Journal of Machine Learning Research 20 (2019) 1-43 Submitted 7/18; Revised 10/19; Published 10/19

Collective Matrix Completion

Mokhtar Z. Alaya MOKHTARZAHDI.ALAYA@GMAIL.COM
Modal'X, UPL, Univ Paris Nanterre,
F92000 Nanterre France

Olga Klopp KLOPPOLGA@MATH.CNRS.FR
ESSEC Business School and CREST,
F95021 Cergy France

Editor: Xiaotong Shen

Abstract

Matrix completion aims to reconstruct a data matrix based on observations of a small number of its entries. Usually in matrix completion a single matrix is considered, which can be, for example, a rating matrix in recommendation system. However, in practical situations, data is often obtained from multiple sources which results in a collection of matrices rather than a single one. In this work, we consider the problem of collective matrix completion with multiple and heterogeneous matrices, which can be count, binary, continuous, etc. We first investigate the setting where, for each source, the matrix entries are sampled from an exponential family distribution. Then, we relax the assumption of exponential family distribution for the noise. In this setting, we do not assume any specific model for the observations. The estimation procedures are based on minimizing the sum of a goodness-of-fit term and the nuclear norm penalization of the whole collective matrix. We prove that the proposed estimators achieve fast rates of convergence under the two considered settings and we corroborate our results with numerical experiments.

Keywords: High-dimensional prediction; Exponential families; Low-rank matrix estimation; Nuclear norm minimization; Low-rank optimization;

1. Introduction

Completing large-scale matrices has recently attracted great interest in machine learning and data mining since it appears in a wide spectrum of applications such as recommender systems (Koren et al., 2009; Bobadilla et al., 2013), collaborative filtering (Netflix challenge) (Goldberg et al., 1992; Rennie and Srebro, 2005), sensor network localization (So and Ye, 2005; Drineas et al., 2006; Oh et al., 2010), system identification (Liu and Vandenberghe, 2009), image processing (Hu et al., 2013), among many others. The basic principle of matrix completion consists in recovering all the entries of an unknown data matrix from incomplete and noisy observations of its entries.

To address the high-dimensionality in matrix completion problem, statistical inference based on low-rank constraint is now an ubiquitous technique for recovering the underlying data matrix. Thus, matrix completion can be formulated as minimizing the rank of the matrix given a random sample of its entries. However, this rank minimization problem is in general NP-hard due to the combinatorial nature of the rank function (Fazel et al., 2001; Fazel, 2002). To alleviate this problem and make it tractable, convex relaxation strategies

were proposed, e.g., the nuclear norm relaxation (Srebro et al., 2005; Candes and Tao, 2010; Recht et al., 2010; Negahban and Wainwright, 2011; Klopp, 2014) or the max-norm relaxation (Cai and Zhou, 2016). Among those surrogate approximations, nuclear norm, which is defined as the sum of the singular values of the matrix or the ℓ_1-norm of its spectrum, is probably the most widely used penalty for low-rank matrix estimation, since it is the tightest convex lower bound of the rank (Fazel et al., 2001).

Motivations. Classical matrix completion focus on a single matrix, whereas in practical situations data is often obtained from a collection of matrices that may cover multiple and heterogeneous sources. For example, in e-commerce users express their feedback for different items such as books, movies, music, etc. In social networks like Facebook and Twitter users often share their opinions and interests on a variety of topics (politics, social events, health). In this examples, informations from multiple sources can be viewed as a collection of matrices coupled through a common set of users.

Rather than exploiting user preference data from each source independently, it may be beneficial to leverage all the available user data provided by various sources in order to generate more encompassing user models (Cantador et al., 2015). For instance, some recommender system runs into the so-called *cold-start problem* (Lam et al., 2008). A user is new or "cold" in a source when he has few to none rated items. Such user may have a rating history in auxiliary sources and we can use his profile in the auxiliary sources to recommend relevant items in the target source. For example, a user's favorite movie genres may be derived from his favorite book genres. Therefore, this shared structure among the sources can be useful to get better predictions (Singh and Gordon, 2008; Bouchard et al., 2013; Gunasekar et al., 2016).

More generally speaking, collective matrix completion finds a natural application in the problem of recommender system with side information. In this problem, in addition to the conventional user-item matrix, it is assumed that we have side information about each user (Chiang et al., 2015; Jain and Dhillon, 2013; Fithian and Mazumder, 2018; Agarwal et al., 2011). For example, in blog recommendation task, we may have access to user generated content (images, tags and text) or user activity (e.g., likes and reblogs). Such side information may be used to improve the quality of recommendation of blogs of interest (Shin and Lee, 2015).

Based on the type of available side information, various methods for recommender systems with side information have been proposed. It can be user generated content (Armentano et al., 2013; Hannon et al., 2010), user/item profile or attribute (Agarwal et al., 2011), social network (Jamali and Ester, 2010; Ma et al., 2011) and context information (Natarajan et al., 2013). A very interesting surveys of the state-of-the-art methods can be found in (Fithian and Mazumder, 2018; Natarajan et al., 2013).

On the other hand, our framework includes the model of Mixed Data Frames with missing observations (Pagès, 2014; Udell et al., 2016). Here matrices collect categorical, numerical and count observations. They appear in numerous applications including in ecology, patient records in health care (Gunasekar et al., 2016), quantitative gene expression values (Natarajan and Dhillon, 2014; Zitnik and Zupan, 2014, 2015), and also in recommender systems and survey data.

Main contributions and related literature. In this paper, we extend the theory of low-rank matrix completion to a collection of multiple and heterogeneous matrices. We first consider general matrix completion setting where we assume that for each matrix its entries are sampled from natural exponential distributions (Lehmann and Casella, 1998). In this setting, we may have Gaussian distribution for continuous data; Bernoulli for binary data; Poisson for count-data, etc. In a second part, we relax the assumption of exponential family distribution for the noise and we do not assume any specific model for the observations. This approach is more popular and widely used in machine learning. The proposed estimation procedure is based on minimizing the sum of a goodness-of-fit term and the nuclear norm penalization of the whole collective matrix. The key challenge in our analysis is to use joint low-rank structure and our algorithm is far from the trivial one which consists in estimating each source matrix separately. We provide theoretical guarantees on our estimation method and show that the collective approach provides faster rate of convergences. We further corroborate our theoretical findings through simulated experiments.

Previous works on collective matrix completion are mainly based on matrix factorization (Srebro et al., 2005). In a nutshell, this approach fits the target matrix as the product of two low-rank matrices. Matrix factorization gives rise to non-convex optimization problems and its theoretical understanding is quite limited. For example, Singh and Gordon (2008) proposed the collective matrix factorization that jointly factorizes multiple matrices sharing latent factors. As in our setting, each matrix can have a different value type and error distribution. In Singh and Gordon (2008), the authors use Bregman divergences to measure the error and extend standard alternating projection algorithms to this setting. They consider a quite general setting which includes as a particular case the nuclear norm penalization approach that we study in the present paper. They do not provide any theoretical guarantee. A Bayesian model for collective matrix factorization was proposed in Singh and Gordon (2010). Horii et al. (2014) and Xu et al. (2016) also consider collective matrix factorization and investigate the strength of the relation among the source matrices. Their estimation procedure is based on penalization by the sum of the nuclear norms of the sources. The convex formulation for collective matrix factorization was proposed in Bouchard et al. (2013) where the authors consider a general situation when the set of matrices do not necessarily have a common set of rows/columns. When this is the case, the estimator proposed in Bouchard et al. (2013) is quite similar to ours. Their algorithm is based on the iterative Singular Value Thresholding and the authors conduct empirical evaluations of this approach on two real data sets.

Most of the previous papers focus on the algorithmic side without providing theoretical guarantees for the collective approach. One exception is the paper by Gunasekar et al. (2015) where the authors prove consistency of the estimate under two observation models: noise-free and additive noise models. Their estimation procedure is based on minimizing the least squares loss penalized by the nuclear norm. To prove the consistency of their estimator, Gunasekar et al. (2015) assume that all the source matrices share the same low-rank factor. They consider the uniform sampling scheme for the observations (see Assumptions 1 and 4 in Gunasekar et al. (2015)). Uniform sampling is an usual assumption in matrix completion literature (see, e.g., (Candes and Tao, 2010; Candès and Recht, 2009; Davenport et al., 2014)). This assumption is restrictive in many applications such as recommendations

systems. The theoretical analysis in the present paper is carried out for general sampling distributions.

Similar to our setting, matrix completion with side information explores the available user data provided by various sources. For instance Jain and Dhillon (2013) and Xu et al. (2013) introduce the so-called Inductive Matrix Completion (IMC). It models side information as knowledge of feature spaces. They show that if the features are perfect (e.,g., see Definition 1 in Chiang et al. (2018) for perfect side information), the sample complexity can be reduced. More precisely, in works on matrix completion with side information, it is usually assumed that one has partially observed low-rank matrix of interest $\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2}$ and, additionally, one has access to two matrices of features $\boldsymbol{A} \in \mathbb{R}^{d_1 \times r_1}$ and $\boldsymbol{B} \in \mathbb{R}^{d_2 \times r_2}$ where each row of $\boldsymbol{A}$ (or $\boldsymbol{B}$) denotes the feature of the i-th row (or column) entity of $\boldsymbol{M}$, $r_i < d_i$ for $i = 1, 2$ and $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{Z}\boldsymbol{B}^T$. The main difference with our setting is that, here, $\boldsymbol{A}$ and $\boldsymbol{B}$ are assumed to be fully observed while our model allows also missing observations for the set of features. The perfect side information assumption is strong and hard to meet in practice. Chiang et al. (2015) relaxed it by assuming that the side information may be noisy (not perfect). In this approach, referred as DirtyIMC, they assume that the unknown matrix is modeled as $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{Z}\boldsymbol{B}^T + \boldsymbol{N}$ where the residual matrix $\boldsymbol{N}$ models imperfections and noise in the features.

Several works consider matrix completion with side information. For example, Chiang et al. (2015) proposes a method based on penalization by the sum of the nuclear norms of $\boldsymbol{M}$ and of each feature. Our method is based on the penalization by the nuclear norm of the whole matrix built of the matrix $\boldsymbol{M}$ and the features $\boldsymbol{A}$ and $\boldsymbol{B}$. In Jain and Dhillon (2013), the authors study the problem of low-rank matrix estimation using rank one measurements. In the noise-free setting, they assume that all the features are known and that the matrices of features are incoherent. The method proposed in Jain and Dhillon (2013) is based on non-convex matrix factorization. In Fithian and Mazumder (2018), the authors consider a general framework for reduced-rank modeling of matrix-valued data. They use a generalized weighted nuclear norm penalty where the matrix is multiplied by positive semidefinite matrices P and Q which depend on the matrix of features. In Agarwal et al. (2011), the authors introduce a per-item user covariate logistic regression model augmenting with user-specific random effects. Their approach is based on a multilevel hierarchical model.

In the case of the heterogeneous data coming from different sources, these approaches can be applied for recovering each source separately. In contrast, our approach aims at collecting all the available information in a single matrix which results in faster rates of convergence. On the other hand, popular algorithms for matrix completion with side information, such as Maxide in Xu et al. (2013) and AltMin in Jain and Dhillon (2013), are based on the least square loss which could be not suitable for data coming from non-Gaussian distributions.

If we consider a single matrix, our model includes as particular case 1-bit matrix completion and, more generally, matrix completion with exponential family noise. 1-bit matrix completion was first studied in Davenport et al. (2014), where the observed entries are assumed to be sampled uniformly at random. This problem was also studied among others by (Cai and Zhou, 2013; Klopp et al., 2015; Alquier et al., 2017). Matrix completion with exponential family noise (for a single matrix) was previously considered in Lafond (2015) and Gunasekar et al. (2014). In these papers authors assume sampling with replacement

where there can be multiple observations for the same entry. In the present paper, we consider more natural setting for matrix completion where each entry may be observed at most once. Our result improves the known results on 1-bit matrix completion and on matrix completion with exponential family noise. In particular, we obtain exact minimax optimal rate of convergence for 1-bit matrix completion and matrix completion with exponential noise which was known up to a logarithmic factor (for more details see Remark 9 in Section 3).

Organization of the paper. The remainder of the paper is organized as follow. In Section 1.1, we introduce basic notation and definitions. Section 2 sets up the formalism for the collective matrix completion. In Section 3, we investigate the exponential family noise model. In Section 4, we study distribution-free setup and we provide the upper bound on the excess risk. To verify the theoretical findings, we corroborate our results with numerical experiments in Section 5, where we present an efficient iterative algorithm that solves the maximum likelihood approximately. The proofs of the main results and key technical lemmas are postponed to the appendices.

1.1. Preliminaries

For the reader's convenience, we provide a brief summary of the standard notation and the definitions that will be frequently used throughout the paper.

Notation. For any positive integer m, we use $[m]$ to denote $\{1, \dots, m\}$. We use capital bold symbols such as $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{A}$, to denote matrices. For a matrix $\boldsymbol{A}$, we denote its (i,j)-th entry by A_{ij}. As usual, let $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ be the Frobenius norm and let $\|\boldsymbol{A}\|_\infty = \max_{i,j} |A_{ij}|$ denote the elementwise ℓ_∞-norm. Additionally, $\|\boldsymbol{A}\|_*$ stands for the nuclear norm (trace norm), that is $\|\boldsymbol{A}\|_* = \sum_i \sigma_i(\boldsymbol{A})$ where $\sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots$ are singular values of $\boldsymbol{A}$, and $\|\boldsymbol{A}\| = \sigma_1(\boldsymbol{A})$ to denote the operator norm. The inner product between two matrices is denoted by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \sum_{ij} A_{ij} B_{ij}$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. We write $\partial\Psi$ the subdifferential mapping of a convex functional Ψ. Given two real numbers a and b, we write $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. The symbols $\mathbb{P}$ and $\mathbb{E}$ denote generic probability and expectation operators whose distribution is determined from the context. The notation c will be used to denote positive constant, that might change from one instance to the other.

Definition 1 *A distribution of a random variable X is said to belong to the natural exponential family, if its probability density function characterized by the parameter η is given by:*

$$X|\eta \sim f_{h,G}(x|\eta) = h(x) \exp\big(\eta x - G(\eta)\big),$$

where h is a nonnegative function, called the base measure function, which is independent of the parameter η. The function $G(\eta)$ is strictly convex, and is called the log*-partition function, or the cumulant function. This function uniquely defines a particular member distribution of the exponential family, and can be computed as: $G(\eta) = \log\big(\int h(x) \exp(\eta x) dx\big)$.*

If G is smooth enough, we have that $\mathbb{E}[X] = G'(\eta)$ and $\mathbb{V}ar[X] = G''(\eta)$, where G' stands for the derivative of G. The exponential family encompasses a wide large of standard distributions such as:

- Normal, $\mathcal{N}(\mu, \sigma^2)$ (known σ), is typically used to model continuous data, with natural parameter $\eta = \frac{\mu}{\sigma^2}$ and $G(\eta) = \frac{\sigma^2}{2}\eta^2$.
- Gamma, $\Gamma(\lambda, \alpha)$ (known α), is often used to model positive valued continuous data, with natural parameter $\eta = -\lambda$ and $G(\eta) = -\alpha \log(-\eta)$.
- Negative binomial, $\mathcal{NB}(p, r)$ (known r), is a popular distribution to model overdispersed count data, whose variance is larger than their mean, with natural parameter $\eta = \log(1-p)$ and $G(\eta) = -r\log(1-\exp(\eta))$.
- Binomial, $\mathcal{B}(p, N)$ (known N), is used to model number of successes in N trials, with natural parameter $\eta = \log(\frac{p}{1-p})$ (logit function) and $G(\eta) = N\log(1+\exp(\eta))$.
- Poisson, $\mathcal{P}(\lambda)$, is used to model count data, with natural parameter $\eta = \log(\lambda)$ and $G(\eta) = \exp(\eta)$.

Exponential, chi-squared, Rayleigh, Bernoulli and geometric distributions are special cases of the above five distributions.

Definition 2 *Let S be a closed convex subset of $\mathbb{R}^m$ and $\Phi : S \subset \boldsymbol{dom}(\Phi) \to \mathbb{R}$ a continuously-differentiable and strictly convex function. The Bregman divergence associated with Φ (Bregman, 1967; Censor and Zenios, 1997) $d_\Phi : S \times S \to [0, \infty)$ is defined as*

$$d_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle x - y, \nabla\Phi(y)\rangle,$$

where $\nabla\Phi(y)$ represents the gradient vector of Φ evaluated at y.

The value of the Bregman divergence $d_\Phi(x, y)$ can be viewed as the difference between the value of Φ at x and the first Taylor expansion of Φ around y evaluated at point x. For exponential family distributions, the Bregman divergence corresponds to the Kullback-Leibler divergence (Banerjee et al., 2005) with $\Phi = G$.

2. Collective matrix completion

Assume that we observe a collection of matrices $\boldsymbol{\mathcal{X}} = (\boldsymbol{X}^1, \ldots, \boldsymbol{X}^V)$. In this collection components $\boldsymbol{X}^v \in \mathbb{R}^{d_u \times d_v}$ have a common set of rows. This common set of rows corresponds, for example, to a common set of users in a recommendation system. The set of columns of each matrix $\boldsymbol{X}^v$ corresponds to a different type of entity. In the case of recommender system it can be books, films, video game, etc. Then, the entries of each matrix $\boldsymbol{X}^v$ corresponds to the user's rankings for this particular type of products.

We assume that the distribution of each matrix $\boldsymbol{X}^v$ depends on the matrix of parameters $\boldsymbol{M}^v$. This distribution can be different for different v. For instance, we can have binary observations for one matrix $\boldsymbol{X}^{v_1}$ with entries which correspond, for example, to like/dislike labels for a certain type of products, multinomial for another matrix $\boldsymbol{X}^{v_2}$ with ranking going from 1 to 5 and Gaussian for a third matrix $\boldsymbol{X}^{v_3}$.

As it happens in many applications, we assume that for each matrix $\boldsymbol{X}^v$ we observe only a small subset of its entries. We consider the following model: for $v \in [V]$ and $(i, j) \in [d_u] \times [d_v]$, let B_{ij}^v be independent Bernoulli random variables with parameter π_{ij}^v.

We suppose that B_{ij}^v are independent from X_{ij}^v. Then, we observe $Y_{ij}^v = B_{ij}^v X_{ij}^v$. We can think of the B_{ij}^v as masked variables. If $B_{ij}^v = 1$, we observe the corresponding entry of $\boldsymbol{X}^v$, and when $B_{ij}^v = 0$, we have a missing observation.

