Main effects and interactions in mixed and incomplete data frames

Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, Robert Tibshirani

▶ **To cite this version:**

Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, Robert Tibshirani. Main effects and interactions in mixed and incomplete data frames. Journal of the American Statistical Association, 2019, 10.1080/01621459.2019.1623041 . hal-02423445

HAL Id: hal-02423445
https://hal.science/hal-02423445

Submitted on 24 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Main effects and interactions in mixed and incomplete data frames

Geneviève Robin[1,2], Olga Klopp[3,4], Julie Josse[1,2], Éric Moulines, *[1,2,5], and Robert Tibshirani[6,7]

[1]Centre de Mathématiques Appliquées, École Polytechnique, France
[2]Projet XPOP, INRIA
[3]ESSEC Business School
[4]CREST, ENSAE
[5]National Research University Higher School of Economics, Russia
[6]Department of Statistics, Stanford University
[7]Department of Biomedical Data Science, Stanford University

March 26, 2019

Abstract

A mixed data frame (MDF) is a table collecting categorical, numerical and count observations. The use of MDF is widespread in statistics and the applications are numerous from abundance data in ecology to recommender systems. In many cases, an MDF exhibits simultaneously *main effects*, such as row, column or group effects and *interactions*, for which a low-rank model has often been suggested. Although the literature on low-rank approximations is very substantial, with few exceptions, existing methods do not allow to incorporate main effects and interactions while providing statistical guarantees. The present work fills this gap.

*This work has been funded by the DataScience Inititiative (Ecole Polytechnique) and the Russian Academic Excellence Project '5-100'

We propose an estimation method which allows to recover simultaneously the main effects and the interactions. We show that our method is near optimal under conditions which are met in our targeted applications. We also propose an optimization algorithm which provably converges to an optimal solution. Numerical experiments reveal that our method, `mimi`, performs well when the main effects are sparse and the interaction matrix has low-rank. We also show that `mimi` compares favorably to existing methods, in particular when the main effects are significantly large compared to the interactions, and when the proportion of missing entries is large. The method is available as an R package on the Comprehensive R Archive Network.

Keywords: Low-rank matrix completion, missing values, heterogeneous data

1 Introduction

Mixed data frames (MDF) (see Pagès (2015); Udell et al. (2016)) are tables collecting categorical, numerical and count data. In most applications, each row is an example or a subject and each column is a feature or an attribute. A distinctive characteristic of MDF is that column entries may be of different types and most often many entries are missing. MDF appear in numerous applications including patient records in health care (survival values at different time points, quantitative and categorical clinical features like blood pressure, gender, disease stage, see, *e.g.*, Murdoch and Detsky (2013)), survey data (Heeringa et al., 2010, Chapters 5 and 6), abundance tables in ecology (Legendre et al., 1997), and recommendation systems (Agarwal et al., 2011).

1.1 Main effects and interactions

In all these applications, data analysis is often made in the light of additional information, such as sites and species traits in ecology, or users and items characteristics in recommendation systems. This caused the introduction of the two central concepts of interest in this article: *main effects* and *interactions*. This terminology is classically used to distinguish between effects of covariates on the observations which are independent of the other covariates (main effects), and effects of covariates on the observations which depend on the value of one or more other covariates (interactions). For example, in health care, a treatment might extend survival for all patients – this is a main effect – or extend survival for young patients but shorten it for older patients – this is an interaction.

Many statistical models have been developed to analyze such types of data. Abundance tables counting species across environments are for instance classically analyzed using the log-linear model (Agresti, 2013, Chapter 4). This model decomposes the logarithms of the

expected abundances into the sum of species (rows) and environment (columns) effects, plus a low-rank interaction term. Other examples include multilevel models (Gelman and Hill, 2007) to analyze hierarchically structured data where examples (patients, students, etc.) are nested within groups (hospitals, schools, etc.).

1.2 Generalized low-rank models

At the same time, low-rank models, which embed rows and columns into low-dimensional spaces, have been widely used for exploratory analysis of MDF (Kiers, 1991; Pagès, 2015; Udell et al., 2016). Despite the abundance of results in low-rank matrix estimation (see Kumar and Schneider (2017) for a literature survey), to the best of our knowledge, most of the existing methods for MDF analysis do not provide a statistically sound way to account for *main effects* in the data. In most applications, estimation of main effects in MDF has been done heuristically as a preprocessing step (Hastie et al., 2015; Udell et al., 2016; Landgraf and Lee, 2015). Fithian and Mazumder (2018) incorporate row and column covariates in their model, but mainly focus on optimization procedures and did not provide statistical guarantees concerning the main effects. Mao et al. (2018) propose a procedure to estimate jointly main effects and a low-rank structure – which can be interpreted as interactions –, but the procedure is based on a least squares loss, and is therefore not suitable to mixed data types.

On the other hand, several approaches to model non-Gaussian, and particularly discrete data are available in the matrix completion literature, but they do not consider main effects. Davenport et al. (2012) introduced one-bit matrix completion, where the observations are binary such as yes/no answers, and provide nearly optimal upper and lower bounds on the mean square error of estimation. One-bit matrix completion was also studied in Cai and Zhou (2013). In Klopp et al. (2015), the authors introduce multinomial matrix completion,

where the observations are allowed to take more than two values, such as ratings in recommendation systems, and propose a minimax optimal estimator. Unbounded non-Gaussian observations have also been studied before. For instance, Cao and Xie (2016) extended the approach of Davenport et al. (2012) to Poisson matrix completion, and Gunasekar et al. (2014) and Lafond (2015) both studied exponential family matrix completion.

1.3 Contributions

In the present paper we propose a new framework for incomplete and mixed data which allows to account for main effects and interactions. Before introducing a general model for MDF with sparse main effects and low-rank interactions, we start in Section 2 with a concrete example from survey data analysis. Then, we propose in Section 3 an estimation procedure based on the minimization of a doubly penalized negative quasi log-likelihood. We also propose a block coordinate gradient descent algorithm to compute our estimator, and prove its convergence result. In Section 4.1 we discuss the statistical guarantees of our procedure and provide upper bounds on the estimation errors of the sparse and low-rank components. To assess the tightness of our convergence rates, in Section 4.2, we derive lower bounds and show that, in a number of situations, our upper bounds are near optimal. In Section 4.3, we specialize our results to three examples of interest in applications.

To support our theoretical claims, numerical results are presented in Section 5. In Section 5.1, we provide the results of our experiments that show that our method "mimi" (**m**ain effects and **i**nteractions in **m**ixed and **i**ncomplete data frames) performs well when the main effects are sparse and the interactions are low-rank. In case of model mis-specification, mimi gives similar results to a two-step procedure where main effects and interactions are estimated separately. Then, in Section 5.2, we compare mimi to existing methods for mixed data imputation. Our experiments reveal that mimi compares favorably to competitors, in

particular, when the main effects are significantly large compared to the interactions, and when the proportion of missing entries is large. Finally, in Section 5.3, we illustrate the method with the analysis of a census data set. The method is implemented in the R (R Core Team, 2017) package available on the Comprehensive R Archive Network; the proofs and additional experiments are postponed to the supplementary material.

Notation We denote the Frobenius norm on $\mathbb{R}^{m_1 \times m_2}$ by $\|\cdot\|_F$, the operator norm by $\|\cdot\|$, the nuclear norm by $\|A\|_*$ and the sup norm $\|\cdot\|_\infty$. $\|\cdot\|_2$ is the usual Euclidean norm, $\|\cdot\|_0$ the number of non zero coefficients, and $\|\cdot\|_\infty$ the infinity norm. For $n \in \mathbb{N}$, denote $[\![n]\!] = \{1, \dots, n\}$. We denote the support of $\alpha \in \mathbb{R}^N$ by $\mathrm{supp}(\alpha) = \{k \in [\![N]\!], \alpha_k \neq 0\}$. For $I \subseteq [\![1]\!] m_1$, we denote $\mathbb{1}_I$, defined by $\mathbb{1}_I(i) = 1$ if $i \in I$ and 0 otherwise, the indicator of set I_h.

2 General model and examples

2.1 American Community Survey

Before introducing our general model, we start by giving a concrete example. The American Community Survey[1] (ACS) provides detailed information about the American people on a yearly basis. Surveyed households are asked to answer 150 questions about their employment, income, housing, etc. As shown in Table 1, this results in a highly heterogeneous and incomplete data collection.

Here, the Family Employment Status (FES) variable categorizes the surveyed population in groups, depending on whether the household contains a couple or a single person, and whether the householders are employed. In an exploratory data analysis perspective,

[1] https://www.census.gov/programs-surveys/acs/about.html

ID	Nb. people	Electricity bill ($)	Food Stamps	Family Employment Status	Allocation
1	2	160	No	Married couple, neither employed	Yes
2	1	390	No	**NA**	No
3	4	**NA**	No	Married couple, husband employed	No
4	2	260	No	Married couple, neither employed	No
5	2	100	No	Married couple, husband employed	No
6	2	130	No	**NA**	No

Table 1: American Community Survey: Excerpt of the 2016 public use microsample data.

a question of interest is: does the household category influence the value of the other variables? For example income, food stamps allocation, etc. Furthermore, as we do not expect the group effects to be sufficient to explain the observations, can we also model residuals, or *interactions*?

Denote $Y = (Y_{ij})$ the data frame containing the households in rows and the questions in columns. If the j-th column is continuous (electricity bill for instance), one might model the group effects and interactions as follows:

$$\mathbb{E}[Y_{ij}] = \alpha^0_{c(i)j} + L^0_{ij},$$

where $c(i)$ indicates the group to which individual i belongs, and $\alpha^0_{c(i)j}$ and L^0_{ij} are fixed group effects and interactions respectively. This corresponds to the so-called *multilevel regression* framework (Gelman and Hill, 2007). If the j-th column is binary (food stamps allocation for instance), one might model

$$\mathbb{P}(Y_{ij} = \text{“Yes”}) = \frac{\mathrm{e}^{X^0_{ij}}}{1 + \mathrm{e}^{X^0_{ij}}}, \quad X^0_{ij} = \alpha^0_{c(i)j} + L^0_{ij},$$

corresponding to a logistic regression framework.

The goal is then, from the mixed and incomplete data frame Y, to estimate simultaneously the vector of group effects α^0 and the matrix of interactions L^0. We propose a method

assuming the vector of main effects α^0 is sparse and the matrix of interactions L^0 has low-rank. The sparsity assumption means that groups affect a small number of variables. On the other hand, the low-rank assumption means the population can be represented by a few archetypical individuals and summary features (Udell et al., 2016, Section 5.4), which interact in a multiplicative manner. In fact, if L^0 is of rank r, then it can be decomposed as the sum of r rank-1 matrices as follows:

$$L^0 = \sum_{k=1}^{r} u_k v_k^\top,$$

where u_k (resp. v_k) is a vector of $\mathbb{R}^{m_1}$ (resp. $\mathbb{R}^{m_2}$). Thus, using the above example, we obtain

$$\mathbb{E}[Y_{ij}] = \alpha^0_{c(i)j} + \sum_{k=1}^{r} u_{ik} v_{jk},$$

where the last term $\sum_{k=1}^{r} u_{ik} v_{jk}$ can be interpreted as the sum of multiplicative interaction terms between latent individual types and features.

