9-2021

Orthogonal inductive matrix completion

Antoine LEDENT
Singapore Management University, aledent@smu.edu.sg

Rrodrigo ALVES
TU Kaiserslautern

Marius KLOFT
TU Kaiserslautern

Orthogonal Inductive Matrix Completion

Antoine Ledent*, Rodrigo Alves*, and Marius Kloft, *Senior Member, IEEE*

***Abstract*—We propose orthogonal inductive matrix completion (OMIC), an interpretable approach to matrix completion based on a sum of multiple orthonormal side information terms, together with nuclear-norm regularization. The approach allows us to inject prior knowledge about the singular vectors of the ground truth matrix. We optimize the approach by a provably converging algorithm, which optimizes all components of the model simultaneously. We study the generalization capabilities of our method in both the distribution-free setting and in the case where the sampling distribution admits uniform marginals, yielding learning guarantees that improve with the quality of the injected knowledge in both cases. As particular cases of our framework, we present models which can incorporate user and item biases or community information in a joint and additive fashion. We analyse the performance of OMIC on several synthetic and real datasets. On synthetic datasets with a sliding scale of user bias relevance, we show that OMIC better adapts to different regimes than other methods. On real-life datasets containing user/items recommendations and relevant side information, we find that OMIC surpasses the state-of-the-art, with the added benefit of greater interpretability.**

I. Introduction and related work

Matrix completion, the problem of recovering the missing entries of a partially observed matrix, has found application in a wide range of domains. As examples consider the following. (1) A streaming provider recommends movies to its users, based on an incomplete database of user-movie ratings. (2) A social network wants to find missing links in their friendship network. (3) A chemical producer wants to predict interactions of chemical compounds from a subset of known pairwise interactions. These examples—from the domains of recommender systems [1], [2], social network analysis [3], and chemical engineering [4]—highlight the wide range of applications of matrix competition. For simplicity, we use movie recommendation as running example here, so the data consists of user-movie ratings. It should be clear that, more generally, we can work with type1-type2 pairs, depending on the application, e.g., user-book, user-user, compound-compound, etc.

To recover missing entries of a matrix, it is necessary to make an assumption on the structure of the ground truth matrix. The most common assumption is that the matrix is of low rank. However, optimally approximating the observed entries whilst minimizing the rank is NP-hard. The SoftImpute algorithm [5] bypasses this difficulty by using the nuclear norm as a regularizer. Not only does SoftImpute work well in practice, it also enjoys favorable theoretical guarantees: only a small number of known entries is required to recover the underlying low-rank matrix exactly [6], [7] or approximately [8], [9] from noisy entries.

In practical applications, the following refinements may help the performance of classic recommender systems. (1) *Incorporating bias* [10], [11]. Some users may generally be more critical than others. This means they tend to give lower ratings than other users, regardless of the movie. Moreover, some movies are intrinsically better than others, so they receive more favorable ratings. Previous work incorporated user and movie bias in a pre-processing step, and then trained matrix completion on the residuals. (2) *Incorporating side information*. For movies, we find plenty of side information (genre, staring actors, director, reviews, etc.) on the web, and we might have access to user attributes (age, gender, etc.), from which we can derive clusters of users (community information). Inductive matrix completion (IMC) [12] uses such side information to guide the prediction of the user-movie ratings. IMC, which is backed up by well-developed learning theory [12], [13], [14], [15], [16], can be applied also to new users with no ratings, but for which side information is given.

Our aim in this work is to create a generic model that can incorporate all the improvements mentioned above into a single flexible framework with a well-principled optimization procedure. To the best of our knowledge, the only work which attempted to incorporate some of the above improvements into a single jointly trained model is [17], which considers

*The first two authors contributed equally to this work
The authors are with TU Kaiserslautern
{ledent,alves,kloft}@cs.uni-kl.de
Code repository: http://github.com/rasalves/OMIC
Accepted for publication in Transactions of Neural Networks and Learning Systems (TNNLS), DOI: 10.1109/TNNLS.2021.3106155

outputs of the form

$$f_{i,j} = \mathbf{x}_i^\top M \mathbf{y}_j + z_{i,j}, \tag{1}$$

with nuclear-norm regularization imposed on both Z and M. The model is trained with gradient descent. The incorporation of both the standard low-rank term Z and the IMC term $XMY^\top$ allows the model to capture both generic low-rank phenomena, as well as any behavior related to the side information. The hyperparameters involved in the nuclear-norm constraints can be optimized through cross-validation and allow the model to decide how relevant the side information is. However, a single given matrix $f_{\bullet,\bullet}$ can correspond to several possible choices of M and Z, thereby limiting the interpretability of the model and the individual terms of the sum (1). Furthermore, the model does not capture user and item biases.

