## SUPPLEMENTARY MATERIAL

# A Imputation error by data type and timing results

In this section we provide more details on the simulations of Section 5.2. Table 6 presents the imputation errors of the compared methods for quantitative variables only, and Table 7 for binary variables. For the quantitative variables, mimi and MLFAMD, which both model main group effects, perform best. As already noticed in Section 5.2, mimi has smaller imputation errors than other methods when the size of the main effects compared to the interactions, and the proportion of missing entries, are both large. For the binary variables, suprisingly, softImpute outperforms consistently the other methods, although it is not designed for mixed data. Finally, Table 8 shows the average computational times of the different compared methods. We observe that the computational times of mimi, GLRM, FAMD and MLFAMD are of comparable order. The aforementioned methods are an order of magnitude slower than softImpute and mice.

| **% missing** | **20** | | | **40** | | | **60** | | |
|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | **0.2** | **1** | **5** | **0.2** | **1** | **5** | **0.2** | **1** | **5** |
| **mean** | 20.7(1.3) | 19.8(0.7) | 19.6(0.6) | 28.0(2.6) | 28.2(1.3) | 26.9(1.1) | 35.5(1.6) | 34.2(1.3) | 34.1(0.5) |
| **mimi** | 13.0(0.4) | **12.3**(0.4) | **11.4**(0.3) | 19.8(1.1) | **19.0**(0.7) | **16.1**(0.5) | 27.1(1.0) | **24.3**(1.1) | **20.2**(0.4) |
| **GLRM** | 16.1(1.0) | 16.9(0.7) | 13.8(0.4) | 24.0(5.3) | 24.5(1.5) | 23.4(1.1) | 36.5(12.3) | 41.9(18.0) | 44.1(3.7) |
| **softImpute** | 14.0(0.5) | 14.0(0.4) | 13.3(0.4) | 20.3(1.2) | 20.9(0.7) | 18.5(0.8) | 27.3(1.2) | 27.4(1.0) | 24.4(0.5) |
| **FAMD** | 12.7(0.5) | 12.9(0.6) | 12.1(0.3) | **19.2**(1.3) | 20.2(0.6) | 17.3(0.6) | 26.9(1.8) | 31.2(1.0) | 22.7(0.4) |
| **MLFAMD** | **12.6**(0.6) | 13.7(0.6) | 12.2(0.4) | 18.8(1.0) | 19.7(0.6) | 17.6(0.7) | **25.4**(1.5) | 26.2(1.2) | 23.5(0.6) |
| **mice** | 17.3(0.8) | 17.2(1.0) | 16.9(0.6) | 25.1(1.2) | 26.0(0.7) | 23.1(1.0) | 40.7(2.8) | 40.1(0.9) | 36.8(1.8) |

Table 1: Quantitative variables: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for different percentages of missing entries (20%, 40%, 60%) and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|L^0\|_F$ (0.2, 1, 5). The values are averaged across 100 replications and the standard deviation is given between parenthesis.

# B Proof of Theorem 1

To prove global convergence of the BCGD algorithm, we use a result from (Tseng and Yun, 2009, Theorem 1) summarized below in Theorem 5, combined with the compacity of the level sets of the objective $F$, proved using Lemma 3 and Lemma 4.

| % missing | 20 | | | 40 | | | 60 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | **0.2** | **1** | **5** | **0.2** | **1** | **5** | **0.2** | **1** | **5** |
| **mean** | 13.0(0.3) | 12.4(0.3) | **11.8**(0.4) | 18.33(0.4) | 17.4(0.3) | 16.9(0.3) | 22.6(0.5) | 22.0(0.6) | **20.8**(0.6) |
| **mimi** | 13.5(0.3) | 13.5(0.3) | 13.5(0.3) | 18.9(0.5) | 19.1(0.3) | 18.9(0.6) | 23.7(0.6) | 23.4(0.5) | 23.1(0.4) |
| **GLRM** | 14.2(0.4) | 14.1(0.6) | 14.2(0.5) | 20.0(0.4) | 20.2(0.4) | 20.4(0.3) | 24.9(0.5) | 25.1(0.6) | 24.9(0.3) |
| **softImpute** | **12.2**(0.1) | **12.0**(0.3) | 12.0(0.6) | **17.0**(0.3) | **16.7**(0.2) | **16.6**(0.4) | **21.6**(0.4) | **21.6**(0.3) | 21.0(0.5) |
| **FAMD** | 13.6(0.4) | 13.8(0.4) | 13.5(0.3) | 19.2(0.5) | 19.8(0.3) | 18.8(0.6) | 24.0(0.5) | 25.0(0.4) | 23.6(0.4) |
| **MLFAMD** | 13.6(0.5) | 13.5(0.4) | 13.6(0.4) | 19.4(0.5) | 19.5(0.4) | 19.6(0.5) | 24.0(0.5) | 24.1(0.4) | 23.9(0.4) |
| **mice** | 14.6(0.3) | 14.5(0.4) | 14.4(0.4) | 20.5(0.4) | 20.3(0.2) | 20.5(0.4) | 25.7(0.4) | 25.7(0.6) | 25.3(0.2) |

Table 2: Binary variables: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for different percentages of missing entries (20%, 40%, 60%) and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|L^0\|_F$ (0.2, 1, 5). The values are averaged across 100 replications and the standard deviation is given between parenthesis.

| **method** | **mean** | **mimi** | **GLRM** | **softImpute** | **FAMD** | **MLFAMD** | **mice** |
|---|---|---|---|---|---|---|---|
| **time (s)** | 1.7e-4 | 6.6 | 5.5 | 0.1 | 2.6 | 3.5 | 0.2 |

Table 3: Computation time of the seven compared methods (averaged across 100 simulations).

**Theorem 1.** *Let* $\{(\alpha^{[k]}, L^{[k]})\}$ *be the current iterates,* $\{(d_\alpha^{[k]}, d_L^{[k]})\}$ *the descent directions and* $\{(\Gamma_\alpha^{[k]}, \Gamma_L^{[k]})\}$ *the functionals generated by the BCGD algorithm. Then the following results hold.*

*(a)* $\{F(\alpha^{[k]}, L^{[k]})\}$ *is nonincreasing and for all* $k$, $(\Gamma_\alpha^{[k]}, \Gamma_L^{[k]})$ *satisfies*

$$-\Gamma_\alpha^{[k]} \geq (1-\theta)\nu\|d_\alpha^{[k]}\|_2^2 \text{ and } -\Gamma_L^{[k]} \geq (1-\theta)\nu\|d_L^{[k]}\|_F^2.$$

*(b) Every cluster point of* $\{(\alpha^{[k]}, L^{[k]})\}$ *is a stationary point of* $F$.

Assumptions **H**1 and 2, combined with the separability of the $\ell_1$ and nuclear norm penalties, guarantee that the conditions of (Tseng and Yun, 2009, Theorem 1) are satisfied. We now show that the data-fitting term $\mathcal{L}(\mathsf{f}_U(\alpha) + L; Y, \Omega)$ is lower-bounded.

**Lemma 1.** *There exists a constant* $c > -\infty$ *such that, for all* $X \in \mathbb{R}^{m_1 \times m_2}$, $\mathcal{L}(X; Y, \Omega) \geq c$.

*Proof.* Recall that $\mathcal{L}(X; Y, \Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \Omega\{-Y_{ij}X_{ij} + g_j(X_{ij})\}$. Thus, we only need to prove that for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, the function $x \mapsto -Y_{ij}x + g_j(x)$ is lower bounded by a constant $c_{ij} > -\infty$. Assume that this is not the case; by the convexity of $x \mapsto -Y_{ij}x + g_j(x)$ we have that either $-Y_{ij}x + g_j(x) \underset{x\to+\infty}{\to} -\infty$ or $-Y_{ij}x + g_j(x) \underset{x\to-\infty}{\to} -\infty$. Assume without

loss of generality that $-Y_{ij}x + g_j(x) \underset{x\to+\infty}{\to} -\infty$. Then, there exists $x_0 \in \mathbb{R}$ such that for all $x \geq x_0$, $-Y_{ij}x + g_j(x) < \log \int_{\substack{y\in\mathcal{Y}_j \\ y\geq Y_{ij}}} h_j(y)\mu_j(d_y)$. Thus, for all $x \geq \max(x_0, 0)$, we have that

$$\begin{aligned}\int_{y\in\mathcal{Y}_j} h_j(y)\mathrm{e}^{yx-g_j(x)}\mu_j(d_y) &= \int_{\substack{y\in\mathcal{Y}_j \\ y<Y_{ij}}} h_j(y)\mathrm{e}^{yx-g_j(x)}\mu_j(d_y) + \int_{\substack{y\in\mathcal{Y}_j \\ y\geq Y_{ij}}} h_j(y)\mathrm{e}^{yx-g_j(x)}\mu_j(d_y) \\ &> \int_{\substack{y\in\mathcal{Y}_j \\ y<Y_{ij}}} h_j(y)\mathrm{e}^{yx-g_j(x)}\mu_j(d_y) + 1 > 1,\end{aligned}$$

contradicting normality of the density $h_j(y)\mathrm{e}^{yx-g_j(x)}$. Thus, there exists $c_{ij} > -\infty$, such that for all $x \in \mathbb{R}$, $-Y_{ij}x + g_j(x) \geq c_{ij}$. Finally we obtain that $\mathcal{L}(X; Y, \Omega) \geq c = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} c_{ij}$. □

Finally, we use Lemma 3 to show the compactness of the level sets of the objective function $F$, defined for $C \in \mathbb{R}$ by

$$L_C = \{(\alpha, L) \in \mathbb{R}^N \times \mathbb{R}^{m_1\times m_2}; F(\alpha, L) \leq C\}.$$

**Lemma 2.** *The level sets of the objective function $F$ are compact.*

*Proof.* For all $(\alpha, L) \in \mathbb{R}^N \times \mathbb{R}^{m_1\times m_2}$, $F(\alpha, L) \geq c + \lambda_1\|L\|_* + \lambda_2\|\alpha\|_1$, where $c$ is the constant defined in Lemma 3. Thus, for all $C \in \mathbb{R}$, the level set $L_C$ is included in the compact set

$$\left\{(\alpha, L) \in \mathbb{R}^N \times \mathbb{R}^{m_1\times m_2}; \|L\|_* \leq \frac{C-c}{2\lambda_1} \text{ and } \|\alpha\|_1 \leq \frac{C-c}{2\lambda_2}\right\}.$$

Furthermore, by the continuity of $F$, the level set $L_C$ is also a closed set. Thus we obtain that for all $C \in \mathbb{R}$, the level set $L_C$ is compact. □

We can now combine Theorem 5, Lemma 3 and Lemma 4 to prove Theorem 1. Let $(\alpha^{[0]}, L^{[0]})$ be an initialization point. Theorem 5 (a) implies that the sequence $(\alpha^{[k]}, L^{[k]})$ generated by the BCGD algorithm lies in the level set of $F$

$$L_{F(\alpha^{[0]},L^{[0]})} = \left\{(\alpha, L) \in \mathbb{R}^N \times \mathbb{R}^{m_1\times m_2}; F(\alpha, L) \leq F(\alpha^{[0]}, L^{[0]})\right\}.$$

Furthermore, $L_{F(\alpha^{[0]},L^{[0]})}$ is compact by Lemma 4, showing that the sequence $(\alpha^{[k]}, L^{[k]})$ has at least one accumulation point. Combined with Theorem 5 (b) and the convexity of $F$,

this shows Theorem 1 (a).