2.2 General model

We now introduce a new framework generalizing the above example to other types of data and main effects. Consider an MDF $Y = (Y_{ij})$ of size $m_1 \times m_2$. The entries in each column $j \in [\![m_2]\!]$ belong to an observation space, denoted $\mathbb{Y}_j$. For example, for numerical data, the observation space is $\mathbb{Y}_j = \mathbb{R}$, and for count data, $\mathbb{Y}_j = \mathbb{N}$ is the set of natural integers. For binary data, the observation space is $\mathbb{Y}_j = \{0, 1\}$. In the entire paper, we assume that the random variables (Y_{ij}) are independent and that for each $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, $Y_{ij} \in \mathbb{Y}_j$ and $\mathbb{E}\left[|Y_{ij}|\right] < \infty$. Furthermore, we will assume that Y_{ij} is sub-exponential with scale γ and variance σ^2: for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$ and $|z| < \gamma$, $\mathbb{E}\left[\mathrm{e}^{z(Y_{ij} - \mathbb{E}[Y_{ij}])}\right] \leq \mathrm{e}^{\sigma^2 z^2/2}$.

In our estimation procedure, we will use a data-fitting term based on heterogeneous exponential family quasi-likelihoods. Let $(\mathbb{Y}, \mathcal{Y}, \mu)$ be a measurable space, $h : \mathbb{Y} \to \mathbb{R}_+$, and $g : \mathbb{R} \to \mathbb{R}$ be functions. Denote by $\mathrm{Exp}^{(h,g)} = \{f_x^{(h,g)} : x \in \mathbb{R}\}$ the canonical exponential family. Here, h is the base function, g is the link function, and $f_x^{(h,g)}$ is the density with respect to the base measure μ given by

$$f_x^{(h,g)}(y) = h(y) \exp\left(yx - g(x)\right), \tag{1}$$

for $y \in \mathbb{Y}$. For simplicity, we assume $\int h(y) \exp(yx) \mu(\mathrm{d}y) < \infty$ for all $x \in \mathbb{R}$.

The exponential family is a flexible framework for different data types. For example, for numerical data, we set $g(x) = x^2\sigma^2/2$ and $h(y) = (2\pi\sigma^2)^{-1/2} \exp(-y^2/\sigma^2)$. In this case, $\mathrm{Exp}^{(h,g)}$ is the family of Gaussian distributions with mean $\sigma^2 x$ and variance σ^2. For count data, we set $g(x) = \exp(ax)$ and $h(y) = 1/y!$, where $a \in \mathbb{R}$. In this case, $\mathrm{Exp}^{(h,g)}$ is the family of Poisson distributions with intensity $\exp(ax)$. For binary data, $g(x) = \log(1 + \exp(x))$ and $h(y) = 1$. Here, $\mathrm{Exp}^{(h,g)}$ is the family of Bernoulli distributions with success probability $1/(1 + \exp(-x))$.

In our estimation procedure, we choose a collection $\{(g_j, h_j),\ j \in [\![m_2]\!]\}$ of link functions and base functions corresponding to the observation spaces $\{(\mathbb{Y}_j, \mathcal{Y}_j, \mu_j)\,,\ j \in [\![m_2]\!]\}$. For each $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, we denote by X_{ij}^0 the value of the parameter minimizing the divergence between the distribution of Y_{ij} and the exponential family $\mathrm{Exp}^{(h_j,g_j)}$, $j \in [\![m_2]\!]$:

$$X_{ij}^0 = \mathrm{argmin}_{x \in \mathbb{R}} \left\{-\mathbb{E}\left[Y_{ij}\right] x + g_j(x)\right\}. \tag{2}$$

To model main effects and interactions we assume the matrix of parameters $X^0 = (X_{ij}^0) \in$

$\mathbb{R}^{m_1 \times m_2}$ can be decomposed as the sum of sparse main effects and low-rank interactions:

$$X^0 = \sum_{k=1}^{N} \alpha_k^0 U^k + L^0. \tag{3}$$

Here, $\mathcal{U} = (U^1, \ldots, U^N)$ is a fixed dictionary of $m_1 \times m_2$ matrices, α^0 is a sparse vector with unknown support $\mathcal{I} = \{k \in [\![N]\!]; \alpha_k^0 \neq 0\}$ and L^0 is an $m_1 \times m_2$ matrix with low-rank. The decomposition introduced in (3) is a general model combining regression on a dictionary and low-rank design.

2.3 Low-rank plus sparse matrix decomposition

Such decompositions have been studied before in the literature. In particular, a large body of work has tackled the problem of reconstructing a sparse and a low-rank terms exactly from the observation of their sum. Chandrasekaran et al. (2011) derived identifiability conditions under which exact reconstruction is possible when the sparse component is *entry-wise sparse*; the same model was also studied in Hsu et al. (2011). Candès et al. (2011) proved a similar results for entry-wise sparsity, when the location of the non-zero entries are chosen uniformly at random. Xu et al. (2010) extended the model to study *column-wise sparsity*. Mardani et al. (2013) studied an even broader framework with general sparsity pattern and determined conditions under which exact recovery is possible.

In the present paper, we consider the problem of estimating a (general) sparse component and a low-rank term from noisy and incomplete observation of their sum, when the noise is heterogeneous and in the exponential family. Because of this noisy setting, we can not recover the two components exactly. Thus, we do not require strong identifiability conditions as those derived in (Chandrasekaran et al., 2011; Hsu et al., 2011; Candès et al., 2011; Xu et al., 2010; Mardani et al., 2013). However, since decomposition (3) may not be unique, we restrict our model to the following class of possible decompositions, to which

our estimator will be the closest. From all possible decompositions (α, L), consider (α^1, L^1) such that

$$(\alpha^1, L^1) \in \operatorname{argmin}_{X^0=\sum \alpha_k U^k + L} \{\|\alpha\|_0 + \operatorname{rank} L\}. \tag{4}$$

Let $s^1 = \|\alpha^1\|_0 + \operatorname{rank} L^1$. Finally let

$$(\alpha^0, L^0) \in \operatorname*{argmin}_{\substack{X^0=\sum \alpha_k U^k + L \\ \|\alpha\|_0 + \operatorname{rank} L = s^1}} \|\alpha\|_0 . \tag{5}$$

The decomposition satisfying (4) and (5) may also not be unique. Assume that there exists a pair $(\alpha^\star, L^\star) \neq (\alpha^0, L^0)$ satisfying (4) and (5). Then,

$$\left\|L^\star - L^0\right\|_F = \left\|\sum_k \alpha_k^\star U^k - \sum_k \alpha_k^0 U^k\right\|_F \leq 2a \left\|\alpha^0\right\|_0 \max_k \|U_k\|_2 = R,$$

with a an upper bound on $\|\alpha^0\|_\infty$. This implies that for all such possible decompositions $(\alpha^\star, L^\star)$ we have that $L^\star$ and $\sum_k \alpha_k^\star U^k$ are in the small balls of radius R and centered at L^0 and $\sum_k \alpha_k^0 U^k$ respectively. Our statistical guarantees in Section 4 show that our estimators of L^0 and $\sum_k \alpha_k^0 U^k$ are in balls of radius at least R, and also centered at L^0 and $\sum_k \alpha_k^0 U^k$. Moreover, we also show that this error bound is minimax optimal in several situations. To summarize, in our model the decomposition may not be unique, but all the possible decompositions are in a neighborhood of radius smaller than the optimal convergence rate.

2.4 Examples

We now provide three examples of dictionaries which can be used to model classical main effects.

Example 1. Group effects We assume the m_1 individuals are divided into H groups. For $h \in [\![H]\!]$ denote by $I_h \subset [\![m_1]\!]$ the h-th group containing n_h individuals. The

size of the dictionary is $N = Hm_2$ and its elements are, for all $(h, q) \in [\![H]\!] \times [\![m_2]\!]$, $U_{h,q} = (\mathbb{1}_{I_h}(i)\mathbb{1}_{\{q\}}(j))_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}$. This example corresponds to the model discussed in Section 2.1; we develop it further in Section 5 with simulations and a survey data analysis.

Example 2. Row and column effects (see *e.g.* (Agresti, 2013, Chapter 4)) Another classical model is the log-linear model for count data analysis. Here, Y is a matrix of counts. Assuming a Poisson model, the parameter matrix X^0, which satisfies $\mathbb{E}[Y_{ij}] = \exp(X^0_{ij})$ for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, is assumed to be decomposed as follows:

$$X^0_{ij} = (\alpha^0_r)_i + (\alpha^0_c)_j + L^0_{ij}, \tag{6}$$

where $\alpha^0_r \in \mathbb{R}^{m_1}$, $\alpha^0_c \in \mathbb{R}^{m_2}$ and $L^0 \in \mathbb{R}^{m_1\times m_2}$ is low-rank. This model is often used to analyze abundance tables of species across environments (see, *e.g.*, ter Braak et al. (2017)). In this case the low-rank structure of L^0 reflects the presence of groups of similar species and environments. Model (6) can be re-written in our framework as

$$X^0 = \sum_{k=1}^{N} \alpha^0_k U_k + L^0,$$

with $\alpha^0 = (\alpha^0_r, \alpha^0_c)$, $N = m_1 + m_2$ and where for $i \in [\![m_1]\!]$ and $j \in [\![m_2]\!]$ we have $U_i = (\mathbb{1}_{\{i\}}(k))_{(k,l)\in[\![m_1]\!]\times[\![m_2]\!]}$ and $U_{m_1+j} = (\mathbb{1}_{\{j\}}(l))_{(k,l)\in[\![m_1]\!]\times[\![m_2]\!]}$.

Example 3. Corruptions Our framework also embeds the well-known robust matrix completion problem (Hsu et al., 2011; Candès et al., 2011; Klopp et al., 2017) which is of interest, for instance, in recommendation systems.In this application, malicious users coexist with normal users, and introduce spurious perturbations.Thus, in robust matrix completion, we observe noisy and incomplete realizations of a low-rank matrix L^0 of fixed rank and containing zeros at the locations of malicious users, perturbed by corruptions. The sparse component corresponding to corruptions is denoted $\sum_{(i,j)\in\mathcal{I}} \alpha^0_k U_{i,j}$,

where the $U_{i,j}$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, are the matrices of the canonical basis of $\mathbb{R}^{m_1 \times m_2}$ $U_{i,j} = (\mathbb{1}_{\{i\}}(k)\mathbb{1}_{\{j\}}(l))_{(k,l)\in[\![m_1]\!]\times[\![m_1]\!]}$ and $\mathcal{I}$ is the set of indices of corrupted entries. Thus, the non-zero components of α^0 correspond to the locations where the malicious users introduced the corruptions.For this example, the particular case of quadratic link functions $g_j(x) = x^2/2$ was studied in Klopp et al. (2017). We generalize these results in two directions: we consider mixed data types and general main effects.

