log 2)/t_2} \\
&\vee \left(\tilde{\sigma}_i^{(k)2}||\tilde{\mathbf{X}}^{(k)}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}\tilde{\mathbf{X}}^{(k)T}_{\mathcal{M}_i^{(k)}}||_{op} \times (f^{(k)} + \log 2)/t_2\right),
\end{aligned}
$$

and combining (5), (6), (7), and (8), we obtain that, with probability at least $1 - e^{-f^{(k)}}/2$,

$$
T_{26} = \mathcal{O}\left(d \vee \sqrt{df^{(k)}} \vee f^{(k)}\right). \tag{9}
$$

Similarly, removing $\mathbf{X}^{(k)T}_{\mathcal{M}_i^{(k)}}\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$, we can obtain that, concurring with (9),

$$
T_{23} = \mathcal{O}\left(\big(d \vee \sqrt{d\, f^{(k)}} \vee f^{(k)}\big)/n^{(k)}\right). \tag{10}
$$

Collecting the bounds (1), (3), (9) and noting the definition of $\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}$ and $\boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}$, we conclude there exists some constant $C_2^{(k)} > 0$ such that, with probability at least $1 - e^{-f^{(k)}}$,

$$
\frac{1}{n^{(k)}}||\mathbf{X}^{(k)}(\hat{\boldsymbol{\pi}}^{(k)} - \boldsymbol{\pi}^{(k)})||_2^2 = \frac{1}{n^{(k)}}||\mathbf{X}^{(k)}_{\mathcal{M}_i^{(k)}}(\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}})||_2^2 \le C_2^{(k)}\frac{r_i^{(k)} \vee d \vee f^{(k)}}{n^{(k)}}.
$$

Similarly, collecting the bound (2), (4) and (10), we conclude there exists some constant $C_1^{(k)} > 0$ such that, with probability at least $1 - e^{-f^{(k)}}$,

$$
||\hat{\boldsymbol{\pi}}^{(k)} - \boldsymbol{\pi}^{(k)}||_2^2 = ||\hat{\boldsymbol{\pi}}^{(k)}_{\mathcal{M}_i^{(k)}} - \boldsymbol{\pi}^{(k)}_{\mathcal{M}_i^{(k)}}||_2^2 \le C_1^{(k)}\frac{r_i^{(k)} \vee d \vee f^{(k)}}{n^{(k)}}.
$$

This concludes the proof of Lemma 1. □

To bound the estimation loss, we write

$$
||\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j||_2^2 = ||\hat{\boldsymbol{\pi}}^{(1)}_{j|p} - \boldsymbol{\pi}^{(1)}_{j|p}||_2^2 + ||\hat{\boldsymbol{\pi}}^{(2)}_{j|p} - \boldsymbol{\pi}^{(2)}_{j|p}||_2^2,
$$

where $\boldsymbol{\pi}^{(k)}_{j|p}$ and $\hat{\boldsymbol{\pi}}^{(k)}_{j|p}$ are the $j|p$ columns of $\boldsymbol{\pi}^{(k)}$ and $\hat{\boldsymbol{\pi}}^{(k)}$, respectively. Following the bounds in Lemma 1 for both networks, we obtain the overall estimation bound as, with probability at least $1 - e^{-f^{(1)}} - e^{-f^{(2)}}$,

$$
\begin{aligned}
||\hat{\boldsymbol{\Pi}}_j - \boldsymbol{\Pi}_j||_2^2 &\le C_1^{(1)}\frac{r^{(1)}_{j|p} \vee d \vee f^{(1)}}{n^{(1)}} + C_1^{(2)}\frac{r^{(2)}_{j|p} \vee d \vee f^{(2)}}{n^{(2)}} \\
&\le \big(C_1^{(1)} + C_1^{(2)}\big)\frac{\big(r^{(2)}_{j|p} \vee d \vee f^{(2)}\big) \vee \big(r^{(2)}_{j|p} \vee d \vee f^{(2)}\big)}{n^{(1)} \wedge n^{(2)}} \\
&= C_1\frac{d \vee \big(r^{(1)}_{j|p} \vee r^{(2)}_{j|p}\big) \vee \big(f^{(1)} \vee f^{(2)}\big)}{n^{(1)} \wedge n^{(2)}} \le C_1\frac{d \vee r_{\max} \vee f_{\max}}{n^{(1)} \wedge n^{(2)}},
\end{aligned}
$$

where $C_1 = C_1^{(1)} + C_1^{(2)}$. Similarly, we write the prediction bound as, with probability at least $1 - e^{-f^{(1)}} - e^{-f^{(2)}}$,

$$\begin{aligned}||\mathbf{X}(\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j)||_2^2 &= ||X^{(1)}(\hat{\boldsymbol{\pi}}_{j|p}^{(1)} - \boldsymbol{\pi}_{j|p}^{(1)})||_2^2 + ||X^{(2)}(\hat{\boldsymbol{\pi}}_{j|p}^{(2)} - \boldsymbol{\pi}_{j|p}^{(2)})||_2^2 \\ &\leq C_2^{(1)} \left\{ r_{j|p}^{(1)} \vee d \vee f^{(1)} + C_2^{(2)} r_{j|p}^{(2)} \vee d \vee f^{(2)} \right\} \\ &\leq C_2 \left\{ d \vee \left(r_{j|p}^{(1)} \vee r_{j|p}^{(2)} \right) \vee \left(f^{(1)} \vee f^{(2)} \right) \right\} \leq C_2 \left\{ d \vee r_{\max} \vee f_{\max} \right\},\end{aligned}$$

where $C_2 = C_2^{(1)} + C_2^{(2)}$ and $r_{\text{max}} = \max_{1 \leq i \leq p} (r_i^{(1)} \vee r_i^{(2)})$. This concludes the proof of Theorem 2.