Theorem 5 (a) and Lemma 3 combined imply that the sequence $\{F(\alpha^{[k]}, L^{[k]})\}$ converges to a limit $F^*$. Furthermore, Theorem 1 (a) and the continuity of $F$ imply that there exists a sub-sequence $\{F(\alpha^{[k]}, L^{[k]})\}_{k\in\mathcal{K}}$ such that $\{F(\alpha^{[k]}, L^{[k]})\}_{k\in\mathcal{K}} \to F(\hat{\alpha}, \hat{L})$. Thus, $F^* = F(\hat{\alpha}, \hat{L})$, which proves Theorem 1 (b).

# C Proof of Theorem 2

Let $\Pi = (\pi_{ij})_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}$ be the distribution of the mask $\Omega$. For $B \in \mathbb{R}^{m_1\times m_2}$ we denote $B_\Omega$ the projection of $B$ on the set of observed entries. We define $\|B\|_\Omega^2 = \|B_\Omega\|_F^2$, and $\|B\|_\Pi^2 = \mathbb{E}\left[\|B\|_\Omega^2\right]$, where the expectation is taken with respect to $\Pi$. The proof of Theorem 2 will follow the subsequent two steps. We first derive an upper bound on the Frobenius error restricted to the observed entries $\|\Delta X\|_\Omega^2$, then show that the expected Frobenius error $\|\Delta X\|_\Pi^2$ is upper bounded by $\|\Delta X\|_\Omega^2$ with high probability, and up to a residual term defined later on.

Let us derive the upper bound on $\|\Delta X\|_\Omega^2$. By definition of $\hat{L}$ and $\hat{\alpha}$: $\mathcal{L}(\hat{X}; Y, \Omega) - \mathcal{L}(X^0; Y, \Omega) \le \lambda_1 \left(\|L^0\|_* - \|\hat{L}\|_*\right) + \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right)$. Recall that, for $\alpha \in \mathbb{R}^N$, we use the notation $\mathsf{f}_U(\alpha) = \sum_{k=1}^N \alpha_k U^k$. Adding $\langle \nabla\mathcal{L}(X^0; Y, \Omega), \Delta X\rangle$ on both sides of the last inequality, we get

$$\begin{aligned}\mathcal{L}(\hat{X}; Y, \Omega) - \mathcal{L}(X^0; Y, \Omega) + \langle \nabla\mathcal{L}(X^0; Y, \Omega), \Delta X\rangle \le \\ \lambda_1 \left(\|L^0\|_* - \|\hat{L}\|_*\right) - \langle \nabla\mathcal{L}(X^0; Y, \Omega), \Delta L\rangle \\ + \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right) - \langle \nabla\mathcal{L}(X^0; Y, \Omega), \mathsf{f}_U(\Delta\alpha)\rangle. \quad (1)\end{aligned}$$

Assumption **H**2 implies that for any pair of matrices $X^1$ and $X^2$ in $\mathbb{R}^{m_1\times m_2}$ satisfying $\|X^1\|_\infty \vee \|X^2\|_\infty \le (1 + æ)a$, the two following inequalities hold for all $\Omega$:

$$\mathcal{L}(X; Y, \Omega) - \mathcal{L}(\tilde{X}; Y, \Omega) - \langle \nabla\mathcal{L}(\tilde{X}; Y, \Omega), X - \tilde{X}\rangle \ge \frac{\sigma_-^2}{2}\|X - \tilde{X}\|_\Omega^2, \quad (2)$$

$$\|\nabla\mathcal{L}(X; Y, \Omega) - \nabla\mathcal{L}(\tilde{X}; Y, \Omega)\|_F \le \sigma_+^2\|X - \tilde{X}\|_\Omega. \quad (3)$$

Plugging (33) into (32) allows to construct a lower bound on the left hand side term and

obtain $\sigma_-^2\|\Delta X\|_\Omega^2/2 \leq \mathsf{A}_1 + \mathsf{A}_2$,

$$\begin{aligned} \mathsf{A}_1 &= \lambda_1 \left( \|L^0\|_* - \|\hat{L}\|_* \right) + \left| \langle \nabla \mathcal{L}(X^0; Y, \Omega), \Delta L \rangle \right|, \\ \mathsf{A}_2 &= \lambda_2 \left( \|\alpha^0\|_1 - \|\hat{\alpha}\|_1 \right) + \left| \langle \nabla \mathcal{L}(X^0; Y, \Omega), \mathsf{f}_U(\Delta\alpha) \rangle \right|. \end{aligned} \tag{4}$$

Let us upper bound $\mathsf{A}_1$. The duality of the norms $\|\cdot\|_*$ and $\|\cdot\|$ implies that

$$\left| \langle \nabla \mathcal{L}(X^0; Y, \Omega), \Delta L \rangle \right| \leq \|\nabla \mathcal{L}(X^0; Y, \Omega)\| \|\Delta L\|_*.$$

Denote by $S_1$ and $S_2$ the linear subspaces spanned respectively by the left and right singular vectors of $L^0$, and $P_{S_1^\perp}$ and $P_{S_2^\perp}$ the orthogonal projectors on the orthogonal of $S_1$ and $S_2$, $P_{L^{0\perp}} : X \mapsto P_{S_1^\perp} X P_{S_2^\perp}$ and $P_{L^0} : X \mapsto X - P_{S_1^\perp} X P_{S_2^\perp}$. The triangular inequality yields

$$\|\hat{L}\|_* = \|L^0 - P_{L^{0\perp}}(\Delta L) - P_{L^0}(\Delta L)\|_* \geq \|L^0 + P_{L^{0\perp}}(\Delta L)\|_* - \|P_{L^0}(\Delta L)\|_*. \tag{5}$$

Moreover, by definition of $P_{L^{0\perp}}$, the left and right singular vectors of $P_{L^{0\perp}}(\Delta L)$ are respectively orthogonal to the left and right singular spaces of $L^0$, implying $\|L^0 + P_{L^{0\perp}}(\Delta L)\|_* = \|L^0\|_* + \|P_{L^{0\perp}}(\Delta L)\|_*$. Plugging this identity into (36) we obtain

$$\|L^0\|_* - \|\hat{L}\|_* \leq \|P_{L^0}(\Delta L)\|_* - \|P_{L^{0\perp}}(\Delta L)\|_*, \tag{6}$$

and $\mathsf{A}_1 \leq \lambda_1 \left( \|P_{L^0}(\Delta L)\|_* - \|P_{L^{0\perp}}(\Delta L)\|_* \right) + \|\nabla \mathcal{L}(X^0; Y, \Omega)\| \|\Delta L\|_*$.

Using $\|\Delta L\|_* \leq \|P_{L^0}(\Delta L)\|_* + \|P_{L^{0\perp}}(\Delta L)\|_*$ and the assumption $\lambda_1 \geq 2\|\nabla \mathcal{L}(X^0; Y, \Omega)\|$ we get $\mathsf{A}_1 \leq 3\lambda_1 \|P_{L^0}(\Delta L)\|_*/2$. In addition, $\|P_{L^0}(\Delta L)\|_* \leq \sqrt{\operatorname{rank}\left(P_{L^0}(\Delta L)\right)} \|P_{L^0}(\Delta L)\|_F$, and $\operatorname{rank}\left(P_{L^0}(\Delta L)\right) \leq 2\operatorname{rank}\left(L^0\right)$ (see, *e.g.* (Klopp, 2014, Theorem 3)). Together with $\|P_{L^0}(\Delta L)\|_F \leq \|\Delta L\|_F$, this finally implies the following upper bound:

$$\mathsf{A}_1 \leq \frac{3\lambda_1}{2} \sqrt{2r} \|\Delta L\|_F. \tag{7}$$

We now derive an upper bound for $\mathsf{A}_2$. The duality between $\|\cdot\|_1$ and $\|\cdot\|_\infty$ ensures

$$\left| \langle \nabla \mathcal{L}(X^0; Y, \Omega), \mathsf{f}_U(\Delta\alpha) \rangle \right| \leq \|\Delta\alpha\|_1 \|\nabla \mathcal{L}(X^0; Y, \Omega)\|_\infty u. \tag{8}$$

The assumption $\lambda_2 \geq 2\|\nabla \mathcal{L}(X^0; Y, \Omega)\|_\infty u$ in conjunction with (39) and the triangular inequality $\|\Delta\alpha\|_1 \leq \|\alpha^0\|_1 + \|\hat{\alpha}\|_1$ yield

$$\mathsf{A}_2 \leq \frac{3\lambda_2}{2} \|\alpha^0\|_1. \tag{9}$$

Combining inequalities (35), (38) and (40) we obtain

$$\|\Delta X\|_\Omega^2 \leq \frac{3\lambda_1}{\sigma_-^2}\sqrt{2r}\|\Delta L\|_F + \frac{3\lambda_2}{\sigma_-^2}\|\alpha^0\|_1. \tag{10}$$