In the simplest situation each coefficient is observed with the same probability, i.e. for every $v \in [V]$ and $(i,j) \in [d_u] \times [d_v], \pi_{ij}^v = \pi$. In many practical applications, this assumption is not realistic. For example, for a recommendation system, some users are more active than others and some items are more popular than others and thus rated more frequently. Hence, the sampling distribution is in fact non-uniform. In the present paper, we consider general sampling model where we only assume that each entry is observed with a positive probability:

Assumption 1 *Assume that there exists a positive constant* $0 < p < 1$ *such that*

$$\min_{v\in[V]} \min_{(i,j)\in[d_u]\times[d_v]} \pi_{ij}^v \geq p.$$

Let Π denotes the joint distribution of the Bernoulli variables $\big\{B_{ij}^v : (i,j) \in [d_u] \times [d_v], v \in [V]\big\}$. For any matrix $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_u \times D}$ where $D = \sum_{v\in[V]} d_v$, we define the weighted Frobenius norm

$$\|\boldsymbol{\mathcal{A}}\|_{\Pi,F}^2 = \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \pi_{ij}^v (A_{ij}^v)^2.$$

Assumption 1 implies $\|\boldsymbol{\mathcal{A}}\|_{\Pi,F}^2 \geq p\|\boldsymbol{\mathcal{A}}\|_F^2$. For each $v \in [V]$ let us denote $\pi_{i\bullet}^v = \sum_{j=1}^{d_v} \pi_{ij}^v$ and $\pi_{\bullet j}^v = \sum_{i=1}^{d_u} \pi_{ij}^v$. Note we can easily get an estimations of $\pi_{i\bullet}^v$ and $\pi_{\bullet j}^v$ using the empirical frequencies:

$$\widehat{\pi_{i\bullet}^v} = \sum_{j\in[d_v]} B_{ij}^v \quad \text{and} \quad \widehat{\pi_{\bullet j}^v} = \sum_{i\in[d_u]} B_{ij}^v.$$

Let $\pi_{i\bullet} = \sum_{v\in[V]} \pi_{i\bullet}^v$, $\pi_{\bullet j} = \max_{v\in[V]} \pi_{\bullet j}^v$, and μ be an upper bound of its maximum, that is

$$\max_{(i,j)\in[d_u]\times[d_v]} (\pi_{i\bullet}, \pi_{\bullet j}) \leq \mu. \tag{1}$$

3. Exponential family noise

In this section we assume that for each v distribution of $\boldsymbol{X}^v$ belongs to the exponential family, that is

$$X_{ij}^v | M_{ij}^v \sim f_{h^v,G^v}(X_{ij}^v | M_{ij}^v) = h^v(X_{ij}^v) \exp\big(X_{ij}^v M_{ij}^v - G^v(M_{ij}^v)\big).$$

We denote $\boldsymbol{\mathcal{M}} = (\boldsymbol{M}^1, \ldots, \boldsymbol{M}^V)$ and let γ be an upper bound on the sup-norm of $\boldsymbol{\mathcal{M}}$, that is $\gamma = |\gamma_1| \vee |\gamma_2|$, where $\gamma_1 \leq M_{ij}^v \leq \gamma_2$ for every $v \in [V]$ and $(i,j) \in [d_u] \times [d_v]$. Hereafter, we denote by $\mathscr{C}_\infty(\gamma) = \big\{\boldsymbol{\mathcal{W}} \in \mathbb{R}^{d_u \times D} : \|\boldsymbol{\mathcal{W}}\|_\infty \leq \gamma\big\}$, the ℓ_∞-norm ball with radius γ in the space $\mathbb{R}^{d_u \times D}$. We need the following assumptions on densities f_{h^v,G^v}:

Assumption 2 *For each $v \in [V]$, we assume that the function $G^v(\cdot)$ is twice differentiable and there exits two constants L^2_γ, U^2_γ satisfying:*

$$\sup_{\eta \in [-\gamma - \frac{1}{K}, \gamma + \frac{1}{K}]} (G^v)''(\eta) \leq U^2_\gamma, \tag{2}$$

and

$$\inf_{\eta \in [-\gamma - \frac{1}{K}, \gamma + \frac{1}{K}]} (G^v)''(\eta) \geq L^2_\gamma, \tag{3}$$

for some $K > 0$.

The first statement, (2), in Assumption 2 ensures that the distributions of X^v_{ij} have uniformly bounded variances and sub-exponential tails (see Lemma 28 in Appendix C). The second one, (3), is the strong convexity condition satisfied by the log-partition function G^v. This assumption is satisfied for most standard distributions presented in the previous section. In Table 1, we list the corresponding constants in Assumption 2.

Model	$(G^v)'(\eta)$	$(G^v)''(\eta)$	L^2_γ	U^2_γ
Normal	$\sigma^2\eta$	σ^2	σ^2	σ^2
Binomial	$\frac{Ne^\eta}{1+e^\eta}$	$\frac{Ne^\eta}{(1+e^\eta)^2}$	$\frac{Ne^{-(\gamma+\frac{1}{K})}}{(1+e^{\gamma+\frac{1}{K}})^2}$	$\frac{N}{4}$
Gamma (if $\gamma_1\gamma_2 > 0$)	$-\frac{\alpha}{\eta}$	$\frac{\alpha}{\eta^2}$	$\frac{\alpha}{(\gamma+\frac{1}{K})^2}$	$\frac{\alpha}{(\lvert\gamma_1\rvert\wedge\lvert\gamma_2\rvert)^2}$
Negative binomial	$\frac{re^\eta}{1-e^\eta}$	$\frac{re^\eta}{(1-e^\eta)^2}$	$\frac{re^{-(\gamma+\frac{1}{K})}}{(1-e^{-(\gamma+\frac{1}{K})})^2}$	$\frac{re^{(\gamma+\frac{1}{K})}}{(1-e^{\gamma+\frac{1}{K}})^2}$
Poisson	e^η	e^η	$e^{-(\gamma+\frac{1}{K})}$	$e^{(\gamma+\frac{1}{K})}$

Table 1: Examples of the corresponding constants L^2_γ and U^2_γ from Assumption 2.

3.1. Estimation procedure

To estimate the collection of matrices of parameters $\boldsymbol{\mathcal{M}} = (\boldsymbol{M}^1, \ldots, \boldsymbol{M}^V)$, we use penalized negative log-likelihood. Let $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{d_u \times D}$, we divide it in V blocks $\boldsymbol{W}^v \in \mathbb{R}^{d_u \times d_v}$: $\boldsymbol{\mathcal{W}} = (\boldsymbol{W}^1, \ldots, \boldsymbol{W}^V)$. Given observations $\boldsymbol{\mathcal{Y}} = (\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^V)$, we write the negative log-likelihood as

$$\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}) = -\frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B^v_{ij} \big(Y^v_{ij} W^v_{ij} - G^v(W^v_{ij}) \big).$$

The nuclear norm penalized estimator $\widehat{\boldsymbol{\mathcal{M}}}$ of $\boldsymbol{\mathcal{M}}$ is defined as follows:

$$\widehat{\boldsymbol{\mathcal{M}}} = (\widehat{\boldsymbol{M}}^1, \ldots, \widehat{\boldsymbol{M}}^V) = \operatorname*{argmin}_{\boldsymbol{\mathcal{W}} \in \mathscr{C}_\infty(\gamma)} \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}) + \lambda \|\boldsymbol{\mathcal{W}}\|_*, \tag{4}$$

where λ is a positive regularization parameter that balances the trade-off between model fit and privileging a low-rank solution. Namely, for large value of λ the rank of the estimator $\widehat{\boldsymbol{\mathcal{M}}}$ is expected to be small.

Let the collection of matrices $(E^v_{11}, \ldots, E^v_{d_u d_v})$ form the canonical basis in the space of matrices of size $d_u \times d_v$. The entry of (E^v_{ij}) is 0 everywhere except for the (i,j)-th entry

where it equals to 1. For (ε_{ij}^v), an *i.i.d* Rademacher sequence, we define $\mathbf{\Sigma}_R = (\mathbf{\Sigma}_R^1, \dots, \mathbf{\Sigma}_R^V)$ where for all $v \in [V]$

$$\mathbf{\Sigma}_R^v = \frac{1}{d_u D} \sum_{(i,j)\in[d_u]\times[d_v]} \varepsilon_{ij}^v B_{ij}^v E_{ij}^v.$$

We now state the main result concerning the recovery of $\boldsymbol{\mathcal{M}}$. Theorem 3 gives a general upper bound on the estimation error of $\widehat{\boldsymbol{\mathcal{M}}}$ defined by (4). Its proof is postponed in Appendix A.1.

Theorem 3 *Assume that Assumptions 1 and 2 hold, and* $\lambda \geq 2\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\|$. *Then, with probability exceeding* $1 - 4/(d_u + D)$ *we have*

$$\frac{1}{d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_{\Pi,F}^2 \leq \frac{c}{p} \max\Big\{ d_u D \,\mathrm{rank}(\boldsymbol{\mathcal{M}}) \Big(\frac{\lambda^2}{L_\gamma^4} + \gamma^2 (\mathbb{E}[\|\mathbf{\Sigma}_R\|])^2 \Big), \frac{\gamma^2 \log(d_u + D)}{d_u D} \Big\},$$

where c *is a numerical constant.*

Using Assumption 1, Theorem 3 implies the following bound on the estimation error measured in normalized Frobenius norm.

Corollary 4 *Under assumptions of Theorem 3 and with probability exceeding* $1-4/(d_u+D)$, *we have*

$$\frac{1}{d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_{F}^2 \leq \frac{c}{p^2} \max\Big\{ d_u D \,\mathrm{rank}(\boldsymbol{\mathcal{M}}) \Big(\frac{\lambda^2}{L_\gamma^4} + \gamma^2 (\mathbb{E}[\|\mathbf{\Sigma}_R\|])^2 \Big), \frac{\gamma^2 \log(d_u + D)}{d_u D} \Big\}.$$

In order to get a bound in a closed form we need to obtain a suitable upper bounds on $\mathbb{E}[\|\mathbf{\Sigma}_R\|]$ and on $\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\|$ with high probability. Therefore we use the following two lemmas.

Lemma 5 *There exists an absolute constant* c *such that*

$$\mathbb{E}[\|\mathbf{\Sigma}_R\|] \leq c\Big(\frac{\sqrt{\mu} + \sqrt{\log(d_u \wedge D)}}{d_u D} \Big).$$

Lemma 6 *Let Assumption 2 holds. Then, there exists an absolute constant* c *such that, with probability at least* $1 - 4/(d_u + D)$, *we have*

$$\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\| \leq c\Bigg(\frac{(U_\gamma \vee K)\big(\sqrt{\mu} + (\log(d_u \vee D))^{3/2}\big)}{d_u D} \Bigg).$$

The proofs of Lemmas 5 and 6 are postponed to Appendices A.2 and A.3. Recall that the condition on λ in Theorem 3 is that $\lambda \geq 2\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\|$. Using Lemma 6, we can choose

$$\lambda = 2c \frac{(U_\gamma \vee K)\big(\sqrt{\mu} + (\log(d_u \vee D))^{3/2}\big)}{d_u D}.$$

With this choice of λ, we obtain the following theorem:

Theorem 7 *Let Assumptions 1 and 2 be satisfied. Then, with probability exceeding* $1 - 4/(d_u + D)$ *we have*

$$\frac{1}{d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \leq \frac{c\,\mathrm{rank}(\boldsymbol{\mathcal{M}})}{p d_u D}\Big(\gamma^2 + \frac{(U_\gamma \vee K)^2}{L_\gamma^4}\Big)\big(\mu + \log^3(d_u \vee D)\big),$$

and

$$\frac{1}{d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{F} \leq \frac{c\,\mathrm{rank}(\boldsymbol{\mathcal{M}})}{p^2 d_u D}\Big(\gamma^2 + \frac{(U_\gamma \vee K)^2}{L_\gamma^4}\Big)\big(\mu + \log^3(d_u \vee D)\big),$$

where c *is an absolute constant.*

Remark 8 *Note that the rate of convergence in Theorem 7 has the following dominant term:*

$$\frac{1}{d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{F} \lesssim \frac{\mathrm{rank}(\boldsymbol{\mathcal{M}})\mu}{p^2 d_u D},$$

where the symbol $\lesssim$ *means that the inequality holds up to a multiplicative constant. If we assume that the sampling distribution is close to the uniform one, that is that there exists positive constants* c_1 *and* c_2 *such that for every* $v \in [V]$ *and* $(i,j) \in [d_u] \times [d_v]$ *we have* $c_1 p \leq \pi^v_{ij} \leq c_2 p$*, then Theorem 7 yields*

$$\frac{1}{d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{F} \lesssim \frac{\mathrm{rank}(\boldsymbol{\mathcal{M}})}{p(d_u \wedge D)}.$$

If we complete each matrix separately, the error will be of the order $\sum_{v=1}^V \mathrm{rank}(\boldsymbol{M}^v)/p(d_u \wedge D)$*. As* $\mathrm{rank}(\boldsymbol{\mathcal{M}}) \leq \sum_{v=1}^V \mathrm{rank}(\boldsymbol{M}^v)$*, the rate of convergence achieved by our estimator is faster compared to the penalization by the sum-nuclear-norm.*

In order to get a small estimation error, p *should be larger than* $\mathrm{rank}(\boldsymbol{\mathcal{M}})/(d_u \wedge D)$*. We denote* $n = \sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} \pi^v_{ij}$*, the expected number of observations. Then, we get the following condition on* n*:*

$$n \geq c\,\mathrm{rank}(\boldsymbol{\mathcal{M}})(d_u \vee D).$$

Remark 9 *In 1-bit matrix completion (Davenport et al., 2014; Klopp et al., 2015; Alquier et al., 2017), instead of observing the actual entries of the unknown matrix* $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{d\times D}$*, for a random subset of its entries* Ω *we observe* $\{Y_{ij} \in \{+1,-1\} : (i,j) \in \Omega\}$*, where* $Y_{ij} = 1$ *with probability* $f(M_{ij})$ *for some link-function* f*. In Davenport et al. (2014) the parameter* $\boldsymbol{\mathcal{M}}$ *is estimated by minimizing the negative log-likelihood under the constraints* $\|\boldsymbol{\mathcal{M}}\|_\infty \leq \gamma$ *and* $\|\boldsymbol{\mathcal{M}}\|_* \leq \gamma\sqrt{rdD}$ *for some* $r > 0$*. Under the assumption that* $\mathrm{rank}(\boldsymbol{\mathcal{M}}) \leq r$*, the authors prove that*

$$\frac{1}{dD}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{F} \leq c_\gamma\sqrt{\frac{r(d \vee D)}{n}}, \tag{5}$$

where c_γ *is a constant depending on* γ *(see Theorem 1 in Davenport et al. (2014)). A similar result using max-norm minimization was obtained in Cai and Zhou (2013). In (Klopp et al.,*

2015) the authors prove a faster rate. Their upper bound (see Corollary 2 in Klopp et al. (2015)) is given by

$$\frac{1}{dD}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_F^2 \leq c_\gamma \frac{\operatorname{rank}(\boldsymbol{\mathcal{M}})(d \vee D)\log(d \vee D)}{n}. \tag{6}$$

In the particular case of 1*-bit matrix completion for a single matrix under uniform sampling scheme, Theorem 7 implies the following bound:*

$$\frac{1}{dD}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_F^2 \leq c_\gamma \frac{\operatorname{rank}(\boldsymbol{\mathcal{M}})(d \vee D)}{n},$$

which improves (6) *by a logarithmic factor. Furthermore, Klopp et al. (2015) provide* $\operatorname{rank}(M)(d \vee D)/n$ *as the lower bound for* 1*-bit matrix completion (see Theorem 3 in Klopp et al. (2015)). So our result answers the important theoretical question what is the exact minimax rate of convergence for* 1*-bit matrix completion which was previously known up to a logarithmic factor.*

In a more general setting of matrix completion with exponential family noise, the minimax optimal rate of convergence was also known only up to logarithmic factor (see Lafond (2015)). Our result provides the exact minimax optimal rate in this more general setting too. It is easy to see, by inspection of the proof of the lower bound in Lafond (2015), that the upper bound provided by Theorem 7 is optimal for the collective matrix completion.

Remark 10 *Note that our estimation method is based on the minimization of the nuclear-norm of the whole collective matrix* $\boldsymbol{\mathcal{M}}$*. Another possibility is to penalize by the sum of the nuclear norms* $\sum_{v\in[V]} \|\boldsymbol{M}^v\|_*$ *(see, e.g., Klopp et al. (2015)). This approach consists in estimating each component matrix independently.*

4. General losses

In the previous section we assume that the link functions G^v are known. This assumption is not realistic in many applications. In this section we relax this assumption in the sense that we do not assume any specific model for the observations. Recall that our observations are a collection of partially observed matrices $\boldsymbol{Y}^v = (B^v_{ij}X^v_{i,j}) \in \mathbb{R}^{d_u\times d_v}$ for $v = 1,\dots,V$ and $\boldsymbol{X}^v = (X^v_{ij}) \in \mathbb{R}^{d_u\times d_v}$. We are interested in the problem of prediction of the entries of the collective matrix $\boldsymbol{\mathcal{X}} = (\boldsymbol{X}^1,\dots,\boldsymbol{X}^V)$. We consider the risk of estimating $\boldsymbol{X}^v$ with a loss function ℓ^v, which measures the discrepancy between the predicted and actual value with respect to the given observations. We focus on non-negative convex loss functions that are Lipschitz:

Assumption 3 *(Lipschitz loss function) For every* $v \in [V]$*, we assume that the loss function* $\ell^v(y,\cdot)$ *is* ρ_v*-Lipschitz in its second argument:* $|\ell^v(y,x) - \ell^v(y,x')| \leq \rho_v|x - x'|$.

Some examples of the loss functions that are 1-Lipschitz are: hinge loss $\ell(y,y') = \max(0, 1 - yy')$, logistic loss $\ell(y,y') = \log(1 + \exp(-yy'))$, and quantile regression loss $\ell(y,y') = \ell_\tau(y' - y)$ where $\tau \in (0,1)$ and $\ell_\tau(z) = z(\tau - \mathbb{1}(z \leq 0))$.

For a matrix $\boldsymbol{\mathcal{M}} = (\boldsymbol{M}^1, \ldots, \boldsymbol{M}^V) \in \mathbb{R}^{d_u \times D}$, we define the empirical risk as

$$R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}}) = \frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B_{ij}^v \ell^v(Y_{ij}^v, M_{ij}^v).$$

We define the oracle as:

$$\overset{\star}{\boldsymbol{\mathcal{M}}} = \big(\overset{\star}{\boldsymbol{M}}^1, \ldots, \overset{\star}{\boldsymbol{M}}^V\big) = \underset{\boldsymbol{\mathcal{Q}} \in \mathscr{C}_\infty(\gamma)}{\operatorname{argmin}} R(\boldsymbol{\mathcal{Q}}) \tag{7}$$

where $R(\boldsymbol{\mathcal{Q}}) = \mathbb{E}[R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}})]$. Here the expectation is taken over the joint distribution of $\{(Y_{ij}^v, B_{ij}^v) : (i,j) \in [d_u] \times [d_v] \text{ and } v \in [V]\}$. We use machine learning approach and will provide an estimator $\widehat{\boldsymbol{\mathcal{M}}}$ that predicts almost as well as $\overset{\star}{\boldsymbol{\mathcal{M}}}$. Thus we will consider excess risk $R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})$. By construction, the excess risk is always positive.

