An Ising Similarity Regression Model for Modeling Multivariate Binary Data

Zhi Yang Tho, Francis K. C. Hui, and Tao Zou

The Australian National University

Supplementary Material

This supplementary material consists of nine parts.

1. Section S1 presents important notations used in the manuscript.

2. Section S2 provides the sample versions of Conditions 1 and 3.

3. Section S3 introduces seven propositions that are useful for the proofs of the theorems.

4. Section S4 presents the proofs of Theorems 1 – 2 based on the propositions.

5. Section S5 provides the proofs of the propositions, together with eight lemmas and their corresponding proofs.

6. Section S6 presents the inference method based on the proposed regularized pseudo-likelihood estimator.

7. Section S7 contains additional details and results of the simulation study in Section 4, along with comparison to other estimators.

8. Section S8 presents supplementary details of the real data application in Section 5.

9. Section S9 provides an application of the Ising similarity regression model to the Scotland Carabidae ground beetle dataset.

Throughout the supplementary material, for a generic m-dimensional vector $\boldsymbol{x} = (x_1, \cdots, x_m)^\top$ and subset of indices $\mathcal{T} \subset \{1, \cdots, m\}$, we let $\boldsymbol{x}_\mathcal{T}$ denote the subvector of $\boldsymbol{x}$ consisting of the elements indexed by $\mathcal{T}$. Furthermore, let $\boldsymbol{\mathcal{X}}^{(i)}_{\mathcal{V}}$ denote the submatrix of $\boldsymbol{\mathcal{X}}^{(i)}$ consisting of columns indexed by $\mathcal{V} \subset \{1, \cdots, K\}$ for $i = 1, \cdots, n$, where $\boldsymbol{\mathcal{X}}^{(i)}$ is defined in equation (3.2). Finally, recall $\bar{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}}\{-l(\boldsymbol{\alpha})\}$ is the unregularized pseudo-likelihood estimator, and $\boldsymbol{\alpha}^{(0)}$ is the true regression coefficient vector. With these in mind, we also define $\boldsymbol{\delta} = \boldsymbol{\alpha} - \boldsymbol{\alpha}^{(0)}$, and $\bar{\boldsymbol{\delta}} = \bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}$.

S1 Important Notations

Table S1 provides definition of key notations used throughout the manuscript.

Table S1: Key notations used in the manuscript along with their definitions.

Notation	Definition
n	Number of multivariate binary response vectors
p	Dimension of multivariate binary response vectors
K	Number of similarity matrices
$\boldsymbol{\Theta}, \theta_{jj'}$	Ising model interaction matrix, and its (j, j')-th element
$\boldsymbol{y}_i, y_{ij}$	i-th multivariate binary response vector, and its j-th element
$\boldsymbol{W}_k, w_{jj'}^{(k)}$	k-th similarity matrix, and its (j, j')-th element
$\boldsymbol{\alpha}, \alpha_k$	Vector of regression coefficients for the similarity matrices, and its k-th element associated with the k-th similarity matrix
$\boldsymbol{\vartheta}$	Parameter vector of the Ising similarity regression model consisting of main effect parameters and regression coefficients
$f(\cdot; \boldsymbol{\vartheta})$	Pmf of the Ising similarity regression model
$Z(\boldsymbol{\vartheta})$	Intractable normalization constant in the pmf of the Ising similarity regression model
$l(\boldsymbol{\alpha}), l_i(\boldsymbol{\alpha})$	Normalized log pseudo-likelihood of n response vectors, log pseudo-likelihood of i-th response vector
$\hat{\boldsymbol{\alpha}}$	Regularized pseudo-likelihood estimator of regression coefficients
$\bar{\boldsymbol{\alpha}}$	Unregularized pseudo-likelihood estimator of regression coefficients
w_k, λ	Adaptive weights and tuning parameter for adaptive lasso penalty
$\boldsymbol{\mathcal{X}}^{(i,j)}$	K-dimensional vector with the k-th element being $\boldsymbol{W}_{j\cdot}^{(k)\top} \boldsymbol{y}_i$
$\boldsymbol{\mathcal{X}}^{(i)}$	$p \times K$ matrix with the j-th row being $\boldsymbol{\mathcal{X}}^{(i,j)}$
$\boldsymbol{\mathcal{X}}$	$np \times K$ matrix with the $((j-1)n+i)$-th row being $\boldsymbol{\mathcal{X}}^{(i,j)}$
$\boldsymbol{\alpha}^{(0)}$	True value of the regression coefficient vector
S	Index set of non-zero regression coefficients in $\boldsymbol{\alpha}^{(0)}$
S^c	Index set of zero regression coefficients in $\boldsymbol{\alpha}^{(0)}$
$\boldsymbol{U}^0$	Expected value of the matrix $\sum_{i=1}^n \boldsymbol{\mathcal{X}}^{(i)\top} \boldsymbol{\mathcal{X}}^{(i)}/(np)$
$\boldsymbol{M}^0$	Expected value of the matrix $-\nabla^2 l(\boldsymbol{\alpha}^{(0)})$

S2 Sample Versions of Conditions 1 and 3

To establish the proofs of the theorems in Section 3, we introduce sample counterparts of Conditions 1 and 3 on $\boldsymbol{M}^n = -\nabla^2 l(\boldsymbol{\alpha}^{(0)})$ and $\boldsymbol{U}^n = \sum_{i=1}^n \boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\mathcal{X}}^{(i)}/(np)$, which are parallel to $\boldsymbol{M}^0$ and $\boldsymbol{U}^0$ defined below equation (3.2), respectively.

Condition 1′. *There exist finite positive constants* $C_{\min}$ *and* $C_{\max}$ *such that* $\Lambda_{\min}(\boldsymbol{M}^n) \geq C_{\min}$ *and* $\Lambda_{\max}(\boldsymbol{U}^n) \leq C_{\max}$.

Condition 3′. *There exists a constant* $C'_M \in (0,1)$ *such that* $\left\|\boldsymbol{M}^n_{S^c,S}(\boldsymbol{M}^n_{S,S})^{-1}\right\|_\infty \leq 1 - C'_M$.

Sample Condition 1′ implies the submatrices $\boldsymbol{M}^n_{S,S}$ and $\boldsymbol{U}^n_{S,S}$ satisfy $\Lambda_{\min}(\boldsymbol{M}^n_{S,S}) \geq C_{\min}$ and $\Lambda_{\max}(\boldsymbol{U}^n_{S,S}) \leq C_{\max}$, respectively. Both conditions are used to develop the propositions in Section S3.

S3 Seven Propositions

Before presenting the propositions, we first introduce a restricted version of criterion (3.1), where S is assumed to be known in this criterion and all the regression coefficients indexed by S^c are set to zero. That is,

$$\min_{\boldsymbol{\alpha}^{[S]}} \left\{ -l(\boldsymbol{\alpha}^{[S]}) + \lambda \sum_{k\in S} w_k \left|\alpha_k^{[S]}\right| \right\}, \tag{S3.1}$$

where $\boldsymbol{\alpha}^{[S]} = (\alpha_1^{[S]}, \cdots, \alpha_K^{[S]})^\top \in \mathbb{R}^K$ with $\alpha_k^{[S]} = 0$ for $k \in S^c$, and $w_k = 1/|\bar{\alpha}_k|$ are the adaptive weights with $\bar{\alpha}_k$ being the k-th element in the vector of the unregularized pseudo-likelihood estimator $\bar{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}}\{-l(\boldsymbol{\alpha})\}$. The estimator that minimizes criterion (S3.1) is denoted as $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \cdots, \tilde{\alpha}_K)^\top$, and we further define $\tilde{\boldsymbol{\delta}} = \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}$. The asymptotic properties of $\tilde{\boldsymbol{\alpha}}$ are discussed in Propositions 3, 4 and 7 below, and we subsequently make use of these results for $\tilde{\boldsymbol{\alpha}}$ to facilitate the proof of Theorem 2 in Section S4. Note the proof can be generalized to incorporate other choices of adaptive weights, although we focus on $w_k = 1/|\bar{\alpha}_k|$ in this article.

Proposition 1. *Assume sample Condition 1′ and Condition 2 are satisfied. If* $K\sqrt{\log(p)/n} = o(1)$ *and there exists a finite positive constant* C_∇ *such that* $K = o(p^{C_\nabla^2/(8C_W^2)})$ *as* $n, p \to \infty$*, then with probability tending to one it holds that*

$$\left\|\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\right\|_2 \leq \bar{M}\sqrt{\frac{K\log(p)}{n}},$$

for a finite positive constant $\bar{M} > 4C_\nabla/C_{\min}$.

Proposition 2. *Assume Conditions 1 and 2 are satisfied. Then for any* $\epsilon > 0$*, it holds that*

(i) $\mathrm{P}\{\Lambda_{\min}(\boldsymbol{M}^n) \leq C_{\min} - \epsilon\} \leq 2\exp\{-\epsilon^2 n/(8C_W^4K^2) + 2\log(K)\}$;

(ii) $\mathrm{P}\{\Lambda_{\max}(\boldsymbol{U}^n) \geq C_{\max} + \epsilon\} \leq 2\exp\{-\epsilon^2 n/(8C_W^4K^2) + 2\log(K)\}$.

Proposition 3. *Assume sample Condition 1′, Conditions 2 and 4 are satisfied. If* $K\sqrt{\log(p)/n} = o(1)$ *and there exists a finite positive constant* C_∇ *such that* $K = o(p^{C_\nabla^2/(8C_W^2)})$ *as* $n, p \to \infty$*, then with probability tending to one it holds that*

$$\left\|\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\right\|_2 \leq M\sqrt{\frac{K_0 \log(p)}{n}},$$

for a finite positive constant $M > 4/C_{\min}$.

Proposition 4. *Assume sample Condition 1′, Conditions 2 and 4 are satisfied. If* $K\sqrt{\log(p)/n} = o(1)$ *and there exists a finite positive constant* C_∇ *such that* $K = o(p^{C_\nabla^2/(8C_W^2)})$ *as* $n, p \to \infty$*, then with probability tending to one it holds that* $\tilde{\alpha}_k \neq 0$ *for all* $k \in S$ *and* $\tilde{\alpha}_k = 0$ *for all* $k \in S^c$.

Proposition 5. *Assume Conditions 1 – 3 are satisfied. Then for any* $\epsilon > 0$*, there exists a finite positive constant* $C = \min\left\{C_{\min}^2 C_M^2/\{1152(1-C_M)^2 C_W^4\}, C_{\min}^2 C_M^2/(288C_W^4), C_M/(48C_W^4), C_{\min}^4 C_M/(192C_W^4)\right\}$*, such that*

$$\mathrm{P}\left\{\left\|\boldsymbol{M}_{S^c,S}^n(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_\infty \geq 1 - \frac{C_M}{2}\right\} \leq 12\exp\left\{-C\frac{n}{K_0^3} + 2\log(K)\right\}.$$

Proposition 6. *The sufficient and necessary conditions for* $\hat{\boldsymbol{\alpha}}$ *to be a minimizer*

of criterion (3.1) *are*

$$\frac{\partial l(\hat{\boldsymbol{\alpha}})}{\partial \alpha_k} = \frac{\lambda}{|\bar{\alpha}_k|}\text{sign}(\hat{\alpha}_k), \text{ if } \hat{\alpha}_k \neq 0, \quad \text{and} \quad \left|\frac{\partial l(\hat{\boldsymbol{\alpha}})}{\partial \alpha_k}\right| < \frac{\lambda}{|\bar{\alpha}_k|}, \text{ if } \hat{\alpha}_k = 0.$$

Moreover, this minimizer is unique due to the strict convexity of criterion (3.1).

Proposition 7. *Assume sample Conditions 1′ and 3′, and Conditions 2 and 4 are satisfied. If* $K\sqrt{\log(p)/n} = o(1)$ *and there exists a finite positive constant* C_∇ *such that* $K = o(p^{C_\nabla^2/(8C_W^2)})$ *as* $n, p \to \infty$*, then with probability tending to one it holds that*

(i) $\frac{\partial l(\tilde{\boldsymbol{\alpha}})}{\partial \alpha_k} = \frac{\lambda}{|\bar{\alpha}_k|}\text{sign}(\tilde{\alpha}_k)$*, for all* $k \in S$*;*

(ii) $\left|\frac{\partial l(\tilde{\boldsymbol{\alpha}})}{\partial \alpha_k}\right| < \frac{\lambda}{|\bar{\alpha}_k|}$*, for all* $k \in S^c$*.*

S4 Proofs of Theorems 1 – 2

Proof of Theorem 1. We first show the conditions $K\sqrt{\log(p)/n} = o(1)$ and $K = o(p^{C_\nabla^2/(8C_W^2)})$, as $n, p \to \infty$, imply $-\epsilon^2 n/(8C_W^4K^2) + 2\log(K) \to -\infty$ for any $\epsilon > 0$. The condition $K\sqrt{\log(p)/n} = o(1)$ implies $K^2\log(p)/n \to 0$ and subsequently $n/K^2 \to \infty$, and the condition $K = o(p^{C_\nabla^2/(8C_W^2)})$ implies

$$\frac{\log(K)}{\log(p)} = o\left(\frac{C_\nabla^2}{8C_W^2}\right) = o(1),$$

which leads to

$$\frac{\log(K)K^2}{n} = \frac{\log(p)K^2}{n}\frac{\log(K)}{\log(p)} \to 0.$$

Then for any $\epsilon > 0$, we have

$$-\frac{\epsilon^2 n}{8C_W^4 K^2} + 2\log(K) = -\frac{n}{K^2}\left\{\frac{\epsilon^2}{8C_W^4} - \frac{2\log(K)K^2}{n}\right\} \to -\infty.$$

Therefore, by Proposition 2, sample Condition 1′ holds with probability tending to one. This implies Proposition 1 holds and thus, with probability tending to one, the unregularized pseudo-likelihood estimator $\bar{\boldsymbol{\alpha}}$ satisfies

$$\left\|\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\right\|_2 \leq \bar{M}\sqrt{\frac{K\log(p)}{n}},$$

for a finite positive constant $\bar{M} > 4C_\nabla/C_{\min}$, which completes the proof. □

Proof of Theorem 2. Similar to the proof of Theorem 1, we obtain $-\epsilon^2 n/(8C_W^4K^2) + 2\log(K) \to -\infty$ for any $\epsilon > 0$ under the conditions $K\sqrt{\log(p)/n} = o(1)$ and $K = o(p^{C_\nabla^2/(8C_W^2)})$, as $n, p \to \infty$. Consequently, sample Condition 1′ holds with probability tending to one by Proposition 2. This also implies Propositions 3 – 4 hold and therefore, with probability tending to one, the minimizer $\tilde{\boldsymbol{\alpha}}$ of criterion (S3.1) satisfies

$$\left\|\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\right\|_2 \leq M\sqrt{\frac{K_0\log(p)}{n}},$$

for a finite positive constant $M > 4/C_{\min}$, and

$$\tilde{\alpha}_k \neq 0 \text{ for all } k \in S, \text{ and } \tilde{\alpha}_k = 0 \text{ for all } k \in S^c.$$

Additionally, the condition $K\sqrt{\log(p)/n} = o(1)$ implies $\log(K)/n = o(1)$, and subsequently $-Cn/K_0^3 + 2\log(K) = n\{-C/K_0^3 + 2\log(K)/n\} \to -\infty$ where C is defined in Proposition 5. Therefore, sample Condition 3′ holds with high probability by Proposition 5, and hence Proposition 7 holds under the conditions of Theorem 2. This implies with probability tending to one that $\tilde{\boldsymbol{\alpha}}$ satisfies the necessary and sufficient conditions provided in Proposition 6 to be the minimizer of criterion (3.1). Since criterion (3.1) is strictly convex, we can conclude $\tilde{\boldsymbol{\alpha}}$ is a unique minimizer i.e., $\hat{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}$. This completes the proof. □

S5 Lemmas and Proofs of Propositions

Recall

$$l_i(\boldsymbol{\alpha}) = \sum_{j=1}^{p} \left[y_{ij} \left(\sum_{k=1}^{K} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) - \log \left\{ 1 + \exp \left(\sum_{k=1}^{K} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) \right\} \right].$$

We first introduce additional notations that will be used in the following lemmas and proofs. The gradient vector of $l_i(\boldsymbol{\alpha})$ is denoted as $\nabla l_i(\boldsymbol{\alpha}) = (\partial l_i(\boldsymbol{\alpha})/\partial\alpha_1, \cdots,$

$\partial l_i(\boldsymbol{\alpha})/\partial \alpha_K)^\top$, where

$$\frac{\partial l_i(\boldsymbol{\alpha})}{\partial \alpha_k} = \sum_{j=1}^{p} \left(\sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) \left\{ y_{ij} - \frac{\exp\left(\sum_{l=1}^{K} \alpha_l \sum_{j' \neq j} w_{jj'}^{(l)} y_{ij'} \right)}{1 + \exp\left(\sum_{l=1}^{K} \alpha_l \sum_{j' \neq j} w_{jj'}^{(l)} y_{ij'} \right)} \right\},$$

for $i = 1, \cdots, n$ and $k = 1, \cdots, K$. The Hessian matrix of $l_i(\boldsymbol{\alpha})$ is given by $\nabla^2 l_i(\boldsymbol{\alpha}) = -\boldsymbol{\mathcal{X}}^{(i)\top} \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}) \boldsymbol{\mathcal{X}}^{(i)}$, where $\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}) = \text{diag}\{\eta_{11}^{(i)}(\boldsymbol{\alpha}), \cdots, \eta_{pp}^{(i)}(\boldsymbol{\alpha})\}$ is a $p \times p$ diagonal matrix with diagonal elements

$$\eta_{jj}^{(i)}(\boldsymbol{\alpha}) = \frac{\exp\left(\sum_{k=1}^{K} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right)}{\left\{ 1 + \exp\left(\sum_{k=1}^{K} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) \right\}^2},$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, p$. The gradient vector of $\eta_{jj}^{(i)}(\boldsymbol{\alpha})$ can then be obtained as $\nabla \eta_{jj}^{(i)}(\boldsymbol{\alpha}) = \varepsilon_j^{(i)}(\boldsymbol{\alpha}) \boldsymbol{\mathcal{X}}^{(i,j)}$, where $\boldsymbol{\mathcal{X}}^{(i,j)}$ is defined in equation (3.2) and

$$\varepsilon_j^{(i)}(\boldsymbol{\alpha}) = \frac{\exp\left(\sum_{k=1}^{K} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) \left\{ 1 - \exp\left(\sum_{k=1}^{K} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) \right\}}{\left\{ 1 + \exp\left(\sum_{k=1}^{K} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) \right\}^3},$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, p$. It can be verified that $|\eta_{jj}^{(i)}(\boldsymbol{\alpha})| \leq 1$ and $|\varepsilon_j^{(i)}(\boldsymbol{\alpha})| \leq 1$ for all $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Recalling $l(\boldsymbol{\alpha}) = \sum_{i=1}^{n} l_i(\boldsymbol{\alpha})/(np)$, we thus obtain $\nabla l(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \nabla l_i(\boldsymbol{\alpha})/(np)$ and $\nabla^2 l(\boldsymbol{\alpha}) =$

$\sum_{i=1}^{n} \nabla^2 l_i(\boldsymbol{\alpha})/(np)$. Therefore, we can write

$$\begin{aligned}
\boldsymbol{M}^0 &= E\left\{-\nabla^2 l(\boldsymbol{\alpha}^{(0)})\right\} \\
&= E\left\{-\frac{1}{np}\sum_{i=1}^{n} \nabla^2 l_i(\boldsymbol{\alpha}^{(0)})\right\} \\
&= E\left\{\frac{1}{np}\sum_{i=1}^{n} \boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)})\boldsymbol{\mathcal{X}}^{(i)}\right\},
\end{aligned}$$

and $\boldsymbol{M}^n = -\nabla^2 l(\boldsymbol{\alpha}^{(0)}) = \sum_{i=1}^{n} \boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)})\boldsymbol{\mathcal{X}}^{(i)}/(np)$.

Lemma 1. *Assume Condition 2 is satisfied. Then with probability tending to one, it holds that*

(i) $\left\|\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S\right\|_\infty \leq \sqrt{\log(p)/n}$;

(ii) $\left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty \leq C_\nabla\sqrt{\log(p)/n}$ *for a finite positive constant* C_∇ *that satisfies* $K = o(p^{C_\nabla^2/(8C_W^2)})$ *as* $p \to \infty$.

Proof of Lemma 1. Under Condition 2, we have

$$\begin{aligned}
&\left|\frac{\partial l_i(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right| \\
&= \left|\sum_{j=1}^{p}\left(\sum_{j'\neq j} w_{jj'}^{(k)} y_{ij'}\right)\left\{y_{ij} - \frac{\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j} w_{jj'}^{(l)} y_{ij'}\right)}{1+\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j} w_{jj'}^{(l)} y_{ij'}\right)}\right\}\right| \\
&\leq \sum_{j=1}^{p}\left(\sum_{j'\neq j}\left|w_{jj'}^{(k)}\right||y_{ij'}|\right)\left\{|y_{ij}| + \left|\frac{\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j} w_{jj'}^{(l)} y_{ij'}\right)}{1+\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j} w_{jj'}^{(l)} y_{ij'}\right)}\right|\right\}
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\sum_{j=1}^{p}\left(\sum_{j'\neq j}\left|w_{jj'}^{(k)}\right|\right)\\
&\leq 2\sum_{j=1}^{p}\|\boldsymbol{W}_k\|_{\infty}\\
&= 2\sum_{j=1}^{p}\|\boldsymbol{W}_k\|_{1}\\
&\leq 2\sum_{j=1}^{p}C_W = 2C_W p,
\end{aligned}
$$

for $i = 1, \cdots, n$ and $k = 1, \cdots, K$. By denoting $E(y_{ij}|\boldsymbol{y}_{i\setminus j})$ as p_{ij}, it can also be verified that

$$
p_{ij} = \frac{\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j}w_{jj'}^{(l)}y_{ij'}\right)}{1+\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j}w_{jj'}^{(l)}y_{ij'}\right)},
$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, p$. This, together with law of iterated expectation, implies

$$
\begin{aligned}
&E\left\{\frac{\partial l_i(\boldsymbol{\alpha}^{(0)})}{\partial\alpha_k}\right\}\\
&= \sum_{j=1}^{p}E\left[\left(\sum_{j'\neq j}w_{jj'}^{(k)}y_{ij'}\right)\left\{y_{ij}-\frac{\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j}w_{jj'}^{(l)}y_{ij'}\right)}{1+\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j}w_{jj'}^{(l)}y_{ij'}\right)}\right\}\right]\\
&= \sum_{j=1}^{p}E\left(E\left[\left(\sum_{j'\neq j}w_{jj'}^{(k)}y_{ij'}\right)\left\{y_{ij}-\frac{\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j}w_{jj'}^{(l)}y_{ij'}\right)}{1+\exp\left(\sum_{l=1}^{K}\alpha_l^{(0)}\sum_{j'\neq j}w_{jj'}^{(l)}y_{ij'}\right)}\right\}\middle|\boldsymbol{y}_{i\setminus j}\right]\right)\\
&= \sum_{j=1}^{p}E\left\{\left(\sum_{j'\neq j}w_{jj'}^{(k)}y_{ij'}\right)(p_{ij}-p_{ij})\right\}
\end{aligned}
$$

$= 0,$

for $i = 1, \cdots, n$ and $k = 1, \cdots, K$. Hence, by applying Azuma-Hoeffding inequality (Wainwright, 2019), we obtain

$$\begin{aligned}
\mathrm{P}\left\{\left|\frac{\partial l(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right| \geq t\right\} &= \mathrm{P}\left\{\left|\frac{1}{np}\sum_{i=1}^{n}\frac{\partial l_i(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right| \geq t\right\} \\
&= \mathrm{P}\left\{\left|\sum_{i=1}^{n}\frac{\partial l_i(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right| \geq npt\right\} \\
&= \mathrm{P}\left(\left|\sum_{i=1}^{n}\left[\frac{\partial l_i(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k} - E\left\{\frac{\partial l_i(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right\}\right]\right| \geq npt\right) \\
&\leq 2\exp\left\{\frac{-2n^2p^2t^2}{\sum_{i=1}^{n}(2C_W p + 2C_W p)^2}\right\} \\
&= 2\exp\left(\frac{-nt^2}{8C_W^2}\right),
\end{aligned}$$

for $k = 1, \cdots, K$. This, together with the union sum inequality, implies

$$\begin{aligned}
\mathrm{P}\left[\left\|\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S\right\|_\infty \geq \sqrt{\frac{\log(p)}{n}}\right] &= \mathrm{P}\left[\bigcup_{k\in S}\left\{\left|\frac{\partial l(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right| \geq \sqrt{\frac{\log(p)}{n}}\right\}\right] \\
&\leq \sum_{k\in S}\mathrm{P}\left\{\left|\frac{\partial l(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right| \geq \sqrt{\frac{\log(p)}{n}}\right\} \\
&\leq \sum_{k\in S} 2\exp\left\{\frac{-n\frac{\log(p)}{n}}{8C_W^2}\right\} \\
&= 2K_0\exp\left\{\frac{-\log(p)}{8C_W^2}\right\} \\
&= 2\exp\left\{\frac{-\log(p)}{8C_W^2} + \log(K_0)\right\}
\end{aligned}$$

$$\to 0,$$

as $p \to \infty$, where K_0 is finite. We thus obtain

$$\mathrm{P}\left[\left\|\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S\right\|_\infty \leq \sqrt{\frac{\log(p)}{n}}\right] \to 1,$$

which proves claim (i).

