NETWORK MODEL AVERAGING PREDICTION FOR LATENT SPACE MODELS BY K-FOLD EDGE CROSS-VALIDATION

Yan Zhang[1*], Jun Liao[2*], Xinyan Fan[2], Kuangnan Fang[3†], Yuhong Yang[4]

[1]*Shanghai University of International Business and Economics*

[2]*Renmin University of China,* [3]*Xiamen University,* [4]*Tsinghua University*

Supplementary Material

The supplementary material provides additional details including theoretical proofs, assumption verifications, algorithmic procedures, and extended simulation and empirical results. Specifically, Section S1 provides the theoretical results for multi-layer networks. Section S2 contains the proofs of Theorems 1–3 and S1–S3. Section S3 provides detailed verifications of Assumptions 1–6. Section S4 introduces the projected gradient descent algorithm for parameter estimation in multi-layer networks. Section S5 describes the procedure for selecting model weights in the multi-layer setting. Section S6 presents the simulation results for both single-layer and multi-layer networks. Finally, Section S7 reports the result of weight allocation and empirical analysis based on the ResearchGate dataset.

*Yan Zhang and Jun Liao are co-first authors and contributed equally to this work.

†Corresponding author. E-mail address: xmufkn@xmu.edu.cn.

S1 Theoretical property of Multi-layer Networks

For multi-layer networks, we also provide three theoretical properties for our method. Theorem S1 shows that the empirical K-fold cross-validation weights asymptotically minimize $L_\dagger(w)$. The assumptions required for Theorem S1 are discussed as follows.

Assumption S1. Suppose that $M \leq N$. For $(i,j) \in \Psi_1$, there exist limiting values $P^{(t)*}_{(m),ij}$ for $\widehat{P}^{(t)}_{(m),ij}$ such that $\sum_{t=1}^{T} \sum_{(i,j)\in\Psi_1} \left(\widehat{P}^{(t)}_{(m),ij} - P^{(t)*}_{(m),ij}\right)^2 = O_p(NMT)$ uniformly for $m = 1, \ldots, M$.

Assumption S1 implies that the estimator of $P^{(t)}_{(m),ij}$ of layer t in the mth candidate model has a limit $P^{(t)*}_{(m),ij}$. Assume that $\|\Theta^{(t)}_{(m)}\|_{\max} \leq \mu$ uniformly for $t = 1, \ldots, T$ and $m = 1, \ldots, M$, where $\Theta^{(t)}_{(m)} = \text{logit}(P^{(t)}_{(m)})$, and $\|\cdot\|_{\max}$ represents the maximum absolute value of entries in a matrix. According to the proof of Theorem 1 in Zhang et al. (2020), when a candidate model has large enough latent dimensions, we have $\sum_{t=1}^{T} \|\widehat{\Theta}^{(t)}_{(m)} - \Theta^{(t)}_{(m)}\|_F^2 = O_p(C_2 NT + C_3 m^2(2N+T))$, where C_2 and C_3 depend on μ. Next, we can obtain $\sum_{t=1}^{T} \sum_{(i,j)\in\Psi_1} (\widehat{P}^{(t)}_{(m),ij} - P^{(t)*}_{(m),ij})^2 \leq \sum_{t=1}^{T} \|\widehat{\Theta}^{(t)}_{(m)} - \Theta^{(t)}_{(m)}\|_F^2/16 = O_p(C_2' NT + C_3' m^2(2N+T))$, where $C_2' = C_2/16$ and $C_3' = C_3/16$. When we assume μ and M to be fixed constants, and $M \leq T \leq N$, we have $\sum_{t=1}^{T} \sum_{(i,j)\in\Psi_1} (\widehat{P}^{(t)}_{(m),ij} - P^{(t)*}_{(m),ij})^2 = O_p(NMT)$ uniformly for $m = 1, \ldots, M$.

Additionally, we define an upper bound on the expected nodal degree over the whole layers $D_\dagger$, which satisfies $D_\dagger = N \max_t \max_{ij} P_{ij}^{(t)}$.

Assumption S2. $TN^{-C_{0\dagger}} = o(1)$ for some positive constant $C_{0\dagger}$.

Assumption S2 is imposed to get the uniform order of $\|A^{(t)} - P^{(t)}\|_2$ for $t = 1, \ldots, T$, where $\|\cdot\|_2$ denotes the spectral norm of a matrix. Next, we denote the loss function for multi-layer networks based on the limiting value as $L_\dagger^*(w) = \sum_{t=1}^T \sum_{(i,j)\in\Psi_1} \{P_{ij}^{(t)*}(w) - P_{ij}^{(t)}\}^2$, where $P_{ij}^{(t)*}(w) = \sum_{m=1}^M w_m P_{(m),ij}^{(t)*}$ and $P_{(m),ij}^{(t)*} = f_2(\alpha_{(m),i}^* + \alpha_{(m),j}^* + z_{(m),i}^{(t)*\top} z_{(m),j}^{(t)*})$. The minimum loss in the class of averaging estimators based on the limiting value is denoted as $\xi_\dagger^* = \inf_{w\in\mathcal{W}} L_\dagger^*(w)$.

Assumption S3. $NMT\xi_\dagger^{*-1} = o(1)$ and $N\max\{D_\dagger, \log N\}T\xi_\dagger^{*-1} = O(1)$.

Assumption S3 requires that $\xi_\dagger^*$ grows faster than NMT, and $N\max\{D_\dagger, \log N\}T\xi_\dagger^{*-1}$ is finite. It is important to note that Assumption S3 does not hold when the set of candidate models includes models with large enough latent dimensions. Specifically, if the m^0th model has large enough latent dimensions, then we have $P_{(m),ij}^{(t)*} = P_{ij}^{(t)}$. This leads to the prediction error $\xi_\dagger^*$ being zero:

$$\xi_\dagger^* = \inf_{w\in\mathcal{W}} \sum_{t=1}^T \sum_{(i,j)\in\Psi_1} \left\{P_{ij}^{(t)*}(w) - P_{ij}^{(t)}\right\}^2 = \sum_{t=1}^T \sum_{(i,j)\in\Psi_1} \left\{P_{(m),ij}^{(t)*} - P_{ij}^{(t)}\right\}^2 = 0.$$

A value of $\xi_\dagger^* = 0$ directly violates Assumption S3. On the other hand,

Assumption S3 holds when all candidate models are misspecified and $M = o(N)$. Consider a scenario where the true latent dimension $d_{0,t}$ for each layer exceeds the maximum candidate dimension M (i.e., $d_{0,t} > M$). For the first part, if the latent factors in the omitted $d_{0,t} - M$ dimensions represent non-negligible structural signals, the squared bias $\xi_\dagger^*$ will be of order N^2T. Given $M = o(N)$, the ratio $NMT\xi_\dagger^{*-1} = o(1)$. For the second part, since the expected nodal degree $D_\dagger \leq N$, the ratio $N \max\{D_\dagger, \log N\}T\xi_\dagger^{*-1}$ remains bounded.

Theorem S1. *Under Assumptions S1-S3, we have*

$$\frac{L_\dagger(\widehat{w}_\dagger)}{\inf_{w\in\mathcal{W}} L_\dagger(w)} \to 1$$

in probability.

Theorem S1 implies that the prediction loss $L_\dagger(w)$ is asymptotically minimized by the K-fold edge cross-validation weights $\widehat{w}_\dagger$. The proof of Theorem S1 is given in Section S2.4 of the Supplementary Material. In scenarios where the candidate set includes models with large enough latent dimensions, we introduce the following assumption.

Assumption S4. $NMT\{\inf_{w\in\mathcal{W}^s} L_\dagger^*(w)\}^{-1} = o(1)$ and $NT \max\{D_\dagger, \log N\}$ $\{\inf_{w\in\mathcal{W}^s} L_\dagger^*(w)\}^{-1} = O(1)$.

Assumption S4 is similar to Assumption 3 in single-layer networks. Denote $\widehat{\zeta}_\dagger = \sum_{m\in\mathcal{T}_\dagger} \widehat{w}_{\dagger,m}$, where $\widehat{w}_{\dagger,m}$ is the mth element of $\widehat{w}_\dagger$, and $\mathcal{T}_\dagger$ is the index set of the models with large enough dimensions.

Theorem S2. *Under Assumptions S1, S2 and S4, if $\mathcal{T}_\dagger$ is not empty, we have $\widehat{\zeta}_\dagger \to 1$ in probability.*

Theorem S2 shows that the proposed method for multi-layer networks asymptotically assigns all weights to the models with large enough latent dimensions if the candidate set includes them. The proof of Theorem S2 can be found in Section S2.5 of the Supplementary Material. In the following content, we provide the convergence rate of the K-fold edge cross-validation-based weights. Firstly, we give some notations. For $t = 1, \ldots, T$, denote the squared risk for multi-layer networks as $R_\dagger(w) = \sum_{t=1}^{T} \sum_{(i,j)\in\Psi_1} E(\widehat{P}_{ij}^{(t)}(w) - P_{ij}^{(t)})$. Let $\xi_\dagger = \inf_{w\in\mathcal{W}} R_\dagger(w)$ and $w_\dagger^0 = \operatorname{argmin}_{w\in\mathcal{W}} R_\dagger(w)$. Arrange $\{A_{ij}^{(t)},\ (i,j)\in\Psi_1\}$, $\{P_{ij}^{(t)},\ (i,j)\in\Psi_1\}$, $\{\sum_{k=1}^{K}\widetilde{P}_{ij}^{(t)[k]}(w),\ (i,j)\in\Psi_1\}$, $\{\widetilde{P}_{(m),ij}^{(t)},\ (i,j)\in\Psi_1\}$, $\{\widehat{P}_{ij}^{(t)}(w),\ (i,j)\in\Psi_1\}$, $\{\widehat{P}_{(m),ij}^{(t)},\ (i,j)\in\Psi_1\}$, $\{P_{ij}^{(t)*}(w),\ (i,j)\in\Psi_1\}$ and $\{P_{(m),ij}^{(t)*},\ (i,j)\in\Psi_1\}$ in same particular order. Then denote them as vector $a_1^{(t)}\in\mathbb{R}^{|\Psi_1|}$, $p_1^{(t)}\in\mathbb{R}^{|\Psi_1|}$, $\widetilde{p}_1^{(t)}(w)\in\mathbb{R}^{|\Psi_1|}$, $\widetilde{p}_{1(m)}^{(t)}\in\mathbb{R}^{|\Psi_1|}$, $\widehat{p}_1^{(t)}(w)\in\mathbb{R}^{|\Psi_1|}$, $\widehat{p}_{1(m)}^{(t)}\in\mathbb{R}^{|\Psi_1|}$, $p_1^{(t)*}(w)\in\mathbb{R}^{|\Psi_1|}$ and $p_{1(m)}^{(t)*}\in\mathbb{R}^{|\Psi_1|}$, respectively. We next introduce two matrices $A^{(t)\circ}\in\mathbb{R}^{N\times N}$ and $P^{(t)\circ}\in\mathbb{R}^{N\times N}$, where the (i,j)th element of $A^{(t)\circ}$ is defined as $A_{ij}^{(t)\circ} = A_{ij}^{(t)}$ if

$(i,j) \in \Psi_1$, otherwise $A_{ij}^{(t)\circ} = 0$. Similarly, the (i,j)th element of $P^{(t)\circ}$ is defined as $P_{ij}^{(t)\circ} = P_{ij}^{(t)}$ if $(i,j) \in \Psi_1$, otherwise $P_{ij}^{(t)\circ} = 0$. Denote $\Lambda_1^{(t)} = (\widehat{p}_{1(1)}^{(t)}, \ldots, \widehat{p}_{1(M)}^{(t)})$, $\Omega_1^{(t)} = (p_1^{(t)} - \widehat{p}_{1(1)}^{(t)}, \ldots, p_1^{(t)} - \widehat{p}_{1(M)}^{(t)})$, $\Lambda^{(t)} = \Lambda_1^{(t)\top}\Lambda_1^{(t)}$ and $\Omega^{(t)} = \Omega_1^{(t)\top}\Omega_1^{(t)}$. Theorem S3 shows the rate of the $\widehat{w}_\dagger$ tending to the infeasible optimal weight vector $w_\dagger^0$. The following assumptions are needed to show this theorem.

Assumption S5. There are two positive constants $\rho_{1\dagger}$ and $\rho_{2\dagger}$, such that $0 < \rho_{1\dagger} < \lambda_{\min}\{\sum_{t=1}^{T}\Lambda^{(t)}/(T|\Psi_1|)\} \leq \lambda_{\max}\{\sum_{t=1}^{T}\Lambda^{(t)}/(T|\Psi_1|)\} \leq M$, in probability tending to 1.

Assumption S6. $\lambda_{\max}\{\sum_{t=1}^{T}\Omega^{(t)}/(T|\Psi_1|)\} = O_p(M)$.

Assumption S7. $N^{1-4\kappa_\dagger}\max\{D_\dagger, \log N\}MT\xi_\dagger^{-1} = o(1)$, $N^{1-4\kappa_\dagger}M^2T\xi_\dagger^{-1} = o(1)$, and $M = o(N^{\min\{4\kappa_\dagger, 1/3\}})$, where $\kappa_\dagger \in (0, 1/2)$.

Assumptions S5 and S6 are similar to Assumptions 4 and 5, respectively. Assumption S7 involves replacing ξ^{-1} with $T\xi_\dagger^{-1}$ in comparison to Assumption 6.

Theorem S3. *If $w_\dagger^0$ is an interior point of $\mathcal{W}$ and Assumptions S1, S2 and S5-S7 are satisfied, then the global minimizer $\widehat{w}_\dagger$ satisfies $\|\widehat{w}_\dagger - w_\dagger^0\| = O_p(\xi_\dagger^{1/2}T^{-1/2}|\Psi_1|^{-1/2+\kappa_\dagger})$, where $\kappa_\dagger$ is defined in Assumption S7.*

Theorem S3 shows the convergence rate of $\widehat{w}_\dagger$ towards $w_\dagger^0$. Compared

to Theorem 3, the convergence rate in the case of multi-layer networks not only relies on $\xi_\dagger$ and $|\Psi_1|$ but also depends on T. The proof of Theorem S3 is shown in Section S2.6 of the Supplementary Material.

S2 Proof of Theoretical Results

S2.1 Proof of Theorem 1

This subsection gives the proof of Theorem 1. We need the following lemmas.

Lemma 1 (Gao et al. (2019); Zhang and Liu (2023)). *Let*

$$\widetilde{w} = \underset{w \in \mathcal{W}}{\operatorname{argmin}} \left\{ L(w) + a_n(w) + b_n \right\},$$

where $a_n(w)$ is a term related to w, and b_n is a term unrelated to w. If

$$\sup_{w \in \mathcal{W}} |a_n(w)| / L^*(w) = o_p(1), \quad \sup_{w \in \mathcal{W}} |L^*(w) - L(w)| / L^*(w) = o_p(1),$$

and there exists a constant c and a positive integer N^* so that when $n \geq N^*, \inf_{w \in \mathcal{W}} L^*(w) \geq c > 0$ almost surely, then $L(\widetilde{w}) / \inf_{w \in \mathcal{W}} L(w) \to 1$ in probability.

Lemma 2 (Lei and Rinaldo (2015); Ma et al. (2020)). *Let A be the sym-*

metric adjacency matrix of a random graph on N nodes with independent edges. Set $E(A_{ij}) = P_{ij}$ for all $i \neq j$ and $P_{ii} \in [0,1]$. Assume that $N \max_{i,j} P_{ij} \leq D$. Then for any C_0, there exists a constant $C = C(C_0)$ such that

$$\|A - P\|_2 \leq C\sqrt{D + \log N}$$

with probability at least $1 - N^{-C_0}$.

Let $CV^*(w) = CV(w) - |\Psi_1|^{-1} \sum_{k=1}^{K} (\|A^{[k]}\|_F^2 - \|P^{[k]}\|_F^2)$, where $P^{[k]} = P \circ S^{[k]}$, and the second term is unrelated to w. Therefore,

$$\widehat{w} = \text{argmin}_{w \in \mathcal{W}} CV(w) = \text{argmin}_{w \in \mathcal{W}} |\Psi_1| CV^*(w).$$

To apply Lemma 1, we set $b_n = 0$ as the constant terms. The term $a_n(w)$ is defined as $a_n(w) = |\Psi_1|CV^*(w) - L(w)$. This difference can be decomposed into two parts: $|\Psi_1|CV^*(w) - L^*(w)$ and $L^*(w) - L(w)$. Thus, according to Lemma 1, Theorem 1 is valid if the following conditions hold:

$$\sup_{w \in \mathcal{W}} \frac{|L(w) - L^*(w)|}{L^*(w)} = o_p(1), \tag{S2.1}$$

and

$$\sup_{w \in \mathcal{W}} \frac{||\Psi_1|CV^*(w) - L^*(w)|}{L^*(w)} = o_p(1). \tag{S2.2}$$

We first consider (S2.1). It can be shown that

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}}\frac{|L(w)-L^*(w)|}{L^*(w)}=\sup_{w\in\mathcal{W}}\left[\left|\sum_{(i,j)\in\Psi_1}\left\{\left(\widehat{P}_{ij}(w)-P_{ij}\right)^2-\left(P^*_{ij}(w)-P_{ij}\right)^2\right\}\right|/L^*(w)\right]\\
\leq&\sup_{w\in\mathcal{W}}\left[\left|\sum_{(i,j)\in\Psi_1}\left\{\left(\widehat{P}_{ij}(w)-P^*_{ij}(w)\right)^2+2\left(\widehat{P}_{ij}(w)-P^*_{ij}(w)\right)\left(P^*_{ij}(w)-P_{ij}\right)\right\}\right|/L^*(w)\right]\\
\leq&\xi^{*-1}\sup_{w\in\mathcal{W}}\sum_{(i,j)\in\Psi_1}\left\{\sum_{m=1}^{M}w_m\left(\widehat{P}_{(m),ij}-P^*_{(m),ij}\right)\right\}^2\\
&+2\sup_{w\in\mathcal{W}}\left[\frac{1}{L^*(w)}\left\{\sum_{(i,j)\in\Psi_1}\left(\widehat{P}_{ij}(w)-P^*_{ij}(w)\right)^2\right\}^{1/2}\left\{\sum_{(i,j)\in\Psi_1}\left(P^*_{ij}(w)-P_{ij}\right)^2\right\}^{1/2}\right]\\
\leq&\xi^{*-1}\sup_{w\in\mathcal{W}}\sum_{(i,j)\in\Psi_1}\sum_{m=1}^{M}w_m\left(\widehat{P}_{(m),ij}-P^*_{(m),ij}\right)^2+2\xi^{*-1/2}\sup_{w\in\mathcal{W}}\left\{\sum_{(i,j)\in\Psi_1}\left(\widehat{P}_{ij}(w)-P^*_{ij}(w)\right)^2\right\}^{1/2}\\
=&\xi^{*-1}O_p\left(NM\right)+\xi^{*-1/2}O_p\left(N^{1/2}M^{1/2}\right)=o_p(1),
\end{aligned}
\tag{S2.3}
$$

where the fifth step uses Assumption 1, and the sixth step uses Assumption 2. Thus, we obtain (S2.1).