4 Proof of Theorem 3

Let $c_{\text{max}} = c_1^{(1)} \vee c_1^{(2)}$, and further denote

$$g_n = C_2 \frac{d \vee r_{\max} \vee f_{\max}}{n} + 2c_{\text{max}} C_2 ||\mathbf{\Pi}||_1 \sqrt{\frac{d \vee r_{\max} \vee f_{\max}}{n}}.$$

Lemma 2. *Suppose that, for node i,*

$$\sqrt{(d \vee r_{\max} \vee f_{\max})/n} + c_{\max} ||\mathbf{\Pi}||_1 \leq \sqrt{c_{\max}^2 ||\mathbf{\Pi}||_1^2 + \phi_0^2/(64 C_2 |\mathcal{S}_i|)}. \tag{11}$$

Under Assumptions 1-3, we have $\phi_{re}(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{-i}, \mathcal{S}_i) \geq \boldsymbol{\phi}_0/2$ with probability at least $1 - e^{-f^{(1)} + \log p} - e^{-f^{(2)} + \log p}$.

Proof of Lemma 2. The inequality (11) implies that $g_n \leq \phi_0^2/(64|\mathcal{S}_i|)$.

For any index set $\mathcal{S}_i$ and vector δ, note the definition of $\boldsymbol{\phi}_{\text{re}}(\cdot)$, then, we have that $||\delta||_1^2 \leq (||\delta_{\mathcal{S}_i^c}||_1 + ||\delta_{\mathcal{S}_i}||_1)^2 \leq (3\sqrt{|\mathcal{S}_i|}||\delta_{\mathcal{S}_i}||_2 + \sqrt{|\mathcal{S}_i|}||\delta_{\mathcal{S}_i}||_2)^2 = 16|\mathcal{S}_i|||\delta_{\mathcal{S}_i}||_2^2$. we also have

$$\begin{aligned}&\frac{\delta^T((\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{-i})^T(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{-i}) - (\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{-i})^T(\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{-i}))\delta}{n||\delta_{\mathcal{S}_i}||_2^2} \\ \leq\ & \frac{||\delta||_1^2}{n||\delta_{\mathcal{S}_i}||_2^2} \max_{j_1, j_2} |(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{j_1})^T(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{j_2}) - (\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_1})^T(\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_2})| \\ \leq\ & \frac{16|\mathcal{S}_i|}{n} \max_{j_1, j_2} |(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{j_1})^T(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{j_2}) - (\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_1})^T(\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_2})|. \end{aligned} \tag{12}$$

Note that,

$$\begin{aligned}&(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{j_1})^T(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{j_2}) - (\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_1})^T(\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_2}) \\ &= \underbrace{(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})^T \mathbf{X}^T \mathbf{H}_i \mathbf{X} (\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})}_{T_{31}} + \underbrace{(\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})^T \mathbf{X}^T \mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_2}}_{T_{32}} + \underbrace{(\mathbf{X} \mathbf{\Pi}_{j_1})^T \mathbf{H}_i \mathbf{X} (\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})}_{T_{33}}.\end{aligned}$$

We will derive the bounds for each of these three terms separately. With $\mathbf{H}_i$ a projection matrix, we have $\lambda_{max}(\mathbf{H}_i) = 1$. We can obtain that

$$\begin{aligned}|T_{31}| &\leq ||\mathbf{H}_i \mathbf{X} (\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2 \times ||\mathbf{H}_i \mathbf{X} (\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})||_2 \\ &\leq \lambda_{max}(\mathbf{H}_i) ||\mathbf{X} (\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2 \times ||\mathbf{X} (\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})||_2 \\ &= ||\mathbf{X} (\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2 \times ||\mathbf{X} (\hat{\mathbf{\Pi}}_{j_2} - \mathbf{\Pi}_{j_2})||_2.\end{aligned}$$

Note that $|T_{32}| \leq ||\mathbf{X}\, \mathbf{\Pi}_{j_2}||_2 ||\mathbf{H}_i \mathbf{X} (\hat{\mathbf{\Pi}}_{j_1} - \mathbf{\Pi}_{j_1})||_2$, and

$$\begin{aligned}||\mathbf{X}\, \mathbf{\Pi}_{j_2}||_2^2 &= ||X^{(1)} \boldsymbol{\pi}_{j|p}^{(1)}||_2^2 + ||X^{(2)} \boldsymbol{\pi}_{j|p}^{(2)}||_2^2 \\ &\leq (c_1^{(1)})^2 n^{(1)} ||\boldsymbol{\pi}_{j|p}^{(1)}||_2^2 + (c_1^{(2)})^2 n^{(2)} ||\boldsymbol{\pi}_{j|p}^{(2)}||_2^2 \\ &\leq c_{\text{max}}^2 n (||\boldsymbol{\pi}_{j|p}^{(1)}||_2^2 + ||\boldsymbol{\pi}_{j|p}^{(2)}||_2^2) \\ &\leq c_{\text{max}}^2 n \left(||\boldsymbol{\pi}_{j|p}^{(1)}||_2 + ||\boldsymbol{\pi}_{j|p}^{(2)}||_2 \right)^2 \\ &\leq c_{\text{max}}^2 ||\mathbf{\Pi}||_1^2.\end{aligned}$$

Therefore,

$$|T_{32}| \leq ||\mathbf{X}\mathbf{\Pi}_{j_2}||_2||\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1}-\mathbf{\Pi}_{j_1})||_2 \leq c_{\max}\sqrt{n}||\mathbf{\Pi}||_1||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_1}-\mathbf{\Pi}_{j_1})||_2. \tag{13}$$

Similarly, we can have

$$|T_{33}| \leq c_{\max}\sqrt{n}||\mathbf{\Pi}||_1||\mathbf{X}(\hat{\mathbf{\Pi}}_{j_2}-\mathbf{\Pi}_{j_2})||_2. \tag{14}$$