For a tuning parameter $\Lambda > 0$, the nuclear norm penalized estimator $\widehat{\boldsymbol{\mathcal{M}}}$ is defined as

$$\widehat{\boldsymbol{\mathcal{M}}} \in \underset{\boldsymbol{\mathcal{Q}} \in \mathscr{C}_\infty(\gamma)}{\operatorname{argmin}} \big\{ R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) + \Lambda \|\boldsymbol{\mathcal{Q}}\|_* \big\}. \tag{8}$$

We next turn to the assumption needed to establish an upper bound on the performance of the estimator $\widehat{\boldsymbol{\mathcal{M}}}$ defined in (8).

Assumption 4 *Assume that there exists a constant $\varsigma > 0$ such that for every $\boldsymbol{\mathcal{Q}} \in \mathscr{C}_\infty(\gamma)$, we have*

$$R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \geq \frac{\varsigma}{d_u D} \|\boldsymbol{\mathcal{Q}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_{\Pi,F}^2.$$

This assumption has been extensively studied in the learning theory literature (Mendelson, 2008; Zhang, 2004; Bartlett et al., 2004; Alquier et al., 2017; Elsener and van de Geer, 2018), and it is called "Bernstein" condition. It is satisfied in various cases of loss function (Alquier et al., 2017) and it ensures a sufficient convexity of the risk around the oracle defined in (7). Note that when the loss function ℓ^v is strongly convex, the risk function inherits this property and automatically satisfies the margin condition. In other cases, this condition requires strong assumptions on the distribution of the observations, for instance for hinge loss or quantile loss (see Section 6 in Alquier et al. (2017)). The following result gives an upper bound on the excess risk of the estimator $\widehat{\boldsymbol{\mathcal{M}}}$.

Theorem 11 *Let Assumptions 1, 3 and 4 hold and set $\rho = \max_{v \in [V]} \rho_v$. Suppose that $\Lambda \geq 2 \sup\{\|\boldsymbol{\mathcal{G}}\| : \boldsymbol{\mathcal{G}} \in \partial R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\}$. Then, with probability at least $1 - 4/(d_u + D)$, we have*

$$R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \leq \frac{c}{p} \max \Big\{ \operatorname{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}}) d_u D \Big(\rho^{3/2} \sqrt{\gamma/\varsigma} (\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2 + \frac{\Lambda^2}{\varsigma} \Big), \\ \frac{\big(\rho\gamma + \rho^{3/2}\sqrt{\gamma/\varsigma}\big) \log(d_u + D)}{d_u D} \Big\}.$$

Theorem 11 gives a general upper bound on the prediction error of the estimator $\widehat{\boldsymbol{\mathcal{M}}}$. Its proof is presented in Appendix A.4. In order to get a bound in a closed form we need to obtain a suitable upper bounds on $\sup\{\|\boldsymbol{\mathcal{G}}\| : \boldsymbol{\mathcal{G}} \in \partial(R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}}))\}$ with high probability.

Lemma 12 *Let Assumption 3 holds. Then, there exists an absolute constant c such that, with probability at least $1 - 4/(d_u + D)$, we have*

$$\|\mathcal{G}\| \leq c\frac{\rho\big(\sqrt{\mu} + \sqrt{\log(d_u \vee D)}\big)}{d_u D},$$

for all $\mathcal{G} \in \partial R_{\mathcal{Y}}(\overset{\star}{\mathcal{M}})$.

The proof of Lemma 12 is given in Appendix A.5. Using Lemma 12 , we can choose

$$\Lambda = 2c\frac{\rho\big(\sqrt{\mu} + \sqrt{\log(d_u \vee D)}\big)}{d_u D}$$

and with this choice of Λ and Lemma 5, we obtain the following theorem:

Theorem 13 *Let Assumptions 1, 3 and 4 hold. Then, we have*

$$R(\widehat{\mathcal{M}}) - R(\overset{\star}{\mathcal{M}}) \leq \frac{c}{p}\operatorname{rank}(\overset{\star}{\mathcal{M}})\frac{(\rho^2 + \rho^{3/2}\sqrt{\gamma/\varsigma})(\mu + \log(d_u \vee D))}{d_u D},$$

with probability at least $1 - 4/(d_u + D)$.

Using Assumption 4, we get the following corollary:

Corollary 14 *With probability at least $1 - 4/(d_u + D)$, we have*

$$\frac{1}{d_u D}\|\widehat{\mathcal{M}} - \overset{\star}{\mathcal{M}}\|_F^2 \leq \frac{c}{p^2\varsigma}\operatorname{rank}(\overset{\star}{\mathcal{M}})\frac{(\rho^2 + \rho^{3/2}\sqrt{\gamma/\varsigma})(\mu + \log(d_u \vee D))}{d_u D}.$$

1-bit matrix completion. In 1-bit matrix completion with logistic (resp. hinge) loss, the Bernstein assumption is satisfied with $\varsigma = 1/(4e^{2\gamma})$ (resp. $\varsigma = 2\tau$, for some τ that verifies $|\overset{\star}{M}{}^v_{ij} - 1/2| \geq \tau, \forall v \in [V], (i,j) \in [d_u] \times [d_v]$). More details for these constants can be found in Propositions 6.1 and 6.3 in Alquier et al. (2017). Then, the excess risk with respect to these two losses under the uniform sampling is given by:

Corollary 15 *With probability at least $1 - 4/(d_u + D)$, we have*

$$R(\widehat{\mathcal{M}}) - R(\overset{\star}{\mathcal{M}}) \leq c\frac{\operatorname{rank}(\overset{\star}{\mathcal{M}})}{p(d_u \wedge D)}.$$

These results are obtained without a logarithmic factor, and it improves the ones given in Theorems 4.2 and 4.4 in Alquier et al. (2017). The natural loss in this context is the 0/1 loss which is often replaced by the hinge or the logistic loss. We assume without loss of generality that $\gamma = 1$, since the Bayes classifier has its entries in $[-1, 1]$, and we define the classification excess risk by:

$$R_{0/1}(\mathcal{M}) = \frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} \pi^v_{ij}\mathbb{P}[X^v_{ij} \neq \operatorname{sign}(M^v_{ij})],$$

for all $\mathcal{M} \in \mathbb{R}^{d_u \times D}$. Using Theorem 2.1 in Zhang (2004), we have

$$R_{0/1}(\widehat{\mathcal{M}}) - R_{0/1}(\overset{\star}{\mathcal{M}}) \leq c\sqrt{\frac{\operatorname{rank}(\overset{\star}{\mathcal{M}})}{p(d_u \wedge D)}}.$$

5. Numerical experiments

In this section, we first provide algorithmic details of the numerical procedure for solving the problem (4), then we conduct experiments on synthetic data to further illustrate the theoretical results of the collective matrix completion.

5.1. Algorithm

The collective matrix completion problem (4) is a semidefinite program (SDP), since it is a nuclear norm minimization problem with a convex feasible domain (Fazel et al., 2001; Srebro et al., 2005). We may solve it, for example, via the interior-point method (Liu and Vandenberghe, 2010). However, SDP solvers can handle a moderate dimensions, thus such formulation is not scalable due to the storage and computation complexity in low-rank matrix completion tasks. In the following, we present an algorithm that solves the problem (4) approximately and in a more efficient way than solving it as SDP.

Proximal Gradient. Problem (4) can be solved by first-order optimization methods such as proximal gradient (PG) which has been popularly used for optimizations problems of the form of (4) (Beck and Teboulle, 2009; Nesterov, 2013; Parikh and Boyd, 2014; Ji and Ye, 2009b; Mazumder et al., 2010; Yao and Kwok, 2015). When $\mathscr{L}_{\boldsymbol{\mathcal{Y}}}$ has L-Lipschitz continuous gradient, that is $\|\nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}) - \nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}})\|_F \leq L\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{Q}}\|_F$, the PG generates a sequence of estimates $\{\boldsymbol{\mathcal{W}}_t\}$ as

$$\begin{aligned}\boldsymbol{\mathcal{W}}_{t+1} &= \operatorname*{argmin}_{\boldsymbol{\mathcal{W}}} \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}) + (\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{W}}_t)^\top \nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}_t) + \frac{L}{2}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{W}}_t\|_F^2 + \lambda\|\boldsymbol{\mathcal{W}}\|_* \\ &= \operatorname{prox}_{\frac{\lambda}{L}\|\cdot\|_*}(\boldsymbol{\mathcal{Z}}_t), \text{ where } \boldsymbol{\mathcal{Z}}_t = \boldsymbol{\mathcal{W}}_t - \frac{1}{L}\nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}_t) \end{aligned} \tag{9}$$

and for any convex function $\Psi : \mathbb{R}^{d_u \times D} \mapsto \mathbb{R}$, the associated proximal operator at $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{d_u \times D}$ is defined as

$$\operatorname{prox}_{\Psi}(\boldsymbol{\mathcal{W}}) = \operatorname{argmin}\big\{\frac{1}{2}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{Q}}\|_F^2 + \Psi(\boldsymbol{\mathcal{Q}}) : \boldsymbol{\mathcal{Q}} \in \mathbb{R}^{d_u \times D}\big\}.$$

The proximal operator of the nuclear norm at $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{d_u \times D}$ corresponds to the singular value thresholding (SVT) operator of $\boldsymbol{\mathcal{W}}$ (Cai. et al., 2010). That is, assuming a singular value decomposition $\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{U}}\boldsymbol{\Sigma}\boldsymbol{\mathcal{V}}^\top$, where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{d_u \times r}$, $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{D \times r}$ have orthonormal columns, $\boldsymbol{\Sigma} = (\sigma_1, \ldots, \sigma_r)$, with $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $r = \operatorname{rank}(\boldsymbol{\mathcal{W}})$, we have

$$\operatorname{SVT}_{\lambda/L}(\boldsymbol{\mathcal{W}}) = \boldsymbol{\mathcal{U}}\operatorname{diag}((\sigma_1 - \lambda/L)_+, \ldots, (\sigma_r - \lambda/L)_+)\boldsymbol{\mathcal{V}}^\top, \tag{10}$$

where $(a)_+ = \max(a, 0)$.

Although PG can be implemented easily, it converges slowly when the Lipschitz constant L is large. In such scenarios, the rate is $\mathcal{O}(1/T)$, where T is the number of iterations (Parikh and Boyd, 2014). Nevertheless, it can be accelerated by replacing $\boldsymbol{\mathcal{Z}}_t$ in (9) with

$$\boldsymbol{\mathcal{Q}}_t = (1 + \theta_t)\boldsymbol{\mathcal{W}}_t - \theta_t\boldsymbol{\mathcal{W}}_{t-1}, \quad \boldsymbol{\mathcal{Z}}_t = \boldsymbol{\mathcal{Q}}_t - \eta\nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}_t). \tag{11}$$

Several choices for θ_t can be used. The resultant accelerated proximal gradient (APG) (see Algorithm 1) converges with the optimal $\mathcal{O}(1/T^2)$ rate (Nesterov, 2013; Ji and Ye, 2009a).

Algorithm 1: APG for Collective Matrix Completion

1. **initialize:** $\boldsymbol{\mathcal{W}}_0 = \boldsymbol{\mathcal{W}}_1 = \boldsymbol{\mathcal{Y}}$, and $\alpha_0 = \alpha_1 = 1$.
2. **for** $t = 1, \ldots, T$ **do**
3. $\quad \boldsymbol{\mathcal{Q}}_t = \boldsymbol{\mathcal{W}}_t + \frac{\alpha_{t-1}-1}{\alpha_t}(\boldsymbol{\mathcal{W}}_t - \boldsymbol{\mathcal{W}}_{t-1})$;
4. $\quad \boldsymbol{\mathcal{W}}_{t+1} = \mathrm{SVT}_{\frac{\lambda}{L}}(\boldsymbol{\mathcal{Q}}_t - \frac{1}{L}\nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}_t))$;
5. $\quad \alpha_{t+1} = \frac{1}{2}(\sqrt{4\alpha_t^2 + 1} + 1)$;
6. **return** $\boldsymbol{\mathcal{W}}_{T+1}$.

Approximate SVT (Yao and Kwok, 2015). To compute $\boldsymbol{\mathcal{W}}_{t+1}$ in the proximal step (SVT) in Algorithm 1, we need first perform SVD of $\boldsymbol{\mathcal{Z}}_t$ given in (11). In general, obtaining the SVD of $d_u \times D$ matrix $\boldsymbol{\mathcal{Z}}_t$ requires $\mathcal{O}((d_u \wedge D)d_u D)$ operations, because its most expensive steps are computing matrix-vector multiplications. Since the computation of the proximal operator of the nuclear norm given in (10) does not require to do the full SVD, only a few singular values of $\boldsymbol{\mathcal{Z}}_t$ which are larger than λ/L are needed. Assume that there are $\hat{k}$ such singular values. As $\boldsymbol{\mathcal{W}}_t$ converges to a low-rank solution $\boldsymbol{\mathcal{W}}_*$, $\hat{k}$ will be small during iterating. The power method (Halko et al., 2011) at Algorithm 2 is a simple and efficient to capture subspace spanned by top-k singular vectors for $\hat{k} \geq k$. Additionally, the power method also allows warm-start, which is particularly useful because the iterative nature of APG algorithm. Once an approximation $\boldsymbol{\mathcal{Q}}$ is found, we have $\mathrm{SVT}_{\lambda/L}(\boldsymbol{\mathcal{Z}}_t) = \boldsymbol{\mathcal{Q}}\mathrm{SVT}_{\lambda/L}(\boldsymbol{\mathcal{Q}}^\top \boldsymbol{\mathcal{Z}}_t)$ (see Proposition 3.1 in Yao and Kwok (2015)). We therefore reduce the time complexity on SVT from $\mathcal{O}((d_u \wedge D)d_u D)$ to $\mathcal{O}(\hat{k}d_u D)$ which is much cheaper.

Algorithm 2: Power Method: `PowerMethod`$(\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{R}}, \epsilon)$

1. **input:** $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{d_u \times D}$, initial $\boldsymbol{\mathcal{R}} \in \mathbb{R}^{D \times k}$ for warm-start, tolerance δ;
2. **initialize** $\boldsymbol{\mathcal{W}}_1 = \boldsymbol{\mathcal{Z}}\boldsymbol{\mathcal{R}}$;
3. **for** $t = 1, 2, \ldots,$ **do**
4. $\quad \boldsymbol{\mathcal{Q}}_{t+1} = \mathrm{QR}(\boldsymbol{\mathcal{W}}_t)$;// QR denotes the QR factorization
5. $\quad \boldsymbol{\mathcal{W}}_{t+1} = \boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}^\top \boldsymbol{\mathcal{Q}}_{t+1})$;
6. $\quad$ **if** $\|\boldsymbol{\mathcal{Q}}_{t+1}\boldsymbol{\mathcal{Q}}_{t+1}^\top - \boldsymbol{\mathcal{Q}}_t\boldsymbol{\mathcal{Q}}_t^\top\|_F \leq \delta$ **then** break;
7. **return** $\boldsymbol{\mathcal{Q}}_{t+1}$.

Algorithm 3 shows how to approximate $\mathrm{SVT}_{\lambda/L}(\boldsymbol{\mathcal{Z}}_t)$. Let the target (exact) rank-k SVD of $\boldsymbol{\mathcal{Z}}_t$ be $\boldsymbol{\mathcal{U}}_k\boldsymbol{\Sigma}_k\boldsymbol{\mathcal{V}}_k^\top$. Step 1 first approximates $\boldsymbol{\mathcal{U}}_k$ by the power method. In steps 2 to 5, a less expensive $\mathrm{SVT}_{\lambda/L}(\boldsymbol{\mathcal{Q}}^\top\boldsymbol{\mathcal{Z}}_t)$ is obtained from (10). Finally, $\mathrm{SVT}_{\lambda/L}(\boldsymbol{\mathcal{Z}}_t)$ is recovered.

Hereafter, we denote the objective function in (4) by $\mathscr{F}_\lambda(\boldsymbol{\mathcal{W}})$, that is $\mathscr{F}_\lambda(\boldsymbol{\mathcal{W}}) = \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}) + \lambda\|\boldsymbol{\mathcal{W}}\|_*$, for any $\boldsymbol{\mathcal{W}} \in \mathscr{C}(\gamma)$. Recall that the gradient of the likelihood $\mathscr{L}_{\boldsymbol{\mathcal{Y}}}$ is written as

$$\nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}) = -\frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} B_{ij}^v(Y_{ij}^v - (G^v)'(W_{ij}^v))E_{ij}^v.$$

Algorithm 3: Approximate SVT: `Approx-SVT`$(\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{R}}, \lambda, \delta)$

input: $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{d_u \times D}, \boldsymbol{\mathcal{R}} \in \mathbb{R}^{D \times k}$, thresholds λ and δ;
$\boldsymbol{\mathcal{Q}} = \texttt{PowerMethod}(\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{R}}, \delta)$;
$[\boldsymbol{\mathcal{U}}, \boldsymbol{\Sigma}, \boldsymbol{\mathcal{V}}] = \mathrm{SVD}(\boldsymbol{\mathcal{Q}}^\top \boldsymbol{\mathcal{Z}})$;
$\boldsymbol{\mathcal{U}} = \{u_i | \sigma_i > \lambda\}$;
$\boldsymbol{\mathcal{V}} = \{v_i | \sigma_i > \lambda\}$;
$\boldsymbol{\Sigma} = \max(\boldsymbol{\Sigma} - \lambda \boldsymbol{\mathcal{I}}, \mathbf{0})$; // ($\boldsymbol{\mathcal{I}}$ denotes the identity matrix)
return $\boldsymbol{\mathcal{Q}}\boldsymbol{\mathcal{U}}, \boldsymbol{\Sigma}, \boldsymbol{\mathcal{V}}$.

By Assumption 2, we have for any $\boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{Q}} \in \mathbb{R}^{d_u \times D}$

$$\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}) - \nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}})\|_F^2 = \frac{1}{(d_u D)^2} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} \{B_{ij}^v ((G^v)'(W_{ij}^v) - (G^v)'(Q_{ij}^v))\}^2$$
$$\leq \frac{U_\gamma^2}{(d_u D)^2} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{Q}}\|_F^2.$$

This yields that $\mathscr{L}_{\boldsymbol{\mathcal{Y}}}$ has L-Lipschitz continuous gradient with $L = U_\gamma/(d_u D) \leq 1$. In the following algorithm and the experimental setup, we choose to work with $L = 1$.