We now show that when the errors $\Delta L$ and $\Delta\alpha$ belong to a subspace $\mathcal{C}$ and for a residual $\mathsf{D}$ - both defined later on - the following holds with high probability:

$$\|\Delta X\|_\Omega^2 \geq \|\Delta X\|_\Pi^2 - \mathsf{D}. \tag{11}$$

We start by defining our constrained set and prove that it contains the errors $\Delta L$ and $\Delta\alpha$ with high probability (Lemma 5-6); then we show that restricted strong convexity holds on this subspace (Lemma 7). For non-negative constants $d_1$, $d_\Pi$, $\rho < m$ and $\varepsilon$ that will be specified later on, define the two following sets where $\Delta\alpha$ and $\Delta L$ should lie:

$$\mathcal{A}(d_1, d_\Pi) = \left\{\alpha \in \mathbb{R}^N : \|\alpha\|_1 \leq d_1,\ \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq d_\Pi\right\}. \tag{12}$$

$$\begin{aligned}\mathcal{L}(\rho,\varepsilon) \quad = \Bigg\{ & L \in \mathbb{R}^{m_1\times m_2}, \alpha \in \mathbb{R}^N : \|L + \mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{72\log(d)}{p\log(6/5)}, \\ & \|L + \mathsf{f}_U(\alpha)\|_\infty \leq 1, \|L\|_* \leq \sqrt{\rho}\|L\|_F + \varepsilon \Bigg\}\end{aligned} \tag{13}$$

If $\|\Delta X\|_\Pi^2$ is too small, the right hand side of (42) is negative. The first inequality in the definition of $\mathcal{L}(\rho,\varepsilon)$ prevents from this. Condition $\|L\|_* \leq \sqrt{\rho}\|L\|_F + \varepsilon$ is a relaxed form of the condition $\|L\|_* \leq \sqrt{\rho}\|L\|_F$ satisfied for matrices of rank $\rho$. Finally, we define the constrained set of interest:

$$\mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) = \mathcal{L}(\rho,\varepsilon) \cap \left\{\mathbb{R}^{m_1\times m_2} \times \mathcal{A}(d_1, d_\Pi)\right\}.$$

Recall $u = \max_k \|U_k\|_1$ and let

$$d_1 = 4\|\alpha^0\|_1, \text{ and } d_\Pi = \frac{3\lambda_2}{\sigma_-^2}\|\alpha^0\|_1 + 64a^2u\mathbb{E}\left[\|\Sigma_R\|_\infty\right]\|\alpha^0\|_1 + 3072a^2p^{-1} + \frac{72a^2\log(d)}{\log(6/5)}.$$

**Lemma 3.** *Let $\lambda_2 \geq 2u\left(\|\nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty + 2\sigma_+^2(1+u)a\right)$ and assume **H**1-2 hold. Then, with probability at least $1 - 8d^{-1}$, $\Delta\alpha \in \mathcal{A}(d_1, d_\Pi)$.*

*Proof.* See Appendix E. □

Lemma 5 implies the upper bound on $\|\Delta\alpha\|_2^2$ of Theorem 2. Thus, we only need to prove the upper bound on $\|\Delta L\|_F^2$. Let $\rho = 32r$ and $\varepsilon = 3\lambda_2/\lambda_1\|\alpha^0\|_1$.

**Lemma 4.** *Assume **H**2 and let*

$$\lambda_1 \geq 2\|\nabla\mathcal{L}(X^0; Y, \Omega)\|, \quad \lambda_2 \geq 2u\left(\|\nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty + 2\sigma_+^2(1+u)a\right).$$

*Then* $\|\Delta L\|_* \leq \sqrt{\rho}\|\Delta L\|_F + \varepsilon$.

*Proof.* See Appendix F □

As a consequence, under the conditions on the regularization parameters $\lambda_1$ and $\lambda_2$ given in Lemma 6 and whenever $\|\Delta L + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq 72\log(d)/(p\log(6/5))$, the error terms $(\Delta L, \Delta\alpha)$ belong to the constrained set $\mathcal{C}(d_1, d_\Pi, \rho, \varepsilon)$ with high probability.

**Case 1:** Suppose $\|\Delta L + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 < 72\log(d)/(p\log(6/5))$. Then, Lemma 5 combined with the fact that $\|M\|_F^2 \leq p^{-1}\|M\|_\Pi^2$ for all $M$, and the identity $(a+b)^2 \geq a^2/4 - 4b^2$ ensures that $\|\Delta L\|_F^2 \leq 4\|\Delta L + \mathsf{f}_U(\Delta\alpha)\|_F^2 + 16\|\mathsf{f}_U(\Delta\alpha)\|_F^2$. Therefore we obtain (ii) of Theorem 2:

$$\|\Delta L\|_F^2 \leq \frac{288a^2\log(d)}{\log(6/5)} + 16\frac{\|\alpha^0\|_1}{p}\Theta_1.$$

**Case 2:** Suppose $\|\Delta L + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq 72\log(d)/(p\log(6/5))$. Then, Lemma 5 and 6 yield that with probability at least $1 - 8d^{-1}$,

$$\left(\frac{\Delta L}{2(1+æ)a}, \frac{\Delta\alpha}{2(1+æ)a}\right) \in \mathcal{C}(d_1', d_\Pi', \rho', \varepsilon'), \text{where}$$

$$d_1' = \frac{d_1}{2(1+æ)a}, \quad d_\Pi' = \frac{d_\Pi}{4(1+æ)^2a^2}, \quad \rho' = \rho, \quad \varepsilon' = \frac{\varepsilon}{2(1+æ)a},$$

and where $d_1, d_\Pi, \rho$ and $\varepsilon$ are the same as in Lemma 5 and 6. We use the following result, proven in Appendix G. Recall that we assume for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\mathbb{P}(\Omega_{ij} = 1) \geq p$ and define:

$$\tilde{\mathcal{A}}(d_1) = \left\{\alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{18\log(d)}{p\log(6/5)}\right\},$$

$$\begin{aligned} \mathsf{D}_\alpha &= 8æd_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1}, \\ \mathsf{D}_X &= \frac{112\rho}{p}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8æ\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8æd_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}. \end{aligned} \tag{14}$$

**Lemma 5.** *(i) For any $\alpha \in \tilde{\mathcal{A}}(d_1)$, with probability at least $1 - 8d^{-1}$,*

$$\|\mathsf{f}_U(\alpha)\|_\Omega^2 \geq \frac{1}{2}\|\mathsf{f}_U(\alpha)\|_\Pi^2 - \mathsf{D}_\alpha.$$

*(ii) For any pair $(L, \alpha) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon)$, with probability at least $1 - 8d^{-1}$*

$$\|L + \mathsf{f}_U(\alpha)\|_\Omega^2 \geq \frac{1}{2}\|L + \mathsf{f}_U(\alpha)\|_\Pi^2 - \mathsf{D}_X. \tag{15}$$

*Proof.* See Appendix G. □

Lemma 7 (ii) applied to $\left(\frac{\Delta L}{2(1+æ)a}, \frac{\Delta\alpha}{2(1+æ)a}\right)$ implies that with probability at least $1-8d^{-1}$, $\|\Delta X\|_\Pi^2 \leq 2\|\Delta X\|_\Omega^2 + 4(1+æ)a\mathsf{D}_X$. Combined with (41), $\|\Delta X\|_F^2 \leq p^{-1}\|\Delta X\|_\Pi^2$, $\|\Delta X\|_F^2 \geq \|\Delta L\|_F^2/2 - \|\mathsf{f}_U(\Delta\alpha)\|_F^2$ and $6\sqrt{2r}\lambda_1/(p\sigma_-^2)\|\Delta L\|_F \leq \|\Delta L\|_F^2/4 + 288r\lambda_1^2/(p^2\sigma_-^4)$, we obtain the result of Theorem 2 (ii):

$$\|\Delta L\|_F^2 \leq \frac{1152r\lambda_1^2}{p^2\sigma_-^4} + \frac{24\lambda_2\|\alpha^0\|_1}{p\sigma_-^2} + 4(1+æ)a\mathsf{D}_X + 4\frac{\|\alpha^0\|}{p}\Theta_1.$$

# D Proof of Theorem 4

We will establish separately two lower bounds of order $rM/p$ and $s/p$ respectively. Define

$$\tilde{\mathcal{L}} = \left\{\tilde{L} \in \mathbb{R}^{m_1 \times r} : \tilde{L}_{ij} \in \left\{0, \eta\min(a, \sigma_+)\left(\frac{r}{pm}\right)^{1/2}\right\}, \forall (i,j) \in [\![m_1]\!] \times [\![r]\!]\right\},$$

where $0 \leq \eta \leq 1$ will be chosen later. Define also the associated set of block matrices

$$\mathcal{L} = \left\{L = (\tilde{L}|\ldots|\tilde{L}|O) \in \mathbb{R}^{m_1 \times m_2} : \tilde{L} \in \tilde{\mathcal{L}}\right\},$$

where $O$ denotes the $m_1 \times (m_2 - r\lfloor m_2/r \rfloor)$ null matrix and, for some $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the integer part of $x$. We also define the following set of vectors

$$\mathcal{A} = \left\{\alpha = (\tilde{O}|\tilde{\alpha}) \in \mathbb{R}^N, \tilde{\alpha}_k \in \{0, \tilde{\eta}\min(a, \sigma_+)\}\ \forall 1 \leq k \leq s\right\},$$

with $\tilde{O} \in \mathbb{R}^{m_2 - s}$ denoting the null vector. Finally, we set

$$\mathcal{X} = \left\{X = L + \mathsf{f}_U(\alpha) \in \mathbb{R}^{m_1 \times m_2}, \alpha \in \mathcal{A}, L \in \mathcal{L}\right\}.$$

For any $X \in \mathcal{X}$ there exists a matrix $L \in \mathcal{L}$ of rank at most $r$ and a vector $\alpha$ with at most $s$ non-zero components satisfying $X = L + \mathsf{f}_U(\alpha)$. Furthermore, for any $\tilde{X} \in \mathcal{X}$ there exists a matrix $\tilde{L} \in \mathcal{L}$ of rank at most $r$ and a vector $\tilde{\alpha}$ with at most $s$ non-zero components satisfying $X - \tilde{X} = \tilde{L} + \mathsf{f}_U(\tilde{\alpha})$. Finally, for all $X \in \mathcal{X}$ and $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $0 \leq X_{ij} \leq (1 + æ)a$. Thus, $\mathcal{X} \subset \mathcal{F}(r,s)$, where $\mathcal{F}(r,s)$ is defined in (29).

