Supplemental Document for "Matrix Completion with Model-free Weighting"

Jiayi Wang [1] **Raymond K. W. Wong** [1] **Xiaojun Mao** [2] **Kwun Chuen Gary Chan** [3]

A. Proof of Lemma 1

Proof of Lemma 1. We first list two properties of max norm as follows. (i) As shown in Srebro & Shraibman (2005), $\|\boldsymbol{B}\|_* \leq \sqrt{n_1 n_2}\|\boldsymbol{B}\|_{\max}$. (ii) By an equivalent definition of max norm due to Lee et al. (2008) (also see equation (8) in Jalali & Srebro (2012)), we have $\|\boldsymbol{C} \circ \boldsymbol{B}\| \leq \|\boldsymbol{C}\|\|\boldsymbol{B}\|_{\max}$. Together with the duality of nuclear norm, we can show that

$$\begin{aligned} |\langle \boldsymbol{C} \circ \boldsymbol{B}, \boldsymbol{B}\rangle| &\leq \|\boldsymbol{C} \circ \boldsymbol{B}\|\|\boldsymbol{B}\|_* \leq \|\boldsymbol{C}\|\|\boldsymbol{B}\|_{\max}\|\boldsymbol{B}\|_* \\ &\leq \sqrt{n_1 n_2}\|\boldsymbol{C}\|\|\boldsymbol{B}\|_{\max}^2. \end{aligned} \tag{S1}$$

□

B. Proofs of Theorems 1, 2 and 3

Let $\boldsymbol{e}_i(n) \in \mathbb{R}^n$ be the canonical basis vector, i.e., the i-th element of $\boldsymbol{e}_i(n)$ is 1 and the remaining elements are 0. We can define similar standard basis elements for n_1-by-n_2 matrices: $\boldsymbol{J}_{ij} = \boldsymbol{e}_i(n_1)\boldsymbol{e}_j^\intercal(n_2)$, which will be used in the applications of matrix Bernstein inequality in our proofs. For any $\beta \geq 0$, define the class of matrices $\mathcal{B}_{\max}(\beta)$ to be the max-norm ball with radius β, i.e.,

$$\mathcal{B}_{\max}(\beta) = \{\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2} : \|\boldsymbol{A}\|_{\max} \leq \beta\}.$$

We also define

$$\mathcal{F} = \{\boldsymbol{u}\boldsymbol{v}^T : \boldsymbol{u} \in \{-1,+1\}^{n_1}, \boldsymbol{v} \in \{-1,+1\}^{n_2}\},$$

the set of rank-one sign matrices. Denote by $K_G \in (1.67, 1.79)$ the Grothendieck's constant. From Srebro & Shraibman (2005),

$$\text{conv}\mathcal{F} \subseteq \mathcal{B}_{\max}(1) \subseteq K_G \text{conv}\mathcal{F}. \tag{S2}$$

Moreover, the cardinality of $\mathcal{F}$ is $|\mathcal{F}| = 2^{n_1+n_2-1}$.

Lemma S1. *Suppose Assumption 1 hold. Let $\boldsymbol{W}_\diamond = (w_{\diamond,i,j}) \in \mathbb{R}^{n_1,n_2}$ where $W_{\diamond,i,j} = \pi_{i,j}^{-1}$. There exists a constant $C_1 \geq 0$ such that with probability at least $1 - 1/(n_1+n_2)$,*

$$\frac{1}{n_1 n_2}\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| \leq C_1 \min\left\{\frac{\log^{1/2}(n_1+n_2)}{\sqrt{\pi_L (n_1 \wedge n_2) n_1 n_2}}, \frac{\sqrt{n_1+n_2}}{\pi_L n_1 n_2}\right\}.$$

Proof of Lemma S1. We use two different proof techniques to show the bounds. Depending on the rate of π_L, one of these two bounds is faster.

First, we show the proof for deriving the second bound. As $\{T_{ij}\}$ are Bernoulli random variables, each entry $T_{ij}W_{\diamond,i,j} - 1$ of matrix $\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}$ is sub-Gaussian random variable. Thus according to the definition of the ψ_2 norm, we have

$$\mathsf{E}\exp\{\log(2) \cdot (T_{ij}W_{\diamond,i,j} - 1)^2/(\pi_{ij}^{-1} - 1)^2\} \leq 2,$$

which implies that $\|T_{ij}W_{\diamond,i,j} - 1\|_{\psi_2} \leq \log^{-1/2} 2 \cdot (\pi_{ij}^{-1} - 1) \leq 2(\pi_{ij}^{-1} - 1)$.

By Theorem S2 in Section D , taking $K = \max_{i,j}\|T_{ij}W_{\diamond,i,j} - 1\|_{\psi_2} \leq 2\pi_L^{-1}$ and $t = (n_1+n_2)^{1/2}$ in Theorem S2, there exists an absolute constant $C_1 > 0$ such that

$$\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| \leq \frac{C_1\sqrt{n_1+n_2}}{\pi_L},$$

with probability at least $1 - 2 \cdot \exp(-(n_1 + n_2))$.

Next, we consider applying the Matrix Bernstein inequality to derive the first bound. For $(n_1 n_2)^{-1} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| = \| \sum_{i,j} (T_{ij} W_{\diamond,i,j} - 1) \boldsymbol{J}_{ij} / (n_1 n_2) \|$, where $\boldsymbol{J}_{ij}$ has 1 for $(i,j)-$th, but 0 for all the remaining entries, let $\boldsymbol{M}_{i,j} = (T_{ij} W_{\diamond,i,j} - 1) \boldsymbol{J}_{ij}$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$, then $(n_1 n_2)^{-1} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| = \|(n_1 n_2)^{-1} \sum_{i,j} \boldsymbol{M}_{i,j}\|$. We can easily verify that $\mathsf{E}(\boldsymbol{M}_{i,j}) = \boldsymbol{0}$ and $\|\boldsymbol{M}_{i,j}\| \le \max\{\pi_L^{-1} - 1, 1\}$ for each i, j by Assumption 1.

Since $\mathsf{E}(T_{ij} W_{\diamond,i,j} - 1)^2 = \pi_{ij}^{-1} - 1$, we can show that

$$
\begin{aligned}
&\left\| \frac{1}{n_1 n_2} \sum_{i,j} \mathsf{E}\left(\boldsymbol{M}_{i,j} \boldsymbol{M}_{i,j}^\intercal\right) \right\| = \left\| \frac{1}{n_1 n_2} \sum_{i,j} \mathsf{E}\left(\boldsymbol{M}_{i,j}^\intercal \boldsymbol{M}_{i,j}\right) \right\| \\
\le & \frac{1}{n_1 n_2} \max \left\{ \max_{1 \le i \le n_1} \sum_{j=1}^{n_2} |1/\pi_{ij} - 1|, \max_{1 \le j \le n_2} \sum_{i=1}^{n_1} |1/\pi_{ij} - 1| \right\} \\
\le & \frac{1}{n_1 \wedge n_2} |1/\pi_L - 1|,
\end{aligned}
$$

where the first inequality comes from Corollary 2.3.2 in Golub & Van Loan (1996).