Penalized Likelihood Accelerated Inexact Soft Impute (PLAIS-Impute). We present here the main algorithm in this paper, referred to as PLAIS-Impute, which is tailored to solving our collective matrix completion problem. The PLAIS-Impute is an adaption of the AIS-Impute algorithm in Yao and Kwok (2015) to the penalized likelihood completion problems. Note that AIS-Impute is an accelerated proximal gradient algorithm with further speed up based on approximate SVD. However, it is dedicated only to square-loss goodness-of-fitting. The PLAIS-Impute is summarized in Algorithm 4. The core steps are 10-12, where an approximate SVT is performed. Steps 10 and 11 use the column space of the last iterations ($\boldsymbol{\mathcal{V}}_t$ and $\boldsymbol{\mathcal{V}}_{t-1}$) to warm-start the power method. For further speed up, a continuation strategy is employed in which λ_t is initialized to a large value and then decreases gradually. The algorithm is restarted (at the step 14) if the objective function $\mathscr{F}_\lambda$ starts to increase. As AIS-Impute, PLAIS-Impute shares both low-iteration complexity and fast $\mathcal{O}(1/T^2)$ convergence rate (see Theorem 3.4 in Yao and Kwok (2015)).

5.2. Synthetic datasets

Software. The implementation of Algorithm 4 for the nuclear norm penalized estimator (4) was done in MATLAB R2017b on a desktop computer with macOS system, Intel i7 Core 3.5 GHz CPU and 16GB of RAM. For fast computation of SVD and sparse matrix computations, the experiments call an external package called PROPACK (Larsen, 1998) implemented in C and Fortran. The code that generates all figures given below is available from `https://github.com/mzalaya/collectivemc`.

Experimental setup. In our experiments we focus on square matrices. We set the number of the source matrices $V = 3$, then, for each $v \in \{1, 2, 3\}$, the low-rank ground truth parameter matrices $\boldsymbol{M}^v \in \mathbb{R}^{d \times d_v}$ are created with sizes $d \in \{3000, 6000, 9000\}$ and

Algorithm 4: PLAIS-Impute for Collective Matrix Completion

input: observed collective matrix $\boldsymbol{\mathcal{Y}}$, parameter λ, decay parameter $\nu \in (0,1)$, tolerance ε;
$[\boldsymbol{\mathcal{U}}_0, \lambda_0, \boldsymbol{\mathcal{V}}_0] = \text{rank-1 SVD}(\boldsymbol{\mathcal{Y}})$;
initialize $c = 1$, $\delta_0 = \|\boldsymbol{\mathcal{Y}}\|_F$, $\boldsymbol{\mathcal{W}}_0 = \boldsymbol{\mathcal{W}}_1 = \lambda_0 \boldsymbol{\mathcal{U}}_0 \boldsymbol{\mathcal{V}}_0^\top$;
for $t = 1, \ldots, T$ **do**
  $\delta_t = \nu^t \delta_0$;
  $\lambda_t = \nu^t(\lambda_0 - \lambda) + \lambda$;
  $\theta_t = (c-1)/(c+2)$;
  $\boldsymbol{\mathcal{Q}}_t = (1+\theta_t)\boldsymbol{\mathcal{W}}_t - \theta_t \boldsymbol{\mathcal{W}}_{t-1}$;
  $\boldsymbol{\mathcal{Z}}_t = \nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}_t))$;
  $\boldsymbol{\mathcal{V}}_{t-1} = \boldsymbol{\mathcal{V}}_{t-1} - \boldsymbol{\mathcal{V}}_t(\boldsymbol{\mathcal{V}}_t^\top \boldsymbol{\mathcal{V}}_{t-1})$;
  $\boldsymbol{\mathcal{R}}_t = \text{QR}([\boldsymbol{\mathcal{V}}_t, \boldsymbol{\mathcal{V}}_{t-1}])$;
  $[\boldsymbol{\mathcal{U}}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \boldsymbol{\mathcal{V}}_{t+1}] = \texttt{Approx-SVT}(\boldsymbol{\mathcal{Z}}_t, \boldsymbol{\mathcal{R}}_t, \lambda_t, \delta_t)$;
  if $\mathscr{F}_\lambda(\boldsymbol{\mathcal{U}}_{t+1}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\mathcal{V}}_{t+1}^\top) > \mathscr{F}_\lambda(\boldsymbol{\mathcal{U}}_t\boldsymbol{\Sigma}_t\boldsymbol{\mathcal{V}}_t^\top)$ **then**
    $c = 1$;
  else
    $c = c + 1$;
  if $|\mathscr{F}_\lambda(\boldsymbol{\mathcal{U}}_{t+1}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\mathcal{V}}_{t+1}^\top) - \mathscr{F}_\lambda(\boldsymbol{\mathcal{U}}_t\boldsymbol{\Sigma}_t\boldsymbol{\mathcal{V}}_t^\top)| \leq \varepsilon$ **then**
    break;
return $\boldsymbol{\mathcal{W}}_{T+1}$.

$d_v \in \{1000, 2000, 3000\}$ (hence $d = D = \sum_{v=1}^3 d_v$). Each source matrix $\boldsymbol{M}^v$ is constructed as $\boldsymbol{M}^v = \boldsymbol{L}^v \boldsymbol{R}^{v\top}$ where $\boldsymbol{L}^v \in \mathbb{R}^{d \times r_v}$ and $\boldsymbol{R}^v \in \mathbb{R}^{d_v \times r_v}$. This gives a random matrix of rank at most r_v. The parameter r_v is set to $\{5, 10, 15\}$. A fraction of the entries of $\boldsymbol{M}^v$ is removed uniformly at random with probability $p \in [0,1]$. Then, the matrices $\boldsymbol{M}^v$ are scaled so that $\|\boldsymbol{M}^v\|_\infty = \gamma = 1$.

For $\boldsymbol{M}^1$, the elements of $\boldsymbol{L}^1$ and $\boldsymbol{R}^1$ are sampled i.i.d. from the normal distribution $\mathcal{N}(0.5, 1)$. For $\boldsymbol{M}^2$, the entries of $\boldsymbol{L}^2$ and $\boldsymbol{R}^2$ are i.i.d. according to Poisson distribution with parameter 0.5. Finally, for $\boldsymbol{M}^3$, the entries of $\boldsymbol{L}^3$ and $\boldsymbol{R}^3$ are i.i.d. sampled from Bernoulli distribution with parameter 0.5. The collective matrix $\boldsymbol{\mathcal{M}}$ is constructed by concatenation of the three sources $\boldsymbol{M}^1, \boldsymbol{M}^2$ and $\boldsymbol{M}^3$, namely $\boldsymbol{\mathcal{M}} = (\boldsymbol{M}^1, \boldsymbol{M}^2, \boldsymbol{M}^3)$. All the details of these experiments are given in Table 2.

The details of our experiments are summarized in Figures 1 and 2. In Figure 1, we plot the convergence of the objective function $\mathscr{F}_\lambda$ versus time in the three experiments. Note that PLAIS-Impute inherits the speed of AIS-Impute as it does not require performing SVD and it has both low per-iteration cost and fast convergence rate. In Figure 1, we plot also the convergence of the objective function $\mathscr{F}_\lambda$ versus $-\log(\lambda)$ in the three experiments. The regularization parameter in the PLAIS-Impute is initialized to a large value and decreased gradually. In Figure 2, we illustrate a learning rank curve obtained by PLAIS-Impute, where the green color corresponds to the input rank and the cyan color to the recovered rank of the collective matrix $\boldsymbol{\mathcal{M}}$.

		$\boldsymbol{M}^1$ $(Gaussian)$	$\boldsymbol{M}^2$ $(Poisson)$	$\boldsymbol{M}^3$ $(Bernoulli)$	$\boldsymbol{\mathcal{M}}$ $(Collective)$
`exp.1`	*dimension*	3000×1000	3000×1000	3000×1000	3000×3000
	rank	5	5	5	*unknown*
`exp.2`	*dimension*	6000×2000	6000×2000	6000×2000	6000×6000
	rank	10	10	10	*unknown*
`exp.3`	*dimension*	9000×3000	9000×3000	9000×3000	9000×9000
	rank	15	15	15	*unknown*

Table 2: Details of the synthetic data in the three experiments.

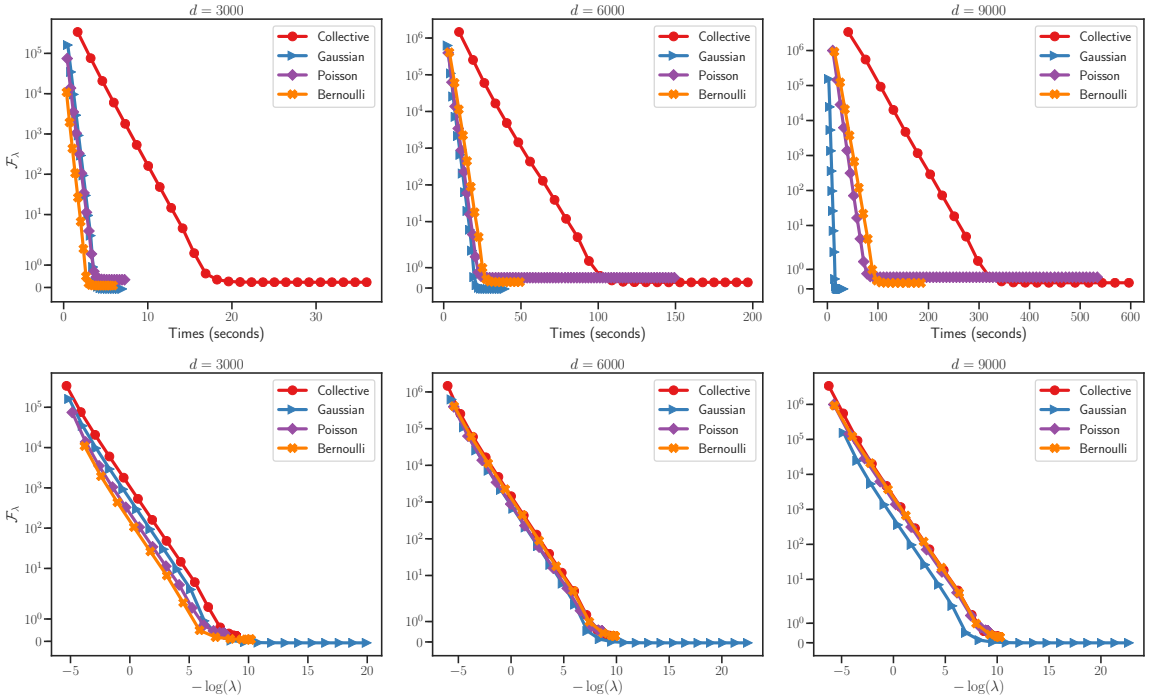

Figure 1: Convergence of the objective function $\mathscr{F}_\lambda$ in problem (4) versus time (top) and versus $-\log(\lambda)$ (bottom) in the three experiments with $p = 0.6$; left for `exp.1`; middle for `exp.2`; right for `exp.3`. Note that the objective functions for Gaussian, Poisson and Bernoulli distributions are calculated separately by the algorithm.

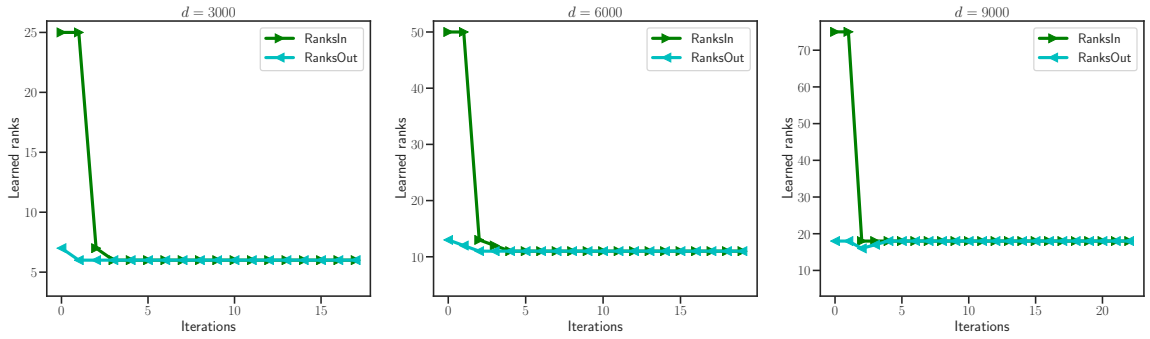

Figure 2: Learning ranks curve versus iterations in the three experiments with $p = 0.6$; left for `exp.1`; middle for `exp.2`; right for `exp.3`. We initialize the algorithm by setting a rank $r_0 = 5r$ where $r \in \{5, 10, 15\}$. The green color corresponds to the input rank while the cyan to the recovered rank of the collective matrix at each iteration. As can be seen, the two ranks gradually converge to the final recovered rank.

Evaluation. In our experiments, the PLAIS-Impute algorithm terminates when the absolute difference in the cost function values between two consecutive iterations is less than $\epsilon = 10^{-6}$. We set the regularization parameter $\lambda \propto \|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\|$ as given by Theorem 3. Note that in step 12 of PLAIS-Impute, the threshold in SVT is given by λ_t (defined in step 6), which is decreasing from one iteration to another. This allows to tune the first regularization parameter λ in the program (4). We randomly sample 80% of the observed entries for training, and the rest for testing.

In order to measure the the accuracy of our estimator, we employ the relative error (as, e.g., in Cai. et al. (2010); Davenport et al. (2014); Cai and Zhou (2013)) which is widely used metric in matrix completion and is defined by

$$\mathrm{RE}(\widehat{\boldsymbol{W}}, \boldsymbol{W}) = \frac{\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^o\|_F}{\|\boldsymbol{W}^o\|_F},$$

where $\widehat{\boldsymbol{W}}$ is the recovered matrix and $\boldsymbol{W}^o$ is the original full data matrix.

We run the PLAIS-Impute algorithm in each experiment by varying the percentage of known entries p from 0 to 1. In Figure 3, we plot the relative errors as a functions of p. We observe in Figure 3 that the relative errors are decaying with p. Note that for each $v \in \{1, 2, 3\}$, the estimator $\widehat{\boldsymbol{M}^v}$ is calculated separately using the same program (4). The results shown in Figure 3 confirm that collective matrix completion approach outperforms the approach that consists in estimating each component source independently.

Cold-Start problem. To simulate cold-start scenarios, we choose one of the source matrices $\boldsymbol{M}^v$ to be "cold" by increasing its sparsity. More precisely, we proceed in the following way: we extract vector of known entries of the chosen matrix and we set the first $1/5$ fraction of its entries to be equal to 0. We denote the obtained matrix by $\boldsymbol{M}^v_{\text{cold}}$ and the collective matrix by $\boldsymbol{\mathcal{M}}^v_{\text{cold}}$. In `exp.1`, `exp.2` and `exp.3`, we increase the sparsity of $\boldsymbol{M}^1$, $\boldsymbol{M}^2$, and $\boldsymbol{M}^3$, respectively. Hence, we get the "cold" collective matrices $\boldsymbol{\mathcal{M}}^1_{\text{cold}} = (\boldsymbol{M}^1_{\text{cold}}, \boldsymbol{M}^2, \boldsymbol{M}^3)$, $\boldsymbol{\mathcal{M}}^2_{\text{cold}} = (\boldsymbol{M}^1, \boldsymbol{M}^2_{\text{cold}}, \boldsymbol{M}^3)$, and $\boldsymbol{\mathcal{M}}^3_{\text{cold}} = (\boldsymbol{M}^1, \boldsymbol{M}^2, \boldsymbol{M}^3_{\text{cold}})$.

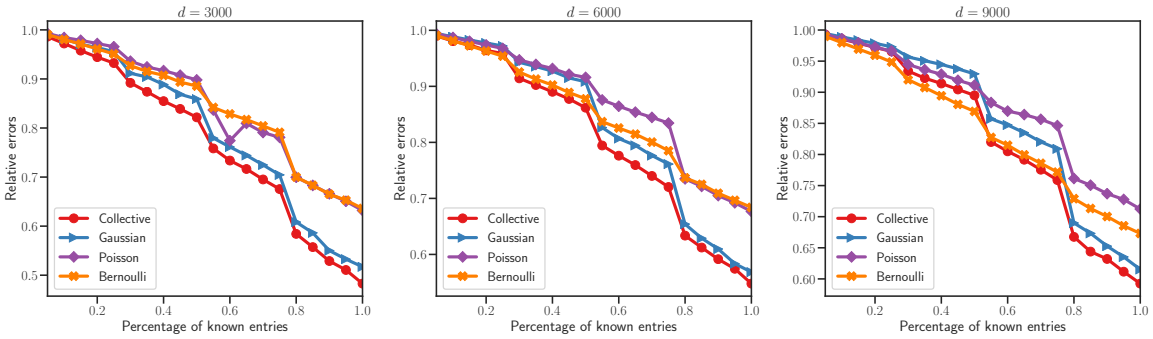

Figure 3: Performance on the synthetic data in terms of relative errors between the target and the estimator matrices as a function of the percentage of known entries p from 0 to 1.

We run 10 times the PLAIS-Impute algorithm for recovering the source $\boldsymbol{M}^v_{\text{cold}}$ and the collective $\boldsymbol{\mathcal{M}}^v_{\text{cold}}$ for each $v = 1, 2, 3$. We denote by $\widehat{\boldsymbol{M}^v_{\text{comp}}}$ the estimator of $\boldsymbol{M}^v_{\text{cold}}$ obtained by running the PLAIS-Impute algorithm only for this component. Analogously, we denote $\widehat{\boldsymbol{M}^v_{\text{collect}}}$ the estimator of $\boldsymbol{M}^v_{\text{cold}}$ obtained by extracting the v-th source of the collective estimator $\widehat{\boldsymbol{\mathcal{M}}^v_{\text{cold}}}$.

In Figure 4, we report the relative errors $\text{RE}(\widehat{\boldsymbol{M}^v_{\text{comp}}}, \boldsymbol{M}^v_{\text{cold}})$ and $\text{RE}(\widehat{\boldsymbol{M}^v_{\text{collect}}}, \boldsymbol{M}^v_{\text{cold}})$ in the three experiments. We see that, the collective matrix completion approach compensates the lack of informations in the "cold" source matrix. Therefore, this shared structure among the sources is useful to get better predictions.

