Supplemental Document for "Matrix Completion with Covariate Information"

Xiaojun Mao*, Song Xi Chen† and Raymond K. W. Wong‡

September 24, 2017

Abstract

This document provides supplementary material to the article "Matrix Completion with Covariate Information" written by the same authors.

S1 Proof of Propositions

Proof of Proposition 1. We have

$$\begin{aligned}
&\mathsf{E}\,\|\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{B}-\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\|_F^2\\
=&\|\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{B}\|_F^2-2\left\langle\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{B},\mathsf{E}\left(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\right)\right\rangle+\mathsf{E}\,\|\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\|_F^2\\
=&\|(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{B})-(\boldsymbol{X}\boldsymbol{\beta}_0+\boldsymbol{B}_0)\|_F^2-\|\boldsymbol{X}\boldsymbol{\beta}_0+\boldsymbol{B}_0\|_F^2+\sum_{ij}\frac{\left((X\beta_0)_{ij}+B_{0ij}\right)^2+\sigma_{ij}^2}{\theta_{ij}},
\end{aligned}$$

due to Conditions C1(a) and C4. For any minimizer $(\boldsymbol{\beta}_s,\boldsymbol{B}_s)$ of R, we have $\boldsymbol{X}\boldsymbol{\beta}_0+\boldsymbol{B}_0=\boldsymbol{X}\boldsymbol{\beta}_s+\boldsymbol{B}_s$, which implies $\boldsymbol{X}(\boldsymbol{\beta}_0-\boldsymbol{\beta}_s)=\boldsymbol{B}_s-\boldsymbol{B}_0$. Since $\boldsymbol{B}_s-\boldsymbol{B}_0\in\mathcal{N}(\boldsymbol{X})$, we can conclude both $\boldsymbol{X}\boldsymbol{\beta}_s=\boldsymbol{X}\boldsymbol{\beta}_0$

*Xiaojun Mao is Ph.D. candidate, Department of Statistics, Iowa State University, Ames, IA 50011, USA (Email: mxjki@iastate.edu).

†Author of Correspondence. Song Xi Chen is Chair Professor, Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing 100651, China (Email: csx@gsm.pku.edu.cn). His research is partially supported by Chinas National Key Research Special Program Grants 2016YFC0207701 and 2015CB856000, and National Natural Science Foundation of China grants 11131002, 71532001 and 71371016.

‡Raymond K. W. Wong is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843, USA (Email: raywong@stat.tamu.edu). His research is partially supported by the National Science Foundation under Grants DMS-1612985 and DMS-1711952 (subcontract).

and $\boldsymbol{B}_s = \boldsymbol{B}_0$. As matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$ is invertible, we know that $\boldsymbol{\beta}_s = \boldsymbol{\beta}_0$. This also implies that $(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ is the unique minimizer. □

Proof of Proposition 2. By operator inequality and matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$ is invertible, we have $\|\boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B}\|_* \leq \|\boldsymbol{P}_{\boldsymbol{X}}^{\perp}\| \|\boldsymbol{B}\|_* \leq \|\boldsymbol{B}\|_*$. For any $\boldsymbol{B} \in \mathbb{R}^{n_1 \times n_2}$,

$$
\begin{aligned}
&\frac{1}{n_1 n_2} \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_2 \left(\alpha \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} \right\|_* + (1-\alpha) \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} \right\|_F^2 \right) \\
\leq & \frac{1}{n_1 n_2} \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \frac{1}{n_1 n_2} \|\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{B}\|_F^2 + \lambda_2 \left(\alpha \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} \right\|_* + (1-\alpha) \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} \right\|_F^2 \right) \\
&+ \lambda_2 (1-\alpha) \|\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{B}\|_F^2 \\
\leq & \frac{1}{n_1 n_2} \left\| \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_2 \left(\alpha \|\boldsymbol{B}\|_* + (1-\alpha) \|\boldsymbol{B}\|_F^2 \right),
\end{aligned}
$$

where the first inequality is strict whenever $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{B} \neq \boldsymbol{0}$. Therefore the solution of (3.7) belongs to $\mathcal{N}(\boldsymbol{X})$ and hence it is also a solution of (3.5). □

S2 Benefit of Covariate Information

Before discussing the benefit of using covariates, we need the following proposition which describes the relationship between $\|\boldsymbol{A}_0\|_*$ and $\|\boldsymbol{B}_0\|_*$.

Proposition S2.1. *Let $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$, where $\boldsymbol{B}_0 \in \mathcal{N}(\boldsymbol{X})$, we have $\|\boldsymbol{B}_0\|_* \leq \|\boldsymbol{A}_0\|_*$. If $\mathcal{R}(\boldsymbol{\beta}_0) \nsubseteq \mathcal{R}(\boldsymbol{B}_0)$, once $\boldsymbol{\beta}_0 \neq \boldsymbol{0}^{m \times n_2}$, we have $\|\boldsymbol{B}_0\|_* < \|\boldsymbol{A}_0\|_*$. Here $\mathcal{R}(\boldsymbol{Y})$ is the row space of a matrix $\boldsymbol{Y}$.*

Proof. For any $\boldsymbol{Z} \in \partial \|\boldsymbol{B}_0\|_*$, we have $\|\boldsymbol{A}_0\|_* \geq \|\boldsymbol{B}_0\|_* + \langle \boldsymbol{Z}, \boldsymbol{X}\boldsymbol{\beta}_0 \rangle$. Write the SVD of $\boldsymbol{B}_0$ as $\sum_{i=1}^{r_{\boldsymbol{B}_0}} \sigma_i(\boldsymbol{B}_0) \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T}$. Let $\mathcal{B}_{\boldsymbol{u}}$ be the linear span of $\boldsymbol{u}_{\boldsymbol{B}_0}^{(1)}, \ldots, \boldsymbol{u}_{\boldsymbol{B}_0}^{(r_{\boldsymbol{B}_0})}$ and $\mathcal{B}_{\boldsymbol{v}}$ be the linear span of $\boldsymbol{v}_{\boldsymbol{B}_0}^{(1)}, \ldots, \boldsymbol{v}_{\boldsymbol{B}_0}^{(r_{\boldsymbol{B}_0})}$. We have the fact that the sub-differential of the convex function $\boldsymbol{B}_0 \mapsto \|\boldsymbol{B}_0\|_*$ is the following set of matrices:

$$
\partial \|\boldsymbol{B}_0\|_* = \left\{ \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T} + \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}} \boldsymbol{Z} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}} : \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}} \boldsymbol{Z} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}} \right\| \leq 1 \right\}.
$$

On the other hand, by Lemma 3.2 in Candès and Recht (2009), there exist matrix $\bar{\boldsymbol{Z}}$ with $\|\bar{\boldsymbol{Z}}\| = 1$ such that $\langle \bar{\boldsymbol{Z}}, \boldsymbol{X}\boldsymbol{\beta}_0 \rangle = \|\bar{\boldsymbol{Z}}\| \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* = \|\boldsymbol{X}\boldsymbol{\beta}_0\|_*$. Pick $\boldsymbol{Z} \in \partial \|\boldsymbol{B}_0\|_*$ such that $\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \boldsymbol{Z} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} = \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \bar{\boldsymbol{Z}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}$, then we have

$$\begin{aligned}
\langle \boldsymbol{Z}, \boldsymbol{X}\boldsymbol{\beta}_0 \rangle &= \left\langle \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T} + \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \bar{\boldsymbol{Z}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}, \boldsymbol{X}\boldsymbol{\beta}_0 \right\rangle \\
&= 0 + \left\langle \bar{\boldsymbol{Z}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}, \boldsymbol{X}\boldsymbol{\beta}_0 \right\rangle = \left\langle \bar{\boldsymbol{Z}}, \boldsymbol{X}\boldsymbol{\beta}_0 \right\rangle - \left\langle \bar{\boldsymbol{Z}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}, \boldsymbol{X}\boldsymbol{\beta}_0 \right\rangle \\
&\geq \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* - \left\| \bar{\boldsymbol{Z}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}} \right\| \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* \geq \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* - \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* = 0.
\end{aligned}$$

Thus we show that $\|\boldsymbol{B}_0\|_* \leq \|\boldsymbol{A}_0\|_*$.

If $\mathcal{R}(\boldsymbol{\beta}_0) \nsubseteq \mathcal{R}(\boldsymbol{B}_0)$, it implies that $\boldsymbol{\beta}_0 \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}} \neq \boldsymbol{\beta}_0$. Thus for the inequality above, we always have $\langle \boldsymbol{Z}, \boldsymbol{X}\boldsymbol{\beta}_0 \rangle > 0$ which implies $\|\boldsymbol{A}_0\|_* > \|\boldsymbol{B}_0\|_*$. □

S2.1 Compare the Upper Bounds

As for $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0)$, it follows from the closed form of $\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}$ that

$$\begin{aligned}
\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}} - \boldsymbol{X}\boldsymbol{\beta}_0 =& \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1 \boldsymbol{I}_{m\times m})^{-1} n_1^{-1} \boldsymbol{X}^\intercal \left(\frac{n_1 n_2}{N} \boldsymbol{W} \circ \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 \right) \\
&- \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1 \boldsymbol{I}_{m\times m})^{-1} n_2\lambda_1 n_1^{-1} \boldsymbol{X}\boldsymbol{\beta}_0.
\end{aligned}$$

Take $\lambda_1 = o(n_2^{-1})$, $n_2\lambda_1 = o(1)$, we have $\boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1 \boldsymbol{I}_{m\times m})^{-1} n_1^{-1} \boldsymbol{X}^\intercal = \boldsymbol{P}_{\boldsymbol{X}}(1 + o(1))$. It implies that,

$$\frac{1}{n_1 n_2} \left\| \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}} - \boldsymbol{X}\boldsymbol{\beta}_0 \right\|_F^2 \leq \frac{1}{n_1 n_2} \left\| \boldsymbol{P}_{\boldsymbol{X}} \left(\frac{n_1 n_2}{N} \boldsymbol{W} \circ \boldsymbol{Y} - \boldsymbol{A}_0 \right) \right\|_F^2 (1 + o(1)) + n_2^2 \lambda_1^2 \left\| \boldsymbol{X}\boldsymbol{\beta}_0 \right\|_F^2 (1 + o(1)).$$

Let $\boldsymbol{P}_{\boldsymbol{X}} = (s_{ij})$, $\mathsf{E}\|\boldsymbol{P}_{\boldsymbol{X}}(\frac{n_1 n_2}{N}\boldsymbol{W} \circ \boldsymbol{Y} - \boldsymbol{A}_0)\|_F^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \mathsf{E}(\sum_{k=1}^{n_1} s_{ik}(n_1 n_2 \omega_{kj} Y_{kj}/N - A_{0kj}))^2 \leq 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\sum_{k=1}^{n_1} s_{ik}^2 \mathsf{E}(n_1 n_2 \omega_{kj} A_{0kj}/N - A_{0kj})^2 + \sum_{k=1}^{n_1} s_{ik}^2 \mathsf{E}(n_1 n_2 \omega_{kj}\epsilon_{kj}/N)^2)$. Due to Condition C1 and C4, we have $\max \mathsf{E}\epsilon_{ij}^2 \leq c_\sigma^2$ and $\|\boldsymbol{A}_0\|_\infty \leq \sqrt{\log(n)} a_1$. Since $\omega_{kj} \overset{\text{i.i.d}}{\sim} \mathrm{Bern}(\theta_0)$, we have

$$\begin{aligned}
\mathsf{E}\left(\frac{\omega_{kj}}{N}\right) =& \mathsf{E}\left(\frac{\omega_{kj}}{\omega_{kj} + \sum_{(s,t)\neq(k,j)} \omega_{st}} \right) = \mathsf{E}\left\{ \mathsf{E}\left(\frac{\omega_{kj}}{\omega_{kj} + c} \middle| \sum_{(s,t)\neq(k,j)} \omega_{st} = c \right) \right\} \\
=& \mathsf{E}\left\{ \sum_{c=0}^{n_1 n_2 - 1} \frac{\theta_0}{1 + c} \right\} = \frac{1}{n_1 n_2}\left(1 - (1 - \theta_0)^{n_1 n_2}\right) \leq \frac{1}{n_1 n_2},
\end{aligned}$$

and similarly,

$$\mathsf{E}\left(\frac{\omega_{kj}}{N^2}\right) = \mathsf{E}\left\{\frac{\omega_{kj}}{\left(\omega_{kj}+\sum_{(s,t)\neq(k,j)}\omega_{st}\right)^2}\right\} = \mathsf{E}\left\{\sum_{c=0}^{n_1n_2-1}\frac{\theta_0}{(1+c)^2}\right\} \leq \frac{2}{n_1n_2\left(n_1n_2+1\right)\theta_0}.$$

Combine the above two results together, we have

$$\begin{aligned}\mathsf{E}\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\frac{n_1n_2}{N}\boldsymbol{W}\circ\boldsymbol{Y}-\boldsymbol{A}_0\right)\right\|_F^2 \leq & 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left\{\left(\frac{2n_1n_2}{\left(n_1n_2+1\right)\theta_0}+2+1\right)\left\{\log\left(n\right)\right\}a_1^2\sum_{k=1}^{n_1}s_{ik}^2+\right.\\ & \left.\frac{2n_1n_2c_\sigma^2}{\left(n_1n_2+1\right)\theta_0}\sum_{k=1}^{n_1}s_{ik}^2\right\}\\ \leq & 2\left\{\left(\frac{2n_1n_2}{\left(n_1n_2+1\right)\theta_0}+3\right)\left\{\log\left(n\right)\right\}a_1^2+\frac{2n_1n_2c_\sigma^2}{\left(n_1n_2+1\right)\theta_0}\right\}n_2m.\end{aligned}$$

Take $\lambda_1 = o\{n_1^{-1}n_2^{-3/2}\log^{-1}(n)\}$, we have $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0) = O_p(n_1^{-1})$.