**Lower bound of order $rM/p$.** Consider the set

$$\mathcal{X}_L = \{X = L + \mathsf{f}_U(\alpha) \in \mathcal{X}; \alpha = 0\}.$$

Lemma 2.9 in Tsybakov (2008) (Varshamov Gilbert bound) implies that there exists a subset $\mathcal{X}_L^0 \subset \mathcal{X}_L$ satisfying $\mathrm{Card}(\mathcal{X}_L^0) \geq 2^{rM/8} + 1$, such that the zero $m_1 \times m_2$ matrix $\mathbf{0} \in \mathcal{X}_L^0$, and that for any two $X$ and $X'$ in $\mathcal{X}_L^0$, $X \neq X'$ we have

$$\|X - X'\|_F^2 \geq \frac{Mr}{8}\left(\eta^2 \min(a, \sigma_+)^2 \frac{r}{pm}\left\lfloor \frac{m_2}{r} \right\rfloor\right) \geq \frac{\eta^2}{16}\min(a^2, \sigma_+^2)\frac{rM}{p}. \tag{16}$$

For $X \in \mathcal{X}_L^0$ we compute the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X)$ between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_X$. Using Assumption **H**2 we obtain

$$\mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X) = \sum_{i,j} \pi_{ij}\left(g_j(X_{ij}) - g_j(0) - g_j'(0)X_{ij}\right) \leq \frac{\sigma_+^2 \eta^2 \min(a, \sigma_+)^2 Mr}{2}. \tag{17}$$

Inequality (48) implies that

$$\frac{1}{\mathrm{Card}(\mathcal{X}_L^0) - 1} \sum_{X \in \mathcal{X}_L^0} \mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X) \leq \frac{1}{16}\log(\mathrm{Card}(\mathcal{X}_L^0) - 1) \tag{18}$$

is satisfied for $\tilde{\eta} = \min\left\{1, (8\sigma_+ \min(a, \sigma_+))^{-1}\right\}$. Then, conditions (47) and (48) guarantee that we can apply Theorem 2.5 from Tsybakov (2008). We obtain that for some constant $\delta > 0$ and with $\Psi_1 = C \min\left(\sigma_+^{-2}, \min(a, \sigma_+^2)\right)$:

$$\inf_{\hat{L}, \hat{\alpha}} \sup_{(L^0, \alpha^0) \in \mathcal{E}} \mathbb{P}_{X^0}\left(\|\Delta L\|_F^2 + \|\Delta\alpha\|_2^2 > \frac{\Psi_1 rM}{p}\right) \geq \delta, \tag{19}$$

**Lower bound of order $s/p$.** Using again the Varshamov-Gilbert bound (Tsybakov (2008), Lemma 2.9) we obtain that there exists a subset $\mathcal{A}^0 \in \mathcal{A}$ satisfying $\mathrm{Card}(\mathcal{A}^0) \geq 2^{s/8}+1$ and containing the null vector $\mathbf{0} \in \mathbb{R}^N$ and such that, for any $\alpha$ and $\alpha'$ of $\mathcal{A}^0$, $\alpha \neq \alpha'$,

$$\|\alpha - \alpha'\|_2^2 \geq \frac{s}{8}\tilde{\eta}^2 \min(a, \sigma_+)^2. \tag{20}$$

Define $\mathcal{X}_\alpha \subset \mathcal{X}$ the set of matrices $X = \mathsf{f}_U(\alpha)$ such that $\alpha \in \mathcal{A}^0$ and $L = 0$. For any $X \in \mathcal{X}_\alpha$ we compute the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X)$ between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_X$

$$\mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X) = \sum_{i,j} \pi_{ij}(g_j(X_{ij}) - g_j(0) - g_j'(0)X_{ij} \leq \sigma_+^2 \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq \sigma_+^2 p \|\mathsf{f}_U(\alpha)\|_F^2. \tag{21}$$

Using Assumption **H**2

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) &\leq \sigma_+^2 p \left(\max_k \|U^k\|_F^2 + 2\tau\right) \|\alpha\|_2^2 \\ &\leq s\sigma_+^2 p \left(\max_k \|U^k\|_F^2 + 2\tau\right) \tilde{\eta}^2 \min(a, \sigma_+)^2. \end{aligned} \tag{22}$$

From (53) we deduce that

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0) - 1} \sum_{\mathcal{A}^0} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) \leq sp \left(\max_k \|U^k\|_F^2 + 2\tau\right) \sigma_+^2 \tilde{\eta}^2 \min(a, \sigma_+)^2. \tag{23}$$

Choosing $\tilde{\eta} = \min\left\{1, \left(\sqrt{p}\sigma_+ \max_k(\|U^k\|_F + 2\tau) \min(a, \sigma_+)\right)^{-1}\right\}$, we now use Tsybakov (2008), Theorem 2.5 which implies for some constant $\delta > 0$

$$\inf_{\hat{L},\hat{\alpha}} \sup_{(L^0,\alpha^0)\in\mathcal{E}} \mathbb{P}_{X^0}\left\{\|\Delta L\|_F^2 + \|\sum_{k=1}^N (\alpha_k^0 - \hat{\alpha}_k)U^k\|_F^2 > \Psi_2 \frac{s\kappa^2}{p}\right\} \geq \delta, \tag{24}$$

$$\Psi_2 = C\left(\frac{1}{\sigma_+^2\left(\max_k \|U^k\|_F^2 + 2\tau\right)} \wedge (a \wedge \sigma_+)^2\right),$$

where we have used that $\|\sum_{k=1}^N (\alpha_k^0 - \hat{\alpha}_k)U^k\|_F^2 \geq \kappa^2 \|\hat{\alpha} - \alpha^0\|_2^2$. We finally obtain the result by combining (50) and (55).

# E Proof of Lemma 5

We start by proving $\|\Delta\alpha\|_1 \leq 4\|\alpha^0\|_1$. By the optimality conditions over a convex set (Aubin and Ekeland, 1984, Chapter 4, Section 2, Proposition 4), there exist two subgradients $\hat{f}_L$ in the subdifferential of $\|\cdot\|_*$ taken at $\hat{L}$ and $\hat{f}_\alpha$ in the subdifferential of $\|\cdot\|_1$ taken at $\hat{\alpha}$, such that for all feasible pairs $(L, \alpha)$ we have

$$\langle \nabla\mathcal{L}(\hat{X}; Y, \Omega), L - \hat{L} + \sum_{k=1}^{N}(\alpha_k - \hat{\alpha}_k)U^k\rangle + \lambda_1\langle \hat{f}_L, L - \hat{L}\rangle + \lambda_2\langle \hat{f}_\alpha, \alpha - \hat{\alpha}\rangle \geq 0. \tag{25}$$

Applying inequality (56) to the pair $(\hat{L}, \alpha^0)$ we obtain $\langle \nabla\mathcal{L}(\hat{X}; Y, \Omega), \sum_{k=1}^{N}\Delta\alpha_k U^k\rangle + \lambda_2\langle \hat{f}_\alpha, \Delta\alpha\rangle \geq 0$. Denote $\tilde{X} = \hat{L} + \sum_{k=1}^{N}\alpha_k^0 U^k$. The last inequality is equivalent to

$$\underbrace{\langle \nabla\mathcal{L}(X^0; Y, \Omega), \sum_{k=1}^{N}\Delta\alpha_k U^k\rangle}_{\mathsf{B}_1} + \underbrace{\langle \nabla\mathcal{L}(\tilde{X}; Y, \Omega) - \nabla\mathcal{L}(X^0; Y, \Omega), \sum_{k=1}^{N}\Delta\alpha_k U^k\rangle}_{\mathsf{B}_2}$$
$$+ \underbrace{\langle \nabla\mathcal{L}(\hat{X}; Y, \Omega) - \nabla\mathcal{L}(\tilde{X}; Y, \Omega), \sum_{k=1}^{N}\Delta\alpha_k U^k\rangle}_{\mathsf{B}_3} + \lambda_2\langle \hat{f}_\alpha, \Delta\alpha\rangle \geq 0.$$

We now derive upper bounds on the three terms $\mathsf{B}_1$, $\mathsf{B}_2$ and $\mathsf{B}_3$ separately. Recall that we denote $u = \max_k \|U^k\|_1$ and use (39) to bound $\mathsf{B}_1$:

$$\mathsf{B}_1 \leq \|\Delta\alpha\|_1\|\nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty u. \tag{26}$$

The duality between $\|\cdot\|_\infty$ and $\|\cdot\|_1$ gives $\mathsf{B}_2 \leq \|\Delta\alpha\|_1\|\nabla\mathcal{L}(\tilde{X}; Y, \Omega) - \nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty u$. Moreover, $\nabla\mathcal{L}(\tilde{X}; Y, \Omega) - \nabla\mathcal{L}(X^0; Y, \Omega)$ is a matrix with entries $g_j'(\tilde{X}_{ij}) - g_j'(X^0_{ij})$, therefore assumption **H**2 ensures $\|\nabla\mathcal{L}(\tilde{X}; Y, \Omega) - \nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty \leq 2\sigma_+^2(1 + æ)a$, and finally we obtain

$$\mathsf{B}_2 \leq \|\Delta\alpha\|_1 2\sigma_+^2(1 + æ)au. \tag{27}$$

We finally bound $\mathsf{B}_3$ as follows. We have that $\mathsf{B}_3 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\Omega_{ij}(g_j'(\hat{X}_{ij}) - g_j'(\tilde{X}_{ij}))(\tilde{X}_{ij} - \hat{X}_{ij})$. Now, for all $j \in [\![m_2]\!]$, $g_j'$ is increasing therefore $(g_j'(\hat{X}_{ij}) - g_j'(\tilde{X}_{ij}))(\tilde{X}_{ij} - \hat{X}_{ij}) \leq 0$, which implies $\mathsf{B}_3 \leq 0$. Combined with (57) and (58) this yields

$$\lambda_2\langle \hat{f}_\alpha, \hat{\alpha} - \alpha^0\rangle \leq \|\Delta\alpha\|_1 u\left(\|\nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty + 2\sigma_+^2(1 + æ)a\right).$$

Besides, the convexity of $\|\cdot\|_1$ gives $\langle \hat{f}_\alpha, \hat{\alpha} - \alpha^0 \rangle \geq \|\hat{\alpha}\|_1 - \|\alpha^0\|_1$, therefore

$$\left\{\lambda_2 - u\left(\|\nabla \mathcal{L}(X^0; Y, \Omega)\|_\infty + 2\sigma_+^2(1+æ)a\right)\right\}\|\hat{\alpha}\|_1 \leq \left\{\lambda_2 + u\left(\|\nabla \mathcal{L}(X^0; Y, \Omega)\|_\infty + 2\sigma_+^2(1+æ)a\right)\right\}\|\alpha^0\|_1,$$

and the condition $\lambda_2 \geq 2\left\{u\left(\|\nabla \mathcal{L}(X^0; Y, \Omega)\|_\infty + 2\sigma_+^2(1+æ)a\right)\right\}$ gives $\|\hat{\alpha}\|_1 \leq 3\|\alpha^0\|_1$ and finally

$$\|\Delta\alpha\|_1 \leq 4\|\alpha^0\|_1. \tag{28}$$

**Case 1:** $\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 < 72a^2\log(d)/(p\log(6/5))$. Then the result holds trivially.