We prove claim (ii) in a similar way. By the union sum inequality, we obtain

$$\begin{aligned}
\mathrm{P}\left\{\left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty \geq C_\nabla\sqrt{\frac{\log(p)}{n}}\right\} &= \mathrm{P}\left[\bigcup_{k=1}^{K}\left\{\left|\frac{\partial l(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right| \geq C_\nabla\sqrt{\frac{\log(p)}{n}}\right\}\right] \\
&\leq \sum_{k=1}^{K}\mathrm{P}\left\{\left|\frac{\partial l(\boldsymbol{\alpha}^{(0)})}{\partial \alpha_k}\right| \geq C_\nabla\sqrt{\frac{\log(p)}{n}}\right\} \\
&\leq \sum_{k=1}^{K} 2\exp\left\{\frac{-n\frac{C_\nabla^2\log(p)}{n}}{8C_W^2}\right\} \\
&= 2K\exp\left\{\frac{-C_\nabla^2\log(p)}{8C_W^2}\right\} \\
&= 2\exp\left\{\frac{-C_\nabla^2\log(p)}{8C_W^2} + \log(K)\right\} \\
&= 2\exp\left\{\log\left(\frac{K}{p^{\frac{C_\nabla^2}{8C_W^2}}}\right)\right\} \\
&= 2\left(\frac{K}{p^{\frac{C_\nabla^2}{8C_W^2}}}\right) \\
&\to 0,
\end{aligned}$$

due to $K = o(p^{C_\nabla^2/(8C_W^2)})$ as $p \to \infty$. This leads to

$$\mathrm{P}\left\{\left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty \leq C_\nabla \sqrt{\frac{\log(p)}{n}}\right\} \to 1,$$

which proves claim (ii). □

Lemma 2. *Assume sample Condition 1′ and Condition 2 are satisfied. Suppose* $\|\boldsymbol{\delta}\|_2 = \bar{M}\sqrt{K\log(p)/n}$ *for a finite positive constant* $\bar{M} > 4C_\nabla/C_{\min}$ *and* $K\sqrt{\log(p)/n} = o(1)$ *as* $n, p \to \infty$. *Then for any* $\bar{\tau} \in [0,1]$, *with probability tending to one it holds that*

$$-\frac{1}{2}\boldsymbol{\delta}^\top \left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \bar{\tau}\boldsymbol{\delta})\right\}\boldsymbol{\delta} \geq \frac{C_{\min}}{4}\|\boldsymbol{\delta}\|_2^2.$$

Proof of Lemma 2. Recall $\nabla^2 l(\boldsymbol{\alpha}) = -\sum_{i=1}^n \boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha})\boldsymbol{\mathcal{X}}^{(i)}/(np)$. By applying the mean value theorem, we obtain $\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}\boldsymbol{\delta}) = \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)})+\bar{\tau}\nabla\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}^*\boldsymbol{\delta})^\top\boldsymbol{\delta}$, for some constant $\bar{\tau}^* \in (0,\bar{\tau})$, $i = 1,\cdots,n$ and $j = 1,\cdots,p$, which leads to

$$\begin{aligned}
&-\frac{1}{2}\boldsymbol{\delta}^\top \left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \bar{\tau}\boldsymbol{\delta})\right\}\boldsymbol{\delta} \\
&= \frac{1}{2np}\sum_{i=1}^n \boldsymbol{\delta}^\top \left\{\boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)} + \bar{\tau}\boldsymbol{\delta})\boldsymbol{\mathcal{X}}^{(i)}\right\}\boldsymbol{\delta}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2np}\sum_{i=1}^{n}\left(\boldsymbol{\mathcal{X}}^{(i)}\boldsymbol{\delta}\right)^{\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}\boldsymbol{\delta})\left(\boldsymbol{\mathcal{X}}^{(i)}\boldsymbol{\delta}\right)\\
&= \frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}\boldsymbol{\delta})\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^{2}\\
&= \frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)})\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^{2}\\
&\quad+\frac{\bar{\tau}}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\nabla\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}^{*}\boldsymbol{\delta})^{\top}\boldsymbol{\delta}\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^{2}. \qquad \text{(S5.1)}
\end{aligned}
$$

By sample Condition 1′, the first term of (S5.1) is bounded from below by

$$
\begin{aligned}
\frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)})\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^{2} &= \frac{1}{2np}\sum_{i=1}^{n}\left(\boldsymbol{\mathcal{X}}^{(i)}\boldsymbol{\delta}\right)^{\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)})\left(\boldsymbol{\mathcal{X}}^{(i)}\boldsymbol{\delta}\right)\\
&= \frac{1}{2}\boldsymbol{\delta}^{\top}\left\{\frac{1}{np}\sum_{i=1}^{n}\boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)})\boldsymbol{\mathcal{X}}^{(i)}\right\}\boldsymbol{\delta}\\
&= \frac{1}{2}\boldsymbol{\delta}^{\top}\boldsymbol{M}^{n}\boldsymbol{\delta}\\
&\geq \frac{1}{2}\Lambda_{\min}\left(\boldsymbol{M}^{n}\right)\|\boldsymbol{\delta}\|_{2}^{2}\\
&\geq \frac{C_{\min}}{2}\|\boldsymbol{\delta}\|_{2}^{2}.
\end{aligned}
$$

As for the second term of (S5.1), we have

$$
\begin{aligned}
&\frac{\bar{\tau}}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\nabla\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}^{*}\boldsymbol{\delta})^{\top}\boldsymbol{\delta}\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^{2}\\
&= \frac{\bar{\tau}}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\varepsilon_{j}^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}^{*}\boldsymbol{\delta})\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^{2}
\end{aligned}
$$

$$\geq -\frac{\bar{\tau}}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left|\varepsilon_j^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}^*\boldsymbol{\delta})\right|\left|\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right|\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^2$$
$$\geq -\frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left|\varepsilon_j^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}^*\boldsymbol{\delta})\right|\left|\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right|\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^2.$$

By Condition 2, we obtain

$$\begin{aligned}\left|\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right| &= \left|\sum_{k=1}^{K}\boldsymbol{W}_{j\cdot}^{(k)\top}\boldsymbol{y}_i\delta_k\right| \\ &\leq \sum_{k=1}^{K}\left|\boldsymbol{W}_{j\cdot}^{(k)\top}\boldsymbol{y}_i\right||\delta_k| \\ &\leq \sum_{k=1}^{K}\|\boldsymbol{W}_k\|_\infty|\delta_k| \\ &= \sum_{k=1}^{K}\|\boldsymbol{W}_k\|_1|\delta_k| \\ &\leq C_W\|\boldsymbol{\delta}\|_1,\end{aligned}$$

for all $i = 1, \cdots, n$ and $j = 1, \cdots, p$, where δ_k is the k-th element of $\boldsymbol{\delta}$. Combining this together with sample Condition 1′, we obtain

$$\begin{aligned}&\frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left|\varepsilon_j^{(i)}(\boldsymbol{\alpha}^{(0)}+\bar{\tau}^*\boldsymbol{\delta})\right|\left|\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right|\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^2 \\ &\leq \frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}C_W\|\boldsymbol{\delta}\|_1\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^2 \\ &= \frac{1}{2np}C_W\|\boldsymbol{\delta}\|_1\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^2\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2np} C_W \left\|\boldsymbol{\delta}\right\|_1 \sum_{i=1}^{n} \left(\boldsymbol{\mathcal{X}}^{(i)}\boldsymbol{\delta}\right)^\top \left(\boldsymbol{\mathcal{X}}^{(i)}\boldsymbol{\delta}\right) \\
&= \frac{1}{2} C_W \left\|\boldsymbol{\delta}\right\|_1 \boldsymbol{\delta}^\top \left(\frac{1}{np}\sum_{i=1}^{n} \boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\mathcal{X}}^{(i)}\right)\boldsymbol{\delta} \\
&= \frac{1}{2} C_W \left\|\boldsymbol{\delta}\right\|_1 \boldsymbol{\delta}^\top \boldsymbol{U}^n \boldsymbol{\delta} \\
&\leq \frac{1}{2} C_W \left\|\boldsymbol{\delta}\right\|_1 \Lambda_{\max}(\boldsymbol{U}^n) \left\|\boldsymbol{\delta}\right\|_2^2 \\
&\leq \frac{1}{2} C_W C_{\max} \left\|\boldsymbol{\delta}\right\|_1 \left\|\boldsymbol{\delta}\right\|_2^2 .
\end{aligned}
$$

Based on the conditions $\left\|\boldsymbol{\delta}\right\|_2 = \bar{M}\sqrt{K\log(p)/n}$ for a finite positive constant $\bar{M} > 4C_\nabla/C_{\min}$ and $K\sqrt{\log(p)/n} = o(1)$ as $n, p \to \infty$, we can apply inequality 4.67(c) in Seber (2008) to obtain

$$
\begin{aligned}
\left\|\boldsymbol{\delta}\right\|_1 &\leq \sqrt{K}\left\|\boldsymbol{\delta}\right\|_2 \\
&= \bar{M}\sqrt{K}\sqrt{\frac{K\log(p)}{n}} \\
&= \bar{M}K\sqrt{\frac{\log(p)}{n}} \\
&= o(1).
\end{aligned}
$$

Therefore, when n and p are large enough, we obtain $\left\|\boldsymbol{\delta}\right\|_1 \leq C_{\min}/(2C_{\max}C_W)$, which implies

$$
\frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p} \left|\varepsilon_j^{(i)}(\boldsymbol{\alpha}^{(0)} + \bar{\tau}^*\boldsymbol{\delta})\right| \left|\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right| \left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\boldsymbol{\delta}\right)^2
$$

$$
\begin{aligned}
&\leq \frac{1}{2} C_W C_{\max} \left\|\boldsymbol{\delta}\right\|_1 \left\|\boldsymbol{\delta}\right\|_2^2 \\
&\leq \frac{1}{2} C_W C_{\max} \frac{C_{\min}}{2 C_{\max} C_W} \left\|\boldsymbol{\delta}\right\|_2^2 \\
&= \frac{C_{\min}}{4} \left\|\boldsymbol{\delta}\right\|_2^2 ,
\end{aligned}
$$

and thus

$$
\begin{aligned}
&\frac{\bar{\tau}}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \eta_{jj}^{(i)} (\boldsymbol{\alpha}^{(0)} + \bar{\tau}^* \boldsymbol{\delta})^\top \boldsymbol{\delta} \left(\boldsymbol{\mathcal{X}}^{(i,j)\top} \boldsymbol{\delta} \right)^2 \\
&\geq -\frac{1}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left| \varepsilon_j^{(i)} (\boldsymbol{\alpha}^{(0)} + \bar{\tau}^* \boldsymbol{\delta}) \right| \left| \boldsymbol{\mathcal{X}}^{(i,j)\top} \boldsymbol{\delta} \right| \left(\boldsymbol{\mathcal{X}}^{(i,j)\top} \boldsymbol{\delta} \right)^2 \\
&\geq -\frac{C_{\min}}{4} \left\|\boldsymbol{\delta}\right\|_2^2 .
\end{aligned}
$$

Finally, combining the lower bounds of the first and second terms of (S5.1), we obtain

$$
-\frac{1}{2} \boldsymbol{\delta}^\top \left\{ \nabla^2 l(\boldsymbol{\alpha}^{(0)} + \bar{\tau} \boldsymbol{\delta}) \right\} \boldsymbol{\delta} \geq \frac{C_{\min}}{2} \left\|\boldsymbol{\delta}\right\|_2^2 - \frac{C_{\min}}{4} \left\|\boldsymbol{\delta}\right\|_2^2 = \frac{C_{\min}}{4} \left\|\boldsymbol{\delta}\right\|_2^2 ,
$$

as required. □

Proof of Proposition 1. We start by defining

$$
\bar{G}(\boldsymbol{\delta}) = -\left\{ l(\boldsymbol{\alpha}^{(0)} + \boldsymbol{\delta}) - l(\boldsymbol{\alpha}^{(0)}) \right\} .
$$

Based on the definition of the unregularized pseudo-likelihood estimator $\bar{\boldsymbol{\alpha}}$, it can be seen that $\bar{\boldsymbol{\delta}} = \bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}$ minimizes $\bar{G}(\boldsymbol{\delta})$. Also, we have $\bar{G}(\mathbf{0}_K) = 0$, which implies $\bar{G}(\bar{\boldsymbol{\delta}}) \leq 0$, where $\mathbf{0}_K$ is a K-dimensional vector of zeros. Therefore, it suffices to show $\bar{G}(\cdot)$ is strictly positive everywhere on the boundary $\partial\bar{\mathcal{A}} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 = \bar{M}\sqrt{K\log(p)/n}\}$ of the ball $\bar{\mathcal{A}} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq \bar{M}\sqrt{K\log(p)/n}\}$ (Rothman et al., 2008).

By using a first order Taylor expansion on $l(\boldsymbol{\alpha}^{(0)} + \boldsymbol{\delta})$ at the point $\boldsymbol{\alpha}^{(0)}$, we obtain

$$l\left(\boldsymbol{\alpha}^{(0)} + \boldsymbol{\delta}\right) - l\left(\boldsymbol{\alpha}^{(0)}\right) = \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}^\top \boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^\top \left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \bar{\tau}\boldsymbol{\delta})\right\}\boldsymbol{\delta},$$

for some constant $\bar{\tau} \in [0, 1]$. Therefore, we can decompose

$$\bar{G}(\boldsymbol{\delta}) = -\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}^\top \boldsymbol{\delta} - \frac{1}{2}\boldsymbol{\delta}^\top \left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \bar{\tau}\boldsymbol{\delta})\right\}\boldsymbol{\delta} = \bar{I}_1 + \bar{I}_2, \quad \text{(S5.2)}$$

where $\bar{I}_1 = -\{\nabla l(\boldsymbol{\alpha}^{(0)})\}^\top\boldsymbol{\delta}$ and $\bar{I}_2 = -(1/2)\boldsymbol{\delta}^\top\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \bar{\tau}\boldsymbol{\delta})\}\boldsymbol{\delta}$. Since C_∇ satisfies $K = o(p^{C_\nabla^2/(8C_W^2)})$, we can apply Lemma 1(ii) along with inequality 4.56(c) in Seber (2008) to obtain, with probability tending to one,

$$\begin{aligned}\left|\bar{I}_1\right| &= \left|-\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}^\top \boldsymbol{\delta}\right| \\ &= \left|\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}^\top \boldsymbol{\delta}\right|\end{aligned}$$

$$
\begin{aligned}
&\leq \left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_{\infty} \|\boldsymbol{\delta}\|_1 \\
&\leq \left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_{\infty} \sqrt{K}\, \|\boldsymbol{\delta}\|_2 \\
&\leq C_{\nabla} \sqrt{\frac{\log(p)}{n}} \sqrt{K} \bar{M} \sqrt{\frac{K \log(p)}{n}} \\
&= C_{\nabla} \bar{M} K \frac{\log(p)}{n},
\end{aligned}
$$

which leads to

$$
\bar{I}_1 \geq -C_{\nabla} \bar{M} K \frac{\log(p)}{n}. \tag{S5.3}
$$

Next, by Lemma 2, with probability tending to one it holds that

$$
\begin{aligned}
\bar{I}_2 &= -\frac{1}{2} \boldsymbol{\delta}^{\top} \left\{ \nabla^2 l(\boldsymbol{\alpha}^{(0)} + \bar{\tau}\boldsymbol{\delta}) \right\} \boldsymbol{\delta} \\
&\geq \frac{C_{\min}}{4} \|\boldsymbol{\delta}\|_2^2 \\
&= \frac{C_{\min}}{4} \bar{M}^2 \frac{K \log(p)}{n}.
\end{aligned} \tag{S5.4}
$$

Using the lower bounds on $\bar{I}_1$ and $\bar{I}_2$ from (S5.3) and (S5.4), respectively, we obtain

$$
\begin{aligned}
\bar{G}(\boldsymbol{\delta}) &= \bar{I}_1 + \bar{I}_2 \\
&\geq -C_{\nabla} \bar{M} K \frac{\log(p)}{n} + \frac{C_{\min}}{4} \bar{M}^2 \frac{K \log(p)}{n} \\
&= \bar{M}^2 K \frac{\log(p)}{n} \left(\frac{C_{\min}}{4} - \frac{C_{\nabla}}{\bar{M}} \right)
\end{aligned}
$$

$$> 0,$$

where the last inequality holds due to the condition $\bar{M} > 4C_{\nabla}/C_{\min}$. Hence with probability tending to one, we obtain

$$\left\|\bar{\boldsymbol{\delta}}\right\|_2 = \left\|\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\right\|_2 \leq \bar{M}\sqrt{\frac{K\log(p)}{n}},$$

as required. □

Lemma 3. *Assume Condition 2 is satisfied. Then for any $\epsilon > 0$, it holds that*

(i) $\mathrm{P}(\|\boldsymbol{M}^n - \boldsymbol{M}^0\|_\infty \geq \epsilon) \leq 2\exp\{-\epsilon^2 n/(8C_W^4K^2) + 2\log(K)\}$;

(ii) $\mathrm{P}(\|\boldsymbol{U}^n - \boldsymbol{U}^0\|_\infty \geq \epsilon) \leq 2\exp\{-\epsilon^2 n/(8C_W^4K^2) + 2\log(K)\}$.

Proof of Lemma 3. We first prove claim (i). Note the (l_1, l_2)-th element of matrix $\boldsymbol{M}^n - \boldsymbol{M}^0$ is given by $\sum_{i=1}^n v_{l_1,l_2}^{(i)}/(np)$, where

$$v_{l_1,l_2}^{(i)} = (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)})\boldsymbol{\mathcal{X}}_{l_2}^{(i)} - E\left\{(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)})\boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right\},$$

for $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Also, recall $\boldsymbol{\mathcal{X}}_l^{(i)}$ denotes the l-th column

of $\boldsymbol{\mathcal{X}}^{(i)}$ for $i = 1, \cdots, n$ and $l = 1, \cdots, K$. That is,

$$\boldsymbol{\mathcal{X}}_l^{(i)} = \begin{pmatrix} \boldsymbol{W}_{1\cdot}^{(l)\top} \boldsymbol{y}_i \\ \vdots \\ \boldsymbol{W}_{p\cdot}^{(l)\top} \boldsymbol{y}_i \end{pmatrix}.$$

Therefore, it can be seen that $v_{l_1,l_2}^{(i)}$ has mean zero for $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Under Condition 2 then, we have for all $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$,

$$\begin{aligned}
\left| (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)} \right| &= \left| \sum_{j=1}^{p} \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}) \left(\boldsymbol{W}_{j\cdot}^{(l_1)\top} \boldsymbol{y}_i \right) \left(\boldsymbol{W}_{j\cdot}^{(l_2)\top} \boldsymbol{y}_i \right) \right| \\
&\leq \sum_{j=1}^{p} \left| \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}) \right| \left| \boldsymbol{W}_{j\cdot}^{(l_1)\top} \boldsymbol{y}_i \right| \left| \boldsymbol{W}_{j\cdot}^{(l_2)\top} \boldsymbol{y}_i \right| \\
&\leq \sum_{j=1}^{p} (1) C_W C_W \\
&= p C_W^2,
\end{aligned}$$

which implies

$$\begin{aligned}
\left| v_{l_1,l_2}^{(i)} \right| &= \left| (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)} - E\left\{ (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)} \right\} \right| \\
&\leq \left| (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)} \right| + \left| E\left\{ (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)} \right\} \right| \\
&\leq \left| (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)} \right| + E\left\{ \left| (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)} \right| \right\}
\end{aligned}$$

$$\leq pC_W^2 + pC_W^2$$

$$= 2pC_W^2.$$

Therefore, we can apply the Azuma-Hoeffding inequality to obtain

$$\begin{aligned}
\mathrm{P}\left(\left|\frac{1}{np}\sum_{i=1}^{n} v_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon}{K}\right) &= \mathrm{P}\left(\left|\sum_{i=1}^{n} v_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon np}{K}\right) \\
&= \mathrm{P}\left[\left|\sum_{i=1}^{n}\left\{v_{l_1,l_2}^{(i)} - E\left(v_{l_1,l_2}^{(i)}\right)\right\}\right| \geq \frac{\epsilon np}{K}\right] \\
&\leq 2\exp\left\{\frac{\frac{-2\epsilon^2 n^2 p^2}{K^2}}{\sum_{i=1}^{n}\left(2pC_W^2 + 2pC_W^2\right)^2}\right\} \\
&= 2\exp\left(\frac{-2\epsilon^2 n^2 p^2}{16nK^2 p^2 C_W^4}\right) \\
&= 2\exp\left(\frac{-\epsilon^2 n}{8C_W^4 K^2}\right),
\end{aligned}$$

for all $l_1, l_2 = 1, \cdots, K$ and $\epsilon > 0$. It follows that

$$\begin{aligned}
\mathrm{P}\left(\left\|\boldsymbol{M}^n - \boldsymbol{M}^0\right\|_\infty \geq \epsilon\right) &= \mathrm{P}\left\{\bigcup_{l_1=1}^{K}\left(\sum_{l_2=1}^{K}\left|\frac{1}{np}\sum_{i=1}^{n} v_{l_1,l_2}^{(i)}\right| \geq \epsilon\right)\right\} \\
&\leq \sum_{l_1=1}^{K}\mathrm{P}\left(\sum_{l_2=1}^{K}\left|\frac{1}{np}\sum_{i=1}^{n} v_{l_1,l_2}^{(i)}\right| \geq \epsilon\right) \\
&\leq \sum_{l_1=1}^{K}\mathrm{P}\left(\bigcup_{l_2=1}^{K}\left|\frac{1}{np}\sum_{i=1}^{n} v_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon}{K}\right) \\
&\leq \sum_{l_1=1}^{K}\sum_{l_2=1}^{K}\mathrm{P}\left(\left|\frac{1}{np}\sum_{i=1}^{n} v_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon}{K}\right)
\end{aligned}$$

$$\leq K^2 2\exp\left(\frac{-\epsilon^2 n}{8C_W^4 K^2}\right)$$
$$= 2\exp\left\{\frac{-\epsilon^2 n}{8C_W^4 K^2} + 2\log(K)\right\},$$

for all $\epsilon > 0$, which proves Lemma 3(i).