Proofs of Theorem 3. Under Condition C3, we have $\|\boldsymbol{A}_0\|_* = O\{\sqrt{n_1n_2\log(n)}\}$ and $\|\boldsymbol{B}_0\|_* = O\{\sqrt{n_1n_2\log(n)}\}$. Under the low rank condition that $r_{\boldsymbol{A}_0} = r_{\boldsymbol{B}_0} + m = o\{\theta_0^{1/2}(n_1\wedge n_2)^{1/2}\}$, we have $\lambda_{\mathrm{KLT}}n_1n_2r_{\boldsymbol{A}_0} = o(\|\boldsymbol{A}_0\|_*)$ and $\lambda_2n_1n_2r_{\boldsymbol{B}_0} = o(\|\boldsymbol{B}_0\|_*)$ since $\lambda_2 = \lambda_{\mathrm{KLT}} \asymp \theta_0^{-1/2}(n_1\wedge n_2)^{-1/2}(n_1n_2)^{-1/2}\log^{1/2}(n)$. Namely, both the first terms in U_{KLT} and U_{UNI} dominate and we compare the second terms. As $r_{\boldsymbol{A}_0} = r_{\boldsymbol{B}_0} + m$, we can claim that $U_{\mathrm{UNI}} < U_{\mathrm{KLT}}$.

For the high rank case, i.e the second term dominates or of the same order as the first term, the first terms in U_{KLT} and U_{UNI} are the smaller order. If $\mathcal{R}(\boldsymbol{\beta}_0) \nsubseteq \mathcal{R}(\boldsymbol{B}_0)$, once there exists the covariate effect, i.e $\boldsymbol{\beta}_0 \neq \boldsymbol{0}^{m\times n_2}$, as given in Proposition S2.1, $\|\boldsymbol{B}_0\|_* < \|\boldsymbol{A}_0\|_*$ which implies $U_{\mathrm{UNI}} < U_{\mathrm{KLT}}$. For the remaining cases, we obtain the result $U_{\mathrm{UNI}} \leq U_{\mathrm{KLT}}$ by $\|\boldsymbol{B}_0\|_* \leq \|\boldsymbol{A}_0\|_*$.

□

S2.2 Proofs of Theorem 4

Proof. For some constant $0 \leq \gamma \leq 1$, if $n_1 \geq n_2$, define

$$\mathcal{C}_1 = \left\{\tilde{\boldsymbol{B}} = (B_{ij}) \in \mathbb{R}^{n_1\times r_{\boldsymbol{B}_0}} : B_{ij} \in \left\{0, \gamma\left(\sigma\wedge a_1\right)\left(\frac{r_{\boldsymbol{B}_0}}{\left(n_1\wedge n_2\right)\theta_0}\right)^{1/2}\right\}, \forall 1\leq i\leq n_1, 1\leq j\leq r_{\boldsymbol{B}_0}\right\},$$

and consider the associated set of block matrices

$$\mathcal{A}\left(\mathcal{C}_1\right) = \left\{\boldsymbol{A} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \left(\tilde{\boldsymbol{B}}|\ldots|\tilde{\boldsymbol{B}}|\boldsymbol{0}\right) \in \mathbb{R}^{n_1\times n_2} : \tilde{\boldsymbol{\beta}} \in \beta\left(a_1\right), \tilde{\boldsymbol{B}} \in \mathcal{C}_1\right\},$$

where $\mathbf{0}$ denotes the $n_1 \times (n_2 - r_{\boldsymbol{B}_0}\lfloor n_2/r_{\boldsymbol{B}_0}\rfloor)$ zero matrix.

It is easy to see that any element of $\mathcal{B}(\mathcal{C}_1)$ and the difference of any two elements of $\mathcal{B}(\mathcal{C}_1)$ has rank at most $r_{\boldsymbol{B}_0}$. The entries of any matrix in $\mathcal{B}(\mathcal{C}_1)$ are within $[0, a_1]$. Due to Lemma 2.9 in Tsybakov (2009), there exists a subset $\mathcal{B}^0 \subset \mathcal{B}(\mathcal{C}_2)$ containing the zero $n_1 \times n_2$ matrix $\mathbf{0}$ where $\mathrm{Card}(\mathcal{B}^0) \geq 2^{r_{\boldsymbol{B}_0} n_1/2} + 1$ and for any two distinct elements $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ of $\mathcal{B}^0$,

$$\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F^2 \geq \frac{n_1 r_{\boldsymbol{B}_0}}{8}\left(\gamma^2 (\sigma \wedge a_1)^2 \left(\frac{r_{\boldsymbol{B}_0}}{(n_1 \wedge n_2)\,\theta_0}\right)\left\lfloor \frac{n_2}{r_{\boldsymbol{B}_0}} \right\rfloor\right) \geq \frac{\gamma^2}{16}(\sigma \wedge a_1)^2 \left(\frac{n_1 n_2 r_{\boldsymbol{B}_0}}{(n_1 \wedge n_2)\,\theta_0}\right).$$

For $0 \leq l \leq r_{\boldsymbol{B}_0}$, take $\boldsymbol{\beta}^0 \subset \beta(a_1)$ such that

$$\boldsymbol{\beta}^0 = \left\{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{m \times n_2} : (X\beta)_{ij} = \gamma(\sigma \wedge a_1)\left(\frac{l}{(n_1 \wedge n_2)\,\theta_0}\right)^{1/2}, \forall 1 \leq i \leq n_1, 1 \leq j \leq n_2\right\}.$$

For any $\boldsymbol{A} \in \mathcal{A}^0 = \boldsymbol{\beta}^0 \cup \mathcal{B}^0$, the Kullback-Leibler divergence $K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_{\boldsymbol{A}})$ between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_{\boldsymbol{A}}$ satisfies

$$K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_{\boldsymbol{A}}) = \mathsf{E}_{\mathbb{P}_{\mathbf{0}}}\left(\sum_{ij} \omega_{ij} \frac{A_{0ij}^2 - 2A_{0ij}Y_{0ij}}{2\sigma^2}\right) = \theta_0 \frac{\|\boldsymbol{A}\|_F^2}{2\sigma^2} \leq \frac{\gamma^2 (r_{\boldsymbol{B}_0} + l)\, n_1 n_2}{n_1 \wedge n_2}.$$

It is easy to know that $\mathrm{Card}(\mathcal{A}^0) = \mathrm{Card}(\mathcal{B}^0) \geq 2^{r_{\boldsymbol{B}_0} n_1/2} + 1$. From above we deduce the condition

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0) - 1} \sum_{\boldsymbol{A} \in \mathcal{A}^0} K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_{\boldsymbol{A}}) \leq \alpha \log\left(\mathrm{Card}(\mathcal{A}^0) - 1\right) \tag{S2.1}$$

is satisfied for any $\alpha > 0$ if $0 < \gamma < \sqrt{\alpha}/2$ and $l \leq r_{\boldsymbol{B}_0}$. The result now follows by application of Theorem 2.5 in Tsybakov (2009).

For $n_1 \leq n_2$, similarly, define

$$\mathcal{C}_2 = \left\{\tilde{\boldsymbol{B}} = (B_{ij}) \in \mathbb{R}^{r_{\boldsymbol{B}_0} \times n_2} : B_{ij} \in \left\{0, \gamma(\sigma \wedge a_1)\left(\frac{r_{\boldsymbol{B}_0}}{(n_1 \wedge n_2)\,\theta_0}\right)^{1/2}\right\}, \forall 1 \leq i \leq r_{\boldsymbol{B}_0}, 1 \leq j \leq n_2\right\},$$

and consider the associated set of block matrices

$$\mathcal{A}(\mathcal{C}_2) = \left\{\boldsymbol{A} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \left(\tilde{\boldsymbol{B}}|\ldots|\tilde{\boldsymbol{B}}|\mathbf{0}\right)^T \in \mathbb{R}^{n_1 \times n_2} : \tilde{\boldsymbol{\beta}} \in \beta(a_1), \tilde{\boldsymbol{B}} \in \mathcal{C}_2\right\},$$

where $\mathbf{0}$ denotes the $(n_1 - r_{\boldsymbol{B}_0}\lfloor n_1/r_{\boldsymbol{B}_0}\rfloor) \times n_2$ zero matrix here. Follow the same proof, we have the same result. □

S2.3 Non-Uniform Missing

For the non-uniform missing, we assume that the missing probability $\boldsymbol{\Theta} = (\theta_{ij})$ is known. Namely, we know $\boldsymbol{\Theta}^* = (1/\theta_{ij})$ in the risk function (3.1). Thus

$$\hat{\boldsymbol{B}}^{\text{NON-UNI}} = \underset{\boldsymbol{B}\in\mathbb{R}^{n_1\times n_2}}{\arg\min} \left\{ \frac{1}{n_1 n_2} \|\boldsymbol{B} - \boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\|_F^2 + \lambda_2 \|\boldsymbol{B}\|_* \right\}. \tag{S2.2}$$

Follow the same proof of Theorem 3 of Koltchinskii et al. (2011), we have that

Theorem S2.1. *Assume Conditions C1-C4, if* $\lambda_2 \geq 2\|\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0\|$, *then*

$$d^2\left(\hat{\boldsymbol{B}}^{\text{NON-UNI}}, \boldsymbol{B}_0\right) \leq \lambda_2 \min\left\{ 2\|\boldsymbol{B}_0\|_*, \left(\frac{1+\sqrt{2}}{2}\right)^2 \lambda_2 n_1 n_2 r_{\boldsymbol{B}_0} \right\}.$$

As for $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{NON-UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0)$, it follows from the closed form of $\hat{\boldsymbol{\beta}}$ that

$$\begin{aligned}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{NON-UNI}} - \boldsymbol{X}\boldsymbol{\beta}_0 =& \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1} n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0) \\ & - \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1} n_2\lambda_1 n_1^{-1}\boldsymbol{X}\boldsymbol{\beta}_0.\end{aligned}$$

Take $\lambda_1 = o(n_2^{-1})$, $n_2\lambda_1 = o(1)$, we have $\boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1} n_1^{-1}\boldsymbol{X}^\intercal = \boldsymbol{P}_{\boldsymbol{X}}(1+o(1))$. It implies that,

$$\frac{1}{n_1 n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{NON-UNI}} - \boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2 \leq \frac{1}{n_1 n_2}\|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0)\|_F^2 (1+o(1)) + n_2^2\lambda_1^2\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F^2 (1+o(1)).$$

It is not hard to show that $\mathsf{E}\|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0)\|_F^2 \leq \{(1/\theta_L - 1)\log(n)a_1^2 + c_\sigma^2/\theta_L\} n_2 m$. Then take $\lambda_1 = o(n_1^{-1}n_2^{-3/2}\log^{-1}(n))$, we have $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{NON-UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0) = O_p(n_1^{-1})$.

The lower bound can be given in the following theorem.

Theorem S2.2. *Assume Condition C4, fix* $a_1 > 0$, *for* $r_{\boldsymbol{B}_0}$ *such that* $1 \leq r_{\boldsymbol{B}_0} \leq \min(n_1, n_2) - m$, $(n_1 \vee n_2) r_{\boldsymbol{B}_0} \leq n_1 n_2 \theta_L$. *Let the variables* ϵ_{ij} *be Gaussian* $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$ *for* $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$. *Then there exist absolute constants* $\alpha \in (0,1)$, $c > 0$ *and* $0 \leq l \leq r_{\boldsymbol{B}_0}$, *such that*

$$\inf_{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{B}}\boldsymbol{\beta}_0 \in \beta(a_1), \boldsymbol{B}_0 \in \mathcal{B}\left(r_{\boldsymbol{B}_0}, a_1\right)} \sup P\left(d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0) > c(\sigma \wedge a_1)^2 \frac{(n_1 \vee n_2)(r_{\boldsymbol{B}_0} + l)}{n_1 n_2 \theta_L} \right) \geq \alpha.$$

S3 Justification of Condition C5(b)

Sweeting (1980) presented a very general result concerning the uniform asymptotic normality of the MLEs. In this section, we want to verify Condition C5(b) under the logistic sampling model given in (4.1) by applying Sweeting's results. A natural estimator of $\boldsymbol{\gamma}_{\cdot j}$ is the conditional MLE $\hat{\boldsymbol{\gamma}}_{\cdot j}$, denoted as that maximizes the log-likelihood,

$$\ell_{n_1}\left(\boldsymbol{\gamma}_{\cdot j}\right) = \sum_{i=1}^{n_1} \left\{\omega_{ij} \log \theta_{ij} + (1-\omega_{ij}) \log\left(1-\theta_{ij}\right)\right\}.$$

We know that the MLE $\hat{\boldsymbol{\gamma}}_{\cdot j}$ of $\boldsymbol{\gamma}_{\cdot j}$ is a consistent estimator and the asymptotic normality of $\boldsymbol{\gamma}_{\cdot j}$ for each $j = 1, \ldots, n_2$ under some regularity conditions. Then we apply Sweeting's result to show the uniform asymptotic normality of these MLEs.

The conditional Fisher information matrix is

$$I_{n_1}\left(\boldsymbol{\gamma}_{\cdot j}\right) = \mathsf{E}\left(-\frac{\partial^2 \ell_{n_1}\left(\boldsymbol{\gamma}_{\cdot j}\right)}{\partial \boldsymbol{\gamma}_{\cdot j}^2}\right) = \sum_{i=1}^{n_1} \theta_{ij}\left(1-\theta_{ij}\right) \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^{\mathsf{T}}. \tag{S3.1}$$

Let $\bar{\boldsymbol{x}}_c = \lim_{n_1 \to \infty} n_1^{-1} \sum_{i=1}^{n_1} \boldsymbol{x}_i$ and $\tilde{\boldsymbol{S}}_x = \begin{bmatrix} 1 & \bar{\boldsymbol{x}}_c^{\mathsf{T}} \\ \bar{\boldsymbol{x}}_c & \boldsymbol{S}_x \end{bmatrix}$. To guarantee the sum of squared errors in Condition C5(b), we require the following conditions for the sampling model:

CA(a). (i) There exists a universal upper bound $\theta_U \in (0,1)$, where θ_U is allowed to depend on n_1 and n_2, such that $\max_{i,j}\{\theta_{ij}\} \leq \theta_U < 1$ uniformly. (ii) $0 < \|\tilde{\boldsymbol{S}}_x\| < \infty$ and $\tilde{\boldsymbol{S}}_x > 0$.