2.5 Missing values

Finally, we consider a setting with missing observations. Let $\Omega = (\Omega_{ij})$ be an observation mask with $\Omega_{ij} = 1$ if Y_{ij} is observed and $\Omega_{ij} = 0$ otherwise. We assume that Ω and Y are independent, i.e. a Missing Completely At Random (MCAR) scenario (Little and Rubin, 2002): (Ω_{ij}) are independent Bernoulli random variables with probabilities π_{ij}, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$. Furthermore for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, we assume there exists $p > 0$ allowed to vary with m_1 and m_2, such that

$$\pi_{ij} \geq p. \tag{7}$$

For $j \in [\![m_2]\!]$, denote by $\pi_{\cdot j} = \sum_{i=1}^{m_1} \pi_{ij}$, $j \in [\![m_2]\!]$ the probability of observing an element in the j-th column. Similarly, for $i \in [\![m_1]\!]$, denote by $\pi_{i\cdot} = \sum_{j=1}^{m_2} \pi_{ij}$ the probability of observing an element in the i-th row. We define the following upper bound:

$$\max_{i,j}(\pi_{i\cdot}, \pi_{\cdot j}) \leq \beta. \tag{8}$$

3 Estimation procedure

Consider the data-fitting term defined by the heterogeneous exponential family negative quasi log-likelihood

$$\mathcal{L}(X; Y, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{ij} \left\{ -Y_{ij} X_{ij} + g_j(X_{ij}) \right\}, \tag{9}$$

and define the function

$$f(\alpha, L) = \mathcal{L}(\mathsf{f}_U(\alpha) + L; Y, \Omega), \tag{10}$$

where for $\alpha \in \mathbb{R}^N$, $\mathsf{f}_U(\alpha) = \sum_{k=1}^N \alpha_k U_k$. We assume $\|\alpha^0\|_\infty \leq a$ and $\|L^0\|_\infty \leq a$ where $a > 0$ is a known upper bound. We use the nuclear norm $\|\cdot\|_*$ (the sum of singular values) and ℓ_1 norm $\|\cdot\|_1$ penalties as convex relaxations of the rank and sparsity constraints respectively:

$$(\hat{\alpha}, \hat{L}) \in \operatorname{argmin}_{(\alpha, L)} F(\alpha, L) \tag{11}$$

$$\text{s. t. } \|\alpha\|_\infty \leq a, \|L\|_\infty \leq a, \tag{12}$$

$$F(\alpha, L) = f(\alpha, L) + \lambda_1 \|L\|_* + \lambda_2 \|\alpha\|_1, \tag{13}$$

with $\lambda_1 > 0$ and $\lambda_2 > 0$. In the sequel, for all $(\hat{\alpha}, \hat{L})$ in the set of solutions, we denote by $\hat{X} = \mathsf{f}_U(\hat{\alpha}) + \hat{L}$.

3.1 Block coordinate gradient descent (BCGD)

To solve (11) we develop a block coordinate gradient descent algorithm where the two components α and L are updated alternatively in an iterative procedure. At every iteration, we compute a (strictly convex) quadratic approximation of the data fitting term and apply block coordinate gradient descent to generate a search direction. This BCGD algorithm is a special instance of the coordinate gradient descent method for non-smooth separable minimization developed in Tseng and Yun (2009).

Note that the upper bound on $\|\alpha\|_\infty$ and $\|L\|_\infty$ is required to derive the statistical guarantees and, for simplicity, we did not implement it in practice. That is, we solve the following relaxed problem:

$$(\hat{\alpha}, \hat{L}) \in \operatorname{argmin}_{(\alpha,L)} F(\alpha, L). \tag{14}$$

Quadratic approximation. For any $(\alpha, L) \in \mathbb{R}^N \times \mathbb{R}^{m_1 \times m_2}$ and for any direction $(d_\alpha, d_L) \in \mathbb{R}^N \times \mathbb{R}^{m_1 \times m_2}$, consider the following local approximation of the data fitting term

$$f(\alpha + d_\alpha, L + d_L) = f(\alpha, L) + \mathcal{A}(\mathsf{f}_U(\alpha) + L, d_\alpha, d_L) + o(\|d_\alpha\|_2^2 + \|d_L\|_F^2) \ , \tag{15}$$

where we have set

$$\begin{aligned}\mathcal{A}(X, d_\alpha, d_L) = -2 \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_{ij}[X_{ij}] Z_{ij}[X_{ij}] (\mathsf{f}_U(d_\alpha)_{ij} + d_{L ij}) \\ + \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_{ij}[X_{ij}] (\mathsf{f}_U(d_\alpha)_{ij} + d_{L ij})^2 + \nu \|d_\alpha\|_2^2 + \nu \|d_L\|_F^2. \end{aligned} \tag{16}$$

In (16), $\nu > 0$ is a positive constant and for $x \in \mathbb{R}$ and $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$,

$$w_{ij}[x] = \Omega_{ij} g_j''(x)/2 \ , \quad Z_{ij}[x] = (Y_{ij} - g_j'(x))/g_j''(x) \ . \tag{17}$$

Note that the approximation (16) is simply a Taylor expansion of $\mathcal{L}$ around X, with an additional quadratic term $\nu\|d_\alpha\|_2^2 + \nu\|d_L\|_F^2$ ensuring its strong convexity. Denote by $(\alpha^{[t]}, L^{[t]})$ the fit of the parameter at iteration t and set $X^{[t]} = \mathsf{f}_U(\alpha^{[t]}) + L^{[t]}$. We update α and L alternatively as follows.

α-Update. We first solve

$$d_\alpha^{[t]} \in \operatorname{argmin}_{d \in \mathbb{R}^N} \left\{ \mathcal{A}(X^{[t]}, d, 0) + \lambda_2 \|\alpha^{[t]} + d\|_1 \right\} \ . \tag{18}$$

Problem (18) may be rewritten as a weighted Lasso problem:

$$\text{argmin}_{\alpha\in\mathbb{R}^d} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[X_{ij}^{[t]}](Z_{ij}^{[t]} - [\mathsf{f}_U(\alpha)]_{ij})^2 + \nu\|\alpha^{[t]} - \alpha\|_2^2 + \lambda_2\|\alpha\|_1 \,,$$

where for $i, j \in [\![m_1]\!] \times [\![m_2]\!]$ we have set $Z_{ij}^{[t]} := Z_{ij}[X_{ij}^{[t]}] + \mathsf{f}_U(\alpha^{[t]})$. Efficient numerical solutions to this problem are available (see, e.g., Friedman et al. (2010)). To update $\alpha^{[t]}$, we select a step size with an Armijo line search. The procedure goes as follows. We choose $\tau_{\text{init}} > 0$ and we let $\tau_\alpha^{[t]}$ be the largest element of $\{\tau_{\text{init}}\beta^j\}_{j=0}^\infty$ satisfying

$$f(\alpha^{[t]} + \tau_\alpha^{[t]} d^{[t]}, L^{[t]}) + \lambda_2\|\alpha^{[t]} + \tau_\alpha^{[t]} d^{[t]}\|_1 \leq f(\alpha^{[t]}, L^{[t]}) + \lambda_2\|\alpha^{[t]}\|_1 + \tau_\alpha^{[t]}\zeta\Gamma_\alpha^{[t]},$$

where $0 < \beta < 1$, $0 < \zeta < 1$, $0 \leq \theta < 1$, and

$$\Gamma_\alpha^{[t]} := -2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[X_{ij}^{[t]}]Z_{ij}[X_{ij}^{[t]}][\mathsf{f}_U(d^{[t]})]_{ij} + \theta\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[X_{ij}^{[t]}]\left[\mathsf{f}_U(d^{[t]})\right]_{ij}^2 + \nu\|d^{[t]}\|_2^2 + \lambda_2\left\{\|\alpha^{[t]} + d^{[t]}\|_1 - \|\alpha^{[t]}\|_1\right\}.$$

We set $\alpha^{[t+1]} = \alpha^{[t]} + \gamma^{[t]} d_\alpha^{[t]}$ and $X^{[t+1/2]} = \mathsf{f}_U(\alpha^{[t+1]}) + L^{[t]}$.

L-Update. We first solve

$$d_L^{[t]} := \text{argmin}_{d\in\mathbb{R}^{m_1\times m_2}} \left\{\mathcal{A}(X^{[t+1/2]}, 0, d) + \lambda_1\|L^{[t]} + d\|_*\right\}, \tag{19}$$

which is equivalent to

$$\text{argmin}_{L\in\mathbb{R}^{m_1\times m_2}} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(\nu + w_{ij}[X_{ij}^{[t+1/2]}])(Z_{ij}^{[t+1/2]} - L_{ij})^2 + \lambda_1\|L\|_*, \tag{20}$$

where for $i, j \in [\![m_1]\!] \times [\![m_2]\!]$ we have set

$$Z_{ij}^{[t+1/2]} = \frac{w_{ij}[X_{ij}^{[t+1/2]}](Z_{ij}[X^{[t+1/2]}] + L_{ij}^{[t]}) + \nu L_{ij}^{[t]}}{\nu + w_{ij}[X_{ij}^{[t+1/2]}]}.$$

The minimisation problem (20) may be seen as a weighted version of `softImpute` (Hastie et al., 2015). Srebro and Jaakkola (2003) proposed to solve (20) using an EM algorithm where the weights in $(0, 1]$ are viewed as frequencies of observations in a missing value framework (see also Mazumder et al. (2010)). We use this procedure, which involves soft-thresholding of the singular values of L, by adapting the `softImpute` package (Hastie et al., 2015). To update $L^{[t]}$, we choose the step size using again the Armijo line search. We set $\tau_{\text{init}} > 0$ and let $\tau_L^{[t]}$ be the largest element of $\{\tau_{\text{init}}\beta^j\}_{j=0}^{\infty}$ satisfying

$$f(\alpha^{[t+1]}, L^{[t]} + \tau_L^{[t]} d_L^{[t]}) + \lambda_1 \|L^{[t]} + \gamma^{[t]} d_L^{[t]}\|_* \leq f(\alpha^{[t+1]}, L^{[t]}) + \lambda_1 \|L^{[t]}\|_* + \tau_L^{[t]} \zeta \Gamma_L^{[t]},$$

$$\Gamma_L^{[t]} := -2 \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_{ij} [X_{ij}^{[t+1/2]}] Z_{ij} [X_{ij}^{[t+1/2]}] d_{L\,ij}^{[t]} + \theta \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_{ij} [X_{ij}^{[t+1/2]}] d_{L\,ij}^{[t]\,2} + \lambda_1 \left\{ \|L^{(t)} + d_L^{[t]}\|_* - \|d_L^{[t]}\|_* \right\} .$$

We finally set $L^{[t+1]} = L^{[t]} + \tau_L^{[t]} d_L^{[t]}$.