By Theorem S3 in Section D , with probability at least $1 - 1/(n_1 + n_2)$, we have

$$
\frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| \le 2 \max \left\{ \sqrt{\frac{2\,|1/\pi_L - 1| \log(n_1 + n_2)}{(n_1 \wedge n_2)\, n_1 n_2}}, 2 \max\left\{ \frac{1}{\pi_L} - 1, 1 \right\} \frac{\log^{3/2}(n_1 + n_2)}{n_1 n_2} \right\}.
$$

Overall, the conclusion follows. □

Lemma S2. *Suppose Assumption 1 holds. With probability at least* $1 - \exp\{-2^{-1} (\log 2) \pi_L^2 \sum_{i,j} \pi_{ij}^{-1}\}$,

$$
\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2 \le 2 \sum_{i,j} \pi_{ij}^{-1}.
$$

In particular, the probability is lower bounded by $1 - \exp\{-2^{-1} (\log 2) n_1 n_2 \pi_L^2 \pi_U^{-1}\}$.

Proof of Lemma S2. Note that $\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2 = \sum_{i,j} T_{ij} \pi_{ij}^{-2}$. Let $\xi > 0$. By Markov inequality, for any $t \ge 0$,

$$
\begin{aligned}
\Pr\left(\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2 \ge t\right) = \Pr\left\{\exp(\xi \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2) \ge \exp(\xi t)\right\} &\le \exp(-\xi t) \mathsf{E} \exp\left(\xi \sum_{i,j} T_{ij} \pi_{ij}^{-2}\right) \\
&= \exp(-\xi t) \prod_{i,j} \mathsf{E} \exp(\xi T_{ij} \pi_{ij}^{-2}).
\end{aligned}
$$

For each (i,j), due to the inequality $1 + x \le \exp(x)$ for $x \ge 0$,

$$
\mathsf{E} \exp(\xi T_{ij} \pi_{ij}^{-2}) = 1 + \{\exp(\xi \pi_{ij}^{-2}) - 1\} \pi_{ij} \le \exp[\{\exp(\xi \pi_{ij}^{-2}) - 1\} \pi_{ij}].
$$

Combining with the above result and taking $t = 2 \sum_{i,j} \pi_{ij}^{-1}$,

$$
\begin{aligned}
\Pr\left(\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2 \ge 2 \sum_{i,j} \pi_{ij}^{-1}\right) &\le \exp\left[-2\xi \sum_{i,j} \pi_{ij}^{-1} + \sum_{i,j} \{\exp(\xi \pi_{ij}^{-2}) - 1\} \pi_{ij}\right] \\
&= \exp\left[-\sum_{i,j} \pi_{ij} \left\{1 + 2\xi \pi_{ij}^{-2} - \exp(\xi \pi_{ij}^{-2})\right\}\right].
\end{aligned}
$$

Note the above inequality holds for any $\xi > 0$.

Next, we focus on the term $g(\xi\pi_{ij}^{-2})$ where $g(x) = 1 + 2x - \exp(x)$ for $x \geq 0$. It is easy to show that g attains its maximum at $x = \log 2$, and $g(\log 2) = 2\log 2 - 1 > 0$. Also, $g(x)$ is increasing for $0 \leq x \leq \log 2$.

Take $\xi = (\log 2)\pi_L^2$. Then $0 \leq \xi\pi_{ij}^{-2} \leq \log 2$, and hence $g(\xi\pi_{ij}^{-2}) > 0$, for all i, j. The lower bound of $g(\xi\pi_{ij}^{-2})$ is crucial in determining the order of the probability bound. Since $g(x) \geq x/2$ for $0 \leq x \leq \log 2$,

$$g(\xi\pi_{ij}^{-2}) = g(\pi_L^2\pi_{ij}^{-2}\log 2) \geq \frac{\log 2}{2}\pi_L^2\pi_{ij}^{-2}, \quad \forall i, j$$

We conclude that

$$\sum_{i,j}\pi_{ij}\left\{1 + 2\xi\pi_{ij}^{-2} - \exp(\xi\pi_{ij}^{-2})\right\} \geq \frac{\log 2}{2}\pi_L^2\sum_{i,j}\pi_{ij}^{-1} \geq \frac{\log 2}{2}n_2n_2\pi_L^2\pi_U^{-1},$$

which leads to the desired result. □

With these two lemmas, we are posed to prove Theorem 1.

Proof of Theorem 1. By Lemma S2, we can show that with probability at least $1 - \exp\{-2^{-1}(\log 2)\pi_L^2\sum_{i,j}\pi_{ij}^{-1}\}$, $\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F \leq (2\sum_{i,j}\pi_{ij}^{-1})^{1/2}$ and hence $\boldsymbol{W}_\diamond$ is feasible for the constrained optimization (5).

Based on the definition of the proposed estimator $\widehat{\boldsymbol{W}}$, we have

$$\begin{aligned} S\left(\widehat{\boldsymbol{W}}, \boldsymbol{\Delta}\right) &= \frac{1}{n_1n_2}\left|\left\langle \boldsymbol{\Delta}, \left(\boldsymbol{T} \circ \widehat{\boldsymbol{W}} - \boldsymbol{J}\right) \circ \boldsymbol{\Delta}\right\rangle\right| \\ &\leq \frac{1}{\sqrt{n_1n_2}}\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\|\,\|\boldsymbol{\Delta}\|_{\max}^2 \\ &\leq \frac{\beta'^2}{\sqrt{n_1n_2}}\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\|. \end{aligned}$$

The desired result then follows from Lemma S1. □

Our theoretical result of the final estimator $\widehat{\boldsymbol{A}}$ will be based on a key lemma (Lemma S4), which establishes the dual of max norm of random matrix $\boldsymbol{\epsilon}$ with general entry-wise scaling. Before we prove Lemma S4, we now show a comparison theorem between sub-Gaussian complexity and Gaussian complexity. This result (Lemma S3) extends Theorem 8 in Banerjee et al. (2014) to allow arbitrary entrywise scaling.

Define the Gaussian width and Gaussian complexity of the set $\mathcal{A}$ respectively as

$$w(\mathcal{A}) = \mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{A}\in\mathcal{A}}\langle \boldsymbol{A}, \boldsymbol{G}\rangle\right] \quad \text{and} \quad \tilde{w}(\mathcal{A}) = \mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{A}\in\mathcal{A}}|\langle \boldsymbol{A}, \boldsymbol{G}\rangle|\right],$$

where $\boldsymbol{G} = (G_{ij})$ and each $\{G_{ij}\}$ are independent standard Gaussian random variables. In our study, $\mathcal{A}$ is a max-norm ball, and so is symmetric. Therefore Gaussian width and Gaussian complexity are equivalent.

Lemma S3 (Extension of Theorem 8 in Banerjee et al. (2014))**.** *Suppose Assumption 2 holds. Let $\boldsymbol{B} = (B_{ij}) \in \mathbb{R}^{n_1\times n_2}$ be a fixed matrix such that $B_{ij} \geq 0$ for each i, j. Then*

$$\mathsf{E}\left[\|\boldsymbol{B} \circ \boldsymbol{\epsilon}\|_{\max}^*\right] \leq \eta_0\tau\mathsf{E}\left[\|\boldsymbol{B} \circ \boldsymbol{G}\|_{\max}^*\right],$$

where $\|\cdot\|_{\max}^*$ is the dual norm of max norm, $\boldsymbol{G} = (G_{ij})$ has independent standard Gaussian entries which are also independent of the random errors $\{\epsilon_{ij}\}$, and $\eta_0 > 0$ is an absolute constant.