Next, we prove claim (ii). Note the (l_1, l_2)-th element of matrix $\boldsymbol{U}^n - \boldsymbol{U}^0$ is given by $\sum_{i=1}^n h_{l_1,l_2}^{(i)}/(np)$, where

$$h_{l_1,l_2}^{(i)} = (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)} - E\left\{(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right\},$$

for $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. It can be seen that $h_{l_1,l_2}^{(i)}$ has mean zero for all $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Furthermore, under Condition 2, we have for all $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$,

$$\left|(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right| = \left|\sum_{j=1}^p \left(\boldsymbol{W}_{j\cdot}^{(l_1)\top} \boldsymbol{y}_i\right)\left(\boldsymbol{W}_{j\cdot}^{(l_2)\top} \boldsymbol{y}_i\right)\right|$$
$$\leq \sum_{j=1}^p \left|\boldsymbol{W}_{j\cdot}^{(l_1)\top} \boldsymbol{y}_i\right| \left|\boldsymbol{W}_{j\cdot}^{(l_2)\top} \boldsymbol{y}_i\right|$$
$$\leq \sum_{j=1}^p C_W C_W$$
$$= pC_W^2,$$

which implies

$$\begin{aligned}
\left|h_{l_1,l_2}^{(i)}\right| &= \left|(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)} - E\left\{(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right\}\right| \\
&\leq \left|(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right| + \left|E\left\{(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right\}\right| \\
&\leq \left|(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right| + E\left\{\left|(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right|\right\} \\
&\leq pC_W^2 + pC_W^2 \\
&= 2pC_W^2.
\end{aligned}$$

Thus, we can again apply the Azuma-Hoeffding inequality to obtain

$$\begin{aligned}
\mathrm{P}\left(\left|\frac{1}{np}\sum_{i=1}^{n} h_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon}{K}\right) &= \mathrm{P}\left(\left|\sum_{i=1}^{n} h_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon np}{K}\right) \\
&= \mathrm{P}\left[\left|\sum_{i=1}^{n}\left\{h_{l_1,l_2}^{(i)} - E\left(h_{l_1,l_2}^{(i)}\right)\right\}\right| \geq \frac{\epsilon np}{K}\right] \\
&\leq 2\exp\left\{\frac{\frac{-2\epsilon^2 n^2 p^2}{K^2}}{\sum_{i=1}^{n}\left(2pC_W^2 + 2pC_W^2\right)^2}\right\} \\
&= 2\exp\left(\frac{-2\epsilon^2 n^2 p^2}{16nK^2 p^2 C_W^4}\right) \\
&= 2\exp\left(\frac{-\epsilon^2 n}{8C_W^4 K^2}\right),
\end{aligned}$$

for all $l_1, l_2 = 1, \cdots, K$ and $\epsilon > 0$. It follows that

$$\mathrm{P}\left(\left\|\boldsymbol{U}^n - \boldsymbol{U}^0\right\|_\infty \geq \epsilon\right) = \mathrm{P}\left\{\bigcup_{l_1=1}^{K}\left(\sum_{l_2=1}^{K}\left|\frac{1}{np}\sum_{i=1}^{n} h_{l_1,l_2}^{(i)}\right| \geq \epsilon\right)\right\}$$

$$
\begin{aligned}
&\leq \sum_{l_1=1}^{K} \mathrm{P}\left(\sum_{l_2=1}^{K} \left| \frac{1}{np} \sum_{i=1}^{n} h_{l_1,l_2}^{(i)} \right| \geq \epsilon \right) \\
&\leq \sum_{l_1=1}^{K} \mathrm{P}\left(\bigcup_{l_2=1}^{K} \left| \frac{1}{np} \sum_{i=1}^{n} h_{l_1,l_2}^{(i)} \right| \geq \frac{\epsilon}{K} \right) \\
&\leq \sum_{l_1=1}^{K} \sum_{l_2=1}^{K} \mathrm{P}\left(\left| \frac{1}{np} \sum_{i=1}^{n} h_{l_1,l_2}^{(i)} \right| \geq \frac{\epsilon}{K} \right) \\
&\leq K^2 2 \exp\left(\frac{-\epsilon^2 n}{8 C_W^4 K^2} \right) \\
&= 2 \exp\left\{ \frac{-\epsilon^2 n}{8 C_W^4 K^2} + 2 \log(K) \right\},
\end{aligned}
$$

for all $\epsilon > 0$, which proves Lemma 3(ii). □

Proof of Proposition 2. We first prove claim (i). By Condition 1 and inequality 4.67(e) in Seber (2008), we have

$$
\begin{aligned}
\Lambda_{\min}(\boldsymbol{M}^n) &= \min_{\|\boldsymbol{x}\|_2=1} \left\{ \boldsymbol{x}^\top \boldsymbol{M}^0 \boldsymbol{x} + \boldsymbol{x}^\top \left(\boldsymbol{M}^n - \boldsymbol{M}^0 \right) \boldsymbol{x} \right\} \\
&\geq \Lambda_{\min}(\boldsymbol{M}^0) - \left\| \boldsymbol{M}^n - \boldsymbol{M}^0 \right\|_2 \\
&\geq C_{\min} - \left\| \boldsymbol{M}^n - \boldsymbol{M}^0 \right\|_\infty .
\end{aligned}
$$

This implies $\{\Lambda_{\min}(\boldsymbol{M}^n) \leq C_{\min} - \epsilon\} \subseteq \{\|\boldsymbol{M}^n - \boldsymbol{M}^0\|_\infty \geq \epsilon\}$ for all $\epsilon > 0$, due to

$$
C_{\min} - \left\| \boldsymbol{M}^n - \boldsymbol{M}^0 \right\|_\infty \leq \Lambda_{\min}(\boldsymbol{M}^n) \leq C_{\min} - \epsilon \implies \left\| \boldsymbol{M}^n - \boldsymbol{M}^0 \right\|_\infty \geq \epsilon.
$$

Therefore, by Lemma 3(i), we obtain for all $\epsilon > 0$,

$$\begin{aligned}\mathrm{P}\left\{\Lambda_{\min}(\boldsymbol{M}^n) \leq C_{\min} - \epsilon\right\} &\leq \mathrm{P}\left(\left\|\boldsymbol{M}^n - \boldsymbol{M}^0\right\|_\infty \geq \epsilon\right) \\ &\leq 2\exp\left\{-\frac{\epsilon^2 n}{8C_W^4 K^2} + 2\log(K)\right\},\end{aligned}$$

which proves claim (i).

Next, we prove claim (ii). By Condition 1 and inequality 4.67(e) in Seber (2008), we have

$$\begin{aligned}\Lambda_{\max}(\boldsymbol{U}^n) &= \max_{\|\boldsymbol{x}\|_2=1}\left\{\boldsymbol{x}^\top \boldsymbol{U}^0 \boldsymbol{x} + \boldsymbol{x}^\top\left(\boldsymbol{U}^n - \boldsymbol{U}^0\right)\boldsymbol{x}\right\} \\ &\leq \Lambda_{\max}(\boldsymbol{U}^0) + \left\|\boldsymbol{U}^n - \boldsymbol{U}^0\right\|_2 \\ &\leq C_{\max} + \left\|\boldsymbol{U}^n - \boldsymbol{U}^0\right\|_\infty.\end{aligned}$$

This implies $\{\Lambda_{\max}(\boldsymbol{U}^n) \geq C_{\max} + \epsilon\} \subseteq \{\|\boldsymbol{U}^n - \boldsymbol{U}^0\|_\infty \geq \epsilon\}$ for all $\epsilon > 0$, due to

$$C_{\max} + \epsilon \leq \Lambda_{\max}(\boldsymbol{U}^n) \leq C_{\max} + \left\|\boldsymbol{U}^n - \boldsymbol{U}^0\right\|_\infty \implies \left\|\boldsymbol{U}^n - \boldsymbol{U}^0\right\|_\infty \geq \epsilon.$$

Therefore, by Lemma 3(ii), we obtain for all $\epsilon > 0$,

$$\mathrm{P}\left\{\Lambda_{\max}(\boldsymbol{U}^n) \geq C_{\max} + \epsilon\right\} \leq \mathrm{P}\left(\left\|\boldsymbol{U}^n - \boldsymbol{U}^0\right\|_\infty \geq \epsilon\right)$$

$$\leq 2\exp\left\{-\frac{\epsilon^2 n}{8C_W^4 K^2}+2\log(K)\right\},$$

which proves claim (ii). □

Lemma 4. *Assume sample Condition 1′ and Condition 2 are satisfied. Suppose* $\|\boldsymbol{\delta}_S\|_2 = M\sqrt{K_0\log(p)/n}$ *for a finite positive constant* $M > 4/C_{\min}$ *and* $K_0\sqrt{\log(p)/n} = o(1)$ *as* $n, p \to \infty$. *Then for any* $\tau \in [0,1]$, *with probability tending to one it holds that*

$$-\frac{1}{2}\boldsymbol{\delta}_S^\top\left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)}+\tau\boldsymbol{\delta}^{[S]})\right\}_{S,S}\boldsymbol{\delta}_S \geq \frac{C_{\min}}{4}\|\boldsymbol{\delta}_S\|_2^2,$$

where $\boldsymbol{\delta}^{[S]} = (\delta_1^{[S]},\cdots,\delta_K^{[S]})$ *with* $\delta_k^{[S]} = \delta_k$ *for* $k \in S$ *and* $\delta_k^{[S]} = 0$ *for* $k \in S^c$.

Proof of Lemma 4. The proof of this lemma follows along similar lines as that of Lemma 2. Recall $\nabla^2 l(\boldsymbol{\alpha}) = -\sum_{i=1}^n \boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha})\boldsymbol{\mathcal{X}}^{(i)}/(np)$. By applying the mean value theorem, we obtain $\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}+\tau\boldsymbol{\delta}^{[S]}) = \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}) + \tau\nabla\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)} + \tau^*\boldsymbol{\delta}^{[S]})^\top\boldsymbol{\delta}^{[S]}$, for some constant $\tau^* \in (0,\tau)$, $i = 1,\cdots,n$ and $j = 1,\cdots,p$, which leads to

$$\begin{aligned}
&-\frac{1}{2}\boldsymbol{\delta}_S^\top\left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)}+\tau\boldsymbol{\delta}^{[S]})\right\}_{S,S}\boldsymbol{\delta}_S \\
&= \frac{1}{2np}\sum_{i=1}^n \boldsymbol{\delta}_S^\top\left\{\boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}+\tau\boldsymbol{\delta}^{[S]})\boldsymbol{\mathcal{X}}^{(i)}\right\}_{S,S}\boldsymbol{\delta}_S \\
&= \frac{1}{2np}\sum_{i=1}^n \left(\boldsymbol{\mathcal{X}}_S^{(i)}\boldsymbol{\delta}_S\right)^\top\boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}+\tau\boldsymbol{\delta}^{[S]})\left(\boldsymbol{\mathcal{X}}_S^{(i)}\boldsymbol{\delta}_S\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)} + \tau \boldsymbol{\delta}^{[S]}) \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&= \frac{1}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}) \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&\quad + \frac{\tau}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \nabla \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)} + \tau^* \boldsymbol{\delta}^{[S]})^\top \boldsymbol{\delta}^{[S]} \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 . \qquad \text{(S5.5)}
\end{aligned}
$$

By sample Condition 1′, the first term of (S5.5) is bounded from below by

$$
\begin{aligned}
&\frac{1}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}) \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&= \frac{1}{2np} \sum_{i=1}^{n} \left(\boldsymbol{\mathcal{X}}_S^{(i)} \boldsymbol{\delta}_S \right)^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \left(\boldsymbol{\mathcal{X}}_S^{(i)} \boldsymbol{\delta}_S \right) \\
&= \frac{1}{2} \boldsymbol{\delta}_S^\top \left\{ \frac{1}{np} \sum_{i=1}^{n} \boldsymbol{\mathcal{X}}_S^{(i)\top} \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_S^{(i)} \right\} \boldsymbol{\delta}_S \\
&= \frac{1}{2} \boldsymbol{\delta}_S^\top \left\{ \frac{1}{np} \sum_{i=1}^{n} \boldsymbol{\mathcal{X}}^{(i)\top} \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}^{(i)} \right\}_{S,S} \boldsymbol{\delta}_S \\
&= \frac{1}{2} \boldsymbol{\delta}_S^\top \boldsymbol{M}_{S,S}^n \boldsymbol{\delta}_S \\
&\geq \frac{1}{2} \Lambda_{\min} \left(\boldsymbol{M}_{S,S}^n \right) \| \boldsymbol{\delta}_S \|_2^2 \\
&\geq \frac{C_{\min}}{2} \| \boldsymbol{\delta}_S \|_2^2 .
\end{aligned}
$$

As for the second term of (S5.5), we have

$$
\frac{\tau}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \nabla \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)} + \tau^* \boldsymbol{\delta}^{[S]})^\top \boldsymbol{\delta}^{[S]} \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2
$$

$$
\begin{aligned}
&= \frac{\tau}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \varepsilon_j^{(i)} (\boldsymbol{\alpha}^{(0)} + \tau^* \boldsymbol{\delta}^{[S]}) \boldsymbol{\mathcal{X}}^{(i,j)\top} \boldsymbol{\delta}^{[S]} \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&= \frac{\tau}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \varepsilon_j^{(i)} (\boldsymbol{\alpha}^{(0)} + \tau^* \boldsymbol{\delta}^{[S]}) \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right) \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&\geq -\frac{\tau}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left| \varepsilon_j^{(i)} (\boldsymbol{\alpha}^{(0)} + \tau^* \boldsymbol{\delta}^{[S]}) \right| \left| \boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right| \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&\geq -\frac{1}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left| \varepsilon_j^{(i)} (\boldsymbol{\alpha}^{(0)} + \tau^* \boldsymbol{\delta}^{[S]}) \right| \left| \boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right| \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 .
\end{aligned}
$$

By Condition 2 and similar techniques used in the proof of Lemma 2, we obtain $|\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S| \leq C_W \|\boldsymbol{\delta}_S\|_1$ for all $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Combining this with sample Condition 1′, we obtain

$$
\begin{aligned}
&\frac{1}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left| \varepsilon_j^{(i)} (\boldsymbol{\alpha}^{(0)} + \tau^* \boldsymbol{\delta}^{[S]}) \right| \left| \boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right| \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&\leq \frac{1}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} C_W \|\boldsymbol{\delta}_S\|_1 \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&= \frac{1}{2np} C_W \|\boldsymbol{\delta}_S\|_1 \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&= \frac{1}{2np} C_W \|\boldsymbol{\delta}_S\|_1 \sum_{i=1}^{n} \left(\boldsymbol{\mathcal{X}}_S^{(i)} \boldsymbol{\delta}_S \right)^\top \left(\boldsymbol{\mathcal{X}}_S^{(i)} \boldsymbol{\delta}_S \right) \\
&= \frac{1}{2} C_W \|\boldsymbol{\delta}_S\|_1 \boldsymbol{\delta}_S^\top \left(\frac{1}{np} \sum_{i=1}^{n} \boldsymbol{\mathcal{X}}_S^{(i)\top} \boldsymbol{\mathcal{X}}_S^{(i)} \right) \boldsymbol{\delta}_S \\
&= \frac{1}{2} C_W \|\boldsymbol{\delta}_S\|_1 \boldsymbol{\delta}_S^\top \boldsymbol{U}_{S,S}^n \boldsymbol{\delta}_S \\
&\leq \frac{1}{2} C_W \|\boldsymbol{\delta}_S\|_1 \Lambda_{\max}(\boldsymbol{U}_{S,S}^n) \|\boldsymbol{\delta}_S\|_2^2
\end{aligned}
$$

$$\leq \frac{1}{2} C_W C_{\max} \left\| \boldsymbol{\delta}_S \right\|_1 \left\| \boldsymbol{\delta}_S \right\|_2^2 .$$

Based on the conditions $\|\boldsymbol{\delta}_S\|_2 = M\sqrt{K_0 \log(p)/n}$ for a finite positive constant $M > 4/C_{\min}$ and $K_0\sqrt{\log(p)/n} = o(1)$ as $n, p \to \infty$, we can then apply inequality 4.67(c) in Seber (2008) to obtain

$$\begin{aligned}
\left\| \boldsymbol{\delta}_S \right\|_1 &\leq \sqrt{K_0} \left\| \boldsymbol{\delta}_S \right\|_2 \\
&= M\sqrt{K_0}\sqrt{\frac{K_0 \log(p)}{n}} \\
&= MK_0\sqrt{\frac{\log(p)}{n}} \\
&= o(1).
\end{aligned}$$

Thus, when n and p are large enough, we have $\|\boldsymbol{\delta}_S\|_1 \leq C_{\min}/(2C_{\max}C_W)$, which implies

$$\begin{aligned}
&\frac{1}{2np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left| \varepsilon_j^{(i)}(\boldsymbol{\alpha}^{(0)} + \tau^* \boldsymbol{\delta}^{[S]}) \right| \left| \boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right| \left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top} \boldsymbol{\delta}_S \right)^2 \\
&\leq \frac{1}{2} C_W C_{\max} \left\| \boldsymbol{\delta}_S \right\|_1 \left\| \boldsymbol{\delta}_S \right\|_2^2 \\
&\leq \frac{1}{2} C_W C_{\max} \frac{C_{\min}}{2C_{\max}C_W} \left\| \boldsymbol{\delta}_S \right\|_2^2 \\
&= \frac{C_{\min}}{4} \left\| \boldsymbol{\delta}_S \right\|_2^2 ,
\end{aligned}$$

and hence

$$\begin{aligned}
&\frac{\tau}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}+\tau^{*}\boldsymbol{\delta}^{[S]})^{\top}\boldsymbol{\delta}^{[S]}\left(\boldsymbol{\mathcal{X}}_{S}^{(i,j)\top}\boldsymbol{\delta}_{S}\right)^{2}\\
&\geq-\frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left|\varepsilon_{j}^{(i)}(\boldsymbol{\alpha}^{(0)}+\tau^{*}\boldsymbol{\delta}^{[S]})\right|\left|\boldsymbol{\mathcal{X}}_{S}^{(i,j)\top}\boldsymbol{\delta}_{S}\right|\left(\boldsymbol{\mathcal{X}}_{S}^{(i,j)\top}\boldsymbol{\delta}_{S}\right)^{2}\\
&\geq-\frac{C_{\min}}{4}\left\|\boldsymbol{\delta}_{S}\right\|_{2}^{2}.
\end{aligned}$$

Finally, combining the lower bounds of the first and second terms of (S5.5), we obtain

$$-\frac{1}{2}\boldsymbol{\delta}_{S}^{\top}\left\{\nabla^{2}l(\boldsymbol{\alpha}^{(0)}+\tau\boldsymbol{\delta}^{[S]})\right\}_{S,S}\boldsymbol{\delta}_{S}\geq\frac{C_{\min}}{2}\left\|\boldsymbol{\delta}_{S}\right\|_{2}^{2}-\frac{C_{\min}}{4}\left\|\boldsymbol{\delta}_{S}\right\|_{2}^{2}=\frac{C_{\min}}{4}\left\|\boldsymbol{\delta}_{S}\right\|_{2}^{2},$$

as required. □

Proof of Proposition 3. The proof of this proposition follows along similar lines as that of Proposition 1. We start by defining

$$G(\boldsymbol{\delta}_{S})=-\left\{l(\boldsymbol{\alpha}^{(0)}+\boldsymbol{\delta}^{[S]})-l(\boldsymbol{\alpha}^{(0)})\right\}+\sum_{k\in S}\frac{\lambda}{|\bar{\alpha}_{k}|}\left(|\alpha_{k}^{(0)}+\delta_{k}^{[S]}|-|\alpha_{k}^{(0)}|\right),$$

where $\boldsymbol{\delta}^{[S]}=(\delta_{1}^{[S]},\cdots,\delta_{K}^{[S]})$ with $\delta_{k}^{[S]}=\delta_{k}$ for $k\in S$ and $\delta_{k}^{[S]}=0$ for $k\in S^{c}$. Based on the definition of $\tilde{\boldsymbol{\alpha}}$ as the minimizer of criterion (S3.1), it can be seen that $\tilde{\boldsymbol{\delta}}_{S}=\tilde{\boldsymbol{\alpha}}_{S}-\boldsymbol{\alpha}_{S}^{(0)}$ minimizes $G(\boldsymbol{\delta}_{S})$. Also, we have $G(\mathbf{0}_{K_{0}})=0$, which implies $G(\tilde{\boldsymbol{\delta}}_{S})\leq 0$. Therefore, it suffices to show $G(\cdot)$ is strictly positive

everywhere on the boundary $\partial\mathcal{A} = \{\boldsymbol{\delta}_S : \|\boldsymbol{\delta}_S\|_2 = M\sqrt{K_0 \log(p)/n}\}$ of the ball $\mathcal{A} = \{\boldsymbol{\delta}_S : \|\boldsymbol{\delta}_S\|_2 \leq M\sqrt{K_0 \log(p)/n}\}$ (Rothman et al., 2008).

Using a first order Taylor expansion on $l(\boldsymbol{\alpha}^{(0)} + \boldsymbol{\delta}^{[S]})$ at the point $\boldsymbol{\alpha}^{(0)}$, we obtain

$$\begin{aligned} l\left(\boldsymbol{\alpha}^{(0)} + \boldsymbol{\delta}^{[S]}\right) - l\left(\boldsymbol{\alpha}^{(0)}\right) &= \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}^\top \boldsymbol{\delta}^{[S]} + \frac{1}{2}\boldsymbol{\delta}^{[S]\top}\left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \tau\boldsymbol{\delta}^{[S]})\right\}\boldsymbol{\delta}^{[S]} \\ &= \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S^\top \boldsymbol{\delta}_S + \frac{1}{2}\boldsymbol{\delta}_S^\top\left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \tau\boldsymbol{\delta}^{[S]})\right\}_{S,S}\boldsymbol{\delta}_S, \end{aligned}$$

for some constant $\tau \in [0, 1]$. Therefore, we can decompose

$$\begin{aligned} &G(\boldsymbol{\delta}_S) \\ &= -\left\{l\left(\boldsymbol{\alpha}^{(0)} + \boldsymbol{\delta}^{[S]}\right) - l(\boldsymbol{\alpha}^{(0)})\right\} + \sum_{k\in S}\frac{\lambda}{|\bar{\alpha}_k|}\left(\left|\alpha_k^{(0)} + \delta_k^{[S]}\right| - \left|\alpha_k^{(0)}\right|\right) \\ &= -\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S^\top \boldsymbol{\delta}_S - \frac{1}{2}\boldsymbol{\delta}_S^\top\left\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \tau\boldsymbol{\delta}^{[S]})\right\}_{S,S}\boldsymbol{\delta}_S \\ &\quad + \sum_{k\in S}\frac{\lambda}{|\bar{\alpha}_k|}\left(\left|\alpha_k^{(0)} + \delta_k^{[S]}\right| - \left|\alpha_k^{(0)}\right|\right) \\ &= I_1 + I_2 + I_3, \end{aligned}$$

where $I_1 = -\{\nabla l(\boldsymbol{\alpha}^{(0)})\}_S^\top\boldsymbol{\delta}_S$, $I_2 = -(1/2)\boldsymbol{\delta}_S^\top\{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \tau\boldsymbol{\delta}^{[S]})\}_{S,S}\boldsymbol{\delta}_S$, and $I_3 = \sum_{k\in S}(\lambda/|\bar{\alpha}_k|)(|\alpha_k^{(0)} + \delta_k^{[S]}| - |\alpha_k^{(0)}|)$. By Lemma 1(i) and inequality 4.56(c)

in Seber (2008), we obtain with probability tending to one,

$$\begin{aligned}
|I_1| &= \left| - \left\{ \nabla l(\boldsymbol{\alpha}^{(0)}) \right\}_S^\top \boldsymbol{\delta}_S \right| \\
&= \left| \left\{ \nabla l(\boldsymbol{\alpha}^{(0)}) \right\}_S^\top \boldsymbol{\delta}_S \right| \\
&\leq \left\| \left\{ \nabla l(\boldsymbol{\alpha}^{(0)}) \right\}_S \right\|_\infty \left\| \boldsymbol{\delta}_S \right\|_1 \\
&\leq \left\| \left\{ \nabla l(\boldsymbol{\alpha}^{(0)}) \right\}_S \right\|_\infty \sqrt{K_0} \left\| \boldsymbol{\delta}_S \right\|_2 \\
&= \left\| \left\{ \nabla l(\boldsymbol{\alpha}^{(0)}) \right\}_S \right\|_\infty \sqrt{K_0} M \sqrt{\frac{K_0 \log(p)}{n}} \\
&\leq \sqrt{\frac{\log(p)}{n}} \sqrt{K_0} M \sqrt{\frac{K_0 \log(p)}{n}} \\
&= M K_0 \frac{\log(p)}{n},
\end{aligned}$$

which leads to

$$I_1 \geq -M K_0 \frac{\log(p)}{n}. \tag{S5.6}$$

Since $K\sqrt{\log(p)/n} = o(1)$ implies $K_0\sqrt{\log(p)/n} = o(1)$, then we can apply Lemma 4 and obtain with high probability,

$$\begin{aligned}
I_2 &= -\frac{1}{2} \boldsymbol{\delta}_S^\top \left\{ \nabla^2 l(\boldsymbol{\alpha}^{(0)} + \tau \boldsymbol{\delta}^{[S]}) \right\}_{S,S} \boldsymbol{\delta}_S \\
&\geq \frac{C_{\min}}{4} \left\| \boldsymbol{\delta}_S \right\|_2^2 \\
&= \frac{C_{\min}}{4} M^2 \frac{K_0 \log(p)}{n}.
\end{aligned} \tag{S5.7}$$

Applying the triangle inequality, we obtain

$$\begin{aligned}
|I_3| &= \left| \sum_{k \in S} \frac{\lambda}{|\bar{\alpha}_k|} \left(\left| \alpha_k^{(0)} + \delta_k^{[S]} \right| - \left| \alpha_k^{(0)} \right| \right) \right| \\
&\leq \sum_{k \in S} \frac{\lambda}{|\bar{\alpha}_k|} \left| |\alpha_k^{(0)} + \delta_k^{[S]}| - |\alpha_k^{(0)}| \right| \\
&\leq \frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}} \sum_{k \in S} \left| |\alpha_k^{(0)} + \delta_k^{[S]}| - |\alpha_k^{(0)}| \right| \\
&\leq \frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}} \sum_{k \in S} \left| \alpha_k^{(0)} + \delta_k^{[S]} - \alpha_k^{(0)} \right| \\
&= \frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}} \sum_{k \in S} \left| \delta_k^{[S]} \right| \\
&= \frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}} \left\| \boldsymbol{\delta}_S \right\|_1 \\
&\leq \frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}} \sqrt{K_0} \left\| \boldsymbol{\delta}_S \right\|_2 \\
&= \frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}} \sqrt{K_0} M \sqrt{\frac{K_0 \log(p)}{n}},
\end{aligned}$$

which implies

$$I_3 \geq -\frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}} \sqrt{K_0} M \sqrt{\frac{K_0 \log(p)}{n}}. \tag{S5.8}$$

Thus, using the lower bounds on I_1, I_2 and I_3 from (S5.6), (S5.7) and (S5.8), respectively, we obtain

$$G(\boldsymbol{\delta}_S) = I_1 + I_2 + I_3$$

$$
\begin{aligned}
&\geq -MK_0\frac{\log(p)}{n} + \frac{C_{\min}}{4}M^2K_0\frac{\log(p)}{n} \\
&\quad - \frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}}\sqrt{K_0}M\sqrt{\frac{K_0\log(p)}{n}} \\
&= M^2K_0\frac{\log(p)}{n}\left\{\frac{C_{\min}}{4} - \frac{1}{M} - \frac{1}{M}\frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}}\sqrt{\frac{n}{\log(p)}}\right\} \\
&> 0,
\end{aligned}
$$

where the last inequality holds due to the condition $M > 4/C_{\min}$ and

$$
\frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}}\sqrt{\frac{n}{\log(p)}} = o_{\mathrm{P}}(1). \tag{S5.9}
$$

To see why (S5.9) holds, we note all the conditions in Proposition 1 holds, such that we obtain for all $k \in S$,

$$
|\bar{\alpha}_k| = |\alpha_k^{(0)}| + o_{\mathrm{P}}(1),
$$

since $O_{\mathrm{P}}(\sqrt{K\log(p)/n}) = o_{\mathrm{P}}(1)$ due to the condition $K\sqrt{\log(p)/n} = o(1)$. Therefore, we have

$$
\min\{|\bar{\alpha}_k| : k \in S\} = \min\{|\alpha_k^{(0)}| : k \in S\} + o_{\mathrm{P}}(1).
$$

This, together with $\lambda\sqrt{n}/[\min\{|\alpha_k^{(0)}| : k \in S\}\sqrt{\log(p)}] \to 0$ in Condition 4, implies (S5.9).