**Case 2:** $\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq 72a^2\log(d)/(p\log(6/5))$. For $d_1 > 0$ recall the definition of the set

$$\tilde{\mathcal{A}}(d_1) = \left\{\alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{18\log(d)}{p\log(6/5)}\right\}.$$

Inequality (59) and $\|\Delta\alpha\|_\infty \leq 2a$ imply that $\Delta\alpha/(2a) \in \tilde{\mathcal{A}}(2\|\alpha^0\|_1/a)$. Therefore we can apply Lemma 7(i) and obtain that with probability at least $1 - 8d^{-1}$,

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \leq 2\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 + 64æa\|\alpha^0\|_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 3072a^2p^{-1}. \tag{29}$$

We now must upper bound the quantity $\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2$. Recall that $\tilde{X} = \sum_{k=1}^N \alpha_k^0 U^k + \hat{X}$. By definition, $\mathcal{L}(\hat{X}; Y, \Omega) + \lambda_1\|\hat{L}\|_* + \lambda_2\|\hat{\alpha}\|_1 \leq \mathcal{L}(\tilde{X}; Y, \Omega) + \lambda_1\|\hat{L}\|_* + \lambda_2\|\alpha^0\|_1$, i.e.

$$\mathcal{L}(\hat{X}; Y, \Omega) - \mathcal{L}(\tilde{X}; Y, \Omega) \leq \lambda_2\left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right).$$

Substracting $\langle \nabla\mathcal{L}(\tilde{X}; Y, \Omega), \hat{X} - \tilde{X}\rangle$ on both sides and using the restricted strong convexity ((33)), we obtain

$$\begin{aligned}\frac{\sigma_-^2}{2}\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 &\leq \lambda_2\left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right) + \langle \nabla\mathcal{L}(\tilde{X}; Y, \Omega), \mathsf{f}_U(\Delta\alpha)\rangle \\ &\leq \lambda_2\left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right) + \underbrace{\left|\langle \nabla\mathcal{L}(X^0; Y, \Omega), \mathsf{f}_U(\Delta\alpha)\rangle\right|}_{\mathsf{C}_1} \\ &\quad + \underbrace{\left|\langle \nabla\mathcal{L}(X^0; Y, \Omega) - \nabla\mathcal{L}(\tilde{X}; Y), \mathsf{f}_U(\Delta\alpha)\rangle\right|}_{\mathsf{C}_2}.\end{aligned} \tag{30}$$

The duality of $\|\cdot\|_1$ and $\|\cdot\|_\infty$ yields $\mathsf{C}_1 \leq \|\nabla\mathcal{L}(X^0;Y,\Omega)\|_\infty u\|\Delta\alpha\|_1$, and

$$\mathsf{C}_2 \leq \|\nabla\mathcal{L}(X^0;Y,\Omega) - \nabla\mathcal{L}(\tilde{X};Y,\Omega)\|_\infty u\|\Delta\alpha\|_1.$$

Furthermore, $\|\nabla\mathcal{L}(X^0;Y,\Omega) - \nabla\mathcal{L}(\tilde{X};Y,\Omega)\|_\infty \leq 2\sigma_+^2 a$, since for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ $|\tilde{X}_{ij} - X^0_{ij}| \leq 2a$ and $g_j''(\tilde{X}_{ij}) \leq \sigma_+^2$. The last three inequalities plugged in (61) give

$$\frac{\sigma_-^2}{2}\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 \quad \leq \lambda_2\left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right) + u\|\Delta\alpha\|_1\left\{\|\nabla\mathcal{L}(X^0;Y,\Omega)\|_\infty + 2\sigma_+^2 a\right\}.$$

The triangular inequality gives

$$\begin{aligned}\frac{\sigma_-^2}{2}\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 \quad &\leq \left\{u\left(\|\nabla\mathcal{L}(X^0;Y,\Omega)\|_\infty + 2\sigma_+^2 a\right) + \lambda_2\right\}\|\alpha^0\|_1 \\ &\quad + \left\{u\left(\|\nabla\mathcal{L}(X^0;Y,\Omega)\|_\infty + 2\sigma_+^2 a\right) - \lambda_2\right\}\|\hat{\alpha}\|_1.\end{aligned}$$

Then, the assumption $\lambda_2 \geq 2u\left(\|\nabla\mathcal{L}(X^0;Y,\Omega)\|_\infty + 2\sigma_+^2(1+æ)a\right)$ gives

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 \leq \frac{3\lambda_2}{\sigma_-^2}\|\alpha^0\|_1.$$

Plugged into (60), this last inequality implies that with probability at least $1 - 8d^{-1}$

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \leq \frac{3\lambda_2}{\sigma_-^2}\|\alpha^0\|_1 + 64æa\|\alpha^0\|_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 3072a^2p^{-1}. \tag{31}$$

Combining (59) and (62) gives the result.

# F Proof of Lemma 6

Using (56) for $L = L^0$ and $\alpha = \alpha^0$ we obtain

$$\langle\nabla\mathcal{L}(\hat{X};Y,\Omega), \Delta L + \sum_{k=1}^N(\Delta\alpha_k)U^k\rangle + \lambda_1\langle\hat{f}_L, \Delta L\rangle + \lambda_2\langle\hat{f}_\alpha, \Delta\alpha\rangle \geq 0.$$

Then, the convexity of $\|\cdot\|_*$ and $\|\cdot\|_1$ imply that $\|L^0\|_* \geq \|\hat{L}\|_* + \langle \partial\|\hat{L}\|_*, \Delta L\rangle$ and $\|\alpha^0\|_1 \geq \|\hat{\alpha}\|_* + \langle \partial\|\hat{\alpha}\|_1, \Delta\alpha\rangle$. The last three inequalities yield

$$\begin{aligned}\lambda_1\left(\|\hat{L}\|_* - \|L^0\|_*\right) + \lambda_2\left(\|\hat{\alpha}\|_1 - \|\alpha^0\|_1\right) &\leq \langle \nabla\mathcal{L}(\hat{X}; Y, \Omega), \Delta L\rangle \\ &\quad + \langle \nabla\mathcal{L}(\hat{X}; Y, \Omega), \sum_{k=1}^{N}(\Delta\alpha_k)U^k\rangle \\ &\leq \|\nabla\mathcal{L}(\hat{X}; Y, \Omega)\|\|\Delta L\|_* + u\|\nabla\mathcal{L}(\hat{X}; Y, \Omega)\|_\infty\|\Delta\alpha\|_1.\end{aligned}$$

Using (37) and the conditions

$$\lambda_1 \geq 2\|\nabla\mathcal{L}(X^0; Y, \Omega)\|, \quad \lambda_2 \geq 2u\left\{\|\nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty + 2\sigma_+^2(1+æ)a\right\},$$

we get

$$\begin{aligned}\lambda_1\left(\|P_{L^0}^{\perp}(\Delta L)\|_* - \|P_{L^0}(\Delta L)\|_*\right) + \lambda_2\left(\|\hat{\alpha}\|_1 - \|\alpha^0\|_1\right) \leq \\ \frac{\lambda_1}{2}\left(\|P_{L^0}^{\perp}(\Delta L)\|_* + \|P_{L^0}(\Delta L)\|_*\right) + \frac{\lambda_2}{2}\|\Delta\alpha\|_1,\end{aligned}$$

which implies $\|P_{L^0}^{\perp}(\Delta L)\|_* \leq 3\|P_{L^0}(\Delta L)\|_* + 3\lambda_2/\lambda_1\|\alpha^0\|_1$. Now, using

$$\|\Delta L\|_* \leq \|P_{L^0}^{\perp}(\Delta L)\|_* + \|P_{L^0}(\Delta L)\|_*, \quad \|P_{L^0}(\Delta L)\|_F \leq \|\Delta L\|_F$$

and $\text{rank}(P_{L^0}(\Delta L)) \leq 2r$, we get $\|\Delta L\|_* \leq \sqrt{32r}\|\Delta L\|_F + 3\lambda_2/\lambda_1\|\alpha^0\|_1$. This completes the proof of Lemma 6.

# G Proof of Lemma 7

**Proof of (i):** Recall $\mathsf{D}_\alpha = 8æd_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1}$ and

$$\tilde{\mathcal{A}}(d_1) = \left\{\alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{18\log(d)}{p\log(6/5)}\right\}.$$

We will show that the probability of the following event is small:

$$\mathcal{B} = \left\{\exists\alpha \in \tilde{\mathcal{A}}(d_1) \text{ such that } \left|\|\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathsf{f}_U(\alpha)\|_\Pi^2\right| > \frac{1}{2}\|\mathsf{f}_U(\alpha)\|_\Pi^2 + \mathsf{D}_\alpha\right\}.$$

Indeed, $\mathcal{B}$ contains the complement of the event we are interested in. We use a peeling argument to upper bound the probability of event $\mathcal{B}$. Let $\nu = 18\log(d)/(p\log(6/5))$ and $\eta = 6/5$. For $l \in \mathbb{N}$ set

$$\mathcal{S}_l = \left\{\alpha \in \tilde{\mathcal{A}}(d_1): \quad \eta^{l-1}\nu \leq \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq \eta^l \nu\right\}.$$

Under the event $\mathcal{B}$, there exists $l \geq 1$ and $\alpha \in \tilde{\mathcal{A}}(d_1) \cap S_l$ such that

$$\left|\|\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathsf{f}_U(\alpha)\|_\Pi^2\right| \quad > \frac{1}{2}\|\mathsf{f}_U(\alpha)\|_\Pi^2 + \mathsf{D}_\alpha > \frac{1}{2}\eta^{l-1}\nu + \mathsf{D}_\alpha = \frac{5}{12}\eta^l\nu + \mathsf{D}_\alpha. \tag{32}$$

For $T > \nu$, consider the set of vectors

$$\tilde{\mathcal{A}}(d_1, T) = \left\{\alpha \in \tilde{\mathcal{A}}(d_1) : \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq T\right\}$$

and the event

$$\mathcal{B}_l = \left\{\exists \alpha \in \tilde{\mathcal{A}}(d_1, \eta^l\nu) : \left|\|\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathsf{f}_U(\alpha)\|_\Pi^2\right| > \frac{5}{12}\eta^l\nu + \mathsf{D}_\alpha\right\}.$$

If $\mathcal{B}$ holds, then (63) implies that $\mathcal{B}_l$ holds for some $l \leq 1$. Therefore ,$\mathcal{B} \subset \cup_{l=1}^{+\infty}\mathcal{B}_l$, and it is enough to estimate the probability of the events $\mathcal{B}_l$ and then apply the union bound. Such an estimation is given in the following lemma, adapted from Lemma 10 in Klopp (2015).