Theorem 2 leads to the following, with probability at least $1-e^{-f^{(1)}+\log(p)}-e^{-f^{(2)}+\log(p)}$,

$$\begin{cases} \dfrac{|T_{31}|}{n} \leq C_2\dfrac{d\vee r_{\max}\vee f_{\max}}{n}, \\ \dfrac{|T_{32}|}{n} \leq c_{\max}C_2||\mathbf{\Pi}||_1\sqrt{\dfrac{d\vee r_{\max}\vee f_{\max}}{n}}, \\ \dfrac{|T_{33}|}{n} \leq c_{\max}C_2||\mathbf{\Pi}||_1\sqrt{\dfrac{d\vee r_{\max}\vee f_{\max}}{n}}. \end{cases} \tag{15}$$

Putting the above three inequalities together, we have,

$$\begin{aligned} &\frac{\delta^T((\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i})^T(\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-i})-(\mathbf{H}_i\mathbf{X}\mathbf{\Pi}_{-i})^T(\mathbf{H}_i\mathbf{X}\mathbf{\Pi}_{-i}))\delta}{n||\delta_{\mathcal{S}_i}||_2^2} \\ \leq\quad & 16|\mathcal{S}_i|\times\frac{|T_{31}|+|T_{32}|+|T_{33}|}{n} = 16|\mathcal{S}_i|g_n \leq 16|\mathcal{S}_i|\frac{\boldsymbol{\phi}_0^2}{64|\mathcal{S}_i|} = \frac{\boldsymbol{\phi}_0^2}{4}. \end{aligned} \tag{16}$$

Together with Assumption 4, we have $\phi_{\text{re}}(\mathbf{H}_i\mathbf{X}\hat{\mathbf{\Pi}}_{-k},\mathcal{S}_k) \geq \boldsymbol{\phi}_0/2$. This concludes the proof of Lemma 2. □

Lemma 3. *(Basic Inequality) Let* $\boldsymbol{\eta}_i = 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - 2n^{-1}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i}-\mathbf{Z}_{-i})\boldsymbol{\beta}_i$ *and*

$$\mathscr{E}(\lambda_i) = \left\{||W_i^{-1}\boldsymbol{\eta}_i||_\infty \leq \lambda_i/2\right\},$$

for λ_i specified in Theorem 3. Under Assumptions 1-2, with h_n defined in Theorem 3, there exit a positive constant $C_3>0$ such that

$$\mathbb{P}(\mathscr{E}(\lambda_i)) \geq 1-e^{-C_3h_n+\log(4q)}-e^{-f^{(1)}+\log(p)}-e^{-f^{(2)}+\log(p)}.$$

Concurring with event $\mathscr{E}(\lambda_i)$, we have the following basic inequality,

$$n^{-1}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2+\lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\beta}}_i|_1 \leq \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1+\boldsymbol{\eta}_i^T(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i). \tag{17}$$

Proof of Lemma 3. Letting

$$\boldsymbol{\xi}_{-i} = \begin{pmatrix} \boldsymbol{\xi}_{-i}^{(1)} & \boldsymbol{\xi}_{-i}^{(1)} \\ \boldsymbol{\xi}_{-i}^{(2)} & -\boldsymbol{\xi}_{-i}^{(2)} \end{pmatrix}, \tag{18}$$

we have $\mathbf{Z}_{-i} = \mathbf{X}\mathbf{\Pi}_{-i}+\boldsymbol{\xi}_{-i}$. With $\hat{\mathbf{Z}}_{-i} = \mathbf{X}\hat{\mathbf{\Pi}}_{-i}$, we get

$$\begin{aligned} \boldsymbol{\eta}_i &= \frac{2}{n}\hat{\mathbf{\Pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i - \frac{2}{n}\hat{\mathbf{\Pi}}_{-i}^T\mathbf{X}^T\mathbf{H}_i(\mathbf{X}\hat{\mathbf{\Pi}}_{-i}-\mathbf{X}\mathbf{\Pi}_{-i}-\boldsymbol{\xi}_{-i})\boldsymbol{\beta}_i \\ &= \underbrace{\frac{2}{n}(\hat{\mathbf{\Pi}}_{-i}-\mathbf{\Pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i}_{T_{34}} + \underbrace{\frac{2}{n}\mathbf{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\epsilon}_i}_{T_{35}} \\ &\quad + \underbrace{\frac{2}{n}(\hat{\mathbf{\Pi}}_{-i}-\mathbf{\Pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\xi}_{-i}\boldsymbol{\beta}_i}_{T_{36}} + \underbrace{\frac{2}{n}\mathbf{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\boldsymbol{\xi}_{-i}\boldsymbol{\beta}_i}_{T_{37}} \\ &\quad - \underbrace{\frac{2}{n}(\hat{\mathbf{\Pi}}_{-i}-\mathbf{\Pi}_{-i})^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{-i}-\mathbf{\Pi}_{-i})\boldsymbol{\beta}_i}_{T_{38}} - \underbrace{\frac{2}{n}\mathbf{\Pi}_{-i}^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\mathbf{\Pi}}_{-i}-\mathbf{\Pi}_{-i})\boldsymbol{\beta}_i}_{T_{39}}. \end{aligned}$$

We aim to bound each of these six terms by $\lambda_i/12$ either probabilistically or deterministically.

Firstly, for some constant $t_\lambda > 0$, we choose the adaptive lasso tuning parameter as below,

$$\lambda_i = t_\lambda \|\boldsymbol{\omega}_i\|_{-\infty}^{-1} ||\mathbf{B}||_1 ||\mathbf{\Pi}||_1 \sqrt{\frac{(d \vee r_{\max} \vee f_{\max}) \log(p)}{n_{\min}}}. \tag{19}$$

Denoting the j-th column of $\mathbf{X}$ by $X_{\cdot j}$, we have $X_{\cdot j}^T X_{\cdot j} = n^{(k)}$ for $k \in \{1, 2\}$ due to standardization. Furthermore,

$$\operatorname{var}\left(\frac{1}{n} X_{\cdot j}^T \mathbf{H}_i \boldsymbol{\epsilon}_i\right) \leq \frac{1}{n^2} X_{\cdot j}^T \mathbf{H}_i X_{\cdot j} \sigma_{p\max}^2 \leq \frac{n^{(k)}}{n^2} \sigma_{p\max}^2 \leq \frac{1}{n} \sigma_{p\max}^2.$$