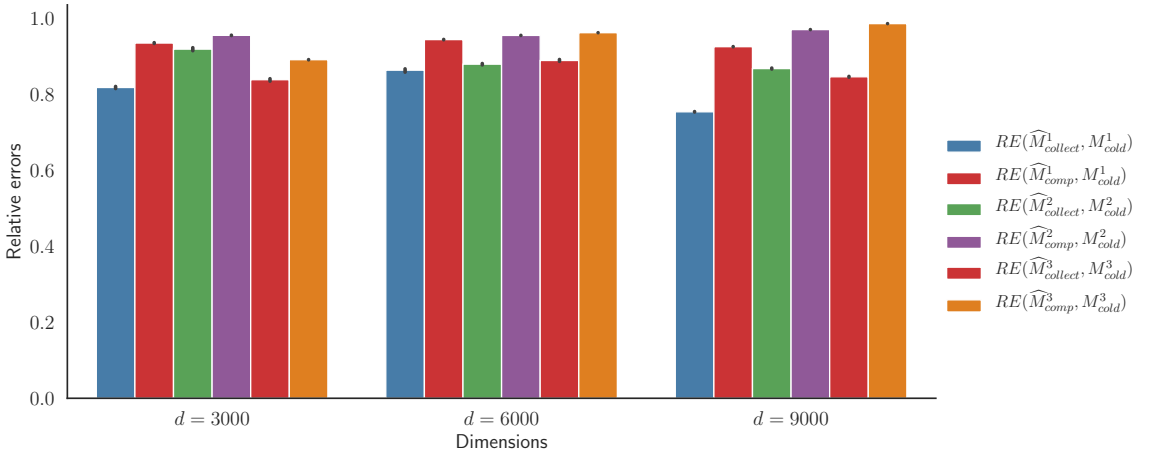

Figure 4: Relative errors over a set of 10 randomly generated datasets according to the cold-start scenarios (with the black lines representing $\pm$ the standard deviation) between the target and the estimator matrices.

6. Conclusion

This paper studies the problem of recovering a low-rank matrix when the data are collected from multiple and heterogeneous source matrices. We first consider the setting where, for each source, the matrix entries are sampled from an exponential family distribution. We then relax this assumption. The proposed estimators are based on minimizing the sum of a goodness-of-fit term and the nuclear norm penalization of the whole collective matrix. Allowing for non-uniform sampling, we establish upper bounds on the prediction risk of our estimator. As a by-product of our results, we provide exact minimax optimal rate of convergence for 1-bit matrix completion which previously was known upto a logarithmic factor. We present the proximal algorithm PLAIS-Impute to solve the corresponding convex programs. The empirical study provides evidence of the efficiency of the collective matrix completion approach in the case of joint low-rank structure compared to estimate each source matrices separately.

Acknowledgment

We would like to thank the Associated Editor and the two anonymous Referees for extremely valuable comments and remarks that helped us greatly to improve the paper. This work was supported by grants from DIM Math Innov Rgion Ile-de-France `https://www.dim-mathinnov.fr`

Appendix A. Proofs

We provide proofs of the main results, Theorems 3 and 11, in this section. The proofs of a few technical lemmas including Lemmas 5, 6 and 12 are also given. Before that, we recall some basic facts about matrices.

Basic facts about matrices. The singular value decomposition (SVD) of $\boldsymbol{A}$ has the form $\boldsymbol{A} = \sum_{l=1}^{\mathrm{rank}(\boldsymbol{A})} \sigma_l(\boldsymbol{A}) u_l(\boldsymbol{A}) v_l^\top(\boldsymbol{A})$ with orthonormal vectors $u_1(\boldsymbol{A}), \ldots, u_{\mathrm{rank}(\boldsymbol{A})}(\boldsymbol{A})$, orthonormal vectors $v_1(\boldsymbol{A}), \ldots, v_{\mathrm{rank}(\boldsymbol{A})}(\boldsymbol{A})$, and real numbers $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_{\mathrm{rank}(\boldsymbol{A})}(\boldsymbol{A}) > 0$ (the singular values of $\boldsymbol{A}$). Let $(\mathcal{S}_1(\boldsymbol{A}), \mathcal{S}_2(\boldsymbol{A}))$ be the pair of linear vectors spaces, where $\mathcal{S}_1(\boldsymbol{A})$ is the linear span space of $\{u_1(\boldsymbol{A}), \ldots, u_{\mathrm{rank}(\boldsymbol{A})}(\boldsymbol{A})\}$, and $\mathcal{S}_2(\boldsymbol{\mathcal{A}})$ is the linear span space of $\{v_1(\boldsymbol{A}), \ldots, v_{\mathrm{rank}(\boldsymbol{A})}(\boldsymbol{A})\}$. We denote by $\mathcal{S}_j^\perp(\boldsymbol{A})$ the orthogonal complements of $\mathcal{S}_j(\boldsymbol{A})$, for $j = 1, 2$ and by $P_{\mathcal{S}}$ the projector on the linear subspace $\mathcal{S}$ of $\mathbb{R}^n$ or $\mathbb{R}^m$.

For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we set $\mathscr{P}_{\boldsymbol{A}}^\perp(\boldsymbol{B}) = P_{\mathcal{S}_1^\perp(\boldsymbol{A})} \boldsymbol{B} P_{\mathcal{S}_2^\perp(\boldsymbol{A})}$ and $\mathscr{P}_{\boldsymbol{A}}(\boldsymbol{B}) = \boldsymbol{B} - \mathscr{P}_{\boldsymbol{A}}^\perp(\boldsymbol{B})$. Since $\mathscr{P}_{\boldsymbol{A}}(\boldsymbol{B}) = P_{\mathcal{S}_1(\boldsymbol{A})} \boldsymbol{B} + P_{\mathcal{S}_1^\perp(\boldsymbol{A})} \boldsymbol{B} P_{\mathcal{S}_2(\boldsymbol{A})}$, and $\mathrm{rank}(P_{\mathcal{S}_j(\boldsymbol{A})} \boldsymbol{B}) \leq \mathrm{rank}(\boldsymbol{A})$, we have that

$$\mathrm{rank}(\mathscr{P}_{\boldsymbol{A}}(\boldsymbol{B})) \leq 2\,\mathrm{rank}(\boldsymbol{A}). \tag{12}$$

It is easy to see that for two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ (Klopp, 2014)

$$\|\boldsymbol{A}\|_* - \|\boldsymbol{B}\|_* \leq \|\mathscr{P}_{\boldsymbol{A}}(\boldsymbol{A} - \boldsymbol{B})\|_* - \|\mathscr{P}_{\boldsymbol{A}}^\perp(\boldsymbol{A} - \boldsymbol{B})\|_*. \tag{13}$$

Finally, we recall the well-known trace duality property: for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times m}$, we have

$$|\langle \boldsymbol{A}, \boldsymbol{B} \rangle| \leq \|\boldsymbol{B}\| \|\boldsymbol{A}\|_*.$$

A.1. Proof of Theorem 3

First, noting that $\widehat{\boldsymbol{\mathcal{M}}}$ is optimal and $\boldsymbol{\mathcal{M}}$ is feasible for the convex optimization problem (4), we thus have the basic inequality that

$$\frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B_{ij}^v \big(G^v(\hat{M}_{ij}^v) - Y_{ij}^v \hat{M}_{ij}^v \big) + \lambda \|\widehat{\boldsymbol{\mathcal{M}}}\|_* \\ \leq \frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B_{ij}^v \big(G^v(M_{ij}^v) - Y_{ij}^v M_{ij}^v \big) + \lambda \|\boldsymbol{\mathcal{M}}\|_*.$$

It yields

$$\frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B_{ij}^v \Big(\big(G^v(\hat{M}_{ij}^v) - G^v(M_{ij}^v) \big) - Y_{ij}^v \big(\hat{M}_{ij}^v - M_{ij}^v \big) \Big) \leq \lambda (\|\boldsymbol{\mathcal{M}}\|_* - \|\widehat{\boldsymbol{\mathcal{M}}}\|_*).$$

Using the Bregman divergence associated to each G^v, we get

$$\frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B_{ij}^v d_{G^v}(\hat{M}_{ij}^v, M_{ij}^v) \\ \leq \lambda (\|\boldsymbol{\mathcal{M}}\|_* - \|\widehat{\boldsymbol{\mathcal{M}}}\|_*) - \frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B_{ij}^v \big((G^v)'(M_{ij}^v) - Y_{ij}^v \big) \big(\hat{M}_{ij}^v - M_{ij}^v \big).$$

Therefore, using the duality between $\|\cdot\|_*$ and $\|\cdot\|$, we arrive at

$$\frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} B_{ij}^v d_{G^v}(\hat{M}_{ij}^v, M_{ij}^v) \leq \lambda(\|\boldsymbol{\mathcal{M}}\|_* - \|\widehat{\boldsymbol{\mathcal{M}}}\|_*) - \langle \nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}}), \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\rangle$$
$$\leq \lambda(\|\boldsymbol{\mathcal{M}}\|_* - \|\widehat{\boldsymbol{\mathcal{M}}}\|_*) + \|\nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\|\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_*.$$

Besides, using the assumption $\lambda \geq 2\|\nabla\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\|$ and inequality (13) lead to

$$\frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} B_{ij}^v d_{G^v}(\hat{M}_{ij}^v, M_{ij}^v) \leq \frac{3\lambda}{2}\|\mathscr{P}_{\boldsymbol{\mathcal{M}}}(\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}})\|_*.$$

Since $\|\mathscr{P}_{\boldsymbol{\mathcal{A}}}(\boldsymbol{\mathcal{B}})\|_* \leq \sqrt{2\,\mathrm{rank}(\boldsymbol{\mathcal{A}})}\|\boldsymbol{\mathcal{B}}\|_F$ for any two matrices $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$, we obtain

$$\frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} B_{ij}^v d_{G^v}(\hat{M}_{ij}^v, M_{ij}^v) \leq \frac{3\lambda}{2}\sqrt{2\,\mathrm{rank}(\boldsymbol{\mathcal{M}})}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_F. \tag{14}$$

Now, Assumption 2 implies that the Bregman divergence satisfies $L_\gamma^2(x-y)^2 \leq 2d_G^v(x,y) \leq U_\gamma^2(x-y)^2$, then we get

$$\Delta^2_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}}) \leq \frac{2}{L_\gamma^2}\frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} B_{ij}^v d_{G^v}(\hat{M}_{ij}^v, M_{ij}^v), \tag{15}$$

where

$$\Delta^2_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}}) = \frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} B_{ij}^v(\hat{M}_{ij}^v - M_{ij}^v)^2.$$

Combining (14) and (15), we arrive at

$$\Delta^2_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}}) \leq \frac{3\lambda}{L_\gamma^2}\sqrt{2\,\mathrm{rank}(\boldsymbol{\mathcal{M}})}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_F. \tag{16}$$

Let us now define the threshold $\beta = \frac{946\gamma^2\log(d_u+D)}{p d_u D}$ and distinguish the two following cases that allows us to obtain an upper bound for the estimation error:
Case 1: if $(d_u D)^{-1}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} < \beta$, then the statement of Theorem 3 is true.
Case 2: it remains to consider the case $(d_u D)^{-1}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \geq \beta$. Lemma 18 in Appendix B.1 implies $\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_F \geq \frac{1}{4\sqrt{2\,\mathrm{rank}(\boldsymbol{\mathcal{M}})}}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_*$, then we obtain

$$\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_* \leq \sqrt{32\,\mathrm{rank}(\boldsymbol{\mathcal{M}})}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_F.$$

This leads to $\widehat{\boldsymbol{\mathcal{M}}} \in \mathscr{C}\big(\beta, 32\,\mathrm{rank}(\boldsymbol{\mathcal{M}})\big)$, where the set

$$\mathscr{C}(\beta, r) = \Big\{\boldsymbol{\mathcal{W}} \in \mathscr{C}_\infty(\gamma) : \|\boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{W}}\|_* \leq \sqrt{r}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|_F \text{ and } (d_u D)^{-1}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \geq \beta\Big\}. \tag{17}$$

Using Lemma 19 in Appendix B.1, we have

$$\Delta^2_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}}) \geq \frac{\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F}}{2d_u D} - 44536\,\mathrm{rank}(\boldsymbol{\mathcal{M}})\gamma^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2 - \frac{5567\gamma^2}{d_u D p}. \tag{18}$$

Together (18) and (16) imply

$$\begin{aligned}\frac{1}{2d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} &\leq \frac{3\lambda}{L^2_\gamma}\sqrt{2\,\mathrm{rank}(\boldsymbol{\mathcal{M}})}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|_F \\ &\quad + 44536\,\mathrm{rank}(\boldsymbol{\mathcal{M}})\gamma^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2 + \frac{5567\gamma^2}{pd_u D} \\ &\leq \frac{18\lambda^2 d_u D}{pL^4_\gamma}\mathrm{rank}(\boldsymbol{\mathcal{M}}) + \frac{1}{4d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \\ &\quad + 44536\,\mathrm{rank}(\boldsymbol{\mathcal{M}})\gamma^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2 + \frac{5567\gamma^2}{pd_u D}.\end{aligned}$$

Then,

$$\begin{aligned}\frac{1}{4d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} &\leq \frac{18\lambda^2 d_u D}{pL^4_\gamma}\mathrm{rank}(\boldsymbol{\mathcal{M}}) \\ &\quad + 44536p^{-1}d_u D\,\mathrm{rank}(\boldsymbol{\mathcal{M}})\gamma^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2 + \frac{5567\gamma^2}{d_u D p},\end{aligned}$$

and,

$$\frac{1}{d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \leq p^{-1}\max\left(d_u D\,\mathrm{rank}(\boldsymbol{\mathcal{M}})\left(\frac{c_1\lambda^2}{L^4_\gamma} + c_2\gamma^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2\right), \frac{c_3\gamma^2}{d_u D}\right),$$

where c_1, c_2 and c_3 are numerical constants. This concludes the proof of Theorem 3.

A.2. Proof of Lemma 5

We use the following result:

Proposition 16 *(Corollary 3.3 in Bandeira and van Handel (2016)) Let $\boldsymbol{W}$ be the $n \times m$ rectangular matrix whose entries W_{ij} are independent centered bounded random variables. Then there exists a universal constant $\boldsymbol{c}$ such that*

$$\mathbb{E}[\|\boldsymbol{W}\|] \leq \boldsymbol{c}\Big(\kappa_1 \vee \kappa_2 + \kappa_* \sqrt{\log(n \wedge m)}\Big),$$

where we have defined

$$\kappa_1 = \max_{i\in[n]}\sqrt{\sum_{j\in[m]}\mathbb{E}[W^2_{i,j}]}, \quad \kappa_2 = \max_{j\in[m]}\sqrt{\sum_{i\in[n]}\mathbb{E}[W^2_{i,j}]}, \quad \textit{and} \quad \kappa_* = \max_{(i,j)\in[n]\times[m]}|W_{ij}|.$$

We apply Proposition 16 to $\boldsymbol{\Sigma}_R = \frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]}\varepsilon^v_{ij}B^v_{ij}E^v_{ij}$. We compute

$$\begin{aligned}\kappa_1 = \frac{1}{d_u D}\max_{i\in[d_u]}\sqrt{\sum_{v\in[V]}\sum_{j\in[d_v]}\mathbb{E}[(\varepsilon^v_{ij})^2(B^v_{ij})^2]} &= \frac{1}{d_u D}\max_{i\in[d_u]}\sqrt{\sum_{v\in[V]}\sum_{j\in[d_v]}\pi^v_{ij}} \\ &= \frac{1}{d_u D}\max_{i\in[d_u]}\sqrt{\pi_{i\bullet}},\end{aligned}$$

$$\kappa_2 = \frac{1}{d_u D} \max_{v\in[V]} \max_{j\in[d_v]} \sqrt{\sum_{i\in[d_u]} \mathbb{E}[(\varepsilon_{ij}^v)^2 (B_{ij}^v)^2]} = \frac{1}{d_u D} \max_{v\in[V]} \max_{j\in[d_v]} \sqrt{\sum_{i\in[d_u]} \pi_{ij}^v}$$
$$\leq \frac{1}{d_u D} \max_{j\in[d_v]} \sqrt{\max_{v\in[V]} \sum_{i\in[d_u]} \pi_{ij}^v}$$
$$\leq \frac{1}{d_u D} \max_{j\in[d_v]} \sqrt{\pi_{\bullet j}},$$

and $\kappa_* = \frac{1}{d_u D} \max_{v\in[V]} \max_{(i,j)\in[d_u]\times[d_v]} |\varepsilon_{ij}^v B_{ij}| \leq \frac{1}{d_u D}$. Using inequality (1), we have $\kappa_1 \leq \frac{\sqrt{\mu}}{d_u D}$ and $\kappa_2 \leq \frac{\sqrt{\mu}}{d_u D}$. Then, $\kappa_1 \vee \kappa_2 \leq \frac{\sqrt{\mu}}{d_u D}$, which establishes Lemma 5.

A.3. Proof of Lemma 6

We write $\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}}) = -\frac{1}{d_u D} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} H_{ij}^v E_{ij}^v$, with $H_{ij}^v = B_{ij}^v \big(X_{ij}^v - (G^v)'(M_{ij}^v)\big)$. For a truncation level $T > 0$ to be chosen, we decompose $\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}}) = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$, where

$$\boldsymbol{\Sigma}_1 = -\frac{1}{d_u D} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \big(H_{ij}^v \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) \leq T)} - \mathbb{E}\big[H_{ij}^v \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) \leq T)}\big]\big) E_{ij}^v,$$

and

$$\boldsymbol{\Sigma}_2 = -\frac{1}{d_u D} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \big(H_{ij}^v \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) > T)} - \mathbb{E}\big[H_{ij}^v \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) > T)}\big]\big) E_{ij}^v,$$

then, the triangular inequality implies $\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\| \leq \|\boldsymbol{\Sigma}_1\| + \|\boldsymbol{\Sigma}_2\|$. Then, the proof is divided on two steps:

Step 1: control of $\|\boldsymbol{\Sigma}_1\|$. In order to control $\|\boldsymbol{\Sigma}_1\|$, we use the following bound on the spectral norms of random matrices. It is obtained by extension to rectangular matrices via self-adjoint dilation of Corollary 3.12 and Remark 3.13 in Bandeira and van Handel (2016).