**Lemma 6.** *Define* $Z_T = \sup_{\alpha \in \tilde{\mathcal{A}}(d_1,T)} \left|\|\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathsf{f}_U(\alpha)\|_\Pi^2\right|$. *Then,*

$$\mathbb{P}\left(Z_T \geq \mathsf{D}_\alpha + \frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/18}.$$

*Proof.* By definition,

$$Z_T = \sup_{\alpha \in \tilde{\mathcal{A}}(d_1,T)} \left|\sum_{(i,j)} \Omega_{ij}\mathsf{f}_U(\alpha)_{ij}^2 - \mathbb{E}\left[\sum_{(i,j)} \Omega_{ij}\mathsf{f}_U(\alpha)_{ij}^2\right]\right|.$$

We use the following Talagrand's concentration inequality, proven in Talagrand (1996) and Chatterjee (2015).

**Lemma 7.** *Assume* $f : [-1,1]^n \mapsto \mathbb{R}$ *is a convex Lipschitz function with Lipschitz constant* $L$. *Let* $\Xi_1, \ldots, \Xi_n$ *be independent random variables taking values in* $[-1,1]$. *Let* $Z := f(\Xi_1, \ldots, \Xi_n)$. *Then, for any* $t \geq 0$, $\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq 16L + t\right) \leq 4\mathrm{e}^{-t^2/2L^2}$.

We apply this result to the function

$$f(x_{11},\dots,x_{m_1m_2}) = \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}(x_{ij}-\pi_{ij})\mathsf{f}_U(\alpha)_{ij}^2\right|,$$

which is Lipschitz with Lipschitz constant $\sqrt{p^{-1}T}$. Indeed, for any $(x_{11},\dots,x_{m_1m_2}) \in \mathbb{R}^{m_1\times m_2}$ and $(z_{11},\dots,z_{m_1m_2}) \in \mathbb{R}^{m_1\times m_2}$:

$$\begin{aligned}
&|f(x_{11},\dots,x_{m_1m_2}) - f(z_{11},\dots,z_{m_1m_2})| \\
&= \left|\sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}(x_{ij}-\pi_{ij})\mathsf{f}_U(\alpha)_{ij}^2\right| - \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}(z_{ij}-\pi_{ij})\mathsf{f}_U(\alpha)_{ij}^2\right|\right| \\
&\le \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\left|\sum_{(i,j)}(x_{ij}-\pi_{ij})\mathsf{f}_U(\alpha)_{ij}^2\right| - \left|\sum_{(i,j)}(z_{ij}-\pi_{ij})\mathsf{f}_U(\alpha)_{ij}^2\right|\right| \\
&\le \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}(x_{ij}-\pi_{ij})\mathsf{f}_U(\alpha)_{ij}^2 - \sum_{(i,j)}(z_{ij}-\pi_{ij})\mathsf{f}_U(\alpha)_{ij}^2\right| \\
&\le \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}(x_{ij}-z_{ij})\mathsf{f}_U(\alpha)_{ij}^2\right| \\
&\le \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\sqrt{\sum_{(i,j)}\pi_{ij}^{-1}(x_{ij}-z_{ij})^2}\sqrt{\sum_{(i,j)}\pi_{ij}\mathsf{f}_U(\alpha)_{ij}^4} \\
&\le \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\sqrt{p^{-1}}\sqrt{\sum_{(i,j)}(x_{ij}-z_{ij})^2}\sqrt{\sum_{(i,j)}\pi_{ij}\mathsf{f}_U(\alpha)_{ij}^2} \\
&\le \sqrt{p^{-1}T}\sqrt{\sum_{(i,j)}(x_{ij}-z_{ij})^2},
\end{aligned}$$

where we used $||a|-|b|| \le |a-b|$,$\|\mathsf{f}_U(\alpha)\|_\infty \le 1$ and $\|A\|_\Pi^2 \le T$. Thus, Lemma 9 and the identity $\sqrt{p^{-1}T} \le \frac{96p^{-1}}{2} + \frac{T}{2\times 96}$ imply

$$\mathbb{P}\left(|Z-\mathbb{E}[Z]| \ge 768p^{-1} + \frac{1}{12}T + t\right) \le 4e^{-t^2p/2T}.$$

Taking $t = T/3$ we get

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq 768p^{-1} + \frac{5}{12}T\right) \leq 4e^{-pT/18}. \tag{33}$$

Now we must bound the expectation $\mathbb{E}[Z_T]$. To do so, we use a symmetrization argument (Ledoux, 2001) which gives

$$\begin{aligned}\mathbb{E}[Z_T] &= \mathbb{E}\left[\sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}\Omega_{ij}\mathsf{f}_U(\alpha)_{ij}^2 - \mathbb{E}[]\sum_{(i,j)}\Omega_{ij}\mathsf{f}_U(\alpha)_{ij}^2\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}\epsilon_{ij}\Omega_{ij}\mathsf{f}_U(\alpha)_{ij}^2\right|\right],\end{aligned}$$

where $\{\epsilon_{ij}\}$ is an i.i.d. Rademacher sequence independent of $\{\Omega_{ij}\}$. We apply an extension Talagrand's contraction inequality to Lipschitz functions (see Koltchinskii (2011), Theorem 2.2) and obtain

$$\begin{aligned}\mathbb{E}[Z_T] = \mathbb{E}\left[\sup_{A\in\mathcal{T}}\left|\sum_{i,j}\epsilon_{ij}\Omega_{ij}A_{ij}^2\right|\right] &\leq 4æ\mathbb{E}\left[\sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}\epsilon_{ij}\Omega_{ij}A_{ij}\right|\right] \\ &= 4æ\mathbb{E}\left[\sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}|\langle\Sigma_R, \mathsf{f}_U(\alpha)\rangle|\right],\end{aligned}$$

where $\Sigma_R = \sum_{(i,j)}\epsilon_{ij}\Omega_{ij}E_{ij}$. Moreover, for $\alpha \in \tilde{\mathcal{A}}(d_1, T)$ we have

$$|\langle\Sigma_R, \mathsf{f}_U(\alpha)\rangle| = \left|\langle\Sigma_R, \sum_{k=1}^{N}\alpha_k U^k\rangle\right| \leq \|\alpha\|_1 u\|\Sigma_R\|_\infty.$$

Finally, we get $\mathbb{E}[Z_T] \leq 4æd_1u\mathbb{E}[\|\Sigma_R\|_\infty]$. Combining this with the concentration inequality (64) we complete the proof of Lemma 8:

$$\mathbb{P}\left(Z_T \geq 8æd_1u\mathbb{E}[\|\Sigma_R\|_\infty] + 768p^{-1} + \frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/18}.$$

□

Lemma 8 gives that $\mathbb{P}(\mathcal{B}_l) \leq 4\exp(-p\eta^l\nu/18)$. Applying the union bound we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{B}) &\leq \sum_{l=1}^{\infty}\mathbb{P}(\mathcal{B}_l) \leq 4\sum_{l=1}^{\infty}\exp(-p\eta^l\nu/18)\\ &\leq 4\sum_{l=1}^{\infty}\exp(-p\log(\eta)l\nu/18),\end{aligned}$$

where we used $e^x \geq x$. Finally, for $\nu = 18\log(d)/(p\log(6/5))$ we obtain

$$\mathbb{P}(\mathcal{B}) \leq \frac{4\exp(-p\nu\log(\eta)/18)}{1-\exp(-p\nu\log(\eta)/18)} \leq \frac{4\exp(-\log(d))}{1-\exp(-\log(d))} \leq \frac{8}{d},$$

since $d-1 \geq d/2$, which concludes the proof of (i).

**Proof of (ii):** The proof is similar to that of (i); we recycle some of the notations for simplicity. Recall $\mathsf{D}_X = 112\rho p^{-1}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8\text{æ}\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8\text{æ}d_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}$, and let

$$\begin{aligned}\mathcal{B} = \Big\{&\exists (L,\alpha) \in \mathcal{C}(d_1,d_\Pi,\rho,\varepsilon);\\ &\left|\|L+\mathsf{f}_U(\alpha)\|_\Omega^2 - \|L+\mathsf{f}_U(\alpha)\|_\Pi^2\right| > \frac{1}{2}\|L+\mathsf{f}_U(\alpha)\|_\Pi^2 + \mathsf{D}_X\Big\},\end{aligned}$$

$\nu = 72\log(d)/(p\log(6/5))$, $\eta = 6/5$ and for $l \in \mathbb{N}$

$$\mathcal{S}_l = \left\{(L,\alpha) \in \mathcal{C}(d_1,d_\Pi,\rho,\varepsilon): \quad \eta^{l-1}\nu \leq \|L+\mathsf{f}_U(\alpha)\|_\Pi^2 \leq \eta^l\nu\right\}.$$

As before, if $\mathcal{B}$ holds, then there exist $l \geq 2$ and $(L,\alpha) \in \mathcal{C}(d_1,d_\Pi,\rho,\varepsilon) \cap S_l$ such that

$$\left|\|L+\mathsf{f}_U(\alpha)\|_\Omega^2 - \|L+\mathsf{f}_U(\alpha)\|_\Pi^2\right| \quad > \frac{5}{12}\eta^l\nu + \mathsf{D}_X. \tag{34}$$

For $T > \nu$, consider the set $\tilde{\mathcal{C}}(T) = \{(L,\alpha) \in \mathcal{C}(d_1,d_\Pi,\rho,\varepsilon) : \|L+\mathsf{f}_U(\alpha)\|_\Pi^2 \leq T\}$, and the event

$$\mathcal{B}_l = \left\{\exists (L,\alpha) \in \tilde{\mathcal{C}}(\eta^l\nu): \quad \left|\|L+\mathsf{f}_U(\alpha)\|_\Omega^2 - \|L+\mathsf{f}_U(\alpha)\|_\Pi^2\right| > \frac{5}{12}\eta^l\nu + \mathsf{D}_X\right\}.$$

Then, (65) implies that $\mathcal{B}_l$ holds and $\mathcal{B} \subset \cup_{l=1}^{+\infty}\mathcal{B}_l$. Thus, we estimate in Lemma 10 the probability of the events $\mathcal{B}_l$, and then apply the union bound.