3.2 Convergence of the BCGD algorithm

The algorithm described in Section 3.1 is a particular case of the coordinate gradient descent method for nonsmooth minimisation introduced in Tseng and Yun (2009). In the aforementioned paper, the authors studied the convergence of the iterate sequence to a stationary point of the objective function. Here, we apply their general result (Tseng and Yun, 2009, Theorem 1) to our problem to obtain global convergence guarantees. Consider the following assumption on the dictionary $\mathcal{U}$.

H1. *For all $k \in [\![N]\!]$ and $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, $U_{ij}^k \in [-1, 1]$ and there exists $\varrho > 0$ such that for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\sum_{k=1}^{N} |U_{ij}^k| \leq \varrho$.*

Assumption **H**1 is satisfied in the three models introduced in Examples 1, 2 and 3: for group effects and corruptions with $\varrho = 1$ and for row and column effects with $\varrho = 2$. In particular, it guarantees that $X^0 = \mathsf{f}_U(\alpha^0) + L^0$ satisfies $\|X^0\|_\infty \leq (1+\varrho)a$. Plugging this in the definition of X^0 in (2), this assumption also implies that $\mathbb{E}[Y_{ij}] \in g_j'([-(1+\varrho)a, (1+\varrho)a])$ for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$. Note that **H**1 can be relaxed by $\|U_k\|_\infty \leq \rho$, with ρ an arbitrary constant. Consider also the following assumption on the link functions.

H2. *For all $j \in [\![m_2]\!]$ the functions g_j are twice continuously differentiable. Moreover, there exist $0 < \sigma_-, \sigma_+ < +\infty$ such that for all $|x| \leq (1+\varrho)a$ and $j \in [\![m_2]\!]$, $\sigma_-^2 \leq g_j''(x) \leq \sigma_+^2$.*

Assumptions **H**1–2 imply that the data-fitting term has Lipschitz gradient. Furthermore, the quadratic approximation defined in (16) is strictly convex at every iteration. We obtain the following convergence result.

Theorem 1. *Assume **H**1–2 and let $\{(\alpha^{[k]}, L^{[k]})\}$ be the iterate sequence generated by the BCGD algorithm. Then the following results hold.*

(a) $\{(\alpha^{[k]}, L^{[k]})\}$ has at least one accumulation point. Furthermore, all the accumulation points of $\{(\alpha^{[k]}, L^{[k]})\}$ are global optima of F.

(b) $\{F(\alpha^{[k]}, L^{[k]})\} \to F(\hat{\alpha}, \hat{L})$.

Proof. See Appendix B. □

4 Statistical guarantees

We now state our main statistical results. Denote by $\langle \cdot, \cdot \rangle$ the usual trace scalar product in $\mathbb{R}^{m_1 \times m_2}$. For $a \geq 0$ and a sparsity pattern $\mathcal{I} \subset [\![N]\!]$, define the following sets

$$\begin{aligned} \mathcal{E}_1(a, \mathcal{I}) &= \left\{ \alpha \in \mathbb{R}^N, \|\alpha\|_\infty \leq a, \operatorname{supp}(\alpha) \subset \mathcal{I} \right\}, \\ \mathcal{E}_2(a, \mathcal{I}) &= \left\{ L \in \mathbb{R}^{m_1 \times m_2}, \|L\|_\infty \leq a, \max_{k \in \mathcal{I}} |\langle L, U_k \rangle| = 0 \right\}, \\ \mathcal{X}(a, \mathcal{I}) &= \left\{ X = \mathsf{f}_U(\alpha) + L; (\alpha, L) \in \mathcal{E}_1(a, \mathcal{I}) \times \mathcal{E}_2(a, \mathcal{I}) \right\}. \end{aligned} \tag{21}$$

H3. *There exist* $a > 0$ *and* $\mathcal{I} \subset [\![N]\!]$ *such that* $(\alpha^0, L^0) \in \mathcal{E}_1(a, \mathcal{I}) \times \mathcal{E}_2(a, \mathcal{I})$.

Assumption **H**3 can be relaxed to allow upper bounds to depend on the entries of α^0 and L^0, but we stick to **H**3 for simplicity.

4.1 Upper bounds

We now derive upper bounds for the Frobenius and ℓ_2 norms of the estimation errors $L^0 - \hat{L}$ and $\alpha^0 - \hat{\alpha}$ respectively. In Theorem 2 we give a general result under conditions on the regularization parameters λ_1 and λ_2, which depend on the random matrix $\nabla \mathcal{L}(X^0; Y, \Omega)$. Then, Lemma 1 and 2 allow us to compute values of λ_1 and λ_2 that satisfy the assumptions of Theorem 2 with high probability. Finally we combine these results in Theorem 3.

We denote $\vee$ and $\wedge$ the max and min operators respectively, $M = m_1 \vee m_2$, $m = m_1 \wedge m_2$ and $d = m_1 + m_2$. We also define $r = \operatorname{rank}(L^0)$, $s = \|\alpha^0\|_0$ and $u = \max_k \|U_k\|_1$. Let $(E_{ij})_{(i,j) \in [\![m_1]\!] \times [\![m_2]\!]}$ be the canonical basis of $\mathbb{R}^{m_1 \times m_2}$ and $\{\epsilon_{ij}\}$ an i.i.d. Rademacher sequence independent of Y and Ω. Define

$$\Sigma_R = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{ij} \epsilon_{ij} E_{ij} \quad \text{and} \quad \nabla \mathcal{L}(X; Y, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{ij} \left\{ -Y_{ij} + g_j'(X_{ij}) \right\} E_{ij}. \tag{22}$$

Σ_R is a random matrix associated with the missingness pattern and $\nabla\mathcal{L}(X;Y,\Omega)$ is the gradient of $\mathcal{L}$ with respect to X. Define also

$$\begin{aligned}\Theta_1 &= \frac{\lambda_2}{\sigma_-^2} + a^2 u \mathbb{E}\left[\|\Sigma_R\|_\infty\right] + \frac{p}{\|\alpha^0\|_1}\left(\frac{a}{p}\right)^2 \log(d),\\ \Theta_2 &= \lambda_1^2 + (1+\varrho)a\mathbb{E}\left[\|\Sigma_R\|^2\right],\\ \Theta_3 &= \frac{\lambda_2}{\lambda_1} + 2\Theta_1.\end{aligned}$$

Theorem 2. *Assume **H**1-3 and let*

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

Then, with probability at least $1-8d^{-1}$,

$$\|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2 \leq \frac{as}{p}C_1\Theta_1 \quad \textit{and} \quad \|L^0 - \hat{L}\|_F^2 \leq \frac{r}{p^2}C_2\Theta_2 + \frac{as}{p}C_3\Theta_3, \tag{23}$$

where C_1, C_2 *and* C_3 *are numerical constants independent of* m_1, m_2 *and* p.

Proof. See Appendix C. □

We now give deterministic upper bounds on $\mathbb{E}\left[\|\Sigma_R\|\right]$ and $\mathbb{E}\left[\|\Sigma_R\|_\infty\right]$ in Lemma 1, and probabilistic upper bounds on $\|\nabla\mathcal{L}(X^0;Y,\Omega)\|$ and $\|\nabla\mathcal{L}(X^0;Y,\Omega)\|_\infty$ in Lemma 2. We will use them to select values of λ_1 and λ_2 which satisfy the assumptions of Theorem 2 and compute the corresponding upper bounds.

Lemma 1. *There exists an absolute constant* C^* *such that the two following inequalities hold*

$$\mathbb{E}\left[\|\Sigma_R\|_\infty\right] \leq 1 \quad \textit{and} \quad \mathbb{E}\left[\|\Sigma_R\|\right] \leq C^*\left\{\sqrt{\beta} + \sqrt{\log m}\right\}.$$

Proof. See Appendix H □

Lemma 2. *Assume **H**1-3. Then, there exists an absolute constant c^* such that the following two inequalities hold with probability at least $1 - d^{-1}$:*

$$\begin{aligned} \|\nabla \mathcal{L}(X^0; Y, \Omega)\|_\infty &\leq 6 \max\left\{\sigma_+\sqrt{\log d}, \frac{\log d}{\gamma}\right\}, \\ \|\nabla \mathcal{L}(X^0; Y, \Omega)\| &\leq c^* \max\left\{\sigma_+\sqrt{\beta \log d}, \frac{\log d}{\gamma} \log\left(\frac{1}{\sigma_-}\sqrt{\frac{m_1 m_2}{\beta}}\right)\right\}, \end{aligned} \tag{24}$$

*where $d = m_1 + m_2$, σ_+ and γ are defined in **H** 2, and β in (8).*

Proof. See Appendix I. □

We now combine Theorem 2, Lemma 1 and 2 with a union bound argument to derive upper bounds on $\|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2$ and $\|L^0 - \hat{L}\|_F^2$. We assume that $M = (m_1 \vee m_2)$ is large enough, that is

$$M \geq \max\left\{\frac{4\sigma_+^2}{\gamma^6}\log^2\left(\frac{\sqrt{m}}{p\gamma\sigma_-}\right), 2\exp\left(\sigma_+^2/\gamma^2 \vee \sigma_+^2\gamma(1+\varrho a)\right)\right\}.$$

Define

$$\begin{aligned} \Phi_1 &= a^2 + \frac{\log(d)}{u\sigma_-^2\gamma} + \frac{a^2\log(d)}{pu\|\alpha^0\|_1}, \\ \Phi_2 &= \frac{\sigma_+^2}{\sigma_-^4}\log(d) + (1+\varrho)a\left(1 \vee (\log m/\beta)\right), \\ \Phi_3 &= \frac{12p\sqrt{\log(d)}}{\gamma(1+\varrho)a\sigma_+\sqrt{\beta}} + \frac{1}{\sigma_-^2}\left(\frac{\log d}{\gamma}\right) + \frac{p}{u}\frac{\log(d)}{u\sigma_-^2\gamma} + \frac{a^2\log(d)}{pu\|\alpha^0\|_1}, \end{aligned}$$

and recall that $s = \|\alpha^0\|_0$, $r = \mathrm{rank}(L^0)$, $\beta \geq \max_{i,j}\left(\sum_{l=1}^{m_2}\pi_{il}, \sum_{k=1}^{m_1}\pi_{kj}\right)$ and that the entries Y_{ij} are sub-exponential with scale parameter γ.