For T_{34}, via the classical Gaussian tail inequality, we have

$$\begin{aligned}
\mathbb{P}\left(||W_i^{-1} T_{34}||_\infty \geq \frac{\lambda_i}{12}\right) &\leq \mathbb{P}\left(||\frac{2}{n}(\hat{\mathbf{\Pi}}_{-i} - \mathbf{\Pi}_{-i})^T \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right) \\
&\leq \mathbb{P}\left(||(\hat{\mathbf{\Pi}}_{-i} - \mathbf{\Pi}_{-i})^T||_\infty ||\frac{2}{n} \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right) \\
&\leq \mathbb{P}\left(||\frac{2}{n} \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12\delta_\Pi}\right) \leq 2q \exp\left\{-\frac{n\lambda_i^2 \|\boldsymbol{\omega}_i\|_{-\infty}^2}{1152\sigma_{p\max}^2 \delta_\Pi^2}\right\} \\
&\leq 2q \cdot p^{-\frac{n}{d} t_1 ||\mathbf{B}||_1^2 ||\mathbf{\Pi}||_1^2} \leq 2q \cdot p \cdot p^{-t_1 ||\mathbf{B}||_1^2 \frac{n}{d} ||\mathbf{\Pi}||_1^2},
\end{aligned} \tag{20}$$

where $t_1 = t_\lambda^2 / (2304 C_1 \sigma_{p\max}^2)$, and δ_Π is the maximum estimation loss of the first stage. The last inequality is obtained based on the following bound of δ_Π. Following Theorem 2, δ_Π satisfies the following inequality with probability at least $1 - e^{-f^{(1)} + \log(p)} - e^{-f^{(2)} + \log(p)}$,

$$\delta_\Pi^2 = \max_{1 \leq j \leq 2p} ||\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j||_1^2 \leq \max_{1 \leq j \leq 2p} \left(2d||\hat{\mathbf{\Pi}}_j - \mathbf{\Pi}_j||_2^2\right) \leq 2C_1 d \left\{\frac{d \vee r_{\max} \vee f_{\max}}{n_{\min}}\right\}. \tag{21}$$

Note that the first inequality of (21) holds, since $\hat{\mathbf{\Pi}}$ and $\mathbf{\Pi}$ have at most $2d$ non-zeros based on our assumptions and the screening in the calibration step.

Similarly, for the second term T_{35}, we have that, with $t_2 = \frac{(t_\lambda)^2}{1152\sigma_{p\max}^2}$,

$$\begin{aligned}
\mathbb{P}\left(||W_i^{-1} T_{35}||_\infty \geq \frac{\lambda_i}{12}\right) &\leq \mathbb{P}\left(||\frac{2}{n} \mathbf{\Pi}_{-i}^T \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right) \\
&\leq \mathbb{P}\left(||\mathbf{\Pi}_{-i}^T||_\infty ||\frac{2}{n} \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right) \\
&\leq \mathbb{P}\left(||\frac{2}{n} \mathbf{X}^T \mathbf{H}_i \boldsymbol{\epsilon}_i||_\infty \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12||\mathbf{\Pi}_{-i}^T||_\infty}\right) \leq 2q \exp\left\{-\frac{n\lambda_i^2 \|\boldsymbol{\omega}_i\|_{-\infty}^2}{1152\sigma_{p\max}^2 ||\mathbf{\Pi}_{-i}^T||_\infty^2}\right\} \\
&= 2q \cdot p^{-t_2 ||\mathbf{B}||_1^2 (d \vee r_{\max} \vee f_{\max}) n / n_{\min}} \leq 2q \cdot p \cdot p^{-t_2 ||\mathbf{B}||_1^2 (d \vee r_{\max} \vee f_{\max}) n / n_{\min}}.
\end{aligned} \tag{22}$$

For the third term T_{36}, we write

$$\begin{aligned}
\mathbb{P}\left(||W_i^{-1} T_{36}||_\infty \geq \frac{\lambda_i}{12}\right) &\leq \mathbb{P}\left(||(\hat{\mathbf{\Pi}}_{-i} - \mathbf{\Pi}_{-i})^T||_\infty ||\frac{2}{n} \mathbf{X}^T \mathbf{H}_i \boldsymbol{\xi}_{-i} \boldsymbol{\beta}_i||_1 \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right) \\
&\leq \mathbb{P}\left(\delta_\Pi \times \max_{j_1, j_2} |\frac{2}{n} X_{\cdot j_1}^T \mathbf{H}_i \boldsymbol{\xi}_{j_2}| \times ||\boldsymbol{\beta}_i||_1 \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12}\right) \\
&\leq \mathbb{P}\left(\max_{j_1, j_2} |\frac{2}{n} X_{\cdot j_1}^T \mathbf{H}_i \boldsymbol{\xi}_{j_2}| \geq \frac{\lambda_i \|\boldsymbol{\omega}_i\|_{-\infty}}{12\delta_\Pi ||\boldsymbol{\beta}_i||_1}\right) \leq 2q \cdot 2p \exp\left\{-\frac{n\lambda_i^2 \|\boldsymbol{\omega}_i\|_{-\infty}^2}{1152\sigma_{q\max}^2 \delta_\Pi^2 ||\boldsymbol{\beta}_i||_1^2}\right\} \\
&= 4q \cdot p \cdot p^{-t_3 ||\mathbf{\Pi}||_1^2 n / d},
\end{aligned} \tag{23}$$

where $\sigma^2_{q\max} = \max_{j_1,j_2} \mathrm{var}(\frac{1}{n} X^T_{\cdot j_1}\mathbf{H}_i\boldsymbol{\xi}_{j_2})$ and $t_3 = \frac{t_\lambda^2}{2304 C_1 \sigma^2_{q\max}}$. Similarly, with $t_4 = \frac{t_\lambda^2}{1152\sigma^2_{q\max}}$, we write T_{37} term as

$$\begin{aligned}\mathbb{P}\left(||W_i^{-1}T_{37}||_\infty \geq \frac{\lambda_i}{12}\right) &\leq 2q\cdot 2p\cdot \exp\left\{-\frac{n\lambda_i^2\|\boldsymbol{\omega}_i\|^2_{-\infty}}{1152\sigma^2_{q\max}||\boldsymbol{\Pi}^T_{-i}||^2_\infty||\boldsymbol{\beta}_i||_1^2}\right\} \\ &= 4q\cdot p\cdot p^{-t_4(d\vee r_{\max}\vee f_{\max})n/n_{\min}}. \end{aligned} \tag{24}$$