Proof of Lemma S3. Since the desired result obviously holds if $\boldsymbol{B} = \boldsymbol{0}$, we assume $\boldsymbol{B} \neq \boldsymbol{0}$ in the rest of this proof. By definition, $\|\boldsymbol{C}\|_{\max}^* = \sup_{\|\boldsymbol{X}\|_{\max}\leq 1}\langle \boldsymbol{X}, \boldsymbol{C}\rangle$ for any $\boldsymbol{C} \in \mathbb{R}^{n_1\times n_2}$. Therefore our goal is to bound a scaled sub-Gaussian complexity via the corresponding scaled Gaussian complexity. We now extend the proof of Theorem 8 of Banerjee et al. (2014) to allow an additional entrywise scaling parameter $\boldsymbol{B}$. We start with considering the sub-Gaussian process $Y_{\boldsymbol{X}} = \langle \boldsymbol{X}, \boldsymbol{B} \circ \boldsymbol{\epsilon} \rangle$ and the Gaussian process $Z_{\boldsymbol{X}} = \langle \boldsymbol{X}, \boldsymbol{B} \circ \boldsymbol{G} \rangle$, both indexed by $\boldsymbol{X} \in \mathcal{B}_{\max}(1)$. For any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)$, by the general Hoeffding's inequality given in Theorem 2.6.3 of Vershynin (2018), we have

$$\Pr\left(|Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}| \geq t\right) \leq 2 \cdot \exp\left(-\frac{C_1 t^2}{\tau^2 \left\|\boldsymbol{B} \circ (\boldsymbol{X}_1 - \boldsymbol{X}_2)\right\|_F^2}\right), \quad t > 0, \tag{S3}$$

where $C_1 > 0$ is an absolute constant. One can show that $\mathsf{E}(Z_{\boldsymbol{X}_1} - Z_{\boldsymbol{X}_2})^2 = \|\boldsymbol{B} \circ (\boldsymbol{X}_1 - \boldsymbol{X}_2)\|_F^2$. According to Theorem 2.1.5 of Talagrand (2006), we can apply the generic chaining argument for upper bounds on the empirical processes $\sqrt{c} Y_{\boldsymbol{X}}/\tau$ and $Z_{\boldsymbol{X}}$. This yields

$$\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} |Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}|\right] \leq \eta_1 \tau \mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Z_{\boldsymbol{X}_1}\right] = \eta_1 \tau w(\mathcal{B}_{\max}(1)), \tag{S4}$$

where η_1 is an absolute constant. Further, we can see that if $\boldsymbol{X} \in \mathcal{B}_{\max}(1)$, then $-\boldsymbol{X} \in \mathcal{B}_{\max}(1)$. Then we have

$$\begin{aligned}\sup_{\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} |Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}| &= \sup_{\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} (Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}) = \sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}_1} + \sup_{\boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} (-Y_{\boldsymbol{X}_2}) \\ &= \sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}_1} + \sup_{-\boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} (\langle -\boldsymbol{X}_2, \boldsymbol{B} \circ \boldsymbol{\epsilon} \rangle) = 2 \sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}_1}.\end{aligned}$$

By taking the expectation on $\boldsymbol{\epsilon}$ on both side, we have

$$\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} |Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}|\right] = 2\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}_1}\right]. \tag{S5}$$

As a result, with $\eta_0 = \eta_1/2$, we have

$$\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} \langle \boldsymbol{B} \circ \boldsymbol{\epsilon}, \boldsymbol{X} \rangle\right] = \mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}}\right] \leq \eta_0 \tau w(\mathcal{B}_{\max}(1)). \tag{S6}$$

That completes the proof. □

Lemma S4. *Suppose Assumption 2 holds. Let $\boldsymbol{B} = (B_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ be a fixed matrix such that $B_{ij} \geq 0$ for each i, j. There exists an absolute constant $C_2 > 0$ such that, with probability at least $1 - 2\exp\{-(n_1 + n_2)\}$,*

$$\|\boldsymbol{B} \circ \boldsymbol{\epsilon}\|_{\max}^* \leq C_2 \tau \|\boldsymbol{B}\|_F \sqrt{n_1 + n_2}.$$

Proof of Lemma S4. Define the set

$$\widetilde{\mathcal{B}}_{\max}(\beta) = \{\boldsymbol{B} \circ \boldsymbol{X} : \boldsymbol{X} \in \mathcal{B}_{\max}(\beta)\} \subset \mathbb{R}^{n_1 \times n_2}.$$

Note that we have

$$\mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} \langle \boldsymbol{B} \circ \boldsymbol{G}, \boldsymbol{X} \rangle\right] = \mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} \langle \boldsymbol{G}, \boldsymbol{B} \circ \boldsymbol{X} \rangle\right] = w(\widetilde{\mathcal{B}}_{\max}(1)).$$

Write $\widetilde{\mathcal{F}} = \{\boldsymbol{B} \circ \boldsymbol{X} : \boldsymbol{X} \in \mathcal{F}\}$. By the the relationship (S2), we have

$$\widetilde{\mathcal{F}} \subseteq \widetilde{\mathcal{B}}_{\max}(1) \subseteq \{\boldsymbol{B} \circ \boldsymbol{X} : \boldsymbol{X} \in K_G \operatorname{conv}(\mathcal{F})\} = K_G\{\boldsymbol{B} \circ \boldsymbol{X} : \boldsymbol{X} \in \operatorname{conv}(\mathcal{F})\} = K_G \operatorname{conv}(\widetilde{\mathcal{F}}).$$

Due to the properties of Gaussian width (see, e.g., Appendix A.1 of Banerjee et al., 2014), we have

$$w(\widetilde{\mathcal{B}}_{\max}(1)) \leq w(K_G \operatorname{conv}(\widetilde{\mathcal{F}})) = K_G w(\operatorname{conv}(\widetilde{\mathcal{F}})) = K_G w(\widetilde{\mathcal{F}}).$$

As for any $\boldsymbol{X} \in \mathcal{F}$, we have

$$\|\boldsymbol{B} \circ \boldsymbol{X}\|_F = \|\boldsymbol{B}\|_F,$$

and so $\langle \boldsymbol{G}, \boldsymbol{B} \circ \boldsymbol{X} \rangle \sim \mathcal{N}(0, \|\boldsymbol{B}\|_F^2)$. Recall that the $|\mathcal{F}| = 2^{n_1+n_2-1}$. By Proposition 3.1(ii) of Koltchinskii (2011), we have

$$w(\widetilde{\mathcal{F}}) \leq C_3 \|\boldsymbol{B}\|_F \sqrt{n_1+n_2}.$$

where C_3 is a absolute constant. By Lemma S3, we conclude that

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X}\in\mathcal{B}_{\max}(1)} \langle \boldsymbol{B}\circ\boldsymbol{\epsilon}, \boldsymbol{X}\rangle\right] &\leq \eta_0\tau\mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{X}\in\mathcal{B}_{\max}(1)} \langle \boldsymbol{B}\circ\boldsymbol{G}, \boldsymbol{X}\rangle\right] \\ &= \eta_0\tau w(\widetilde{\mathcal{B}}_{\max}(1)) \leq K_G\eta_0\tau w(\widetilde{\mathcal{F}}) \\ &\leq C_3K_G\eta_0\tau\|\boldsymbol{B}\|_F\sqrt{n_1+n_2}. \end{aligned} \tag{S7}$$