Hence with probability tending to one, we obtain

$$\left\|\tilde{\boldsymbol{\delta}}_S\right\|_2 = \left\|\tilde{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S^{(0)}\right\|_2 = \left\|\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\right\|_2 \leq M\sqrt{\frac{K_0 \log(p)}{n}}.$$

□

Proof of Proposition 4. Since $\tilde{\boldsymbol{\alpha}}$ is the minimizer of criterion (S3.1), then we have $\tilde{\alpha}_k = 0$ for $k \in S^c$. The result in Proposition 3 holds since all the required conditions are satisfied. Therefore, by Proposition 3, we can apply inequality 4.56(b) in Seber (2008) to obtain, with probability tending to one,

$$\left\|\tilde{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S^{(0)}\right\|_\infty \leq \left\|\tilde{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S^{(0)}\right\|_2 \leq M\sqrt{\frac{K_0 \log(p)}{n}} \leq \frac{\min\{|\alpha_k^{(0)}| : k \in S\}}{2},$$

for a finite positive constant $M > 4/C_{\min}$, where the final inequality is due to

$$\min\{|\alpha_k^{(0)}| : k \in S\} \geq 2M\sqrt{\frac{K_0 \log(p)}{n}}. \tag{S5.10}$$

To see why (S5.10) holds, we first note Condition 4 implies

$$\frac{\min\{|\alpha_k^{(0)}| : k \in S\}}{\lambda\sqrt{\frac{n}{\log(p)}}} \to \infty,$$

and

$$\frac{\lambda n}{\sqrt{K_0 \log(p)}} = \frac{\lambda \sqrt{\frac{n}{\log(p)}}}{\sqrt{\frac{K_0 \log(p)}{n}}} \to \infty.$$

These lead to

$$\frac{\min\{|\alpha_k^{(0)}| : k \in S\}}{\lambda \sqrt{\frac{n}{\log(p)}}} \frac{\lambda \sqrt{\frac{n}{\log(p)}}}{\sqrt{\frac{K_0 \log(p)}{n}}} = \frac{\min\{|\alpha_k^{(0)}| : k \in S\}}{\sqrt{\frac{K_0 \log(p)}{n}}} \to \infty,$$

which implies (S5.10).

Finally, we can apply the triangle inequality and obtain for $k \in S$,

$$\begin{aligned}
|\tilde{\alpha}_k| &\geq \left|\alpha_k^{(0)}\right| - \left|\tilde{\alpha}_k - \alpha_k^{(0)}\right| \\
&\geq \min\{|\alpha_k^{(0)}| : k \in S\} - \left\|\tilde{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S^{(0)}\right\|_\infty \\
&\geq \min\{|\alpha_k^{(0)}| : k \in S\} - \frac{\min\{|\alpha_k^{(0)}| : k \in S\}}{2} \\
&= \frac{\min\{|\alpha_k^{(0)}| : k \in S\}}{2} > 0,
\end{aligned}$$

which completes the proof. □

Lemma 5. *Assume Condition 2 is satisfied. Then for any $\epsilon > 0$, it holds that*

(i) $\mathrm{P}(\|\boldsymbol{M}_{S^c,S}^n - \boldsymbol{M}_{S^c,S}^0\|_\infty \geq \epsilon) \leq 2\exp\left\{-\epsilon^2 n/(8C_W^4 K_0^2) + \log(K_0) + \log(K - K_0)\right\}$;

(ii) $\mathrm{P}(\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\|_\infty \geq \epsilon) \leq 2\exp\left\{-\epsilon^2 n/(8C_W^4 K_0^2) + 2\log(K_0)\right\}$;

(iii) $\mathrm{P}(\|\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\|_\infty \geq \epsilon) \leq 2\exp\left\{-\epsilon^2 n/(8C_W^4 K_0^2) + 2\log(K_0)\right\}$.

Proof of Lemma 5. The proof of this lemma follows along similar lines as that of Lemma 3. We first prove claim (i). First recall the (l_1, l_2)-th element of matrix $\boldsymbol{M}^n - \boldsymbol{M}^0$ is given by $\sum_{i=1}^n v_{l_1,l_2}^{(i)}/(np)$, where

$$v_{l_1,l_2}^{(i)} = (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)} - E\left\{(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right\},$$

for $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. It follows that $v_{l_1,l_2}^{(i)}$ has mean zero for $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Furthermore, under Condition 2, we have $|(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \boldsymbol{\mathcal{X}}_{l_2}^{(i)}| \leq pC_W^2$ for all $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Using similar techniques to those in the proof of Lemma 3, this implies $|v_{l_1,l_2}^{(i)}| \leq 2pC_W^2$ for all $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Thus, we can apply the Azuma-Hoeffding inequality to obtain

$$\begin{aligned}
\mathrm{P}\left(\left|\frac{1}{np}\sum_{i=1}^n v_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon}{K_0}\right) &= \mathrm{P}\left(\left|\sum_{i=1}^n v_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon np}{K_0}\right) \\
&= \mathrm{P}\left[\left|\sum_{i=1}^n \left\{v_{l_1,l_2}^{(i)} - E\left(v_{l_1,l_2}^{(i)}\right)\right\}\right| \geq \frac{\epsilon np}{K_0}\right] \\
&\leq 2\exp\left\{\frac{\frac{-2\epsilon^2 n^2 p^2}{K_0^2}}{\sum_{i=1}^n \left(2pC_W^2 + 2pC_W^2\right)^2}\right\} \\
&= 2\exp\left(\frac{-2\epsilon^2 n^2 p^2}{16nK_0^2 p^2 C_W^4}\right) \\
&= 2\exp\left(\frac{-\epsilon^2 n}{8C_W^4 K_0^2}\right),
\end{aligned}$$

for all $l_1, l_2 = 1, \cdots, K$ and $\epsilon > 0$. It follows that

$$\begin{aligned}
\mathrm{P}\left(\left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \geq \epsilon\right) &= \mathrm{P}\left\{\bigcup_{l_1\in S^c}\left(\sum_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n v^{(i)}_{l_1,l_2}\right| \geq \epsilon\right)\right\} \\
&\leq \sum_{l_1\in S^c}\mathrm{P}\left(\sum_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n v^{(i)}_{l_1,l_2}\right| \geq \epsilon\right) \\
&\leq \sum_{l_1\in S^c}\mathrm{P}\left(\bigcup_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n v^{(i)}_{l_1,l_2}\right| \geq \frac{\epsilon}{K_0}\right) \\
&\leq \sum_{l_1\in S^c}\sum_{l_2\in S}\mathrm{P}\left(\left|\frac{1}{np}\sum_{i=1}^n v^{(i)}_{l_1,l_2}\right| \geq \frac{\epsilon}{K_0}\right) \\
&\leq (K-K_0)(K_0)2\exp\left(\frac{-\epsilon^2 n}{8C_W^4 K_0^2}\right) \\
&= 2\exp\left\{\frac{-\epsilon^2 n}{8C_W^4 K_0^2} + \log(K_0) + \log(K-K_0)\right\},
\end{aligned}$$

for all $\epsilon > 0$, which proves claim (i).

Similarly, we obtain

$$\begin{aligned}
\mathrm{P}\left(\left\|\boldsymbol{M}^n_{S,S} - \boldsymbol{M}^0_{S,S}\right\|_\infty \geq \epsilon\right) &= \mathrm{P}\left\{\bigcup_{l_1\in S}\left(\sum_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n v^{(i)}_{l_1,l_2}\right| \geq \epsilon\right)\right\} \\
&\leq \sum_{l_1\in S}\mathrm{P}\left(\sum_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n v^{(i)}_{l_1,l_2}\right| \geq \epsilon\right) \\
&\leq \sum_{l_1\in S}\mathrm{P}\left(\bigcup_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n v^{(i)}_{l_1,l_2}\right| \geq \frac{\epsilon}{K_0}\right) \\
&\leq \sum_{l_1\in S}\sum_{l_2\in S}\mathrm{P}\left(\left|\frac{1}{np}\sum_{i=1}^n v^{(i)}_{l_1,l_2}\right| \geq \frac{\epsilon}{K_0}\right) \\
&\leq (K_0)^2 2\exp\left(\frac{-\epsilon^2 n}{8C_W^4 K_0^2}\right)
\end{aligned}$$

$$= 2\exp\left\{\frac{-\epsilon^2 n}{8C_W^4 K_0^2} + 2\log(K_0)\right\},$$

for all $\epsilon > 0$, which proves claim (ii).

Next, we prove claim (iii). Recall the (l_1, l_2)-th element of matrix $\boldsymbol{U}^n - \boldsymbol{U}^0$ is given by $\sum_{i=1}^n h_{l_1,l_2}^{(i)}/(np)$, where

$$h_{l_1,l_2}^{(i)} = (\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)} - E\left\{(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}\right\},$$

for $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. It follows that $h_{l_1,l_2}^{(i)}$ has mean zero for $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Furthermore, under Condition 2, we have $|(\boldsymbol{\mathcal{X}}_{l_1}^{(i)})^\top \boldsymbol{\mathcal{X}}_{l_2}^{(i)}| \leq pC_W^2$ for all $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Using similar techniques to those in the proof of Lemma 3, this implies $|h_{l_1,l_2}^{(i)}| \leq 2pC_W^2$ for all $i = 1, \cdots, n$ and $l_1, l_2 = 1, \cdots, K$. Therefore, we can apply the Azuma-Hoeffding inequality to obtain

$$\begin{aligned}
\mathrm{P}\left(\left|\frac{1}{np}\sum_{i=1}^n h_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon}{K_0}\right) &= \mathrm{P}\left(\left|\sum_{i=1}^n h_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon np}{K_0}\right) \\
&= \mathrm{P}\left[\left|\sum_{i=1}^n \left\{h_{l_1,l_2}^{(i)} - E\left(h_{l_1,l_2}^{(i)}\right)\right\}\right| \geq \frac{\epsilon np}{K_0}\right] \\
&\leq 2\exp\left\{\frac{\frac{-2\epsilon^2 n^2 p^2}{K_0^2}}{\sum_{i=1}^n \left(2pC_W^2 + 2pC_W^2\right)^2}\right\} \\
&= 2\exp\left(\frac{-2\epsilon^2 n^2 p^2}{16nK_0^2 p^2 C_W^4}\right)
\end{aligned}$$

$$= 2\exp\left(\frac{-\epsilon^2 n}{8C_W^4 K_0^2}\right),$$

for all $l_1, l_2 = 1, \cdots, K$ and $\epsilon > 0$. It follows that

$$\begin{aligned}
\mathrm{P}\left(\left\|\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\right\|_\infty \geq \epsilon\right) &= \mathrm{P}\left\{\bigcup_{l_1\in S}\left(\sum_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n h_{l_1,l_2}^{(i)}\right| \geq \epsilon\right)\right\} \\
&\leq \sum_{l_1\in S}\mathrm{P}\left(\sum_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n h_{l_1,l_2}^{(i)}\right| \geq \epsilon\right) \\
&\leq \sum_{l_1\in S}\mathrm{P}\left(\bigcup_{l_2\in S}\left|\frac{1}{np}\sum_{i=1}^n h_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon}{K_0}\right) \\
&\leq \sum_{l_1\in S}\sum_{l_2\in S}\mathrm{P}\left(\left|\frac{1}{np}\sum_{i=1}^n h_{l_1,l_2}^{(i)}\right| \geq \frac{\epsilon}{K_0}\right) \\
&\leq (K_0)^2 2\exp\left(\frac{-\epsilon^2 n}{8C_W^4 K_0^2}\right) \\
&= 2\exp\left\{\frac{-\epsilon^2 n}{8C_W^4 K_0^2} + 2\log(K_0)\right\},
\end{aligned}$$

for all $\epsilon > 0$, which proves claim (iii). □

Lemma 6. *Assume Conditions 1 and 2 are satisfied. Then for any $\epsilon > 0$, it holds that*

(i) $\mathrm{P}\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^n) \leq C_{\min} - \epsilon\} \leq 2\exp\{-\epsilon^2 n/(8C_W^4 K_0^2) + 2\log(K_0)\}$;

(ii) $\mathrm{P}\{\Lambda_{\max}(\boldsymbol{U}_{S,S}^n) \geq C_{\max} + \epsilon\} \leq 2\exp\{-\epsilon^2 n/(8C_W^4 K_0^2) + 2\log(K_0)\}$.

Proof of Lemma 6. The proof of this lemma follows along similar lines as that of Proposition 2. We first prove claim (i). By Condition 1 and inequality 4.67(e)

in Seber (2008), we have

$$\begin{aligned}\Lambda_{\min}(\boldsymbol{M}_{S,S}^n) &= \min_{\|\boldsymbol{x}\|_2=1} \left\{\boldsymbol{x}^\top \boldsymbol{M}_{S,S}^0 \boldsymbol{x} + \boldsymbol{x}^\top \left(\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right) \boldsymbol{x}\right\} \\ &\geq \Lambda_{\min}(\boldsymbol{M}_{S,S}^0) - \left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_2 \\ &\geq C_{\min} - \left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_\infty .\end{aligned}$$

This implies $\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^n) \leq C_{\min} - \epsilon\} \subseteq \{\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\|_\infty \geq \epsilon\}$ for all $\epsilon > 0$, due to

$$\begin{aligned}&C_{\min} - \left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_\infty \leq \Lambda_{\min}(\boldsymbol{M}_{S,S}^n) \leq C_{\min} - \epsilon \\ &\implies \left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_\infty \geq \epsilon.\end{aligned}$$

Therefore, by Lemma 5(ii), we obtain for all $\epsilon > 0$,

$$\begin{aligned}\mathrm{P}\left\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^n) \leq C_{\min} - \epsilon\right\} &\leq \mathrm{P}\left(\left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_\infty \geq \epsilon\right) \\ &\leq 2\exp\left\{-\frac{\epsilon^2 n}{8C_W^4 K_0^2} + 2\log(K_0)\right\},\end{aligned}$$

which proves claim (i).

Next, we prove claim (ii). By Condition 1 and inequality 4.67(e) in Seber

(2008), we have

$$\begin{aligned}\Lambda_{\max}(\boldsymbol{U}_{S,S}^n) &= \max_{\|\boldsymbol{x}\|_2=1} \left\{\boldsymbol{x}^\top \boldsymbol{U}_{S,S}^0 \boldsymbol{x} + \boldsymbol{x}^\top \left(\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\right) \boldsymbol{x}\right\} \\ &\leq \Lambda_{\max}(\boldsymbol{U}_{S,S}^0) + \left\|\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\right\|_2 \\ &\leq C_{\max} + \left\|\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\right\|_\infty .\end{aligned}$$

This implies $\{\Lambda_{\max}(\boldsymbol{U}_{S,S}^n) \geq C_{\max} + \epsilon\} \subseteq \{\|\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\|_\infty \geq \epsilon\}$ for all $\epsilon > 0$, due to

$$C_{\max} + \epsilon \leq \Lambda_{\max}(\boldsymbol{U}_{S,S}^n) \leq C_{\max} + \left\|\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\right\|_\infty \implies \left\|\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\right\|_\infty \geq \epsilon.$$

Therefore, by Lemma 5(iii), we obtain for all $\epsilon > 0$,

$$\begin{aligned}\mathrm{P}\left\{\Lambda_{\max}(\boldsymbol{U}_{S,S}^n) \geq C_{\max} + \epsilon\right\} &\leq \mathrm{P}\left(\left\|\boldsymbol{U}_{S,S}^n - \boldsymbol{U}_{S,S}^0\right\|_\infty \geq \epsilon\right) \\ &\leq 2\exp\left\{-\frac{\epsilon^2 n}{8C_W^4 K_0^2} + 2\log(K_0)\right\},\end{aligned}$$

which proves claim (ii).

□

Lemma 7. *Assume Conditions 1 – 2 are satisfied. Then for any $\bar{\epsilon} > 0$, it holds*

that

$$\mathrm{P}\left\{\left\|(\boldsymbol{M}_{S,S}^{n})^{-1}-(\boldsymbol{M}_{S,S}^{0})^{-1}\right\|_{\infty}\geq\bar{\epsilon}\right\}\leq 2\exp\left\{-\frac{C_{\min}^{2}n}{32C_{W}^{4}K_{0}^{2}}+2\log(K_{0})\right\}+2\exp\left\{-\frac{C_{\min}^{4}\bar{\epsilon}^{2}n}{32C_{W}^{4}K_{0}^{3}}+2\log(K_{0})\right\}.$$

Proof of Lemma 7. Note $(\boldsymbol{M}_{S,S}^{n})^{-1}-(\boldsymbol{M}_{S,S}^{0})^{-1}=(\boldsymbol{M}_{S,S}^{0})^{-1}(\boldsymbol{M}_{S,S}^{0}-\boldsymbol{M}_{S,S}^{n})(\boldsymbol{M}_{S,S}^{n})^{-1}$ and $\|(\boldsymbol{M}_{S,S}^{0})^{-1}\|_{2}=\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^{0})\}^{-1}$. By inequalities 4.57(d) and 4.57(e) in Seber (2008), we have

$$\begin{aligned}
&\left\|(\boldsymbol{M}_{S,S}^{n})^{-1}-(\boldsymbol{M}_{S,S}^{0})^{-1}\right\|_{\infty}\\
&\leq\sqrt{K_{0}}\left\|(\boldsymbol{M}_{S,S}^{n})^{-1}-(\boldsymbol{M}_{S,S}^{0})^{-1}\right\|_{2}\\
&=\sqrt{K_{0}}\left\|(\boldsymbol{M}_{S,S}^{0})^{-1}(\boldsymbol{M}_{S,S}^{0}-\boldsymbol{M}_{S,S}^{n})(\boldsymbol{M}_{S,S}^{n})^{-1}\right\|_{2}\\
&\leq\sqrt{K_{0}}\left\|(\boldsymbol{M}_{S,S}^{0})^{-1}\right\|_{2}\left\|\boldsymbol{M}_{S,S}^{0}-\boldsymbol{M}_{S,S}^{n}\right\|_{2}\left\|(\boldsymbol{M}_{S,S}^{n})^{-1}\right\|_{2}\\
&\leq\sqrt{K_{0}}\left\|(\boldsymbol{M}_{S,S}^{0})^{-1}\right\|_{2}\left\|\boldsymbol{M}_{S,S}^{0}-\boldsymbol{M}_{S,S}^{n}\right\|_{\infty}\left\|(\boldsymbol{M}_{S,S}^{n})^{-1}\right\|_{2}\\
&=\frac{\sqrt{K_{0}}}{\Lambda_{\min}(\boldsymbol{M}_{S,S}^{0})}\left\|\boldsymbol{M}_{S,S}^{0}-\boldsymbol{M}_{S,S}^{n}\right\|_{\infty}\left\|(\boldsymbol{M}_{S,S}^{n})^{-1}\right\|_{2}\\
&\leq\left\{\frac{\left\|(\boldsymbol{M}_{S,S}^{n})^{-1}\right\|_{2}}{C_{\min}}\right\}\left(\sqrt{K_{0}}\left\|\boldsymbol{M}_{S,S}^{0}-\boldsymbol{M}_{S,S}^{n}\right\|_{\infty}\right),
\end{aligned}$$

where the last inequality holds due to Condition 1. Furthermore, by noting $\left\|(\boldsymbol{M}_{S,S}^{n})^{-1}\right\|_{2}=\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^{n})\}^{-1}$, then we can set $\epsilon=C_{\min}/2$ in Lemma 6(i)

to obtain

$$\begin{aligned}
\mathrm{P}\left\{\frac{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2}{C_{\min}} \geq \frac{2}{C_{\min}^2}\right\} &= \mathrm{P}\left\{\frac{1}{C_{\min}\Lambda_{\min}(\boldsymbol{M}_{S,S}^n)} \geq \frac{2}{C_{\min}^2}\right\} \\
&= \mathrm{P}\left\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^n) \leq \frac{C_{\min}}{2}\right\} \\
&= \mathrm{P}\left\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^n) \leq C_{\min} - \frac{C_{\min}}{2}\right\} \\
&\leq 2\exp\left\{-\frac{\left(\frac{C_{\min}}{2}\right)^2 n}{8C_W^4 K_0^2} + 2\log(K_0)\right\} \\
&= 2\exp\left\{-\frac{C_{\min}^2 n}{32C_W^4 K_0^2} + 2\log(K_0)\right\}.
\end{aligned}$$

Also, by setting $\epsilon = C_{\min}^2\bar{\epsilon}/(2\sqrt{K_0})$ in Lemma 5(ii), we have

$$\begin{aligned}
\mathrm{P}\left(\sqrt{K_0}\left\|\boldsymbol{M}_{S,S}^0 - \boldsymbol{M}_{S,S}^n\right\|_\infty \geq \frac{C_{\min}^2\bar{\epsilon}}{2}\right) &= \mathrm{P}\left(\left\|\boldsymbol{M}_{S,S}^0 - \boldsymbol{M}_{S,S}^n\right\|_\infty \geq \frac{C_{\min}^2\bar{\epsilon}}{2\sqrt{K_0}}\right) \\
&\leq 2\exp\left\{-\frac{\left(\frac{C_{\min}^2\bar{\epsilon}}{2\sqrt{K_0}}\right)^2 n}{8C_W^4 K_0^2} + 2\log(K_0)\right\} \\
&= 2\exp\left\{-\frac{C_{\min}^4\bar{\epsilon}^2 n}{32C_W^4 K_0^3} + 2\log(K_0)\right\},
\end{aligned}$$

for all $\bar{\epsilon} > 0$.