We next verify (S2.2). Observe that

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}}\left\{\left||\Psi_1|CV^*(w)-L^*(w)\right|/L^*(w)\right\}\\
=&\sup_{w\in\mathcal{W}}\left\{\left|\|a_1-\widetilde{p}_1(w)\|^2-(a_1-p_1)^\top(a_1+p_1)-\|p^*_1(w)-p_1\|^2\right|/L^*(w)\right\}\\
\leq&\sup_{w\in\mathcal{W}}\left\{\left|\|a_1-p^*_1(w)\|^2-(a_1-p_1)^\top(a_1+p_1)-\|p^*_1(w)-p_1\|^2\right|/L^*(w)\right\}\\
&+\sup_{w\in\mathcal{W}}\left\{\left|\|a_1-\widetilde{p}_1(w)\|^2-\|a_1-p^*_1(w)\|^2\right|/L^*(w)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \sup_{w\in\mathcal{W}} \left\{ \left| \|p_1^*(w) - p_1 - a_1 + p_1\|^2 - (a_1 - p_1)^\top (a_1 + p_1) - \|p_1^*(w) - p_1\|^2 \right| / L^*(w) \right\} \\
&\quad + \sup_{w\in\mathcal{W}} \left\{ \left| \|a_1 - \widetilde{p}_1(w)\|^2 - \|a_1 - p_1^*(w)\|^2 \right| / L^*(w) \right\} \\
&\leq 2\xi^{*-1} \sup_{w\in\mathcal{W}} \left| p_1^*(w)^\top (a_1 - p_1) \right| + \sup_{w\in\mathcal{W}} \left\{ \left| \|a_1 - \widetilde{p}_1(w)\|^2 - \|a_1 - p_1^*(w)\|^2 \right| / L^*(w) \right\}.
\end{aligned} \tag{S2.4}
$$

Similar to the derivation in (S2.3), based on Assumptions 1 and 2, we can obtain

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}} \left\{ \left| \|a_1 - \widetilde{p}_1(w)\|^2 - \|a_1 - p_1^*(w)\|^2 \right| / L^*(w) \right\} \\
&= \sup_{w\in\mathcal{W}} \left\{ \left| \|p_1^*(w) - \widetilde{p}_1(w)\|^2 + 2\big(a_1 - p_1^*(w)\big)^\top \big(p_1^*(w) - \widetilde{p}_1(w)\big) \right| / L^*(w) \right\} \\
&\leq \xi^{*-1} \sup_{w\in\mathcal{W}} \left\{ \sum_{(i,j)\in\Psi_1} \left(P_{ij}^*(w) - \widetilde{P}_{ij}(w) \right)^2 \right\} + \\
&\quad 2 \sup_{w\in\mathcal{W}} \left\{ \left| \sum_{(i,j)\in\Psi_1} \big(A_{ij} - P_{ij}^*(w)\big) \left(P_{ij}^*(w) - \widetilde{P}_{ij}(w) \right) \right| / L^*(w) \right\} \\
&\leq \xi^{*-1} \sup_{w\in\mathcal{W}} \left[\sum_{(i,j)\in\Psi_1} \left\{ \sum_{m=1}^{M} w_m \left(P_{(m),ij}^* - \widetilde{P}_{(m),ij} \right) \right\}^2 \right] \\
&\quad + 2 \sup_{w\in\mathcal{W}} \left[\frac{1}{L^*(w)} \left\{ \sum_{(i,j)\in\Psi_1} \big(A_{ij} - P_{ij}^*(w)\big)^2 \right\}^{1/2} \left\{ \sum_{(i,j)\in\Psi_1} \left(P_{ij}^*(w) - \widetilde{P}_{ij}(w) \right)^2 \right\}^{1/2} \right] \\
&\leq \xi^{*-1} \sup_{w\in\mathcal{W}} \left\{ \sum_{(i,j)\in\Psi_1} \sum_{m=1}^{M} w_m \left(P_{(m),ij}^* - \widetilde{P}_{(m),ij} \right)^2 \right\}
\end{aligned}
$$

$$
\begin{aligned}
&+2\sup_{w\in\mathcal{W}}\left[\left\{\frac{1}{L^*(w)}\sum_{(i,j)\in\Psi_1}\left(A_{ij}-P_{ij}^*(w)\right)^2\right\}^{1/2}\left\{\frac{1}{L^*(w)}\sum_{(i,j)\in\Psi_1}\left(P_{ij}^*(w)-\widetilde{P}_{ij}(w)\right)^2\right\}^{1/2}\right]\\
\leq&\xi^{*-1}O_p(NM)+2\left\{\xi^{*-1}O_p\left(NM\right)\right\}^{1/2}\sup_{w\in\mathcal{W}}\left\{\frac{1}{L^*(w)}\sum_{(i,j)\in\Psi_1}\left(A_{ij}-P_{ij}^*(w)\right)^2\right\}^{1/2}\\
=&o_p(1)+2o_p(1)\sup_{w\in\mathcal{W}}\left\{\frac{1}{L^*(w)}\sum_{(i,j)\in\Psi_1}\left(A_{ij}-P_{ij}^*(w)\right)^2\right\}^{1/2}. \qquad \text{(S2.5)}
\end{aligned}
$$

From Lemma 2, we have $P\big(\|A-P\|_2 > C\sqrt{D+\log N}\big) \leq N^{-C_0}$ for any $C_0 > 0$. Then $\|A-P\|_2 = O_p(\max\{D,\log N\}^{1/2})$. Subsequently, we can obtain

$$
\|p_1-a_1\|^2 \leq \|A-P\|_F^2 \leq \operatorname{rank}(A-P)\|A-P\|_2^2 = O_p\left(N\max\{D,\log N\}\right). \qquad \text{(S2.6)}
$$

Thus, together with Assumption 2, we have

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}}\left\{\frac{1}{L^*(w)}\sum_{(i,j)\in\Psi_1}\left(A_{ij}-P_{ij}^*(w)\right)^2\right\}^{1/2}\\
\leq&\sup_{w\in\mathcal{W}}\left[\frac{1}{L^*(w)}\sum_{(i,j)\in\Psi_1}\left\{2\left(A_{ij}-P_{ij}\right)^2+2\left(P_{ij}-P_{ij}^*(w)\right)^2\right\}\right]^{1/2}\\
\leq&\sqrt{2}\xi^{*-1/2}O_p\left(N^{1/2}\max\{D,\log N\}^{1/2}\right)+\sqrt{2}=O_p(1). \qquad \text{(S2.7)}
\end{aligned}
$$

According to (S2.5) and (S2.7), we have

$$
\sup_{w\in\mathcal{W}}\left\{\left|\|a_1-\widetilde{p}_1(w)\|^2-\|a_1-p_1^*(w)\|^2\right|/L^*(w)\right\}=o_p(1). \qquad \text{(S2.8)}
$$

For the first summand in (S2.4), we first denote the lth element of a_1, p_1 and $p^*_{1(m)}$ to be $a_{1,l}$, $p_{1,l}$ and $p^*_{1(m),l}$, respectively. Then it follows that for any $\delta > 0$,

$$
\begin{aligned}
& P\left\{\xi^{*-1} \sup_{w\in\mathcal{W}} \left|p_1^{*\top}(w)(a_1 - p_1)\right| > \delta\right\} \\
=& P\left\{\xi^{*-1} \sup_{w\in\mathcal{W}} \left| \sum_{l=1}^{|\Psi_1|} \left\{p^*_{1,l}(w)(a_{1,l} - p_{1,l})\right\} \right| > \delta\right\} \\
=& P\left\{\xi^{*-1} \sup_{w\in\mathcal{W}} \left| \sum_{l=1}^{|\Psi_1|} \left\{ \sum_{m=1}^{M} w_m p^*_{1(m),l}(a_{1,l} - p_{1,l})\right\}\right| > \delta\right\} \\
\leq& \sum_{m=1}^{M} P\left\{ \left| \sum_{l=1}^{|\Psi_1|} \left(p^*_{1(m),l}a_{1,l} - p^*_{1(m),l}p_{1,l}\right) \right| > \xi^*\delta\right\} \\
\leq& \xi^{*-2}\delta^{-2} \sum_{l=1}^{|\Psi_1|}\sum_{m=1}^{M} \mathrm{var}\left(p^*_{1(m),l}a_{1,l} - p^*_{1(m),l}p_{1,l}\right) \\
\leq& \xi^{*-2}\delta^{-2} M|\Psi_1| = o(1), \qquad \text{(S2.9)}
\end{aligned}
$$

where the third step uses Boole's inequality and the fourth step uses Chebyshev's Inequality. For the fifth step, since $p^*_{1(m),l}$, $p_{1,l} \in [0,1]$ and $a_{1,l} \in \{0,1\}$ for any $m = 1,\ldots,M$ and $l = 1,\ldots,|\Psi_1|$, we have $\left(p^*_{1(m),l}a_{1,l} - p^*_{1(m),l}p_{1,l}\right)^2 \leq 1$. Therefore, we can get $\mathrm{var}\left(p^*_{1(m),l}a_{1,l} - p^*_{1(m),l}p_{1,l}\right) = E\left(p^*_{1(m),l}a_{1,l} - p^*_{1(m),l}p_{1,l}\right)^2 \leq 1$. The sixth step uses the fact that $|\Psi_1|$ has the same order of N^2 and Assumption 2. By (S2.4), (S2.8) and (S2.9), we obtain (S2.2). This completes the proof of Theorem 1.

S2.2 Proof of Theorem 2

This subsection gives the proof of Theorem 2. Similar to (S2.4), we have

$$||\Psi_1|CV^*(w) - L^*(w)| \leq |p_1^*(w)^\top(a_1 - p_1)| + \left|\|a_1 - \widetilde{p}_1(w)\|^2 - \|a_1 - p_1^*(w)\|^2\right|. \tag{S2.10}$$

From (S2.5), (S2.7), and Assumption 1, we can get

$$\begin{aligned}&\left|\|a_1 - \widetilde{p}_1(w)\|^2 - \|a_1 - p_1^*(w)\|^2\right| \\ =&O_p(NM) + O_p(NM^{1/2}\max\{D, \log N\}^{1/2}) + (L^*(w))^{1/2}O_p(N^{1/2}M^{1/2}).\end{aligned} \tag{S2.11}$$

In addition, similar to (S2.9), it is seen that

$$\sup_{w\in\mathcal{W}} \left|p_1^{*\top}(w)(a_1 - p_1)\right| = O_p(NM^{1/2}). \tag{S2.12}$$

According to (S2.10), (S2.11) and (S2.12), we have

$$\begin{aligned}|\Psi_1|CV^*(w) =&L^*(w) + O_p(NM) + O_p(NM^{1/2}\max\{D, \log N\}^{1/2}) \\ &+ (L^*(w))^{1/2}O_p(N^{1/2}M^{1/2}).\end{aligned} \tag{S2.13}$$

Denote $\nu \in \mathbb{R}^M$ to be a weight vector with $\nu_m = 0$ if $m \in \mathcal{T}$ and $\nu_m = w_m/(1-\zeta)$ if $m \notin \mathcal{T}$. It is obvious that if $m \in \mathcal{T}$, we have $P^*_{(m),ij} = P_{(m),ij}$.

In view of Zhang and Liu (2023), we can obtain

$$\begin{aligned} L^*(w) &= \sum_{(i,j)\in\Psi_1} \left\{ \sum_{m=1}^{M} w_m \left(P^*_{(m),ij} - P_{ij} \right) \right\}^2 \\ &= \sum_{(i,j)\in\Psi_1} \left\{ \sum_{m\notin\mathcal{T}} w_m \left(P^*_{(m),ij} - P_{ij} \right) \right\}^2 \\ &= (1-\zeta)^2 \sum_{(i,j)\in\Psi_1} \left\{ \sum_{m\notin\mathcal{T}} \frac{w_m}{1-\zeta} \left(P^*_{(m),ij} - P_{ij} \right) \right\}^2 \\ &= (1-\zeta)^2 \sum_{(i,j)\in\Psi_1} \left\{ \sum_{m=1}^{M} \nu_m \left(P^*_{(m),ij} - P_{ij} \right) \right\}^2 = (1-\zeta)^2 L^*(\nu). \end{aligned} \tag{S2.14}$$

Combining (S2.13) and (S2.14), we have

$$\begin{aligned} |\Psi_1| CV^*(\widehat{w}) =& (1-\widehat{\zeta})^2 L^*(\widehat{\nu}) + (1-\widehat{\zeta})(L^*(\widehat{\nu}))^{1/2} O_p(N^{1/2}M^{1/2}) \\ &+ O_p(NM) + O_p(NM^{1/2}\max\{D, \log N\}^{1/2}), \end{aligned} \tag{S2.15}$$

where $\widehat{\nu} = (\widehat{\nu}_1, \ldots, \widehat{\nu}_M)^\top \in \mathbb{R}^M$ and $\widehat{\nu}_m = \widehat{w}_m/(1-\widehat{\zeta})$ for $m = 1, \ldots, M$. Let $\overline{w} \in \mathbb{R}^M$ be a weight vector with $\sum_{m\in\mathcal{T}} \overline{w}_m = 1$. By (S2.14), we have $L^*(\overline{w}) = 0$. Therefore, we have

$$|\Psi_1| CV^*(\overline{w}) = O_p(NM) + O_p(NM^{1/2}\max\{D, \log N\}^{1/2}). \tag{S2.16}$$

Since $\widehat{w}$ minimizes $CV^*(w)$, by (S2.15) and (S2.16), we have

$$(1-\widehat{\zeta})^2 L^*(\widehat{\nu}) + (1-\widehat{\zeta})(L^*(\widehat{\nu}))^{1/2} O_p(N^{1/2}M^{1/2}) + O_p(NM)$$
$$+ O_p(NM^{1/2}\max\{D, \log N\}^{1/2}) \leq O_p(NM) + O_p(NM^{1/2}\max\{D, \log N\}^{1/2}).$$

It follows that

$$(1-\widehat{\zeta})^2 \left(\inf_{w\in\mathcal{W}^s} L^*(w)\right) + (1-\widehat{\zeta}) \left(\inf_{w\in\mathcal{W}^s} L^*(w)\right)^{1/2} O_p(N^{1/2}M^{1/2}) + O_p(NM)$$
$$+ O_p(NM^{1/2}\max\{D, \log N\}^{1/2}) \leq O_p(NM) + O_p(NM^{1/2}\max\{D, \log N\}^{1/2}). \tag{S2.17}$$

Dividing both sides of the (S2.17) by $\inf_{w\in\mathcal{W}^s} L^*(w)$, we obtain

$$\begin{aligned}
&(1-\widehat{\zeta})^2 + (1-\widehat{\zeta}) \left(\inf_{w\in\mathcal{W}^s} L^*(w)\right)^{-1/2} O_p(N^{1/2}M^{1/2}) + \left(\inf_{w\in\mathcal{W}^s} L^*(w)\right)^{-1} O_p(NM) \\
&+ \left(\inf_{w\in\mathcal{W}^s} L^*(w)\right)^{-1/2} O_p(N^{1/2}M^{1/2}) \left(\inf_{w\in\mathcal{W}^s} L^*(w)\right)^{-1/2} O_p(N^{1/2}\max\{D, \log N\}^{1/2}) \\
&\leq \left(\inf_{w\in\mathcal{W}^s} L^*(w)\right)^{-1} \left\{O_p(NM) + O_p(NM^{1/2}\max\{D, \log N\}^{1/2})\right\}. \quad \text{(S2.18)}
\end{aligned}$$

Together with Assumption 3, we obtain $\widehat{\zeta} \to 1$ in probability. This completes the proof of Theorem 2.

S2.3 Proof of Theorem 3

This subsection gives the proof of Theorem 3. Denote $\epsilon = \xi^{1/2}|\Psi_1|^{-1/2+\kappa}$. To verify Theorem 3, following Liao and Zou (2020) and Liao et al. (2021), it suffices to show that there exists a constant C_1 such that, for the $M \times 1$ vector $r = (r_1, \ldots, r_M)^\top$,

$$\lim_{|\Psi_1|\to\infty} P\left\{ \inf_{\|r_0\|=C_1, (w^0+\epsilon r)\in\mathcal{W}} CV(w^0+\epsilon r) > CV(w^0) \right\} = 1, \qquad \text{(S2.19)}$$

which indicates that there exists a minimizer $\widehat{w}$ in the bounded closed domain $\{w^0 + \epsilon r : \|r\| \le C_1, (w^0 + \epsilon r) \in \mathcal{W}\}$ such that $\|\widehat{w} - w^0\| = O_p(\epsilon)$.

Denote $\Phi_1 = (\widehat{p}_{1(1)} - \widetilde{p}_{1(1)}, \ldots, \widehat{p}_{1(M)} - \widetilde{p}_{1(M)})$. Then we can write $\|\widehat{p}_1(w) - \widetilde{p}_1(w)\|^2 = w^\top \Phi_1^\top \Phi_1 w$, and $(a_1 - \widehat{p}_1(w))^\top (\widehat{p}_1(w) - \widetilde{p}_1(w)) = w^\top \Phi_1^\top e_1 + w^\top \Phi_1^\top \Omega_1 w$, where $e_1 = a_1 - p_1$. Thus, we can decompose $CV(w)$ as

$$\begin{aligned} CV(w) &= \frac{1}{|\Psi_1|} \sum_{k=1}^{K} \left\| A^{[k]} - \widetilde{P}^{[k]}(w) \right\|_F^2 = \|a_1 - \widetilde{p}_1(w)\|^2 / |\Psi_1| \\ &= \left\{ \|a_1 - \widehat{p}_1(w)\|^2 + \|\widehat{p}_1(w) - \widetilde{p}_1(w)\|^2 + 2\left(a_1 - \widehat{p}_1(w)\right)^\top \left(\widehat{p}_1(w) - \widetilde{p}_1(w)\right) \right\} / |\Psi_1| \\ &= \left\{ \|a_1 - \widehat{p}_1(w)\|^2 + w^\top \Phi_1^\top \Phi_1 w + 2 w^\top \Phi_1^\top e_1 + 2 w^\top \Phi_1^\top \Omega_1 w \right\} / |\Psi_1|. \end{aligned}$$

Note that

$$\begin{aligned}
&\|a_1 - \widehat{p}_1(w^0 + \epsilon r)\|^2 - \|a_1 - \widehat{p}_1(w^0)\|^2 \\
=&\widehat{p}_1(w^0 + \epsilon r)^\top \widehat{p}_1(w^0 + \epsilon r) - \widehat{p}_1(w^0)^\top \widehat{p}_1(w^0) - 2a_1^\top \left\{\widehat{p}_1(w^0 + \epsilon r) - \widehat{p}_1(w^0)\right\} \\
=&(w^0 + \epsilon r)^\top \Lambda_1^\top \Lambda_1 (w^0 + \epsilon r) - w^{0\top} \Lambda_1^\top \Lambda_1 w^0 - 2a_1^\top \left\{\Lambda_1(w^0 + \epsilon r) - \Lambda_1 w^0\right\} \\
=&\epsilon^2 r^\top \Lambda r + 2(\Lambda_1 w^0 - a_1)^\top \Lambda_1 \epsilon r.
\end{aligned}$$

As a result,

$$\begin{aligned}
&CV(w^0 + \epsilon r) - CV(w^0) \\
=&\left\{\|a_1 - \widehat{p}_1(w^0 + \epsilon r)\|^2 - \|a_1 - \widehat{p}_1(w^0)\|^2 + \epsilon^2 r^\top \Phi_1^\top \Phi_1 r + 2\epsilon r^\top \Phi_1^\top e_1 \right. \\
&\left. + 2\epsilon^2 r^\top \Phi_1^\top \Omega_1 r + 2\epsilon r^\top \Phi_1^\top \Omega_1 w^0 + 2\epsilon w^{0\top} \Phi_1^\top \Omega_1 r + 2\epsilon r^\top \Phi_1^\top \Phi_1 w^0 \right\} / |\Psi_1| \\
=&\left\{\epsilon^2 r^\top \Lambda r + 2\epsilon(\Lambda_1 w^0 - a_1)^\top \Lambda_1 r + \epsilon^2 r^\top \Phi_1^\top \Phi_1 r + 2\epsilon r^\top \Phi_1^\top e_1 + 2\epsilon^2 r^\top \Phi_1^\top \Omega_1 r \right. \\
&\left. + 2\epsilon r^\top \Phi_1^\top \Omega_1 w^0 + 2\epsilon w^{0\top} \Phi_1^\top \Omega_1 r + 2\epsilon r^\top \Phi_1^\top \Phi_1 w^0 \right\} / |\Psi_1|. \qquad \text{(S2.20)}
\end{aligned}$$

It is obvious that $\epsilon^2 r^\top \Phi_1^\top \Phi_1 r \geq 0$ for any r. Besides, under Assumption 4, we have

$$\epsilon^2 r^\top \Lambda r > \rho_1 \epsilon^2 |\Psi_1| \|r\|^2 > 0 \qquad \text{(S2.21)}$$

in probability tending to 1. In the following, we show that the remaining six terms of (S2.20) are asymptotically dominated by $\epsilon^2 r^\top \Lambda r$.