**Lemma 8.** *Let* $W_T = \sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} \left| \|L + \mathsf{f}_U(\alpha)\|_\Omega^2 - \|L + \mathsf{f}_U(\alpha)\|_\Pi^2 \right|$.

$$\mathbb{P}\left( W_T \geq \mathsf{D}_X + \frac{5}{12}T \right) \leq 4\mathrm{e}^{-pT/72}.$$

*Proof.* The proof is two-fold: first we show that $W_T$ concentrates around its expectation, then bound its expectation. By definition,

$$W_T = \sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} \left| \sum_{(i,j)} \Omega_{ij}(L_{ij} + \mathsf{f}_U(\alpha)_{ij})^2 - \mathbb{E}\left[ \sum_{(i,j)} \Omega_{ij}(L_{ij} + \mathsf{f}_U(\alpha)_{ij})^2 \right] \right|.$$

The concentration proof is exactly similar to the proof in Lemma 8, but we choose $t = T/6$, and we obtain

$$\mathbb{P}\left( |W_T - \mathbb{E}[W_T]| \geq 768p^{-1} + \frac{3}{12}T \right) \leq 4e^{-pT/72}. \tag{35}$$

Let us now bound the expectation $\mathbb{E}[W_T]$. Again, we use a standard symmetrization argument (Ledoux, 2001) which gives

$$\mathbb{E}[W_T] \quad \leq 2\mathbb{E}\left[ \sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} \left| \sum_{(i,j)} \epsilon_{ij}\Omega_{ij}(L_{ij} + \mathsf{f}_U(\alpha)_{ij})^2 \right| \right],$$

where $\{\epsilon_{ij}\}$ is an i.i.d. Rademacher sequence independent of $\Omega_{ij}$. Then, the contraction inequality (see Koltchinskii (2011), Theorem 2.2) yields

$$\mathbb{E}[W_T] \leq 4æ\mathbb{E}\left[ \sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} |\langle \Sigma_R, L + \mathsf{f}_U(\alpha)\rangle| \right],$$

where $\Sigma_R = \sum_{(i,j)} \epsilon_{ij}\Omega_{ij}E_{ij}$. Moreover

$$\begin{aligned} |\langle \Sigma_R, L + \mathsf{f}_U(\alpha)\rangle| \quad &\leq |\langle \Sigma_R, L\rangle| + |\langle \Sigma_R, \mathsf{f}_U(\alpha)\rangle| \\ &\leq \|L\|_*\|\Sigma_R\| + \|\alpha\|_1 u\|\Sigma_R\|_\infty. \end{aligned}$$

For $(L,\alpha) \in \tilde{\mathcal{C}}(T)$ we have by assumption $\|\alpha\|_1 \leq d_1$, $\|\mathsf{f}_U(\alpha)\|_\Pi \leq \sqrt{d_\Pi}$ and $\|L\|_* \leq \sqrt{\rho}\|L\|_F + \varepsilon$. We obtain

$$\begin{aligned} \|L\|_* \quad &\leq \sqrt{\frac{\rho}{p}}\|L\|_\Pi + \varepsilon \leq \sqrt{\frac{\rho}{p}}\left(\|L + \mathsf{f}_U(\alpha)\|_\Pi + \|\mathsf{f}_U(\alpha)\|_\Pi\right) + \varepsilon \\ &\leq \sqrt{\frac{\rho}{p}}\left(\sqrt{T} + \sqrt{d_\Pi}\right) + \varepsilon. \end{aligned}$$

This gives

$$\begin{aligned}\mathbb{E}\left[W_T\right] &\leq 4æ\left\{\sqrt{\frac{\rho}{p}}\left(\sqrt{T}+\sqrt{d_\Pi}\right)+\varepsilon\right\}\|\Sigma_R\|+4æd_1u\|\Sigma_R\|_\infty \\ &\leq \frac{T}{12}+\frac{d_\Pi}{2}+56æ^2\frac{\rho}{p}\|\Sigma_R\|^2+4æ\varepsilon\|\Sigma_R\|+4æd_1u\|\Sigma_R\|_\infty.\end{aligned}$$

Combining this with the concentration inequality (66) we finally obtain:

$$\mathbb{P}\left(W_T \geq \mathsf{D}_X+\frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/72}.$$

□

Lemma 10 gives that $\mathbb{P}\left(\mathcal{B}_l\right) \leq 4\exp(-p\eta^l\nu/72)$. Applying the union bound we obtain

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}\right) &\leq \sum_{l=1}^{\infty}\mathbb{P}\left(\mathcal{B}_l\right) \leq 4\sum_{l=1}^{\infty}\exp(-p\eta^l\nu/72) \\ &\leq 4\sum_{l=1}^{\infty}\exp(-p\log(\eta)l\nu/72),\end{aligned}$$

where we used $e^x \geq x$. Finally, for $\nu = 72\log(d)/(p\log(6/5))$ we obtain

$$\mathbb{P}\left(\mathcal{B}\right) \leq \frac{4\exp(-p\nu\log(\eta)/72)}{1-\exp(-p\nu\log(\eta)/72)} \leq \frac{4\exp(-\log(d))}{1-\exp(-\log(d))} \leq 8d^{-1},$$

since $d-1 \geq d/2$, which concludes the proof of (ii).

# H Proof of Lemma 1

The first inequality is trivially true using that $\|\Sigma\|_\infty = \max_{i,j}|\Omega_{ij}\epsilon_{ij}| \leq 1$. We prove the second inequality using an extension to rectangular matrices via self-adjoint dilation of Corollary 3.3 in Bandeira and van Handel (2016).

**Proposition 1.** *Let $A$ be an $m_1 \times m_2$ rectangular matrix with $A_{ij}$ independent centered bounded random variables. then, there exists a universal constant $C^*$ such that*

$$\mathbb{E}\left[\|A\|\right] \leq C^*\left\{\sigma_1 \vee \sigma_2 + \sigma_*\sqrt{\log(m_1 \wedge m_2)}\right\},$$

$$\sigma_1 = \max_i \sqrt{\sum_j \mathbb{E}\left[A_{ij}^2\right]}, \quad \sigma_2 = \max_j \sqrt{\sum_i \mathbb{E}\left[A_{ij}^2\right]}, \quad \sigma_* = \max_{i,j} |A_{ij}|.$$

Applying Proposition 1 to $\Sigma_R$ with $\sigma_1 \vee \sigma_2 \leq \sqrt{\beta}$ and $\sigma_* \leq 1$ we obtain

$$\mathbb{E}\left[\|\Sigma_R\|\right] \leq C^* \left\{\sqrt{\beta} + \sqrt{\log(m_1 \wedge m_2)}\right\}.$$

# I Proof of Lemma 2

Denote $\Sigma = \nabla \mathcal{L}(X^0; Y, \Omega)$. Definition (2) implies that $\mathbb{E}\left[Y_{ij}\right] = g_j'(X_{ij}^0)$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$. Combined with the sub-exponentiality of the entries $Y_{ij}$, we obtain that for all $i, j$, $Y_{ij} - g_j'(X_{ij}^0)$ is sub-exponential with scale and variance parameters $1/\gamma$ and $\sigma_+^2$ respectively. Then, noticing that $|\Omega_{ij}| \leq 1$ implies that for all $t \geq 0$,

$$\mathbb{P}\left\{\left|\Omega_{ij}\left(Y_{ij} - g_j'(X_{ij}^0)\right)\right| \geq t\right\} \leq \mathbb{P}\left\{\left|Y_{ij} - g_j'(X_{ij}^0)\right| \geq t\right\},$$

we obtain that the random variables $\Sigma_{ij} = \Omega_{ij}\left(Y_{ij} - g_j'(X_{ij}^0)\right)$ are also sub-exponential. Thus, for all $i, j$ and for all $t \geq 0$ we have that $|\Sigma_{ij}| \leq t$ with probability at least $1 - \max\left\{2e^{-t^2/2\sigma_+^2}, 2e^{-\gamma t/2}\right\}$. A union bound argument then yields

$$\|\Sigma\|_\infty \leq t \quad \text{w. p. at least } 1 - \max\left\{2m_1 m_2 e^{-t^2/2\sigma_+^2}, 2m_1 m_2 e^{-\gamma t/2}\right\},$$

where $\gamma$ and $\sigma_+$ are defined in **H**2. Using $\log(m_1 m_2) \leq 2 \log d$, where $d = m_1 + m_2$ and setting $t = 6 \max\left\{\sigma_+ \sqrt{\log d}, \gamma^{-1} \log d\right\}$, we obtain that with probability at least $1 - d^{-1}$,

$$\|\Sigma\|_\infty \leq 6 \max\left\{\sigma_+ \sqrt{\log d}, \gamma^{-1} \log d\right\},$$

which proves the first inequality. Now we prove the second inequality using the following result obtained by extension of Theorem 4 in Tropp (2012) to rectangular matrices.