Finally, we have

$$\left\{\left\|(\boldsymbol{M}_{S,S}^n)^{-1} - (\boldsymbol{M}_{S,S}^0)^{-1}\right\|_\infty \geq \bar{\epsilon}\right\}$$

$$\subseteq \left\{ \left\{ \frac{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2}{C_{\min}} \right\} \left(\sqrt{K_0} \left\| \boldsymbol{M}_{S,S}^0 - \boldsymbol{M}_{S,S}^n \right\|_\infty \right) \geq \left(\frac{2}{C_{\min}^2} \right) \left(\frac{C_{\min}^2 \bar{\epsilon}}{2} \right) \right\}$$
$$\subseteq \left\{ \frac{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2}{C_{\min}} \geq \frac{2}{C_{\min}^2} \right\} \bigcup \left\{ \sqrt{K_0} \left\| \boldsymbol{M}_{S,S}^0 - \boldsymbol{M}_{S,S}^n \right\|_\infty \geq \frac{C_{\min}^2 \bar{\epsilon}}{2} \right\},$$

for all $\bar{\epsilon} > 0$. Thus, by the union-sum inequality, we obtain

$$\begin{aligned}
&\mathrm{P}\left\{ \left\|(\boldsymbol{M}_{S,S}^n)^{-1} - (\boldsymbol{M}_{S,S}^0)^{-1}\right\|_\infty \geq \bar{\epsilon} \right\} \\
&\leq \mathrm{P}\left[\left\{ \frac{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2}{C_{\min}} \geq \frac{2}{C_{\min}^2} \right\} \bigcup \left\{ \sqrt{K_0} \left\| \boldsymbol{M}_{S,S}^0 - \boldsymbol{M}_{S,S}^n \right\|_\infty \geq \frac{C_{\min}^2 \bar{\epsilon}}{2} \right\} \right] \\
&\leq \mathrm{P}\left\{ \frac{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2}{C_{\min}} \geq \frac{2}{C_{\min}^2} \right\} + \mathrm{P}\left(\sqrt{K_0} \left\| \boldsymbol{M}_{S,S}^0 - \boldsymbol{M}_{S,S}^n \right\|_\infty \geq \frac{C_{\min}^2 \bar{\epsilon}}{2} \right) \\
&\leq 2\exp\left\{ -\frac{C_{\min}^2 n}{32 C_W^4 K_0^2} + 2\log(K_0) \right\} + 2\exp\left\{ -\frac{C_{\min}^4 \bar{\epsilon}^2 n}{32 C_W^4 K_0^3} + 2\log(K_0) \right\},
\end{aligned}$$

for all $\bar{\epsilon} > 0$, which completes the proof. □

Proof of Proposition 5. We first write $\boldsymbol{M}_{S^c,S}^n (\boldsymbol{M}_{S,S}^n)^{-1} = \boldsymbol{T}_1 + \boldsymbol{T}_2 + \boldsymbol{T}_3 + \boldsymbol{T}_4$, where

$$\boldsymbol{T}_1 = \boldsymbol{M}_{S^c,S}^0 \left\{ (\boldsymbol{M}_{S,S}^n)^{-1} - (\boldsymbol{M}_{S,S}^0)^{-1} \right\},$$

$$\boldsymbol{T}_2 = (\boldsymbol{M}_{S^c,S}^n - \boldsymbol{M}_{S^c,S}^0)(\boldsymbol{M}_{S,S}^0)^{-1},$$

$$\boldsymbol{T}_3 = (\boldsymbol{M}_{S^c,S}^n - \boldsymbol{M}_{S^c,S}^0) \left\{ (\boldsymbol{M}_{S,S}^n)^{-1} - (\boldsymbol{M}_{S,S}^0)^{-1} \right\}, \text{ and}$$

$$\boldsymbol{T}_4 = \boldsymbol{M}_{S^c,S}^0 (\boldsymbol{M}_{S,S}^0)^{-1}.$$

By Condition 3, we have $\|\boldsymbol{T}_4\|_\infty \leq 1 - C_M$, which implies

$$\begin{aligned}
&\left\{ \left\| \boldsymbol{M}^n_{S^c,S} (\boldsymbol{M}^n_{S,S})^{-1} \right\|_\infty \geq 1 - \frac{C_M}{2} \right\} \\
&\subseteq \left\{ \|\boldsymbol{T}_1\|_\infty \geq \frac{C_M}{6} \right\} \cup \left\{ \|\boldsymbol{T}_2\|_\infty \geq \frac{C_M}{6} \right\} \cup \left\{ \|\boldsymbol{T}_3\|_\infty \geq \frac{C_M}{6} \right\},
\end{aligned}$$

due to

$$\begin{aligned}
&1 - \frac{C_M}{2} \leq \left\| \boldsymbol{M}^n_{S^c,S} (\boldsymbol{M}^n_{S,S})^{-1} \right\|_\infty \leq \|\boldsymbol{T}_1\|_\infty + \|\boldsymbol{T}_2\|_\infty + \|\boldsymbol{T}_3\|_\infty + \|\boldsymbol{T}_4\|_\infty \\
&\leq \|\boldsymbol{T}_1\|_\infty + \|\boldsymbol{T}_2\|_\infty + \|\boldsymbol{T}_3\|_\infty + 1 - C_M \\
&\implies \|\boldsymbol{T}_1\|_\infty + \|\boldsymbol{T}_2\|_\infty + \|\boldsymbol{T}_3\|_\infty \geq \frac{C_M}{2} \\
&\implies \|\boldsymbol{T}_1\|_\infty \geq \frac{C_M}{6} \text{ or } \|\boldsymbol{T}_2\|_\infty \geq \frac{C_M}{6} \text{ or } \|\boldsymbol{T}_3\|_\infty \geq \frac{C_M}{6}.
\end{aligned}$$

Next, we proceed to obtain the tail bounds for $\boldsymbol{T}_1, \boldsymbol{T}_2$ and $\boldsymbol{T}_3$ in the following three steps.

Step 1. We start by rewriting $\boldsymbol{T}_1 = \boldsymbol{M}^0_{S^c,S} (\boldsymbol{M}^0_{S,S})^{-1} (\boldsymbol{M}^0_{S,S} - \boldsymbol{M}^n_{S,S}) (\boldsymbol{M}^n_{S,S})^{-1}$. By Condition 3 and inequality 4.67(d) in Seber (2008), we obtain

$$\begin{aligned}
\|\boldsymbol{T}_1\|_\infty &\leq \left\| \boldsymbol{M}^0_{S^c,S} (\boldsymbol{M}^0_{S,S})^{-1} \right\|_\infty \left\| \boldsymbol{M}^0_{S,S} - \boldsymbol{M}^n_{S,S} \right\|_\infty \left\| (\boldsymbol{M}^n_{S,S})^{-1} \right\|_\infty \\
&\leq \left\| \boldsymbol{M}^0_{S^c,S} (\boldsymbol{M}^0_{S,S})^{-1} \right\|_\infty \left\| \boldsymbol{M}^0_{S,S} - \boldsymbol{M}^n_{S,S} \right\|_\infty \sqrt{K_0} \left\| (\boldsymbol{M}^n_{S,S})^{-1} \right\|_2 \\
&\leq (1 - C_M) \left(\sqrt{K_0} \left\| \boldsymbol{M}^n_{S,S} - \boldsymbol{M}^0_{S,S} \right\|_\infty \right) \left\{ \left\| (\boldsymbol{M}^n_{S,S})^{-1} \right\|_2 \right\},
\end{aligned}$$

which implies

$$\begin{aligned}
&\left\{\|\boldsymbol{T}_1\|_\infty \geq \frac{C_M}{6}\right\} \\
&\subseteq \left\{\sqrt{K_0}\left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_\infty \geq \frac{C_{\min}C_M}{12(1-C_M)}\right\} \cup \left\{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2 \geq \frac{2}{C_{\min}}\right\}.
\end{aligned}$$

The above is due to

$$\begin{aligned}
&\frac{C_M}{6} \leq \|\boldsymbol{T}_1\|_\infty \leq (1-C_M)\left(\sqrt{K_0}\left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_\infty\right)\left\{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2\right\} \\
&\implies \left(\sqrt{K_0}\left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_\infty\right)\left\{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2\right\} \geq \frac{C_M}{6(1-C_M)} \\
&\qquad = \left\{\frac{C_{\min}C_M}{12(1-C_M)}\right\}\left(\frac{2}{C_{\min}}\right) \\
&\implies \sqrt{K_0}\left\|\boldsymbol{M}_{S,S}^n - \boldsymbol{M}_{S,S}^0\right\|_\infty \geq \frac{C_{\min}C_M}{12(1-C_M)} \text{ or } \left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2 \geq \frac{2}{C_{\min}}.
\end{aligned}$$

By setting $\epsilon = C_{\min}/2$ in Lemma 6(i), we have

$$\begin{aligned}
\mathrm{P}\left\{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2 \geq \frac{2}{C_{\min}}\right\} &= \mathrm{P}\left\{\frac{1}{\Lambda_{\min}(\boldsymbol{M}_{S,S}^n)} \geq \frac{2}{C_{\min}}\right\} \\
&= \mathrm{P}\left\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^n) \leq \frac{C_{\min}}{2}\right\} \\
&= \mathrm{P}\left\{\Lambda_{\min}(\boldsymbol{M}_{S,S}^n) \leq C_{\min} - \frac{C_{\min}}{2}\right\} \\
&\leq 2\exp\left\{-\frac{C_{\min}^2 n}{32C_W^4K_0^2} + 2\log(K_0)\right\}.
\end{aligned}$$

Furthermore, by setting $\epsilon = C_{\min}C_M/\{12(1-C_M)\sqrt{K_0}\}$ in Lemma 5(ii), we

obtain

$$\begin{aligned}
&\mathrm{P}\left\{\sqrt{K_0}\left\|\boldsymbol{M}_{S,S}^n-\boldsymbol{M}_{S,S}^0\right\|_\infty \geq \frac{C_{\min}C_M}{12(1-C_M)}\right\}\\
&=\mathrm{P}\left\{\left\|\boldsymbol{M}_{S,S}^n-\boldsymbol{M}_{S,S}^0\right\|_\infty \geq \frac{C_{\min}C_M}{12(1-C_M)\sqrt{K_0}}\right\}\\
&\leq 2\exp\left\{-\frac{C_{\min}^2C_M^2n}{1152(1-C_M)^2C_W^4K_0^3}+2\log(K_0)\right\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathrm{P}\left(\|\boldsymbol{T}_1\|_\infty \geq \frac{C_M}{6}\right)\\
&\leq \mathrm{P}\left\{\sqrt{K_0}\left\|\boldsymbol{M}_{S,S}^n-\boldsymbol{M}_{S,S}^0\right\|_\infty \geq \frac{C_{\min}C_M}{12(1-C_M)}\right\}+\mathrm{P}\left\{\left\|(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_2 \geq \frac{2}{C_{\min}}\right\}\\
&\leq 2\exp\left\{-\frac{C_{\min}^2C_M^2n}{1152(1-C_M)^2C_W^4K_0^3}+2\log(K_0)\right\}\\
&\quad+2\exp\left\{-\frac{C_{\min}^2n}{32C_W^4K_0^2}+2\log(K_0)\right\}.
\end{aligned}$$

Step 2. By Condition 1 and inequality 4.67(d) Seber (2008), we have

$$\begin{aligned}
\|\boldsymbol{T}_2\|_\infty &\leq \left\|\boldsymbol{M}_{S^c,S}^n-\boldsymbol{M}_{S^c,S}^0\right\|_\infty \left\|(\boldsymbol{M}_{S,S}^0)^{-1}\right\|_\infty\\
&\leq \left\|\boldsymbol{M}_{S^c,S}^n-\boldsymbol{M}_{S^c,S}^0\right\|_\infty \sqrt{K_0}\left\|(\boldsymbol{M}_{S,S}^0)^{-1}\right\|_2\\
&= \left\|\boldsymbol{M}_{S^c,S}^n-\boldsymbol{M}_{S^c,S}^0\right\|_\infty \sqrt{K_0}\frac{1}{\Lambda_{\min}(\boldsymbol{M}_{S,S}^0)}\\
&\leq \left\|\boldsymbol{M}_{S^c,S}^n-\boldsymbol{M}_{S^c,S}^0\right\|_\infty \frac{\sqrt{K_0}}{C_{\min}},
\end{aligned}$$

which implies

$$\left\{\|\boldsymbol{T}_2\|_\infty \geq \frac{C_M}{6}\right\} \subseteq \left\{\left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \geq \frac{C_M C_{\min}}{6\sqrt{K_0}}\right\}.$$

Note the above is due to

$$\begin{aligned}\frac{C_M}{6} \leq \|\boldsymbol{T}_2\|_\infty &\leq \frac{\sqrt{K_0}}{C_{\min}} \left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \\ \implies \left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty &\geq \frac{C_M C_{\min}}{6\sqrt{K_0}}.\end{aligned}$$

By setting $\epsilon = C_M C_{\min}/(6\sqrt{K_0})$ in Lemma 5(i), we obtain

$$\begin{aligned}\mathrm{P}\left(\|\boldsymbol{T}_2\|_\infty \geq \frac{C_M}{6}\right) &\leq \mathrm{P}\left(\left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \geq \frac{C_M C_{\min}}{6\sqrt{K_0}}\right) \\ &\leq 2\exp\left\{-\frac{C_M^2 C_{\min}^2 n}{288 C_W^4 K_0^3} + \log(K_0) + \log(K - K_0)\right\}.\end{aligned}$$

Step 3. We have

$$\|\boldsymbol{T}_3\|_\infty \leq \left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \left\|(\boldsymbol{M}^n_{S,S})^{-1} - (\boldsymbol{M}^0_{S,S})^{-1}\right\|_\infty,$$

which implies

$$\begin{aligned}&\left\{\|\boldsymbol{T}_3\|_\infty \geq \frac{C_M}{6}\right\} \\ &\subseteq \left\{\left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \geq \sqrt{\frac{C_M}{6}}\right\} \cup \left\{\left\|(\boldsymbol{M}^n_{S,S})^{-1} - (\boldsymbol{M}^0_{S,S})^{-1}\right\|_\infty \geq \sqrt{\frac{C_M}{6}}\right\},\end{aligned}$$

where the above is due to

$$\begin{aligned}
&\frac{C_M}{6} \le \|\boldsymbol{T}_3\|_\infty \le \left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \left\|(\boldsymbol{M}^n_{S,S})^{-1} - (\boldsymbol{M}^0_{S,S})^{-1}\right\|_\infty \\
&\implies \left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \left\|(\boldsymbol{M}^n_{S,S})^{-1} - (\boldsymbol{M}^0_{S,S})^{-1}\right\|_\infty \ge \frac{C_M}{6} = \sqrt{\frac{C_M}{6}}\sqrt{\frac{C_M}{6}} \\
&\implies \left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \ge \sqrt{\frac{C_M}{6}} \text{ or } \left\|(\boldsymbol{M}^n_{S,S})^{-1} - (\boldsymbol{M}^0_{S,S})^{-1}\right\|_\infty \ge \sqrt{\frac{C_M}{6}}.
\end{aligned}$$

Thus by setting $\epsilon = \sqrt{C_M/6}$ in Lemma 5(i), we obtain

$$\begin{aligned}
&\mathrm{P}\left(\left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \ge \sqrt{\frac{C_M}{6}}\right) \\
&\le 2\exp\left\{-\frac{C_M n}{48 C_W^4 K_0^2} + \log(K_0) + \log(K - K_0)\right\}.
\end{aligned}$$

Moreover, by setting $\bar{\epsilon} = \sqrt{C_M/6}$ in Lemma 7 we obtain

$$\begin{aligned}
&\mathrm{P}\left\{\left\|(\boldsymbol{M}^n_{S,S})^{-1} - (\boldsymbol{M}^0_{S,S})^{-1}\right\|_\infty \ge \sqrt{\frac{C_M}{6}}\right\} \\
&\le 2\exp\left\{-\frac{C_{\min}^2 n}{32 C_W^4 K_0^2} + 2\log(K_0)\right\} + 2\exp\left\{-\frac{C_{\min}^4 C_M n}{192 C_W^4 K_0^3} + 2\log(K_0)\right\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathrm{P}\left(\|\boldsymbol{T}_3\|_\infty \ge \frac{C_M}{6}\right) \\
&\le \mathrm{P}\left(\left\|\boldsymbol{M}^n_{S^c,S} - \boldsymbol{M}^0_{S^c,S}\right\|_\infty \ge \sqrt{\frac{C_M}{6}}\right) + \mathrm{P}\left\{\left\|(\boldsymbol{M}^n_{S,S})^{-1} - (\boldsymbol{M}^0_{S,S})^{-1}\right\|_\infty \ge \sqrt{\frac{C_M}{6}}\right\} \\
&\le 2\exp\left\{-\frac{C_M n}{48 C_W^4 K_0^2} + \log(K_0) + \log(K - K_0)\right\} + 2\exp\left\{-\frac{C_{\min}^2 n}{32 C_W^4 K_0^2} + 2\log(K_0)\right\}
\end{aligned}$$

$$+ 2\exp\left\{-\frac{C_{\min}^4 C_M n}{192 C_W^4 K_0^3} + 2\log(K_0)\right\}.$$

Note $\log(K_0) \leq \log(K)$, $\log(K - K_0) \leq \log(K)$ and $-n/K_0^2 \leq -n/K_0^3$. Then, by combining the tail bounds from Step 1 to Step 3, we obtain

$$\begin{aligned}
&\mathrm{P}\left\{\left\|\boldsymbol{M}_{S^c,S}^n(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_\infty \geq 1 - \frac{C_M}{2}\right\} \\
&\leq \mathrm{P}\left(\|\boldsymbol{T}_1\|_\infty \geq \frac{C_M}{6}\right) + \mathrm{P}\left(\|\boldsymbol{T}_2\|_\infty \geq \frac{C_M}{6}\right) + \mathrm{P}\left(\|\boldsymbol{T}_3\|_\infty \geq \frac{C_M}{6}\right) \\
&\leq 2\exp\left\{-\frac{C_{\min}^2 C_M^2}{1152(1-C_M)^2 C_W^4}\frac{n}{K_0^3} + 2\log(K)\right\} + 2\exp\left\{-\frac{C_{\min}^2}{32 C_W^4}\frac{n}{K_0^3} + 2\log(K)\right\} \\
&\quad + 2\exp\left\{-\frac{C_M^2 C_{\min}^2}{288 C_W^4}\frac{n}{K_0^3} + 2\log(K)\right\} + 2\exp\left\{-\frac{C_M}{48 C_W^4}\frac{n}{K_0^3} + 2\log(K)\right\} \\
&\quad + 2\exp\left\{-\frac{C_{\min}^2}{32 C_W^4}\frac{n}{K_0^3} + 2\log(K)\right\} + 2\exp\left\{-\frac{C_{\min}^4 C_M}{192 C_W^4}\frac{n}{K_0^3} + 2\log(K)\right\} \\
&\leq 12\exp\left\{-C\frac{n}{K_0^3} + 2\log(K)\right\},
\end{aligned}$$

where

$$\begin{aligned}
C &= \min\left\{\frac{C_{\min}^2 C_M^2}{1152(1-C_M)^2 C_W^4}, \frac{C_{\min}^2}{32 C_W^4}, \frac{C_{\min}^2 C_M^2}{288 C_W^4}, \frac{C_M}{48 C_W^4}, \frac{C_{\min}^4 C_M}{192 C_W^4}\right\} \\
&= \min\left\{\frac{C_{\min}^2 C_M^2}{1152(1-C_M)^2 C_W^4}, \frac{C_{\min}^2 C_M^2}{288 C_W^4}, \frac{C_M}{48 C_W^4}, \frac{C_{\min}^4 C_M}{192 C_W^4}\right\}.
\end{aligned}$$

This completes the proof. □

Lemma 8. *Assume sample Condition 1′ and Condition 2 are satisfied. Then for*

any $\iota \in [0,1]$*, it holds that*

$$\left\| \left\{ \nabla^2 l(\boldsymbol{\alpha}^{(0)} + \iota\tilde{\boldsymbol{\delta}}) - \nabla^2 l(\boldsymbol{\alpha}^{(0)}) \right\} \tilde{\boldsymbol{\delta}} \right\|_\infty \leq C_W C_{\max} \left\| \tilde{\boldsymbol{\delta}}_S \right\|_2^2 .$$

Proof of Lemma 8. The proof of this lemma follows along similar lines as those of Lemmas 2 and 4. Applying the mean value theorem, we have $\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}+\iota\tilde{\boldsymbol{\delta}}) = \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}) + \iota\nabla\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)} + \iota^*\tilde{\boldsymbol{\delta}})^\top\tilde{\boldsymbol{\delta}}$, for some constant $\iota^* \in (0,\iota)$, $i = 1,\cdots,n$ and $j = 1,\cdots,p$. Then under sample Condition 1′ and Condition 2, we obtain for all $k = 1,\cdots,K$,

$$\begin{aligned}
&\left| \left[\left\{ \nabla^2 l(\boldsymbol{\alpha}^{(0)} + \iota\tilde{\boldsymbol{\delta}}) \right\}_{k,S} - \left\{ \nabla^2 l(\boldsymbol{\alpha}^{(0)}) \right\}_{k,S} \right] \tilde{\boldsymbol{\delta}}_S \right| \\
&= \left| -\frac{1}{np} \left[\sum_{i=1}^n \left(\boldsymbol{\mathcal{X}}_k^{(i)} \right)^\top \left\{ \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)} + \iota\tilde{\boldsymbol{\delta}}) - \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \right\} \boldsymbol{\mathcal{X}}_S^{(i)} \right] \tilde{\boldsymbol{\delta}}_S \right| \\
&= \left| \frac{1}{np} \left[\sum_{i=1}^n \left(\boldsymbol{\mathcal{X}}_k^{(i)} \right)^\top \left\{ \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)} + \iota\tilde{\boldsymbol{\delta}}) - \boldsymbol{\eta}^{(i)}(\boldsymbol{\alpha}^{(0)}) \right\} \boldsymbol{\mathcal{X}}_S^{(i)} \right] \tilde{\boldsymbol{\delta}}_S \right| \\
&= \left| \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \mathcal{X}_k^{(i,j)} \left\{ \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)} + \iota\tilde{\boldsymbol{\delta}}) - \eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)}) \right\} \boldsymbol{\mathcal{X}}_S^{(i,j)\top} \tilde{\boldsymbol{\delta}}_S \right| \\
&= \left| \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \mathcal{X}_k^{(i,j)} \left\{ \iota\nabla\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)} + \iota^*\tilde{\boldsymbol{\delta}})^\top\tilde{\boldsymbol{\delta}} \right\} \boldsymbol{\mathcal{X}}_S^{(i,j)\top} \tilde{\boldsymbol{\delta}}_S \right| \\
&\leq \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \left| \mathcal{X}_k^{(i,j)} \right| |\iota| \left| \nabla\eta_{jj}^{(i)}(\boldsymbol{\alpha}^{(0)} + \iota^*\tilde{\boldsymbol{\delta}})^\top\tilde{\boldsymbol{\delta}} \right| \left| \boldsymbol{\mathcal{X}}_S^{(i,j)\top} \tilde{\boldsymbol{\delta}}_S \right| \\
&\leq \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \left| \boldsymbol{W}_{j\cdot}^{(k)\top} \boldsymbol{y}_i \right| \left| \varepsilon_j^{(i)}(\boldsymbol{\alpha}^{(0)} + \iota^*\tilde{\boldsymbol{\delta}}) \left(\boldsymbol{\mathcal{X}}^{(i,j)\top}\tilde{\boldsymbol{\delta}} \right) \right| \left| \boldsymbol{\mathcal{X}}_S^{(i,j)\top} \tilde{\boldsymbol{\delta}}_S \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} C_W\left(\boldsymbol{\mathcal{X}}_S^{(i,j)\top}\tilde{\boldsymbol{\delta}}_S\right)^2 \\
&= C_W\left\{\frac{1}{np}\sum_{i=1}^{n}\left(\boldsymbol{\mathcal{X}}_S^{(i)}\tilde{\boldsymbol{\delta}}_S\right)^\top\left(\boldsymbol{\mathcal{X}}_S^{(i)}\tilde{\boldsymbol{\delta}}_S\right)\right\} \\
&= C_W\tilde{\boldsymbol{\delta}}_S^\top\left(\frac{1}{np}\sum_{i=1}^{n}\boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\mathcal{X}}^{(i)}\right)_{S,S}\tilde{\boldsymbol{\delta}}_S \\
&\leq C_W\Lambda_{\max}(\boldsymbol{U}_{S,S}^n)\left\|\tilde{\boldsymbol{\delta}}_S\right\|_2^2 \\
&\leq C_W C_{\max}\left\|\tilde{\boldsymbol{\delta}}_S\right\|_2^2,
\end{aligned}
$$

where $\mathcal{X}_k^{(i,j)}$ is the k-th element of the vector $\boldsymbol{\mathcal{X}}^{(i,j)}$ for $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Since $\|\{\nabla^2 l(\boldsymbol{\alpha}^{(0)}+\iota\tilde{\boldsymbol{\delta}})-\nabla^2 l(\boldsymbol{\alpha}^{(0)})\}\tilde{\boldsymbol{\delta}}\|_\infty = \max_{1\leq k\leq K}\{|[\{\nabla^2 l(\boldsymbol{\alpha}^{(0)}+\iota\tilde{\boldsymbol{\delta}})\}_{k,S} - \{\nabla^2 l(\boldsymbol{\alpha}^{(0)})\}_{k,S}]\tilde{\boldsymbol{\delta}}_S|\}$, this completes the proof of the lemma. □

Proof of Proposition 7. Since all conditions in Proposition 4 are satisfied, then we obtain with probability tending to one that $\tilde{\alpha}_k \neq 0$ for all $k \in S$. Furthermore, recall $\tilde{\boldsymbol{\alpha}}$ is the minimizer of criterion (S3.1). Hence, with probability tending to one, $\partial l(\tilde{\boldsymbol{\alpha}})/\partial\alpha_k = (\lambda/|\bar{\alpha}_k|)\text{sign}(\tilde{\alpha}_k)$ for all $k \in S$. This proves claim (i).