For the deterministic term T_{38}, choosing $t_\lambda \geq 12C_2||\boldsymbol{\Pi}||_1^{-1}\sqrt{(d\vee r_{\max}\vee f_{\max})/(n\log(p))}$, along with *Cauchy-Schwarz Inequality*, we have

$$\begin{aligned}||W_i^{-1}T_{38}||_\infty &\leq \frac{||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|^{-1}_{-\infty}}{n}\max_{j_1,j_2}|(\hat{\boldsymbol{\Pi}}_{j_1}-\boldsymbol{\Pi}_{j_1})^T\mathbf{X}^T\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_2}-\boldsymbol{\Pi}_{j_2})| \\ &\leq \frac{||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|^{-1}_{-\infty}}{n}\max_{j_1,j_2}\left\{||\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_1}-\boldsymbol{\Pi}_{j_1})||_2||\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_2}-\boldsymbol{\Pi}_{j_2})||_2\right\} \\ &\leq \frac{||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|^{-1}_{-\infty}}{n}\max_{j_1,j_2}\left\{\lambda_{\max}(\mathbf{H}_i)||\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_1}-\boldsymbol{\Pi}_{i_1})||_2||\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_2}-\boldsymbol{\Pi}_{j_2})||_2\right\} \\ &\leq \frac{||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|^{-1}_{-\infty}}{n}\max_{j_1,j_2}\left\{||\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_1}-\boldsymbol{\Pi}_{j_1})||_2||\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_2}-\boldsymbol{\Pi}_{j_2})||_2\right\} \\ &\leq ||\boldsymbol{\beta}_i||_1\|\boldsymbol{\omega}_i\|^{-1}_{-\infty}C_2\frac{d\vee r_{\max}\vee f_{\max}}{n} \leq \frac{\lambda_i}{12}\times\left(\frac{12C_2}{t_\lambda||\boldsymbol{\Pi}||_1}\sqrt{\frac{d\vee r_{\max}\vee f_{\max}}{n\log(p)}}\right)\leq\frac{\lambda_i}{12}.\end{aligned}$$

Similarly, we choose $t_\lambda \geq 24\sqrt{C_2 n_{\min}/(n\log(p))}$, and take Theorem 2 to obtain

$$\begin{aligned}||W_i^{-1}T_{39}||_\infty &\leq 2\frac{||\boldsymbol{\beta}_i||_1||\boldsymbol{\Pi}^T_{-i}||_\infty\|\boldsymbol{\omega}_i\|^{-1}_{-\infty}}{n}\max_{j_1,j_2}|X^T_{\cdot j_1}\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_2}-\boldsymbol{\Pi}_{j_2})| \\ &\leq 2\frac{||\boldsymbol{\beta}_i||_1||\boldsymbol{\Pi}^T_{-i}||_\infty\|\boldsymbol{\omega}_i\|^{-1}_{-\infty}}{\sqrt{n}}\max_{j_2}||\mathbf{H}_i\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_2}-\boldsymbol{\Pi}_{j_2})||_2 \\ &\leq 2\frac{||\boldsymbol{\beta}_i||_1||\boldsymbol{\Pi}^T_{-i}||_\infty\|\boldsymbol{\omega}_i\|^{-1}_{-\infty}}{\sqrt{n}}\max_{j_2}||\mathbf{X}(\hat{\boldsymbol{\Pi}}_{j_2}-\boldsymbol{\Pi}_{j_2})||_2 \leq \frac{\lambda_i}{12}\times\left(\frac{24}{t_\lambda}\sqrt{\frac{C_2 n_{\min}}{n\log(p)}}\right)\leq\frac{\lambda_i}{12}.\end{aligned}$$

Note that $n \geq n_{\min}$. Putting together the probabilistic bounds (20), (21), (22), (23) and (24), along with union bound, there exist a constant $C_3 > 0$ such that

$$\mathbb{P}(\mathscr{E}(\lambda_i)) \geq 1 - 3e^{-C_3 h_n + \log(4pq)} - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}.$$

Next we will establish the basic inequality, concurring with the event $\mathscr{E}(\lambda_i)$.

Since the estimator $\hat{\boldsymbol{\beta}}_i$ from the adaptive lasso minimizes the corresponding objective function, we have

$$\frac{1}{n}||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i||_2 + \lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\beta}}_i|_1 \leq \frac{1}{n}||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i||_2 + \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1. \tag{25}$$

Because $\mathbf{H}_i\mathbf{Y}_i = \mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i$, we can rewrite

$$\begin{aligned}&||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i||_2^2 \\ &= ||\mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i||_2^2 \\ &= ||\mathbf{H}_i\boldsymbol{\epsilon}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i - \mathbf{Z}_{-i}\boldsymbol{\beta}_i) + ||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i - \mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i||_2 \\ &= ||\mathbf{H}_i\boldsymbol{\epsilon}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i - \mathbf{Z}_{-i}\boldsymbol{\beta}_i) + ||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)||_2^2 + ||\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i||_2^2 \\ &\quad +2\boldsymbol{\beta}_i^T(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})^T\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i).\end{aligned} \tag{26}$$

Similarly we can rewrite

$$\begin{aligned}||\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i||_2^2 &= ||\mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i||_2^2 \\ &= ||\mathbf{H}_i\boldsymbol{\epsilon}_i||_2^2 + ||\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i||_2^2 - 2\boldsymbol{\epsilon}_i^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i.\end{aligned} \tag{27}$$