Condition CA(a) is a mild condition. The upper bound θ_U and the lower bound θ_L in C5(a) are considered together to ensure the invertibility of $I_{n_1}(\boldsymbol{\gamma}_{\cdot j})$.

Denote the parameter space $\boldsymbol{\Xi}$ is a bounded subset of $\mathbb{R}^{m+1}$ which covers the parameters $\boldsymbol{\gamma}_{\cdot j}$ for $j = 1, \ldots, n_2$. Let $P_{\boldsymbol{\xi}}$, $P_{n_1,\boldsymbol{\xi}}$, $n_1 \geq 1$, be probability measures of random variables $\boldsymbol{A}(\boldsymbol{\xi})$, $\boldsymbol{A}_{n_1}(\boldsymbol{\xi})$, $n_1 \geq 1$ defined on the Borel subset of a metric space depending on a $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, and let $C(\mathbb{R}^{m+1})$ be the space of real bounded uniformly continuous functions, $\boldsymbol{A}_{n_1}(\boldsymbol{\xi}) \overset{u}{\Rightarrow} \boldsymbol{A}(\boldsymbol{\xi})$ in $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ if and only if

$$\sup_{\boldsymbol{\xi} \in \boldsymbol{\Xi}} \left|P_{n_1,\boldsymbol{\xi}}\left(\boldsymbol{S}\right) - P_{\boldsymbol{\xi}}\left(\boldsymbol{S}\right)\right| \to 0, \text{as } n_1 \to \infty,$$

for any Borel set $\boldsymbol{S}$ with $P_{\boldsymbol{\xi}}\left(\partial \boldsymbol{S}\right) = 0$.

In order to show the uniform weak convergence of MLEs, Sweeting proposed additional two regularity conditions in Sweeting (1980), which we present in a form that would connect well to the logistic regression model setting.

CA(b). There exist nonrandom square matrices $\boldsymbol{D}_{n_1}(\boldsymbol{\xi})$, continuous in $\boldsymbol{\xi}$, satisfying $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\|_F \to 0$, as $n_1 \to \infty$, such that

$$\boldsymbol{W}_{n_1}(\boldsymbol{\xi}) \equiv \boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\, I_{n_1}(\boldsymbol{\xi}) \left\{\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\right\}^{\intercal} \overset{u}{\Rightarrow} \boldsymbol{W}(\boldsymbol{\xi}),$$

and $\mathsf{P}(\boldsymbol{W}(\boldsymbol{\xi}) > 0) = 1$.

CA(c). For all $\epsilon > 0$, (i) $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sup_{\boldsymbol{\xi}'\in\mathcal{A}(\boldsymbol{\xi},\epsilon)} \|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\boldsymbol{D}_{n_1}(\boldsymbol{\xi}') - \boldsymbol{I}_{m+1}\|_F \to 0$, where $\mathcal{A}(\boldsymbol{\xi},\epsilon) = \{\boldsymbol{\xi}' \in \boldsymbol{\Xi} : \|\boldsymbol{D}_{n_1}^{\intercal}(\boldsymbol{\xi})(\boldsymbol{\xi}' - \boldsymbol{\xi})\|_F \le \epsilon\}$, and

(ii)$\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sup_{\boldsymbol{\xi}^k\in\mathcal{A}(\boldsymbol{\xi},\epsilon),1\le k\le(m+1)} \|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\{(I_{n_1}(\boldsymbol{\xi}^1)_{1.}^{\intercal},\ldots,I_{n_1}(\boldsymbol{\xi}^{m+1})_{(m+1).}^{\intercal}) - I_{n_1}(\boldsymbol{\xi})\}\{\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\}^{\intercal}\|_F \to 0$, where $I_{n_1}(\boldsymbol{\xi}^k)_{k.}$ is the $k-$th row of $I_{n_1}(\boldsymbol{\xi}^k)$ for $1 \le k \le m+1$.

Under growth and convergence Condition CA(b) and continuity Condition CA(c), Corollary 1 of Sweeting (1980) showed that the MLE of $\hat{\boldsymbol{\xi}}$ is asymptotic normal uniformly with respect to $\boldsymbol{\xi} \in \boldsymbol{\Xi}$,

$$\boldsymbol{W}_{n_1}^{1/2}(\boldsymbol{\xi})\, \boldsymbol{D}_{n_1}(\boldsymbol{\xi}) \left(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right) \overset{u}{\Rightarrow} \boldsymbol{Z}, \text{as } n_1 \to \infty,$$

where $\boldsymbol{Z}$ is the standard normal random vector in $\mathbb{R}^{m+1}$ and independent of $\boldsymbol{W}(\boldsymbol{\xi})$.

In the case of the logistic regression model, the parameter space $\boldsymbol{\Xi}$ is an open subset of $\mathbb{R}^{m+1}$ such that for any $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ and $\theta_{i\boldsymbol{\xi}} = \exp(\boldsymbol{x}_i^{\intercal}\boldsymbol{\xi})/\{1+\exp(\boldsymbol{x}_i^{\intercal}\boldsymbol{\xi})\}$, $0 < \theta_L \le \min_{i,j}\{\theta_{i\boldsymbol{\xi}}\} \le \max_{i,j}\{\theta_{i\boldsymbol{\xi}}\} \le \theta_U < 1$. Let $\pi_{\boldsymbol{\xi}} = n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})$, $\boldsymbol{D}_{n_1}(\boldsymbol{\xi}) = (n_1\pi_{\boldsymbol{\xi}})^{1/2}\boldsymbol{I}_{m+1}$ and $\boldsymbol{W}(\boldsymbol{\xi}) = \tilde{\boldsymbol{S}}_x$, thus $\boldsymbol{W}_{n_1}(\boldsymbol{\xi}) \equiv \boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})I_{n_1}(\boldsymbol{\xi})\{\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\}^{\intercal} = (n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}(\boldsymbol{\xi})$, where $I_{n_1}(\boldsymbol{\xi})$ is defined as the Fisher matrix in (S3.1). The justifications of Conditions CA(b) and CA(c) on any $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ are given in the following.

Justification of Condition CA(b). For any $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, since $\boldsymbol{\Xi}$ is a bounded subset of $\mathbb{R}^{m+1}$, then $\pi_{\boldsymbol{\xi}} = \sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})/n_1 \in \{\min\{\theta_U(1-\theta_U), \theta_L(1-\theta_L)\}, 1\}$. It is easy to see that $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\|_F = \sqrt{m+1}(n_1\pi_{\boldsymbol{\xi}})^{-1/2} \to 0$ under the case $\theta_L > (n_1 n_2)^{-1}(n_1 \vee n_2)\log(n)$ as $n_1 \to \infty$.

Under Condition C2, there exist a positive constant a_x such that $\|\boldsymbol{X}\|_\infty < a_x$, $\lim_{n_1\to\infty} n_1^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{X} =$

$\lim_{n_1\to\infty} n_1^{-1}\sum_{i=1}^{n_1}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T} = \boldsymbol{S}_x$. Also we have $\bar{\boldsymbol{x}}_c = \lim_{n_1\to\infty} n_1^{-1}\sum_{i=1}^{n_1}\boldsymbol{x}_i$, thus

$$\begin{aligned} n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\boldsymbol{x}_i - \pi_{\boldsymbol{\xi}}\bar{\boldsymbol{x}}_c =& n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\boldsymbol{x}_i - n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\bar{\boldsymbol{x}}_c \\ &\leq \left| n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_c\right)\right| \to 0. \end{aligned}$$

Similarly, we have $n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\boldsymbol{x}_i \to \pi_{\boldsymbol{\xi}}\bar{\boldsymbol{x}}_c$ and $n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T} \to \pi_{\boldsymbol{\xi}}\boldsymbol{S}_x$. These imply, $(n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}(\boldsymbol{\xi}) = (n_1\pi_{\boldsymbol{\xi}})^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\mathsf{T} \to \tilde{\boldsymbol{S}}_x$.

Since $\boldsymbol{D}_{n_1}(\boldsymbol{\xi}) = (n_1\pi_{\boldsymbol{\xi}})^{1/2}\boldsymbol{I}_{m+1}$, $\boldsymbol{W}(\boldsymbol{\xi}) = \tilde{\boldsymbol{S}}_x$,

$$\boldsymbol{W}_{n_1}(\boldsymbol{\xi}) \equiv (n_1\pi_{\boldsymbol{\xi}})^{-1} I_{n_1}(\boldsymbol{\xi}) \overset{u}{\Rightarrow} \tilde{\boldsymbol{S}}_x,$$

Here $\boldsymbol{W}(\boldsymbol{\xi}) = \tilde{\boldsymbol{S}}_x$ and $\mathsf{P}(\tilde{\boldsymbol{S}}_x > 0) = 1$. □

Justification of Condition CA(c). For Condition CA(c)(i), for $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, the set $\mathcal{A}(\boldsymbol{\xi},\epsilon) = \|\boldsymbol{D}_{n_1}^\mathsf{T}(\boldsymbol{\xi})(\boldsymbol{\xi}'-\boldsymbol{\xi})\|_F \leq \epsilon$ implies

$$\mathrm{tr}\left\{\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\mathsf{T}\boldsymbol{D}_{n_1}(\boldsymbol{\xi})\boldsymbol{D}_{n_1}^\mathsf{T}(\boldsymbol{\xi})\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\right\} = (n_1\pi_{\boldsymbol{\xi}})\,\mathrm{tr}\left\{\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\mathsf{T}\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\right\} \leq \epsilon^2.$$

Let $\theta_{i\boldsymbol{\xi}'} = \exp(\tilde{\boldsymbol{x}}_i^\mathsf{T}\boldsymbol{\xi}')/\{1+\exp(\tilde{\boldsymbol{x}}_i^\mathsf{T}\boldsymbol{\xi}')\}$ and $\pi_{\boldsymbol{\xi}'} = \sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}'}(1-\theta_{i\boldsymbol{\xi}'})$. Since we have $\theta_{i\boldsymbol{\xi}'} - \theta_{i\boldsymbol{\xi}} = \theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\tilde{\boldsymbol{x}}_i^\mathsf{T}(\boldsymbol{\xi}'-\boldsymbol{\xi}) + (\boldsymbol{\xi}'-\boldsymbol{\xi})^\mathsf{T}\tilde{\boldsymbol{x}}_i(1-2\theta_{i\boldsymbol{\xi}^*})\theta_{i\boldsymbol{\xi}^*}(1-\theta_{i\boldsymbol{\xi}^*})\tilde{\boldsymbol{x}}_i^\mathsf{T}(\boldsymbol{\xi}'-\boldsymbol{\xi})$ for $\boldsymbol{\xi}^* \in \mathcal{B}^{m+1}(\boldsymbol{\xi}, d(\boldsymbol{\xi},\boldsymbol{\xi}'))$, where $\mathcal{B}^{m+1}(\boldsymbol{\xi}, d(\boldsymbol{\xi},\boldsymbol{\xi}'))$ is the ball belongs to $\mathbb{R}^{m+1}$ with center at $\boldsymbol{\xi}$ and radius $d(\boldsymbol{\xi},\boldsymbol{\xi}')$, $d(\boldsymbol{\xi},\boldsymbol{\xi}')$ is euclidean distance between the vector $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$. Since $\boldsymbol{\xi}^* \in \boldsymbol{\Xi}$, we have $|(1-2\theta_{i\boldsymbol{\xi}^*})\theta_{i\boldsymbol{\xi}^*}(1-\theta_{i\boldsymbol{\xi}^*})| < 2$. Combining the fact that there exist a positive constant a_x such that $\|\boldsymbol{X}\|_\infty < a_x$, $\|\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\mathsf{T}\| < \infty$, we can say that $\theta_{i\boldsymbol{\xi}'} - \theta_{i\boldsymbol{\xi}} = \theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\tilde{\boldsymbol{x}}_i^\mathsf{T}(\boldsymbol{\xi}'-\boldsymbol{\xi}) + o((\boldsymbol{\xi}'-\boldsymbol{\xi}))$ and $(\theta_{i\boldsymbol{\xi}'}-\theta_{i\boldsymbol{\xi}})^2 = \theta_{i\boldsymbol{\xi}}^2(1-\theta_{i\boldsymbol{\xi}})^2\mathrm{tr}(\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\mathsf{T}(\boldsymbol{\xi}'-$

$\boldsymbol{\xi})(\boldsymbol{\xi}'-\boldsymbol{\xi})^\intercal)+o((\boldsymbol{\xi}'-\boldsymbol{\xi})^\intercal(\boldsymbol{\xi}'-\boldsymbol{\xi}))$. It implies that

$$
\begin{aligned}
(n_1\pi_{\boldsymbol{\xi}})^{1/2}\left|\pi_{\boldsymbol{\xi}'}-\pi_{\boldsymbol{\xi}}\right| =& (n_1\pi_{\boldsymbol{\xi}})^{1/2}\left|n_1^{-1}\sum_{i=1}^{n_1}\left\{\theta_{i\boldsymbol{\xi}'}\left(1-\theta_{i\boldsymbol{\xi}'}\right)-\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\right\}\right| \leq n_1^{-1/2}\pi_{\boldsymbol{\xi}}^{1/2}\sum_{i=1}^{n_1}3\left|\theta_{i\boldsymbol{\xi}'}-\theta_{i\boldsymbol{\xi}}\right| \\
\leq& 3n_1^{-1/2}\pi_{\boldsymbol{\xi}}^{1/2}\sqrt{\frac{n_1}{n_1\pi_{\boldsymbol{\xi}}}\sum_{i=1}^{n_1}\left\{n_1\pi_{\boldsymbol{\xi}}\left(\theta_{i\boldsymbol{\xi}'}-\theta_{i\boldsymbol{\xi}}\right)^2\right\}} \\
\leq& 3n_1^{-1/2}\sqrt{\sum_{i=1}^{n_1}n_1\pi_{\boldsymbol{\xi}}\mathrm{tr}\left\{\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\intercal\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\intercal+o\left(\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\intercal\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\right)\right\}} \\
\leq& 3n_1^{-1/2}\sqrt{2n_1\mathrm{tr}\left\{\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\intercal\right)n_1\pi_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\intercal\right\}} \leq 3\sqrt{2\left\|\tilde{\boldsymbol{S}}_x\right\|}\epsilon,
\end{aligned}
$$

which implies $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sup_{\boldsymbol{\xi}'\in\mathcal{A}(\boldsymbol{\xi},\epsilon)}\|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\boldsymbol{D}_{n_1}(\boldsymbol{\xi}')-\boldsymbol{I}_{m+1}\|_F=(m+1)|\pi_{\boldsymbol{\xi}'}/\pi_{\boldsymbol{\xi}}-1|=(m+1)(n_1\pi_{\boldsymbol{\xi}})^{-1/2}(n_1\pi_{\boldsymbol{\xi}})^{1/2}|\pi_{\boldsymbol{\xi}'}-\pi_{\boldsymbol{\xi}}|\to 0$ as $n_1\to\infty$.