We first consider the item $|\epsilon(\Lambda_1 w^0 - a_1)^\top \Lambda_1 r|$. According to the def-

inition of ξ and w^0, we have $E(\|\Lambda_1 w^0 - p_1\|^2) = \xi$. Therefore, $\|\Lambda_1 w^0 - p_1\| = O_p(\xi^{1/2})$. From Assumption 4, it is clear that $\|\Lambda_1\|_2 = \lambda_{\max}^{1/2}(\Lambda) = O_p(|\Psi_1|^{1/2} M^{1/2})$. Besides, from (S2.6), we have

$$\|e_1\|^2 = \|a_1 - p_1\|^2 = O_p\left(N \max\{D, \log N\}\right). \tag{S2.22}$$

Thus,

$$\begin{aligned}
&|\epsilon(\Lambda_1 w^0 - a_1)^\top \Lambda_1 r| \\
=&|\epsilon(\Lambda_1 w^0 - p_1 + p_1 - a_1)^\top \Lambda_1 r| \\
\leq&|\epsilon(\Lambda_1 w^0 - p_1)^\top \Lambda_1 r| + |\epsilon(p_1 - a_1)^\top \Lambda_1 r| \\
\leq&\epsilon\|\Lambda_1 w^0 - p_1\|\|\Lambda_1\|_2\|r\| + \epsilon\|p_1 - a_1\|\|\Lambda_1\|_2\|r\| \\
=&O_p\left(\epsilon\xi^{1/2}|\Psi_1|^{1/2}M^{1/2}\right)\|r\| + O_p\left(\epsilon N^{1/2}\max\{D, \log N\}^{1/2}|\Psi_1|^{1/2}M^{1/2}\right)\|r\|.
\end{aligned} \tag{S2.23}$$

Notice that

$$\begin{aligned}
&\|\Phi_1\|_2^2 \leq \mathrm{tr}(\Phi_1^\top \Phi_1) = \sum_{m=1}^{M} \left\|\widetilde{p}_{1(m)} - \widehat{p}_{1(m)}\right\|^2 \\
\leq&\sum_{m=1}^{M} 2\left(\left\|\widetilde{p}_{1(m)} - p^*_{1(m)}\right\|^2 + \left\|p^*_{1(m)} - \widehat{p}_{1(m)}\right\|^2\right) = O_p\left(NM^2\right),
\end{aligned} \tag{S2.24}$$

where the last step is due to Assumption 1. Based on (S2.24) and (S2.22),

we can obtain

$$|\epsilon r^\top \Phi_1^\top e_1| \leq \epsilon \|e_1\| \|\Phi_1\|_2 \|r\| = O_p\left(\epsilon N \max\{D, \log N\}^{1/2} M\right) \|r\|. \quad \text{(S2.25)}$$

Observe from Assumption 5 that $\lambda_{\max}^2(\Omega_1) = \lambda_{\max}(\Omega_1^\top \Omega_1) = \lambda_{\max}(\Omega) = O_p(|\Psi_1|M)$, and hence,

$$|\epsilon^2 r^\top \Phi_1^\top \Omega_1 r| \leq \epsilon^2 \lambda_{\max}(\Omega_1) \|\Phi_1\|_2 \|r\|^2 = O_p\left(\epsilon^2 |\Psi_1|^{1/2} N^{1/2} M^{3/2}\right) \|r\|^2. \quad \text{(S2.26)}$$

Noting that $E(\|\Omega_1 w^0\|^2) = E(\|p_1 - \widehat{p}_1(w^0)\|^2) = \xi$, we have $\|\Omega_1 w^0\| = O_p\left(\xi^{1/2}\right)$, which implies that

$$|\epsilon r^\top \Phi_1^\top \Omega_1 w^0| \leq \epsilon \|\Phi_1\|_2 \|\Omega_1 w^0\| \|r\| = O_p\left(\epsilon N^{1/2} M \xi^{1/2}\right) \|r\|. \quad \text{(S2.27)}$$

Similar to the process of getting (S2.24), we can obtain $\|\widehat{p}_{1(m)} - \widetilde{p}_{1(m)}\|^2 = O_p(NM)$, uniformly in m $(m = 1, ..., M)$. Thus,

$$\|\Phi_1 w^0\|^2 = \left\| \sum_{m=1}^{M} w_m^0 (\widehat{p}_{1(m)} - \widetilde{p}_{1(m)}) \right\|^2 \leq \sum_{m=1}^{M} w_m^0 \|\widehat{p}_{1(m)} - \widetilde{p}_{1(m)}\|^2 = O_p(NM), \quad \text{(S2.28)}$$

which indicates that

$$|\epsilon w^{0\top}\Phi_1^\top\Omega_1 r| \le \epsilon\lambda_{\max}(\Omega_1)\|\Phi_1 w^0\|\|r\| = O_p\left(\epsilon|\Psi_1|^{1/2}N^{1/2}M\right)\|r\|. \quad \text{(S2.29)}$$

Similarly, utilizing (S2.24) and (S2.28) again, we obtain

$$|\epsilon r^\top\Phi_1^\top\Phi_1 w^0| \le \epsilon\|\Phi_1 w^0\|\|\Phi_1\|_2\|r\| = O_p\left(\epsilon NM^{3/2}\right)\|r\|. \quad \text{(S2.30)}$$

Finally, recalling $\epsilon = \xi^{1/2}|\Psi_1|^{-1/2+\kappa}$, since $|\Psi_1|$ has the order of N^2, by Assumption 6, we see that

$$\frac{\epsilon\xi^{1/2}|\Psi_1|^{1/2}M^{1/2}}{\epsilon^2|\Psi_1|} = \frac{M^{1/2}}{|\Psi_1|^\kappa} = o(1), \quad \text{(S2.31)}$$

$$\frac{\epsilon N^{1/2}\max\{D,\log N\}^{1/2}|\Psi_1|^{1/2}M^{1/2}}{\epsilon^2|\Psi_1|} = \frac{N^{1/2}\max\{D,\log N\}^{1/2}M^{1/2}}{\xi^{1/2}|\Psi_1|^\kappa} = o(1), \quad \text{(S2.32)}$$

$$\frac{\epsilon^2|\Psi_1|^{1/2}N^{1/2}M^{3/2}}{\epsilon^2|\Psi_1|} = \frac{N^{1/2}M^{3/2}}{|\Psi_1|^{1/2}} = o(1), \quad \text{(S2.33)}$$

$$\frac{\epsilon N^{1/2}M\xi^{1/2}}{\epsilon^2|\Psi_1|} = \frac{N^{1/2}M}{|\Psi_1|^{1/2+\kappa}} = o(1), \quad \text{(S2.34)}$$

$$\frac{\epsilon|\Psi_1|^{1/2}N^{1/2}M}{\epsilon^2|\Psi_1|} = \frac{N^{1/2-2\kappa}M}{\xi^{1/2}} = o(1), \quad \text{(S2.35)}$$

$$\frac{\epsilon N\max\{D,\log N\}^{1/2}M}{\epsilon N^{1/2}\max\{D,\log N\}^{1/2}|\Psi_1|^{1/2}M^{1/2}} = \frac{N^{1/2}M^{1/2}}{|\Psi_1|^{1/2}} = o(1), \quad \text{(S2.36)}$$

and

$$\frac{\epsilon N M^{3/2}}{\epsilon |\Psi_1|^{1/2} N^{1/2} M} = \frac{N^{1/2} M^{1/2}}{|\Psi_1|^{1/2}} = o(1). \tag{S2.37}$$

Therefore, from (S2.23), (S2.25), (S2.26), (S2.27), (S2.29) and (S2.30), and using (S2.21), (S2.31)- (S2.37) we can prove that the remaining six terms of (S2.20) are asymptotically dominated by $\epsilon^2 r^\top \Lambda r$. Since $CV(w)$ is strictly convex over $\mathcal{W}$ under Assumption 4, the minimizer within $\{w^0 + \epsilon r : \|r\| \leq C_1, (w^0 + \epsilon r) \in \mathcal{W}\}$ is precisely the unique global minimizer of $CV(w)$ over $\mathcal{W}$. This completes the proof of Theorem 3.

S2.4 Proof of Theorem S1

This subsection gives the proof of Theorem S1. Let

$$CV_\dagger^*(w) = CV_\dagger(w) - \sum_{t=1}^{T} \sum_{k=1}^{K} \left(\|A^{(t)[k]}\|_F^2 - \|P^{(t)[k]}\|_F^2 \right) / (T|\Psi_1|),$$

where $P^{(t)[k]} = P^{(t)} \circ S^{[k]}$ and the second term is unrelated to w. Therefore,

$$\widehat{w}_\dagger = \text{argmin}_{w \in \mathcal{W}} CV_\dagger(w) = \text{argmin}_{w \in \mathcal{W}} T|\Psi_1| CV_\dagger^*(w).$$

According to Lemma 1, Theorem S1 is valid if the following conditions hold:

$$\sup_{w \in \mathcal{W}} \frac{|L_\dagger(w) - L_\dagger^*(w)|}{L_\dagger^*(w)} = o_p(1), \tag{S2.38}$$

and

$$\sup_{w\in\mathcal{W}}\frac{|T|\Psi_1|CV_\dagger^*(w)-L_\dagger^*(w)|}{L_\dagger^*(w)}=o_p(1). \tag{S2.39}$$

We first consider (S2.38). It can be shown that

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}\frac{|L_\dagger(w)-L_\dagger^*(w)|}{L_\dagger^*(w)}\\
=&\sup_{w\in\mathcal{W}}\left[\left|\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left\{\left(\widehat{P}_{ij}^{(t)}(w)-P_{ij}^{(t)}\right)^2-\left(P_{ij}^{(t)*}(w)-P_{ij}^{(t)}\right)^2\right\}\right|/L_\dagger^*(w)\right]\\
\le&\sup_{w\in\mathcal{W}}\left[\left|\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left\{\left(\widehat{P}_{ij}^{(t)}(w)-P_{ij}^{(t)*}(w)\right)^2+2\left(\widehat{P}_{ij}^{(t)}(w)-P_{ij}^{(t)*}(w)\right)\left(P_{ij}^{(t)*}(w)-P_{ij}^{(t)}\right)\right\}\right|/L_\dagger^*(w)\right]\\
\le&\xi_\dagger^{*-1}\sup_{w\in\mathcal{W}}\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left\{\sum_{m=1}^{M}w_m\left(\widehat{P}_{(m),ij}^{(t)}-P_{(m),ij}^{(t)*}\right)\right\}^2\\
&+2\sup_{w\in\mathcal{W}}\left[\frac{1}{L_\dagger^*(w)}\left\{\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(\widehat{P}_{ij}^{(t)}(w)-P_{ij}^{(t)*}(w)\right)^2\right\}^{1/2}\left\{\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(P_{ij}^{(t)*}(w)-P_{ij}^{(t)}\right)^2\right\}^{1/2}\right]\\
\le&\xi_\dagger^{*-1}\sup_{w\in\mathcal{W}}\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\sum_{m=1}^{M}w_m\left(\widehat{P}_{(m),ij}^{(t)}-P_{(m),ij}^{(t)*}\right)^2\\
&+2\xi_\dagger^{*-1/2}\sup_{w\in\mathcal{W}}\left\{\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(\widehat{P}_{ij}^{(t)}(w)-P_{ij}^{(t)*}(w)\right)^2\right\}^{1/2}\\
=&\xi_\dagger^{*-1}O_p(NMT)+2\xi_\dagger^{*-1/2}O_p(N^{1/2}M^{1/2}T^{1/2})=o_p(1),
\end{aligned}\tag{S2.40}$$

where the fifth step uses Assumption S1, and the sixth step uses Assumption

S3. Thus, we obtain (S2.38). We next verify (S2.39). Observe that

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}}\left\{\left|T|\Psi_1|CV_\dagger^*(w)-L_\dagger^*(w)\right|/L_\dagger^*(w)\right\}\\
=&\sup_{w\in\mathcal{W}}\left[\frac{1}{L_\dagger^*(w)}\left|\sum_{t=1}^{T}\left\{\left\|a_1^{(t)}-\widetilde{p}_1^{(t)}(w)\right\|^2-\left(a_1^{(t)}-p_1^{(t)}\right)^\top\left(a_1^{(t)}+p_1^{(t)}\right)-\left\|p_1^{(t)*}(w)-p_1^{(t)}\right\|^2\right\}\right|\right]\\
\le&\sup_{w\in\mathcal{W}}\left[\frac{1}{L_\dagger^*(w)}\left|\sum_{t=1}^{T}\left\{\left\|a_1^{(t)}-p_1^{(t)*}(w)\right\|^2-\left(a_1^{(t)}-p_1^{(t)}\right)^\top\left(a_1^{(t)}+p_1^{(t)}\right)-\left\|p_1^{(t)*}(w)-p_1^{(t)}\right\|^2\right\}\right|\right]\\
&+\sup_{w\in\mathcal{W}}\left[\frac{1}{L_\dagger^*(w)}\left|\sum_{t=1}^{T}\left\{\left\|a_1^{(t)}-\widetilde{p}_1^{(t)}(w)\right\|^2-\left\|a_1^{(t)}-p_1^{(t)*}(w)\right\|^2\right\}\right|\right]\\
\le&2\xi_\dagger^{*-1}\sup_{w\in\mathcal{W}}\left|\sum_{t=1}^{T}p_1^{(t)*}(w)^\top\left(a_1^{(t)}-p_1^{(t)}\right)\right|\\
&+\sup_{w\in\mathcal{W}}\left[\frac{1}{L_\dagger^*(w)}\left|\sum_{t=1}^{T}\left\{\left\|a_1^{(t)}-\widetilde{p}_1^{(t)}(w)\right\|^2-\left\|a_1^{(t)}-p_1^{(t)*}(w)\right\|^2\right\}\right|\right]. \quad \text{(S2.41)}
\end{aligned}
$$

By Assumptions 7 and 9, we have

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}}\left[\frac{1}{L_\dagger^*(w)}\left|\sum_{t=1}^{T}\left\{\left\|a_1^{(t)}-\widetilde{p}_1^{(t)}(w)\right\|^2-\left\|a_1^{(t)}-p_1^{(t)*}(w)\right\|^2\right\}\right|\right]\\
=&\sup_{w\in\mathcal{W}}\left[\frac{1}{L_\dagger^*(w)}\left|\sum_{t=1}^{T}\left\{\|p_1^{(t)*}(w)-\widetilde{p}_1^{(t)}(w)\|^2+2\big(a_1^{(t)}-p_1^{(t)*}(w)\big)^\top\big(p_1^{(t)*}(w)-\widetilde{p}_1^{(t)}(w)\big)\right\}\right|\right]\\
\le&\xi_\dagger^{*-1}\sup_{w\in\mathcal{W}}\left\{\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(P_{ij}^{(t)*}(w)-\widetilde{P}_{ij}^{(t)}(w)\right)^2\right\}\\
&+2\sup_{w\in\mathcal{W}}\left|\frac{1}{L_\dagger^*(w)}\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(A_{ij}^{(t)}-P_{ij}^{(t)*}(w)\right)\left(P_{ij}^{(t)*}(w)-\widetilde{P}_{ij}^{(t)}(w)\right)\right|\\
\le&\xi_\dagger^{*-1}\sup_{w\in\mathcal{W}}\left[\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left\{\sum_{m=1}^{M}w_m\left(P_{(m),ij}^{(t)*}-\widetilde{P}_{(m),ij}^{(t)}\right)\right\}^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+2\sup_{w\in\mathcal{W}}\left[\frac{1}{L_\dagger^*(w)}\left\{\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(A_{ij}^{(t)}-P_{ij}^{(t)*}(w)\right)^2\right\}^{1/2}\left\{\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(P_{ij}^{(t)*}(w)-\widetilde{P}_{ij}^{(t)}(w)\right)^2\right\}^{1/2}\right]\\
\le&\xi_\dagger^{*-1}\sup_{w\in\mathcal{W}}\left[\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\sum_{m=1}^{M}w_m\left(P_{(m),ij}^{(t)*}-\widetilde{P}_{(m),ij}^{(t)}\right)^2\right]\\
&+2\sup_{w\in\mathcal{W}}\left[\left\{\frac{1}{L_\dagger^*(w)}\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(A_{ij}^{(t)}-P_{ij}^{(t)*}(w)\right)^2\right\}^{1/2}\left\{\frac{1}{L_\dagger^*(w)}\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(P_{ij}^{(t)*}(w)-\widetilde{P}_{ij}^{(t)}(w)\right)^2\right\}^{1/2}\right]\\
\le&\xi_\dagger^{*-1}O_p(NMT)+2\left\{\xi_\dagger^{*-1}O_p\left(NMT\right)\right\}^{1/2}\sup_{w\in\mathcal{W}}\left\{\frac{1}{L_\dagger^*(w)}\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(A_{ij}^{(t)}-P_{ij}^{(t)*}(w)\right)^2\right\}^{1/2}\\
=&o_p(1)+2o_p(1)\sup_{w\in\mathcal{W}}\left\{\frac{1}{L_\dagger^*(w)}\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left(A_{ij}^{(t)}-P_{ij}^{(t)*}(w)\right)^2\right\}^{1/2}. \quad \text{(S2.42)}
\end{aligned}
$$