Plugging equations (26) and (27) into (25), we then have

$$\begin{aligned}
&\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2+\lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\beta}}_i|_1\\
&\leq \quad \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1+\left(\frac{2}{n}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i\boldsymbol{\epsilon}_i-\frac{2}{n}\hat{\mathbf{Z}}_{-i}^T\mathbf{H}_i(\hat{\mathbf{Z}}_{-i}-\mathbf{Z}_{-i})\boldsymbol{\beta}_i\right)^T(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)\\
&= \quad \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1+\boldsymbol{\eta}_i^T(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i).
\end{aligned}$$

Thus, the basic inequality is established. This concludes the proof of Lemma 3. □

Conditioning on the event $\mathscr{E}(\lambda_i)$, we remove the random term $\boldsymbol{\eta}_i$ from the basic inequality as

$$\begin{aligned}
&\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2\\
&\leq \quad \lambda_i\boldsymbol{\omega}_i^T|\boldsymbol{\beta}_i|_1-\lambda_i\boldsymbol{\omega}_i^T|\hat{\boldsymbol{\beta}}_i|_1+\boldsymbol{\eta}_i^T(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)\\
&\leq \quad \lambda_i\boldsymbol{\omega}_{\mathcal{S}_i}^T|\boldsymbol{\beta}_{\mathcal{S}_i}|_1-\lambda_i\boldsymbol{\omega}_{\mathcal{S}_i}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}|_1-\lambda_i\boldsymbol{\omega}_{\mathcal{S}_i^c}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}|_1+\boldsymbol{\eta}_{\mathcal{S}_i^c}^T(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c})+\boldsymbol{\eta}_{\mathcal{S}_i}^T(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i})\\
&\leq \quad \lambda_i\boldsymbol{\omega}_{\mathcal{S}_i}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}|_1-\lambda_i\boldsymbol{\omega}_{\mathcal{S}_i^c}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}|_1+\frac{\lambda_i}{2}\boldsymbol{\omega}_{\mathcal{S}_i^c}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}|_1+\frac{\lambda_i}{2}\boldsymbol{\omega}_{\mathcal{S}_i}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}|_1\\
&\leq \quad \frac{3}{2}\lambda_i\boldsymbol{\omega}_{\mathcal{S}_i}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}|_1-\frac{1}{2}\lambda_i\boldsymbol{\omega}_{\mathcal{S}_i^c}^T|\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}|_1\\
&\leq \quad \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}||_1-\frac{1}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}||_1.
\end{aligned} \tag{28}$$

The last inequality implies that

$$\begin{aligned}
&\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2\\
&\leq \quad \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}||_1\leq\frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty\sqrt{|\mathcal{S}_i|}||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}||_2\\
&\leq \quad \frac{3}{2}\lambda_i\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty\sqrt{|\mathcal{S}_i|}\frac{2||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2}{\sqrt{n}\boldsymbol{\phi}_0},
\end{aligned} \tag{29}$$

where the last inequality follows Assumption 4 and Lemma 2. The above inequality leads to that,

$$\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2\leq\frac{9(\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty)^2}{\boldsymbol{\phi}_0^2}|\mathcal{S}_i|\lambda_i^2.$$

Plugging in (19), and letting $C_4=3t_\lambda$, we obtain that

$$\frac{1}{n}||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2\leq\frac{C_4^2\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty^2||\mathbf{B}||_1^2||\boldsymbol{\Pi}||_1^2}{\boldsymbol{\phi}_0^2\|\boldsymbol{\omega}_i\|_{-\infty}^2}|\mathcal{S}_i|\frac{(d\vee r_{\max}\vee f_{\max})\log(p)}{n_{\min}}. \tag{30}$$

The fact that $||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2^2$ is always positive in (28) implies that

$$\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}||_1\leq3\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}||_1, \tag{31}$$

which further leads to that

$$||\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i||_1=||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c}||_1+||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}||_1\leq\left(3\frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty}{\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}}+1\right)||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}-\boldsymbol{\beta}_{\mathcal{S}_i}||_1.$$

Noting the inequality (30), we can follow Assumption 4 and Lemma 2 to derive that

$$\begin{aligned}
||\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i||_1 &\leq \left(3\frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty}{\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}}+1\right)\sqrt{|\mathcal{S}_i|}\frac{2||\mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)||_2}{\sqrt{n}\boldsymbol{\phi}_0}\\
&\leq \left(3\frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty}{\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}}+1\right)\sqrt{|\mathcal{S}_i|}\frac{2C_4\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty||\mathbf{B}||_1||\boldsymbol{\Pi}||_1}{\boldsymbol{\phi}_0^2\|\boldsymbol{\omega}_i\|_{-\infty}}\sqrt{|\mathcal{S}_i|}\sqrt{\frac{(d\vee r_{\max}\vee f_{\max})\log(p)}{n_{\min}}}\\
&\leq 8C_4\frac{\|\boldsymbol{\omega}_{\mathcal{S}_i}\|_\infty^2||\mathbf{B}||_1||\boldsymbol{\Pi}||_1}{\boldsymbol{\phi}_0^2\|\boldsymbol{\omega}_i\|_{-\infty}^2}|\mathcal{S}_i|\sqrt{\frac{(d\vee r_{\max}\vee f_{\max})\log(p)}{n_{\min}}},
\end{aligned} \tag{32}$$

where the last inequality comes from the facts that $\|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty} \geq \|\boldsymbol{\omega}_i\|_{-\infty}$ and $\|\boldsymbol{\omega}_i\|_{-\infty} \leq \|\boldsymbol{\omega}_{\mathcal{S}_i}\|_{\infty}$. Since the inequality (28) concurs with the event $\mathscr{E}(\lambda_i)$, the above prediction and estimation bounds hold with probability at least $1 - 3e^{-C_3 h_n + \log(4pq)} - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$. This completes the proof of Theorem 3.