Proposition 17 *(Bandeira and van Handel, 2016) Let* $\boldsymbol{W}$ *be the* $n \times m$ *rectangular matrix whose entries* W_{ij} *are independent centered bounded random variables. Then, for any* $0 \leq \epsilon \leq 1/2$ *there exists a universal constant* c_ϵ *such that for every* $x \geq 0$,

$$\mathbb{P}\big[\|\boldsymbol{W}\| \geq 2\sqrt{2}(1+\epsilon)(\kappa_1 \vee \kappa_2) + x\big] \leq (n \wedge m) \exp\Big(-\frac{x^2}{c_\epsilon \kappa_*^2}\Big),$$

where κ_1, κ_2, *and* κ_* *are defined as in Proposition 16.*

We apply Proposition 17 to $\boldsymbol{\Sigma}_1$. We compute

$$\kappa_1 = \frac{1}{d_u D} \max_{i\in[d_u]} \sqrt{\sum_{v\in[V]} \sum_{j\in[d_v]} \mathbb{E}\Big[\big(H_{ij}^v \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) \leq T)} - \mathbb{E}\big[H_{ij}^v \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) \leq T)}\big]\big)^2\Big]}.$$

Besides, we have

$$\mathbb{E}\Big[\big(H_{ij}^v \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) \leq T)} - \mathbb{E}\big[H_{ij}^v \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) \leq T)}\big]\big)^2\Big] \leq \mathbb{E}\big[(H_{ij}^v)^2 \mathbb{1}_{((X_{ij}^v - \mathbb{E}[X_{ij}^v]) \leq T)}\big],$$

and

$$\begin{aligned}\mathbb{E}\big[(H_{ij}^v)^2\mathbb{1}_{((X_{ij}^v-\mathbb{E}[X_{ij}^v])\le T)}\big] &= \mathbb{E}\big[(B_{ij}^v)^2\big(X_{ij}^v-\mathbb{E}[X_{ij}^v]\big)^2\mathbb{1}_{((X_{ij}^v-\mathbb{E}[X_{ij}^v])\le T)}\big]\\ &\le \pi_{ij}^v\mathbb{V}ar[X_{ij}^v]\\ &= \pi_{ij}^v(G^v)''(M_{ij}^v).\end{aligned}$$

By Assumption 2, we obtain $\mathbb{E}\big[(H_{ij}^v)^2\mathbb{1}_{((X_{ij}^v-\mathbb{E}[X_{ij}^v])\le T)}\big] \le \pi_{ij}^v U_\gamma^2$ for all $v\in[V], (i,j)\in[d_u]\times[d_v]$. Then,

$$\kappa_1 \le \frac{U_\gamma}{d_u D}\max_{i\in[d_u]}\sqrt{\sum_{v\in[V]}\sum_{j\in[d_v]}\pi_{ij}^v} \le \frac{U_\gamma}{d_u D}\max_{i\in[d_u]}\sqrt{\pi_{i\bullet}^v} \le \frac{U_\gamma\sqrt{\mu}}{d_u D},$$

and

$$\kappa_2 \le \frac{U_\gamma}{d_u D}\max_{j\in[d_v]}\sqrt{\max_{v\in[V]}\sum_{i\in[d_u]}\pi_{ij}^v} \le \frac{U_\gamma}{d_u D}\max_{j\in[d_v]}\sqrt{\pi_{\bullet j}} \le \frac{U_\gamma\sqrt{\mu}}{d_u D}.$$

It yields, $\kappa_1\vee\kappa_2 \le \frac{U_\gamma\sqrt{\mu}}{d_u D}$. Moreover, we have $\mathbb{E}\big[H_{ij}^v\mathbb{1}_{((X_{ij}^v-\mathbb{E}[X_{ij}^v])\le T)}\big] \le T$, which entails $\kappa_* \le \frac{2T}{d_u D}$. By choosing $\epsilon = 1/2$ in Proposition 17, we obtain, with probability at least $1-4(d_u\wedge D)e^{-x^2}$,

$$\|\mathbf{\Sigma}_1\| \le \frac{3U_\gamma\sqrt{2\mu}+2\sqrt{c_{1/2}}xT}{d_u D}.$$

Therefore, by setting $x=\sqrt{2\log(d_u+D)}$, we get with probability at least $1-4/(d_u+D)$,

$$\|\mathbf{\Sigma}_1\| \le \frac{3U_\gamma\sqrt{2\mu}+2\sqrt{c_{1/2}}\sqrt{2\log(d_u+D)}T}{d_u D}. \tag{19}$$

Step 2: control of $\|\mathbf{\Sigma}_2\|$. To control $\|\mathbf{\Sigma}_2\|$, we use Chebyshev's inequality, that is

$$\mathbb{P}\big[\|\mathbf{\Sigma}_2\| \ge \mathbb{E}[\|\mathbf{\Sigma}_2\|]+x\big] \le \frac{\mathbb{V}ar[\|\mathbf{\Sigma}_2\|]}{x^2}, \text{ for all } x>0.$$

We start by estimating $\mathbb{E}[\|\mathbf{\Sigma}_2\|]$. We use the fact that $\mathbb{E}[\|\mathbf{\Sigma}_2\|] \le \mathbb{E}[\|\mathbf{\Sigma}_2\|_F]$:

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{\Sigma}_2\|_F^2\big] &= \frac{1}{(d_u D)^2}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]}\mathbb{E}\Big[\big(H_{ij}^v\mathbb{1}_{((X_{ij}^v-\mathbb{E}[X_{ij}^v])>T)}-\mathbb{E}\big[H_{ij}^v\mathbb{1}_{((X_{ij}^v-\mathbb{E}[X_{ij}^v])>T)}\big]\big)^2\Big]\\ &\le \frac{1}{(d_u D)^2}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]}\mathbb{E}\big[(H_{ij}^v)^2\mathbb{1}_{((X_{ij}^v-\mathbb{E}[X_{ij}^v])>T)}\big]\\ &\le \frac{1}{(d_u D)^2}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]}\pi_{ij}^v\mathbb{E}\big[(X_{ij}^v-\mathbb{E}[X_{ij}^v])^2\mathbb{1}_{((X_{ij}^v-\mathbb{E}[X_{ij}^v])>T)}\big]\\ &\le \frac{1}{(d_u D)^2}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]}\pi_{ij}^v\sqrt{\mathbb{E}\big[(X_{ij}^v-\mathbb{E}[X_{ij}^v])^4\big]}\sqrt{\mathbb{P}\big[((X_{ij}^v-\mathbb{E}[X_{ij}^v])>T)\big]}.\end{aligned}$$

By Lemma 28, we have that $X_{ij}^v - \mathbb{E}[X_{ij}^v]$ is an (U_γ, K)-sub-exponential random variable for every $v \in [V]$ and $(i,j) \in [d_u] \times [d_v]$. It yields, using (2) in Theorem 26, that

$$\mathbb{E}\big[(X_{ij}^v - \mathbb{E}[X_{ij}^v])^p\big] \le \mathfrak{c} p^p \|X_{ij}^v\|_{\psi_1}^p, \text{ for every } p \ge 1,$$

and by (1) in Theorem 26

$$\mathbb{P}\big[|X_{ij}^v - \mathbb{E}[X_{ij}^v]| > T\big] \le \exp\Big(1 - \frac{T}{\mathfrak{c}_{\rm se}\|X_{ij}^v\|_{\psi_1}}\Big),$$

where $\mathfrak{c}$ and $\mathfrak{c}_{\rm se}$ are absolute constants. Consequently,

$$\begin{aligned}
\mathbb{E}\big[\|\mathbf{\Sigma}_2\|_F^2\big] &\le \frac{\mathfrak{c}}{(d_u D)^2} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \pi_{ij}^v \sqrt{\|X_{ij}^v\|_{\psi_1}^4} \sqrt{\exp\Big(1 - \frac{T}{\mathfrak{c}_{\rm se}\|X_{ij}^v\|_{\psi_1}}\Big)} \\
&\le \frac{\mathfrak{c}}{(d_u D)^2} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} (U_\gamma \vee K)^2 \pi_{ij}^v \sqrt{\exp\Big(-\frac{T}{\mathfrak{c}_{\rm se} K}\Big)}.
\end{aligned}$$

We choose $T = T_* := 4\mathfrak{c}_{\rm se}(U_\gamma \vee K)\log(d_u \vee D)$. It yields,

$$\begin{aligned}
\mathbb{E}\big[\|\mathbf{\Sigma}_2\|_F^2\big] &\le \frac{\mathfrak{c}}{(d_u D)^2} \frac{1}{(d_u \vee D)^2} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} (U_\gamma \vee K)^2 \pi_{ij}^v \\
&\le \frac{\mathfrak{c}(U_\gamma \vee K)^2}{(d_u D)^2} \frac{1}{(d_u \vee D)^2} \sum_{v\in[V]} \sum_{j\in[d_v]} \pi_{ij}^v \\
&\le \frac{\mathfrak{c}(U_\gamma \vee K)^2}{(d_u D)^2} \frac{1}{(d_u \vee D)^2} (d_u \vee D)\mu \\
&\le \frac{\mathfrak{c}(U_\gamma \vee K)^2 \mu}{(d_u D)^2 d_u \vee D}.
\end{aligned}$$

Using the fact that $x \mapsto \sqrt{x}$ is concave, we obtain

$$\mathbb{E}[\|\mathbf{\Sigma}_2\|] \le \mathbb{E}[\|\mathbf{\Sigma}_2\|_F] \le \sqrt{\mathbb{E}\big[\|\mathbf{\Sigma}_2\|_F^2\big]} \le \sqrt{\frac{\mathfrak{c}(U_\gamma \vee K)^2\mu}{(d_u D)^2 d_u \vee D}} \le \frac{\mathfrak{c}(U_\gamma \vee K)\sqrt{\mu}}{d_u D\sqrt{d_u \vee D}}. \tag{20}$$

Let us now control the variance of $\|\mathbf{\Sigma}_2\|$. We have immediately, using (20),

$$\mathbb{V}ar[\|\mathbf{\Sigma}_2\|] \le \mathbb{E}[\|\mathbf{\Sigma}_2\|^2] \le \mathbb{E}\big[\|\mathbf{\Sigma}_2\|_F^2\big] \le \frac{\mathfrak{c}(U_\gamma \vee K)^2\mu}{(d_u D)^2 d_u \vee D}.$$

By Chebyshev's inequality and using (20), we have, with probability at least $1 - 4/(d_u + D)$,

$$\|\mathbf{\Sigma}_2\| \le \frac{\mathfrak{c}(U_\gamma \vee K)\sqrt{\mu}}{d_u D\sqrt{d_u \vee D}} + \frac{\mathfrak{c}(U_\gamma \vee K)\sqrt{\mu}}{d_u D} \le \frac{\mathfrak{c}(U_\gamma \vee K)\sqrt{\mu}}{d_u D}. \tag{21}$$

Finally, combining (19) and (21), we obtain, with probability at least $1 - 4/(d_u + D)$,

$$\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\| \le \frac{3U_\gamma\sqrt{2\mu} + 8(U_\gamma \vee K)\mathfrak{c}_{\rm se}\sqrt{2\mathfrak{c}_{1/2}\log(d_u + D)}\log(d_u \vee D) + \mathfrak{c}(U_\gamma \vee K)\sqrt{\mu}}{d_u D}$$

Then,

$$\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{M}})\| \leq c\Big(\frac{(U_\gamma \vee K)\big(\sqrt{\mu} + (\log(d_u \vee D))^{3/2}\big)}{d_u D}\Big),$$

where c is an absolute constant. This finishes the proof of Lemma 6.

A.4. Proof of Theorem 11

We start the proof with the following inequality using the fact that $\widehat{\boldsymbol{\mathcal{M}}}$ is the minimizer of the objective function in problem (8)

$$0 \leq -(R_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}) + \Lambda\|\widehat{\boldsymbol{\mathcal{M}}}\|_*) + (R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}}) + \Lambda\|\overset{\star}{\boldsymbol{\mathcal{M}}}\|_*).$$

Then, by adding $R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \geq 0$, we obtain

$$R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \leq -\big\{\big(R_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \big(R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big)\big\} + \Lambda\big(\|\overset{\star}{\boldsymbol{\mathcal{M}}}\|_* - \|\widehat{\boldsymbol{\mathcal{M}}}\|_*\big).$$

(13) implies $\|\boldsymbol{A}\|_* - \|\boldsymbol{B}\|_* \leq \|\mathscr{P}_{\boldsymbol{A}}(\boldsymbol{A} - \boldsymbol{B})\|_*$ and we get

$$\begin{aligned} R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) &\leq -\big\{\big(R_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \big(R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big)\big\} + \Lambda\|\mathscr{P}_{\overset{\star}{\boldsymbol{\mathcal{M}}}}(\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}})\|_* \\ &\leq -\big(R_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) + \big(R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) \\ &\quad + \Lambda\sqrt{2\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})}\|\widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_F. \end{aligned} \tag{22}$$

Let us now define the threshold $\nu = \frac{32\left(1+e\sqrt{3\rho/\varsigma\gamma}\right)\rho\gamma\log(d_u+D)}{3pd_uD}$ and distinguish the two following cases that allows us to obtain an upper bound for the prediction error:

Case 1: if $R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) < \nu$, then the statement of Theorem 11 is true.

Case 2: it remains to consider the case $R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \geq \nu$. Lemma 21 implies

$$\|\widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_* \leq \sqrt{32\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})}\|\widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_F,$$

then $\widehat{\boldsymbol{\mathcal{M}}} \in \mathscr{Q}(\nu, 32\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}}))$ where

$$\mathscr{Q}(\nu, r) = \Big\{\boldsymbol{\mathcal{Q}} \in \mathscr{C}_\infty(\gamma) : \|\boldsymbol{\mathcal{Q}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_* \leq \sqrt{r}\|\boldsymbol{\mathcal{Q}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_F \text{ and } R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \geq \nu\Big\}.$$

Using Lemma 22, we have

$$\begin{aligned} &R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) - \big(R_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) \\ &\quad \leq \frac{R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})}{2} + \frac{c\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})\rho^2\varsigma^{-1}(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{(1/4e) + (1 - 1/\sqrt{4e})\sqrt{3\rho/4\varsigma\gamma}}. \end{aligned} \tag{23}$$

Now, plugging (23) in (22), we get

$$R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \leq \frac{\mathcal{C}\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})\rho^2\varsigma^{-1}(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{(1/4e) + (1-1/\sqrt{4e})\sqrt{3\rho/4\varsigma\gamma}} + 2\Lambda\sqrt{2\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})}\|\widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_F,$$

where $\mathcal{C} = 1024$. Then using the fact that for any $a, b \in \mathbb{R}$, and $\epsilon > 0$, we have $2ab \leq a^2/(2\epsilon) + 2\epsilon b^2$, we get for $\epsilon = p\varsigma/4$

$$\begin{aligned}
R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) &\leq \frac{\mathcal{C}d_u D p^{-1}\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})\rho^2\varsigma^{-1}(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{(1/4e) + (1-1/\sqrt{4e})\sqrt{3\rho/4\varsigma\gamma}} \\
&\quad + \Lambda^2 d_u D(p\varsigma/4)^{-1}\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}}) + \frac{p\varsigma}{2d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_F^2 \\
&\leq \frac{\mathcal{C}d_u D p^{-1}\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})\rho^2\varsigma^{-1}(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{(1/4e) + (1-1/\sqrt{4e})\sqrt{3\rho/4\varsigma\gamma}} \\
&\quad + \Lambda^2 d_u D(p\varsigma/4)^{-1}\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}}) + \frac{\varsigma}{2d_u D}\|\widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_{\Pi,F}^2.
\end{aligned}$$

Using Assumption 4, we obtain

$$\begin{aligned}
R(\widehat{\boldsymbol{\mathcal{M}}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) &\leq \frac{2\mathcal{C}d_u D p^{-1}\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})\rho^2\varsigma^{-1}(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{(1/4e) + (1-1/\sqrt{4e})\sqrt{3\rho/4\varsigma\gamma}} + 8\Lambda^2 d_u D(p\varsigma)^{-1}\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}}) \\
&\leq (p\varsigma)^{-1}\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})d_u D\Big(\frac{\rho^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{(1/4e) + (1-1/\sqrt{4e})\sqrt{3\rho/4\varsigma\gamma}} + 8\Lambda^2\Big).
\end{aligned}$$

This finishes the proof of Theorem 11.

A.5. Proof of Lemma 12

By the nonnegative factor and the sum properties of subdifferential calculus (Boyd and Vandenberghe, 2004), we write

$$\partial R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}}) = \Big\{\boldsymbol{\mathcal{G}} = \frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]} B_{ij}^v G_{ij}^v E_{ij}^v : G_{ij}^v \in \partial\ell^v(Y_{ij}^v, \overset{\star}{M}_{ij}^v)\Big\}$$

Recall that the sudifferential of $\partial\ell^v(Y_{ij}^v, \overset{\star}{M}_{ij}^v)$ at the point $\overset{\star}{M}_{ij}^v$ is defined as

$$\partial\ell^v(Y_{ij}^v, \overset{\star}{M}_{ij}^v) = \{G_{ij}^v : \ell^v(Y_{ij}^v, Q_{ij}^v) \geq \ell^v(Y_{ij}^v, \overset{\star}{M}_{ij}^v) + G_{ij}^v(Q_{ij}^v - \overset{\star}{M}_{ij}^v)\}.$$

Thanks to Assumption 3, we have, for all $G_{ij}^v \in \partial\ell^v(Y_{ij}^v, \overset{\star}{M}_{ij}^v)$

$$|G_{ij}^v(Q_{ij}^v - \overset{\star}{M}_{ij}^v)| \leq |\ell^v(Y_{ij}^v, Q_{ij}^v) - \ell^v(Y_{ij}^v, \overset{\star}{M}_{ij}^v)| \leq \rho_v|Q_{ij}^v - \overset{\star}{M}_{ij}^v|,$$

In particular, with $Q_{ij}^v \neq \overset{\star}{M}_{ij}^v$ for all $v \in [V]$ and $(i,j) \in [d_u]\times[d_v]$, we get $|G_{ij}^v| \leq \rho_v$. Then, any subgradient $\boldsymbol{\mathcal{G}}$ of $R_{\boldsymbol{\mathcal{Y}}}$ has entries bounded by $\rho/(d_u D)$ (recall $\rho = \max_{v\in[V]}\rho_v$). By a

triangular inequality and the convexity of $\|\cdot\|$, we have

$$\begin{aligned}\|\boldsymbol{\mathcal{G}}\| &\leq \|\boldsymbol{\mathcal{G}} - \mathbb{E}[\boldsymbol{\mathcal{G}}]\| + \|\mathbb{E}[\boldsymbol{\mathcal{G}}]\| \\ &\leq \|\boldsymbol{\mathcal{G}} - \mathbb{E}[\boldsymbol{\mathcal{G}}]\| + \mathbb{E}[\|\boldsymbol{\mathcal{G}}\|],\end{aligned}$$

for any subgradient $\boldsymbol{\mathcal{G}}$ of $R_{\boldsymbol{\mathcal{Y}}}$. On the one hand, we use the fact that $\mathbb{E}[\|\boldsymbol{\mathcal{G}}\|] \leq \mathbb{E}[\|\boldsymbol{\mathcal{G}}\|_F] \leq \sqrt{\mathbb{E}[\|\boldsymbol{\mathcal{G}}\|_F^2]}$. Using (1), we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\mathcal{G}}\|_F^2] &\leq \frac{1}{(d_u D)^2} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \rho_v^2 \mathbb{E}[B_{ij}^v] \\ &\leq \frac{\rho^2}{(d_u D)^2} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \pi_{ij}^v \\ &\leq \frac{\rho^2 \mu}{(d_u D)^2}.\end{aligned}$$

Now we apply Proposition 17 to $\boldsymbol{\mathcal{G}} - \mathbb{E}[\boldsymbol{\mathcal{G}}]$. Taking into account (1), we upper bound the constants κ_1, κ_2 and κ_* as follows:

$$\begin{aligned}\kappa_1 &= \frac{1}{d_u D} \max_{i\in[d_u]} \sqrt{\sum_{v\in[V]} \sum_{j\in[d_v]} \mathbb{E}[(B_{ij}^v G_{ij}^v - \mathbb{E}[B_{ij}^v G_{ij}^v])^2]} \\ &\leq \frac{2\rho}{d_u D} \max_{i\in[d_u]} \sqrt{\sum_{v\in[V]} \sum_{j\in[d_v]} \pi_{ij}^v} \\ &\leq \frac{2\rho\sqrt{\mu}}{d_u D},\end{aligned}$$

$$\begin{aligned}\kappa_2 &= \frac{1}{d_u D} \max_{v\in[V]} \max_{j\in[d_v]} \sqrt{\sum_{i\in[d_u]} \mathbb{E}[(B_{ij}^v G_{ij}^v - \mathbb{E}[B_{ij}^v G_{ij}^v])^2]} \\ &\leq \frac{2\rho}{d_u D} \max_{v\in[V]} \max_{j\in[d_v]} \sqrt{\sum_{i\in[d_u]} \pi_{ij}^v} \\ &\leq \frac{2\rho\sqrt{\mu}}{d_u D},\end{aligned}$$

and $\kappa_* = \frac{1}{d_u D} \max_{v\in[V]} \max_{(i,j)\in[d_u]\times[d_v]} |B_{ij}^v G_{ij}^v - \mathbb{E}[B_{ij}^v G_{ij}^v]| \leq \frac{2\rho}{d_u D}$. Now, choose $\epsilon = 1/2$ in Proposition 17, then we obtain, with probability at least $1 - 4(d_u \wedge D)e^{-x^2}$,

$$\|\boldsymbol{\mathcal{G}} - \mathbb{E}[\boldsymbol{\mathcal{G}}]\| \leq \frac{6\rho\sqrt{2\mu} + 2\rho\sqrt{c_{1/2}}x}{d_u D}. \tag{24}$$

Setting $x = \sqrt{2\log(d_u + D)}$ in (24), we get with probability at least $1 - 4/(d_u + D)$,

$$\|\boldsymbol{\mathcal{G}}\| \leq \frac{(1 + 6\sqrt{2})\rho\sqrt{\mu} + 2\rho\sqrt{c_{1/2}}\sqrt{2\log(d_u + D)}}{d_u D}, \tag{25}$$

for any subgradient $\boldsymbol{\mathcal{G}}$ of $R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})$.