From Lemma 2 and Assumption S2, we have $P(\|A^{(t)}-P^{(t)}\|_2 > C_\dagger\sqrt{D_\dagger+\log N}) \le N^{-C_{0\dagger}}$ for any $C_{0\dagger} > 0$. Under Assumption S2, there exists a constant $C_{0\dagger} > 0$ such that

$$
\begin{aligned}
&P\left(\sup_t\left\|A^{(t)}-P^{(t)}\right\|_2 > C_\dagger\sqrt{D_\dagger+\log N}\right)\\
\le&\sum_{t=1}^{T}P\left(\left\|A^{(t)}-P^{(t)}\right\|_2 > C_\dagger\sqrt{D_\dagger+\log N}\right) \le TN^{-C_{0\dagger}} = o(1). \quad \text{(S2.43)}
\end{aligned}
$$

Therefore, we have $\sup_t\left\|A^{(t)}-P^{(t)}\right\|_2 = O_p(\max\{D_\dagger,\log N\}^{1/2})$. Then, we can get

$$
\sup_t\left\|p_1^{(t)}-a_1^{(t)}\right\|^2 \le \sup_t\left\|A^{(t)}-P^{(t)}\right\|_F^2
$$

$$\leq \sup_t \left\{ \operatorname{rank}\left(A^{(t)} - P^{(t)}\right) \left\| A^{(t)} - P^{(t)} \right\|_2^2 \right\} = O_p\left(N \max\{D_\dagger, \log N\}\right).$$

Then we have

$$\sum_{t=1}^{T} \|p_1^{(t)} - a_1^{(t)}\|^2 \leq T \sup_t \left\|p_1^{(t)} - a_1^{(t)}\right\|^2 = O_p\left(N \max\{D_\dagger, \log N\} T\right). \quad \text{(S2.44)}$$

By Assumption S3, we have

$$\begin{aligned}
&\sup_{w \in \mathcal{W}} \left\{ \frac{1}{L_\dagger^*(w)} \sum_{t=1}^{T} \sum_{(i,j)\in\Psi_1} \left(A_{ij}^{(t)} - P_{ij}^{(t)*}(w) \right)^2 \right\}^{1/2} \\
\leq &\sup_{w \in \mathcal{W}} \left[\frac{1}{L_\dagger^*(w)} \sum_{t=1}^{T} \sum_{(i,j)\in\Psi_1} \left\{ 2\left(A_{ij}^{(t)} - P_{ij}^{(t)} \right)^2 + 2\left(P_{ij}^{(t)} - P_{ij}^{(t)*}(w) \right)^2 \right\} \right]^{1/2} \\
\leq &\sqrt{2}\xi_\dagger^{*-1/2} O_p\left(N^{1/2} \max\{D_\dagger, \log N\}^{1/2} T^{1/2}\right) + \sqrt{2} = O(1). \qquad \text{(S2.45)}
\end{aligned}$$

According to (S2.42) and (S2.45), we have

$$\sup_{w \in \mathcal{W}} \left[\frac{1}{L_\dagger^*(w)} \left| \sum_{t=1}^{T} \left\{ \left\| a_1^{(t)} - \widetilde{p}_1^{(t)}(w) \right\|^2 - \left\| a_1^{(t)} - p_1^{(t)*}(w) \right\|^2 \right\} \right| \right] = o_p(1). \quad \text{(S2.46)}$$

Denote the lth element of $a_1^{(t)}$, $p_1^{(t)}$ and $p_1^{(t)*}$ to be $a_{1,l}^{(t)}$, $p_{1,l}^{(t)}$ and $p_{1,l}^{(t)*}$. Then, by Assumption S3 and considering that $|\Psi_1|$ has the order of N^2, for any

$\delta > 0$,

$$
\begin{aligned}
& P\left\{\xi_{\dagger}^{*-1} \sup_{w \in \mathcal{W}} \left| \sum_{t=1}^{T} p_1^{(t)*}(w)^{\top} \left(a_1^{(t)} - p_1^{(t)}\right) \right| > \delta \right\} \\
=& P\left\{\xi_{\dagger}^{*-1} \sup_{w \in \mathcal{W}} \left| \sum_{t=1}^{T} \sum_{l=1}^{|\Psi_1|} \sum_{m=1}^{M} w_m p_{1(m),l}^{(t)*} \left(a_{1,l}^{(t)} - p_{1,l}^{(t)}\right) \right| > \delta \right\} \\
\leq & \sum_{m=1}^{M} P\left\{ \left| \sum_{t=1}^{T} \sum_{l=1}^{|\Psi_1|} \left(p_{1(m),l}^{(t)*} a_{1,l}^{(t)} - p_{1(m),l}^{(t)*} p_{1,l}^{(t)}\right) \right| > \xi_{\dagger}^{*} \delta \right\} \\
\leq & \xi_{\dagger}^{*-2} \delta^{-2} \sum_{t=1}^{T} \sum_{l=1}^{|\Psi_1|} \sum_{m=1}^{M} \operatorname{var}\left(p_{1(m),l}^{(t)*} a_{1,l}^{(t)} - p_{1(m),l}^{(t)*} p_{1,l}^{(t)}\right) \\
\leq & \xi_{\dagger}^{*-2} \delta^{-2} TM|\Psi_1| = o(1).
\end{aligned} \tag{S2.47}
$$

By (S2.41) , (S2.46) and (S2.47), we obtain (S2.39). This completes the proof of Theorem S1.

S2.5 Proof of Theorem S2

This subsection gives the proof of Theorem S2. According to (S2.41), we can get

$$
\begin{aligned}
\left|T|\Psi_1|CV_{\dagger}^{*}(w) - L_{\dagger}^{*}(w)\right| \leq & \left| \sum_{t=1}^{T} p_1^{(t)*}(w)^{\top} \left(a_1^{(t)} - p_1^{(t)}\right) \right| \\
& + \left| \sum_{t=1}^{T} \left\{ \left\| a_1^{(t)} - \widetilde{p}_1^{(t)}(w) \right\|^2 - \left\| a_1^{(t)} - p_1^{(t)*}(w) \right\|^2 \right\} \right|.
\end{aligned} \tag{S2.48}
$$

Under Assumptions 7 and 8, we have

$$\left|\sum_{t=1}^{T}\left\{\left\|a_1^{(t)}-\widetilde{p}_1^{(t)}(w)\right\|^2-\left\|a_1^{(t)}-p_1^{(t)*}(w)\right\|^2\right\}\right|$$
$$=O_p(NMT)+O_p(NM^{1/2}T\max\{D_\dagger,\log N\}^{1/2})+(L_\dagger^*(w))^{1/2}O_p(N^{1/2}M^{1/2}T^{1/2}). \tag{S2.49}$$

Similar to (S2.47), we have

$$\sup_{w\in\mathcal{W}}\left|\sum_{t=1}^{T}p_1^{(t)*}(w)^\top\left(a_1^{(t)}-p_1^{(t)}\right)\right|=O_p(NM^{1/2}T^{1/2}). \tag{S2.50}$$

Based on (S2.48), (S2.49) and (S2.50), we have

$$\begin{aligned}T|\Psi_1|CV_\dagger^*(w)=&L_\dagger^*(w)+O_p(NMT)+O_p(NM^{1/2}T\max\{D_\dagger,\log N\}^{1/2})\\&+(L_\dagger^*(w))^{1/2}O_p(N^{1/2}M^{1/2}T^{1/2}).\end{aligned} \tag{S2.51}$$

It is easy to see that $P_{(m),ij}^{(t)*}=P_{(m),ij}^{(t)}$ if $m\in\mathcal{T}$. Hence, we have

$$\begin{aligned}L_\dagger^*(w)=&\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left\{\sum_{m=1}^{M}w_m\left(P_{(m),ij}^{(t)*}-P_{ij}^{(t)}\right)\right\}^2\\=&\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left\{\sum_{m\notin\mathcal{T}}w_m\left(P_{(m),ij}^{(t)*}-P_{ij}^{(t)}\right)\right\}^2\\=&(1-\zeta)^2\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left\{\sum_{m\notin\mathcal{T}}\frac{w_m}{1-\zeta}\left(P_{(m),ij}^{(t)*}-P_{ij}^{(t)}\right)\right\}^2\end{aligned}$$

$$=(1-\zeta)^2\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_1}\left\{\sum_{m=1}^{M}\nu_m\left(P_{(m),ij}^{(t)*}-P_{ij}^{(t)}\right)\right\}^2=(1-\zeta)^2L_\dagger^*(\nu). \tag{S2.52}$$

Together with (S2.51), we have

$$\begin{aligned}T|\Psi_1|CV_\dagger^*(\widehat{w}_\dagger)=&(1-\widehat{\zeta}_\dagger)^2L_\dagger^*(\widehat{\nu}_\dagger)+(1-\widehat{\zeta}_\dagger)(L_\dagger^*(\widehat{\nu}_\dagger))^{1/2}O_p(N^{1/2}M^{1/2}T^{1/2})\\&+O_p(NMT)+O_p(NM^{1/2}T\max\{D_\dagger,\log N\}^{1/2}),\end{aligned} \tag{S2.53}$$

where $\widehat{\nu}_\dagger=\widehat{w}_\dagger/(1-\widehat{\zeta}_\dagger)$. It is obvious that $L_\dagger^*(\overline{w})=0$. By (S2.51), we have

$$T|\Psi_1|CV_\dagger^*(\overline{w})=O_p(NMT)+O_p(NM^{1/2}T\max\{D_\dagger,\log N\}^{1/2}). \tag{S2.54}$$

Because $\widehat{w}_\dagger$ minimizes $CV_\dagger^*(w)$, according to (S2.53) and (S2.54), we have

$$\begin{aligned}&(1-\widehat{\zeta}_\dagger)^2L_\dagger^*(\widehat{\nu}_\dagger)+(1-\widehat{\zeta}_\dagger)(L_\dagger^*(\widehat{\nu}_\dagger))^{1/2}O_p(N^{1/2}M^{1/2}T^{1/2})\\&+O_p(NMT)+O_p(NM^{1/2}T\max\{D_\dagger,\log N\}^{1/2})\\\le&O_p(NMT)+O_p(NM^{1/2}T\max\{D_\dagger,\log N\}^{1/2}).\end{aligned}$$

It is easy to see that

$$(1-\widehat{\zeta}_\dagger)^2\left(\inf_{w\in\mathcal{W}^s}L_\dagger^*(w)\right)+(1-\widehat{\zeta}_\dagger)\left(\inf_{w\in\mathcal{W}^s}L_\dagger^*(w)\right)^{1/2}O_p(N^{1/2}M^{1/2}T^{1/2})$$

$$+ O_p(NMT) + O_p(NM^{1/2}T\max\{D_\dagger, \log N\}^{1/2})$$

$$\leq O_p(NMT) + O_p(NM^{1/2}T\max\{D_\dagger, \log N\}^{1/2}).$$

Similar to (S2.18), and by Assumption S4, we can obtain $\widehat{\zeta}_\dagger \to 1$ in probability. This completes the proof of Theorem S2.

S2.6 Proof of Theorem S3

This subsection gives the proof of Theorem S3. Denote $\epsilon_\dagger = \xi_\dagger^{1/2} T^{-1/2} |\Psi_1|^{-1/2+\kappa_\dagger}$. To verify Theorem S3, following Liao and Zou (2020) and Liao et al. (2021), we only need to show that there exists a constant $C_{0\dagger}$ such that, for the $M \times 1$ vector $r = (r_1, \ldots, r_M)^\top$,

$$\lim_{|\Psi_1|\to\infty} P\left\{ \inf_{\|r_0\|=C_{0\dagger}, (w_\dagger^0+\epsilon_\dagger r)\in\mathcal{W}} CV_\dagger(w_\dagger^0 + \epsilon_\dagger r) > CV_\dagger(w_\dagger^0) \right\} = 1, \quad \text{(S2.55)}$$

which indicates that there exists a minimizer $\widehat{w}_\dagger$ in the bounded closed domain $\left\{w_\dagger^0 + \epsilon_\dagger r : \|r\| \leq C_{0\dagger}, (w_\dagger^0 + \epsilon_\dagger r) \in \mathcal{W}\right\}$ such that $\left\|\widehat{w}_\dagger - w_\dagger^0\right\| = O_p(\epsilon_\dagger)$. Denote $\Phi_1^{(t)} = \left(\widehat{p}_{1(1)}^{(t)} - \widetilde{p}_{1(1)}^{(t)}, \ldots, \widehat{p}_{1(M)}^{(t)} - \widetilde{p}_{1(M)}^{(t)}\right)$. Then we can write $\left\|\widehat{p}_1^{(t)}(w) - \widetilde{p}_1^{(t)}(w)\right\|^2 = w^\top \Phi_1^{(t)\top} \Phi_1^{(t)} w$, and $\left(a_1^{(t)} - \widehat{p}_1^{(t)}(w)\right)^\top \left(\widehat{p}_1^{(t)}(w) - \widetilde{p}_1^{(t)}(w)\right) = w^\top \Phi_1^{(t)\top} e_1^{(t)} + w^\top \Phi_1^{(t)\top} \Omega_1^{(t)} w$, where $e_1^{(t)} = a_1^{(t)} - p_1^{(t)}$. Thus, we

can decompose $CV_\dagger(w)$ as

$$\begin{aligned}
CV_\dagger(w) =&\frac{1}{T|\Psi_1|}\sum_{t=1}^{T}\sum_{k=1}^{K}\left\|A^{(t)[k]}-\widetilde{P}^{(t)[k]}(w)\right\|_F^2\\
=&\frac{1}{T|\Psi_1|}\sum_{t=1}^{T}\left\|a_1^{(t)}-\widetilde{p}_1^{(t)}(w)\right\|^2\\
=&\frac{1}{T|\Psi_1|}\sum_{t=1}^{T}\left\{\left\|a_1^{(t)}-\widehat{p}_1^{(t)}(w)\right\|^2+\left\|\widehat{p}_1^{(t)}(w)-\widetilde{p}_1^{(t)}(w)\right\|^2\right.\\
&\left.+2\left(a_1^{(t)}-\widehat{p}_1^{(t)}(w)\right)^\top\left(\widehat{p}_1^{(t)}(w)-\widetilde{p}_1^{(t)}(w)\right)\right\}\\
=&\frac{1}{T|\Psi_1|}\sum_{t=1}^{T}\left\{\left\|a_1^{(t)}-\widehat{p}_1^{(t)}(w)\right\|^2+w^\top\Phi_1^{(t)\top}\Phi_1^{(t)}w+2w^\top\Phi_1^{(t)\top}e_1+2w^\top\Phi_1^{(t)\top}\Omega_1^{(t)}w\right\}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\left\|a_1^{(t)}-\widehat{p}_1^{(t)}(w_\dagger^0+\epsilon_\dagger r)\right\|^2-\left\|a_1^{(t)}-\widehat{p}_1(w_\dagger^0)\right\|^2\\
=&\widehat{p}_1^{(t)}(w_\dagger^0+\epsilon_\dagger r)^\top\widehat{p}_1^{(t)}(w_\dagger^0+\epsilon_\dagger r)-\widehat{p}_1^{(t)}(w_\dagger^0)^\top\widehat{p}_1^{(t)}(w_\dagger^0)-2a_1^{(t)\top}\left\{\widehat{p}_1^{(t)}(w_\dagger^0+\epsilon_\dagger r)-\widehat{p}_1^{(t)}(w_\dagger^0)\right\}\\
=&(w_\dagger^0+\epsilon_\dagger r)^\top\Lambda_1^{(t)\top}\Lambda_1^{(t)}(w_\dagger^0+\epsilon_\dagger r)-w_\dagger^{0\top}\Lambda_1^{(t)\top}\Lambda_1^{(t)}w_\dagger^0-2a_1^{(t)\top}\left\{\Lambda_1^{(t)}(w_\dagger^0+\epsilon_\dagger r)-\Lambda_1^{(t)}w_\dagger^0\right\}\\
=&\epsilon_\dagger^2 r^\top\Lambda^{(t)}r+2\left(\Lambda_1^{(t)}w_\dagger^0-a_1^{(t)}\right)^\top\Lambda_1^{(t)}\epsilon_\dagger r.
\end{aligned}$$

As a result,

$$\begin{aligned}
&CV_\dagger(w_\dagger^0+\epsilon_\dagger r)-CV_\dagger(w_\dagger^0)\\
=&\frac{1}{T|\Psi_1|}\sum_{t=1}^{T}\left\{\|a_1^{(t)}-\widehat{p}_1^{(t)}(w_\dagger^0+\epsilon_\dagger r)\|^2-\|a_1^{(t)}-\widehat{p}_1^{(t)}(w_\dagger^0)\|^2+\epsilon_\dagger^2 r^\top\Phi_1^{(t)\top}\Phi_1^{(t)}r+2\epsilon_\dagger r^\top\Phi_1^{(t)\top}e_1^{(t)}\right.\\
&\left.+2\epsilon_\dagger^2 r^\top\Phi_1^{(t)\top}\Omega_1^{(t)}r+2\epsilon_\dagger r^{(t)\top}\Phi_1^{(t)\top}\Omega_1^{(t)}w_\dagger^0+2\epsilon_\dagger w_\dagger^{0\top}\Phi_1^{(t)\top}\Omega_1^{(t)}r+2\epsilon_\dagger r^\top\Phi_1^{(t)\top}\Phi_1^{(t)}w_\dagger^0\right\}
\end{aligned}$$

$$=\frac{1}{T|\Psi_1|}\sum_{t=1}^{T}\Big\{\epsilon_\dagger^2 r^\top\Lambda^{(t)}r+2\epsilon_\dagger\left(\Lambda_1^{(t)}w_\dagger^0-a_1^{(t)}\right)^\top\Lambda_1^{(t)}r+\epsilon_\dagger^2 r^\top\Phi_1^{(t)\top}\Phi_1^{(t)}r+2\epsilon_\dagger r^\top\Phi_1^{(t)\top}e_1^{(t)}$$
$$+2\epsilon_\dagger^2 r^\top\Phi_1^{(t)\top}\Omega_1^{(t)}r+2\epsilon_\dagger r^\top\Phi_1^{(t)\top}\Omega_1^{(t)}w_\dagger^0+2\epsilon_\dagger w_\dagger^{0\top}\Phi_1^{(t)\top}\Omega_1^{(t)}r+2\epsilon_\dagger r^\top\Phi_1^{(t)\top}\Phi_1^{(t)}w_\dagger^0\Big\}. \tag{S2.56}$$

It is obvious that $\epsilon_\dagger^2 r^\top\Phi_1^{(t)\top}\Phi_1^{(t)}r\geq 0$ for any r. Besides, under Assumption S5, we have

$$\sum_{t=1}^{T}\epsilon_\dagger^2 r^\top\Lambda^{(t)}r>\rho_{1\dagger}\epsilon_\dagger^2 T|\Psi_1|\|r\|^2>0 \tag{S2.57}$$

in probability tending to 1. In the following, we show that the remaining six terms of (S2.56) are asymptotically dominated by $\sum_{t=1}^{T}\epsilon_\dagger^2 r^\top\Lambda^{(t)}r$.