5 Proof of Theorem 4

Lemma 4. *Suppose that, for node i,*

$$\sqrt{(d \vee r_{\max} \vee f_{\max})/n} + c_{\max}||\mathbf{\Pi}||_1 \leq \sqrt{c_{\max}^2 ||\mathbf{\Pi}||_1^2 + \min(\phi_0^2/64, \tau(4-\tau)^{-1}\|\boldsymbol{\omega}_i\|_{-\infty}/\psi_i)/(C_2|\mathcal{S}_i|)}. \quad (33)$$

Under Assumptions 1-5, we have $||W_{\mathcal{S}_i^c}^{-1}(\hat{\mathcal{I}}_{i,21}\hat{\mathcal{I}}_{i,11}^{-1})W_{\mathcal{S}_i}||_\infty \leq 1-\tau/2$ *with the probability at least* $1 - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$.

Proof of Lemma 4. The inequality (33) implies that $\psi_i \|\boldsymbol{\omega}_i\|_{-\infty}^{-1} |\mathcal{S}_i| g_n \leq \frac{\tau}{4-\tau}$.

By the inequalities (15) and (16) in the proof of Lemma 2 and union bound, we have that, with probability at least $1 - e^{-f^{(1)}+\log(p)} - e^{-f^{(2)}+\log(p)}$,

$$\max_{j_1, j_2} \left\{ \frac{1}{n} |(\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{j_1})^T (\mathbf{H}_i \mathbf{X} \hat{\mathbf{\Pi}}_{j_2}) - (\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_1})^T (\mathbf{H}_i \mathbf{X} \mathbf{\Pi}_{j_2})| \right\} \leq g_n.$$

With the definitions of infinity norm $||\cdot||_\infty$, $\hat{\mathcal{I}}_{i,11}$, and $\mathcal{I}_{i,11}$, we can obtain the following inequality indexed by set $\mathcal{S}_i$,

$$\psi_i ||W_{\mathcal{S}_i}^{-1}(\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11})||_\infty \leq \psi_i \|\boldsymbol{\omega}_{\mathcal{S}_i}\|_{-\infty}^{-1} ||\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11}||_\infty \leq \psi_i \|\boldsymbol{\omega}_{\mathcal{S}_i}\|_{-\infty}^{-1} |\mathcal{S}_i| g_n \leq \frac{\tau}{4-\tau}. \quad (34)$$

Similarly we can obtain the following bound indexed by the complement set $\mathcal{S}_i^c$,

$$\psi_i ||W_{\mathcal{S}_i^c}^{-1}(\hat{\mathcal{I}}_{i,21} - \mathcal{I}_{i,21})||_\infty \leq \psi_i \|\boldsymbol{\omega}_{\mathcal{S}_i^c}\|_{-\infty}^{-1} |\mathcal{S}_i| g_n \leq \frac{\tau}{4-\tau}. \quad (35)$$

Applying the matrix inversion error bound in Horn and Johnson [2012] and the triangular inequality, we have that

$$\begin{aligned} ||\hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i}||_\infty &\leq ||\mathcal{I}_{i,11}^{-1} W_{\mathcal{S}_i}||_\infty + ||\hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i} - \mathcal{I}_{i,11}^{-1} W_{\mathcal{S}_i}||_\infty \\ &\leq \psi_i + \frac{\psi_i ||W_{\mathcal{S}_i}^{-1}(\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11})||_\infty}{1 - \psi_i ||W_{\mathcal{S}_i}^{-1}(\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11})||_\infty} \psi_i \leq \psi_i + \frac{\tau}{4-2\tau}\psi_i \leq \frac{4-\tau}{4-2\tau}\psi_i. \end{aligned} \quad (36)$$

Also note that we can rewrite

$$\begin{aligned} & W_{\mathcal{S}_i^c}^{-1} \left(\hat{\mathcal{I}}_{i,21} \hat{\mathcal{I}}_{i,11}^{-1} - \mathcal{I}_{i,21} \mathcal{I}_{i,11}^{-1} \right) W_{\mathcal{S}_i} \\ = \;& W_{\mathcal{S}_i^c}^{-1} \left(\hat{\mathcal{I}}_{i,21} - \mathcal{I}_{i,21} \right) \hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i} + W_{\mathcal{S}_i^c}^{-1} \mathcal{I}_{i,21} \mathcal{I}_{i,11}^{-1} W_{\mathcal{S}_i} W_{\mathcal{S}_i}^{-1} \left(\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11} \right) \hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i}. \end{aligned}$$

Then, it follows from (34), (35), (36) and Assumption 5 that

$$\begin{aligned} & ||W_{\mathcal{S}_i^c}^{-1} \left(\hat{\mathcal{I}}_{i,21} \hat{\mathcal{I}}_{i,11}^{-1} - \mathcal{I}_{i,21} \mathcal{I}_{i,11}^{-1} \right) W_{\mathcal{S}_i}||_\infty \\ \leq \;& ||W_{\mathcal{S}_i^c}^{-1} \left(\hat{\mathcal{I}}_{i,21} - \mathcal{I}_{i,21} \right) ||_\infty ||\hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i}||_\infty \\ & + ||W_{\mathcal{S}_i^c}^{-1} \mathcal{I}_{i,21} \mathcal{I}_{i,11}^{-1} W_{\mathcal{S}_i}||_\infty ||W_{\mathcal{S}_i}^{-1} \left(\hat{\mathcal{I}}_{i,11} - \mathcal{I}_{i,11} \right) ||_\infty ||\hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i}||_\infty \leq \tau/2. \end{aligned}$$

Therefore, together with Assumption 5 again, we can conclude that $||W_{\mathcal{S}_i^c}^{-1}(\hat{\mathcal{I}}_{i,21}\hat{\mathcal{I}}_{i,11}^{-1})W_{\mathcal{S}_i}||_\infty \leq 1-\tau/2$.