Let $\varphi(\boldsymbol{Z}) = \sup_{\|\boldsymbol{X}\|_{\max}\leq 1}\langle \boldsymbol{B}\circ\boldsymbol{Z}, \boldsymbol{X}\rangle$ for any $\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}$. We aim to provide the concentration of $\varphi(\boldsymbol{\epsilon})$ to its expectation. For notational simplicity, we will focus on the setting with $B_{ij} > 0$ for all i, j; otherwise, one can reduce the support of φ to those entries corresponding to non-zero B_{ij}. Due to the possibly unbounded support of $\boldsymbol{\epsilon}$, we adopt an extension of McDiarmid's inequality Kontorovich (2014) with unbounded diameter. For any $\boldsymbol{Z}_1 = (Z_{1,ij}), \boldsymbol{Z}_2 = (Z_{2,ij}) \in \mathbb{R}^{n_1\times n_2}$,

$$\begin{aligned}|\varphi(\boldsymbol{Z}_1) - \varphi(\boldsymbol{Z}_2)| &\leq \sup_{\|\boldsymbol{X}\|_{\max}\leq 1} |\langle \boldsymbol{B}\circ\boldsymbol{Z}_1, \boldsymbol{X}\rangle - \langle \boldsymbol{B}\circ\boldsymbol{Z}_2, \boldsymbol{X}\rangle| \\ &\leq \sup_{\|\boldsymbol{X}\|_{\max}\leq 1} \sum_{i,j} B_{ij}|X_{ij}||Z_{1,ij} - Z_{2,ij}| \\ &\leq \sup_{\boldsymbol{X}\in K_G\mathrm{conv}(\mathcal{F})} \sum_{i,j} B_{ij}|X_{ij}||Z_{1,ij} - Z_{2,ij}| \\ &\leq q(\boldsymbol{Z}_1, \boldsymbol{Z}_2),\end{aligned}$$

where $q(\boldsymbol{Z}_1, \boldsymbol{Z}_2) =: \sum_{i,j} q_{ij}(Z_{1,ij}, Z_{2,ij}) =: \sum_{i,j} K_G B_{ij}|Z_{1,ij} - Z_{2,ij}|$ is a metric. Therefore φ is 1-Lipschitz with respect to the metric q. Let ε'_{ij} be an independent copy of ε_{ij}, and γ_{ij} be an independent Rademacher random variable. We can show that the subgaussian norm of $\gamma_{ij}q_{ij}(\varepsilon_{ij}, \varepsilon'_{ij})$ is bounded by $C_4\tau b_{ij}$ for some absolute constant $C_4 > 0$. By Theorem 1 of Kontorovich (2014), we conclude that

$$\mathsf{P}(|\varphi(\boldsymbol{\epsilon}) - \mathsf{E}\varphi(\boldsymbol{\epsilon})| > t) \leq 2\exp\left(-\frac{t^2}{2C_4\tau^2\|\boldsymbol{B}\|_F^2}\right), \quad t \geq 0.$$

Combining with (S7), we achieve the desired result.

□

Lemma S5. *Suppose Assumptions 1 and 2 hold. There exists an absolute constant* $C_2 > 0$ *such that with probability at least* $1 - 2\exp\{-(n_1+n_2)\}$,

$$\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\right\|_{\max}^* \leq C_2\tau\kappa\sqrt{n_1+n_2}.$$

Proof of Lemma S5. Notice that $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{T}$, and $\widehat{\boldsymbol{W}}$ is a function of $\boldsymbol{T}$. By Lemma S4, conditioned on $\boldsymbol{T}$, we have

$$\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\right\|_{\max}^* \leq C_2\tau\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\|_F\sqrt{n_1+n_2}, \tag{S8}$$

with conditional probability at least $1 - 2\exp\{-(n_1+n_2)\}$. Since the probability bound does not depend on $\boldsymbol{T}$, (S8) holds with the same probability bound unconditionally. By construction, $\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\|_F \leq \kappa$, we have the desired result. □

Proof of Theorem 2. It follows from the definition of $\widehat{\boldsymbol{A}}$ that for $\boldsymbol{A}_\star \in \mathbb{R}^{n_1\times n_2}$ with $\|\boldsymbol{A}_\star\|_{\max} \leq \beta$,

$$\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ 1/2}\circ\left(\widehat{\boldsymbol{A}} - \boldsymbol{Y}\right)\right\|_F^2 \leq \frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ 1/2}\circ(\boldsymbol{A}_\star - \boldsymbol{Y})\right\|_F^2 + \mu(\|\boldsymbol{A}_\star\|_* - \|\widehat{\boldsymbol{A}}\|_*). \tag{S9}$$

Since we can rewrite the first term in the left hand side of (S9) as

$$\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ 1/2}\circ\left(\widehat{\boldsymbol{A}} - \boldsymbol{Y}\right)\right\|_F^2 = \frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ 1/2}\circ\left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star + \boldsymbol{A}_\star - \boldsymbol{Y}\right)\right\|_F^2,$$

the inequality (S9) leads to

$$\begin{aligned}\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ 1/2}\circ\left(\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star\right)\right\|_F^2 \leq& \frac{2}{n_1n_2}\left\langle\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ 1/2}\circ\left(\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star\right),\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ 1/2}\circ(\boldsymbol{Y}-\boldsymbol{A}_\star)\right\rangle+\mu(\|\boldsymbol{A}_\star\|_*-\|\widehat{\boldsymbol{A}}\|_*)\\ =& \frac{2}{n_1n_2}\left\langle\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star,\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\right\rangle+\mu(\|\boldsymbol{A}_\star\|_*-\|\widehat{\boldsymbol{A}}\|_*).\end{aligned}$$

Therefore, due to Theorem 1, Lemma S5 and condition of μ, with the property that $\|\boldsymbol{A}_\star\|_*\leq\sqrt{n_1n_2}\|\boldsymbol{A}_\star\|_{\max}$, we have

$$\begin{aligned}\frac{1}{n_1n_2}\left\|\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star\right\|_F^2 &\leq \frac{1}{n_1n_2}\left\langle\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star,\left(\boldsymbol{T}\circ\widehat{\boldsymbol{W}}-\boldsymbol{J}\right)\circ\left(\boldsymbol{A}_\star-\widehat{\boldsymbol{A}}\right)\right\rangle+\frac{1}{n_1n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ(1/2)}\circ(\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star)\|_F^2\\ &\leq S(\widehat{\boldsymbol{W}},\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star)+\left|\frac{2}{n_1n_2}\langle\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star,\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\rangle\right|+\mu(\|\boldsymbol{A}_\star\|_*-\|\widehat{\boldsymbol{A}}\|_*)\\ &\leq S(\widehat{\boldsymbol{W}},\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star)+\frac{2}{n_1n_2}\left\|\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star\right\|_{\max}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\right\|_{\max}^*+\mu\|\boldsymbol{A}_\star\|_*\\ &\leq C_1(\beta^2+\beta)\min\left\{\frac{\log^{1/2}(n_1+n_2)}{\sqrt{\pi_L(n_1\wedge n_2)}},\frac{\sqrt{n_1+n_2}}{\pi_L\sqrt{n_1n_2}}\right\}+\frac{4C_2\beta\tau\kappa\sqrt{n_1+n_2}}{n_1n_2} \qquad \text{(S10)}\\ &\leq C_1(\beta^2)\min\left\{\frac{\log^{1/2}(n_1+n_2)}{\sqrt{\pi_L(n_1\wedge n_2)}},\frac{\sqrt{n_1+n_2}}{\pi_L\sqrt{n_1n_2}}\right\}+\frac{4C_2\beta\tau\kappa\sqrt{n_1+n_2}}{n_1n_2}. \qquad \text{(S11)}\end{aligned}$$

with probability at least $1-\exp\{-2^{-1}(\log 2)\pi_L^2\sum_{i,j}\pi_{ij}^{-1}\}-2\exp\{-(n_1+n_2)\}-1/(n_1+n_2)$. □