We first consider the item $\sum_{t=1}^{T}\left|\epsilon_\dagger(\Lambda_1^{(t)}w_\dagger^0-a_1^{(t)})^\top\Lambda_1^{(t)}r\right|$. According to the definitions of $\xi_\dagger$ and $w_\dagger^0$, we have $E\big(\sum_{t=1}^{T}\|\Lambda_1^{(t)}w_\dagger^0-p_1^{(t)}\|^2\big)=\xi_\dagger$. Further, by Assumption S5, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\left|\epsilon_\dagger\left(\Lambda_1^{(t)}w_\dagger^0-a_1^{(t)}\right)^\top\Lambda_1^{(t)}r\right|\\
\leq&\sum_{t=1}^{T}\left|\epsilon_\dagger\left(\Lambda_1^{(t)}w_\dagger^0-p_1^{(t)}\right)^\top\Lambda_1^{(t)}r\right|+\sum_{t=1}^{T}\left|\epsilon_\dagger\left(p_1^{(t)}-a_1^{(t)}\right)^\top\Lambda_1^{(t)}r\right|\\
\leq&\epsilon_\dagger\left(\sum_{t=1}^{T}\left\|\Lambda_1^{(t)}w_\dagger^0-p_1^{(t)}\right\|^2\right)^{1/2}\left(\sum_{t=1}^{T}\left\|\Lambda_1^{(t)}r\right\|^2\right)^{1/2}+\epsilon_\dagger\left(\sum_{t=1}^{T}\left\|p_1^{(t)}-a_1^{(t)}\right\|^2\right)^{1/2}\left(\sum_{t=1}^{T}\left\|\Lambda_1^{(t)}r\right\|^2\right)^{1/2}\\
\leq&\epsilon_\dagger\left\{\left(\sum_{t=1}^{T}\left\|\Lambda_1^{(t)}w_\dagger^0-p_1^{(t)}\right\|^2\right)^{1/2}+\left(\sum_{t=1}^{T}\left\|p_1^{(t)}-a_1^{(t)}\right\|^2\right)^{1/2}\right\}\left\{r^\top\left(\sum_{t=1}^{T}\Lambda^{(t)}\right)r\right\}^{1/2}
\end{aligned}$$

$$=O_p\left(\epsilon_\dagger \xi_\dagger^{1/2} T^{1/2}|\Psi_1|^{1/2}M^{1/2}\right)\|r\| + O_p\left(\epsilon_\dagger TN^{1/2}\max\{D_\dagger, \log N\}^{1/2}|\Psi_1|^{1/2}M^{1/2}\right)\|r\|. \tag{S2.58}$$

Notice that

$$\begin{aligned}&\sum_{t=1}^{T}\left\|\Phi_1^{(t)}\right\|_2^2 \le \sum_{t=1}^{T}\mathrm{tr}\left(\Phi_1^{(t)\top}\Phi_1^{(t)}\right) = \sum_{t=1}^{T}\sum_{m=1}^{M}\left\|\widetilde{p}_{1(m)}^{(t)} - \widehat{p}_{1(m)}^{(t)}\right\|^2 \\ \le &\sum_{m=1}^{M} 2\left(\sum_{t=1}^{T}\left\|\widetilde{p}_{1(m)}^{(t)} - p_{1(m)}^{(t)*}\right\|^2 + \sum_{t=1}^{T}\left\|p_{1(m)}^{(t)*} - \widehat{p}_{1(m)}^{(t)}\right\|^2\right) = O_p\left(NM^2T\right),\end{aligned} \tag{S2.59}$$

where the last step is due to Assumption S1. Based on (S2.59) and (S2.44), we can obtain

$$\begin{aligned}\sum_{t=1}^{T}\left|\epsilon_\dagger r^\top \Phi_1^{(t)\top} e_1^{(t)}\right| \le &\epsilon_\dagger\|r\|\left(\sum_{t=1}^{T}\left\|\Phi_1^{(t)}\right\|_2^2\right)^{1/2}\left(\sum_{t=1}^{T}\left\|e_1^{(t)}\right\|^2\right)^{1/2} \\ \le &O_p\left(\epsilon_\dagger TMN\max\{D_\dagger, \log N\}^{1/2}\right)\|r\|.\end{aligned} \tag{S2.60}$$

Observe from Assumption S6 that $\lambda_{\max}(\sum_{t=1}^{T}\Omega^{(t)}) = O_p(T|\Psi_1|M)$, and hence,

$$\sum_{t=1}^{T}\left|\epsilon_\dagger^2 r^\top \Phi_1^{(t)\top}\Omega_1^{(t)} r\right| \le \epsilon_\dagger^2\left(\sum_{t=1}^{T}\left\|\Phi_1^{(t)} r\right\|^2\right)^{1/2}\left(\sum_{t=1}^{T}\left\|\Omega_1^{(t)} r\right\|^2\right)^{1/2}$$

$$\leq \epsilon_\dagger^2 \|r\|^2 \left(\sum_{t=1}^{T} \left\| \Phi_1^{(t)} \right\|_2^2 \right)^{1/2} \lambda_{\max}^{1/2} \left(\sum_{t=1}^{T} \Omega^{(t)} \right) = O_p \left(\epsilon_\dagger^2 T M^{3/2} N^{1/2} |\Psi_1|^{1/2} \right) \|r\|^2. \tag{S2.61}$$

Noting that $E\big(\sum_{t=1}^{T} \left\| \Omega_1^{(t)} w_\dagger^0 \right\|^2\big) = E\big(\sum_{t=1}^{T} \left\| p_1^{(t)} - \widehat{p}_1^{(t)}(w_\dagger^0) \right\|^2\big) = \xi_\dagger$, we have $\sum_{t=1}^{T} \left\| \Omega_1^{(t)} w_\dagger^0 \right\|^2 = O_p(\xi_\dagger)$, which implies that

$$\begin{aligned} &\sum_{t=1}^{T} \left| \epsilon_\dagger r^\top \Phi_1^{(t)\top} \Omega_1^{(t)} w_\dagger^0 \right| \leq \epsilon_\dagger \left(\sum_{t=1}^{T} \left\| \Phi_1^{(t)} r \right\|^2 \right)^{1/2} \left(\sum_{t=1}^{T} \left\| \Omega_1^{(t)} w_\dagger^0 \right\|^2 \right)^{1/2} \\ \leq &\epsilon_\dagger \|r\| \left(\sum_{t=1}^{T} \left\| \Phi_1^{(t)} \right\|_2^2 \right)^{1/2} \left(\sum_{t=1}^{T} \left\| \Omega_1^{(t)} w_\dagger^0 \right\|^2 \right)^{1/2} = O_p \left(\epsilon_\dagger \xi_\dagger^{1/2} T^{1/2} M N^{1/2} \right) \|r\|. \end{aligned} \tag{S2.62}$$

Similar to (S2.59), we can obtain $\sum_{t=1}^{T} \left\| \widehat{p}_{1(m)}^{(t)} - \widetilde{p}_{1(m)}^{(t)} \right\|^2 = O_p(NMT)$, uniformly in m ($m = 1, \ldots, M$). Thus,

$$\begin{aligned} &\sum_{t=1}^{T} \left\| \Phi_1^{(t)} w_\dagger^0 \right\|^2 = \sum_{t=1}^{T} \left\| \sum_{m=1}^{M} w_{\dagger m}^0 \left(\widehat{p}_{1(m)}^{(t)} - \widetilde{p}_{1(m)}^{(t)} \right) \right\|^2 \\ \leq &\sum_{t=1}^{T} \sum_{m=1}^{M} w_{\dagger m}^0 \left\| \widehat{p}_{1(m)}^{(t)} - \widetilde{p}_{1(m)}^{(t)} \right\|^2 = O_p(NMT), \end{aligned} \tag{S2.63}$$

which indicates that

$$\sum_{t=1}^{T} \left| \epsilon_\dagger w_\dagger^{0\top} \Phi_1^{(t)\top} \Omega_1^{(t)} r \right| \leq \epsilon_\dagger \left(\sum_{t=1}^{T} \left\| \Phi_1^{(t)} w_\dagger^0 \right\|^2 \right)^{1/2} \left(\sum_{t=1}^{T} \left\| \Omega_1^{(t)} r \right\|^2 \right)^{1/2}$$

$$\begin{aligned}
&\leq \epsilon_\dagger \|r\| \left(\sum_{t=1}^{T} \left\| \Phi_1^{(t)} w_\dagger^0 \right\|^2 \right)^{1/2} \lambda_{\max}^{1/2} \left(\sum_{t=1}^{T} \Omega^{(t)} \right) \\
&= O_p\left(\epsilon_\dagger T M N^{1/2} |\Psi_1|^{1/2} \right) \|r\|. \qquad \text{(S2.64)}
\end{aligned}$$

Similarly, utilizing (S2.59) and (S2.63) again, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \left| \epsilon_\dagger r^\top \Phi_1^{(t)\top} \Phi_1^{(t)} w_\dagger^0 \right| \leq \epsilon_\dagger \left\{ \sum_{t=1}^{T} \left\| \Phi_1^{(t)} r \right\|^2 \right\}^{1/2} \left\{ \sum_{t=1}^{T} \left\| \Phi_1^{(t)} w_\dagger^0 \right\|^2 \right\}^{1/2} \\
\leq &\epsilon_\dagger \|r\| \left\{ \sum_{t=1}^{T} \left\| \Phi_1^{(t)} \right\|_2^2 \right\}^{1/2} \left\{ \sum_{t=1}^{T} \left\| \Phi_1^{(t)} w_\dagger^0 \right\|^2 \right\}^{1/2} = O_p\left(\epsilon_\dagger T M^{3/2} N \right) \|r\|. \qquad \text{(S2.65)}
\end{aligned}$$

Recalling $\epsilon_\dagger = \xi_\dagger^{1/2} T^{-1/2} |\Psi_1|^{-1/2+\kappa_\dagger}$, Assumption S7 ensures that

$$\frac{\epsilon_\dagger \xi_\dagger^{1/2} T^{1/2} |\Psi_1|^{1/2} M^{1/2}}{\epsilon_\dagger^2 T |\Psi_1|} = \frac{M^{1/2}}{|\Psi_1|^{\kappa_\dagger}} = o(1), \qquad \text{(S2.66)}$$

$$\frac{\epsilon_\dagger T N^{1/2} \max\{D_\dagger, \log N\}^{1/2} |\Psi_1|^{1/2} M^{1/2}}{\epsilon_\dagger^2 T |\Psi_1|} = \frac{N^{1/2} T^{1/2} \max\{D_\dagger, \log N\}^{1/2} M^{1/2}}{\xi_\dagger^{1/2} |\Psi_1|^{\kappa_\dagger}} = o(1), \qquad \text{(S2.67)}$$

$$\frac{\epsilon_\dagger^2 T M^{3/2} N^{1/2} |\Psi_1|^{1/2}}{\epsilon_\dagger^2 T |\Psi_1|} = \frac{M^{3/2}}{N}^{1/2} |\Psi_1|^{1/2} = o(1), \qquad \text{(S2.68)}$$

and

$$\frac{\epsilon_\dagger T M N^{1/2} |\Psi_1|^{1/2}}{\epsilon_\dagger^2 T |\Psi_1|} = \frac{(TN)^{1/2} M}{\xi_\dagger^{1/2} |\Psi_1|^{\kappa_\dagger}} = o(1). \qquad \text{(S2.69)}$$

In addition, by Assumption S7, we obtain that

$$\frac{\epsilon_\dagger TMN \max\{D_\dagger, \log N\}^{1/2}}{\epsilon_\dagger TN^{1/2} \max\{D_\dagger, \log N\}^{1/2} |\Psi_1|^{1/2} M^{1/2}} = \frac{M^{1/2}{}^{1/2}}{N} |\Psi_1|^{1/2} = o(1), \quad \text{(S2.70)}$$

$$\frac{\epsilon_\dagger \xi_\dagger^{1/2} T^{1/2} MN^{1/2}}{\epsilon_\dagger \xi_\dagger^{1/2} T^{1/2} |\Psi_1|^{1/2} M^{1/2}} = \frac{M^{1/2}{}^{1/2}}{N} |\Psi_1|^{1/2} = o(1), \quad \text{(S2.71)}$$

and

$$\frac{\epsilon_\dagger TM^{3/2}N}{\epsilon_\dagger TMN^{1/2} |\Psi_1|^{1/2}} = \frac{M^{1/2}{}^{1/2}}{N} |\Psi_1|^{1/2} = o(1). \quad \text{(S2.72)}$$

Therefore, from (S2.58), (S2.60), (S2.61), (S2.62), (S2.64) and (S2.65), and using (S2.57) and (S2.66)-(S2.72), we see that the remaining six terms of (S2.56) are asymptotically dominated by $\sum_{t=1}^{T} \epsilon_\dagger^2 r^\top \Lambda^{(t)} r$. This completes the proof of Theorem S3.

S3 Verifications of Assumptions 1-6

S3.1 Verifications of Assumption 1

Let $\alpha_{(m)}^* \in \mathbb{R}^N$ and $Z_{(m)}^* \in \mathbb{R}^{N \times m}$ denote the limiting values of $\widehat{\alpha}_{(m)}$ and $\widehat{Z}_{(m)}$, respectively. We require that $JZ_{(m)}^* = Z_{(m)}^*$ and $J\widehat{Z}_{(m)} = \widehat{Z}_{(m)}$. Denote $\Theta_{(m)}^* = \alpha_{(m)}^* 1_N^\top + 1_N \alpha_{(m)}^{*\top} + Z_{(m)}^* Z_{(m)}^{*\top}$ and $\widehat{\Theta}_{(m)} = \widehat{\alpha}_{(m)} 1_N^\top + 1_N \widehat{\alpha}_{(m)}^\top + \widehat{Z}_{(m)} \widehat{Z}_{(m)}^\top$. We assume that $\|\Theta\|_{\max} \leq \mu$, $\|\Theta_{(m)}^*\|_{\max} \leq \mu$ and $\|\widehat{\Theta}_{(m)}\|_{\max} \leq \mu$ uniformly for $m = 1, \ldots, M$, where $\|\cdot\|_{\max}$ denotes the maximum abso-

lute value of entries in a matrix. Since $\widehat{\Theta}^0_{(m)} = \text{logit}(\widehat{P}^0_{(m)})$ are obtained by minimizing the following objective function:

$$\mathcal{L}(\Theta) = -\sum_{i,j}\{A(\Psi_1)_{ij}\Theta_{ij} - f(\Theta_{ij})\},$$

we have

$$\mathcal{L}(\widehat{\Theta}^0_{(m)}) - \mathcal{L}(\Theta^{*,0}_{(m)}) \leq 0, \tag{S3.73}$$

where $\Theta^{*,0}_{(m)}$ is the limiting value of $\widehat{\Theta}^0_{(m)}$. Expanding $\mathcal{L}(\widehat{\Theta}^0_{(m)}) - \mathcal{L}(\Theta^{*,0}_{(m)})$, we obtain

$$\begin{aligned}
&\mathcal{L}(\Theta^{*,0}_{(m)}) - \mathcal{L}(\widehat{\Theta}^0_{(m)}) = \sum_{i,j}\left\{A(\Psi_1)_{ij}\left(\widehat{\Theta}^0_{(m),ij} - \Theta^{*,0}_{(m),ij}\right) - \left(f(\widehat{\Theta}^0_{(m),ij}) - f(\Theta^{*,0}_{(m),ij})\right)\right\} \\
=&\sum_{i,j}\left\{\left(A(\Psi_1)_{ij} - f'(\Theta^{*,0}_{(m),ij})\right)\left(\widehat{\Theta}^0_{(m),ij} - \Theta^{*,0}_{(m),ij}\right)\right\} \\
&-\sum_{i,j}\left\{f(\widehat{\Theta}^0_{(m),ij}) - f(\Theta^{*,0}_{(m),ij}) - f'(\Theta^{*,0}_{(m),ij})\left(\widehat{\Theta}^0_{(m),ij} - \Theta^{*,0}_{(m),ij}\right)\right\} \\
=&\sum_{i,j}\left\{\left(A(\Psi_1)_{ij} - f'(\Theta^{*,0}_{(m),ij})\right)\left(\widehat{\Theta}^0_{(m),ij} - \Theta^{*,0}_{(m),ij}\right)\right\} - \sum_{i,j}\frac{1}{2}f''(\bar{\Theta}_{(m),ij})\left(\widehat{\Theta}^0_{(m),ij} - \Theta^{*,0}_{(m),ij}\right)^2 \\
\leq&\sum_{i,j}\left\{\left(A(\Psi_1)_{ij} - f'(\Theta^{*,0}_{(m),ij})\right)\left(\widehat{\Theta}^0_{(m),ij} - \Theta^{*,0}_{(m),ij}\right)\right\} - \frac{1}{2}\min_{|\omega|\leq\mu} f''(\omega)\|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F^2,
\end{aligned} \tag{S3.74}$$

where $\bar{\Theta}_{(m),ij}$ is a value between $\widehat{\Theta}^0_{(m)}$ and $\Theta^{*,0}_{(m)}$. According to (S3.73) and (S3.74), we have

$$\begin{aligned}
&\|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F^2 \leq \frac{2}{\min_{|\omega|\leq\mu} f''(\omega)} \sum_{i,j} \left\{ \left(A(\Psi_1)_{ij} - f'(\Theta^{*,0}_{(m),ij}) \right) \left(\widehat{\Theta}^0_{(m),ij} - \Theta^{*,0}_{(m),ij} \right) \right\} \\
=&\frac{2}{\min_{|\omega|\leq\mu} f''(\omega)} \left\langle A(\Psi_1) - f'(\Theta^{*,0}_{(m)}), \widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)} \right\rangle, \qquad \text{(S3.75)}
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two matrices. For any two matrices A and B, we have $|\langle A, B\rangle| \leq \|A\|_2 \sqrt{\text{rank}(B)} \|B\|_F$. According to the definition of $\widehat{\Theta}_{(m)}$ and $\Theta^*_{(m)}$, we have

$$\begin{aligned}
&\text{rank}(\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}) \\
=&\text{rank}\left((\widehat{\alpha}^0_{(m)} - \alpha^{*,0}_{(m)}) 1_N^\top + 1_N (\widehat{\alpha}^0_{(m)} - \alpha^{*,0}_{(m)})^\top + \widehat{Z}^0_{(m)} \widehat{Z}^{0\top}_{(m)} - Z^{*,0}_{(m)} Z^{*,0\top}_{(m)} \right) \leq 2 + 2m.
\end{aligned}$$