Theorem 3. *Assume **H**1-3 and let*

$$\lambda_1 = 2c^*\sigma_+\sqrt{\beta\log d}, \quad \lambda_2 \geq \frac{24u\log(d)}{\gamma},$$

where c_* is the absolute constant defined in Lemma 2. Then, with probability at least $1 - 10d^{-1}$,

$$\|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2 \leq C\frac{sau}{p}\Phi_1, \text{ and } \|L^0 - \hat{L}\|_F^2 \leq C\left(\frac{r\beta}{p^2}\Phi_2 + \frac{sau}{p}\Phi_3\right), \tag{25}$$

with C an absolute constant.

Denoting by $\lesssim$ the inequality up to constant and logarithmic factors we get:

$$\|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2 \lesssim \frac{su}{p}, \text{ and } \|L^0 - \hat{L}\|_F^2 \lesssim \frac{r\beta}{p^2} + \frac{su}{p},$$

In the case of almost uniform sampling, *i.e.*, for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ and two positive constants c_1 and c_2, $c_1 p \leq \pi_{ij} \leq c_2 p$ we obtain that $\beta \leq c_2 M p$ and the following simplified bound:

$$\|L^0 - \hat{L}\|_F^2 \lesssim \frac{rM}{p} + \frac{su}{p}. \tag{26}$$

The rate given in (26) is the sum of the usual convergence rate of low-rank matrix completion rM/p and of the usual sparse vector convergence rate s (Bühlmann and van de Geer, 2011; Tsybakov, 2008) multiplied by u/p. This additional factor accounts for missing observations (p^{-1}) and interplay between main effects and interactions (u). Furthermore, the estimation risk of $\mathsf{f}_U(\alpha^0)$ is also the usual sparse vector convergence rate, with an additional up^{-1} factor accounting for interactions and missing values.

Note that whenever the dictionary $\mathcal{U}$ is linearly independent, Theorem 3 also provides an upper bound on the estimation error of α^0. Let $G \in \mathbb{R}^{N \times N}$ be the Gram matrix of the dictionary $\mathcal{U}$ defined by $G_{kl} = \langle U_k, U_l \rangle$ for all $(k,l) \in [\![N]\!] \times [\![N]\!]$.

H4. *For $\kappa > 0$ and all $\alpha \in \mathbb{R}^N$, $\alpha^\top G \alpha \geq \kappa^2 \|\alpha\|_2^2$.*

Recall that in the group effects model, we denote by I_h the set of rows which belong to group h. **H**4 is satisfied for the group effects model with $\kappa^2 = \min_h |I_h|$, the row and

column effects model with $\kappa^2 = \min(m_1, m_2)$ and the corruptions model with $\kappa^2 = 1$. If **H**4 is satisfied then, Theorem 3 implies that (up to constant and logarithmic factors):

$$\|\alpha^0 - \hat{\alpha}\|_2^2 \lesssim \frac{su}{p\kappa^2}.$$

4.2 Lower bounds

To characterize the tightness of the convergence rates given in Theorem 3, we now provide lower bounds on the estimation errors. We need three additional assumptions.

H5. *The sampling of entries is uniform, i.e. for all* $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\pi_{ij} = p$.

H6. *There exists* $\mathcal{I} \subset [\![N]\!]$, $a > 0$ *and* $X \in \mathcal{X}_{\mathcal{I},a}$ *such that for all* $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $Y_{ij} \sim \mathrm{Exp}^{(h_j, g_j)}(X_{ij})$.

Denote $\tau = \max_k \sum_{l \neq k} |\langle U_k, U_l \rangle|$. Without loss of generality we assume $m_1 = m_1 \vee m_2 = M$. For all $X \in \mathbb{R}^{m_1 \times m_2}$ we denote $\mathbb{P}_X$ the product distribution of (Y, Ω) satisfying **H**5 and 6. Consider two integers $s \leq (m_1 \wedge m_2)/2$ and $r \leq (m_1 \wedge m_2)/2$. We define the following set

$$\mathcal{F}(r,s) = \bigcup_{|\mathcal{I}| \leq s} \{(\alpha, L) \in \mathcal{E}_1(a, \mathcal{I}) \times \mathcal{E}_2(a, \mathcal{I}); \mathrm{rank}\,(L) \leq r\}\,. \tag{27}$$

Theorem 4. *Assume* ***H****1-5 and* $p \geq \frac{r}{m_1 \wedge m_2}$. *Then, there exists a constant* $\delta > 0$ *such that*

$$\inf_{\hat{L}, \hat{\alpha}} \sup_{(L^0, \alpha^0) \in \mathcal{F}(r,s)} \mathbb{P}_{X^0} \left(\|L^0 - \hat{L}\|_F^2 + \|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2 > \Psi_1 \frac{rM}{p} + \Psi_2 \frac{s\kappa^2}{p} \right) \geq \delta, \tag{28}$$

$$\begin{aligned} \Psi_1 &= C \min\left(\sigma_+^{-2}, \min(a, \sigma_+)^2\right), \\ \Psi_2 &= C \left(\frac{1}{\sigma_+^2 \left(\max_k \|U^k\|_F^2 + 2\tau\right)} \wedge (a \wedge \sigma_+)^2 \right). \end{aligned} \tag{29}$$

Proof. See Appendix D. □

Model	Group effects	Row & col effects	Corruptions
u	$\max_h \lvert I_h \rvert$	M	1
$\lVert \Delta L \rVert_F^2 + \lVert \mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha}) \rVert_F^2$	$rM/p + s \max_h \lvert I_h \rvert / p$	$rM/p + sM/p$	$rM/p + s/p$

Table 2: Order of magnitude of the upper bound for Examples 1, 2 and 3 (up to logarithmic factors).

Model	Group effects	Row & col effects	Corruptions
u	$\max_h \lvert I_h \rvert$	M	1
$\max_k \lVert U_k \rVert_F^2$	$\max_h \lvert I_h \rvert$	M	1
κ^2	$\min_h \lvert I_h \rvert$	m	1
$\lVert \Delta L \rVert_F^2 + \lVert \mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha}) \rVert_F^2$	$rM/p + (s \min_h \lvert I_h \rvert)/(p \max_h \lvert I_h \rvert)$	$rM/p + sm/(pM)$	$rM/p + s/p$

Table 3: Order of magnitude of the lower bound for Examples 1, 2 and 3.

4.3 Examples

We now specialize our theoretical results to Examples 1, 2 and 3 presented in Section 2.2. We compute the values of u, τ and $\max_k \lVert U^k \rVert_F^2$ for the group effects, row and column effects and corruption models, and obtain the rates of Theorem 3 and Theorem 4 for these particular cases. Recall that in the group effects model, we denote by I_h the set of rows which belong to group h. The orders of magnitude are summarized in Table 2 for the upper bound and in Table 3 for the lower bound. Comparing Table 2 and Table 3 we see that the convergence rates obtained in Theorem 3 are minimax optimal across the three examples whenever $s < r$. Furthermore, in the corruptions model our rates are optimal (up to constant and logarithmic factors) for any values of r, s and M, and equal to the minimax rates derived in Klopp et al. (2017). In the case of group effects, the rates are

optimal when $r > s \max_h |I_h|/M$ or when $\max_h |I_h|$ is of the order of a constant. When $s > rM/\max_h |I_h|$, we have an additional factor of the order $(\max_h |I_h|)^2/\min_h |I_h|$ in the upper bound. Note that the bounds have the same dependence in the sparsity pattern s. In the row and column model, when $r < s$, we have an additional factor of the order s/r in the upper bound.

5 Numerical results

5.1 Estimation of main effects and interactions

We start by evaluating our method (referred to as "mimi": **m**ain effects and **i**nteractions in **m**ixed and **i**ncomplete data) in terms of estimation of main effects and interactions. In this experiment, we focus on the group effects model presented in Section 2.1, with $H = 5$ groups of equal size. We select at random s non-zero coefficients in α^0, and construct a matrix L^0 of rank k. Then, $X^0 = \sum_{h=1}^{H} \sum_{j=1}^{m_2} \alpha_{hj}^0 U^{h,j} + L^0$, with $U_{h,j}$, $1 \leq h \leq H$ and $1 \leq j \leq m_2$ defined in Example 1. Finally, every entry of the matrix is observed with probability p.

In this first experiment, we consider only numeric variables to compare mimi to the following two-step method. In this alternative method, the main effects α^0 are estimated by the means of the variables taken by group; this corresponds to the preprocessing step performed in Udell et al. (2016) and Landgraf and Lee (2015) for instance. Then, L^0 is estimated using `softImpute` (Hastie et al., 2015); we refer to this method as "group mean + softImpute". The regularization parameters of both methods are selected with cross-validation.

The results are displayed in Figure 1 where we plot the estimation errors $\|\hat{\alpha} - \alpha^0\|_2^2$ and

in Figure 2 $\|\hat{L} - L^0\|_F^2$ for different levels of sparsity and different ranks.

On Figure 1 we observe that for a fixed rank, mimi has a smaller error ($\|\hat{\alpha} - \alpha^0\|_2^2$) than the two-step procedure for small sparsity levels, and that the difference between the two methods cancels as the sparsity level increases. Furthermore, as the rank also increases (from top to bottom and from left to right), the difference between mimi and the two-step procedure also decreases. Finally, for large ranks and sparsity levels simultaneously, mimi has a large estimation error $\|\hat{\alpha} - \alpha^0\|_2^2$ compared to the two-step procedure which does not assume sparsity. This case can be seen as a model mis-specification setting.

On Figure 2 we observe that mimi has overall smaller errors ($\|\hat{L} - L^0\|_F^2$) than the two-step procedure. The difference between the two methods cancels as the rank increases. We also observe that the level of sparsity has little impact on the results. However, for large ranks and sparsity levels simultaneously, mimi has a larger estimation error $\|\hat{L} - L^0\|_F^2$ than the two-step procedure.

Secondly, we fix the level of sparsity to $s = 5$ and the rank to $k = 5$, and perform the same experiment for increasing problem sizes (150×30, 1500×300 and 1500×3000). The results are given in Figure 3. We observe that the excess risk $\|\hat{L} - L^0\|_F^2$ the two methods are similar. In terms of estimation of α^0, the estimation error of mimi is constant as the problem size increases but the sparsity level of α^0 is kept constant, as predicted by Theorem 3. On the contrary, we observe that estimating α^0 in a preprocessing step yields large errors in high dimensions.