Proof of Theorem 3. Without loss of generality, we assume that $n_1\geq n_2$. For some constant $0\leq\gamma\leq 1$ such that $B=\sigma^{-2}(\sigma\wedge\beta)^2/(\gamma^2)$ is an integer and $B\leq n_2$, define

$$\mathcal{C}_1=\left\{\tilde{\boldsymbol{A}}=(A_{ij})\in\mathbb{R}^{n_1\times B}:A_{ij}\in\{0,\gamma\beta\},\forall 1\leq i\leq n_1,1\leq j\leq B\right\},$$

and consider the associated set of block matrices

$$\mathcal{A}(\mathcal{C}_1)=\left\{\boldsymbol{A}=\left(\widetilde{\boldsymbol{A}}|\dots|\widetilde{\boldsymbol{A}}|\boldsymbol{0}\right)\in\mathbb{R}^{n_1\times n_2}:\widetilde{\boldsymbol{A}}\in\mathcal{C}_1\right\},$$

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

It is easy to see that for any $\boldsymbol{A}\in\mathcal{A}(\mathcal{C}_1)$, we have that $\|\boldsymbol{A}\|_{\max}\leq\sqrt{B}\|\boldsymbol{A}\|_\infty\leq\beta$. Due to Lemma 2.9 in Tsybakov (2009), there exists a subset $\mathcal{A}^0\subset\mathcal{A}(\mathcal{C}_1)$ containing the zero $n_1\times n_2$ matrix $\mathbf{0}$ where $\text{Card}(\mathcal{A}^0)\geq 2^{Bn_1/8}+1$ and for any two distinct elements $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ of $\mathcal{A}^0$,

$$\|\boldsymbol{A}_1-\boldsymbol{A}_2\|_F^2\geq\frac{n_1B}{8}\left\{\gamma^2\beta^2\left\lfloor\frac{n_2}{B}\right\rfloor\right\}\geq\frac{n_1n_2\gamma^2\beta^2}{16}. \qquad \text{(S12)}$$

For any $\boldsymbol{A}\in\mathcal{A}^0$, from the noisy observed model in section 2.2, the probability distribution $\mathbb{P}_{\boldsymbol{A}}=\Pi_{i,j}[(2\pi\sigma^2)^{-1/2}\exp\{-(Y_{ij}-A_{ij})^2/(2\sigma^2)\}]^{T_{ij}}$. Take $\mathbb{P}_{\mathbf{0}}=\Pi_{i,j}[(2\pi\sigma^2)^{-1/2}\exp\{-Y_{ij}^2/(2\sigma^2)\}]^{T_{ij}}$. Thus the Kullback-Leibler divergence $K(\mathbb{P}_{\mathbf{0}},\mathbb{P}_{\boldsymbol{A}})=\mathsf{E}_{\mathbb{P}_{\mathbf{0}}}(\log(\mathbb{P}_{\mathbf{0}}/\mathbb{P}_{\boldsymbol{A}}))$ between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_{\boldsymbol{A}}$ satisfies

$$K\left(\mathbb{P}_{\mathbf{0}},\mathbb{P}_{\boldsymbol{A}}\right)=\mathsf{E}_{\mathbb{P}_{\mathbf{0}}}\left(\sum_{ij}T_{ij}\frac{A_{ij}^2-2A_{ij}Y_{ij}}{2\sigma^2}\right)=\frac{\left\|\boldsymbol{\Pi}^{\circ 1/2}\circ\boldsymbol{A}\right\|_F^2}{2\sigma^2}\leq\frac{\gamma^2\beta^2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\pi_{ij}}{2\sigma^2}\leq C_5\frac{\gamma^2\beta^2n_1n_2\pi_L}{2\sigma^2},$$

for some positive constant C_5. The last inequality is due to the condition that $n_1n_2\pi_L\asymp\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\pi_{ij}$.

From above we deduce the condition

$$\frac{1}{\text{Card}(\mathcal{A}^0)-1}\sum_{\boldsymbol{A}\in\mathcal{A}^0}K\left(\mathbb{P}_{\mathbf{0}},\mathbb{P}_{\boldsymbol{A}}\right)\leq\lambda\log\left(\text{Card}(\mathcal{A}^0)-1\right), \qquad \text{(S13)}$$

The above condition is valid when we take

$$\gamma^2 = C_6 \left(\frac{(\sigma \wedge \beta)^2}{\beta^2 n_2 \pi_L} \right)^{1/2}$$

for some constant C_6 that depends on λ. Also, one can verify that under the conditions $\pi_L^{-1} = \mathcal{O}(\beta^2 (n_1 \wedge n_2)/(\sigma \wedge \beta)^2)$ and $\pi_L^{1/2} = \mathcal{O}((n_1 \wedge n_2)^{1/2} \sigma^2 / [\beta(\sigma \wedge \beta)])$, $\gamma \le 1$ and $B \le n_2$. Then we subsitute γ^2 in the bound of S12 and we achieve the final bound as the one showed in the Theorem.

Together with the similar argument when $n_2 \ge n_1$, the result now follows by application of Theorem 2.5 in Tsybakov (2009). This completes the proof. □

Lemma S6. *Suppose Assumption 2 hold. For a fixed matrix $\boldsymbol{B} = (B_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ where $B_{ij} \ge 0$, there exists an absolute constant $C_5 > 0$ such that, with probability at least $1 - 2\exp(-(n_1 + n_2))$,*

$$\|\boldsymbol{B} \circ \boldsymbol{\epsilon}\| \le C_5 \|\boldsymbol{B}\|_\infty \tau (\sqrt{n_1} + \sqrt{n_2}).$$

Proof. By the definition of $\| \cdot \|_{\psi_2}$,

$$\max_{i,j} \|B_{ij} \epsilon_{ij}\|_{\psi_2} \le \|\boldsymbol{B}\|_\infty \tau.$$

Apply Theorem S2 in Section D , and take $t = \sqrt{n_1} + \sqrt{n_2}$ in Theorem S2. Then conclusion follows.