It implies that

$$\begin{aligned}
&\left| \left\langle A(\Psi_1) - f'(\Theta^{*,0}_{(m)}), \widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)} \right\rangle \right| \\
\leq &\left| \left\langle A(\Psi_1) - pP, \widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)} \right\rangle \right| + \left| \left\langle pP - f'(\Theta^{*,0}_{(m)}), \widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)} \right\rangle \right| \\
\leq &\sqrt{2+2m} \|A(\Psi_1) - pP\|_2 \|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F + \|pP - f'(\Theta^{*,0}_{(m)})\|_F \|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F.
\end{aligned} \qquad \text{(S3.76)}$$

Further, (S3.75) and (S3.76) together imply that

$$\begin{aligned}\|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F \leq & \frac{2\sqrt{2+2m}\|A(\Psi_1) - pP\|_2 + 2\|pP - f'(\Theta^{*,0}_{(m)})\|_F}{\min_{|\omega|\leq\mu} f''(\omega)} \\ \leq & \frac{2\sqrt{2+2m}}{\min_{|\omega|\leq\mu} f''(\omega)}\|A(\Psi_1) - pP\|_2 + \frac{\|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F + \|\widehat{\Theta}^0_{(m)} - \Theta_p\|_F}{8\min_{|\omega|\leq\mu} f''(\omega)},\end{aligned}$$

where $\Theta_p = (\text{logit}(pP_{ij})) \in \mathbb{R}^{N\times N}$. Assume that μ satisfies $\min_{\omega\leq\mu} f''(\omega) > 1/8$, then we have

$$\|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F \leq \frac{16\min_{|\omega|\leq\mu} f''(\omega)\sqrt{2+2m}}{(8\min_{|\omega|\leq\mu} f''(\omega) - 1)\min_{|\omega|\leq\mu} f''(\omega)}\|A(\Psi_1) - pP\|_2 + \frac{\|\widehat{\Theta}^0_{(m)} - \Theta_p\|_F}{8\min_{|\omega|\leq\mu} f''(\omega) - 1}. \tag{S3.77}$$

Since $E(A(\Psi_1)_{ij}) = pP_{ij}$, according to Lemma 2, we have

$$\|A(\Psi_1) - pP\|_2 = O_p(\max\{D, \log N\}^{1/2}). \tag{S3.78}$$

Further assume that $\max_{1\leq i\leq N}\|z_i\|$, and $\|\alpha\|_{\max} \leq \mu_1/2$. Denote the condition number of Z (i.e., the ratio of the largest to the smallest singular values) as κ_Z. When $m \geq d_0$, Theorem 9 of Ma et al. (2020) implies that, under Assumption S2 of Ma et al. (2020) and some constraints on $\|Z\|_2^2$, we have

$$\|\widehat{\Theta}^0_{(m)} - \Theta_p\|_F^2 \leq C_4\kappa_Z^2 e^{2\mu_1} Nm \times \max\left\{e^{\mu}, \frac{\log N}{N}\right\}, \tag{S3.79}$$

with probability at least $1 - N^{-C_5}$, where C_4 and C_5 are some positive constants.

When the dimension of the latent vector is mis-specified, i.e., $m \leq d_0$, Theorem 14 of Ma et al. (2020) gives the upper bound of the estimation error of Θ obtained by the projected gradient descent algorithm. Let $U_m D_m U_m^\top$ be the best rank-m approximation to $ZZ^\top$ and $\bar{F}_m = ZZ^\top - U_m D_m U_m^\top$. Denote the condition number of $U_m D_m^{1/2}$ as $\kappa_\star$. According to this theorem, under Assumption S2 of Ma et al. (2020) and some constraints on $\|ZZ^\top\|_2$, we have

$$\|\widehat{\Theta}^0_{(m)} - \Theta_p\|_F^2 \leq C_4 \kappa_\star^2 \left[e^{2\mu_1} Nm \times \max\left\{ e^\mu, \frac{\log N}{N} \right\} + e^{\mu_1} \|\bar{F}_m\|_F^2 \right], \tag{S3.80}$$

with probability at least $1 - N^{-C_5}$. Furthermore, by Schmidt-Mirsky-Eckart Young Theorem, we have $\|\bar{F}_m\|_F^2 = \lambda_{m+1}^2 + \cdots + \lambda_{d_0}^2$, where λ_r is the rth largest singular value of $ZZ^\top$ for $r = 1, \ldots, d_0$. Consider a special case where each element in $\bar{F}_m$ follows a standard Gaussian distribution. According to Theorem 2.5 in Chafai et al. (2009), the order of λ_{m+1}^2 is N. Assuming that $\lambda_{d_0}^2$ and λ_{m+1}^2 are of the same order, we find that the order of $\|\bar{F}_m\|_F^2$ is $N(d_0 - m)$. When μ, $\kappa_\star$ are finite constants, and $e^{\mu_1} = M/d_0 = O(1)$,

according to (S3.77)-(S3.80), we have

$$\|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F \leq O_p(\sqrt{NM}),$$

when $D = O(N)$. Note that the bound in (S3.78) does not depend on m. The bounds in (S3.79) and (S3.80) each hold for a fixed m with probability at least $1 - N^{-C_5}$. Since μ is a fixed constant, we have $e^\mu \geq e^0 = 1 > \log N/N$ for all sufficiently large N, which implies $\max\{e^\mu, \log N/N\} = e^\mu = O(1)$. Therefore, the bounds in (S3.79) and (S3.80) are both of order $O_p(Nm)$. Applying a union bound over $m = 1, \ldots, M$, the above bound $\max_{1\leq m\leq M} \|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F^2 \leq O_p(NM)$ holds uniformly for all $m = 1, \ldots, M$ with probability at least $1 - MN^{-C_5}$. When M is fixed, $MN^{-C_5} \to 0$ trivially. When M is allowed to diverge with N, the probability $1 - MN^{-C_5}$ still tends to 1 as long as $M = o(N^{C_5})$. Since $C_5 > 0$, Assumption 1 remains valid even when M diverges with N. Then we have

$$\begin{aligned}\max_m \sum_{(i,j)\in\Psi_1} \left(\widehat{P}_{(m),ij} - P^*_{(m),ij}\right)^2 \leq & \max_m \|\widehat{P}_{(m)} - P^*_{(m)}\|_F^2 \\ \leq & \frac{1}{16p^2} \max_m \|\widehat{\Theta}^0_{(m)} - \Theta^{*,0}_{(m)}\|_F^2 \\ \leq & O_p(NM), \end{aligned} \tag{S3.81}$$

when p is a fixed constant.

S3.2 Verifications of Assumption 2

Consider the case where $M < d_0$, which means that the candidate models do not include the models with large enough dimensions. Let $\mathcal{H} = \alpha 1_N^\top + 1_N \alpha^\top$, $\mathcal{H}^*(w) = \sum_{m=1}^M w_m(\alpha_{(m)}^* 1_N^\top + 1_N \alpha_{(m)}^{*\top})$, $\mathcal{I} = ZZ^\top$, $\widehat{\mathcal{I}}_{(m)} = \widehat{Z}_{(m)}\widehat{Z}_{(m)}^\top$, $\widehat{\mathcal{I}}(w) = \sum_{m=1}^M w_m \widehat{\mathcal{I}}_{(m)}$, $\mathcal{I}_{(m)}^* = Z_{(m)}^* Z_{(m)}^{*\top}$ and $\mathcal{I}^*(w) = \sum_{m=1}^M w_m \mathcal{I}_{(m)}^*$. Besides, Θ_1 denotes the matrix obtained from Θ by setting all entries of Θ with indices in Ψ_2 to 0. Similarly, let $\Theta_1^*(w)$, $\mathcal{H}_1$, $\mathcal{H}_1^*(w)$, $\mathcal{I}_1$ and $\mathcal{I}_1^*(w)$ be the matrices obtained from $\Theta^*(w)$, $\mathcal{H}$, $\mathcal{H}^*(w)$, $\mathcal{I}$ and $\mathcal{I}^*(w)$ by setting all entries of them with indices in Ψ_2 to 0, respectively. Then, we have

$$
\begin{aligned}
&\|\Theta_1^*(w) - \Theta_1\|_F^2 \\
=&\|\mathcal{H}_1^*(w) - \mathcal{H}_1 + \mathcal{I}_1^*(w) - \mathcal{I}_1\|_F^2 \\
\geq&\|\mathcal{H}_1^*(w) - \mathcal{H}_1\|_F^2 + \|\mathcal{I}_1^*(w) - \mathcal{I}_1\|_F^2 - 2|\langle \mathcal{H}_1^*(w) - \mathcal{H}_1, \mathcal{I}_1^*(w) - \mathcal{I}_1\rangle| \\
\geq&\|\mathcal{H}_1^*(w) - \mathcal{H}_1\|_F^2 + \|\mathcal{I}_1^*(w) - \mathcal{I}_1\|_F^2 - 2\|\mathcal{H}_1^*(w) - \mathcal{H}_1\|_F \|\mathcal{I}_1^*(w) - \mathcal{I}_1\|_F \\
\geq&\|\mathcal{H}_1^*(w) - \mathcal{H}_1\|_F^2 + \|\mathcal{I}_1^*(w) - \mathcal{I}_1\|_F^2 - c_1^{-1}\|\mathcal{H}_1^*(w) - \mathcal{H}_1\|_F^2 - c_1\|\mathcal{I}_1^*(w) - \mathcal{I}_1\|_F^2 \\
=&(1 - c_1^{-1})\|\mathcal{H}_1^*(w) - \mathcal{H}_1\|_F^2 + (1 - c_1)\|\mathcal{I}_1^*(w) - \mathcal{I}_1\|_F^2, \qquad \text{(S3.82)}
\end{aligned}
$$

where $0 < c_1 < 1$. Next, we introduce some notations. Let $\mathcal{I}_{(m)}^\star = U_m D_m U_m^\top$ be the top-m eigen-decomposition of $\mathcal{I}$, which is the best rank-m approximation to $\mathcal{I}$ according to Schmidt-Mirsky-Eckart Young Theorem, and

$\mathcal{I}^\star(w) = \sum_{m=1}^{M} w_m \mathcal{I}^\star_{(m)}$. Theorem 9 in Ma et al. (2020) assures that $\widehat{\mathcal{I}}_{(m)}$ and $\widehat{\Theta}_{(m)}$ consistently estimates $\mathcal{I}^\star_{(m)}$ and $\mathcal{H}+\mathcal{I}^\star_{(m)}$ under certain conditions. Therefore, we have $\mathcal{I}^*_{(m)} = \mathcal{I}^\star_{(m)}$ and $\mathcal{H}^*_{(m)} = \mathcal{H}$. Then (S3.82) becomes

$$\|\Theta^*_1(w) - \Theta_1\|_F^2 \geq (1 - c_1)\|\mathcal{I}^\star_1(w) - \mathcal{I}_1\|_F^2. \tag{S3.83}$$

Schmidt-Mirsky-Eckart Young Theorem implies that $\|\mathcal{I}^\star_{(m)} - \mathcal{I}\|_F^2 = \lambda^2_{m+1} + \cdots + \lambda^2_{d_0}$. Then, for any $w \in \mathcal{W}$,

$$\|\mathcal{I}^\star(w) - \mathcal{I}\|_F^2 \geq \|\mathcal{I}^\star_{(M)} - \mathcal{I}\|_F^2 = \lambda^2_{M+1} + \cdots + \lambda^2_{d_0}, \tag{S3.84}$$

where the inequality is because $\mathcal{I}^\star_{(M)}$ is the optimal rank-M approximation of $\mathcal{I}$. When each element in $\mathcal{I}^\star_{(M)} - \mathcal{I}$ follows a standard Gaussian distribution, the order of λ^2_{M+1} is N. Assuming that $\lambda^2_{d_0}$ and λ^2_{M+1} are of the same order, the order of $\|\mathcal{I}^\star(w) - \mathcal{I}\|_F^2$ is at least $N(d_0 - M)$. For any $i, j \in 1, \ldots, N$ and $w \in \mathcal{W}$,

$$\begin{aligned} |P^*_{ij}(w) - P_{ij}| &= |f''(\widetilde{\Theta})\left\{\Theta^*_{ij}(w) - \Theta_{ij}\right\}| \\ &= f''(\widetilde{\Theta})|\Theta^*_{ij}(w) - \Theta_{ij}| \\ &\geq \min_{|\omega| \leq \mu} f''(\omega)|\Theta^*_{ij}(w) - \Theta_{ij}|, \end{aligned}$$

where $\widetilde{\Theta}$ is a value between $\Theta^*_{ij}(w)$ and Θ_{ij}. Then, for any $w \in \mathcal{W}$, we have

$$\begin{aligned} L^*(w) &= \sum_{(i,j)\in\Psi_1} \left\{P^*_{ij}(w) - P_{ij}\right\}^2 \\ &\geq \left\{\min_{|\omega|\leq\mu} f''(\omega)\right\}^2 \sum_{(i,j)\in\Psi_1} \left\{\Theta^*_{ij}(w) - \Theta_{ij}\right\}^2 \\ &= \left\{\min_{|\omega|\leq\mu} f''(\omega)\right\}^2 \|\Theta^*_1(w) - \Theta_1\|_F^2. \end{aligned}$$

When μ is a fixed constant, assume that $\|\mathcal{I}^\star_1(w) - \mathcal{I}_1\|_F^2$ has the same order of $\|\mathcal{I}^\star(w) - \mathcal{I}\|_F^2$. Together with (S3.83) and (S3.84), the first part of Assumption 2 holds true under the condition $M = o(d_0 - M)$, and the second part of Assumption 2 holds true when $\max\{D, \log N\} = O(d_0 - M)$.

S3.3 Verifications of Assumption 3

In this subsection, we consider the case where $M \geq d_0$, that is, the candidate set includes models with large enough latent dimensions. The set $\mathcal{W}^s$ assigns all weights to the misspecified models. It means that the weights of candidate models with latent space dimensions not less than d_0 are all 0. Without loss of generality, we assume that $Z^*_{(m)} = U_m D_m^{1/2}$, and $Z^*_{(m)} Z^{*\top}_{(m)} = \sum_{m_0=1}^m Z_{\cdot,m_0} Z^\top_{\cdot,m_0}$, where $Z_{\cdot,m_0}$ is the m_0-th column vector of Z for $m = 1, \ldots, d_0$. For any $w \in \mathcal{W}^s$, we have

$$
\begin{aligned}
L^*(w) &\geq \left\{ \min_{|\omega|\leq\mu} f''(\omega) \right\}^2 \sum_{(i,j)\in\Psi_1} \left\{ \sum_{m=1}^{M} w_m(\alpha^*_{(m),i} + \alpha^*_{(m),j} + z^{*\top}_{(m),i} z^*_{(m),j}) - \alpha_i - \alpha_j - z_i^\top z_j \right\}^2 \\
&= \left\{ \min_{|\omega|\leq\mu} f''(\omega) \right\}^2 \sum_{(i,j)\in\Psi_1} \left\{ \sum_{m=1}^{d_0-1} w_m(\alpha^*_{(m),i} + \alpha^*_{(m),j} + z^{*\top}_{(m),i} z^*_{(m),j}) - \alpha_i - \alpha_j - z_i^\top z_j \right\}^2 \\
&= \left\{ \min_{|\omega|\leq\mu} f''(\omega) \right\}^2 \sum_{(i,j)\in\Psi_1} \left\{ \sum_{m=1}^{d_0-1} w_m z^{*\top}_{(m),i} z^*_{(m),j} - z_i^\top z_j \right\}^2 \\
&= \left\{ \min_{|\omega|\leq\mu} f''(\omega) \right\}^2 \|\mathcal{I}^\star_{d_0-1,1}(w) - \mathcal{I}_1\|_F^2,
\end{aligned}
$$

where $\mathcal{I}^\star_{d_0-1,1}(w)$ denotes the matrices obtained from $\mathcal{I}^\star_{d_0-1}(w) = \sum_{m=1}^{d_0-1} w_m \mathcal{I}^\star_{(m)}$ by setting all entries of $\mathcal{I}^\star_{d_0-1}(w)$ with indices in Ψ_2 to 0. Note that

$$
\|\mathcal{I}^\star_{d_0-1}(w) - \mathcal{I}\|_F^2 \geq \|\mathcal{I}^\star_{d_0-1} - \mathcal{I}\|_F^2 = \|Z_{\cdot,d_0} Z_{\cdot,d_0}^\top\|_F^2.
$$

When μ is a fixed constant, assume that the order of $\|Z_{\cdot,d_0}\|_2^2$ is N, and $\|\mathcal{I}^\star_{d_0-1,1}(w) - \mathcal{I}_1\|_F^2$ has the same order of $\|\mathcal{I}^\star_{d_0-1}(w) - \mathcal{I}\|_F^2$, then the order of $L^*(w)$ is at least N^2 for any $w \in \mathcal{W}^s$. Assumption 3 holds when $M = o(N)$ and $\max\{D, \log N\} = O(N)$.

S3.4 Verification of Assumptions 4-6

Assumptions 4-6 are standard regularity conditions in the model averaging literature. Here we discuss their specific implications and validity under the

latent space model framework.

Assumption 4 requires that the minimum singular value of the matrix $\Lambda/|\Psi_1|$ is bounded away from zero, while the maximum singular value is bounded by M with probability tending to 1. The lower bound condition ensures identifiability and diversity among the candidate models. It implies that the prediction vectors from latent space models with different dimensions are not perfectly linearly dependent. In the context of latent space models, the latent dimension determines the geometric complexity of the network structure. Consequently, a model with dimension d typically yields a probability matrix distinct from one with a different dimension, which prevents multicollinearity. We acknowledge that the lower bound condition on $\lambda_{\min}(\Lambda/|\Psi_1|)$ may weaken when the candidate set includes models with dimensions close to the true dimension d_0. In particular, when both d_0 and d_0+1 are in the candidate set, the prediction vectors $\widehat{p}_{1(d_0)}$ and $\widehat{p}_{1(d_0+1)}$ may converge to similar limiting values, as the additional latent factor in the (d_0+1)-dimensional model captures increasingly negligible signal. In such a transitional regime, $\lambda_{\min}(\Lambda/|\Psi_1|)$ could approach zero, and the convergence rate in Theorem 3 would be affected. Importantly, this does not affect the validity of Theorems 1 and 2, which do not rely on Assumption 4. Moreover, for any finite network size N, the estimation

noise in the PGD algorithm generically prevents exact collinearity among the prediction vectors, so Assumption 4 is satisfied for fixed N and is primarily a concern in the asymptotic limit. We also note that this type of identifiability condition is standard in the model averaging literature (e.g., Condition (C.4) in Liao and Zou (2020)).