For Condition CA(c)(ii), Let $\theta_{i\boldsymbol{\xi}^k}=\exp(\tilde{\boldsymbol{x}}_i^\intercal\boldsymbol{\xi}^k)/\{1+\exp(\tilde{\boldsymbol{x}}_i^\intercal\boldsymbol{\xi}^k)$ and $\pi_{\boldsymbol{\xi}^k}=\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}^k}(1-\theta_{i\boldsymbol{\xi}^k})/n_1$. For any $\boldsymbol{\xi}\in\boldsymbol{\Xi}$, since $(n_1\pi_{\boldsymbol{\xi}^k})^{-1}I_{n_1}(\boldsymbol{\xi}^k)\to\tilde{\boldsymbol{S}}_x$ and $(n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}(\boldsymbol{\xi})\to\tilde{\boldsymbol{S}}_x$ showed as before, we have over the sets $\|\boldsymbol{D}_{n_1}^\intercal(\boldsymbol{\xi})(\boldsymbol{\xi}^k-\boldsymbol{\xi})\|_F\leq\epsilon$, for $1\leq k\leq m+1$, as $n_1\to\infty$,

$$
\begin{aligned}
\left\|(n_1\pi_{\boldsymbol{\xi}})^{-1}\left\{I_{n_1}\left(\boldsymbol{\xi}^k\right)-I_{n_1}\left(\boldsymbol{\xi}\right)\right\}\right\|_F \leq& \left\|\left\{(n_1\pi_{\boldsymbol{\xi}})^{-1}-\left(n_1\pi_{\boldsymbol{\xi}^k}\right)^{-1}\right\}I_{n_1}\left(\boldsymbol{\xi}^k\right)\right\|_F \\
&+\left\|\left(n_1\pi_{\boldsymbol{\xi}^k}\right)^{-1}I_{n_1}\left(\boldsymbol{\xi}^k\right)-(n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}\left(\boldsymbol{\xi}\right)\right\|_F \\
\leq (n_1\pi_{\boldsymbol{\xi}})^{-3/2}(n_1\pi_{\boldsymbol{\xi}})^{1/2}|\pi_{\boldsymbol{\xi}^k}-\pi_{\boldsymbol{\xi}}|&\left\|\left(n_1\pi_{\boldsymbol{\xi}^k}\right)^{-1}I_{n_1}\left(\boldsymbol{\xi}^k\right)\right\|_F+\left\|\left(n_1\pi_{\boldsymbol{\xi}^k}\right)^{-1}I_{n_1}\left(\boldsymbol{\xi}^k\right)-(n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}\left(\boldsymbol{\xi}\right)\right\|_F
\end{aligned}
$$

By the inequalities $\|\tilde{\boldsymbol{S}}_x\|_F\leq\sqrt{m+1}\|\tilde{\boldsymbol{S}}_x\|<\infty$ and $\sqrt{n_1\pi_{\boldsymbol{\xi}}}|\pi_{\boldsymbol{\xi}^k}-\pi_{\boldsymbol{\xi}}|\leq 3\sqrt{2\|\tilde{\boldsymbol{S}}_x\|}\epsilon$, we have $\|(n_1\pi_{\boldsymbol{\xi}})^{-1}\{I_{n_1}(\boldsymbol{\xi}^k)-I_{n_1}(\boldsymbol{\xi})\}\|_F\to 0$.

Thus we have

$$
\begin{aligned}
&\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sup_{\boldsymbol{\xi}^k\in\mathcal{A}(\boldsymbol{\xi},\epsilon)}\left\|\boldsymbol{D}_{n_1}^{-1}\left(\boldsymbol{\xi}\right)\left\{\left(I_{n_1}\left(\boldsymbol{\xi}^1\right)_{1.}^\intercal,\ldots,I_{n_1}\left(\boldsymbol{\xi}^{m+1}\right)_{(m+1).}^\intercal\right)-I_{n_1}\left(\boldsymbol{\xi}\right)\right\}\left\{\boldsymbol{D}_{n_1}^{-1}\left(\boldsymbol{\xi}\right)\right\}^\intercal\right\|_F \\
\leq&\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sum_{k=1}^{m+1}\sup_{\boldsymbol{\xi}^k\in\mathcal{A}(\boldsymbol{\xi},\epsilon)}\left\|(n_1\pi_{\boldsymbol{\xi}})^{-1}\left\{I_{n_1}\left(\boldsymbol{\xi}^k\right)-I_{n_1}\left(\boldsymbol{\xi}\right)\right\}\right\|_F\to 0.
\end{aligned}
$$

□

Applying Corollary 1 in Sweeting (1980) we have that $I^{1/2}(\boldsymbol{\xi})(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi})\overset{u}{\Rightarrow}\boldsymbol{Z}$ for all $\boldsymbol{\gamma}\in\boldsymbol{\Xi}$. Under Condition CA(a), we have the parameters $\boldsymbol{\gamma}_{\cdot j}\in\boldsymbol{\Xi}$ for $j=1,\ldots,n_2$. Namely, $I^{1/2}(\boldsymbol{\gamma}_{\cdot j})(\hat{\boldsymbol{\gamma}}_{\cdot j}-\boldsymbol{\gamma}_{\cdot j})\overset{u}{\Rightarrow}$

$\boldsymbol{Z}$ which implies $I^{1/2}(\boldsymbol{\gamma}_{\cdot j})(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) \xrightarrow{d} \mathcal{N}(0,1)$ for all $j = 1, \ldots, n_2$. For $j = 1, \ldots, n_2$, define $\pi_j = \sum_{i=1}^{n_1} \theta_{ij}(1-\theta_{ij})/n_1$, and $\pi_j^* \in ((1-\theta_U)^2/\theta_L^2, (1-\theta_L)^2/\theta_U^2)$, then we have $\pi_j \in \{\min\{\theta_U(1-\theta_U), \theta_L(1-\theta_L)\}, 1\}$ and $\pi_j^* = \sum_{i=1}^{n_1}\{(1-\theta_{ij})^2/\theta_{ij}^2\}/n_1$. As shown in Justification of Condition CA(b), $(n_1\pi_j)^{-1/2} I^{1/2}(\boldsymbol{\gamma}_{\cdot j}) \to \tilde{\boldsymbol{S}}_x^{\,1/2}$. Thus we have $\sqrt{n_1\pi_j}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) \overset{u}{\Rightarrow} \boldsymbol{Z}_j$, where $\boldsymbol{Z}_j \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{S}}_x^{\,-1})$ for all $1 \le j \le n_2$. For each $j = 1, \ldots, n_2$, $|\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}| = O_p(\boldsymbol{1}/\sqrt{\boldsymbol{n_1\pi_j}})$.

The estimator of θ_{ij} is given by $\hat{\theta}_{ij} = \exp(\tilde{\boldsymbol{x}}_i^{\intercal}\hat{\boldsymbol{\gamma}}_{\cdot j})/\{1+\exp(\tilde{\boldsymbol{x}}_i^{\intercal}\hat{\boldsymbol{\gamma}}_{\cdot j})\}$, thus, for each $j = 1, \ldots, n_2$, $\sup_i |\hat{\theta}_{ij} - \theta_{ij}| = O_p(1/\sqrt{n_1\pi_j})$. Also, we have that for specific $\boldsymbol{\gamma}_{\cdot j}^* \in \mathcal{B}^{m+1}(\boldsymbol{\gamma}_{\cdot j}, d(\boldsymbol{\gamma}_{\cdot j}, \hat{\boldsymbol{\gamma}}_{\cdot j}))$,

$$\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}} = -\frac{1}{\theta_{ij}^2}\left(\frac{\partial\theta_{ij}}{\partial\boldsymbol{\gamma}_{\cdot j}}\right)^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) + (\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})^{\intercal}\frac{\partial^2(1/\theta_{ij})}{\partial\boldsymbol{\gamma}_{\cdot j}^2}\Big|_{\boldsymbol{\gamma}_{\cdot j}^*}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}),$$

which can simplify to be

$$\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}} = -\frac{(1-\theta_{ij})}{\theta_{ij}}\tilde{\boldsymbol{x}}_i^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) + (\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})^{\intercal}\frac{\left(1-\theta_{ij}^*\right)}{\theta_{ij}^*}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}).$$

Since there exist a positive constant a_x such that $\|\boldsymbol{X}\|_\infty < a_x$, we have $\|\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\intercal}\|_\infty < \infty$. Also $\boldsymbol{\gamma}_{\cdot j}^* \in B(\boldsymbol{\gamma}_{\cdot j}, d(\boldsymbol{\gamma}_{\cdot j}, \hat{\boldsymbol{\gamma}}_{\cdot j}))$ and $\|\boldsymbol{X}\|_\infty < a_x$ implies $\theta_{ij}^* \to \theta_{ij}$, as $n_1 \to \infty$. Namely, $(1-\theta_{ij}^*)/\theta_{ij}^* \to (1-\theta_{ij})/\theta_{ij}$, as $n_1 \to \infty$. Once $\theta_{ij} \neq 0$ and $\tilde{\boldsymbol{x}}_i \neq \boldsymbol{0}$, by Taylor expansion and continuous mapping theorem, we can see that:

$$\left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2 = \frac{(1-\theta_{ij})^2}{\theta_{ij}^2}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})^{\intercal}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) + o\left((\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})\right),$$

for $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$. As $n_1 \to \infty$, we have

$$\sum_{i=1}^{n_1}\frac{(1-\theta_{ij})^2}{\theta_{ij}^2}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\intercal}/\left(n_1\pi_j^*\right) \to \tilde{\boldsymbol{S}}_x.$$

By Slutsky theorem,

$$\frac{\pi_j}{\pi_j^*}\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2 \overset{u}{\Rightarrow} \boldsymbol{Z}_j^{\intercal}\left(\tilde{\boldsymbol{S}}_x^{\,-1}\right)^{-1}\boldsymbol{Z}_j^{\intercal},$$

which implies that $\pi_j\pi_j^{*-1}\sum_{i=1}^{n_1}(1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \overset{u}{\Rightarrow} \boldsymbol{U}_j$, where $\boldsymbol{U}_j \sim \chi_{m+1}^2$ for all $j = 1, \ldots, n_2$.

By using Polya's theorem, we have for any $t > \eta_g^{-1}(m+1)$, let $\eta_g = \min\{\pi_j/\pi_j^*\}$, $k_{n_1} = \max_j \sup_t |\mathsf{P}(\sum_i(1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \ge t) - \mathsf{P}(\chi_{m+1}^2 \ge \eta_g t)| \le 1/n_2^2$ there exists a positive integer N_{1/n_2^2},

for $n_1 \geq N_{1/n_2^2}$,

$$\sup_j \mathsf{P}\left\{\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \geq t\right\} \leq \sup_j \mathsf{P}\left\{\frac{\pi_j}{\pi_j^*}\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \geq \eta_g t\right\}$$

$$\leq \sup_j \left\{\mathsf{P}\left(\chi^2_{m+1} \geq \eta_g t\right)\right\} + k_{n_1} \leq \left[\frac{\eta_g t}{m+1}\exp\left\{1-\frac{\eta_g t}{m+1}\right\}\right]^{\frac{m+1}{2}} + k_{n_1}.$$

Take $c_{n_1,n_2} = n_2\log(n_2)/\eta_g$ and $t_0 = (m+3)$, for $t > t_0$, we have

$$\mathsf{P}\left\{\sum_{j=1}^{n_2}\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \geq c_{n_1,n_2}t\right\} \leq \sum_{j=1}^{n_2}\mathsf{P}\left\{\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \geq \frac{c_{n_1,n_2}t}{n_2}\right\}$$

$$= \sum_{j=1}^{n_2}\mathsf{P}\left\{\chi^2_{m+1} \geq \frac{\eta_g c_{n_1,n_2}t}{n_2}\right\} \leq \sum_{j=1}^{n_2}\left[\frac{\eta_g c_{n_1,n_2}t}{n_2(m+1)}\exp\left\{1-\frac{\eta_g c_{n_1,n_2}t}{n_2(m+1)}\right\}\right]^{\frac{m+1}{2}} + n_2 k_{n_1}$$

$$\leq (m+1)^{-(m+1)/2}\exp\left\{\frac{m+1}{2}-\frac{\eta_g c_{n_1,n_2}t}{2n_2}+\log(t)+\frac{m+1}{2}\log\left(\frac{\eta_g c_{n_1,n_2}}{n_2}\right)+\log(n_2)\right\} + n_2 k_{n_1}$$

$$\leq (m+1)^{-(m+1)/2}\exp\left\{\frac{m+1}{2}-\frac{t\log(n_2)}{2}+\log(t)+\frac{m+3}{2}\log(n_2)\right\} + n_2 k_{n_1}$$

$$\leq (m+1)^{-(m+1)/2}\exp\left\{m+2-\frac{t}{2}+\log(t)\right\} + n_2 k_{n_1}.$$

Let $g(t) = (m+1)^{-(m+1)/2}\exp\{m+2-t/2+\log(t)\}$ and $h_{n_1,n_2} = n_2 k_{n_1} \leq 1/n_2$ in Condition C5(b), we have $\lim_{t\to\infty} g(t) = 0$ and $\lim_{n_1,n_2\to\infty} h_{n_1,n_2} = 0$. It satisfies the requirements.