5.2 Imputation of mixed data

To evaluate mimi in a mixed data setting, we compare it in terms of imputation of missing values to five state-of-the-art methods:

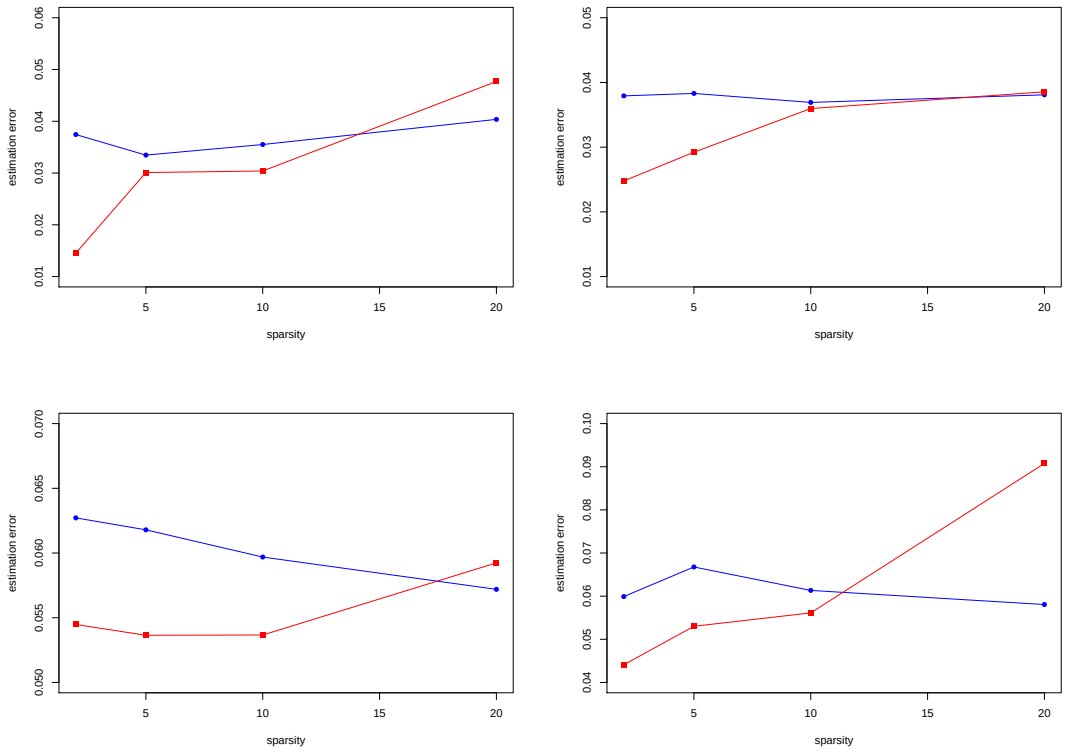

Figure 1: Estimation error $\|\hat{\alpha} - \alpha^0\|_2^2$ of mimi (red squares) and of groups means + softImpute (blue points) for increasing problem sparsity levels and ranks. The sparsity $s = 2, 5, 10, 20$ is indicated in the abscissa and the rank $k = 2, 5, 10, 20$ corresponds to different plots: top left $k = 2$, top right $k = 5$, bottom left $k = 10$, bottom right $k = 20$. The dimensions are fixed to $m_1 = 300$ and $m_2 = 30$ and the proportion of missing entries to $p = 0.2$.

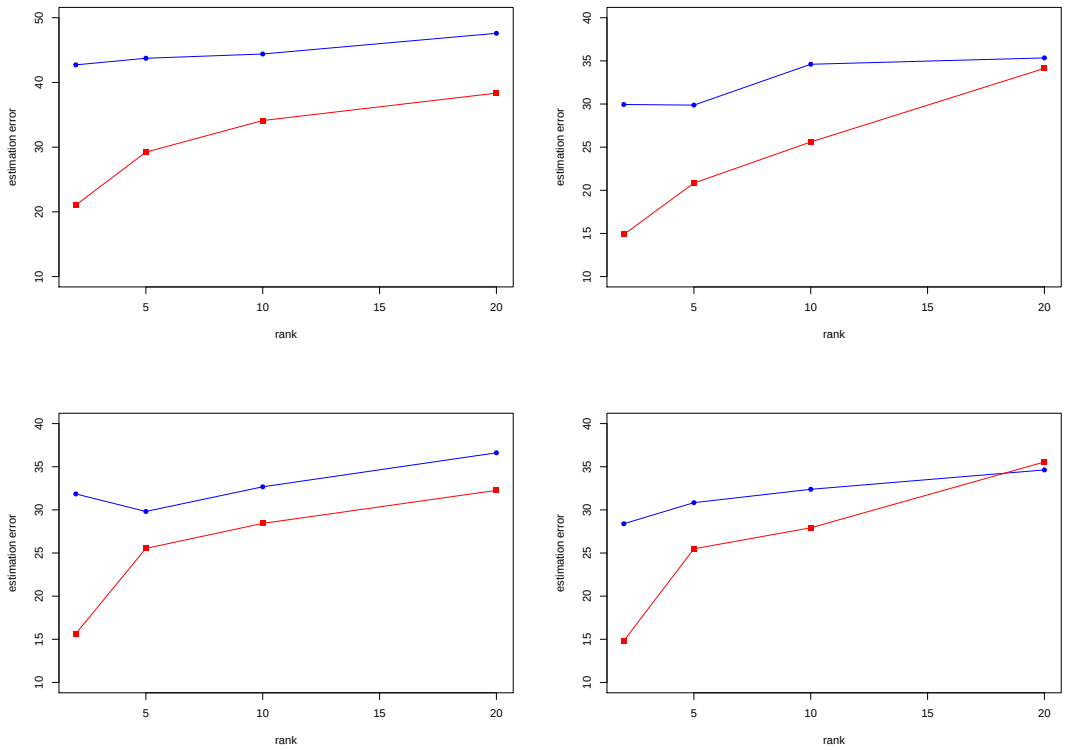

Figure 2: Estimation error $\|\hat{L} - L^0\|_F^2$ of mimi (red squares) and of groups means + softImpute (blue points) for increasing problem sparsity levels and ranks. The rank $k = 2, 5, 10, 20$ is indicated in the abscissa and the sparsity $s = 2, 5, 10, 20$ corresponds to different plots: top left $s = 2$, top right $s = 5$, bottom left $s = 10$, bottom right $s = 20$. The dimensions are fixed to $m_1 = 300$ and $m_2 = 30$ and the proportion of missing entries to $p = 0.2$.

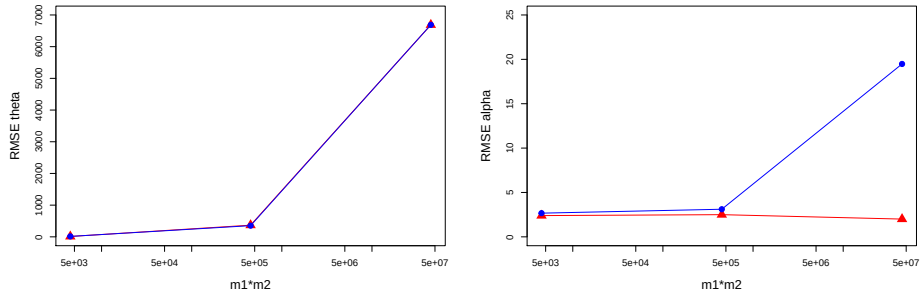

Figure 3: Estimation error of mimi (red triangles) and of groups means + softImpute (blue points) for increasing problem sizes ($m_1 m_2$, in log scale).

- softImpute (Hastie et al., 2015), a method based on soft-thresholding of singular values to impute numeric data implemented in the R package `softImpute`.
- Generalized Low-Rank Model (GLRM, Udell et al. (2016)), a matrix factorization framework for mixed data implemented in R in the `h2o` package.
- Factorial Analysis of Mixed Data (FAMD, Pagès (2015)), a principal component method for mixed data implemented in the R package `missMDA` (Josse and Husson, 2016).
- Multilevel Factorial Analysis of Mixed Data (MLFAMD, Husson et al. (2018)), an extension of FAMD to impute *multilevel* data, i.e. when individual are nested within groups. The method is also implemented in `missMDA`.
- Multivariate Imputation by Chained Equations (mice, van Buuren and Groothuis-Oudshoorn (2011)), an implementation of multiple imputation using Fully Conditional Specification. In the package `mice`, different models can be set for each column to account for mixed data.

Note that we also add a comparison to imputation by the column means, in order to have a baseline reference. We fix a dictionary $\mathcal{U}$ of indicator matrices corresponding to group effects (see Example 1), and generate a parameter matrix satisfying the decomposition (3). Then, columns are sampled from different data types, namely Gaussian and Bernoulli. For varying proportions of missing entries and values of the ratio $\rho = \|\mathsf{f}_U(\alpha^0)\|_F / \|L^0\|_F$, we evaluate the six methods in terms of imputation error of the two different data types. The parameters of all the methods (number of components for GLRM and FAMD and regularization parameters for softImpute and mimi) are selected using cross-validation. In addition, we use an optional ridge regularization in the h2o implementation of the GLRM method, which penalizes the ℓ_2 norm of the left and right principal components (U and V), and improved the imputation in practice. The details are available in the associated code provided as supplementary material.

% missing	**20**			**40**			**60**		
ρ	**0.2**	**1**	**5**	**0.2**	**1**	**5**	**0.2**	**1**	**5**
mean	24.5(0.7)	23.3(0.7)	22.9(0.4)	24.4(1.15)	33.2(1.1)	31.0(1.0)	42.1(1.2)	40.7(1.2)	39.9(0.6)
mimi	18.6(0.4)	**18.3**(0.3)	**17.7**(0.3)	18.8(0.3)	27.0(0.5)	**24.8**(0.6)	36.0(1.0)	**33.7**(0.8)	**30.6**(0.4)
GLRM	21.5(0.7)	22.0(0.8)	19.9(0.5)	21.5(0.7)	31.7(1.2)	31.0(0.9)	44.5(10.8)	49.4(16.2)	50.7(3.2)
softImpute	**18.5**(0.3)	18.5(0.2)	17.9(0.3)	18.6(0.3)	**26.8**(0.6)	24.9(0.5)	34.9(1.0)	34.9(0.8)	32.2(0.5)
FAMD	**18.5**(0.4)	18.9(0.4)	18.1(0.4)	18.7(0.3)	28.3(0.6)	25.6(0.7)	36.0(1.5)	40.6(0.8)	32.7(0.5)
MLFAMD	**18.5**(0.4)	19.2(0.4)	18.3(0.4)	**18.5**(0.5)	27.7(0.6)	26.3(0.5)	**34.9**(1.3)	40.7(1.0)	33.5(0.6)
mice	22.3(0.8)	22.6(0.6)	22.1(0.6)	22.7(0.6)	32.9(0.6)	30.1(0.9)	48.1(2.4)	48.1(0.9)	44.7(1.4)

Table 4: 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. In this simulation $m_1 = 150$, $m_2 = 30$, $s = 3$ and $r = 2$.