The upper bound condition guarantees numerical stability. Since the estimated probabilities are derived from the logistic function, every entry is strictly bounded in the unit interval. We verify the upper bound by exploiting the boundedness of these probability estimates. Specifically, the maximum singular value is bounded by the Frobenius norm as

$$\lambda_{\max}(\Lambda/|\Psi_1|) \leq |\Psi_1|^{-1}\|\Lambda_1\|_F^2 \leq M.$$

Assumption 5 requires the maximum singular value of the error covariance matrix $\Omega/|\Psi_1|$ to be bounded by M. This condition ensures that the variance of prediction errors does not diverge. According to the definition of Ω and Ω_1, the absolute values of the entries in Ω_1 are strictly bounded by 1. Therefore, we have

$$\lambda_{\max}(\Omega/|\Psi_1|) \leq |\Psi_1|^{-1}\|\Omega_1\|_F^2 \leq M.$$

Assumption 6 imposes constraints on the growth rates of the number of candidate models M, the network size N, and the risk ξ. In practice, we typically average over a small finite set of dimensions, so M is usually a small constant relative to the network size N, which satisfies the condition $M = o(N^{\min\{4\kappa,1/3\}})$. Furthermore, recall the decomposition of the risk function $R(w)$:

$$\begin{aligned} R(w) &= \sum_{(i,j)\in\Psi_1} E[\widehat{P}_{ij}(w) - P_{ij}]^2 \\ &= \sum_{(i,j)\in\Psi_1} \big\{(P^*_{ij}(w) - P_{ij})^2 + E[\widehat{P}_{ij}(w) - P^*_{ij}(w)]^2 \\ &\quad + 2(P^*_{ij}(w) - P_{ij})E[\widehat{P}_{ij}(w) - P^*_{ij}(w)]\big\} \end{aligned}$$

According to Assumption 1, we have $E[\sum_{(i,j)\in\Psi_1}(\widehat{P}_{ij}(w) - P^*_{ij}(w))^2] = O(NM)$. Applying Jensen's inequality, we further obtain

$$\sum_{(i,j)\in\Psi_1} (E[\widehat{P}_{ij}(w) - P^*_{ij}(w)])^2 \le E\Big[\sum_{(i,j)\in\Psi_1} (\widehat{P}_{ij}(w) - P^*_{ij}(w))^2\Big] = O(NM).$$

Combining these bounds with the Cauchy–Schwarz inequality for the cross term, it follows that

$$\frac{R(w)}{L^*(w)} \ge 1 + \frac{O(NM)}{L^*(w)} - 2\left\{\frac{O(NM)}{L^*(w)}\right\}^{1/2},$$

By Assumption 2, $NM\xi^{*-1} = o(1)$, which implies $R(w)/L^*(w) \geq 1 + o(1)$. Consequently, $\xi = \inf_{w\in\mathcal{W}} R(w) \geq (1 + o(1))\xi^*$. We now verify Assumption 6. Under Assumption 2, the mis-specified model condition $NM(\xi^*)^{-1} = o(1)$ holds. Substituting $\xi \geq (1 + o(1))\xi^*$ yields

$$\frac{N^{1-4\kappa}M^2}{\xi} \leq \left\{\frac{NM}{\xi^*(1+o(1))}\right\}\left\{\frac{M}{N^{4\kappa}}\right\} = o(1).$$

Similarly, using the condition $N\max\{D, \log N\}(\xi^*)^{-1} = O(1)$ from Assumption 2, we have

$$\frac{N^{1-4\kappa}\max\{D,\log N\}M}{\xi} \leq \frac{1}{1+o(1)}\left\{\frac{M}{N^{4\kappa}}\right\}\left(\frac{N\max\{D,\log N\}}{\xi^*}\right) = o(1).$$

Therefore, Assumption 6 follows directly from Assumption 2.

S4 Algorithm for Parameter Estimation in Multi-layer Networks

Algorithm 1 Projected Gradient Descent Algorithm for Parameter Estimation in Multi-layer Networks

Input: network adjacency matrix $\{A^{(t)}\}_{t=1}^T$; latent space dimension $\{d_t\}_{t=1}^T$; initial values α_0 and $\mathbb{Z}_0 = \{Z_0^{(1)}, \ldots, Z_0^{(T)}\}$; step sizes η_α and η_Z; number of iterations U

Output: $\widehat{\alpha} = \alpha_U$, $\widehat{Z}^{(t)} = Z_U^{(t)}$ for $t = 1, \ldots, T$

1: **for** $u = 0, 1, \ldots, U-1$ **do**
2: **for** $t = 1, \ldots, T$ **do**
3: $\Theta_u^{(t)} = \alpha_u 1_N^\top + 1_N \alpha_u^\top + Z_u^{(t)} Z_u^{(t)\top}$
4: $Z_{u+1}^{(t)} = Z_u^{(t)} + 2\eta_Z \left\{A^{(t)} - \sigma\left(\Theta_u^{(t)}\right)\right\} Z_u^{(t)}$ ▷ $\sigma(x) = 1/\{1+\exp(-x)\}$

$Z_{u+1}^{(t)} = JZ_{u+1}^{(t)}$
end for
$\alpha_{u+1} = \alpha_u + 2\eta_\alpha \left[\sum_{t=1}^{T} \left\{A^{(t)} - \sigma\left(\Theta_u^{(t)}\right)\right\}\right] 1_N$
end for

S5 Procedure for Selecting Model Weights in Multi-Layer Networks

Step 1: Partition the nodal pairs, denoted as $\Psi = \{(i,j) : i,\ j = 1,\ldots,N\}$, into two parts symmetrically, namely Ψ_1 and Ψ_2. The estimator $\widehat{Z}_{(m)}^{(t)} \in \mathbb{R}^{N\times m}$ and $\widehat{\alpha}_{(m)} \in \mathbb{R}^N$ for the mth model is obtained by applying Algorithm 1 to $A^{(t)}(\Psi_1)$. Then we can compute the estimators of $P_{ij}^{(t)}$ for the mth model by $\widehat{P}_{(m),ij}^{(t)} = f_2\left(\widehat{\alpha}_{(m),i} + \widehat{\alpha}_{(m),j} + \widehat{z}_{(m),i}^{(t)\top}\widehat{z}_{(m),j}^{(t)}\right)$, where $f_2(x) = |\Psi|/\{(1+\exp(-x))|\Psi_1|\}$, and $\widehat{z}_{(m),i}^{(t)} \in \mathbb{R}^m$ denotes the ith row of $\widehat{Z}_{(m)}^{(t)}$. Subsequently, the nodal pairs in Ψ_1 are evenly divided into K groups, with $G_k,\ k = 1,\ldots,K$, representing the set of nodal pairs in the kth group.

Step 2: For $k = 1,\ldots,K$,

(a) Exclude the nodal pairs in the kth group from Ψ_1, and utilize the remaining nodal pairs in Ψ_1 to compute the estimators of $Z_{(m)}^{(t)}$ and α in the mth model ($m = 1,\ldots,M$), which are $\widetilde{Z}_{(m)}^{(t)[-k]}$ and $\widetilde{\alpha}_{(m)}^{[-k]}$, respectively.

(b) Calculate the predictions for observations within the kth group for each model. That is, we calculate the prediction of $P^{(t)} \circ S^{[k]}$ by $\widetilde{P}_{(m)}^{(t)[k]} = f_1\left(\widetilde{\alpha}_{(m)}^{[-k]} 1_N^\top + 1_N \widetilde{\alpha}_{(m)}^{[-k]\top} + \widetilde{Z}_{(m)}^{(t)[-k]} \widetilde{Z}_{(m)}^{(t)[-k]\top}\right) \circ S^{[k]}$.

Step 3: Construct the K-fold edge cross-validation criterion $CV_\dagger(w) = \frac{1}{T|\Psi_1|} \sum_{t=1}^{T} \sum_{k=1}^{K} \left\| A^{(t)[k]} - \widetilde{P}^{(t)[k]}(w) \right\|_F^2$, where $A^{(t)[k]} = A^{(t)} \circ S^{[k]}$ and $\widetilde{P}^{(t)[k]}(w) = \sum_{m=1}^{M} w_m \widetilde{P}_{(m)}^{(t)[k]}$.

Step 4: Select the model weights by minimizing the K-fold edge cross-validation criterion, i.e., $\widehat{w}_\dagger = \operatorname{argmin}_{w \in \mathcal{W}} CV_\dagger(w)$, where $\mathcal{W} = \left\{ w \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\}$. Subsequently, construct an averaging prediction for $P_{ij}^{(t)}$, $(i,j) \in \Psi_2$ through $\widehat{P}_{ij}^{(t)}(\widehat{w}_\dagger) = \sum_{m=1}^{M} \widehat{w}_{\dagger m} \widehat{P}_{(m),ij}^{(t)}$.

Minimizing $CV_\dagger(w)$ can be reformulated as a quadratic programming problem concerning w. Therefore, we can use quadratic programming to solve the K-fold edge cross-validation weights. It is worth noting that assuming latent dimensions are the same across networks can be restrictive in practice. One might argue that performing cross-validation separately for each network allows for layer-specific weight selection and fits better when layers are extremely heterogeneous. However, we adopt the joint cross-validation strategy to minimize the aggregated loss across the entire system. By pooling information from all layers to determine a common set

of weights, our method borrows strength across layers and results in more robust and stable weight estimation.

To clarify the nature of this joint strategy, we emphasize that the phrase "borrow strength across layers" refers specifically to the weight estimation stage, not to the model estimation stage. In the model estimation stage, structural heterogeneity across layers is fully respected: for a given candidate dimension m, the PGD algorithm estimates a shared α and layer-specific latent positions $Z_{(m)}^{(t)}$ for each $t = 1, \ldots, T$. The "borrowing" occurs only when determining the shared weight vector $w = (w_1, \ldots, w_M)$ via the joint cross-validation objective $CV_\dagger(w) = (T|\Psi_1|)^{-1} \sum_{t=1}^{T} \sum_{k=1}^{K} \|A^{(t)[k]} - \widetilde{P}^{(t)[k]}(w)\|_F^2$. We note that the model averaging operates at the probability matrix level: $\widehat{P}^{(t)}(w) = \sum_m w_m \widehat{P}_{(m)}^{(t)}$, rather than at the latent vector level. Although different layers may have distinct latent structures $Z^{(t)}$, the shared intercept α provides a common baseline across layers, so that the cross-validation error in each layer primarily reflects how well each candidate dimension m approximates the layer-specific latent structure. Aggregating these cross-validation errors across T layers therefore averages out layer-specific noise in the weight estimation, yielding more stable weight estimates than cross-validating each layer separately. This is reflected theoretically in Theorem S3, where the convergence rate of $\widehat{w}_\dagger$ toward $w_\dagger^0$

includes a factor of $T^{-1/2}$, showing that additional layers accelerate the convergence of the estimated weights.

The shared intercept parameter α plays an important role in this framework. Intuitively, when different layers share the same α, the candidate models' relative predictive performance across layers is driven primarily by how well each candidate dimension m approximates the layer-specific latent structures $Z^{(t)}$. This consistency in the evaluation criterion across layers is what makes the aggregation of cross-validation errors across layers informative. If, instead, the baseline parameters α were allowed to differ across layers (i.e., $\alpha^{(t)}$ varies with t), the model estimation stage can still be extended straightforwardly by fitting layer-specific intercepts. However, the theoretical analysis would require additional regularity conditions to ensure that the candidate models' relative performance remains comparable across layers, and the gain from pooling may diminish if the layers become too heterogeneous. In the extreme case where the layers are completely unrelated, forcing a common weight vector could be suboptimal, and layer-specific cross-validation would be more appropriate. We leave the formal extension to layer-specific $\alpha^{(t)}$ as a direction for future research.

As shown in the results of Case 1 in Section S6.2, where true latent dimensions vary significantly across layers, NetMA consistently outperforms

other methods. This empirical evidence suggests that the gain in estimation stability from the joint approach outweighs the potential limitation of using shared weights.

S6 Simulation Results

S6.1 Single-layer Networks

Case 3 (Network density). In this case, we examine the influence of network density. Specifically, we fix the network size $N = 200$ and the true dimension of latent space $d_0 = 6$. The expected average nodal degree ranges from 10 to 80.

We also design four different settings ($M = 2,\ 4,\ 6,$ and 8) to show the performance of our method. Figure S1 presents the results of the three settings for Case 3. It can be seen from Figure S1 that NetMA performs the best for various M. In addition, when the network density is small, the "oracle" performs very poorly. This also verifies what we mentioned before, i.e., when the network size and the density are small but the real dimension of the latent space is relatively large, it is challenging to estimate the model based on the true dimension. Under this situation, the advantage of NetMA is very obvious.

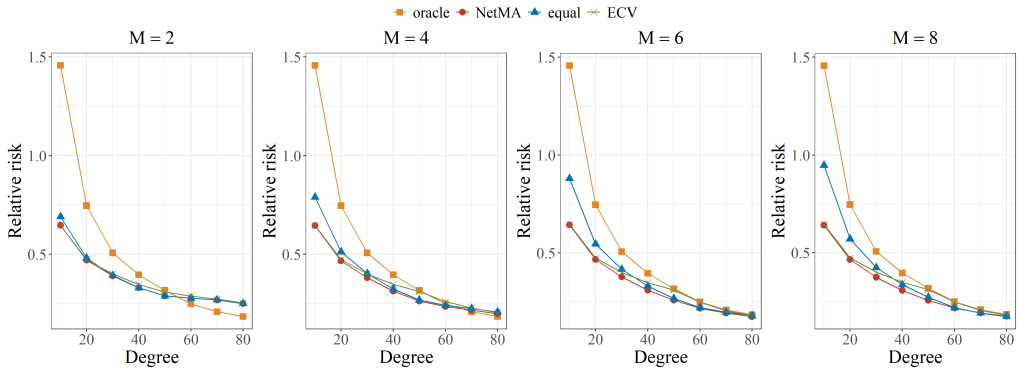

Figure S1: Relative risks of the four methods with different network densities in single-layer networks.

S6.2 Multi-layer Networks

In this section, we conduct some simulations on multi-layer networks, and set $T = 5$. The data-generating process is presented as follows. For the latent vectors, we generate $Z^{(t)} \in \mathbb{R}^{N \times d_{0,t}}$ where $d_{0,t}$ is the true dimension of the latent vector in the tth layer by first generating i.i.d $N(0, 1)$ entries and then transforming it according to the steps introduced in single-layer networks for $t = 1, \ldots, 5$. Generate the degree heterogeneity parameters $\alpha = (\alpha_1, \ldots, \alpha_N)^\top$ with each entry independent and identically generated from Uniform$(-1, 1)$. Then we can calculate the probability matrix $P^{(t)} \in \mathbb{R}^{N \times N}$ by $\text{logit}(P_{ij}^{(t)}) = \alpha_i + \alpha_j + z_i^{(t)\top} z_j^{(t)}$. Similarly, we can transform $P^{(t)}$ by multiplying a constant to control the expected average degree. To evaluate the performance of the estimation methods, we intro-

duce the relative empirical risk function for multi-layer networks, which is defined as follows,

$$\widehat{R}_{\dagger}(\widehat{w}_{\dagger}) = \frac{1}{Q}\sum_{q=1}^{Q}\frac{\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_2}\left\{\widehat{P}_{ij}^{(t)\{q\}}(\widehat{w}_{\dagger}^{\{q\}}) - P_{ij}^{(t)\{q\}}\right\}^2}{\sum_{t=1}^{T}\sum_{(i,j)\in\Psi_2}\left\{P_{ij}^{(t)\{q\}}\right\}^2},$$

where Q is the number of simulation replications. In the following cases, we set $Q = 100$, $K = 5$ and $|\Psi_1|/|\Psi_2| = 7:3$. Next, we design three cases and show the results of the simulations.

Case 1 (dimension of latent space). In this case, we consider a 5-layer network with the number of nodes in each layer $N = 200$. The expected average nodal degree is set to be 60. The number of the candidate models ranges from 1 to 10.

For Case 1, we design three settings to evaluate the performances of the four methods. The true dimensions of the latent vectors in the five layers for the three settings are $\{3, 3, 3, 3, 8\}$, $\{2, 4, 6, 8, 10\}$ and $\{8, 8, 8, 8, 3\}$, respectively. Figure S2 shows the simulation results of the three different settings from left to right. It's easy to see that as the true dimension increases, the risk of estimating the parameters also increases. Additionally, it is obvious that when $M \geq 3$, NetMA performs the best under different settings, followed by simple averaging. Compared to the "oracle" and ECV,

NetMA has a substantial advantage, especially when M is large. Except for the "oracle", the effectiveness of the other three methods is poor when M is small. This is also easy to understand, as in this situation, the candidate models are not sufficient for modelling the network. However, as M gradually increases, the relative risks of these three methods decrease sharply and eventually stabilize. Additionally, it can be observed that in the first two settings, there always exists a certain gap between the relative risks of the ECV method and those of the "oracle".

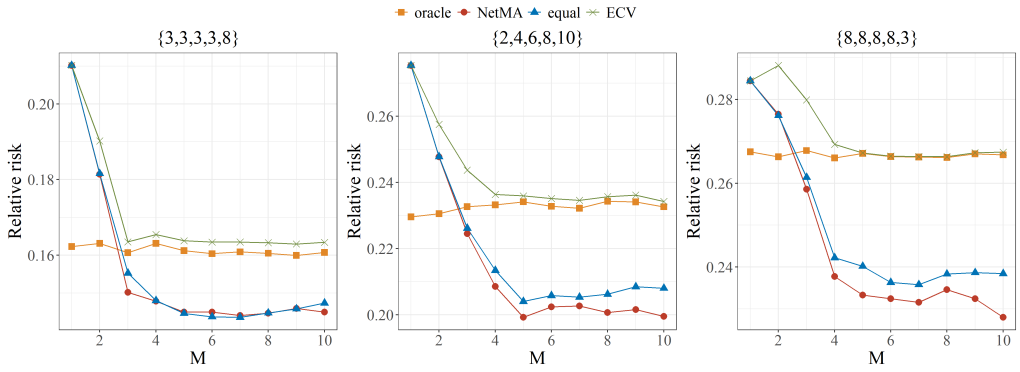

Figure S2: Relative risks of the four methods with different candidate models in multi-layer networks.

Case 2 (Number of nodes). In this case, we explore the performance differences of the four methods as the network size increases. Specifically, we construct a 5-layer network and vary the number of nodes in each layer from 100 to 500. The true dimensions of the latent vectors in the five layers are set to be 2, 4, 6, 8, and 10 respectively. In each layer, the expected

average nodal degree is $0.3(N-1)$.

We set $M = 4$, 6 and 8. The results are shown in Figure S3. As we can see, when the sample size is small ($N = 100, 200$), NetMA almost always outperforms other methods, including the "oracle" for all settings. As the sample size increases (e.g., $N = 300, 400, 500$), if M is small (such as $M = 4$), the "oracle" surpasses NetMA. However, when M increases (e.g., $M = 6, 8$), NetMA outperforms the "oracle". Additionally, NetMA generally outperforms ECV and the "equal" methods.