S4 Lemmas and Proofs

In this section, we provide various results required in the proofs of Theorems 1 and 2, as well as Corollaries 1 and 2. First, we review some basic facts about matrices which will be useful in the following development. For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n_1\times n_2}$, we have

- Trace Duality Property:

$$|\mathrm{tr}(\boldsymbol{A}^\intercal\boldsymbol{B})| \leq \|\boldsymbol{B}\|\,\|\boldsymbol{A}\|_*. \tag{S4.1}$$

- Norm Inequalities:

$$\|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\|_* \leq \sqrt{r_{\boldsymbol{A}}}\,\|\boldsymbol{A}\|_F \quad \text{and} \quad \|\boldsymbol{A}\| \leq \|\boldsymbol{A}\|_F \leq \sqrt{r_{\boldsymbol{A}}}\,\|\boldsymbol{A}\|, \tag{S4.2}$$

where $r_{\boldsymbol{A}}$ is the rank of matrix $\boldsymbol{A}$.

Write $\boldsymbol{J}_{ij} = \boldsymbol{e}_i(n_1)\boldsymbol{e}_j^\intercal(n_2)$, where $\boldsymbol{e}_i(n) \in \mathbb{R}^n$ is the standard basis vector with the i-th element being 1 and the rest being 0. Now we present several lemmas.

Lemma S4.1. *Let $\Psi^{(1)} = \sum_{ij} \omega_{ij}\epsilon_{ij}\boldsymbol{J}_{ij}/(n_1 n_2 \hat{\theta}_{ij})$. Under Conditions C1, C4 and C5, for some positive constants c_σ, η, δ_σ and all $t > t_0$, there exists $\Delta^{(1)}(\delta_\sigma, t)$ such that*

$$\left\|\Psi^{(1)}\right\| \leq \Delta^{(1)}(\delta_\sigma, t) \asymp \max\left\{\frac{\sqrt{(n_1 \vee n_2)\log(n)}}{\sqrt{\theta_L} n_1 n_2}, (n_1 n_2)^{-3/4}\left(c_{n_1,n_2} t\right)^{1/2}\log^{\delta_\sigma/4}(n)\right\}$$

holds with probability at least $1 - 1/n - g(t) - h_{n_1,n_2} - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$.

More specifically, for the uniform missingness, we have $\theta_{ij} \equiv \theta_0$ and $\hat{\theta}_{ij} \equiv N/(n_1 n_2)$ and for some positive constants δ_σ and C_1 such that

$$\left\|\Psi^{(1)}\right\| \leq C_1 \frac{\sqrt{(n_1 \vee n_2)\log(n)}}{\sqrt{\theta_0} n_1 n_2}$$

holds with probability at least $1 - 1/n - \log^{-\delta_\sigma}(n) - 2/(n_1 \vee n_2)$.

To prove Lemma S4.1, we apply Theorem 6.2 which is matrix Bernstein inequality for the sub-exponential case provided by Tropp (2012).

Proof of Lemma S4.1. For any rectangular matrix $\boldsymbol{M}$, let $\mathcal{L}(\boldsymbol{M})$ be the self-adjoint dilation of $\boldsymbol{M}$ defined as

$$\mathcal{L}(\boldsymbol{M}) := \begin{bmatrix} 0 & \boldsymbol{M} \\ \boldsymbol{M}^\intercal & 0 \end{bmatrix}.$$

In our case, for $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$, let

$$\boldsymbol{G}_{n_2(i-1)+j} = \mathcal{L}\left(\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right) \quad \text{and} \quad \boldsymbol{H}_{n_2(i-1)+j} = \mathcal{L}\left(\frac{c_\sigma}{\sqrt{\theta_L}}\boldsymbol{J}_{ij}\right).$$

To apply Theorem 6.2 of Tropp (2012), we verify the conditions needed in the following.

Since ϵ_{ij} is independent of ω_{ij}, we have

$$\mathsf{E}\left(\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right) = \mathsf{E}(\epsilon_{ij})\,\mathsf{E}\left(\frac{\omega_{ij}\boldsymbol{J}_{ij}}{\theta_{ij}}\right) = \boldsymbol{0},$$

which implies $\mathsf{E}\left(\boldsymbol{G}_{n_2(i-1)+j}\right) = \boldsymbol{0}$. Write $\eta_H = \eta/\theta_L$, where η is the constant in Condition C1. Now we want to show that

$$\mathsf{E}\left\{\mathcal{L}\left(\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right)^l\right\} \leq \frac{l!}{2}\cdot\eta_H^{l-2}\mathcal{L}\left(\frac{c_\sigma}{\sqrt{\theta_L}}\boldsymbol{J}_{ij}\right)^2 \quad \text{for } l = 2, 3, 4, \ldots. \tag{S4.3}$$

In our case, under Condition C1 and C4, for a finite constant η, we have

$$\mathsf{E}\left|\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\right|^l = \frac{\mathsf{E}\left|\epsilon_{ij}\right|^l \mathsf{E}\omega_{ij}}{\theta_{ij}^l} \leq \frac{\max \mathsf{E}\left|\epsilon_{ij}\right|^l}{\theta_L^{l-1}} \leq \frac{1}{2} l! \left(\frac{c_\sigma}{\sqrt{\theta_L}}\right)^2 \left(\frac{\eta}{\theta_L}\right)^{l-2}, \; l = 2, 3, \ldots.$$

Thus it suffices to show that $\mathcal{L}^l(\boldsymbol{J}_{ij}) \leq \mathcal{L}^2(\boldsymbol{J}_{ij})$ for $l = 2, 3, 4, \ldots$.

Let $\boldsymbol{K}_{n,i} = \boldsymbol{e}_i(n)\boldsymbol{e}_i^\intercal(n)$, where $\boldsymbol{e}_i(n) \in \mathbb{R}^n$ is the standard basis vector of $\mathbb{R}^n$ with the i-th element being 1 and the rest being 0. By the properties of $\boldsymbol{J}_{ij}$, it is not hard to show that for $l = 2s$ or $2s+1$, we have

$$\mathcal{L}^{2s}\left(\boldsymbol{J}_{ij}\right) = \begin{bmatrix} \boldsymbol{K}_{n_1,i} & 0 \\ 0 & \boldsymbol{K}_{n_2,j} \end{bmatrix} \quad \text{and} \quad \mathcal{L}^{2s+1}\left(\boldsymbol{J}_{ij}\right) = \begin{bmatrix} 0 & \boldsymbol{J}_{ij} \\ \boldsymbol{J}_{ij}^\intercal & 0 \end{bmatrix} = \mathcal{L}\left(\boldsymbol{J}_{ij}\right).$$

Hence (S4.3) is verified as $\begin{bmatrix} \boldsymbol{K}_{n_1,i} & -\boldsymbol{J}_{ij} \\ -\boldsymbol{J}_{ij}^\intercal & \boldsymbol{K}_{n_2,j} \end{bmatrix} \geq 0$.

Set the constant $\sigma_H^2 = \|\sum_{ij} \mathcal{L}(c_\sigma \boldsymbol{J}_{ij}/\sqrt{\theta_L})^2\| = c_\sigma^2 \|\sum_{ij} \mathcal{L}(\boldsymbol{J}_{ij})^2\|/\theta_L$. Since

$$\left\|\sum_{ij} \mathcal{L}\left(\boldsymbol{J}_{ij}\right)^2\right\| = \left\|\begin{bmatrix} \sum_{ij} \boldsymbol{K}_{n_1,i} & 0 \\ 0 & \sum_{ij} \boldsymbol{K}_{n_2,j} \end{bmatrix}\right\| = \max\left\{\left\|\sum_{ij} \boldsymbol{K}_{n_1,i}\right\|, \left\|\sum_{ij} \boldsymbol{K}_{n_2,j}\right\|\right\}$$

$$= \max\left\{\left\|n_2 \boldsymbol{I}_{n_1}\right\|, \left\|n_1 \boldsymbol{I}_{n_2}\right\|\right\} = n_1 \vee n_2,$$

we have $\sigma_H^2 = c_\sigma^2(n_1 \vee n_2)/\theta_L$. By the property of dilation (2.12) of Tropp (2012),

$$\mathsf{P}\left[\lambda_{\max}\left\{\sum_{ij} \mathcal{L}\left(\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right)\right\} \geq t\right] = \mathsf{P}\left(\left\|\sum_{ij} \frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| \geq t\right).$$

By the Matrix Bernstein Inequality in Theorem 6.2 of Tropp (2012), we show that, for all $t_1 > 0$,

$$\mathsf{P}\left(\left\|\sum_{ij} \frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| \geq t_1\right) \leq n \cdot \exp\left\{\frac{-t_1^2/2}{c_\sigma^2\left(n_1 \vee n_2\right)/\theta_L + \eta_H t_1}\right\}$$

$$\leq \begin{cases} n \cdot \exp\left\{\frac{-t_1^2}{4c_\sigma^2(n_1 \vee n_2)/\theta_L}\right\} & \text{for } t_1 \leq c_\sigma^2\left(n_1 \vee n_2\right)/\left(\theta_L \eta_H\right) \\ n \cdot \exp\left\{\frac{-t_1}{4\eta_H}\right\} & \text{for } t_1 \geq c_\sigma^2\left(n_1 \vee n_2\right)/\left(\theta_L \eta_H\right) \end{cases}$$

In other words, for any $s_1 > 0$, with probability at least $1 - \exp\{-s_1\}$, we have

$$\left\|\sum_{ij} \frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| \leq \max\left\{2c_\sigma\sqrt{\frac{(n_1 \vee n_2)\left\{s_1 + \log(n)\right\}}{\theta_L}}, 4\eta_H\left\{s_1 + \log(n)\right\}\right\}$$

where $s_1' = s_1 + \log(n)$. Choose $s_1 = \log(n)$, i.e, $s_1' = 2\log(n)$. With probability at least $1 - 1/n$, we have

$$\frac{1}{n_1 n_2}\left\|\sum_{ij}\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| \leq \frac{2c_\sigma\sqrt{2\left(n_1 \vee n_2\right)\log\left(n\right)}}{\sqrt{\theta_L}n_1 n_2} := \Delta^{(1)'}.$$

We also know that

$$\begin{aligned}\left\|\sum_{ij}\epsilon_{ij}\omega_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)\boldsymbol{J}_{ij}\right\|^2 \leq& \left\|\sum_{ij}\epsilon_{ij}\omega_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)\boldsymbol{J}_{ij}\right\|_F^2 = \sum_{ij}\epsilon_{ij}^2\omega_{ij}^2\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \\ \leq& \sum_{ij}\epsilon_{ij}^2\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \leq \max\epsilon_{ij}^2\sum_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2.\end{aligned}$$

Due to Markov inequality, under Condition C1, we have for any $a > 0$,

$$\mathsf{P}\left(\max\epsilon_{ij}^2 \geq a\right) = \mathsf{P}\left(\max\epsilon_{ij}^4 \geq a^2\right) \leq \frac{\sum_{ij}\mathsf{E}\epsilon_{ij}^4}{a^2} \leq \frac{12n_1n_2c_\sigma^2\eta^2}{a^2}.$$

Take $a = (n_1n_2)^{1/2}\log^{\delta_\sigma/2}(n)$ for a positive constant δ_σ, we have $\max\epsilon_{ij}^2 \leq (n_1n_2)^{1/2}\log^{\delta_\sigma/2}(n)$ with probability at least $1 - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$.

Combining with Condition C5(b), we have for $t > t_0$, with probability at least $1-g(t)-h_{n_1,n_2}-12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$, $\|\sum_{ij}\epsilon_{ij}\omega_{ij}(1/\hat{\theta}_{ij} - 1/\theta_{ij})\boldsymbol{J}_{ij}\| \leq (n_1n_2)^{1/4}(c_{n_1,n_2}t)^{1/2}\log^{\delta_\sigma/4}(n)$.