Before we prove claim (ii), we first prove for all $k \in S$,

$$
\frac{\lambda}{|\bar{\alpha}_k|} < \frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}. \tag{S5.11}
$$

Since all the conditions of Proposition 1 are satisfied, we obtain for all $k \in$

S^c,

$$|\bar{\alpha}_k| = |\bar{\alpha}_k - \alpha_k^{(0)}| + |\alpha_k^{(0)}| = O_{\mathrm{P}}\left(\sqrt{\frac{K\log(p)}{n}}\right),$$

and thus $\max\{|\bar{\alpha}_k| : k \in S^c\} = O_{\mathrm{P}}(\sqrt{K\log(p)/n})$. This, together with $\lambda n/\{\sqrt{K}\log(p)\} \to \infty$ in Condition 4, implies

$$\frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}\sqrt{\frac{n}{\log(p)}} \to \infty. \tag{S5.12}$$

Recalling (S5.9) gives

$$\frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}}\sqrt{\frac{n}{\log(p)}} = o(1),$$

under the assumed conditions, this leads to

$$\frac{\lambda}{\min\{|\bar{\alpha}_k| : k \in S\}} < \frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}, \tag{S5.13}$$

which implies (S5.11).

Then, we let $\boldsymbol{u} = (\max\{|\bar{\alpha}_k| : k \in S^c\}/\lambda)\nabla l(\tilde{\boldsymbol{\alpha}})$. From (S5.11), we have for all $k \in S$,

$$\frac{\max\{|\bar{\alpha}_k| : k \in S^c\}}{\lambda}\left|\frac{\partial l(\tilde{\boldsymbol{\alpha}})}{\partial \alpha_k}\right| < \frac{|\bar{\alpha}_k|}{\lambda}\left|\frac{\partial l(\tilde{\boldsymbol{\alpha}})}{\partial \alpha_k}\right|.$$

Hence, by (i), we obtain, with probability tending to one,

$$\begin{aligned}\|\boldsymbol{u}_S\|_\infty &= \max_{k\in S}\left\{\frac{\max\{|\bar{\alpha}_k| : k \in S^c\}}{\lambda}\left|\frac{\partial l(\tilde{\boldsymbol{\alpha}})}{\partial \alpha_k}\right|\right\} \\ &< \max_{k\in S}\left\{\frac{|\bar{\alpha}_k|}{\lambda}\left|\frac{\partial l(\tilde{\boldsymbol{\alpha}})}{\partial \alpha_k}\right|\right\} \\ &= \max_{k\in S}\{|\text{sign}(\tilde{\alpha}_k)|\} \\ &= 1.\end{aligned}$$

Then, by the mean value theorem and recalling $\tilde{\boldsymbol{\delta}} = \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}$, we obtain for some constant $\iota \in (0,1)$,

$$\begin{aligned}&\frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}\boldsymbol{u} - \nabla l(\boldsymbol{\alpha}^{(0)}) \\ &= \nabla l(\tilde{\boldsymbol{\alpha}}) - \nabla l(\boldsymbol{\alpha}^{(0)}) \\ &= \nabla^2 l(\boldsymbol{\alpha}^{(0)} + \iota\tilde{\boldsymbol{\delta}})\tilde{\boldsymbol{\delta}} \\ &= \nabla^2 l(\boldsymbol{\alpha}^{(0)} + \iota\tilde{\boldsymbol{\delta}})\tilde{\boldsymbol{\delta}} + \nabla^2 l(\boldsymbol{\alpha}^{(0)})\tilde{\boldsymbol{\delta}} - \nabla^2 l(\boldsymbol{\alpha}^{(0)})\tilde{\boldsymbol{\delta}} \\ &= -\boldsymbol{M}^n\tilde{\boldsymbol{\delta}} + \boldsymbol{r}^n,\end{aligned}$$

where $\boldsymbol{r}^n = \{\nabla^2 l(\boldsymbol{\alpha}^{(0)} + \iota\tilde{\boldsymbol{\delta}}) - \nabla^2 l(\boldsymbol{\alpha}^{(0)})\}\tilde{\boldsymbol{\delta}}$. Rearranging the equation gives

$$\boldsymbol{M}^n\tilde{\boldsymbol{\delta}} = -\frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}\boldsymbol{u} + \nabla l(\boldsymbol{\alpha}^{(0)}) + \boldsymbol{r}^n,$$

which can be written in block matrix form as

$$\begin{pmatrix} \boldsymbol{M}^n_{S,S} & \boldsymbol{M}^n_{S,S^c} \\ \boldsymbol{M}^n_{S^c,S} & \boldsymbol{M}^n_{S^c,S^c} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\delta}}_S \\ \tilde{\boldsymbol{\delta}}_{S^c} \end{pmatrix} = \begin{pmatrix} -\frac{\lambda}{\max\{|\bar{\alpha}_k|:k\in S^c\}}\boldsymbol{u}_S \\ -\frac{\lambda}{\max\{|\bar{\alpha}_k|:k\in S^c\}}\boldsymbol{u}_{S^c} \end{pmatrix} + \begin{pmatrix} \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S \\ \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_{S^c} \end{pmatrix} + \begin{pmatrix} \boldsymbol{r}^n_S \\ \boldsymbol{r}^n_{S^c} \end{pmatrix}.$$

Since $\tilde{\boldsymbol{\delta}}_{S^c} = \tilde{\boldsymbol{\alpha}}_{S^c} - \boldsymbol{\alpha}^{(0)}_{S^c} = \mathbf{0}_{K-K_0}$, then we obtain

$$\boldsymbol{M}^n_{S,S}\tilde{\boldsymbol{\delta}}_S = -\frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}\boldsymbol{u}_S + \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S + \boldsymbol{r}^n_S, \tag{S5.14}$$

and

$$\boldsymbol{M}^n_{S^c,S}\tilde{\boldsymbol{\delta}}_S = -\frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}\boldsymbol{u}_{S^c} + \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_{S^c} + \boldsymbol{r}^n_{S^c}. \tag{S5.15}$$

Sample Condition 1′ implies $\boldsymbol{M}^n_{S,S}$ is invertible, so we can substitute (S5.14) into (S5.15) and obtain

$$\begin{aligned} &\boldsymbol{M}^n_{S^c,S}(\boldsymbol{M}^n_{S,S})^{-1}\left[-\frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}\boldsymbol{u}_S + \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S + \boldsymbol{r}^n_S\right] \\ &\quad = -\frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}\boldsymbol{u}_{S^c} + \left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_{S^c} + \boldsymbol{r}^n_{S^c}. \end{aligned}$$

Rearranging the equation gives

$$\begin{aligned} \boldsymbol{u}_{S^c} = &\frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_{S^c}}{\lambda} + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\boldsymbol{r}^n_{S^c}}{\lambda} \\ &- \boldsymbol{M}^n_{S^c,S}(\boldsymbol{M}^n_{S,S})^{-1}\left[-\boldsymbol{u}_S + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_S}{\lambda}\right. \end{aligned}$$

$$+\frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\boldsymbol{r}_S^n}{\lambda}\Bigg].$$

Using $\|\boldsymbol{u}_S\|_\infty < 1$ from above and sample Condition 3′, we obtain

$$\begin{aligned}
&\|\boldsymbol{u}_{S^c}\|_\infty \\
&\leq \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\left\|\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_{S^c}\right\|_\infty}{\lambda} + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\|\boldsymbol{r}_{S^c}^n\|_\infty}{\lambda} \\
&\quad + \left\|\boldsymbol{M}_{S^c,S}^n(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_\infty \Bigg[\|\boldsymbol{u}_S\|_\infty + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\left\|\left\{\nabla l(\boldsymbol{\alpha}^{(0)})\right\}_{S}\right\|_\infty}{\lambda} \\
&\quad + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\|\boldsymbol{r}_S^n\|_\infty}{\lambda}\Bigg] \\
&< \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty}{\lambda} + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\|\boldsymbol{r}^n\|_\infty}{\lambda} \\
&\quad + \left\|\boldsymbol{M}_{S^c,S}^n(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_\infty \Bigg\{1 + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty}{\lambda} \\
&\quad + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\|\boldsymbol{r}^n\|_\infty}{\lambda}\Bigg\} \\
&= \left\|\boldsymbol{M}_{S^c,S}^n(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_\infty + \left\{1 + \left\|\boldsymbol{M}_{S^c,S}^n(\boldsymbol{M}_{S,S}^n)^{-1}\right\|_\infty\right\} \\
&\quad \times \left\{\frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty}{\lambda} + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\|\boldsymbol{r}^n\|_\infty}{\lambda}\right\} \\
&\leq 1 - C'_M + (2 - C'_M) \\
&\quad \times \left\{\frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty}{\lambda} + \frac{\max\{|\bar{\alpha}_k| : k \in S^c\}\|\boldsymbol{r}^n\|_\infty}{\lambda}\right\},
\end{aligned}$$

for a constant $C'_M \in (0, 1)$.

Since C_∇ satisfies $K = o(p^{C_\nabla^2/(8C_W^2)})$, then we apply Lemma 1(ii) and obtain, with probability tending to one,

$$\frac{\max\{|\bar{\alpha}_k| : k \in S^c\} \left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty}{\lambda} \leq \frac{\max\{|\bar{\alpha}_k| : k \in S^c\} C_\nabla \sqrt{\frac{\log(p)}{n}}}{\lambda} = C_\nabla \frac{\max\{|\bar{\alpha}_k| : k \in S^c\} \sqrt{\frac{\log(p)}{n}}}{\lambda},$$

where $(\max\{|\bar{\alpha}_k| : k \in S^c\}/\lambda)\sqrt{\log(p)/n} = o(1)$ due to (S5.12). Therefore, when n and p are sufficiently large, we obtain

$$\frac{\max\{|\bar{\alpha}_k| : k \in S^c\} \left\|\nabla l(\boldsymbol{\alpha}^{(0)})\right\|_\infty}{\lambda} \leq \frac{C'_M}{8 - 4C'_M},$$

for a constant $C'_M \in (0, 1)$. By Lemma 8 and Proposition 3, we have

$$\begin{aligned}\frac{\max\{|\bar{\alpha}_k| : k \in S^c\} \left\|\boldsymbol{r}^n\right\|_\infty}{\lambda} &\leq \frac{\max\{|\bar{\alpha}_k| : k \in S^c\} C_W C_{\max} \left\|\tilde{\tilde{\boldsymbol{\delta}}}_S\right\|_2^2}{\lambda} \\ &\leq \frac{\max\{|\bar{\alpha}_k| : k \in S^c\} C_W C_{\max} M^2 K_0 \frac{\log(p)}{n}}{\lambda} \\ &= C_W C_{\max} M^2 \frac{\max\{|\bar{\alpha}_k| : k \in S^c\} \sqrt{\frac{\log(p)}{n}}}{\lambda} K_0 \sqrt{\frac{\log(p)}{n}},\end{aligned}$$

for a finite positive constant $M > 4/C_{\min}$. Since $(\max\{|\bar{\alpha}_k| : k \in S^c\}/\lambda)\sqrt{\log(p)/n} = o(1)$, and $K_0\sqrt{\log(p)/n} = o(1)$, due to the condition $K\sqrt{\log(p)/n} = o(1)$ as

$n, p \to \infty$, then for sufficiently large n and p we have

$$\frac{\max\{|\bar{\alpha}_k| : k \in S^c\} \left\|\boldsymbol{r}^n\right\|_\infty}{\lambda} \leq \frac{C'_M}{8 - 4C'_M},$$

for a constant $C'_M \in (0, 1)$. Thus, we obtain

$$\begin{aligned}
\left\|\boldsymbol{u}_{S^c}\right\|_\infty &< 1 - C'_M + (2 - C'_M)\left(\frac{C'_M}{8 - 4C'_M} + \frac{C'_M}{8 - 4C'_M}\right) \\
&= 1 - C'_M + (2 - C'_M)\left\{\frac{2C'_M}{4(2 - C'_M)}\right\} \\
&= 1 - C'_M + \frac{1}{2}C'_M \\
&= 1 - \frac{1}{2}C'_M \\
&< 1,
\end{aligned}$$

since $C'_M \in (0, 1)$. Finally, by recalling $\boldsymbol{u} = (\max\{|\bar{\alpha}_k| : k \in S^c\}/\lambda)\nabla l(\tilde{\boldsymbol{\alpha}})$, we obtain for all $k \in S^c$,

$$\begin{aligned}
\left|\frac{\partial l(\tilde{\boldsymbol{\alpha}})}{\partial \alpha_k}\right| &\leq \left\|\{\nabla l(\tilde{\boldsymbol{\alpha}})\}_{S^c}\right\|_\infty \\
&= \frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}} \left\|\boldsymbol{u}_{S^c}\right\|_\infty \\
&< \frac{\lambda}{\max\{|\bar{\alpha}_k| : k \in S^c\}}(1) \\
&\leq \frac{\lambda}{|\bar{\alpha}_k|},
\end{aligned}$$

which proves claim (ii). □

S6 Inference Method

In this section, we provide more details on the inference method based on the proposed regularized pseudo-likelihood estimator employed in the real data applications of Sections 5 and S9. We first discuss the calculation of the empirical sandwich covariance matrix, which is used to construct 95% Wald confidence interval for each of the regression coefficients that are estimated to be non-zero using regularized pseudo-likelihood estimation. Specifically, let $\hat{\boldsymbol{\vartheta}} = (\hat{\theta}_{11}, \cdots, \hat{\theta}_{pp}, \hat{\boldsymbol{\alpha}}^\top)^\top$ denote the regularized estimates of the Ising similarity regression model, $\hat{S} = \{k : \hat{\alpha}_k \neq 0, \text{ for } k = 1, \cdots, K\}$ denote the estimated index set of the non-zero regression coefficients, and

$$\tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}}) = \sum_{j=1}^{p} \left[y_{ij} \left(\theta_{jj} + \sum_{k \in \hat{S}} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) - \log \left\{ 1 + \exp \left(\theta_{jj} + \sum_{k \in \hat{S}} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'} \right) \right\} \right],$$

denote the log pseudo-likelihood function of the i-th observation for $i = 1, \cdots, n$ after model selection, where $\boldsymbol{\vartheta}_{\hat{S}} = (\theta_{11}, \cdots, \theta_{pp}, \boldsymbol{\alpha}_{\hat{S}}^\top)^\top$ and $\boldsymbol{\alpha}_{\hat{S}}$ is a subvector of

$\boldsymbol{\alpha}$ that consists of the elements indexed by $\hat{S}$. The empirical sandwich covariance matrix is then given as

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\vartheta}}_{\hat{S}}) = \left\{ -\sum_{i=1}^{n} \nabla^2 \tilde{l}_i(\hat{\boldsymbol{\vartheta}}_{\hat{S}}) \right\}^{-1} \left[\sum_{i=1}^{n} \nabla \tilde{l}_i(\hat{\boldsymbol{\vartheta}}_{\hat{S}}) \left\{ \nabla \tilde{l}_i(\hat{\boldsymbol{\vartheta}}_{\hat{S}}) \right\}^{\top} \right] \left\{ -\sum_{i=1}^{n} \nabla^2 \tilde{l}_i(\hat{\boldsymbol{\vartheta}}_{\hat{S}}) \right\}^{-1},$$

where $\nabla \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})$ and $\nabla^2 \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})$ denote the gradient vector and the Hessian matrix, respectively, of $\tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})$ for $i = 1, \cdots, n$. In particular, the elements of $\nabla \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})$ for $i = 1, \cdots, n$ are given as

$$\frac{\partial \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \theta_{jj}} = y_{ij} - \frac{\exp\left(\theta_{jj} + \sum_{k \in \hat{S}} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'}\right)}{1 + \exp\left(\theta_{jj} + \sum_{k \in \hat{S}} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'}\right)},$$

for $j = 1, \cdots, p$, and

$$\frac{\partial \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \alpha_k} = (\boldsymbol{W}_k \boldsymbol{y}_i)^{\top} \left(\frac{\partial \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \theta_{11}}, \cdots, \frac{\partial \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \theta_{pp}} \right)^{\top},$$

for $k \in \hat{S}$. Furthermore, the elements of $\nabla^2 \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})$ for $i = 1, \cdots, n$ are given by

$$\frac{\partial^2 \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \theta_{jj}^2} = \frac{-\exp\left(\theta_{jj} + \sum_{k \in \hat{S}} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'}\right)}{\left\{1 + \exp\left(\theta_{jj} + \sum_{k \in \hat{S}} \alpha_k \sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'}\right)\right\}^2},$$

for $j = 1, \cdots, p$,

$$\frac{\partial^2 \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \theta_{j'j'} \partial \theta_{jj}} = 0,$$

for $j' \neq j$,

$$\frac{\partial^2 \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \alpha_k \partial \theta_{jj}} = \frac{\partial^2 \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \theta_{jj} \partial \alpha_k} = \frac{-\left(\sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'}\right) \exp\left(\theta_{jj} + \sum_{\ell \in \hat{S}} \alpha_\ell \sum_{j' \neq j} w_{jj'}^{(\ell)} y_{ij'}\right)}{\left\{1 + \exp\left(\theta_{jj} + \sum_{\ell \in \hat{S}} \alpha_\ell \sum_{j' \neq j} w_{jj'}^{(\ell)} y_{ij'}\right)\right\}^2},$$

for $j = 1, \cdots, p$ and $k \in \hat{S}$, and

$$\frac{\partial^2 \tilde{l}_i(\boldsymbol{\vartheta}_{\hat{S}})}{\partial \alpha_k \partial \alpha_{k'}} = \sum_{j=1}^{p} \frac{-\left(\sum_{j' \neq j} w_{jj'}^{(k)} y_{ij'}\right)\left(\sum_{j' \neq j} w_{jj'}^{(k')} y_{ij'}\right) \exp\left(\theta_{jj} + \sum_{\ell \in \hat{S}} \alpha_\ell \sum_{j' \neq j} w_{jj'}^{(\ell)} y_{ij'}\right)}{\left\{1 + \exp\left(\theta_{jj} + \sum_{\ell \in \hat{S}} \alpha_\ell \sum_{j' \neq j} w_{jj'}^{(\ell)} y_{ij'}\right)\right\}^2},$$

for $k, k' \in \hat{S}$. After obtaining the empirical sandwich covariance matrix $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\vartheta}}_{\hat{S}})$, the Wald confidence intervals are then constructed using the general form $\hat{\alpha}_k \pm 1.96 \times \mathrm{SE}(\hat{\alpha}_k)$ for $k \in \hat{S}$, where $\mathrm{SE}(\hat{\alpha}_k)$ is obtained by taking the square root of the corresponding diagonal element from $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\vartheta}}_{\hat{S}})$. These intervals are constructed based on the theory that the coefficients standardized by their corresponding standard errors i.e., $(\hat{\alpha}_k - \alpha_k^{(0)})/\mathrm{SE}(\hat{\alpha}_k)$ are asymptotically standard normal, noting that asymptotic normality result is a difficult problem for Ising model in general, and much of the existing Ising model literature (e.g., Wainwright, Lafferty, and Ravikumar, 2006; Höfling and Tibshirani, 2009; Cheng et al., 2014; Guo et al., 2015) have not explicitly addressed this problem; see also Section 6 for a discussion on future investigation of the asymptotic normality. Furthermore, the post-selection inference problem, to which our inference method pertains, is challenging; see Berk et al. (2013) and Lee et al. (2016) for

related work, noting these are not easily generalizable to our case as they focused on linear regression setting with independent univariate continuous responses as opposed to our multivariate binary response vectors with dependence structure from the Ising model.

S7 Supplementary Details of Simulation Study

S7.1 Simulation Settings of Section 4

This section provide details of the simulation settings used to obtain the numerical results in Section 4 of the main text. For the true values of the main effect parameters $\{\theta_{jj}^{(0)};; j = 1, \cdots, p\}$, we follow Guo et al. (2010) and generate them independently and uniformly in the range $[-1, -0.5] \cup [0.5, 1]$. For the regression coefficients of the similarity matrices, we generate $\alpha_k^{(0)}$ independently and uniformly from $[-0.4, -0.3] \cup [0.3, 0.4]$ for $k = 1, \cdots, 5$ and set $\alpha_k^{(0)} = 0$ for $k = 6, \cdots, 20$, leading to $K_0 = 5$ truly non-zero coefficients out of a total of $K = 20$ coefficients. Turning to the similarity matrices, we follow the framework of Zou et al. (2017) and set the diagonal elements of all the similarity matrices $\{\boldsymbol{W}_k : k = 1, \cdots, K\}$ to zero, while the off-diagonal elements are simulated as $w_{jj'}^{(k)} = w_{j'j}^{(k)} = \exp(-d_{jj'}^{(k)2})$ with $d_{jj'}^{(k)} \overset{i.i.d.}{\sim} U(p^{-1/2}, p^{1/2})$ for $j, j' = 1, \cdots, p$. Based on the above settings, we then construct the true Ising

similarity regression model as in (2.3), and afterward simulate 1000 datasets consisting of n multivariate binary vectors of dimension p using the `R` package `IsingSampler` (Epskamp, 2020), for each combination of n and p.

For all four methods considered in Section 4 (Regularized, Lasso, Unregularized and Oracle), we evaluate the performance of their point estimates for the regression coefficients α_k and main effect parameters θ_{jj} using $\text{MSE}_\alpha = (1/1000)\sum_{l=1}^{1000}\sum_{k=1}^{K}(\hat{\alpha}_{k,\{l\}}-\alpha_k^{(0)})^2/K$ and $\text{MSE}_\theta = (1/1000)\sum_{l=1}^{1000}\sum_{j=1}^{p}(\hat{\theta}_{jj,\{l\}}-\theta_{jj}^{(0)})^2/p$, respectively, noting the MSE is scaled by the number of associated parameters, and $\hat{\alpha}_{k,\{l\}}$ and $\hat{\theta}_{jj,\{l\}}$ generically denote the estimated parameters from the l-th simulated dataset. Additionally, we assess model selection performance for the proposed estimators and lasso-regularized estimators of the regression coefficients based on $\text{TPR} = (1/1000)\sum_{l=1}^{1000}\sum_{k=1}^{5}\mathbf{1}(\hat{\alpha}_{k,\{l\}} \neq 0)/5$ and $\text{FPR} = (1/1000)\sum_{l=1}^{1000}\sum_{k=6}^{K}\mathbf{1}(\hat{\alpha}_{k,\{l\}} \neq 0)/(K-5)$, where $\mathbf{1}(\cdot)$ is the indicator function.

S7.2 Additional Comparison with Other Estimators

In this section, we perform simulation studies to compare the performance of our proposed regularized estimator to other estimators. The simulation settings are similar to those in Section 4 of the main text.

Recalling that the proposed regularized pseudo-likelihood estimator utilizes

a ten-fold cross-validation approach to select the tuning parameter λ in equation (2.6), we additionally consider two widely used methods, AIC and BIC, to select λ. We denote such regularized pseudo-likelihood estimators with λ chosen using AIC and BIC as the AIC and BIC estimators, respectively.

In more detail, the AIC and BIC estimators involve finding an optimal λ that minimizes the criteria $2\hat{K}_\lambda - 2\sum_{i=1}^{n}\sum_{j=1}^{n}\log\{f_j(y_{ij}|\boldsymbol{y}_{i\setminus j;\hat{\boldsymbol{\vartheta}}_\lambda})\}$ and $\log(n)\hat{K}_\lambda - 2\sum_{i=1}^{n}\sum_{j=1}^{n}\log\{f_j(y_{ij}|\boldsymbol{y}_{i\setminus j;\hat{\boldsymbol{\vartheta}}_\lambda})\}$, respectively, where $\hat{\boldsymbol{\vartheta}}_\lambda = (\hat{\theta}_{11,\lambda}, \cdots, \hat{\theta}_{pp,\lambda}, \hat{\boldsymbol{\alpha}}_\lambda^\top)^\top$ denote the estimated values for $\boldsymbol{\vartheta}$ based on minimizing (2.6) by setting the value of the tuning parameter to be λ, and $\hat{K}_\lambda$ denotes the number of non-zero estimated regression coefficients in $\hat{\boldsymbol{\alpha}}_\lambda$. The resulting AIC and BIC estimators are then set as $\hat{\boldsymbol{\vartheta}}_\lambda$ that correspond to their respective optimal λ.

Table S2 presents the TPR and FPR of the AIC and BIC estimators, along with the TPR and FPR of the proposed estimator and lasso-regularized estimator, noting the latter were previously reported in Section 4 but are also included here for ease of comparison. Similar to the proposed estimator, the TPR and FPR of both the AIC and BIC estimators tend to one and zero, respectively, when n and p increase. However, when p is small, the AIC estimator tends to suffer from relatively large FPR albeit with a slightly larger TPR, and such tradeoff between the TPR and FPR of the AIC estimator is expected due to its weaker penalization on model complexity leading to overfitting. On the other hand, the

BIC estimator with a stronger penalization on model complexity is found to have similar selection performance to our proposed regularized estimator.