□

C. Non-asymptotic Error Bound under Low-rank Settings and Asymptotically Homogeneous Missingness

Theorem S1. *Suppose Assumption 2 hold and $\pi_L \asymp \pi_U \asymp \pi$. Assume $\|\boldsymbol{A}_\star\|_{\max} \le \beta$ and $\boldsymbol{A}_\star$ has rank R. If $\kappa' = \kappa - \|\boldsymbol{T}\|_F$ is bounded and $\mu \asymp \sqrt{\{\tau^2 \pi \log(n_1 + n_2)\}\{(n_1 \wedge n_2)\, n_1 n_2\}^{-1}}$, then there exists a constant $C_6 > 0$, such that with probability at least $1 - 3(n_1 + n_2)^{-1}$,*

$$d^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star) \le \frac{C_6 R (\tau^2 \vee \|A_\star\|_\infty^2) \log(n_1 + n_2)}{[\pi(n_1 \wedge n_2)]^{-1}}.$$

Proof outline. From the basic inequality, we have

$$\frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{1/2} \circ (\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star)\|_F^2 \le \frac{2}{n_1 n_2} \langle \widehat{\boldsymbol{A}} - \boldsymbol{A}_\star, \boldsymbol{T} \circ \widehat{\boldsymbol{W}} \circ \boldsymbol{\epsilon} \rangle + \mu \|\boldsymbol{A}_\star\|_* - \mu \|\widehat{\boldsymbol{A}}\|_*.$$

Note that weights are restricted to be greater than 1. We then have

$$\frac{1}{n_1 n_2} \|\boldsymbol{T} \circ (\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star)\|_F^2 \le \frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{1/2} \circ (\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star)\|_F^2 \le \frac{2}{n_1 n_2} \|\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star\|_* \|\boldsymbol{T} \circ \widehat{\boldsymbol{W}} \circ \boldsymbol{\epsilon}\| + \mu(\|\boldsymbol{A}_\star\|_* - \|\widehat{\boldsymbol{A}}\|_*).$$

Due to the constraint of κ', $\|\widehat{\boldsymbol{W}}\|_\infty$ is bounded, we can use Lemma S7 to derive the bound of $\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}} \circ \boldsymbol{\epsilon}\|$. The remaining argument is rather standard and the same as the proof for No-weighted estimators with nuclear norm regularization (Klopp, 2014). □

Lemma S7. *Suppose assumptions in Corollary S1 hold. Then there exists a constant $C_7 > 0$ such that, with probability at least $1 - (n_1 + n_2)$,*

$$\frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \widehat{\boldsymbol{W}} \circ \boldsymbol{\epsilon}\| \le C_7 \sqrt{\frac{\tau^2 \pi \log(n_1 + n_2)}{(n_1 \wedge n_2)\, n_1 n_2}}$$

Proof. The proof is very similar as the proof in Lemma S1.

We consider applying the Matrix Berinstein inequality for random matrices with bounded sub-exponential norm.

Due to the constraint of κ', there exists a constant C_8 such that $\|\hat{\boldsymbol{W}}\|_\infty \le C_8$.

For $(n_1 n_2)^{-1}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\| = \|\sum_{i,j}(T_{ij}\hat{W}_{i,j}\epsilon_{ij})\boldsymbol{J}_{ij}/(n_1 n_2)\|$, where $\boldsymbol{J}_{ij}$ has 1 for $(i,j)-$th, but 0 for all the remaining entries, let $\boldsymbol{M}_{i,j} = (T_{ij}\hat{W}_{i,j}\epsilon_{ij})\boldsymbol{J}_{ij}$, then $(n_1 n_2)^{-1}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\| = \|(n_1 n_2)^{-1}\sum_{i,j}\boldsymbol{M}_{i,j}\|$. We can easily verify that $\mathsf{E}(\boldsymbol{M}_{i,j}) = \mathbf{0}$. Note that ϵ_{ij} are sub-Gaussian random variables and therefore sub-exponential random variables. Then $\max_{i,j}\|\|\boldsymbol{M}_{i,j}\|\|_{\psi_1} \leq \max_{i,j}\|T_{ij}\hat{W}_{i,j}\epsilon_{ij}\|_{\psi_1} \leq C_9\tau$, where $\|\cdot\|_{\psi_1}$ is the sub-exponential norm of a random variable and C_9 is some constant depending on the C_8.

Since $\mathsf{E}(T_{ij}\hat{W}_{i,j}\epsilon_{ij})^2 \leq cC_8^2\pi_{ij}\tau^2$ for some absolute constant c, we can show that

$$\begin{aligned}
&\left\|\frac{1}{n_1 n_2}\sum_{i,j}\mathsf{E}\left(\boldsymbol{M}_{i,j}\boldsymbol{M}_{i,j}^{\intercal}\right)\right\| = \left\|\frac{1}{n_1 n_2}\sum_{i,j}\mathsf{E}\left(\boldsymbol{M}_{i,j}^{\intercal}\boldsymbol{M}_{i,j}\right)\right\| \\
\leq&\frac{1}{n_1 n_2}\max\left\{\max_{1\leq i\leq n_1}\sum_{j=1}^{n_2}c_2\pi_{ij}\tau^2, \max_{1\leq j\leq n_2}\sum_{i=1}^{n_1}c_2\pi_{ij}\tau^2\right\} \\
\leq&\frac{c_3\tau^2}{n_1\wedge n_2}\pi,
\end{aligned}$$

for some constant c_3.

By Proposition 11 in Klopp (2014), there exsits a constant C_7, such that with probability at least $1-1/(n_1+n_2)$,

$$\frac{1}{n_1 n_2}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\right\| \leq C_7\max\left\{\sqrt{\frac{\tau^2\pi\log\left(n_1+n_2\right)}{\left(n_1\wedge n_2\right)n_1 n_2}}, \tau\log(1/\sqrt{\pi})\frac{\log^{3/2}(n_1+n_2)}{n_1 n_2}\right\}.$$

Overall, the conclusion follows. □

D. Useful Results

Theorem S2 (Theorem 4.4.5 of Vershynin (2018))**.** *Let $\boldsymbol{A}$ be an $n_1\times n_2$ random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables. Then, for any $t>0$ we have*

$$\|\boldsymbol{A}\| \leq CK(\sqrt{n_1}+\sqrt{n_2}+t)$$

with probability at least $1-2\exp(-t^2)$. Here $K=\max_{ij}\|A_{ij}\|_{\psi_2}$ and C is an absolute constant.

Proof. The proof can be found on Page 91 in Vershynin (2018). □

Theorem S3 (Proposition 1 of Koltchinskii et al. (2011))**.** *Let $\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_N$ be independent random matrices with dimensions $n_1\times n_2$ that satisfy $\boldsymbol{E}\boldsymbol{Z}_i=0$ and $\|\boldsymbol{Z}_i\|\leq U$ almost surely for some constant U and all $i=1,\ldots,n$. Define*

$$\sigma_Z = \max\left\{\left\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{E}(\boldsymbol{Z}_i\boldsymbol{Z}_i^{\intercal})\right\|^{1/2}, \left\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{E}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{Z}_i)\right\|^{1/2}\right\}.$$

Then, for all $t>0$, with probability at least $1-\exp(-t)$ we have

$$\left\|\frac{\boldsymbol{Z}_1+\cdots+\boldsymbol{Z}_N}{N}\right\| \leq 2\max\left\{\sigma_Z\sqrt{\frac{t+\log(n_1+n_2)}{N}}, U\frac{t+\log(n_1+n_2)}{N}\right\},$$

Proof. The proof can be found on Page 2325 in Koltchinskii et al. (2011). □

E. Algorithm

E.1. Convex Algorithm for Solving (7)

Follow Fang et al. (2018) and Cai & Zhou (2016), we consider an equivalent form objective function in (7) below.