Then for $t > t_0$, with probability at least $1 - 1/n - g(t) - h_{n_1,n_2} - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$, we have

$$\begin{aligned}&\frac{1}{n_1n_2}\left\|\sum_{ij}\frac{\epsilon_{ij}\omega_{ij}}{\hat{\theta}_{ij}}\boldsymbol{J}_{ij}\right\| \leq \frac{1}{n_1n_2}\left\|\sum_{ij}\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| + \frac{1}{n_1n_2}\left\|\sum_{ij}\epsilon_{ij}\omega_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)\boldsymbol{J}_{ij}\right\| \\ \leq& \Delta^{(1)'} + \left(n_1n_2\right)^{-3/4}\left(c_{n_1,n_2}t\right)^{1/2}\log^{\delta_\sigma/4}\left(n\right) \\ :=& \Delta^{(1)}\left(\delta_\sigma, t\right) \asymp \max\left\{\frac{\sqrt{\left(n_1 \vee n_2\right)\log\left(n\right)}}{\sqrt{\theta_L}n_1n_2}, \left(n_1n_2\right)^{-3/4}\left(c_{n_1,n_2}t\right)^{1/2}\log^{\delta_\sigma/4}\left(n\right)\right\}.\end{aligned}$$

As for the uniform missingness, for the first term without the estimators $\hat{\theta}_{ij}$, we have the same upper bound. We also know that for the second term,

$$\begin{aligned}\mathsf{E}\left\|\sum_{ij}\epsilon_{ij}\omega_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)\boldsymbol{J}_{ij}\right\|^2 \leq& \mathsf{E}\left\{\sum_{ij}\epsilon_{ij}^2\omega_{ij}^2\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2\right\} \leq c_\sigma^2\mathsf{E}\left\{\sum_{ij}\omega_{ij}\left(\frac{n_1n_2}{N}-\frac{1}{\theta_0}\right)^2\right\} \\ =& c_\sigma^2\mathsf{E}\left\{N\left(\frac{n_1n_2}{N}-\frac{1}{\theta_0}\right)^2\right\} = c_\sigma^2\left(n_1n_2\right)^2\mathsf{E}\left\{\frac{1}{N}-\frac{1}{n_1n_2\theta_0}\right\}.\end{aligned}$$

Also $\mathsf{E}(N) = n_1 n_2 \theta_0$ and Taylor expansions for the moments of functions of random variables implies that $\mathsf{E}(1/N) = 1/(\theta_0 n_1 n_2) + 1/(\theta_0 n_1 n_2)^3 \mathsf{Var}(N)(1+o(1)) = 1/(\theta_0 n_1 n_2) + (1-\theta_0)/(\theta_0 n_1 n_2)^2 (1+o(1))$ due to the fact that $\mathsf{E}(N - (n_1 n_2)\theta_0)^4 = o(\mathsf{Var}(N))$. We have $\mathsf{E}\|\epsilon_{ij}\omega_{ij}(n_1 n_2/N - 1/\theta_0)\boldsymbol{J}_{ij}\| \leq 2c_\sigma^2(1-\theta_0)/\theta_0^2$.

Due to Markov inequality, we have for $0 < \delta_\sigma < 2$, $\|\epsilon_{ij}\omega_{ij}(n_1 n_2/N - 1/\theta_0)\boldsymbol{J}_{ij}\| \leq c_\sigma^2(1-\theta_0)\log^{\delta_\sigma}(n)/\theta_0^2 \leq c_\sigma^2 \log^{\delta_\sigma}(n)/\theta_0^2$ with probability at least $1 - 2\log^{-\delta_\sigma}(n)$. Since $n_1 n_2 \theta_0 > (n_1 \vee n_2)\log(n)$, we have $\log^{\delta_\sigma}(n)/\theta_0^2 < (n_1 \vee n_2)\log(n)/\theta_0$. Then we have under the uniform missingness, for a positive constant C_1,

$$\left\|\Psi^{(1)}\right\| \leq C_1 \frac{\sqrt{(n_1 \vee n_2)\log(n)}}{\sqrt{\theta_0} n_1 n_2}$$

holds with probability at least $1 - 1/n - 2\log^{-\delta_\sigma}(n)$.

□

Lemma S4.2. *Let* $\Psi^{(2)} = \sum_{ij} A_{0ij}(\omega_{ij}/\theta_{ij} - 1)\boldsymbol{J}_{ij}/(n_1 n_2)$. *Under Conditions C3-C5, there exists* $\Delta^{(2)}$ *such that*

$$\left\|\Psi^{(2)}\right\| \leq \Delta^{(2)} \asymp \frac{\sqrt{|1/\theta_L - 1|\,(n_1 \vee n_2)\log(n)}}{n_1 n_2}$$

holds with probability at least $1 - 1/n$.

To prove Lemma S4.2, we utilize Proposition 1 given by Koltchinskii et al. (2011) as an immediate consequence of the Matrix Bernstein Inequality due to Ahlswede and Winter (2002) and Tropp (2012). For matrix $\boldsymbol{A}_0$, define that:

$$|\boldsymbol{A}_0|^* := \max\left\{ \sqrt{\frac{\max\limits_{1\leq i\leq n_1} \sum_{j=1}^{n_2} |1/\theta_{ij} - 1| A_{0,ij}^2}{n_1 n_2}}, \sqrt{\frac{\max\limits_{1\leq j\leq n_2} \sum_{i=1}^{n_1} |1/\theta_{ij} - 1| A_{0,ij}^2}{n_1 n_2}} \right\}. \tag{S4.4}$$

Proof of Lemma S4.2. Let $\boldsymbol{M}_{n_2(i-1)+j} = A_{0ij}(\omega_{ij}/\theta_{ij} - 1)\boldsymbol{J}_{ij}$. Under Conditions C4 and C5, it is easy to show that $\max\limits_k \|\boldsymbol{M}_k\| \leq \max\{1/\theta_{ij} - 1, 1\}\|\boldsymbol{A}_0\|_\infty \leq \max\{1/\theta_L - 1, 1\}\|\boldsymbol{A}_0\|_\infty$ and

$$\sigma_M = \max\left\{ \frac{1}{n_1 n_2}\left\|\sum_k \mathsf{E}(\boldsymbol{M}_k \boldsymbol{M}_k^\intercal)\right\|^{1/2}, \frac{1}{n_1 n_2}\left\|\sum_k \mathsf{E}(\boldsymbol{M}_k^\intercal \boldsymbol{M}_k)\right\|^{1/2} \right\} \leq |\boldsymbol{A}_0|^*.$$

Take $U_M = \max\{1/\theta_L - 1, 1\}\|\boldsymbol{A}_0\|_\infty$. By Proposition 1 of Koltchinskii et al. (2011), we have, for all $t > 0$,

$$\left\|\Psi^{(2)}\right\| \le 2\max\left\{|\boldsymbol{A}_0|^* \sqrt{\frac{t+\log(n)}{n_1 n_2}}, \max\left\{\frac{1}{\theta_L} - 1, 1\right\}\|\boldsymbol{A}_0\|_\infty \frac{t+\log(n)}{n_1 n_2}\right\}$$

with probability at least $1 - \exp\{-t\}$.

According to (S4.4), under Conditions C3 and C5, we have

$$|\boldsymbol{A}_0|^* \le \sqrt{\frac{|1/\theta_L - 1|}{n_1 n_2}} \max\left\{\|\boldsymbol{A}_0\|_{\infty,2}, \|\boldsymbol{A}_0^\intercal\|_{\infty,2}\right\} \le a_2 \sqrt{\frac{|1/\theta_L - 1|}{n_1 \wedge n_2}}.$$

Under additional Condition C3 and $t = \log(n)$, with probability at least $1 - 1/n$,

$$\left\|\Psi^{(2)}\right\| \le 2(a_1 \vee a_2)\max\left\{\sqrt{\frac{2|1/\theta_L - 1|\log(n)}{(n_1 \wedge n_2) n_1 n_2}}, 2\max\left\{\frac{1}{\theta_L} - 1, 1\right\}\frac{\log^{3/2}(n)}{n_1 n_2}\right\} := \Delta^{(2)},$$

for some positive constants a_1 and a_2 defined in Condition C3.

Since $(n_1 n_2)^{-1}\log^{3/2}(n) = o\{(n_1 \vee n_2)^{1/2}(n_1 n_2)^{-1}\log^{1/2}(n)\}$ and $\sqrt{|1/\theta_L - 1|} = o(\max\{1/\theta_L - 1, 1\})$ when $\theta_L = o(1)$, we have $\Delta^{(2)} \asymp \sqrt{|1/\theta_L - 1|}(n_1 \vee n_2)^{1/2}(n_1 n_2)^{-1}\log^{1/2}(n)$. □

Lemma S4.3. *Let $\Psi^{(3)} = \sum_{ij} A_{0ij}(\omega_{ij}/\hat{\theta}_{ij} - \omega_{ij}/\theta_{ij})\boldsymbol{J}_{ij}/(n_1 n_2)$. Under Conditions C3 and C5, for all $t > t_0$, there exists $\Delta^{(3)}(t)$ such that*

$$\left\|\Psi^{(3)}\right\| \le \Delta^{(3)}(t) \asymp \frac{\sqrt{c_{n_1,n_2} t\log(n)}}{n_1 n_2}$$

holds with probability at least $1 - g(t) - h_{n_1,n_2}$.

More specifically, for the uniform missingness, we have $\theta_{ij} \equiv \theta_0$ and $\hat{\theta}_{ij} \equiv N/(n_1 n_2)$ and for $0 < \delta_\sigma < 2$, such that

$$\left\|\Psi^{(3)}\right\| \le \frac{\sqrt{2(n_1 \vee n_2)\log(n)}a_1}{\sqrt{\theta_0} n_1 n_2}$$

holds with probability at least $1 - 2\log^{-\delta_\sigma}(n)$.

Proof of Lemma S4.3. By the inequality (S4.2), we have

$$\left\|\Psi^{(3)}\right\| \le \frac{1}{n_1 n_2}\left\|\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{A}_0 - \boldsymbol{W} \circ \boldsymbol{\Theta}^* \circ \boldsymbol{A}_0\right\|_F$$

$$= \frac{1}{n_1 n_2}\sqrt{\sum_{ij} A_{0ij}^2 \omega_{ij}^2 \left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2} \le \frac{\|\boldsymbol{A}_0\|_\infty}{n_1 n_2}\sqrt{\sum_{ij}\left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2}.$$

Under Condition C5, $\sqrt{\sum_{ij}(1/\hat{\theta}_{ij}-1/\theta_{ij})^2} \le \sqrt{c_{n_1,n_2}t}$ with probability at least $1-g(t)-h_{n_1,n_2}$. It implies that under Condition C3, with probability at least $1-g(t)-h_{n_1,n_2}$,

$$\begin{aligned}\left\|\Psi^{(3)}\right\| \le & \frac{\sqrt{c_{n_1,n_2}t\log(n)}a_1}{n_1n_2} \le \frac{\sqrt{c_{n_1,n_2}t\log(n)}\,(a_1\vee a_2)}{n_1n_2} \\ :=& \Delta^{(3)}(t) \asymp \frac{\sqrt{c_{n_1,n_2}t\log(n)}}{n_1n_2}.\end{aligned}$$

Since $(n_1n_2)^{-1}\log^{1/2}(n) = o((n_1n_2)^{-3/4}\log^{\delta_\sigma/4}(n))$, we have $\Delta^{(3)}(t) = o(\Delta^{(1)}(\delta_\sigma,t))$.

As for the uniform missingness, similarly as the proof in Lemma S4.1, we have that $\mathsf{E}\{1/N - 1/(n_1n_2\theta_0)\} \le 2(1-\theta_0)/(\theta_0 n_1 n_2)^2$. Then for $0<\delta_\sigma<2$, with probability at least $1-2\log^{-\delta_\sigma}(n)$, $\|\omega_{ij}(n_1n_2/N - 1/\theta_0)\boldsymbol{J}_{ij}\| \le 2(1-\theta_0)\log^{\delta_\sigma}(n)/\theta_0^2 \le 2\log^{\delta_\sigma}(n)/\theta_0^2 \le 2(n_1\vee n_2)\log(n)/\theta_0$ for $n_1n_2\theta_0 > (n_1\vee n_2)\log(n)$. Thus it is not hard to conclude that, for $0<\delta_\sigma<2$, with probability at least $1-2\log^{-\delta_\sigma}(n)$,

$$\left\|\Psi^{(3)}\right\| \le \frac{\sqrt{2\,(n_1\vee n_2)\log(n)}a_1}{\sqrt{\theta_0}n_1n_2}.$$

□

S5 Proofs of Theorem 1 and Corollary 1

Proof of Theorem 1. Under Conditions C1 and C3-C5, Lemmas S4.1-S4.3 show that there exist constants $\Delta^{(1)}(\delta_\sigma,t)$, $\Delta^{(2)}$ and $\Delta^{(3)}(t)$ such that

$$\left\|\Psi^{(1)}\right\| \le \Delta^{(1)}(\delta_\sigma,t), \quad \left\|\Psi^{(2)}\right\| \le \Delta^{(2)}, \quad \left\|\Psi^{(3)}\right\| \le \Delta^{(3)}(t),$$

with probability at least $1-1/n-g(t)-h_{n_1,n_2}-12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$, $1-1/n$ and $1-g(t)-h_{n_1,n_2}$ respectively. As defined in (4.2), $\Delta(\delta_\sigma,t) = \max\{\theta_L^{-1/2}(n_1\vee n_2)^{1/2}(n_1n_2)^{-1}\log^{1/2}(n), (n_1n_2)^{-3/4}(c_{n_1,n_2}t)^{1/2}\log^{\delta_\sigma/4}(n)\}$. We have for a positive constant C_0, $\Delta^{(1)}(\delta_\sigma,t)+\Delta^{(2)}+\Delta^{(3)}(t) \le C_0\Delta(\delta_\sigma,t)$.

It follows from the closed form of $\hat{\boldsymbol{\beta}}$ that

$$\begin{aligned}\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}_0 =& \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_1^{-1}\boldsymbol{X}^\intercal\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0\right) \\ & - \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_2\lambda_1 n_1^{-1}\boldsymbol{X}\boldsymbol{\beta}_0.\end{aligned}$$

Take $\lambda_1 = o(n_2^{-1})$, $n_2\lambda_1 = o(1)$, we have $\boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_1^{-1}\boldsymbol{X}^\intercal = \boldsymbol{P}_{\boldsymbol{X}}(1+o(1))$. It implies that,

$$\begin{aligned}\frac{1}{n_1n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2 \leq & \frac{1}{n_1n_2}\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0\right)\right\|_F^2(1+o(1)) + n_2^2\lambda_1^2\left\|\boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2(1+o(1)) \\ \leq & \frac{m}{n_1n_2}\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0\right)\right\|^2(1+o(1)) + n_2^2\lambda_1^2\left\|\boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2(1+o(1)) \\ \leq & 2mn_1n_2\left(C_0^2\Delta^2(\delta_\sigma,t) + a_1n_2^2\left\{\log(n)\right\}\lambda_1^2\right)\end{aligned}$$

with the probability at least $1 - 2/n - 2g(t) - 2h_{n_1,n_2} - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$.