We also compare the proposed estimator with traditional Ising model estimators that do not incorporate the additional information from similarity matrices $\boldsymbol{W}_k$ and directly estimate the Ising model interaction matrix $\boldsymbol{\Theta}$ (see e.g., Section 3 of Höfling and Tibshirani, 2009). Specifically, we consider an estimator (Ising-Lasso) that minimizes the objective function

$$-\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[y_{ij}(\theta_{jj}+\sum_{j'\neq j}\theta_{jj'}y_{ij'})-\log\left\{1+\exp(\theta_{jj}+\sum_{j'\neq j}\theta_{jj'}y_{ij'})\right\}\right]+\lambda\sum_{j<j'}|\theta_{jj'}|, \tag{S7.1}$$

subject to $\theta_{jj'}=\theta_{j'j}$ for $j\neq j'$, where $\lambda>0$ is a tuning parameter for the lasso penalty. This estimator is computed using the `R` package `glmnet` by properly setting up the model matrix in the above lasso-regularized logistic regression in an analogous way to how our proposed regularized estimator is computed, where the tuning parameter is also selected by a similar ten-fold cross-validation approach. To understand how the model matrix can be set up, we first note that equation (2.2) in the main text includes the traditional Ising model as a special case, by letting $\boldsymbol{W}_1=\boldsymbol{\Delta}_{12}+\boldsymbol{\Delta}_{21},\cdots,\boldsymbol{W}_{p-1}=\boldsymbol{\Delta}_{1p}+\boldsymbol{\Delta}_{p1},\boldsymbol{W}_p=\boldsymbol{\Delta}_{23}+\boldsymbol{\Delta}_{32},\cdots,\boldsymbol{W}_K=\boldsymbol{\Delta}_{(p-1)p}+\boldsymbol{\Delta}_{p(p-1)}$, $\alpha_1=\theta_{12},\cdots,\alpha_{p-1}=\theta_{1p},\alpha_p=$

Table S2: TPR and FPR for the Regularized, Lasso, AIC and BIC estimators of the regression coefficients in the simulation study involving the Ising similarity regression model.

Regularized								
	TPR				FPR			
$p\backslash n$	50	100	200	400	50	100	200	400
10	0.337	0.490	0.654	0.824	0.167	0.202	0.225	0.235
25	0.835	0.970	0.999	1	0.218	0.218	0.197	0.167
50	0.954	0.999	1	1	0.186	0.174	0.154	0.102
100	0.976	0.998	0.999	1	0.043	0.105	0.076	0.009
200	0.925	0.998	1	1	0.043	0.058	0.005	0
Lasso								
	TPR				FPR			
$p\backslash n$	50	100	200	400	50	100	200	400
10	0.307	0.500	0.711	0.888	0.124	0.185	0.237	0.290
25	0.859	0.980	1	1	0.298	0.356	0.392	0.395
50	0.982	0.999	1	1	0.403	0.412	0.426	0.428
100	0.999	1	1	1	0.294	0.327	0.405	0.429
200	0.996	1	1	1	0.254	0.354	0.392	0.433
AIC								
	TPR				FPR			
$p\backslash n$	50	100	200	400	50	100	200	400
10	0.604	0.669	0.774	0.884	0.451	0.435	0.438	0.429
25	0.927	0.990	1	1	0.425	0.400	0.391	0.371
50	0.962	0.999	1	1	0.377	0.390	0.369	0.222
100	0.969	0.997	0.999	1	0.098	0.268	0.157	0.014
200	0.904	0.998	1	1	0.095	0.126	0.009	0
BIC								
	TPR				FPR			
$p\backslash n$	50	100	200	400	50	100	200	400
10	0.388	0.459	0.591	0.769	0.214	0.177	0.151	0.145
25	0.856	0.967	0.998	1	0.237	0.183	0.133	0.084
50	0.958	0.999	1	1	0.201	0.151	0.102	0.067
100	0.969	0.997	0.999	1	0.050	0.090	0.066	0.008
200	0.904	0.998	1	1	0.047	0.062	0.005	0

$\theta_{23}, \cdots, \alpha_K = \theta_{(p-1)p}$, $K = p(p-1)/2$, and recalling $\boldsymbol{\Delta}_{jj'}$ is a $p \times p$ matrix with the (j, j')-th element being one and other elements being zeros for $j, j' = 1, \cdots, p$. These $\boldsymbol{W}_k$ matrices can then be substituted into the expression of $\boldsymbol{\mathcal{X}}$ given in Section 4 of the main text to obtain the required model matrix $(\boldsymbol{I}_p \otimes \mathbf{1}_n, \boldsymbol{\mathcal{X}})$. We also consider an unregularized variant (Ising-Unregularized) of the above estimator by minimizing (S7.1) with λ set to zero. Again, the Ising-Unregularized estimator is computed using the `R` package `glmnet`. As such, the resulting Ising-Unregularized estimator corresponds to the largest possible model selected by `glmnet`, which might still contain some zero estimated $\theta_{jj'}$ under some cases due to the large number (i.e., $p(p+1)/2$) of parameters needing to be estimated. We assess the performance of these two estimators, along with the Oracle, Regularized, Lasso and Unregularized estimators from Section 4 of the main text, in estimating the true Ising model interaction matrix $\boldsymbol{\Theta}^{(0)} = \sum_{j=1}^{p} \theta_{jj}^{(0)} \boldsymbol{\Delta}_{jj} + \sum_{k=1}^{K} \alpha_k^{(0)} \boldsymbol{W}_k$.

Table S3 reports the estimation error measured under the Frobenius norm (Frobenius-error) and the spectral norm (Spectral-error), computed as $(1/1000) \sum_{l=1}^{1000} \|\hat{\boldsymbol{\Theta}}_{\{l\}} - \boldsymbol{\Theta}^{(0)}\|_F$ and $(1/1000) \sum_{l=1}^{1000} \|\hat{\boldsymbol{\Theta}}_{\{l\}} - \boldsymbol{\Theta}^{(0)}\|_2$, respectively, where $\hat{\boldsymbol{\Theta}}_{\{l\}}$ generically denotes the estimated interaction matrix from the l-th simulated dataset. It is clear that the traditional Ising model estimators (Ising-Lasso and Ising-Unregularized) have the worst performance in recovering the interaction matrix,

Table S3: Frobenius-error and spectral-error for the Oracle, Regularized, Lasso, Unregularized, Ising-Lasso, and Ising-Unregularized estimators of the interaction matrix $\boldsymbol{\Theta}$ in the simulation study involving the Ising similarity regression model. Dashes indicate procedures that are not executed due to prohibitively time-intensity.

		Frobenius-error						Spectral-error					
p	n	Oracle	Regularized	Lasso	Unregularized	Ising-Lasso	Ising-Unregularized	Oracle	Regularized	Lasso	Unregularized	Ising-Lasso	Ising-Unregularized
10	50	2.771	3.223	2.861	5.561	3.365	11.447	1.650	1.926	1.729	3.323	2.070	7.668
	100	1.906	2.461	2.274	3.611	3.053	5.838	1.132	1.472	1.386	2.153	1.863	3.611
	200	1.344	1.860	1.729	2.477	2.682	3.704	0.799	1.115	1.057	1.466	1.618	2.244
	400	0.930	1.342	1.266	1.703	2.145	2.486	0.556	0.808	0.771	1.007	1.280	1.483
25	50	3.428	4.605	4.421	5.822	6.967	151.669	1.438	1.880	1.798	2.365	2.861	94.351
	100	2.367	3.139	3.144	3.947	6.638	25.223	0.981	1.277	1.271	1.595	2.692	13.378
	200	1.670	2.120	2.234	2.766	6.123	12.335	0.692	0.863	0.902	1.117	2.496	5.867
	400	1.165	1.442	1.580	1.936	5.182	7.644	0.481	0.588	0.639	0.786	2.138	3.416
50	50	5.060	6.390	6.471	7.433	13.483	263.305	2.433	2.792	2.788	2.987	5.015	93.879
	100	2.920	3.576	3.892	4.641	12.635	253.155	1.024	1.190	1.283	1.441	3.994	129.932
	200	1.991	2.392	2.701	3.198	12.184	47.021	0.675	0.775	0.859	0.977	3.830	21.351
	400	1.415	1.631	1.911	2.247	10.964	22.041	0.473	0.527	0.602	0.683	3.488	9.087
100	50	14.281	12.192	13.973	15.626	27.172	103.003	10.185	6.271	8.461	10.243	12.776	21.443
	100	6.181	6.689	6.797	7.600	23.061	329.099	3.764	3.320	3.102	3.920	7.706	85.681
	200	3.066	3.331	3.666	4.213	—	—	1.167	1.152	1.197	1.310	—	—
	400	2.006	2.110	2.462	2.829	—	—	0.623	0.649	0.707	0.727	—	—
200	50	19.079	22.523	21.985	20.426	42.164	76.856	11.736	13.123	13.964	11.753	13.187	12.780
	100	7.186	7.744	8.399	8.777	37.970	136.217	3.609	3.710	4.111	3.713	7.761	20.779
	200	3.948	4.060	4.603	5.241	—	—	1.153	1.164	1.238	1.254	—	—
	400	2.671	2.806	3.119	3.582	—	—	0.644	0.667	0.702	0.727	—	—

since these estimators do not incorporate the extra information from the similarity matrices $\boldsymbol{W}_k$. In addition, the estimation errors of the proposed estimator tend to be smaller than both the lasso-regularized estimator and the unregularized estimator, and are getting closer to the oracle estimator as n and p increase.

S7.3 Additional Simulation Results for Varying K

We conduct further simulation studies to investigate the effect of the number of similarity matrices K on the empirical performance of various estimators for the Ising similarity regression model. The simulation settings are largely similar to those in Section 4 of the main text, with the only exception being that we now additionally consider $K \in \{10, 40, 80, 200\}$, while holding $K_0 = 5$ as the number of truly non-zero regression coefficients. For this simulation study, we

only consider two combinations of sample size and number of binary responses: $(n, p) = (50, 25)$ and $(n, p) = (400, 50)$.

Table S4 shows the increasingly worse estimation performance for all three estimators of the main effect parameter (Regularized, Lasso and Unregularized) when the number of irrelevant similarity matrices increases. The same conclusion of increasingly worse estimation performance when K increases holds for the estimators of the regression coefficients, despite the seemingly decreasing pattern of the MSEs for the proposed estimator and lasso-regularized estimator of the regression coefficients. Note however that the latter is due primarily to the definition of $\text{MSE}_\alpha = (1/1000)\sum_{l=1}^{1000}\sum_{k=1}^{K}(\hat{\alpha}_{\{l\}} - \alpha_k^{(0)})/K$, whose denominator increases with K. In fact, such worsening estimation performance of the regression coefficients is further reflected in Table S6, which demonstrates the increasingly large error in estimating the interaction matrix for all three estimators as K increases. Furthermore, Tables S4 and S6 show the unregularized estimator consistently has the worst estimation performance, while the proposed estimator performs better than the lasso-regularized estimator when $(n, p) = (400, 50)$.

Table S5 presents the averaged true positives TP = $(1/1000)\sum_{l=1}^{1000}\sum_{k=1}^{5}\mathbf{1}(\hat{\alpha}_{k,\{l\}} \neq 0)$ and averaged false positives FP = $(1/1000)\sum_{l=1}^{1000}\sum_{k=6}^{K}\mathbf{1}(\hat{\alpha}_{k,\{l\}} \neq 0)$, along with the TPR $= (1/1000)\sum_{l=1}^{1000}\sum_{k=1}^{5}\mathbf{1}(\hat{\alpha}_{k,\{l\}} \neq 0)/5$ and FPR $= (1/1000)\sum_{l=1}^{1000}\sum_{k=6}^{K}\mathbf{1}(\hat{\alpha}_{k,\{l\}} \neq 0)/(K-5)$ as defined previously in Section

S7.1. As the number of irrelevant similarity matrices increases, the similarity selection performance of all four estimators of the regression coefficients (Regularized, Lasso, AIC and BIC) becomes worse as indicated by the decreasing and increasing patterns of their TP and FP, respectively. When $(n, p) = (400, 50)$, the TP for all estimators are always equal to the number of truly non-zero regression coefficients i.e., five, under all values of K. Again, the decreasing patterns of the FPR for some of the estimators are due to the denominator in the expression of FPR, which includes a factor of $(K - 5)$. Among the four estimators, the TP of AIC decreases at the slowest rate, while its FP increases at the fastest rate as K increases. This is to be expected given the weaker model complexity penalty in the AIC. The proposed estimator also tends to perform better than the lasso-regularized estimator, especially in terms of much smaller FP in some cases. Finally, the BIC estimator tends to have comparable selection performance to our proposed estimator.

S8 Supplementary Details of Application to U.S. Senate Roll Call Voting Data

We first describe the procedure in constructing the final U.S. Senate roll call voting dataset used for analysis in Section 5 of the main text. Following Guo et al.

Table S4: MSE for the Regularized, Lasso, and Unregularized estimators of the regression coefficients (MSE_α) and main effect parameters (MSE_θ) in the simulation study involving the Ising similarity regression model with $K \in \{10, 20, 40, 80, 200\}$. MSE_α is multiplied by 1000 for clarity.

			$1000 \times \text{MSE}_\alpha$			MSE_θ		
p	n	K	Regularized	Lasso	Unregularized	Regularized	Lasso	Unregularized
25	50	10	22.645	20.630	25.176	0.356	0.321	0.402
		20	14.717	13.474	25.985	0.391	0.351	0.546
		40	10.190	8.386	31.193	0.435	0.366	0.795
		80	7.622	4.990	46.441	0.513	0.377	1.439
		200	4.875	2.340	276.987	0.608	0.407	7.725
50	400	10	0.612	0.836	0.918	0.029	0.033	0.036
		20	0.386	0.625	0.939	0.033	0.038	0.047
		40	0.270	0.433	0.992	0.036	0.041	0.059
		80	0.197	0.279	1.056	0.040	0.044	0.078
		200	0.143	0.147	1.264	0.048	0.049	0.136

(2010), we remove procedural votes with the 'Yea/Nay' proportion falling outside of the interval [0.3,0.7], as these are uncontroversial bills that do not reveal political dynamics in the U.S. Senate. The original dataset also contains a small number of missing values since not all senators vote on every bill. Specifically, 1.44% of all votes are missing, and we choose to impute these with the majority vote of the senator's party on the same bill. Note even though we use party to construct one of the similarity matrices for our model in Section 5, we do not believe this form of imputation impacts our analysis to any great degree, given the percentage of the missing data is very small. As a result, the final dataset is made up of $n = 138$ bills voted by $p = 100$ senators.

Next, we provide a descriptive analysis on the binary votes as well as various attributes of the 100 U.S. senators from the 117-th Congress. Figure S1

Table S5: TPR and FPR along with TP and FP (in parentheses) for the Regularized, Lasso, AIC and BIC estimators of the regression coefficients in the simulation study involving the Ising similarity regression model with $K \in \{10, 20, 40, 80, 200\}$.

			TPR (TP)			
p	n	K	Regularized	Lasso	AIC	BIC
25	50	10	0.855 (4.274)	0.911 (4.555)	0.937 (4.685)	0.875 (4.375)
		20	0.835 (4.174)	0.859 (4.295)	0.927 (4.636)	0.856 (4.282)
		40	0.803 (4.017)	0.785 (3.927)	0.921 (4.603)	0.832 (4.159)
		80	0.75 (3.748)	0.697 (3.484)	0.899 (4.495)	0.785 (3.924)
		200	0.538 (2.688)	0.596 (2.982)	0.836 (4.179)	0.597 (2.986)
50	400	10	1 (5)	1 (5)	1 (5)	1 (5)
		20	1 (5)	1 (5)	1 (5)	1 (5)
		40	1 (5)	1 (5)	1 (5)	1 (5)
		80	1 (5)	1 (5)	1 (5)	1 (5)
		200	1 (5)	1 (5)	1 (5)	1 (5)
			FPR (FP)			
p	n	K	Regularized	Lasso	AIC	BIC
25	50	10	0.252 (1.259)	0.457 (2.283)	0.425 (2.125)	0.262 (1.312)
		20	0.218 (3.265)	0.298 (4.473)	0.425 (6.368)	0.237 (3.557)
		40	0.182 (6.369)	0.181 (6.349)	0.436 (15.255)	0.212 (7.435)
		80	0.161 (12.043)	0.104 (7.775)	0.473 (35.478)	0.194 (14.581)
		200	0.108 (21.010)	0.051 (9.849)	0.643 (125.447)	0.163 (31.702)
50	400	10	0.073 (0.364)	0.617 (3.086)	0.098 (0.492)	0.07 (0.349)
		20	0.102 (1.537)	0.428 (6.417)	0.222 (3.326)	0.067 (1.008)
		40	0.128 (4.477)	0.291 (10.184)	0.347 (12.128)	0.051 (1.784)
		80	0.127 (9.537)	0.193 (14.454)	0.406 (30.485)	0.037 (2.749)
		200	0.118 (23.077)	0.099 (19.337)	0.435 (84.809)	0.023 (4.575)

provides two histograms detailing the average binary votes $\bar{y}_j = \sum_{i=1}^{138} y_{ij}/138$ for each senator and $\bar{y}_i = \sum_{j=1}^{100} y_{ij}/100$ for each bill, where $i = 1, \cdots, 138$ and $j = 1, \cdots, 100$. The mean and median for $\bar{y}_j$ are given as 0.5323 and 0.5072, respectively, while the mean and median for $\bar{y}_i$ are given as 0.5323 and 0.5, respectively. Out of a total of 100 senators, 24 of those are females while the remaining 76 are males. In terms of the party distribution, there are 48 Democrat senators, 50 Republican senators and two independent senators. Out of the

Table S6: Frobenius-error and spectral-error for the Regularized, Lasso, and Unregularized estimators of the interaction matrix $\mathbf{\Theta}$ in the simulation study involving the Ising similarity regression model with $K \in \{10, 20, 40, 80, 200\}$.

			Frobenius-error			Spectral-error		
p	n	K	Regularized	Lasso	Unregularized	Regularized	Lasso	Unregularized
25	50	10	4.171	4.004	4.408	1.723	1.649	1.811
		20	4.605	4.421	5.822	1.880	1.798	2.365
		40	5.148	4.759	8.088	2.093	1.920	3.265
		80	5.914	5.067	12.209	2.397	2.031	4.944
		200	6.844	5.419	31.047	2.773	2.168	14.122
50	400	10	1.519	1.684	1.759	0.501	0.542	0.554
		20	1.631	1.911	2.247	0.527	0.602	0.683
		40	1.790	2.109	2.917	0.565	0.657	0.868
		80	2.020	2.310	3.926	0.628	0.714	1.152
		200	2.465	2.561	6.106	0.746	0.787	1.770

50 U.S. states, there are 21 states with both Democrat senators, 22 states with both Republican senators, 5 states with one Democrat senator and one Republican senator, one state with one democrat senator and one independent senator, and one state with one Republican senator and one independent senator. The number of senators for each of the seven most common occupation that are used in Section 5 for constructing similarity matrices is given in Figure S2, demonstrating that a large number ($>$40) of senators are lawyers. Figure S3 provides the histograms for the age, number of Tweets, and number of Twitter followers of the 100 senators. The age of the senators range from 34 to 87 years old and is centered around 65 years old. The number of Tweets and number of Twitter followers both have a right-skewed distribution, indicating the presence of a small number of senators who are much more active and popular than the other senators on Twitter.

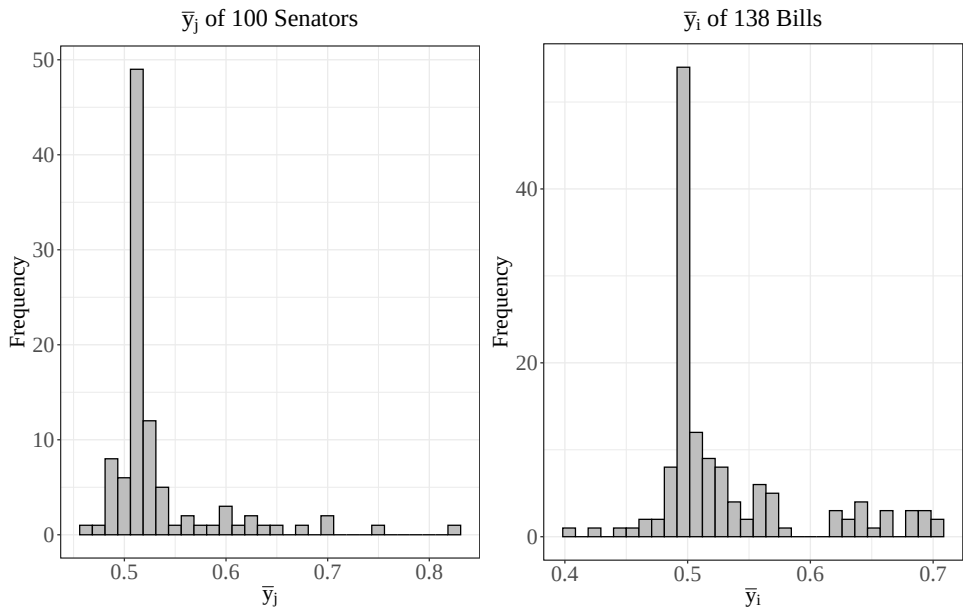

Figure S1: Average binary votes $\{\bar{y}_j : j = 1, \cdots, 100\}$ for each U.S. senator (left) and $\{\bar{y}_i : i = 1, \cdots, 138\}$ for each bill (right) in the 117-th Congress.

For qualitative attributes (state, party, class and gender), we construct similarity matrices $\boldsymbol{W}_k$ by setting $w_{jj'}^{(k)} = 1$ if the j-th and j'-th senators are in the same category and zero otherwise. For the occupation variable, given there is such a large number of categories, we consider the seven most common occupations in the sample (lawyer, executive, businessman, farmer, army, teacher, professor) and construct a separate similarity matrix for each of these occupations. That is, $w_{jj'}^{(k)} = 1$ if the j-th and j'-th senators have the same specific occupation and zero otherwise. As for the similarity matrices based on the quantitative attributes (age, number of tweets, number of Twitter followers), we set

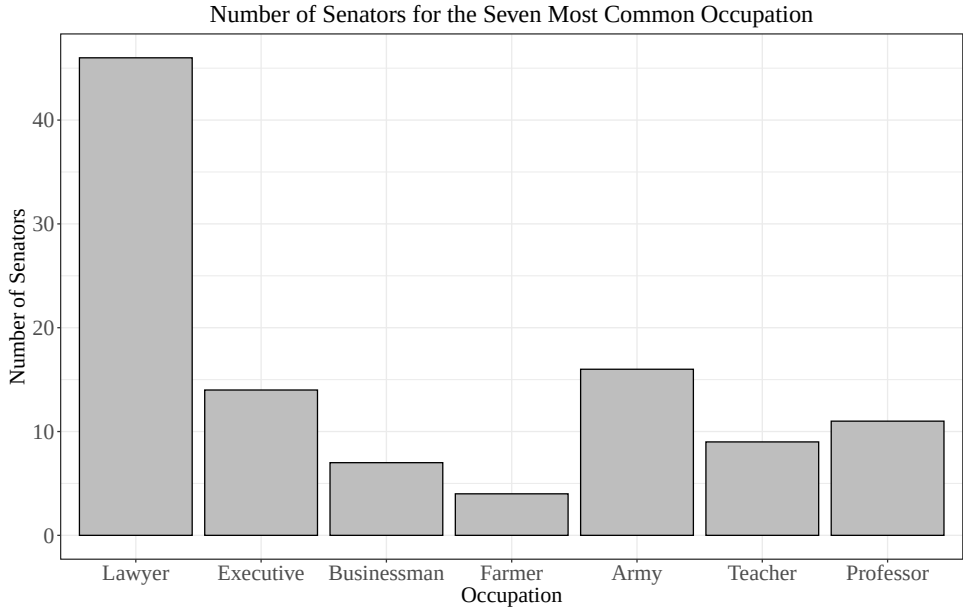

Figure S2: Number of senators for each of the seven most common occupation.

their off-diagonal elements to be $w_{jj'}^{(k)} = \exp(-|z_{jk} - z_{j'k}|^2)$.

Figure S4 presents a histogram summarizing the estimated main effect parameters $\{\hat{\theta}_{jj} : j = 1, \cdots, 100\}$ of all senators. It can be seen that the estimated main effect parameters ranged from -12.5 to -1.5, with most of them concentrated around -10 to -5. Figure S5 presents the graph for the estimated interaction matrix $\hat{\mathbf{\Theta}}$ without any removal of edges for the same subset of 20 senators as in Figure 1 in the main text, which gives qualitatively similar interpretation as Figure 1.