**Proposition 2.** *Let $W_1, \ldots, W_n$ be independent random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}\left[W_i\right] = 0$. Suppose that*

$$\delta_* = \sup_{i \in [\![n]\!]} \inf_{\delta > 0} \left\{\mathbb{E}\left[\exp\left(\|W_i\|/\delta\right)\right] \leq \mathrm{e}\right\} < +\infty. \tag{36}$$

*Then, there exists an absolute constant $c^*$ such that, for all $t > 0$ and with probability at least $1 - \mathrm{e}^{-t}$ we have*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} W_i \right\| \leq c^* \max \left\{ \sigma_W \sqrt{\frac{t + \log d}{n}}, \delta_* \left( \log \frac{\delta_*}{\sigma_W} \right) \frac{t + \log d}{n} \right\},$$

*where*

$$\sigma_W = \max \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[ W_i W_i^\top \right] \right\|^{1/2}, \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[ W_i^\top W_i \right] \right\|^{1/2} \right\}.$$

For all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ define $Z_{ij} = -\Omega_{ij} \left( Y_{ij} - g_j'(X_{ij}^0) \right) E_{ij}$. The sub-exponentiality of the variables $\Omega_{ij} \left( Y_{ij} - g_j'(X_{ij}^0) \right)$ implies that for all $i, j \in [\![m_1]\!] \times [\![m_2]\!]$

$$\delta_{ij} = \inf_{\delta > 0} \quad \left\{ \mathbb{E}\left[ \exp\left( \left| \Omega_{ij} \left( Y_{ij} - g_j'(X_{ij}^0) \right) \right| / \delta \right) \right] \leq \mathrm{e} \right\} \leq \frac{1}{\gamma}.$$

We can therefore apply Proposition 2 to the matrices $Z_{ij}$ defined above, with the quantity

$$\sigma_Z = \max \left\{ \left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[ Z_{ij} Z_{ij}^\top \right] \right\|^{1/2}, \left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[ Z_{ij}^\top Z_{ij} \right] \right\|^{1/2} \right\}. \tag{37}$$

We obtain that for all $t \geq 0$ and with probability at least $1 - \mathrm{e}^{-t}$,

$$\|\Sigma\| \leq c^* \max \left\{ \sigma_Z \sqrt{m_1 m_2 (t + \log d)}, \left( \log \frac{1}{\gamma \sigma_Z} \right) \frac{t + \log d}{\gamma} \right\}.$$

We bound $\sigma_Z$ from above and below as follows.

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[ Z_{ij} Z_{ij}^\top \right] = \sum_{i=1}^{m_1} \left\{ \sum_{j=1}^{m_2} \mathbb{E}\left[ \Omega_{ij}^2 \right] \mathbb{E}\left[ \left( Y_{ij} - g_j'(X_{ij}^0) \right)^2 \right] \right\} E_{ii}(m_1),$$

where $E_{ii}(n)$, $i, n \geq 1$ denotes the $n \times n$ square matrix with 1 in the $(i,i)$-th entry and zero everywhere else. Therefore

$$\left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[ Z_{ij} Z_{ij}^\top \right] \right\|^{1/2} = \sqrt{\frac{1}{m_1 m_2} \max_i \sum_{j=1}^{m_2} \mathbb{E}\left[ \Omega_{ij}^2 \right] \mathbb{E}\left[ \left( Y_{ij} - g_j'(X_{ij}^0) \right)^2 \right]}.$$

Then, assumption **H**2 gives

$$\left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[ Z_{ij} Z_{ij}^\top \right] \right\|^{1/2} \leq \sigma_+ \sqrt{\frac{1}{m_1 m_2} \left( \max_i \sum_{j=1}^{m_2} \mathbb{E}\left[ \Omega_{ij}^2 \right] \right)},$$

and

$$\left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[ Z_{ij} Z_{ij}^\top \right] \right\|^{1/2} \geq \sigma_- \sqrt{\frac{1}{m_1 m_2} \left( \max_i \sum_{j=1}^{m_2} \mathbb{E}\left[ \Omega_{ij}^2 \right] \right)}.$$

Similarly, we obtain

$$\left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[ Z_{ij}^\top Z_{ij} \right] \right\|^{1/2} \leq \sigma_+ \sqrt{\frac{1}{m_1 m_2} \left( \max_j \sum_{i=1}^{m_1} \mathbb{E}\left[ \Omega_{ij}^2 \right] \right)},$$

and

$$\left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[ Z_{ij}^\top Z_{ij} \right] \right\|^{1/2} \geq \sigma_- \sqrt{\frac{1}{m_1 m_2} \left( \max_j \sum_{i=1}^{m_1} \mathbb{E}\left[ \Omega_{ij}^2 \right] \right)}.$$

Combining the last four inequalities, we obtain

$$\sigma_- \sqrt{\frac{\beta}{m_1 m_2}} \leq \sigma_Z \leq \sigma_+ \sqrt{\frac{\beta}{m_1 m_2}},$$

and setting $t = \log d$, we further obtain for all $t \geq 0$ and with probability at least $1 - d^{-1}$:

$$\|\Sigma\| \leq c^* \max \left\{ \sigma_+ \sqrt{2\beta \log d}, \frac{2 \log d}{\gamma} \log \left( \frac{1}{\sigma_-} \sqrt{\frac{m_1 m_2}{\beta}} \right) \right\},$$

which proves the result.

# References


Agarwal, D., L. Zhang, and R. Mazumder (2011, September). Modeling item-–item similarities for personalized recommendations on yahoo! front page. *Ann. Appl. Stat. 5*(3), 1839–1875.

Agresti, A. (2013). *Categorical Data Analysis, 3rd Edition*. Wiley.

Aubin, J.-P. and I. Ekeland (1984). *Applied nonlinear analysis*. Pure and applied mathematics. John Wiley, New-York. A Wiley-Interscience publication.

Bandeira, A. S. and R. van Handel (2016, July). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab. 44*(4), 2479–2506.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cai, T. and W.-X. Zhou (2013, December). A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res. 14*(1), 3619–3647.

Candès, E. J., X. Li, Y. Ma, and J. Wright (2011, June). Robust principal component analysis? *J. ACM 58*(3), 11:1–11:37.

Cao, Y. and Y. Xie (2016, March). Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing 64*(6).

Chandrasekaran, V., S. Sanghavi, P. A. Parrilo, and A. S. Willsky (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization 21*(2), 572–596.

Chatterjee, S. (2015, February). Matrix estimation by universal singular value thresholding. *Ann. Statist. 43*(1), 177–214.

Davenport, M. A., Y. Plan, E. van den Berg, and M. Wootters (2012). 1-bit matrix completion. *CoRR abs/1209.3672*.

Fithian, W. and R. Mazumder (2018, 05). Flexible low-rank statistical modeling with missing data and side information. *Statist. Sci. 33*(2), 238–260.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

Gelman, A. and J. Hill (2007, June). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gunasekar, S., P. Ravikumar, and J. Ghosh (2014). Exponential family matrix completion under structural constraints. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1917–II–1925. JMLR.org.

Hastie, T., R. Mazumder, J. Lee, and R. Zadeh (2015, January). Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *The Journal of Machine Learning Research 16*, 3367–3402.

Heeringa, S., B. West, and P. Berlung (2010). *Applied Survey Data Analysis*. New Yor: Chapman and Hall/CRC.

Hsu, D., S. M. Kakade, and T. Zhang (2011). Robust matrix decomposition with sparse corruptions. *EEE Transactions on Information Theory 57*(11), 7221–7234.

Husson, F., J. Josse, B. Narasimhan, and G. Robin (2018, April). Imputation of mixed data with multilevel singular value decomposition. *arXiv e-prints*, arXiv:1804.11087.

Josse, J. and F. Husson (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software 70*(1), 1–31.

Kiers, H. A. L. (1991, June). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika 56*(2), 197–212.

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

Klopp, O. (2015). Matrix completion by singular value thresholding: sharp bounds. *Electronic journal of statistics 9*(2), 2348–2369.

Klopp, O., J. Lafond, É. Moulines, and J. Salmon (2015). Adaptive multinomial matrix completion. *Electronic Journal of Statistics 9*, 2950–2975.

Klopp, O., K. Lounici, and A. B. Tsybakov (2017, October). Robust matrix completion. *Probability Theory and Related Fields 169*(1), 523–564.

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

Kumar, N. K. and J. Schneider (2017). Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra 65*(11), 2212–2244.

Lafond, J. (2015). Low rank matrix completion with exponential family noise. *Journal of Machine Learning Research: Workshop and Conference Proceedings 40*, 1–18.

Landgraf, A. J. and Y. Lee (2015, June). Generalized principal component analysis: Projection of saturated model parameters. Technical report, The Ohio State University, Department of Statistics.

Ledoux, M. (2001). *The concentration of measure phenomenon*, Volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence.

Legendre, P., R. Galzin, and M. L. Harmelin-Vivien (1997). Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology 78*(2), 547–562.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. New-York: John Wiley & Sons series in probability and statistics.

Mao, X., S. X. Chen, and R. K. W. Wong (2018). Matrix completion with covariate information. *Journal of the American Statistical Association 0*(0), 1–13.

Mardani, M., G. Mateos, and G. B. Giannakis (2013, Aug). Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies. *IEEE Transactions on Information Theory 59*(8), 5186–5205.

Mazumder, R., T. Hastie, and R. Tibshirani (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research 11*, 2287–2322.

Murdoch, T. and A. Detsky (2013). The inevitable application of big data to health care. *JAMA 309*(13), 1351–1352.

Pagès, J. (2015). *Multiple factor analysis by example using R.* Chapman & Hall/CRC the R series (CRC Press). Taylor & Francis Group.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Srebro, N. and T. Jaakkola (2003). Weighted low-rank approximations. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 720–727. AAAI Press.

Talagrand, M. (1996, January). A new look at independence. *Ann. Probab. 24*(1), 1–34.

ter Braak, C. J., P. Peres-Neto, and S. Dray (2017, January). A critical issue in model-based inference for studying trait-based community assembly and a solution. *PeerJ 5*, e2885.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics 12*(4), 389–434.

Tseng, P. and S. Yun (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program. 117*(1-2, Ser. B), 387–423.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation* (1st ed.). Springer Publishing Company, Incorporated.

Udell, M., C. Horn, R. Zadeh, and S. Boyd (2016). Generalized low rank models. *Foundations and Trends in Machine Learning 9*(1).

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles 45*(3), 1–67.

Xu, H., C. Caramanis, and S. Sanghavi (2010). Robust pca via outlier pursuit. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, USA, pp. 2496–2504. Curran Associates Inc.