It follows from the definition of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{B}}$ that

$$\begin{aligned}&\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{A}} - \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right\|_F^2 + \lambda_1\left\|\hat{\boldsymbol{\beta}}\right\|_F^2 + \lambda_2\left(\alpha\left\|\hat{\boldsymbol{B}}\right\|_* + (1-\alpha)\left\|\hat{\boldsymbol{B}}\right\|_F^2\right) \\ \leq & \frac{1}{n_1n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{B}_0 - \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right\|_F^2 + \lambda_1\left\|\hat{\boldsymbol{\beta}}\right\|_F^2 + \lambda_2\left(\alpha\left\|\boldsymbol{B}_0\right\|_* + (1-\alpha)\left\|\boldsymbol{B}_0\right\|_F^2\right). \qquad \text{(S5.1)}\end{aligned}$$

Since we can rewrite the first term in the left hand side of (S5.1) as

$$\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{A}} - \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right\|_F^2 = \frac{1}{n_1n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{B}} - \boldsymbol{B}_0 + \boldsymbol{B}_0 - \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right\|_F^2,$$

the inequality (S5.1) is equivalent to

$$\begin{aligned}\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0\right\|_F^2 \leq & \frac{2}{n_1n_2}\left(\left\langle\hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{\epsilon}\right\rangle + \left\langle\hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{A}_0 - \boldsymbol{A}_0\right\rangle\right. \\ & \left. + \left\langle\hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right\rangle + \left\langle\hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{A}_0 - \boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{A}_0\right\rangle\right) \\ & + \lambda_2\alpha\left(\left\|\boldsymbol{B}_0\right\|_* - \left\|\hat{\boldsymbol{B}}\right\|_*\right) + \lambda_2(1-\alpha)\left(\left\|\boldsymbol{B}_0\right\|_F^2 - \left\|\hat{\boldsymbol{B}}\right\|_F^2\right).\end{aligned}$$

We focus on the bound related to $\|\boldsymbol{B}_0\|_*$ in (4.3), namely,

$$d^2\left(\hat{\boldsymbol{B}}, \boldsymbol{B}_0\right) \leq C'\max\left\{\lambda_2\alpha\left\|\boldsymbol{B}_0\right\|_*, \lambda_2(1-\alpha)\left\|\boldsymbol{B}_0\right\|_F^2, n_1n_2\Delta^2(\delta_\sigma,t)\right\}, \qquad \text{(S5.2)}$$

first. By the trace duality property given in (S4.1), with probability at least $1 - 2/n - 2g(t) - 2h_{n_1,n_2} - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$, we have

$$\begin{aligned}\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0\right\|_F^2 \leq & 2\left\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0\right\|_*\left(\left\|\Psi^{(1)}\right\| + \left\|\Psi^{(2)}\right\| + \left\|\Psi^{(3)}\right\|\right) \\ & + \lambda_2\alpha\left(\left\|\boldsymbol{B}_0\right\|_* - \left\|\hat{\boldsymbol{B}}\right\|_*\right) + \lambda_2(1-\alpha)\left(\left\|\boldsymbol{B}_0\right\|_F^2 - \left\|\hat{\boldsymbol{B}}\right\|_F^2\right) \\ \leq & 2C_0\left(\left\|\hat{\boldsymbol{B}}\right\|_* + \left\|\boldsymbol{B}_0\right\|_*\right)\Delta(\delta_\sigma,t) + \lambda_2\alpha\left(\left\|\boldsymbol{B}_0\right\|_* - \left\|\hat{\boldsymbol{B}}\right\|_*\right) + \lambda_2(1-\alpha)\left(\left\|\boldsymbol{B}_0\right\|_F^2 - \left\|\hat{\boldsymbol{B}}\right\|_F^2\right).\end{aligned}$$

For $0 < \alpha \leq 1$ and $\lambda_2 \alpha \geq 2C_0 \Delta(\delta_\sigma, t)$, we can simplify the inequality to

$$\frac{1}{n_1 n_2} \left\| \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \right\|_F^2 \leq \left(2C_0 \Delta\left(\delta_\sigma, t\right) + \lambda_2 \alpha\right) \left\| \boldsymbol{B}_0 \right\|_* + \lambda_2 \left(1 - \alpha\right) \left\| \boldsymbol{B}_0 \right\|_F^2,$$

with probability at least $1 - 2/n - 2g(t) - 2h_{n_1,n_2} - 12c_\sigma^2 \eta^2 \log^{-\delta_\sigma}(n)$.

Now we focus on the bound related to $r_{\boldsymbol{B}_0}$ in (4.3), namely,

$$d^2\left(\hat{\boldsymbol{B}}, \boldsymbol{B}_0\right) \leq C' \max\left\{ n_1 n_2 r_{\boldsymbol{B}_0} \left(\lambda_2 \alpha\right)^2, \lambda_2 \left(1 - \alpha\right) \left\| \boldsymbol{B}_0 \right\|_F^2 \right\}. \tag{S5.3}$$

To prove the remaining bounds, note that for any $\boldsymbol{Z} \in \partial \|\boldsymbol{B}_0\|_*$, we have $\|\boldsymbol{B}_0\|_* + \langle \boldsymbol{Z}, \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \rangle \leq \|\hat{\boldsymbol{B}}\|_*$. The inequality (S5.1) implies, for any $\boldsymbol{Z} \in \partial \|\boldsymbol{B}_0\|_*$

$$\begin{aligned} &\frac{1}{n_1 n_2} \left\| \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \right\|_F^2 \\ \leq &\frac{2}{n_1 n_2} \left\langle \hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} - \boldsymbol{B}_0 - \boldsymbol{X}\hat{\boldsymbol{\beta}} \right\rangle + \lambda_2 \alpha \left\langle \boldsymbol{Z}, \boldsymbol{B}_0 - \hat{\boldsymbol{B}} \right\rangle + \lambda_2 \left(1 - \alpha\right) \left\| \boldsymbol{B}_0 \right\|_F^2. \end{aligned} \tag{S5.4}$$

On the other hand, by definition of $\partial \|\boldsymbol{B}_0\|_*$, $\boldsymbol{Z} = \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T} + \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \boldsymbol{W} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}$, where $\boldsymbol{W}$ is an arbitrary matrix with $\|\boldsymbol{W}\| \leq 1$. It follows from the trace duality (S4.1) that there exists $\boldsymbol{W}$ with $\|\boldsymbol{W}\| \leq 1$ such that

$$\left\langle \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \boldsymbol{W} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}, \boldsymbol{B}_0 - \hat{\boldsymbol{B}} \right\rangle = - \left\langle \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \boldsymbol{W} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}, \hat{\boldsymbol{B}} \right\rangle = \left\langle \boldsymbol{W}, \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \hat{\boldsymbol{B}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} \right\rangle = \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \hat{\boldsymbol{B}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} \right\|_*.$$

For this particular choice of $\boldsymbol{W}$, (S5.4) implies that

$$\begin{aligned} &\frac{1}{n_1 n_2} \left\| \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \right\|_F^2 + \lambda_2 \alpha \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \hat{\boldsymbol{B}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} \right\|_* \\ \leq &\frac{2}{n_1 n_2} \left\langle \hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} - \boldsymbol{B}_0 - \boldsymbol{X}\hat{\boldsymbol{\beta}} \right\rangle + \lambda_2 \alpha \left\langle \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T}, \boldsymbol{B}_0 - \hat{\boldsymbol{B}} \right\rangle + \lambda_2 \left(1 - \alpha\right) \left\| \boldsymbol{B}_0 \right\|_F^2. \end{aligned} \tag{S5.5}$$

Using the facts that $\| \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T} \| = 1$ and $\langle \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T}, \boldsymbol{B}_0 - \hat{\boldsymbol{B}} \rangle = \langle \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T}, \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}} (\boldsymbol{B}_0 - \hat{\boldsymbol{B}}) \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}} \rangle$, we deduce from (S5.5) that

$$\begin{aligned} &\frac{1}{n_1 n_2} \left\| \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \right\|_F^2 + \lambda_2 \alpha \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \hat{\boldsymbol{B}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} \right\|_* \\ \leq &2 \left\langle \hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{M} \right\rangle + \lambda_2 \alpha \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}} \left(\boldsymbol{B}_0 - \hat{\boldsymbol{B}} \right) \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}} \right\|_* + \lambda_2 \left(1 - \alpha\right) \left\| \boldsymbol{B}_0 \right\|_F^2, \end{aligned} \tag{S5.6}$$

where $\boldsymbol{M} = (\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} - \boldsymbol{B}_0 - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n_1 n_2)$.

To provide an upper bound on $2\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\boldsymbol{M}\rangle$ we use the following decomposition:

$$\begin{aligned}\left\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\boldsymbol{M}\right\rangle=&\left\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\mathcal{P}_{\boldsymbol{B}_0}\left(\boldsymbol{M}\right)\right\rangle+\left\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\rangle\\=&\left\langle\mathcal{P}_{\boldsymbol{B}_0}\left(\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right),\mathcal{P}_{\boldsymbol{B}_0}\left(\boldsymbol{M}\right)\right\rangle+\left\langle\hat{\boldsymbol{B}},\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\rangle,\end{aligned}$$

where $\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})=\boldsymbol{M}-\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}$. Due to the trace duality (S4.1),

$$\begin{aligned}2\left|\left\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\boldsymbol{M}\right\rangle\right|\leq&\Lambda\left\|\mathcal{P}_{\boldsymbol{B}_0}\left(\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right)\right\|_F+\Gamma\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*\\\leq&\Lambda\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F+\Gamma\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*,\end{aligned}$$

where $\Lambda=2\|\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})\|_F$ and $\Gamma=2\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}(\boldsymbol{M})\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\|$. Note that $\Gamma\leq 2\|\boldsymbol{M}\|\leq 2C_0\Delta(\delta_\sigma,t):=\Gamma^\star$.

Since $\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})=\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}+\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}}\boldsymbol{M}$, $\operatorname{rank}(\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}})\leq r_{\boldsymbol{B}_0}$ and $\operatorname{rank}(\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}})\leq r_{\boldsymbol{B}_0}$, we have

$$\Lambda\leq 2\sqrt{\operatorname{rank}\left(\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})\right)}\|\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})\|\leq 2\sqrt{2r_{\boldsymbol{B}_0}}C_0\Delta\left(\delta_\sigma,t\right):=\Lambda^\star.$$

Due to the facts that

$$\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}}\left(\boldsymbol{B}_0-\hat{\boldsymbol{B}}\right)\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}\right\|_*\leq\sqrt{r_{\boldsymbol{B}_0}}\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}}\left(\boldsymbol{B}_0-\hat{\boldsymbol{B}}\right)\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}\right\|_F\leq\sqrt{r_{\boldsymbol{B}_0}}\left\|\boldsymbol{B}_0-\hat{\boldsymbol{B}}\right\|_F,$$

we have

$$\begin{aligned}&\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F^2+\lambda_2\alpha\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*\\&\leq\left(\Lambda+\lambda_2\alpha\sqrt{r_{\boldsymbol{B}_0}}\right)\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F+\Gamma\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*+\lambda_2\left(1-\alpha\right)\|\boldsymbol{B}_0\|_F^2,\end{aligned}$$

which implies

$$\begin{aligned}&\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F^2+\left(\lambda_2\alpha-2C_0\Delta(\delta_\sigma,t)\right)\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*\\&\leq\left(\Lambda+\lambda_2\alpha\sqrt{r_{\boldsymbol{B}_0}}\right)\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F+\lambda_2\left(1-\alpha\right)\|\boldsymbol{B}_0\|_F^2.\end{aligned}$$

Take $\lambda_2\alpha\geq 2C_0\Delta(\delta_\sigma,t)$, we have

$$\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F^2\leq n_1n_2r_{\boldsymbol{B}_0}\left(2\sqrt{2}C_0\Delta\left(\delta_\sigma,t\right)+\lambda_2\alpha\right)^2+2\lambda_2\left(1-\alpha\right)\|\boldsymbol{B}_0\|_F^2.$$

Note that $2C_0\Delta(\delta_\sigma,t)\leq\lambda_2\alpha$, this means (S5.3) holds.

Finally, in Theorem 1, under the choice of parameters $0<\alpha\leq 1$ and $\lambda_2\alpha\geq(2+4m)C_0\Delta(\delta_\sigma,t)$, we have $n_1n_2C_0^2\Delta^2(\delta_\sigma,t)\leq n_1n_2r_{\boldsymbol{B}_0}(\lambda_2\alpha)^2$. Thus (4.3) follows from (S5.2) and (S5.3). □

Proof of Corollary 1 and Corollary 2. For Corollary 1, it is readily shown that $\sqrt{n_1 n_2 \theta_0/(1-\theta_0)}(1/\hat{\theta} - 1/\theta_0) \xrightarrow{d} \mathcal{N}(0,1)$. Since $\mathsf{P}\{(1/\hat{\theta} - 1/\theta_0)^2 \geq (1-\theta_0)t/\theta_0 \leq \mathsf{P}\{\chi_1^2 > t\} + \sup_t |\mathsf{P}\{\chi_1^2 > t\} - \mathsf{P}\{\theta_0(1/\hat{\theta} - 1/\theta_0)^2/(1-\theta_0) \geq t\}|$ where χ_1^2 is the chi-square random variable with one degree of freedom. Choose $c_{n_1,n_2} = (1-\theta_0)/\theta_0$, $t_0 > 0$, $g(t) = \mathsf{P}\{\chi_1^2 > t\}$ and $h_{n_1,n_2} = \sup_t |\mathsf{P}\{\theta_0(1/\hat{\theta} - 1/\theta_0)^2/(1-\theta_0) \geq t\} - g(t)|$ in Condition C5(b). While that $\lim_{t\to\infty} g(t) = 0$ is obvious, by Polya's theorem, $\lim_{n_1,n_2\to\infty} h_{n_1,n_2} = 0$. Thus Condition C5(b) holds for any positive t under the uniform probability of observation model. Under Condition C2 and C3, we have $\|\boldsymbol{B}_0\|_F = O\{\sqrt{n_1 n_2 \log(n)}\}$ and $\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F = O\{\sqrt{n_1 n_2 \log(n)}\}$. Thus the dominate term in the right hand side is $n_1 n_2 r_{\boldsymbol{B}_0} \Delta_1^2$.