The results, presented in Table 4, reveal that mimi, softImpute, FAMD and MLFAMD yield imputation errors of comparable order. In this simulation setting, our method mimi improves on these existing methods when the ratio $\rho = \|\mathsf{f}_U(\alpha^0)\|_F / \|L^0\|_F$ is large, i.e. when the scale of the main effects is large compared to the interactions. The size of this improvement also increases with the amount of missing values. The imputation error by data type (quantitative and qualitative) are given in Appendix A, along with average experimental computational times of all the compared methods.

5.3 American Community Survey

We next apply our method on the American Community Survey data presented in Section 2.1. We use the 2016 survey [2] and restrict ourselves to the population of Alabama (24,614 household units). We focus on twenty variables (11 quantitative and 9 binary), and use mimi to estimate the effect of the Family Employment Status categories on these 20 variables. In other words, we place ourselves in the framework of Section 2.1 and Example 1. We model the quantitative attributes using Gaussian distributions, and the binary attributes with Bernoulli distributions. Using the same notations as in Section 2.1, $c(i)$ denotes the group (the FES) to which household i belongs. Thus, if the j-th column is continuous (income), our model implies:

$$\mathbb{E}[Y_{ij}] = \alpha^0_{c(i)j} + L^0_{ij}.$$

If the j-th column is binary (food stamps allocation for instance), we model

$$\mathbb{P}(Y_{ij} = 1) = \frac{\mathrm{e}^{X^0_{ij}}}{1 + \mathrm{e}^{X^0_{ij}}}, \quad X^0_{ij} = \alpha^0_{c(i)j} + L^0_{ij}.$$

[2] available at https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

In Table 5, we display the value of the parameter $\alpha_{c(i)j}$ for all possible groups $c(i)$ and some variables j corresponding to the number of people in household, food stamps and allocations attributions. The value of $\alpha_{c(i)j}$ is related to the expected value $\mathbb{E}[Y_{ij}]$: everything else being fixed, $\mathbb{E}[Y_{ij}]$ is an increasing function of $\alpha_{c(i)j}$. Thus, in terms of interpretation, the "group effect" $\alpha_{c(i)j}$ indicates (everything else being equal) whether belonging to category $c(i)$ yields larger or smaller values for $\mathbb{E}[Y_{ij}]$ compared to other categories.

We observe that household categories corresponding to married couples and single women have positive group effects on the variable "Number of people", meaning that these categories of households tend to have more children. We also observe that household categories containing employed people tend to receive less food stamps than other categories.

FES	Nb of people	Food stamps (0: no, 1: yes)	Allocations (0: no, 1: yes)
Couple - both in LF	0.38	-1.8	-0.68
Couple - male in LF	0.32	-1.4	-0.39
Couple - female in LF	0	-0.9	0
Couple - neither in LF	0	-1.6	-0.12
Male - in LF	0	0	0
Male - not in LF	0	0	0
Female - in LF	0.28	-0.19	0
Female - not in LF	0.13	0	0

Table 5: Effect of family type and employment status estimated with mimi.

The estimated low-rank component $\hat{L}$ has rank 5, indicating that 5 dimensions, in addition to the family type and employment status covariate, are needed to explain the observed data points.

6 Conclusion

This article introduces a general framework to analyze high-dimensonal, mixed and incomplete data frames with main effects and interactions. Upper bounds on the estimation error of main effects and interactions are derived. These bounds match with the lower-bounds under weak additional assumptions. Our theoretical results are supported by a numerical experiments on synthetic and survey data, showing that the introduced method performs best when the proportion of missing values is large and the main effects and interactions are of comparable size.

Our work opens several directions of future research. A natural extension would be to consider the inference problem, *i.e.*, to derive confidence intervals for the main effects coefficients. Another useful direction would be to consider exponential family distributions with multi-dimensional parameters, for example multinomial, distributions, to incorporate categorical variables with more than two categories. One could also learn the scale parameter (which we currently assume fixed) adaptively.

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.

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.

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.

Theorem 5. *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.*

% missing	20			40			60		
ρ	**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 6: 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.

% missing	20			40			60		
ρ	**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 7: 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.

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

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

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

(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 ℓ_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 3. *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 4. *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 Ω. For $B \in \mathbb{R}^{m_1\times m_2}$ we denote B_Ω 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 Π. 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 $\hat{\alpha}$: $\mathcal{L}(\hat{X}; Y, \Omega) - \mathcal{L}(X^0; Y, \Omega) \leq \lambda_1 \left(\|L^0\|_* - \|\hat{L}\|_*\right) + \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right)$. 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

$$\begin{aligned}\mathcal{L}(\hat{X}; Y, \Omega) - \mathcal{L}(X^0; Y, \Omega) + \langle \nabla\mathcal{L}(X^0; Y, \Omega), \Delta X\rangle \leq \\ \lambda_1 \left(\|L^0\|_* - \|\hat{L}\|_*\right) - \langle \nabla\mathcal{L}(X^0; Y, \Omega), \Delta L\rangle \\ + \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right) - \langle \nabla\mathcal{L}(X^0; Y, \Omega), \mathsf{f}_U(\Delta\alpha)\rangle. \quad (30)\end{aligned}$$

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

$$\mathcal{L}(X; Y, \Omega) - \mathcal{L}(\tilde{X}; Y, \Omega) - \langle \nabla\mathcal{L}(\tilde{X}; Y, \Omega), X - \tilde{X}\rangle \geq \frac{\sigma_-^2}{2}\|X - \tilde{X}\|_\Omega^2, \quad (31)$$

$$\|\nabla\mathcal{L}(X; Y, \Omega) - \nabla\mathcal{L}(\tilde{X}; Y, \Omega)\|_F \leq \sigma_+^2\|X - \tilde{X}\|_\Omega. \quad (32)$$

Plugging (31) into (30) allows to construct a lower bound on the left hand side term and

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

$$\begin{aligned} \mathsf{A}_1 &= \lambda_1 \left(\|L^0\|_* - \|\hat{L}\|_* \right) + \left| \langle \nabla \mathcal{L}(X^0; Y, \Omega), \Delta L \rangle \right|, \\ \mathsf{A}_2 &= \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1 \right) + \left| \langle \nabla \mathcal{L}(X^0; Y, \Omega), \mathsf{f}_U(\Delta\alpha) \rangle \right|. \end{aligned} \tag{33}$$

Let us upper bound 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

$$\|\hat{L}\|_* = \|L^0 - P_{L^{0\perp}}(\Delta L) - P_{L^0}(\Delta L)\|_* \geq \|L^0 + P_{L^{0\perp}}(\Delta L)\|_* - \|P_{L^0}(\Delta L)\|_*. \tag{34}$$

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 (34) we obtain

$$\|L^0\|_* - \|\hat{L}\|_* \leq \|P_{L^0}(\Delta L)\|_* - \|P_{L^{0\perp}}(\Delta L)\|_*, \tag{35}$$

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:

$$\mathsf{A}_1 \leq \frac{3\lambda_1}{2} \sqrt{2r} \|\Delta L\|_F. \tag{36}$$

We now derive an upper bound for A_2. The duality between $\|\cdot\|_1$ and $\|\cdot\|_\infty$ ensures

$$\left| \langle \nabla \mathcal{L}(X^0; Y, \Omega), \mathsf{f}_U(\Delta\alpha) \rangle \right| \leq \|\Delta\alpha\|_1 \|\nabla \mathcal{L}(X^0; Y, \Omega)\|_\infty u. \tag{37}$$

The assumption $\lambda_2 \geq 2\|\nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty u$ in conjunction with (37) and the triangular inequality $\|\Delta\alpha\|_1 \leq \|\alpha^0\|_1 + \|\hat{\alpha}\|_1$ yield

$$\mathsf{A}_2 \leq \frac{3\lambda_2}{2}\|\alpha^0\|_1. \tag{38}$$

Combining inequalities (33), (36) and (38) we obtain

$$\|\Delta X\|_\Omega^2 \leq \frac{3\lambda_1}{\sigma_-^2}\sqrt{2r}\|\Delta L\|_F + \frac{3\lambda_2}{\sigma_-^2}\|\alpha^0\|_1. \tag{39}$$

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

$$\|\Delta X\|_\Omega^2 \geq \|\Delta X\|_\Pi^2 - \mathsf{D}. \tag{40}$$

We start by defining our constrained set and prove that it contains the errors ΔL 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_Π, $\rho < m$ and ε that will be specified later on, define the two following sets where $\Delta\alpha$ and ΔL should lie:

$$\mathcal{A}(d_1, d_\Pi) = \left\{\alpha \in \mathbb{R}^N : \|\alpha\|_1 \leq d_1,\ \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq d_\Pi\right\}. \tag{41}$$

$$\begin{aligned}\mathcal{L}(\rho, \varepsilon) \quad = \Bigg\{ & L \in \mathbb{R}^{m_1 \times m_2}, \alpha \in \mathbb{R}^N : \|L + \mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{72\log(d)}{p\log(6/5)}, \\ & \|L + \mathsf{f}_U(\alpha)\|_\infty \leq 1, \|L\|_* \leq \sqrt{\rho}\|L\|_F + \varepsilon \Bigg\}\end{aligned} \tag{42}$$

If $\|\Delta X\|_\Pi^2$ is too small, the right hand side of (40) 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 ρ. 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 5. *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 6. *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 λ_1 and λ_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 $1-8d^{-1}$,

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

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

and where d_1, d_Π, ρ and ε 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:

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

$$\begin{aligned} \mathsf{D}_\alpha &= 8\varrho d_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1}, \\ \mathsf{D}_X &= \frac{112\rho}{p}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8\varrho\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8\varrho d_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}. \end{aligned} \tag{43}$$

Lemma 7. *(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 $1-8d^{-1}$

$$\|L + \mathsf{f}_U(\alpha)\|_\Omega^2 \geq \frac{1}{2}\|L + \mathsf{f}_U(\alpha)\|_\Pi^2 - \mathsf{D}_X. \tag{44}$$

Proof. See Appendix G. □

Lemma 7 (ii) applied to $\left(\frac{\Delta L}{2(1+\varrho)a}, \frac{\Delta\alpha}{2(1+\varrho)a}\right)$ implies that with probability at least $1-8d^{-1}$, $\|\Delta X\|_\Pi^2 \leq 2\|\Delta X\|_\Omega^2 + 4(1+\varrho)a\mathsf{D}_X$. Combined with (39), $\|\Delta X\|_F^2 \leq p^{-1}\|\Delta X\|_\Pi^2$, $\|\Delta X\|_F^2 \geq \|\Delta L\|_F^2/2 - \|\mathsf{f}_U(\Delta\alpha)\|_F^2$ 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):

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

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 α 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 $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $0 \leq X_{ij} \leq (1 + \varrho)a$. Thus, $\mathcal{X} \subset \mathcal{F}(r,s)$, where $\mathcal{F}(r,s)$ is defined in (27).