This concludes the proof of Lemma 4. $\square$

The optimality of $\hat{\boldsymbol{\beta}}_i$ in the adaptive lasso step and KKT condition lead to

$$-\frac{2}{n}(\mathbf{H}_i\hat{\mathbf{Z}}_{-i})^T(\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i) + \lambda_i W_i \alpha_i = 0, \tag{37}$$

where $\alpha_i \in \mathbb{R}^{2p-2}$, satisfying that $||\alpha_i||_\infty \leq 1$ and $\alpha_{ij} I(\hat{\boldsymbol{\beta}}_{ij} \neq 0) = sign(\hat{\boldsymbol{\beta}}_{ij})$.

Plug in the equation $\mathbf{H}_i\mathbf{Y}_i = \mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i$, we can have that

$$\begin{aligned}
\mathbf{H}_i\mathbf{Y}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i &= \mathbf{H}\mathbf{Z}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\boldsymbol{\epsilon}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i \\
&= \mathbf{H}_i\boldsymbol{\epsilon}_i + \mathbf{H}_i\mathbf{Z}_{-i}\boldsymbol{\beta}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i + \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\boldsymbol{\beta}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}\hat{\boldsymbol{\beta}}_i \\
&= \mathbf{H}_i\boldsymbol{\epsilon}_i - \mathbf{H}_i(\hat{\mathbf{Z}}_{-i} - \mathbf{Z}_{-i})\boldsymbol{\beta}_i - \mathbf{H}_i\hat{\mathbf{Z}}_{-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i).
\end{aligned} \tag{38}$$

This, along with KKT condition (37), leads to

$$2\hat{\mathcal{I}}_i(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) - \boldsymbol{\eta}_i = -\lambda_i W_i \alpha_i, \tag{39}$$

where $\boldsymbol{\eta}_i$ is defined in Lemma 3.

Letting $\hat{\boldsymbol{\beta}}_{\mathcal{S}_i^c} = \boldsymbol{\beta}_{\mathcal{S}_i^c} = 0$, equation (39) can be decomposed as

$$\begin{cases} 2\hat{\mathcal{I}}_{i,11}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}) - \boldsymbol{\eta}_{\mathcal{S}_i} = -\lambda_i W_{\mathcal{S}_i}\alpha_{\mathcal{S}_i}, \\ 2\hat{\mathcal{I}}_{i,21}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}) - \boldsymbol{\eta}_{\mathcal{S}_i^c} = -\lambda_i W_{\mathcal{S}_i^c}\alpha_{\mathcal{S}_i^c}. \end{cases} \tag{40}$$

We can solve for $\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}$ from the first equation of (40) as

$$\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i} = 2^{-1}\hat{\mathcal{I}}_{i,11}^{-1}(\boldsymbol{\eta}_{\mathcal{S}_i} - \lambda_i W_{\mathcal{S}_i}^T \alpha_{\mathcal{S}_i}) = 2^{-1}\hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i}(W_{\mathcal{S}_i}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i} - \lambda_i \alpha_{\mathcal{S}_i}). \tag{41}$$

Following the similar strategy in the proof of Lemma 3, we can prove that there exists a constant $C_5 > 0$ such that $||W_i^{-1}\boldsymbol{\eta}_i||_\infty \leq \frac{\tau}{4-\tau}\lambda_i$ with probability at least $1 - 3e^{-C_5\, h_n + \log(4q) + \log(p)} - e^{-f^{(1)} + \log(p)} - e^{-f^{(2)} + \log(p)}$. Thus, together with $||\alpha_{\mathcal{S}_i}||_\infty \leq 1$, we obtain the infinity norm estimation loss on the true support set $\mathcal{S}_i$

$$\begin{aligned}
||\hat{\boldsymbol{\beta}}_{\mathcal{S}_i} - \boldsymbol{\beta}_{\mathcal{S}_i}||_\infty &\leq 2^{-1}||\hat{\mathcal{I}}_{i,11}^{-1} W_{\mathcal{S}_i}||_\infty(||W_{\mathcal{S}_i}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i}||_\infty + \lambda_i) \\
&\leq 2^{-1}\frac{4-\tau}{4-2\tau}\psi_i\frac{4}{4-\tau}\lambda_i = \frac{\lambda_i\psi_i}{2-\tau} \leq \min_{j\in\mathcal{S}_i}|\boldsymbol{\beta}_{ij}| = b_i,
\end{aligned}$$

where the last inequality comes from the condition on the minimal signal strength b_i. The above inequality implies $sign(\hat{\boldsymbol{\beta}}_{\mathcal{S}_i}) = sign(\boldsymbol{\beta}_{\mathcal{S}_i})$.

Plugging (40) into the left hand side of the second equation in (41), we can verify that

$$\begin{aligned}
&||W_{\mathcal{S}_i^c}^{-1}\hat{\mathcal{I}}_{i,21}(\hat{\mathcal{I}}_{i,11})^{-1}(\boldsymbol{\eta}_{\mathcal{S}_i} - \lambda_i W_{\mathcal{S}_i}\alpha_{\mathcal{S}_i}) - W_{\mathcal{S}_i^c}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i^c}||_\infty \\
\leq\ & ||W_{\mathcal{S}_i^c}^{-1}\hat{\mathcal{I}}_{i,21}\hat{\mathcal{I}}_{i,11}^{-1}W_{\mathcal{S}_i}||_\infty(||W_{\mathcal{S}_i}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i}||_\infty + \lambda_i) + ||W_{\mathcal{S}_i^c}^{-1}\boldsymbol{\eta}_{\mathcal{S}_i^c}||_\infty \\
\leq\ & (1-\tau/2)(4/(4-\tau))\lambda_i + \tau/(4-\tau)\lambda_i = \lambda_i.
\end{aligned}$$

Therefore, we have constructed a solution $\hat{\boldsymbol{\beta}}_i$ which satisfies the KKT condition (39) and $sign(\hat{\boldsymbol{\beta}}_i) = sign(\boldsymbol{\beta}_i)$, that is, $\hat{\mathcal{S}}_i = \mathcal{S}_i$. This completes the proof of Theorem 4.

References

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.