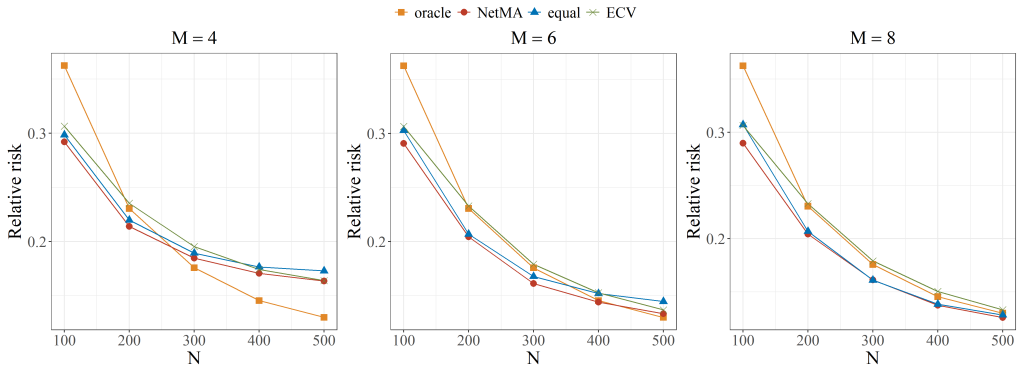

Figure S3: Relative risks of the four methods with different network sizes in multi-layer networks.

Case 3 (Network density). In this case, we study the performance of the four methods under different network densities. The network consists of 5 layers, with 200 nodes in each layer. The true dimensions of the latent vectors in the 5 layers are 2, 4, 6, 8, and 10, respectively. We vary the expected average nodal degree from 20 to 80.

Figure S4 shows the simulation results of Case 3 when $M = 4,\ 6,\ 8$ respectively, which are similar to the results of Figure S1. Specifically, NetMA performs the best and is stable under different M and network densities.

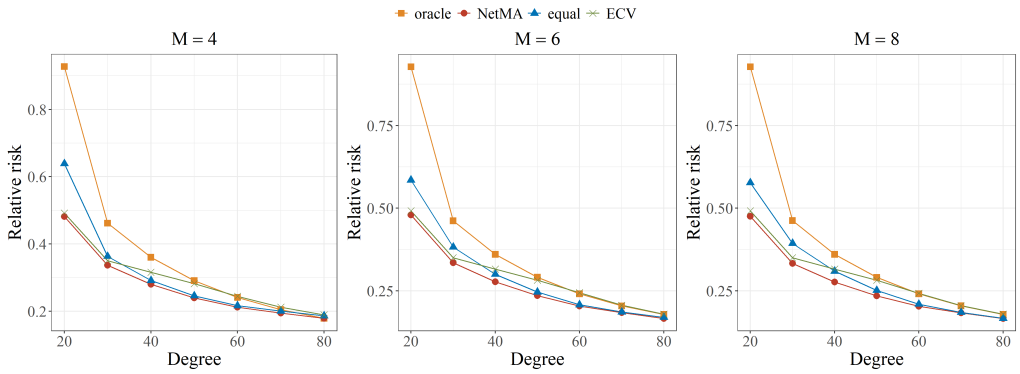

Figure S4: Relative risks of the four methods with different network densities in multilayer networks.

S7 Empirical Results

S7.1 Addtional Results for Virtual World Data

S7.2 Prediction on ResearchGate Data

In this section, we apply the proposed method NetMA to real-world data. In practical scenarios, as we do not know the true value of the latent vectors, it is infeasible to assess the model performance in terms of latent space estimation. Instead, we consider the downstream task based on the estimated

Table S1: Average weights estimated by ECV and NetMA on virtual world data.

	Model	ECV ($K = 5$)	ECV ($K = 10$)	NetMA ($K = 5$)	NetMA ($K = 10$)
$M = 4$	1	0.0000	0.0000	0.2856	0.1905
	2	0.8200	0.5900	0.2794	0.3311
	3	0.1100	0.0700	0.0400	0.0143
	4	0.0700	0.3400	0.3950	0.4641
$M = 6$	1	0.0000	0.0000	0.2793	0.1838
	2	0.6600	0.2700	0.2594	0.3060
	3	0.0800	0.0100	0.0173	0.0041
	4	0.0100	0.0300	0.0212	0.0068
	5	0.0100	0.0100	0.0227	0.0012
	6	0.2400	0.6800	0.4001	0.4981
$M = 8$	1	0.0000	0.0000	0.2584	0.1651
	2	0.3800	0.0400	0.2375	0.2848
	3	0.0600	0.0000	0.0063	0.0008
	4	0.0000	0.0100	0.0020	0.0020
	5	0.0000	0.0000	0.0007	0.0013
	6	0.0300	0.0200	0.0085	0.0071
	7	0.4300	0.7500	0.1353	0.1640
	8	0.1000	0.1800	0.3512	0.3748

latent space representation such as link prediction, community detection and node classification. Link prediction is a popular procedure in network science to predict missing or future links (Lü and Zhou, 2011; Kumar et al., 2020). Here, we compare the performance of NetMA, ECV and a simple averaging method on link prediction. The three methods are examined by (i) changing the number of candidate models, (ii) changing the ratio of $|\Psi_1|$ and $|\Psi_2|$, and (iii) changing the number of folds of cross-validation.

Scientific collaboration refers to the cooperative behaviour between two or more researchers, such as joining together in developing projects or research. One of the key concerns for researchers is how to find suitable collaborators. Scientific and academic social networks, such as LinkedIn[1],

[1]https://www.linkedin.com/

ResearchGate[2], and Mendeley[3], provide scholars with an opportunity to understand the current research directions of others. Roozbahani et al. (2021) collect a dataset called ResearchGate dataset for Rcommending Systems (RGRS) which contains the information of 266,207 members of ResearchGate. The dataset provides information on followers and following. According to this structural information, we construct a network in which the nodes represent researchers and the edges represent the follower/following relationship. Although there is a directed relationship of following or being followed between two scholars, for convenience, we only consider whether there is a follower or following relationship between them. If such a relationship exists, there is an edge between the two scholars; if it does not exist, there is no edge. Thus, the network is undirected and unweighted. Many real networks have core-periphery structures, where the core contains nodes that are intensively connected and the periphery contains nodes that are weakly connected to the core. The core nodes are the main focus of our research. To abstract core nodes, we first remove nodes whose degrees are no more than 30 and the edges connected with them. Then we abstract the 24-core network according to the idea of Wang and Rohe (2016). Specifically, a c-core is obtained by removing nodes whose number of neighbours

[2]https://www.researchgate.net/
[3]https://www.mendeley.com/

is less than c and the edges connected to them. Keep iterating this step until the nodes in the network no longer change. Finally, the core network has 397 nodes and 8,415 edges. The density of the network is 0.107.

We use the model averaging and model selection methods to analyse the core network. The candidate models are latent space models with different latent space dimensions, which are presented in Model (2.1). The dimensions of the latent space for the mth ($m = 1, \ldots, M$) model is m and we let $M = 2,\ 4,\ 6,$ and 8. We consider three scenarios for $\pi_{12} = |\Psi_1|/|\Psi_2|$: 7/3, 8/2, and 9/1. In addition, when selecting weights for NetMA and models for ECV, we take $K = 5$ and $K = 10$ for the K-fold cross-validation. The experiments are replicated 100 times, and we report results for link prediction in Table S2. We use four metrics to compare the prediction performance. Specifically, AUROC is the area under the ROC curve; AUPR is the area under the precision-recall curve; MLogf is the average value of the log-likelihood function; and MSE is the mean squared error. The bold numbers in the table indicate the best results under different metrics in a setting. The results show that the proposed NetMA method performs better than ECV and simple averaging in most scenarios. Besides, the prediction performance of NetMA is quite similar for $K = 5$ and $K = 10$.

Table S2: Evaluation for link prediction on ResearchGate data (standard errors are in brackets).

Metric	Method	M=2			M=4			M=6			M=8		
		$\pi_{12} = 7/3$	8/2	9/1	$\pi_{12} = 7/3$	8/2	9/1	$\pi_{12} = 7/3$	8/2	9/1	$\pi_{12} = 7/3$	8/2	9/1
AUROC	equal	0.7983	0.8022	0.8069	0.8194	0.8235	0.8279	0.8363	0.8415	0.8464	0.8414	0.8476	0.8531
		(0.0053)	(0.0065)	(0.0082)	(0.0047)	(0.0060)	(0.0081)	(0.0038)	(0.0050)	(0.0072)	(0.0036)	(0.0046)	(0.0070)
	ECV (K=5)	0.8009	0.8062	**0.8118**	0.8217	0.8269	0.8331	0.8373	0.8451	0.8516	0.8375	0.8454	0.8522
		(0.0053)	(0.0067)	**(0.0081)**	(0.0055)	(0.0079)	(0.0082)	(0.0033)	(0.0040)	(0.0067)	(0.0033)	(0.0039)	(0.0067)
	ECV (K=10)	0.8009	0.8062	**0.8118**	0.8224	0.8284	**0.8341**	0.8375	0.8451	0.8516	0.8376	0.8455	0.8526
		(0.0053)	(0.0067)	**(0.0081)**	(0.0039)	(0.0061)	**(0.0073)**	(0.0033)	(0.0040)	(0.0067)	(0.0033)	(0.0039)	(0.0065)
	NetMA (K=5)	0.8018	0.8063	0.8111	0.8226	0.8265	0.8317	0.8419	0.8478	0.8526	0.8428	0.8492	0.8544
		(0.0054)	(0.0067)	(0.0081)	(0.0056)	(0.0079)	(0.0081)	(0.0034)	(0.0044)	(0.0070)	(0.0033)	(0.0043)	(0.0069)
	NetMA (K=10)	**0.8020**	**0.8066**	0.8116	**0.8237**	**0.8285**	0.8334	**0.8421**	**0.8484**	**0.8535**	**0.8431**	**0.8498**	**0.8554**
		(0.0054)	**(0.0067)**	(0.0081)	**(0.0040)**	**(0.0063)**	(0.0076)	**(0.0033)**	**(0.0043)**	**(0.0069)**	**(0.0034)**	**(0.0043)**	**(0.007)**
AUPR	equal	0.3812	0.3890	0.3965	0.4140	0.4230	0.4320	0.4423	0.4548	0.4658	0.4530	0.4686	0.4820
		(0.0098)	(0.0139)	(0.0170)	(0.0100)	(0.0143)	(0.0170)	(0.0088)	(0.0131)	(0.0167)	(0.0083)	(0.0127)	(0.0169)
	ECV (K=5)	0.3816	0.3930	0.4039	0.4092	0.4236	0.4383	0.4353	0.4564	0.4754	0.4359	0.4575	0.4773
		(0.0100)	(0.0144)	(0.0166)	(0.0104)	(0.0153)	(0.0180)	(0.0098)	(0.0129)	(0.0180)	(0.0096)	(0.0125)	(0.0183)
	ECV (K=10)	0.3816	0.3930	0.4039	0.4104	0.4256	0.4397	0.4356	0.4564	0.4754	0.4360	0.4579	0.4787
		(0.0100)	(0.0144)	(0.0166)	(0.0088)	(0.0138)	(0.0171)	(0.0096)	(0.0129)	(0.0180)	(0.0094)	(0.0120)	(0.0173)
	NetMA (K=5)	**0.3857**	0.3953	0.4041	0.4174	0.4287	0.4402	**0.4530**	**0.4696**	0.4831	**0.4549**	**0.4731**	0.4880
		(0.0098)	(0.0142)	(0.0169)	(0.0110)	(0.0157)	(0.0171)	**(0.0088)**	**(0.0127)**	(0.0169)	**(0.0086)**	**(0.0122)**	(0.0174)
	NetMA (K=10)	0.3854	**0.3955**	**0.4047**	**0.4179**	**0.4309**	**0.4423**	0.4515	0.4693	**0.4838**	0.4536	0.4730	**0.4889**
		(0.0098)	**(0.0142)**	**(0.0169)**	**(0.0087)**	**(0.0140)**	**(0.0168)**	(0.0089)	(0.0126)	**(0.0171)**	(0.0084)	(0.0123)	**(0.0175)**
Mlogf	equal	**-0.2951**	-0.2902	-0.2855	**-0.2821**	-0.2772	-0.2727	-0.2700	-0.2646	-0.2599	**-0.2669**	-0.2607	-0.2554
		(0.0054)	(0.0063)	(0.0087)	**(0.0048)**	(0.0055)	(0.0079)	(0.0044)	(0.0050)	(0.0075)	**(0.0046)**	(0.0052)	(0.0078)
	ECV (K=5)	-0.3048	-0.2968	-0.2896	-0.2993	-0.2890	-0.2804	-0.3017	-0.2862	-0.2734	-0.3021	-0.2869	-0.2737
		(0.0059)	(0.0072)	(0.0095)	(0.0053)	(0.0065)	(0.0089)	(0.0053)	(0.0059)	(0.0090)	(0.0054)	(0.0063)	(0.0094)
	ECV (K=10)	-0.3048	-0.2968	-0.2896	-0.2992	-0.2886	-0.2801	-0.3017	-0.2862	-0.2734	-0.3023	-0.2873	-0.2737
		(0.0059)	(0.0072)	(0.0095)	(0.0053)	(0.0061)	(0.0090)	(0.0053)	(0.0059)	(0.0090)	(0.0056)	(0.0062)	(0.0093)
	NetMA (K=5)	-0.2962	**-0.2900**	**-0.2846**	-0.2830	**-0.2770**	-0.2713	**-0.2680**	**-0.2609**	**-0.2554**	-0.2673	**-0.2598**	**-0.2539**
		(0.0055)	**(0.0066)**	**(0.0089)**	(0.0051)	**(0.0067)**	(0.0081)	**(0.0043)**	**(0.0046)**	**(0.0075)**	(0.0045)	**(0.0049)**	**(0.0078)**
	NetMA (K=10)	-0.2972	-0.2905	-0.2848	-0.2842	**-0.2770**	**-0.2711**	-0.2703	-0.2619	-0.2557	-0.2695	-0.2607	-0.2541
		(0.0055)	(0.0066)	(0.0089)	(0.0048)	**(0.0058)**	**(0.0082)**	(0.0043)	(0.0047)	(0.0076)	(0.0046)	(0.0050)	(0.0079)
MSE	equal	0.0817	0.0810	0.0802	0.0790	0.0782	0.0774	0.0766	0.0755	0.0746	0.0757	0.0744	0.0732
		(0.0012)	(0.0015)	(0.0023)	(0.0012)	(0.0015)	(0.0022)	(0.0012)	(0.0014)	(0.0021)	(0.0012)	(0.0014)	(0.0022)
	ECV (K=5)	0.0822	0.0812	0.0801	0.0802	0.0788	0.0775	0.0788	0.0767	0.0748	0.0787	0.0766	0.0747
		(0.0013)	(0.0016)	(0.0024)	(0.0012)	(0.0016)	(0.0023)	(0.0012)	(0.0013)	(0.0022)	(0.0012)	(0.0013)	(0.0022)
	ECV (K=10)	0.0822	0.0812	0.0801	0.0801	0.0787	0.0774	0.0788	0.0767	0.0748	0.0788	0.0766	0.0746
		(0.0013)	(0.0016)	(0.0024)	(0.0012)	(0.0015)	(0.0023)	(0.0012)	(0.0013)	(0.0022)	(0.0012)	(0.0013)	(0.0022)
	NetMA (K=5)	**0.0816**	**0.0807**	**0.0798**	**0.0789**	**0.0778**	0.0767	**0.0758**	**0.0743**	**0.0731**	**0.0756**	**0.0740**	**0.0727**
		(0.0012)	**(0.0016)**	**(0.0024)**	**(0.0013)**	**(0.0016)**	(0.0022)	**(0.0011)**	**(0.0013)**	**(0.0021)**	**(0.0012)**	**(0.0013)**	**(0.0022)**
	NetMA (K=10)	0.0817	**0.0807**	**0.0798**	0.0790	**0.0778**	**0.0766**	0.0762	0.0745	**0.0731**	0.0759	0.0741	**0.0727**
		(0.0012)	**(0.0016)**	**(0.0024)**	(0.0012)	**(0.0015)**	**(0.0022)**	(0.0011)	(0.0013)	**(0.0021)**	(0.0012)	(0.0013)	**(0.0022)**

Next, we record the weights of the candidate models obtained by ECV and NetMA in each replication when $\pi_{12} = 9/1$. Table S3 shows the mean of the weights based on 100 replications for each method under $K = 5$ and $K = 10$. It can be seen that for this network, NetMA tends to put more weights on the candidate models with larger dimensions of the latent space, and ECV shows a preference for selecting candidate models with higher dimensional latent spaces.

Table S3: Average weights estimated by ECV and NetMA on ResearchGate data.

	Model	ECV ($K = 5$)	ECV ($K = 10$)	NetMA ($K = 5$)	NetMA ($K = 10$)
$M = 2$	1	0.0000	0.0000	0.0449	0.0077
	2	1.0000	1.0000	0.9551	0.9923
$M = 4$	1	0.0000	0.0000	0.0344	0.0029
	2	0.0000	0.0000	0.1439	0.1094
	3	0.1600	0.0700	0.1530	0.0873
	4	0.8400	0.9300	0.6687	0.8004
$M = 6$	1	0.0000	0.0000	0.0080	0.0000
	2	0.0000	0.0000	0.1177	0.0587
	3	0.0000	0.0000	0.0536	0.0365
	4	0.0000	0.0000	0.1247	0.1330
	5	0.0000	0.0000	0.0949	0.0959
	6	1.0000	1.0000	0.6010	0.6760

Bibliography

Chafai, D., O. Guedon, G. Lecue, and A. Pajor (2009). Singular values of random matrices. In Lecture Notes.

Gao, Y., X. Zhang, S. Wang, T. T.-l. Chong, and G. Zou (2019). Frequentist model averaging for threshold models. Annals of the Institute of Statistical Mathematics 71, 275–306.

Kumar, A., S. S. Singh, K. Singh, and B. Biswas (2020). Link prediction techniques, applications, and performance: A survey. Physica A: Statistical Mechanics and its Applications 553, 124289.

Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models. The Annals of Statistics 43(1), 215 – 237.

Liao, J. and G. Zou (2020). Corrected mallows criterion for model averaging. Computational Statistics & Data Analysis 144, 106902.

Liao, J., G. Zou, Y. Gao, and X. Zhang (2021). Model averaging prediction for time series models with a diverging number of parameters. Journal of Econometrics 223(1), 190–221.

Lü, L. and T. Zhou (2011). Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications 390(6), 1150–1170.

Ma, Z., Z. Ma, and H. Yuan (2020). Universal latent space model fitting for large networks with edge covariates. Journal of Machine Learning Research 21(4), 1–67.

Roozbahani, Z., J. Rezaeenour, R. Shahrooei, and H. Emamgholizadeh (2021). Presenting a dataset for collaborator recommending systems in academic social network: A case study on reseachgate. Journal of Data, Information and Management 3, 29–40.

Wang, S. and K. Rohe (2016). Discussion of "coauthorship and citation networks for statisticians". The Annals of Applied Statistics 10(4), 1820–1826.

Zhang, X. and C.-A. Liu (2023). Model averaging prediction by K-fold cross-validation. Journal of Econometrics 235(1), 280–301.

Zhang, X., S. Xue, and J. Zhu (2020). A flexible latent space model for multilayer networks. In Proceedings of the 37th International Conference on Machine Learning, pp. 11288–11297.