$$\begin{aligned}&\min_{\boldsymbol{X},\boldsymbol{Z}} \frac{1}{n_1 n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ(1/2)}\circ(\boldsymbol{Y}-\boldsymbol{Z}_{12})\|_F^2+\mu\langle\boldsymbol{I},\boldsymbol{X}\rangle,\\ &\text{Subject to } \boldsymbol{X}\succcurlyeq 0,\ \boldsymbol{X}=\boldsymbol{Z},\ \boldsymbol{Z}\in\mathcal{P}_\beta\end{aligned}$$

where $\boldsymbol{Z},\boldsymbol{X}\in\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$, $\mathcal{S}$ is the class of all symmetric matrices in $\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$, $\mathcal{P}_\beta := \{\boldsymbol{C}\in\mathcal{S} : \operatorname{diag}(\boldsymbol{C})\geq 0, \|\boldsymbol{C}\|_\infty\leq\beta\}$, $\boldsymbol{I}$ is an identity matrix and

$$\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{Z}_{11} & \boldsymbol{Z}_{12}\\ \boldsymbol{Z}_{12}^\intercal & \boldsymbol{Z}_{22}\end{bmatrix}, \boldsymbol{Z}_{11}\in\mathbb{R}^{n_1\times n_1}, \boldsymbol{Z}_{22}\in\mathbb{R}^{n_2\times n_2}$$

The derivation of above representation mainly comes from two facts: 1. The nuclear norm of $\boldsymbol{Z}_{12}$ is the the smallest possible sum of elements on the diagonal of $\boldsymbol{Z}$ given $\boldsymbol{Z}\succcurlyeq 0$ (Fazel et al., 2001); 2. The max norm of matrix $\boldsymbol{Z}_{12}$ is the smallest possible maximum element on the diagonal of $\boldsymbol{Z}$ given $\boldsymbol{Z}\succcurlyeq 0$ (Srebro et al., 2005).

The augmented Lagrangian function can be written as

$$\begin{aligned}&\mathcal{L}(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{V})=\frac{1}{n_1 n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ(1/2)}\circ(\boldsymbol{Y}-\boldsymbol{Z}_{12})\|_F^2+\mu\langle\boldsymbol{I},\boldsymbol{X}\rangle+\langle\boldsymbol{V},\boldsymbol{X}-\boldsymbol{Z}\rangle+\frac{\rho}{2}\|\boldsymbol{X}-\boldsymbol{Z}\|_F^2,\\ &\text{Subject } \ \boldsymbol{X}\succcurlyeq 0,\ \boldsymbol{Z}\in\mathcal{P}_\beta,\end{aligned}$$

where $\boldsymbol{V}\in\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ is the dual variable and $\rho>0$ is a hyper-parameter.

Then the alternating direction method of multipliers (ADMM) algorithm solves this optimization problem by minimizing the augmented Lagrangian with respect to different variables alternatingly. More explicitly, at the $(t+1)$-th iteration, the following updates are implemented:

$$\begin{aligned}&\boldsymbol{X}^{t+1}=\Pi\{\boldsymbol{Z}^t+\rho^{-1}(\boldsymbol{V}^t+\mu\boldsymbol{I})\},\\ &\boldsymbol{Z}^{(t+1)}=\arg\min_{\boldsymbol{Z}\in\mathcal{P}_\beta}\frac{1}{n_1 n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ(1/2)}\circ(\boldsymbol{Y}-\boldsymbol{Z}_{12})\|_F^2+\frac{\rho}{2}\|\boldsymbol{Z}-\boldsymbol{X}^{t+1}-\rho^{-1}\boldsymbol{V}^t\|_F^2=\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}\{\boldsymbol{X}^{t+1}+\rho^{-1}\boldsymbol{V}^t\},\\ &\boldsymbol{V}^{t+1}=\boldsymbol{V}^t+\tau\rho(\boldsymbol{X}^{t+1}-\boldsymbol{Z}^{t+1}),\end{aligned}$$

where $\Pi(\cdot)$ is the projection to the space $\{\boldsymbol{C}\in\mathcal{S}:\boldsymbol{C}\succcurlyeq 0\}$, and $\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}$ is defined in Definition S1. Detailed derivation can be found in Fang et al. (2018) and Cai & Zhou (2016).

Definition S1. *We use $\boldsymbol{C}(i,j)$ to represent the element on the i-th row and j-th column of a matrix $\boldsymbol{C}$. For the matrix $\boldsymbol{C}\in\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$, it can be partitioned into*

$$\boldsymbol{C}=\begin{bmatrix}\boldsymbol{C}_{11} & \boldsymbol{C}_{12}\\ \boldsymbol{C}_{12}^\intercal & \boldsymbol{C}_{22}\end{bmatrix}, \boldsymbol{C}_{11}\in\mathbb{R}^{n_1\times n_1}, \boldsymbol{C}_{22}\in\mathbb{R}^{n_2\times n_2}$$

Then

$$\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})=\begin{bmatrix}\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{11} & \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{12}\\ \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{12}^\intercal & \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{22}\end{bmatrix},$$

where

$$\begin{aligned}
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{11}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{11}(i,j), -\beta\}\} \qquad \text{if } i \neq j,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{11}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{11}(i,j), 0\}\} \qquad \text{if } i = j,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{22}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{22}(i,j), -\beta\}\}, \qquad \text{if } i \neq j,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{22}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{22}(i,j), 0\}\} \qquad \text{if } i = j,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{12}(i,j) = \min\left\{\beta, \max\left\{\frac{\boldsymbol{Y}(i,j)\widehat{\boldsymbol{W}}(i,j) + \rho\boldsymbol{C}(i,j)}{\widehat{\boldsymbol{W}}(i,j) + \rho}, -\beta\right\}\right\} \qquad \text{if } \boldsymbol{T}(i,j) = 1,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{12}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{12}(i,j), -\beta\}\} \text{ if } i \neq j \qquad \text{if } \boldsymbol{T}(i,j) = 0.
\end{aligned}$$

We summarize the algorithm in Algorithm 1. Some piratical implementations to adaptively tune ρ and accelerate the computation can be found in Section 3.3 and 3.4 in Fang et al. (2018).

Algorithm 1 ADMM algorithm

```
Input: $\boldsymbol{Y}$, $\boldsymbol{T}$, $\beta$, $\mu$, $\widehat{\boldsymbol{W}}$, $\rho = 0.1$, $\tau = 1.618$, K
Initialize $\boldsymbol{X}^0$, $\boldsymbol{Z}^0$, $\boldsymbol{V}^0$, $R$
for $t = 1$ to $K - 1$ do
  $\boldsymbol{X}^{t+1} \leftarrow \Pi\{\boldsymbol{Z}^t + \rho^{-1}(\boldsymbol{V}^t + \mu\boldsymbol{I})\}$
  $\boldsymbol{Z}^{(t+1)} \leftarrow \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}\{\boldsymbol{X}^{t+1} + \rho^{-1}\boldsymbol{V}^t\}$
  $\boldsymbol{V}^{t+1} \leftarrow \boldsymbol{V}^t + \tau\rho(\boldsymbol{X}^{t+1} - \boldsymbol{Z}^{t+1})$
  Stop if objective value changes less than tolerance
end for
```

E.2. Nonconvex Algorithm for Solving (7)

The nonconvex algorithm for max-norm regularization developed in Lee et al. (2010) base on the equivalent definition of max-norm via matrix factorizations:

$$\|\boldsymbol{C}\|_{\max} := \inf\left\{\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty} : \boldsymbol{C} = \boldsymbol{U}\boldsymbol{V}^{\intercal}\right\},$$

where $\|\cdot\|_{2,\infty}$ denotes the maximum l_2 row norm of a matrix.