For Corollary 2, it is shown in Section S1.4 that by taking $c_{n_1,n_2} = \eta_g^{-1} n_2 \log(n_2)$ and $t_0 = (m+3)$, we have

$$\mathsf{P}\left\{\sum_{j=1}^{n_2}\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2 \geq c_{n_1,n_2} t\right\} \leq (m+1)^{-(m+1)/2} \exp\left\{m + 2 - \frac{t}{2} + \log(t)\right\} + n_2 k_{n_1}$$

where η_g is a constant depend on θ_L, χ_{m+1}^2 is the chi-square random variable with $m+1$ degrees of freedom, and $k_{n_1} = \max_j \sup_t |\mathsf{P}\{\sum_i (1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \geq t\} - \mathsf{P}(\chi_{m+1}^2 \geq t)|$.

Take $g(t) = (m+1)^{-(m+1)/2} \exp\{m+2-t/2+\log(t)\}$, and $h_{n_1,n_2} = n_2 k_{n_1}$. Then, $\lim_{t\to\infty} g(t) = 0$. By Polya's theorem, it is shown in Section S1.4 that there exists a positive integer N such that for $n_1 > N$ and $k_{n_1} < 1/n_2^2$, which implies that $\lim_{n_1,n_2\to\infty} h_{n_1,n_2} = 0$. Thus Condition C5(b) holds for any positive $t > t_0$ for the logistic model. Choose t as (4.5), we have $\sup_t \Delta(\delta_\sigma, t) = \Delta_2(\delta_\sigma) \asymp \eta_g^{-1/2} n_1^{-3/4} n_2^{-1/4} \log^{1/2}(n_2) \log^{\delta_\sigma/3}(n)$. This implies that the convergence rate for $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ given in (4.3) is $\eta_g^{-1} n_1^{-1/2} n_2^{1/2} \log(n_2) \log^{2\delta_\sigma/3}(n)$. Under Condition C2 and C3, we have $\|\boldsymbol{B}_0\|_F = O\{\sqrt{n_1 n_2 \log(n)}\}$ and $\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F = O\{\sqrt{n_1 n_2 \log(n)}\}$. Thus the dominate term in the right hand side is $n_1 n_2 r_{\boldsymbol{B}_0} \Delta_2^2(\delta_\sigma)$.

Assume that $n_1 \asymp \eta_g^2 n_2 \log^{2+2\delta_\sigma}(n_2)$, then right hand side becomes $r_{\boldsymbol{B}_0} \log^{-2\delta_\sigma/3}(n_2)$. □

S6 Proof of Theorem 2

Proof of Theorem 2. Since that $\lambda_1 = o(n_2^{-1})$, $n_2\lambda_1 = o(1)$, we have

$$\left(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m}\right)^{-1} \to \boldsymbol{S}_x^{-1}.$$

We have the estimators $\hat{\theta}_{ij}$ of θ_{ij} satisfy that for $|\hat{\theta}_{ij} - \theta_{ij}| = O_p(n_1^{-1/2})$. Thus for the jth column of matrix $\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}$, we have

$$\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right)_j = \left(\boldsymbol{W}\circ\left(1+O_p\left(n_1^{-1/2}\right)\right)\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\right)_j.$$

Let $\boldsymbol{Z}_j = n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})_j$. Then $Z_{kj} = n_1^{-1}\sum_{i=1}^{n_1} x_{ik}\omega_{ij}Y_{ij}/\theta_{ij}$ for each $k = 1,\ldots,m$. Since $\mathsf{E}(x_{ik}\omega_{ij}Y_{ij}/(n_1\theta_{ij})) = x_{ik}(X\beta_0 + B_0)_{ij}/n_1$, $\mathsf{Var}(x_{ik}\omega_{ij}Y_{ij}/(n_1\theta_{ij})) = x_{ik}^2(1-\theta_{ij})\{(X\beta_0 + B_0)_{ij}^2 + \sigma_{ij}^2\}/(n_1^2\theta_{ij})$, define $s_{n_1}^2 = \sum_{i=1}^{n_1} x_{ik}^2(1-\theta_{ij})\{(X\beta_0+B_0)_{ij}^2+\sigma_{ij}^2\}/(n_1^2\theta_{ij})$. Also we have

$$\begin{aligned}\mathsf{E}\left|x_{ik}\omega_{ij}Y_{ij}/\left(n_1\theta_{ij}\right) - x_{ik}\left(X\beta_0+B_0\right)_{ij}/n_1\right|^3 = \Big(&x_{ik}^3(1-\theta_{ij})\{(X\beta_0+B_0)_{ij}^3 + (X\beta_0+B_0)_{ij}\sigma_{ij}^2\}/\theta_{ij}^2\\ &-3x_{ik}^3\left(1-\theta_{ij}\right)\left\{\left(X\beta_0+B_0\right)_{ij}^3 + \left(X\beta_0+B_0\right)_{ij}\sigma_{ij}^2\right\}/\theta_{ij} + 2x_{ik}^3\left(X\beta_0+B_0\right)_{ij}^3\Big)/n_1^3,\end{aligned}$$

implies the Lyapunovs condition satisfied, namely,

$$\lim_{n_1\to\infty}\frac{1}{s_{n_1}^3}\sum_{i=1}^{n_1}\mathsf{E}\left|x_{ik}\omega_{ij}Y_{ij}/\theta_{ij} - x_{ik}\left(X\beta_0+B_0\right)_{ij}\right|^3 = 0.$$

By Lyapunov Central Limit Theorem, we have

$$\frac{1}{s_{n_1}}\sum_{i=1}^{n_1}\left(x_{ik}\omega_{ij}Y_{ij}/\left(n_1\theta_{ij}\right) - x_{ik}\left(X\beta_0+B_0\right)_{ij}/n_1\right) \xrightarrow{d} \mathcal{N}\left(0,1\right).$$

Combining with $n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}\to\boldsymbol{S}_x$, we have $\boldsymbol{Z}_j = n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})_j = \boldsymbol{S}_x\boldsymbol{\beta}_{0j} + O_p(\mathbf{1}/\sqrt{\boldsymbol{n_1}})$.

For the estimator $\hat{\boldsymbol{\beta}}_j = (n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}+n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y})_j = (1+o(1))\boldsymbol{S}_x^{-1}n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ(1+O_p(n_1^{-1/2}))\boldsymbol{\Theta}^*\circ\boldsymbol{Y})_j$, we have $\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{0j}\xrightarrow{p}0$ and $\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{0j}\|_F^2 = O_p(m/n_1) = O_p(1/n_1)$. This completes the proof of Theorem 2. □

Table S1: Empirical root mean square errors (RMSEs), test errors, estimated ranks and their standard errors (in parentheses) under model $\boldsymbol{A}_0 = \boldsymbol{B}_0$ and uniform observation mechanism (UNI), with (n_1, n_2)=(400,400), (600,600), (800,800), (1000,1000) $m = 20$, and $r = 10$, for two versions of the proposed methods, and the four existing methods (SZ, NW, KLT and MHT).

$n_1 = n_2 = 400$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.0121 (1e-04)	2.2346 (0.015)	2.2354 (0.015)	0.5723 (0.0071)	62.41 (1.59)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.0121 (1e-04)	2.2342 (0.015)	2.2350 (0.015)	0.5721 (0.0071)	62.23 (1.58)
SZ			2.1082 (0.0167)	0.5059 (0.0076)	46.76 (2.74)
NW			2.0417 (0.0172)	0.4722 (0.0076)	94.48 (5.73)
KLT			2.2565 (0.0148)	0.5827 (0.007)	42.07 (1.58)
MHT			2.0550 (0.0171)	0.4796 (0.0076)	51.42 (2.57)
$n_1 = n_2 = 600$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.0147 (1e-04)	2.0246 (0.0104)	2.0257 (0.0104)	0.4540 (0.0044)	75.82 (1.49)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.0147 (1e-04)	2.0206 (0.0105)	2.0217 (0.0105)	0.4521 (0.0044)	74.51 (1.4)
SZ			1.8500 (0.0132)	0.3725 (0.0048)	58.17 (5.15)
NW			1.7794 (0.013)	0.3425 (0.0047)	120.92 (10.29)
KLT			2.0389 (0.0106)	0.4594 (0.0045)	55.49 (1.49)
MHT			1.7902 (0.011)	0.3476 (0.0042)	66.43 (2.46)
$n_1 = n_2 = 800$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.0170 (1e-04)	1.8712 (0.0093)	1.8728 (0.0092)	0.3794 (0.0036)	85.54 (1.38)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.0170 (1e-04)	1.8617 (0.0093)	1.8633 (0.0093)	0.3753 (0.0036)	82.46 (1.19)
SZ			1.6731 (0.0105)	0.2956 (0.0034)	60.91 (5.64)
NW			1.6055 (0.0085)	0.2707 (0.0029)	130.13 (6.05)
KLT			1.8817 (0.0092)	0.3824 (0.0036)	64.86 (1.36)
MHT			1.6107 (0.0099)	0.2734 (0.0032)	80.98 (6.26)
$n_1 = n_2 = 1000$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.0185 (1e-04)	1.7238 (0.0073)	1.7258 (0.0073)	0.3275 (0.0027)	93.03 (1.36)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.0185 (1e-04)	1.7090 (0.0073)	1.7111 (0.0073)	0.3216 (0.0026)	88.14 (1.12)
SZ			1.5076 (0.0069)	0.2435 (0.0023)	72.89 (2.72)
NW			1.4485 (0.0103)	0.2234 (0.0029)	157.62 (18.01)
KLT			1.7317 (0.0073)	0.3291 (0.0027)	72.37 (1.27)
MHT			1.4556 (0.0068)	0.2260 (0.0021)	85.43 (2.48)

S7 (Cont') Simulation study

S8 (Cont') Empirical Study

As suggested at `http://files.grouplens.org/datasets/movielens/ml-1m-README.txt`, we divide age into 7 categories: under 18, $18-24$, $25-34$, $35-44$, $45-49$, $50-55$ and 56+ in the modeling of probability estimator $\hat{\boldsymbol{\Theta}}^*$. However, it will cost much more ranks than keep it as numerical in the covariate $\boldsymbol{X}$ for prediction. To achieve a balance, we merge some age categories to form three to seven categories of the age variable. Specifically, the three categories layout is: under 24, $25-49$ and 50+; the four categories: under 24, $25-34$, $35-49$ and 50+; the five categories: under 24, $25-34$, $35-44$, $45-49$ and 50+; the six categories: under 18, $18-24$, $25-34$, $35-44$, $45-49$ and 50+; and the seven categories:under 18, $18-24$, $25-34$, $35-44$, $45-49$, $50-55$ and 56+. The predictions errors of using the four and five age categories are the best among the choices of three to seven categorization of the age.

Table S2: Root mean square prediction errors (RMSPEs) and ranks of the completed matrix based on Split1 and Split2 for the two versions of the proposed method (SVT-$\hat{\alpha}$-LOG) and ($\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG) and the four existing methods proposed respectively in Sun and Zhang (2012)(SZ), Negahban and Wainwright (2012)(NW), Koltchinskii et al. (2011)(KLT) and Mazumder et al. (2010)(MHT).

		Split1		Split2		Overall
rank(X)		RMSPE	Rank	RMSPE	Rank	RMSPE
2	SVT-$\hat{\alpha}$-LOG	0.9415	47	0.9541	45	0.9478
	$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.9416	45	0.9543	42	0.9480
4	SVT-$\hat{\alpha}$-LOG	0.9420	48	0.9540	42	0.9480
	$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.9423	46	0.9540	42	0.9482
5	SVT-$\hat{\alpha}$-LOG	0.9420	49	0.9544	43	0.9483
	$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.9422	47	0.9544	43	0.9483
	SZ	0.9412	39	0.9563	31	0.9488
	NW	0.9421	269	0.9589	289	0.9506
	KLT	0.9584	1	0.9688	1	0.9636
	MHT	0.9414	56	0.9568	46	0.9491

Table S2 reports the root mean square prediction errors (RMSPEs), estimated ranks and overall

RMSPEs of different estimators for both Split1 and Split2. The result with two categorical covariate $\boldsymbol{X}$ are included. Similarly as the simulation results reported in the Section 6, SVT-$\hat{\alpha}$-LOG and $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG produced highly comparable results, which indicated the applicability of $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG to larger data sets whenever computational resources are scarce. In Split2, the proposed methods outperformed SZ NW, KLT and MHT in terms of smaller RMSPEs and either smaller or more reasonable rank estimation. Although the proposed methods were slightly inferior to SZ and MHT in Split1, they outperformed SZ and MHT significantly in Split2 by having smaller RMSPEs. Among the ten matrix completion methods considered, the six proposed methods and the KLT method offered the most consistent results between Split1 and Split2, while the other three methods exhibited much larger variations, especially in the estimated ranks. Overall speaking, the two proposed methods were among the top two performers of the analysis reported in Table S2.

References

Ahlswede, R. and Winter, A. (2002), "Strong Converse for Identification via Quantum Channels," *IEEE Transactions on Information Theory*, 48, 569–579.

Candès, E. J. and Recht, B. (2009), "Exact Matrix Completion via Convex Optimization," *Foundations of Computational Mathematics*, 9, 717–772.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011), "Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion," *The Annals of Statistics*, 39, 2302–2329.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322.

Negahban, S. and Wainwright, M. J. (2012), "Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds with Noise," *Journal of Machine Learning Research*, 13, 1665–1697.

Sun, T. and Zhang, C.-H. (2012), "Calibrated Elastic Regularization in Matrix Completion," in *Advances in Neural Information Processing Systems*, pp. 863–871.

Sweeting, T. (1980), "Uniform Asymptotic Normality of the Maximum Likelihood Estimator," *The Annals of Statistics*, 8, 1375–1381.

Tropp, J. A. (2012), "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12, 389–434.

Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*, New York: Springer-Verlag New York.