Appendix B. Technical Lemmas

In this section, we provide several technical lemmas, which are used for proving our main results.

B.1. Useful lemmas for the proof of Theorem 3

Lemma 18 *Let $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathscr{C}_\infty(\gamma)$. Assume that $\lambda \geq 2\|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{B}})\|$, and $\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{A}}) + \lambda\|\boldsymbol{\mathcal{A}}\|_* \leq \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{B}}) + \lambda\|\boldsymbol{\mathcal{B}}\|_*$. Then,*

(i) $\|\mathscr{P}^{\perp}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_* \leq 3\|\mathscr{P}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_*$,

(ii) $\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}}\|_* \leq 4\sqrt{2\,\mathrm{rank}(\boldsymbol{\mathcal{B}})}\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}}\|_F$.

Proof We have $\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{B}}) - \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{A}}) \geq \lambda(\|\boldsymbol{\mathcal{A}}\|_* - \|\boldsymbol{\mathcal{B}}\|_*)$. (13) implies

$$\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{B}}) - \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{A}}) \geq \lambda\big(\|\mathscr{P}^{\perp}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_* - \|\mathscr{P}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_*\big).$$

Moreover, by convexity of $\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\cdot)$ and the duality between $\|\cdot\|_*$ and $\|\cdot\|$ we obtain

$$\mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{B}}) - \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{A}}) \leq \langle \nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{B}}), \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{A}}\rangle \leq \|\nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{B}})\|\|\boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{A}}\|_* \leq \frac{\lambda}{2}\|\boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{A}}\|_*.$$

Therefore,

$$\|\mathscr{P}^{\perp}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_* \leq \|\mathscr{P}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_* + \frac{1}{2}\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}}\|_* \tag{26}$$

Using the triangle inequality, we get

$$\|\mathscr{P}^{\perp}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_* \leq 3\|\mathscr{P}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_*,$$

which proves (i). To prove (ii), note that $\|\mathscr{P}_{\boldsymbol{\mathcal{B}}}(\boldsymbol{\mathcal{A}})\|_* \leq \sqrt{2\,\mathrm{rank}(\boldsymbol{\mathcal{B}})}\|\boldsymbol{\mathcal{A}}\|_F$, and (i) imply

$$\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}}\|_* \leq 4\sqrt{2\,\mathrm{rank}(\boldsymbol{\mathcal{B}})}\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}})\|_F.$$

■

Lemma 19 *Let $\beta = \frac{946\gamma^2\log(d_u+D)}{pd_uD}$. Then, for all $\boldsymbol{\mathcal{W}} \in \mathscr{C}(\beta, r)$,*

$$\Big|\Delta^2_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{M}}) - (d_uD)^{-1}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F}\Big| \leq \frac{(d_uD)^{-1}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F}}{2} + 1392r\gamma^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2 + \frac{5567\gamma^2}{d_uDp}$$

with probability at least $1 - 4/(d_u + D)$.

Proof We use a standard peeling argument. For any $\alpha > 1$ and $0 < \eta < 1/2\alpha$, we define

$$\boldsymbol{\kappa} = \frac{1}{1/(2\alpha) - \eta}\Big(128\gamma^2 r(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2 + \frac{512\gamma^2}{d_uDp}\Big)$$

and we consider the event

$$\mathscr{W} = \Big\{ \exists\, \boldsymbol{\mathcal{W}} \in \mathscr{C}(\beta, r) : \Big| \Delta^2_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{M}}) - (d_u D)^{-1} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \Big| > \frac{(d_u D)^{-1} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F}}{2} + \boldsymbol{\kappa} \Big\}.$$

For $s \in \mathbb{N}^*$, set

$$\mathcal{R}_s = \Big\{ \boldsymbol{\mathcal{W}} \in \mathscr{C}(\beta, r) : \alpha^{s-1}\beta \leq (d_u D)^{-1} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \leq \alpha^s \beta \Big\}.$$

If the event $\mathscr{W}$ holds for some matrix $\boldsymbol{\mathcal{W}} \in \mathscr{C}(\beta, r)$, then $\boldsymbol{\mathcal{W}}$ belongs to some $\mathcal{R}_s$ and

$$\begin{aligned} \Big| \Delta^2_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{M}}) - (d_u D)^{-1} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \Big| &\geq \frac{(d_u D)^{-1} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F}}{2} + \boldsymbol{\kappa} \\ &\geq \frac{1}{2\alpha} \alpha^s \beta + \boldsymbol{\kappa}. \end{aligned}$$

For $\theta \geq \beta$ consider the following set of matrices

$$\mathscr{C}(\beta, r, \theta) = \Big\{ \boldsymbol{\mathcal{W}} \in \mathscr{C}(\beta, r) : (d_u D)^{-1} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \leq \theta \Big\},$$

and the following event

$$\mathscr{W}_s = \Big\{ \exists\, \boldsymbol{\mathcal{W}} \in \mathscr{C}(\beta, r, \theta) : \Big| \Delta^2_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{M}}) - (d_u D)^{-1} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \Big| \geq \frac{1}{2\alpha} \alpha^s \beta + \boldsymbol{\kappa} \Big\}.$$

Note that $\boldsymbol{\mathcal{W}} \in \mathscr{W}_s$ implies that $\boldsymbol{\mathcal{W}} \in \mathscr{C}(\beta, r, \alpha^s \beta)$. Then, we get $\mathscr{W} \subset \cup_s \mathscr{W}_s$. Thus, it is enough to estimate the probability of the simpler event $\mathscr{W}_s$ and then apply a the union bound. Such an estimation is given by the following lemma:

Lemma 20 *Let*

$$\boldsymbol{\mathcal{Z}}_\theta = \sup_{\boldsymbol{\mathcal{W}} \in \mathscr{C}(\beta, r, \theta)} \Big| \Delta^2_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{M}}) - (d_u D)^{-1} \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{M}}\|^2_{\Pi,F} \Big|.$$

Then, we have

$$\mathbb{P}\big[\boldsymbol{\mathcal{Z}}_\theta > \frac{\theta}{2\alpha} + \boldsymbol{\kappa} \big] \leq 4 \exp\Big(- \frac{p d_u D \eta^2 \theta}{8\gamma^2} \Big).$$

The proof of Lemma 20 follows along the same lines of Lemma 10 in Klopp (2015). We now apply an union bound argument combined to Lemma 20, we get

$$\begin{aligned} \mathbb{P}[\mathscr{W}] \leq \mathbb{P}[\cup_{s=1}^\infty \mathscr{W}_s] &\leq 4 \sum_{s=1}^\infty \exp\Big(- \frac{p d_u D \eta^2 \alpha^s \beta}{8\gamma^2} \Big) \\ &\leq 4 \sum_{s=1}^\infty \exp\Big(- \frac{p d_u D \eta^2 \beta \log \alpha}{8\gamma^2} s \Big) \\ &\leq \frac{4 \exp\Big(- \frac{p d_u D \eta^2 \beta \log \alpha}{8\gamma^2} \Big)}{1 - \exp\Big(- \frac{p d_u D \eta^2 \beta \log \alpha}{8\gamma^2} \Big)}. \end{aligned}$$

By choosing $\alpha = e, \eta = 1/4e$ and β as stated we get the desired result. ■

B.2. Useful lemmas for the proof of Theorem 11

Lemma 21 *Suppose* $\Lambda \geq 2\sup\{\|\boldsymbol{\mathcal{G}}\| : \boldsymbol{\mathcal{G}} \in \partial R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\}$. *Then*

$$\|\widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_* \leq 4\sqrt{2\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})}\|\widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_F.$$

Proof For any subgradient $\boldsymbol{\mathcal{G}}$ of $R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})$, we have $R_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}) \geq R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}}) + \langle \boldsymbol{\mathcal{G}}, \widehat{\boldsymbol{\mathcal{M}}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\rangle$. Then, the definition of the estimator $\widehat{\boldsymbol{\mathcal{M}}}$, entails $R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}}) - R_{\boldsymbol{\mathcal{Y}}}(\widehat{\boldsymbol{\mathcal{M}}}) \geq \Lambda(\|\widehat{\boldsymbol{\mathcal{M}}}\|_* - \|\overset{\star}{\boldsymbol{\mathcal{M}}}\|_*)$, hence $\langle \boldsymbol{\mathcal{G}}, \overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}}\rangle \geq \Lambda(\|\widehat{\boldsymbol{\mathcal{M}}}\|_* - \|\overset{\star}{\boldsymbol{\mathcal{M}}}\|_*)$. The duality between $\|\cdot\|_*$ and $\|\cdot\|$ yields

$$\Lambda(\|\widehat{\boldsymbol{\mathcal{M}}}\|_* - \|\overset{\star}{\boldsymbol{\mathcal{M}}}_*\|) \leq \|\boldsymbol{\mathcal{G}}\|\|\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}}\|_* \leq \frac{\Lambda}{2}\|\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}}\|_*$$

then $\|\widehat{\boldsymbol{\mathcal{M}}}\|_* - \|\overset{\star}{\boldsymbol{\mathcal{M}}}_*\| \leq \frac{1}{2}\|\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}}\|_*$. Now, (13) implies

$$\|\mathscr{P}^{\perp}_{\overset{\star}{\boldsymbol{\mathcal{M}}}}(\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}})\|_* \leq \|\mathscr{P}_{\overset{\star}{\boldsymbol{\mathcal{M}}}}(\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}})\|_* + \frac{1}{2}\|\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}}\|_* \leq 3\|\mathscr{P}_{\overset{\star}{\boldsymbol{\mathcal{M}}}}(\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}})\|_*.$$

Therefore $\|\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}}\|_* \leq 4\|\mathscr{P}_{\overset{\star}{\boldsymbol{\mathcal{M}}}}(\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}})\|_*$. Since $\|\mathscr{P}_{\overset{\star}{\boldsymbol{\mathcal{M}}}}(\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}})\|_* \leq \sqrt{2\,\mathrm{rank}(\overset{\star}{\boldsymbol{\mathcal{M}}})}\|\overset{\star}{\boldsymbol{\mathcal{M}}} - \widehat{\boldsymbol{\mathcal{M}}}\|_F$, we establish the proof of Lemma 21. ■

Lemma 22 *Let*

$$\nu = \frac{32\big(1 + e\sqrt{3\rho/\varsigma\gamma}\big)\rho\gamma\log(d_u + D)}{3pd_uD},$$

then, with probability at least $1 - 4/(d_u + D)$*, the following holds uniformly over* $\boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r)$

$$\Big|\big(R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \big(R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big)\Big| \leq \frac{R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})}{2} + \frac{16}{(1/4e) + (1 - 1/\sqrt{4e})\sqrt{3\rho/4\varsigma\gamma}} r\rho^2(p\varsigma)^{-1}(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2.$$

Proof The proof is based on the peeling argument. For any $\delta > 1$ and $0 < \vartheta < 1/2\delta$, define

$$\boldsymbol{\zeta} = \frac{16r(p\varsigma)^{-1}\rho^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{(1/2\delta) + \sqrt{3\rho/4\varsigma\gamma} - \Big(\vartheta + \sqrt{3\rho/4\varsigma\gamma\vartheta}\Big)}, \tag{27}$$

and we consider the event

$$\mathscr{A} = \Big\{\exists\, \boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r) : \Big|\big(R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \big(R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big)\Big| > \frac{R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})}{2} + \boldsymbol{\zeta}\Big\}.$$

For $l \in \mathbb{N}^*$, we define the sequence of subsets

$$\mathscr{J}_l = \Big\{\boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r) : \delta^{l-1}\nu \leq R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \leq \delta^l\nu\Big\}.$$

If the event $\mathscr{A}$ holds for some matrix $\boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r)$, then $\boldsymbol{\mathcal{Q}}$ belongs to some $\mathcal{J}_l$ and

$$\begin{aligned}\Big|\big(R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \big(R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big)\Big| &> \frac{R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})}{2} + \boldsymbol{\zeta} \\ &\geq \frac{1}{2\delta}\delta^l\nu + \boldsymbol{\zeta}.\end{aligned}$$

For $\theta \geq \nu$, consider the following set of matrices

$$\mathscr{Q}(\nu, r, \theta) = \Big\{\boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r) : R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}}) \leq \theta\Big\},$$

and the following event

$$\mathscr{A}_l = \Big\{\exists\, \boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r, \theta) : \Big|\big(R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \big(R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big)\Big| \geq \frac{1}{2\delta}\delta^l\nu + \boldsymbol{\zeta}\Big\}.$$

Note that $\boldsymbol{\mathcal{Q}} \in \mathcal{J}_l$ implies that $\boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r, \delta^l\nu)$. Then, we get $\mathscr{A} \subset \cup_l \mathscr{A}_l$. Thus, it is enough to estimate the probability of the simpler event $\mathscr{A}_l$ and then apply a the union bound. Such an estimation is given in Lemma 23, where we derive a concentration inequality for the following supremum of process:

$$\boldsymbol{\Xi}_\theta = \sup_{\boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r, \theta)} \Big|\big(R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \big(R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big)\Big|$$

We now apply an union bound argument combined to Lemma 23, we get

$$\begin{aligned}\mathbb{P}[\mathscr{A}] \leq \mathbb{P}[\cup_{l=1}^{\infty}\mathscr{A}_l] &\leq \sum_{l=1}^{\infty} \exp\Big(-\frac{3d_u D\vartheta\delta^l\nu}{8\rho\gamma}\Big) \\ &\leq \sum_{l=1}^{\infty} \exp\Big(-\frac{3d_u D\vartheta\log(\delta)\nu}{8\rho\gamma}l\Big) \\ &\leq \frac{\exp\Big(-\frac{3d_u D\vartheta\log(\delta)\nu}{8\rho\gamma}\Big)}{1 - \exp\Big(-\frac{3d_u D\vartheta\log(\delta)\nu}{8\rho\gamma}\Big)},\end{aligned}$$

where se used the elementary inequality that $u^s = e^{s\log(u)} \geq s\log(u)$. By choosing $\delta = e, \vartheta = 1/4e$ and ν as stated we get the desired result. ■

Lemma 23 *One has*

$$\mathbb{P}\Big[\boldsymbol{\Xi}_\theta \geq \Big(1 + \delta\sqrt{\frac{3\rho}{\varsigma\gamma}}\Big)\frac{\theta}{2\delta} + \boldsymbol{\zeta}\Big] \leq \exp\Big(-\frac{3d_u D\vartheta\theta}{8\rho\gamma}\Big).$$

Proof The proof of this lemma is based on Bousquet's concentration theorem:

Theorem 24 *(Bousquet, 2002) (see also Corollary 16.1 in van de Geer (2016)) Let $\mathcal{F}$ be a class of real-valued functions. Let $T_1, \dots, T_N$ be independent random variables such that $\mathbb{E}[f(T_i)] = 0$ and $|f(T_i)| \leq \xi$ for all $i = 1, \dots, N$ and for all $f \in \mathcal{F}$. Introduce $Z = \sup_{f\in\mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^N \big(f(T_i) - \mathbb{E}[f(T_i)]\big) \right|$. Assume further that*

$$\frac{1}{N} \sum_{i=1}^N \sup_{f\in\mathcal{F}} \mathbb{E}\big[f^2(T_i)\big] \leq M^2.$$

Then we have for all $t > 0$

$$\mathbb{P}\left[Z \geq 2\mathbb{E}[Z] + M\sqrt{\frac{2t}{N}} + \frac{4t\xi}{3N} \right] \leq e^{-t}.$$