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

$$\|X - X'\|_F^2 \geq \frac{Mr}{8}\left(\eta^2 \min(a, \sigma_+)^2 \frac{r}{pm}\left\lfloor \frac{m_2}{r} \right\rfloor\right) \geq \frac{\eta^2}{16}\min(a^2, \sigma_+^2)\frac{rM}{p}. \tag{45}$$

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

$$\mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X) = \sum_{i,j} \pi_{ij}\left(g_j(X_{ij}) - g_j(0) - g_j'(0)X_{ij}\right) \leq \frac{\sigma_+^2 \eta^2 \min(a, \sigma_+)^2 Mr}{2}. \tag{46}$$

Inequality (46) implies that

$$\frac{1}{\mathrm{Card}(\mathcal{X}_L^0) - 1} \sum_{X \in \mathcal{X}_L^0} \mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X) \leq \frac{1}{16}\log(\mathrm{Card}(\mathcal{X}_L^0) - 1) \tag{47}$$

is satisfied for $\tilde{\eta} = \min\left\{1, (8\sigma_+ \min(a, \sigma_+))^{-1}\right\}$. Then, conditions (45) and (46) guarantee that we can apply Theorem 2.5 from Tsybakov (2008). We obtain that for some constant $\delta > 0$ and with $\Psi_1 = C \min\left(\sigma_+^{-2}, \min(a, \sigma_+^2)\right)$:

$$\inf_{\hat{L}, \hat{\alpha}} \sup_{(L^0, \alpha^0) \in \mathcal{E}} \mathbb{P}_{X^0}\left(\|\Delta L\|_F^2 + \|\Delta\alpha\|_2^2 > \frac{\Psi_1 rM}{p}\right) \geq \delta, \tag{48}$$

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 α and α' of $\mathcal{A}^0$, $\alpha \neq \alpha'$,

$$\|\alpha - \alpha'\|_2^2 \geq \frac{s}{8}\tilde{\eta}^2 \min(a, \sigma_+)^2. \tag{49}$$

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 $\mathbb{P}_X$

$$\mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X) = \sum_{i,j} \pi_{ij}(g_j(X_{ij}) - g_j(0) - g_j'(0)X_{ij} \leq \sigma_+^2 \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq \sigma_+^2 p \|\mathsf{f}_U(\alpha)\|_F^2. \tag{50}$$

Using Assumption **H**2

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) &\leq \sigma_+^2 p \left(\max_k \|U^k\|_F^2 + 2\tau \right) \|\alpha\|_2^2 \\ &\leq s\sigma_+^2 p \left(\max_k \|U^k\|_F^2 + 2\tau \right) \tilde{\eta}^2 \min(a, \sigma_+)^2. \end{aligned} \tag{51}$$

From (51) we deduce that

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0) - 1} \sum_{\mathcal{A}^0} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) \leq sp \left(\max_k \|U^k\|_F^2 + 2\tau \right) \sigma_+^2 \tilde{\eta}^2 \min(a, \sigma_+)^2. \tag{52}$$

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 $\delta > 0$

$$\inf_{\hat{L}, \hat{\alpha}} \sup_{(L^0, \alpha^0) \in \mathcal{E}} \mathbb{P}_{X^0} \left\{ \|\Delta L\|_F^2 + \| \sum_{k=1}^N (\alpha_k^0 - \hat{\alpha}_k) U^k \|_F^2 > \Psi_2 \frac{s\kappa^2}{p} \right\} \geq \delta, \tag{53}$$

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

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 (48) and (53).

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, α) we have

$$\langle \nabla \mathcal{L}(\hat{X}; Y, \Omega), L - \hat{L} + \sum_{k=1}^N (\alpha_k - \hat{\alpha}_k) U^k \rangle + \lambda_1 \langle \hat{f}_L, L - \hat{L} \rangle + \lambda_2 \langle \hat{f}_\alpha, \alpha - \hat{\alpha} \rangle \geq 0. \tag{54}$$

Applying inequality (54) 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 B_1, B_2 and B_3 separately. Recall that we denote $u = \max_k \|U^k\|_1$ and use (37) to bound B_1:

$$\mathsf{B}_1 \leq \|\Delta\alpha\|_1 \|\nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty u. \tag{55}$$

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 $g_j'(\tilde{X}_{ij}) - g_j'(X_{ij}^0)$, therefore assumption **H**2 ensures $\|\nabla\mathcal{L}(\tilde{X}; Y, \Omega) - \nabla\mathcal{L}(X^0; Y, \Omega)\|_\infty \leq 2\sigma_+^2(1+\varrho)a$, and finally we obtain

$$\mathsf{B}_2 \leq \|\Delta\alpha\|_1 2\sigma_+^2(1+\varrho)au. \tag{56}$$

We finally bound 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 (55) and (56) this yields

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

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

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

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

$$\|\Delta\alpha\|_1 \leq 4\|\alpha^0\|_1. \tag{57}$$

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 (57) 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 $1 - 8d^{-1}$,

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \leq 2\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 + 64\varrho a\|\alpha^0\|_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 3072a^2p^{-1}. \tag{58}$$

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 ((31)), we obtain

$$\begin{aligned}
\frac{\sigma_-^2}{2}\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 &\leq \lambda_2\left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right) + \langle\nabla\mathcal{L}(\tilde{X};Y,\Omega), \mathsf{f}_U(\Delta\alpha)\rangle \\
&\leq \lambda_2\left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right) + \underbrace{\left|\langle\nabla\mathcal{L}(X^0;Y,\Omega), \mathsf{f}_U(\Delta\alpha)\rangle\right|}_{\mathsf{C}_1} \\
&\quad + \underbrace{\left|\langle\nabla\mathcal{L}(X^0;Y,\Omega) - \nabla\mathcal{L}(\tilde{X};Y), \mathsf{f}_U(\Delta\alpha)\rangle\right|}_{\mathsf{C}_2}
\end{aligned} \tag{59}$$

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

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

Plugged into (58), this last inequality implies that with probability at least $1 - 8d^{-1}$

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \leq \frac{3\lambda_2}{\sigma_-^2}\|\alpha^0\|_1 + 64\varrho a\|\alpha^0\|_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 3072a^2p^{-1}. \tag{60}$$

Combining (57) and (60) gives the result.

F Proof of Lemma 6

Using (54) 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 (35) and the conditions

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

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 $\mathrm{rank}(P_{L^0}(\Delta L)) \leq 2r$, 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 $\mathsf{D}_\alpha = 8\varrho d_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1}$ 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

$$\left| \|\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathsf{f}_U(\alpha)\|_\Pi^2 \right| \quad > \frac{1}{2}\|\mathsf{f}_U(\alpha)\|_\Pi^2 + \mathsf{D}_\alpha > \frac{1}{2}\eta^{l-1}\nu + \mathsf{D}_\alpha = \frac{5}{12}\eta^l \nu + \mathsf{D}_\alpha. \tag{61}$$

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 (61) 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 8. *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 9. *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 $t \geq 0$, $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq 16L + t) \leq 4\mathrm{e}^{-t^2/2L^2}$.*

We apply this result to the function

$$f(x_{11}, \ldots, x_{m_1 m_2}) = \sup_{\alpha \in \tilde{\mathcal{A}}(d_1,T)} \left| \sum_{(i,j)} (x_{ij} - \pi_{ij}) \mathsf{f}_U(\alpha)_{ij}^2 \right|,$$

which is Lipschitz with Lipschitz constant $\sqrt{p^{-1}T}$. Indeed, for any $(x_{11}, \ldots, x_{m_1 m_2}) \in$

$\mathbb{R}^{m_1\times m_2}$ and $(z_{11},\dots,z_{m_1m_2})\in\mathbb{R}^{m_1\times m_2}$:

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

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

$$
\mathbb{P}\left(|Z-\mathbb{E}[Z]|\ge 768p^{-1}+\frac{5}{12}T\right)\le 4e^{-pT/18}. \tag{62}
$$

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

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

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 $\mathbb{E}[Z_T]\leq 4\varrho d_1 u\mathbb{E}[\|\Sigma_R\|_\infty]$. Combining this with the concentration inequality (62) we complete the proof of Lemma 8:

$$\mathbb{P}\left(Z_T\geq 8\varrho d_1 u\mathbb{E}[\|\Sigma_R\|_\infty] + 768p^{-1} + \frac{5}{12}T\right)\leq 4\mathrm{e}^{-pT/18}.$$

□

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 $\mathsf{D}_X = 112\rho p^{-1}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8\varrho\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8\varrho d_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}$, 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

$$\left|\|L+\mathsf{f}_U(\alpha)\|_\Omega^2 - \|L+\mathsf{f}_U(\alpha)\|_\Pi^2\right| \quad > \frac{5}{12}\eta^l\nu + \mathsf{D}_X. \tag{63}$$

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, (63) 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 10. *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

$$\mathbb{P}\left(|W_T - \mathbb{E}[W_T]| \geq 768 p^{-1} + \frac{3}{12} T \right) \leq 4e^{-pT/72}. \tag{64}$$

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 Ω_{ij}. Then, the contraction inequality (see Koltchinskii (2011), Theorem 2.2) yields

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

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

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

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

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

This gives

$$\begin{aligned}\mathbb{E}\left[W_T\right] &\leq 4\varrho\left\{\sqrt{\frac{\rho}{p}}\left(\sqrt{T}+\sqrt{d_\Pi}\right)+\varepsilon\right\}\|\Sigma_R\| + 4\varrho d_1 u\|\Sigma_R\|_\infty \\ &\leq \frac{T}{12} + \frac{d_\Pi}{2} + 56\varrho^2\frac{\rho}{p}\|\Sigma_R\|^2 + 4\varrho\varepsilon\|\Sigma_R\| + 4\varrho d_1 u\|\Sigma_R\|_\infty.\end{aligned}$$

Combining this with the concentration inequality (64) 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 Σ_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 σ_+^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

$$\|\Sigma\|_\infty \leq t \quad \text{w. p. at least } 1 - \max\left\{2m_1m_2e^{-t^2/2\sigma_+^2}, 2m_1m_2e^{-\gamma t/2}\right\},$$

where γ and σ_+ are defined in **H**2. Using $\log(m_1m_2) \leq 2\log d$, 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 $\mathbb{E}[W_i] = 0$. Suppose that*

$$\delta_* = \sup_{i\in[\![n]\!]} \inf_{\delta>0} \left\{\mathbb{E}\left[\exp\left(\|W_i\|/\delta\right)\right] \leq \mathrm{e}\right\} < +\infty. \tag{65}$$

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

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

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

$$\sigma_Z = \max\left\{ \left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[Z_{ij} Z_{ij}^\top\right] \right\|^{1/2}, \left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[Z_{ij}^\top Z_{ij}\right] \right\|^{1/2} \right\}. \tag{66}$$

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 σ_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.