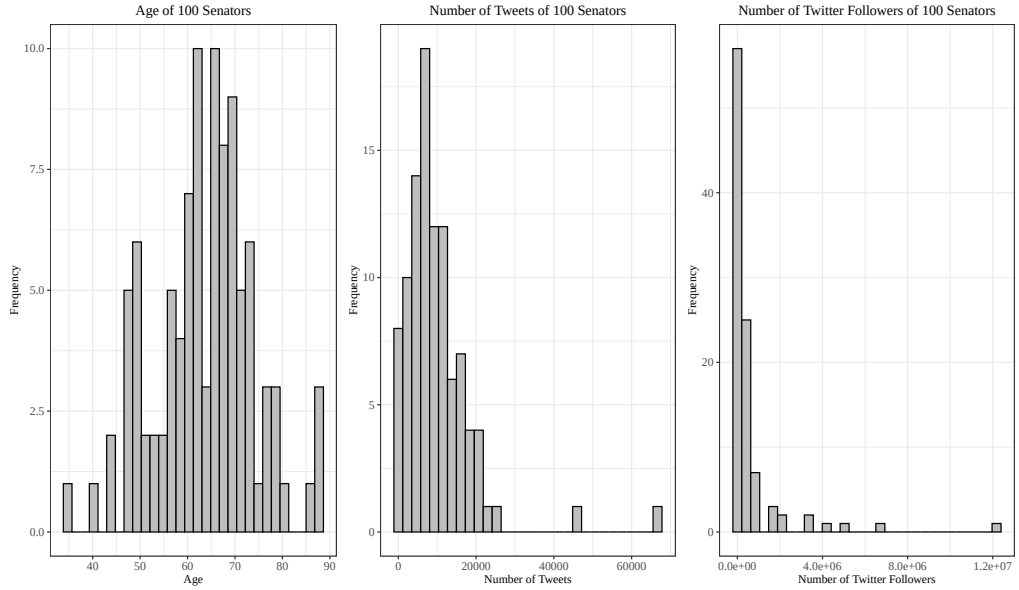

Figure S3: Age, number of Tweets and number of Twitter followers of 100 U.S. Senators in the 117-th Congress.

S9 Application to Scotland Carabidae Ground Beetle Dataset

We apply the Ising similarity regression model to study an ecology dataset provided by Ribera et al. (2001). The original data consists of counts of $p = 68$ Carabidae ground beetle species collected from a total of $n = 87$ sites spread across nine main areas in Scotland using pitfall traps, which we convert into binary presence-absence records. Additional details of the dataset can be found in Ribera et al. (2001), and we note Carabidae ground beetles are commonly used for assessing environmental pollution and characterizing soil-nutrient status (e.g., Szyszko, 1983; Heliövaara and Vaisanen, 2018).

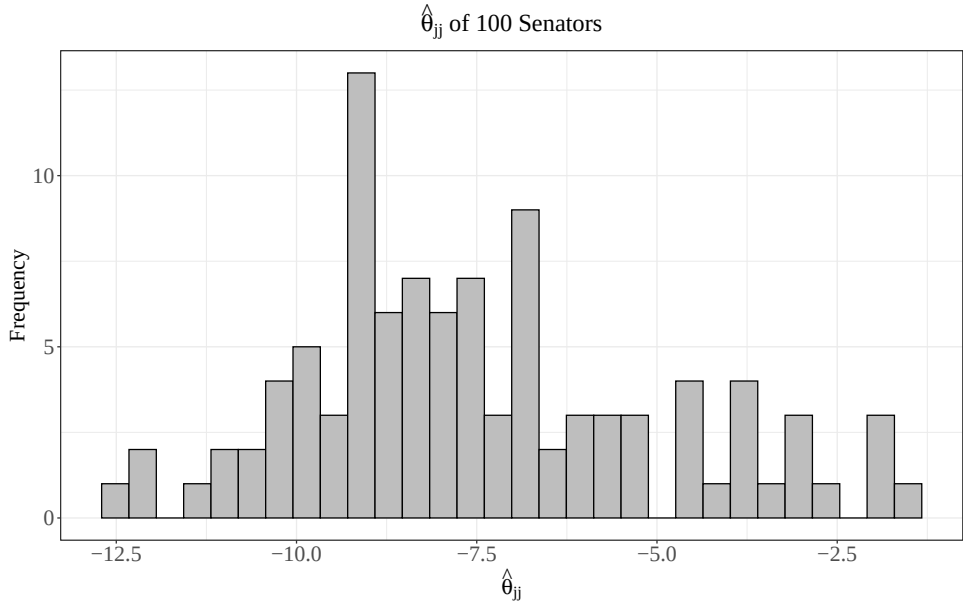

Figure S4: Estimation results for the main effect parameters $\{\hat{\theta}_{jj} : j = 1, \cdots, 100\}$ based on fitting the Ising similarity regression model (2.3) to the U.S. Senate roll call voting data using regularized pseudo-likelihood estimation.

The main purpose of this application is to investigate the relationship between trait similarity and the conditional dependence structure of species presences, by fitting the proposed Ising similarity regression model to the presence-absence records using similarity matrices constructed based on species traits. This allows us to identify species traits that are important in explaining the dependence structure of species occurrences e.g., suggestive of potential biotic interactions between species. The dataset consists of $K = 20$ traits, ten of which are quantitative while the remaining ten are qualitative; see Table S7 for the full list of traits and their corresponding abbreviations. Figure S6 presents

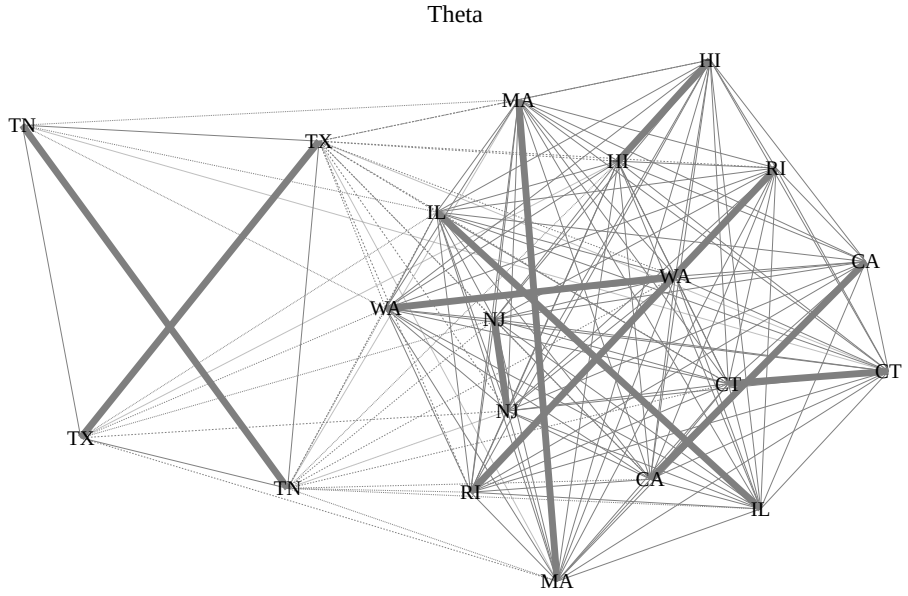

Figure S5: Weighted graph based on the estimated interaction matrix $\hat{\mathbf{\Theta}}$ for a subset of 20 senators with all edges included, where the edge width is proportional to its associated $\hat{\theta}_{jj'}$ and edges between senators with different states and parties are represented as dashed lines. Nodes are labeled with the state abbreviation of each senator. The four nodes on the left, labeled 'TN' or 'TX', are Republicans, while the remaining nodes are Democrats.

histograms summarizing the average binary presences $\bar{y}_j = \sum_{i=1}^{87} y_{ij}/87$ for each species and $\bar{y}_i = \sum_{j=1}^{68} y_{ij}/68$ for each site, where $i = 1, \cdots, 87$ and $j = 1, \cdots, 68$, with the mean and median for $\bar{y}_j$ given as 0.3005 and 0.2471, respectively, while the mean and median for $\bar{y}_i$ are given as 0.3005 and 0.2941, respectively. Figure S7 shows most distributions of the quantitative traits are symmetric, while Table S8 provides a breakdown of the 68 species into different categories of each qualitative trait variable.

Similar to the application in Section 5, we construct a similarity matrix for

Table S7: Trait variables of the Scotland Carabidae ground beetle dataset along with their abbreviations. Adapted from Ribera et al. (2001).

Quantitative (all measures log-transformed)	
LYW	Diameter of the eye, measured from above
LAL	Length of the antenna
LPW	Maximum width of the pronotum
LPH	Maximum width ("vaulting") of the pronotum
LEW	Maximum width of the elytra
LFL	Length of the metafemur (with the articulation segments), from the coxa to the apex
LTR	Length of the metatrochanter
LRL	Length of the metatarsi
LFW	Maximum width of the metafemur
LTL	Total length (length of the pronotum in the medial line plus length of the elytra, from the medial ridge of the scutellum to the apex)
Qualitative	
CLG	Color of the legs (1. pale; 2, black; 3, metallic)
CLB	Color of the body (1. pale; 2, black; 3, metallic)
WIN	Wing development (1, apterous or brachypterous; 2, dimorphic; 3, macropterous)
PRS	Shape of the pronotum (1, oval; 2, cordiform; 3, trapezoidal)
OVE	Overwintering (1, only adults; 2, adults and larvae or only larvae)
FOA	Food of the adult (1, mostly Collembola; 2, generalist predator; 3, mostly plant material)
DAY	Daily activity (1, only diurnal; 2, nocturnal)
BRE	Breeding season (1, spring; 2, summer; 3, autumn or winter)
EME	Main period of emergence of the adults (1, spring; 2, summer; 3, autumn)
ACT	Main period of adult activity (1, autumn; 2, summer only)

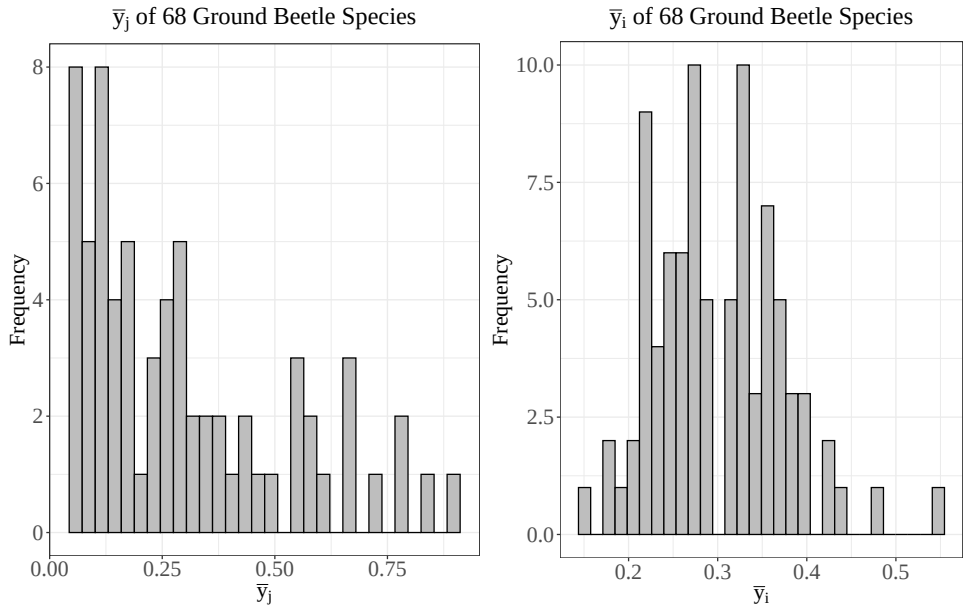

Figure S6: Average binary presences $\{\bar{y}_j : j = 1, \cdots, 68\}$ for each species (left) and $\{\bar{y}_i : i = 1, \cdots, 87\}$ for each site (right) in the Scotland Carabidae ground beetle dataset.

Table S8: Number of species belonging to each category of qualitative trait variables; see Table S7 for definition of each category. Dashes indicate trait variables with only two categories.

Category	CLG	CLB	WIN	PRS	OVE	FOA	DAY	BRE	EME	ACT
1	38	23	18	21	30	12	20	32	6	26
2	22	27	21	26	38	40	48	10	54	42
3	8	18	29	21	—	16	—	26	8	—

each trait, where for qualitative traits we set $w_{jj'}^{(k)} = 1$ if the j-th and j'-th species are in the same category and zero otherwise, while for quantitative traits we consider $w_{jj'}^{(k)} = \exp(-|z_{jk} - z_{j'k}|^2)$. The proposed model is then fitted using the regularized pseudo-likelihood estimator with adaptive lasso penalty, and the tuning parameter λ is selected using ten-fold cross validation where the observations are grouped at the site level. We also construct 95% Wald confidence intervals based on the empirical sandwich-based standard errors in Section S6.

Table S9 presents the estimated regression coefficients associated with the $K = 20$ trait similarity matrices along with their corresponding 95% Wald confidence intervals, while the estimation results for the main effect parameters are given in Figure S4. To summarize, 14 out of the 20 regression coefficients are estimated to be non-zero. Of these, there is statistically clear evidence five traits (LFL, LTR, CLG, WIN and BRE) exhibit positive associations with conditional dependence structure of species presences. For instance, this implies two species with same breeding season (BRE) would have more similar occurrence patterns

Table S9: Point estimates and 95% confidence intervals (in parentheses) for the regression coefficients corresponding to the $K = 20$ similarity matrices, based on fitting the Ising similarity regression model (2.3) to the Scotland Carabidae ground beetle data using regularized pseudo-likelihood estimation. Estimates whose corresponding confidence interval excludes zero are bolded.

Estimation of α_k				
LYW	LAL	LPW	LPH	LEW
-0.058	0	-0.032	0	0
(-0.127,0.012)	0	(-0.111,0.046)	0	0
LFL	LTR	LRL	LFW	LTL
0.110	**0.158**	**-0.176**	0	0
(0.021 , 0.198)	**(0.121,0.194)**	**(-0.280,-0.071)**	0	0
CLG	CLB	WIN	PRS	OVE
0.057	0.025	**0.095**	-0.000	0
(0.020, 0.095)	(-0.029,0.079)	**(0.066, 0.125)**	(-0.047,0.047)	0
FOA	DAY	BRE	EME	ACT
0.038	-0.063	**0.123**	**-0.049**	-0.022
(-0.009,0.085)	(-0.130,0.005)	**(0.088,0.158)**	**(-0.090,-0.008)**	(-0.064,0.021)

across sites, which in turn may suggest this trait plays an important role in mediating positive biotic interactions between the species. On the other hand, two traits (LRL and EME) are found to have a negative effect on the dependence structure among species presences e.g., two species with same main period of adult emergence (EME) are less likely to be present at the same site, indicating this trait could be related to competition between species.

Figure S9 presents graphs of the weighted similarity matrices $\hat{\alpha}_k \boldsymbol{W}_k$ for selected traits, along with the estimated interaction matrix $\hat{\boldsymbol{\Theta}}$ for a subset of ten species. The choice of species to be included is made by beginning with an empty set, and sequentially adding pairs of species which have the largest off-diagonal elements $\hat{\theta}_{jj'}$ in $\hat{\boldsymbol{\Theta}}$ until the set contained the top ten species. It

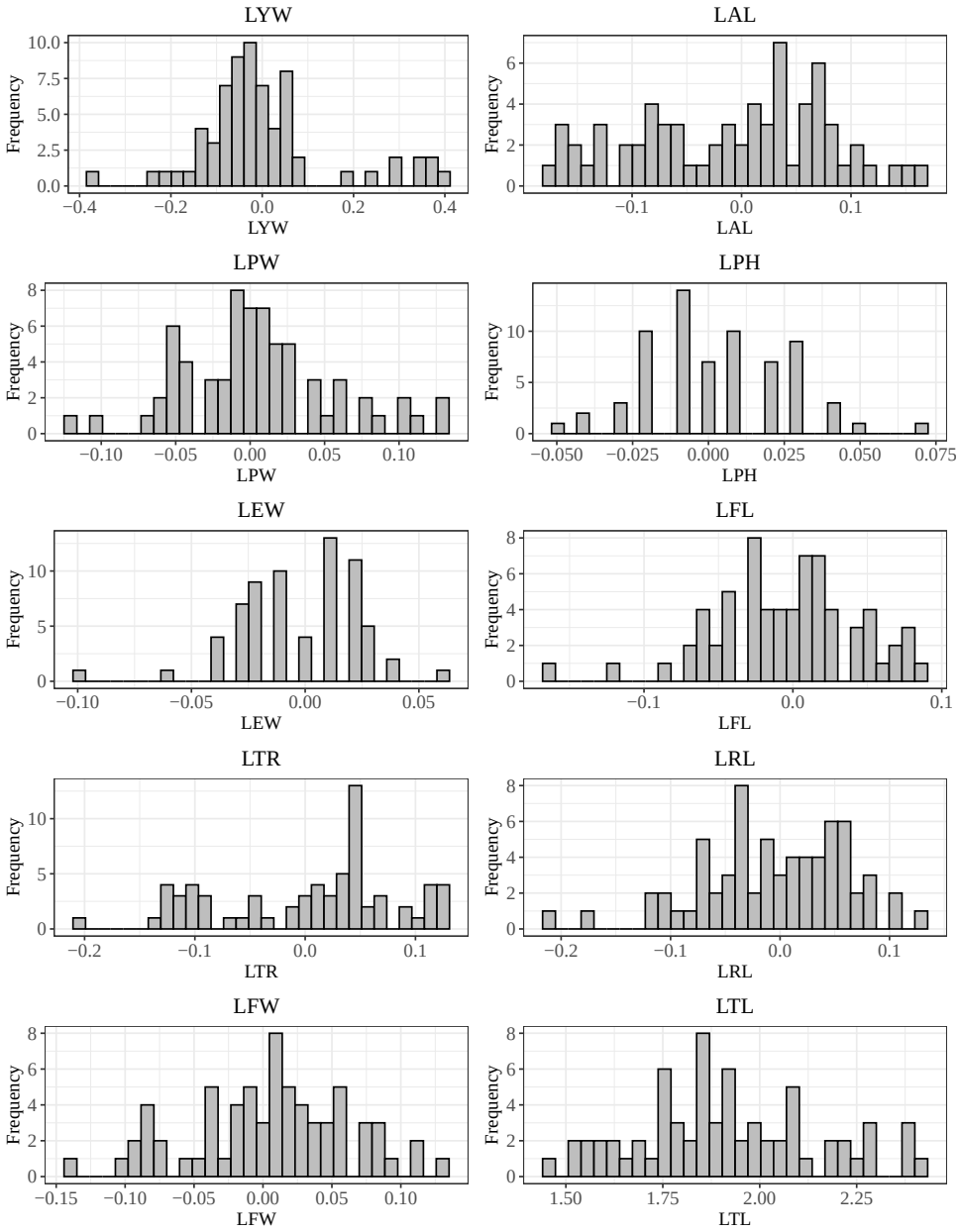

Figure S7: Histograms for ten quantitative traits of 68 ground beetle species.

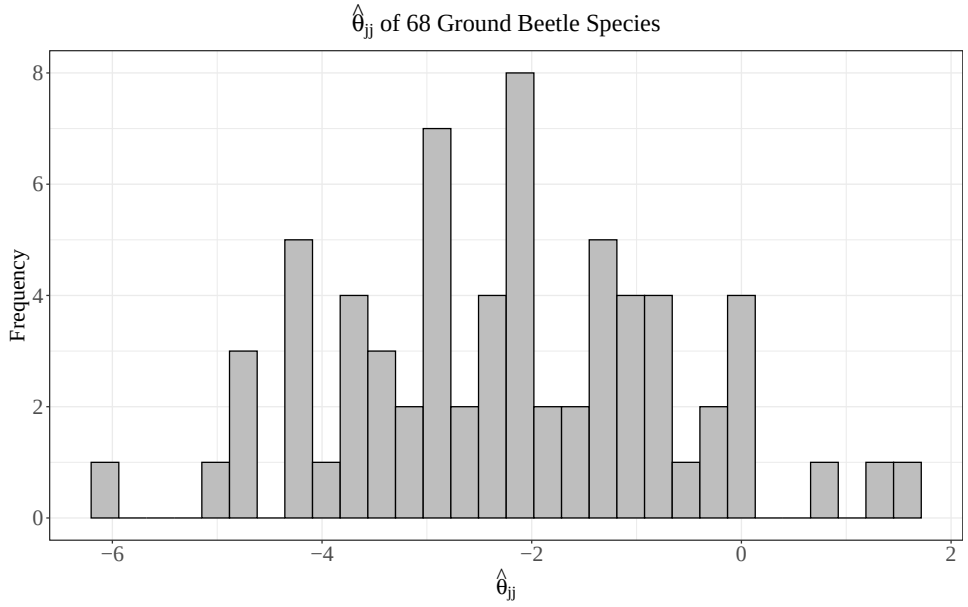

Figure S8: Estimation results for the main effect parameters $\{\hat{\theta}_{jj} : j = 1, \cdots, 68\}$ based on fitting the Ising similarity regression model (2.3) to the Scotland Carabidae ground beetle data using regularized pseudo-likelihood estimation.

can be seen that similarity in terms of breeding season and wing development contributes more than similarity in terms of color of legs to the conditional dependence structure between species presences, due to their larger estimated regression coefficients in Table S9.

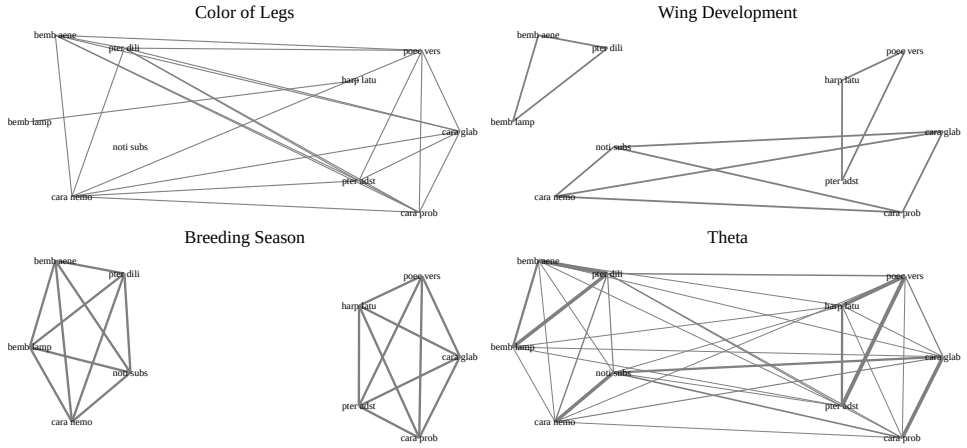

Figure S9: Graphs of weighted similarity matrices $\hat{\alpha}_k \boldsymbol{W}_k$ associated with color of legs, wing development and breeding season for a subset of 10 species, where the edge width is proportional to the estimated $\hat{\alpha}_k$. The bottom right plot presents the weighted graph based on the estimated interaction matrix $\hat{\boldsymbol{\Theta}}$, where the edge width is proportional to its associated $\hat{\theta}_{jj'}$. Nodes are labeled with the code of each species; see Table 4 of Ribera et al. (2001) for the corresponding species names.

References

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics* **41**, 802–837.

Cheng, J., E. Levina, P. Wang, and J. Zhu (2014). A sparse Ising model with covariates. *Biometrics* **70**, 943–953.

Epskamp, S. (2020). *IsingSampler: Sampling methods and distribution functions for the Ising Model*. R package version 0.2.1.

Guo, J., J. Cheng, E. Levina, G. Michailidis, and J. Zhu (2015). Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *The Annals of Applied Statistics* **9**, 821–848.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2010). Joint structure estimation for categorical Markov networks. Unpublished manuscript, `https://www.researchgate.net/publication/266584068_Joint_Structure_Estimation_for_Categorical_Markov_Networks`.

Heliövaara, K. and R. Vaisanen (2018). *Insects and pollution*. CRC Press.

Höfling, H. and R. Tibshirani (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research **10***, 883–906.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44**, 907–927.

Ribera, I., S. Dolédec, I. S. Downie, and G. N. Foster (2001). Effect of land disturbance and stress on

species traits of ground beetle assemblages. *Ecology* **82**, 1112–1129.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.

Seber, G. (2008). *A matrix handbook for statisticians*. Wiley.

Szyszko, J. (1983). *State of Carabidae (Col.) fauna in fresh pine forest and tentative valorisation of this environment*. Warsaw Agricultural University Press.

Wainwright, M. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Wainwright, M. J., J. Lafferty, and P. Ravikumar (2006). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems*, Volume **19**, pp. 1465–1472. MIT Press.

Zou, T., W. Lan, H. Wang, and C.-L. Tsai (2017). Covariance regression analysis. *Journal of the American Statistical Association* **112**, 266–281.