We start by bounding the expectation

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\Xi}_\theta] &= \mathbb{E}\Big[\sup_{\boldsymbol{\mathcal{Q}}\in\mathscr{Q}(\nu,r,\theta)} \Big| \big(R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \big(R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) \Big| \Big] \\
&= \mathbb{E}\Big[\sup_{\boldsymbol{\mathcal{Q}}\in\mathscr{Q}(\nu,r,\theta)} \Big| \big(R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big) - \mathbb{E}\big[R_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}) - R_{\boldsymbol{\mathcal{Y}}}(\overset{\star}{\boldsymbol{\mathcal{M}}})\big] \Big| \Big] \\
&= \mathbb{E}\Big[\sup_{\boldsymbol{\mathcal{Q}}\in\mathscr{Q}(\nu,r,\theta)} \Big| \frac{1}{d_u D} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} B_{ij}^v \big(\ell^v(Y_{ij}^v, Q_{ij}^v) - \ell^v(Y_{ij}^v, \overset{\star}{M}{}_{ij}^v)\big) \\
&\qquad\qquad - \mathbb{E}\big[B_{ij}^v \big(\ell^v(Y_{ij}^v, Q_{ij}^v) - \ell^v(Y_{ij}^v, \overset{\star}{M}{}_{ij}^v)\big)\big] \Big| \Big] \\
&\leq 2\mathbb{E}\Big[\sup_{\boldsymbol{\mathcal{Q}}\in\mathscr{Q}(\nu,r,\theta)} \Big| \frac{1}{d_u D} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \varepsilon_{ij}^v B_{ij}^v \big(\ell^v(Y_{ij}^v, Q_{ij}^v) - \ell^v(Y_{ij}^v, \overset{\star}{M}{}_{ij}^v)\big) \Big| \Big] \\
&\leq 4\rho\mathbb{E}\Big[\sup_{\boldsymbol{\mathcal{Q}}\in\mathscr{Q}(\nu,r,\theta)} \Big| \frac{1}{d_u D} \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \varepsilon_{ij}^v B_{ij}^v (Q_{ij}^v - \overset{\star}{M}{}_{ij}^v) \Big| \Big] \\
&\leq 4\rho\mathbb{E}\Big[\sup_{\boldsymbol{\mathcal{Q}}\in\mathscr{Q}(\nu,r,\theta)} \Big| \langle \boldsymbol{\Sigma}_R, \boldsymbol{\mathcal{Q}} - \overset{\star}{\boldsymbol{\mathcal{M}}} \rangle \Big| \Big] \\
&\leq 4\rho\mathbb{E}\Big[\|\boldsymbol{\Sigma}_R\| \sup_{\boldsymbol{\mathcal{Q}}\in\mathscr{Q}(\nu,r,\theta)} \|\boldsymbol{\mathcal{Q}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_* \Big],
\end{aligned}$$

where the first inequality follows from symmetrization of expectations theorem of van der Vaart and Wellener, the second from contraction principle of Ledoux and Talagrand (see Theorems 14.3 and 14.4 in Bühlmann and van de Geer (2011)), and the third from duality between nuclear and operator norms. We have $\boldsymbol{\mathcal{Q}} \in \mathscr{Q}(\nu, r, \theta)$ then $\|\boldsymbol{\mathcal{Q}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_* \leq \sqrt{r}\|\boldsymbol{\mathcal{Q}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_F$ and using Assumption 4, we have $\|\boldsymbol{\mathcal{Q}} - \overset{\star}{\boldsymbol{\mathcal{M}}}\|_* \leq \sqrt{r(p\varsigma)^{-1}\big(R(\boldsymbol{\mathcal{Q}}) - R(\overset{\star}{\boldsymbol{\mathcal{M}}})\big)} \leq \sqrt{r(p\varsigma)^{-1}\theta}$. Then,

$$\mathbb{E}[\boldsymbol{\Xi}_\theta] \leq 4\sqrt{r(p\varsigma)^{-1}\theta}\rho\mathbb{E}[\|\boldsymbol{\Sigma}_R\|].$$

For the upper bound ξ in Theorem 24, we have that

$$\big|\ell^v(Y_{ij}^v, Q_{ij}^v) - \ell^v(Y_{ij}^v, \overset{\star}{M}{}_{ij}^v)\big| \leq \rho_v |Q_{ij}^v - \overset{\star}{M}{}_{ij}^v| \leq 2\rho_v\gamma \leq 2\rho\gamma.$$

Now we compute M in Theorem 24. Thanks to Assumption 4, we have

$$\begin{aligned}
&\frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]}\mathbb{E}\big[\big(B_{ij}^v\big(\ell^v(Y_{ij}^v,Q_{ij}^v)-\ell^v(Y_{ij}^v,\overset{\star}{M}{}_{ij}^v)\big)\big)^2\big]\\
&\qquad\le \frac{1}{d_u D}\sum_{v\in[V]}\sum_{(i,j)\in[d_u]\times[d_v]}(\rho_v)^2\mathbb{E}\big[B_{ij}^v(Q_{ij}^v-\overset{\star}{M}{}_{ij}^v)^2\big]\\
&\qquad\le \frac{\rho^2}{d_u D}\|\boldsymbol{\mathcal{Q}}-\overset{\star}{\boldsymbol{\mathcal{M}}}\|_{\Pi,F}^2\\
&\qquad\le \frac{\rho^2}{\varsigma}(R(\boldsymbol{\mathcal{Q}})-R(\overset{\star}{\boldsymbol{\mathcal{M}}}))\\
&\qquad\le \frac{\rho^2\theta}{\varsigma}.
\end{aligned}$$

Then, Bousquet's theorem implies that for all $t>0$,

$$\mathbb{P}\Big[\boldsymbol{\Xi}_\theta \ge 2\mathbb{E}[\boldsymbol{\Xi}_\theta]+\sqrt{\frac{2\rho^2\theta t}{\varsigma d_u D}}+\frac{8\rho\gamma t}{3d_u D}\Big]\le e^{-t}.$$

Taking $t=\frac{3d_uD\vartheta\theta}{8\rho\gamma}$, we obtain

$$\mathbb{P}\Big[\boldsymbol{\Xi}_\theta \ge 8\gamma\sqrt{r(p\varsigma)^{-1}\theta}\rho\mathbb{E}[\|\boldsymbol{\Sigma}_R\|]+\Big(\sqrt{\frac{3\rho}{4\varsigma\gamma}\vartheta}+\vartheta\Big)\theta\Big]\le\exp\Big(-\frac{3d_uD\vartheta\theta}{8\rho\gamma}\Big). \tag{28}$$

Using the fact that for any $a,b\in\mathbb{R}$, and $\epsilon>0$, $2ab\le a^2/\epsilon+\epsilon b^2$, we get (for $\epsilon=1/2\delta+\sqrt{3\rho/4\varsigma\gamma}-\Big(\vartheta+\sqrt{3\rho\vartheta/4\varsigma\gamma}\Big)$), we get

$$\begin{aligned}
8\gamma\sqrt{r(p\varsigma)^{-1}\theta}\rho\mathbb{E}[\|\boldsymbol{\Sigma}_R\|]+\Big(\sqrt{\frac{3\vartheta\rho}{\varsigma\gamma}}+\vartheta\Big)\theta &\le \frac{16r(p\varsigma)^{-1}\rho^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{\frac{1}{2\delta}+\sqrt{\frac{3\rho}{4\varsigma\gamma}}-\vartheta-\sqrt{\frac{3\rho}{4\varsigma\gamma}\vartheta}}+\Big(\frac{1}{2\delta}+\sqrt{\frac{3\rho}{4\varsigma\gamma}}\Big)\theta\\
&\le \frac{16r(p\varsigma)^{-1}\rho^2(\mathbb{E}[\|\boldsymbol{\Sigma}_R\|])^2}{\frac{1}{2\delta}+\sqrt{\frac{3\rho}{4\varsigma\gamma}}-\vartheta-\sqrt{\frac{3\rho}{4\varsigma\gamma}\vartheta}}+\Big(1+\delta\sqrt{\frac{3\rho}{\varsigma\gamma}}\Big)\frac{\theta}{2\delta}.
\end{aligned}$$

Using (28), we get $\mathbb{P}\Big[\boldsymbol{\Xi}_\theta\ge\Big(1+\delta\sqrt{\frac{3\rho}{\varsigma\gamma}}\Big)\frac{\theta}{2\delta}+\boldsymbol{\zeta}\Big]\le\exp\Big(-\frac{3d_uD\vartheta\theta}{8\rho\gamma}\Big)$. This finishes the proof of Lemma 23. ■

Appendix C. Sub-exponential random variables

The material here is taken from R.Vershynin (2010).

Definition 25 *A random variable X is sub-exponential with parameters (ω,b) if for all t such that $|t|\le 1/b$,*

$$\mathbb{E}\big[\exp\big(t(X-\mathbb{E}[X])\big)\big]\le\exp\big(\frac{t^2\omega^2}{2}\big). \tag{29}$$

When $b = 0$, we interpret $1/0$ as being the same as ∞, it follows immediately from this definition that any sub-Gaussian random variable is also sub-exponential. There are also a variety of other conditions equivalent to sub-exponentiality, which we relate by defining the sub-exponential norm of random variable. In particular, we define the sub-exponential norm (sometimes known as the ψ_1-Orlicz in the literature) as

$$\|X\|_{\psi_1} := \sup_{q \geq 1} \frac{1}{q} (\mathbb{E}[|X^q|])^{1/q}.$$

Then we have the following lemma which provides several equivalent characterizations of sub-exponential random variables.

Theorem 26 *(Equivalence of sub-exponential properties (R.Vershynin, 2010))*
Let X be a random variable and $\omega > 0$ be a constant. Then, the following properties are all equivalent with suitable numerical constants $K_i > 0, i = 1, \ldots, 4$, that are different from each other by at most an absolute constant $\mathcal{C}$, meaning that if one statement (i) holds with parameter K_i, then the statement (j) holds with parameter $K_j \leq \mathcal{C}K_i$.

(1) sub-exponential tails: $\mathbb{P}[|X| > t] \leq \exp\big(1 - \frac{t}{\omega K_1}\big)$, for all $t \geq 0$.

(2) sub-exponential moments: $(\mathbb{E}[|X^q|])^{1/q} \leq K_2 \omega q$, for all $q \geq 1$.

(3) existence of moment generating function (Mgf): $\mathbb{E}\big[\exp\big(\frac{X}{\omega K_3}\big)\big] \leq e$.

Note that in each of the statements of Theorem 26, we may replace ω by $\|X\|_{\psi_1}$ and, up to absolute constant factors, $\|X\|_{\psi_1}$ is the smallest possible number in these inequalities.

Lemma 27 *(Mgf of sub-exponential random variables (R.Vershynin, 2010)) Let X be a centered sub- exponential random variable. Then, for t such that $|t| \leq c/\|X\|_{\psi_1}$, one has*

$$\mathbb{E}[\exp(tX)] \leq \exp(Ct^2 \|X\|_{\psi_1}^2)$$

where $C, c > 0$ are absolute constants.

Lemma 28 *For all $v \in [V]$ and $(i, j) \in [d_u] \times [d_v]$, the random variable $X_{i,j}^v$ is a sub-exponential with parameters (U_γ, K), where K is defined in Assumption 2. Moreover, we have that $\|X_{i,j}^v\|_{\psi_1} = \mathcal{C}(U_\gamma \vee K)$ for some absolute constant $\mathcal{C}$.*

Proof Let t such that $|t| \leq 1/K$, then

$$\begin{aligned}
&\mathbb{E}[\exp\big(t(X_{ij}^v - \mathbb{E}[X_{ij}^v])\big)] \\
&\qquad = e^{-t(G^v)'(M_{ij}^v)} \int_{\mathbb{R}} h^v(x) \exp\big((t + M_{ij}^v)x - G^v(M_{ij}^v)\big) dx \\
&\qquad = e^{G^v(t+M_{ij}^v) - G^v(M_{ij}^v) - t(G^v)'(M_{ij}^v)} \int_{\mathbb{R}} h^v(x) \exp\big((t + M_{ij}^v)x - G^v(t + M_{ij}^v)\big) dx \\
&\qquad = e^{G^v(t+M_{ij}^v) - G^v(M_{ij}^v) - t(G^v)'(M_{ij}^v)},
\end{aligned}$$

where we used in the last inequality the fact that that $\int_{\mathbb{R}} h^v(x) \exp\big((t + M_{ij}^v)x - G^v(t + M_{ij}^v)\big) dx = \int_{\mathbb{R}} f_{h^v, G^v}(X_{i,j}^v | t + M_{ij}^v) dx = 1$. Therefore, an ordinary Taylor series expansion

of G^v implies that there exists $t_{\gamma,K} \in [-\gamma - \frac{1}{K}, \gamma + \frac{1}{K}]$ such that $G^v(t + M_{ij}^v) - G^v(M_{ij}^v) - t(G^v)'(M_{ij}^v) = (t^2/2)(G^v)''(t_{\gamma,K}^2)$. By Assumption 2, we obtain

$$\mathbb{E}[\exp\left(t(X_{ij}^v - \mathbb{E}[X_{ij}^v])\right)] \leq \exp\Big(\frac{t^2 U_\gamma^2}{2}\Big).$$

Using Lemma 27, we get $\|X_{i,j}^v\|_{\psi_1} = c(U_\gamma \vee K)$ for some absolute constant c. This proves Lemma 28. ■

References

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

P. Alquier, V. Cottet, and G. Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *arXiv:1702.01402*, 2017.

M.G. Armentano, D. Godoy, and A. A. Amandi. Followee recommendation based on text analysis of micro-blogging activity. *Information Systems*, 38(8):1116 – 1127, 2013. ISSN 0306-4379.

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

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.

P. L. Bartlett, M. I. Jordan, and J. D. Mcauliffe. Large margin classifiers: Convex loss, low noise, and convergence rates. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1173–1180. MIT Press, 2004.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, 2009.

J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Know.-Based Syst.*, 46:109–132, 2013.

G. Bouchard, D. Yin, and S. Guo. Convex collective matrix factorization. In *AISTATS*, volume 31 of *JMLR Workshop and Conference Proceedings*, pages 144–152. JMLR.org, 2013.

O. Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495 – 500, 2002.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg, 2011.

J.-F. Cai., E. J. Cands, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

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

T. Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electron. J. Statist.*, 10(1):1493–1525, 2016.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

E. J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

I. Cantador, I. Fernández-Tobías, Shlomo Berkovsky, and P. Cremonesi. *Cross-Domain Recommender Systems*, pages 919–959. Springer US, Boston, MA, 2015.

Y. Censor and S.A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, USA, 1997.

K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon. Matrix completion with noisy side information. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3447–3455, Cambridge, MA, USA, 2015. MIT Press.

K.-Y. Chiang, I. S. Dhillon, and C.-J. Hsieh. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *Journal of Machine Learning Research*, 19(76):1–35, 2018.

M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189, 2014.

P. Drineas, A. Javed, M. Magdon-Ismail, G. Pandurangant, R. Virrankoski, and A. Savvides. Distance matrix reconstruction from incomplete distance information for sensor network localization. In *2006 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks*, volume 2, pages 536–544, 2006.

A. Elsener and S. van de Geer. Robust low-rank matrix estimation. *Ann. Statist.*, 46(6B): 3481–3509, 12 2018.

M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

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

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

D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

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

S. Gunasekar, M. Yamada, D. Yin, and Y. Chang. Consistent collective matrix completion under joint low rank structure. In *AISTATS*, 2015.

S. Gunasekar, J. C. Ho, J. Ghosh, S. Kreml, A. N. Kho, J. C Denny, B. A. Malin, and J. Sun. Phenotyping using structured collective matrix factorization of multi–source ehr data. *preprint arXiv:1609.04466*, 2016.

N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2): 217–288, 2011.

J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 199–206, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-906-0.

S. Horii, T. Matsushima, and S. Hirasawa. A note on the correlated multiple matrix completion based on the convex optimization method. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1618–1623, 2014.

Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2117–2130, 2013.

P. Jain and I. S. Dhillon. Provable inductive matrix completion. *CoRR*, abs/1306.0626, 2013.

M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 135–142, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-906-0.

S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 457–464, New York, NY, USA, 2009a. ACM.

S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 457–464, New York, NY, USA, 2009b. ACM.

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

O. Klopp. Matrix completion by singular value thresholding: Sharp bounds. *Electron. J. Statist.*, 9(2):2348–2369, 2015.

O. Klopp, J. Lafond, E. Moulines, and J. Salmon. Adaptive multinomial matrix completion. *Electron. J. Statist.*, 9(2):2950–2975, 2015.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

J. Lafond. Low rank matrix completion with exponential family noise. In Peter Grnwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1224–1243, Paris, France, 2015. PMLR.

X. N. Lam, T. Vu, T. D. Le, and A. D. Duong. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2Nd International Conference on Ubiquitous Information Management and Communication*, ICUIMC '08, pages 208–211, New York, NY, USA, 2008. ACM.

R. M. Larsen. Lanczos bidiagonalization with partial reorthogonalization, 1998.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, New York, NY, USA, second edition, 1998.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.*, 31(3):1235–1256, 2009.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2010.

H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 287–296, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:2287–2322, 2010.

S. Mendelson. Obtaining fast error rates in nonconvex situations. *Journal of Complexity*, 24(3):380 – 397, 2008.

N. Natarajan and I. S. Dhillon. Inductive matrix completion for predicting genedisease associations. *Bioinformatics*, 30(12):i60–i68, 06 2014. ISSN 1367-4803.

N. Natarajan, D. Shin, and I. S. Dhillon. Which app will you use next?: Collaborative filtering with interactional context. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 201–208, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2409-0.

S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, 39(2):1069–1097, 2011.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

S. Oh, A. Montanari, and A. Karbasi. Sensor network localization from local connectivity: Performance analysis for the mds-map algorithm. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pages 1–5, 2010.

J. Pagès. *Multiple Factor Analysis by Example Using R*. Chapman & Hall/CRC The R Series. Taylor & Francis, 2014. ISBN 9781482205473.

N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, 2014.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

J. D. M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 713–719, New York, NY, USA, 2005. ACM.

R.Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010.

S. Shin, D.and Cetintas and I. S. Lee, K.-C.and Dhillon. Tumblr blog recommendation with boosted inductive matrix completion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 203–212, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3794-6.

A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 650–658, New York, NY, USA, 2008. ACM.

A. P. Singh and G. J. Gordon. A bayesian matrix factorization model for relational data. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pages 556–563, Arlington, Virginia, United States, 2010. AUAI Press.

A. M.-C. So and Y. Ye. Theory of semidefinite programming for sensor network localization. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, pages 405–414, Philadelphia, PA, USA, 2005. Society for Industrial and Applied Mathematics.

N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization, 2005.

M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Found. Trends Mach. Learn.*, 9(1):1–118, June 2016. ISSN 1935-8237.

S. van de Geer. *Estimation and Testing Under Sparsity: École d'Été de Probabilités de Saint-Flour XLV – 2015*. Lecture Notes in Mathematics. Springer International Publishing, 2016.

L. Xu, Z. Chen, Q. Zhou, E. Chen, N. J. Yuan, and X. Xie. Aligned matrix completion: Integrating consistency and independency in multiple domains. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 529–538, 2016.

M. Xu, R. Jin, and Z.-H. Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2301–2309, USA, 2013. Curran Associates Inc.

Q. Yao and J. T. Kwok. Accelerated inexact soft-impute for fast large-scale matrix completion. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 4002–4008. AAAI Press, 2015.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

M. Zitnik and B. Zupan. Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 400–11, 2014.

M. Zitnik and B. Zupan. Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(1):41–53, 2015. ISSN 0162-8828.