To incorporate the nuclear norm regularization, we also notice an equivalent definition of the nuclear norm:

$$\|\boldsymbol{C}\|_* := \inf \frac{1}{2}\left\{\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2 : \boldsymbol{C} = \boldsymbol{U}\boldsymbol{V}^{\intercal}\right\}.$$

Then we have the following relaxation of the objective function in (7). Take

$$f(\boldsymbol{L}, \boldsymbol{R}) = \frac{1}{n_1 n_2}\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ(1/2)} \circ (\boldsymbol{Y} - \boldsymbol{L}\boldsymbol{R}^{\intercal})\|_F^2 + \frac{\mu}{2}(\|\boldsymbol{L}\|_F^2 + \|\boldsymbol{R}\|_F^2),$$

and we obtain

$$\begin{aligned}
&\min_{\boldsymbol{L},\boldsymbol{R}} f(\boldsymbol{L}, \boldsymbol{R}),\\
&\text{Subject to } \max\{\|\boldsymbol{L}\|_{2,\infty}, \|\boldsymbol{R}\|_{2,\infty}\} \leq \beta.
\end{aligned}$$

This optimization form is exactly the one in Lee et al. (2010) except that we add another nuclear penalty in the objective function f.

Like what Lee et al. (2010) considered, the projected gradient descent method can be applied to iteratively solve this problem. We define the project $\mathcal{P}_B$ as the Euclidean projection onto the set $\{\boldsymbol{M} : \|\boldsymbol{M}\|_{2,\infty} \leq B\}$. This projection can be computed by re-scaling the rows of current input matrix whose norms exceed B so their norms equal B. Rows with norms less than B are unchanged by the projection. We summarize the algorithm in Algorithm 2.

Algorithm 2 Projected gradient descent algorithm

```
Input: $\boldsymbol{Y}$, $\boldsymbol{T}$, $\beta$, $\mu$, $\widehat{\boldsymbol{W}}$, step size $\tau$, K
Initialize $\boldsymbol{L}^0$, $\boldsymbol{R}^0$,
for $t = 1$ to $K - 1$ do
    $\boldsymbol{L}^{t+1} \leftarrow \mathcal{P}_\beta \left( \boldsymbol{L} - \tau \frac{\partial f}{\partial \boldsymbol{L}} \right)$
    $\boldsymbol{R}^{t+1} \leftarrow \mathcal{P}_\beta \left( \boldsymbol{R} - \tau \frac{\partial f}{\partial \boldsymbol{R}} \right)$
    Stop if objective value changes less than tolerance
end for
```

F. Additional Simulation Results

The simulation results for SNR = 1 and SNR = 10 are shown in Table S1 and S2 respectively.

Table S1. Similar to Table 1, but for SNR = 1.

	Setting 1		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**1.901(0.004)**	**1.918(0.004)**	13.69(0.097)
`SoftImpute`	1.944(0.004)	1.961(0.004)	19.55(0.092
`CZ`	2.052(0.004)	2.044(0.004)	27.695(0.128)
`FLT`	**1.927(0.004)**	**1.946(0.004)**	15.265(0.105)
`NW`	2.012(0.004)	2.01(0.004)	25.61(0.069)
`KLT`	2.439(0.005)	2.492(0.005)	10.175(0.063)
	Setting 2		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**1.716(0.004)**	**1.669(0.004)**	14.73(0.113)
`SoftImpute`	1.721(0.004)	1.685(0.004)	16.335(0.107)
`CZ`	1.86(0.004)	1.799(0.004)	25.965(0.115)
`FLT`	**1.711(0.004)**	**1.674(0.004)**	14.565(0.102)
`NW`	1.805(0.005)	1.747(0.005)	37.82(0.422)
`KLT`	2.16(0.005)	2.093(0.005)	2.065(0.110)
	Setting 3		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**2.412(0.006)**	**2.586(0.007)**	12.495(0.098)
`SoftImpute`	2.923(0.007)	3.113(0.007)	29.15(0.112)
`CZ`	2.641(0.006)	2.812(0.006)	28.695(0.109)
`FLT`	2.878(0.007)	3.097(0.007)	20.105(0.105)
`NW`	**2.668(0.006)**	**2.779(0.007)**	33.115(0.066)
`KLT`	3.667(0.007)	3.969(0.007)	9.765(0.067)

G. Code

The code for implementing the proposed method can be found in `https://github.com/jiayiwang1017/MC-weighting-code`.

References

Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pp. 1556–1564, 2014.

Cai, T. T. and Zhou, W.-X. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10 (1):1493–1525, 2016.

Fang, E. X., Liu, H., Toh, K.-C., and Zhou, W.-X. Max-norm optimization for robust matrix recovery. *Mathematical Programming*, 167(1):5–35, 2018.

Table S2. Similar to Table 1, but for SNR = 10.

	Setting 1		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
Proposed	**0.425(0.001)**	**0.443(0.001)**	30.565(0.143)
SoftImpute	0.483(0.001)	0.501(0.001)	91.615(1.845)
CZ	0.663(0.001)	0.688(0.001)	54.225(0.178)
FLT	0.427(0.001)	0.446(0.001)	32.105(0.129)
NW	**0.405(0.001)**	**0.422(0.001)**	30.180(0.196)
KLT	1.894(0.003)	1.958(0.003)	8.665(0.059)
	Setting 2		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
Proposed	**0.401(0.001)**	**0.418(0.001)**	29.360(0.136)
SoftImpute	0.475(0.001)	0.484(0.001)	120.670(3.387)
CZ	0.684(0.002)	0.722(0.002)	59.300(0.485)
FLT	**0.409(0.001)**	**0.426(0.001)**	30.380(0.145)
NW	0.496(0.003)	0.510(0.003)	19.055(0.368)
KLT	1.975(0.006)	1.873(0.004)	1.375(0.144)
	Setting 3		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
Proposed	**0.627(0.001)**	**0.688(0.002)**	32.675(0.144)
SoftImpute	0.868(0.002)	0.966(0.003)	77.115(1.343)
CZ	0.999(0.003)	1.105(0.004)	56.890(0.152)
FLT	**0.670(0.002)**	**0.736(0.002)**	39.740(0.820)
NW	0.703(0.003)	0.768(0.003)	21.860(0.614)
KLT	3.157(0.006)	3.460(0.006)	9.535(0.088)

Fazel, M., Hindi, H., and Boyd, S. P. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 6, pp. 4734–4739. IEEE, 2001.

Golub, G. H. and Van Loan, C. F. Matrix computations. johns hopkins studies in the mathematical sciences, 1996.

Jalali, A. and Srebro, N. Clustering using max-norm constrained optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 481–488, 2012.

Klopp, O. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Kontorovich, A. Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pp. 28–36, 2014.

Lee, J. D., Recht, B., Srebro, N., Tropp, J., and Salakhutdinov, R. R. Practical large-scale optimization for max-norm regularization. In *Advances in neural information processing systems*, pp. 1297–1305, 2010.

Lee, T., Shraibman, A., and Špalek, R. A direct product theorem for discrepancy. In *2008 23rd Annual IEEE Conference on Computational Complexity*, pp. 71–80. IEEE, 2008.

Srebro, N. and Shraibman, A. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pp. 545–560. Springer, 2005.

Srebro, N., Rennie, J., and Jaakkola, T. S. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pp. 1329–1336, 2005.

Talagrand, M. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.

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

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.