Our model remedies these failings. Firstly, the corresponding predictors can take the following form as a particular case:

$$f_{i,j} = c + u_i + m_j + \mathbf{x}_i^\top M \mathbf{y}_j + z_{i,j}, \tag{2}$$

where c is an unknown constant (corresponding to a global bias of the model), u_i and m_j are the user bias and movie quality terms (i.e. free parameters which are trained based on the sole constraint that the user bias term u_i can only depend on the user i (but not on the item j) and vice versa), $\mathbf{x}_i$ and $\mathbf{y}_j$ are the known side information vectors of the i-th user and j-th movie, whilst M and $Z = (z_{i,j})$ are parameter matrices to which *nuclear norm regularization* is applied. The nuclear norm of a matrix is defined as the sum of its singular values. Regularizing the nuclear norm has a rank-sparsity inducing effect similar to the sparsity inducing effect of LASSO. Thus, our regularizer (introduced formally in equation (5) below) *both* indirectly encourages M, and therefore the term corresponding to be $\mathbf{x}_i^\top M \mathbf{y}_j$ to be low rank (whilst relying on the side information for prediction), *and* encourages the residual term Z to have low rank. In summary, we are able to model biases, side information terms, as well as residual generic low-rank effects in a single, jointly trained model.

Furthermore, we will impose orthogonality constraints that effectively require each term in the sum in (2) to live in separate, mutually orthogonal subspaces. This has two advantages. First, training can be performed for all components simultaneously, as we will show. Second, the variables in (2) admit interpretation. This is because any ground truth matrix can be represented *uniquely* (thanks to the orthogonality conditions) through (2). We thus interpret the magnitude of the summands in (2) as their relevance to the model. For instance, this implies that the user-movie match Z is free of any behavior that could be interpreted as user bias or movie quality.

We note that several specific variations of our model are possible depending on whether or not side information is present, how many distinct types of side information are present, whether or not we want to include user/item biases, etc. Therefore, we rely on a *unified formal framework* to describe all possible variations of these ideas as follows. Our general model's predictors have the following form:

$$F = (f_{i,j}) = \sum_{k=1}^{K} \sum_{l=1}^{L} X^{(k)} M^{(k,l)} (Y^{(l)})^\top. \tag{3}$$

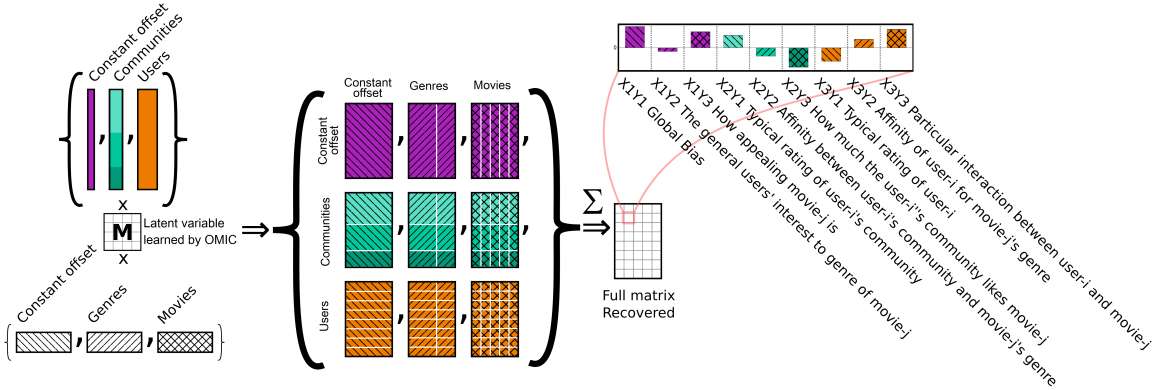

Figure 1: Visualization of orthogonal inductive matrix completion (specifically, the model choice BOMIC+ explained in Section II-D). The model is a sum of matrix terms, each of the form $XMY^\top$. This means each combination of X and Y gives rise to a term in the sum. We interpret the magnitude of this term as its relevance to the prediction.

Here $X^{(1)}, \ldots, X^{(K)}$ and $Y^{(1)}, \ldots, Y^{(L)}$ are so-called "auxiliary matrices", which define the specific model choice with respect to the incorporation (or lack thereof) of biases, side information etc. We now observe that the terms of equation (2) can all be written in the form $X^{(k)}M^{(k,l)}(Y^{(l)})^\top$ for some suitably defined $X^{(k)}, Y^{(l)}$. For instance, if we set X as the identity matrix and Y as a column matrix of all 1's, then the matrix $XMY^\top$ has the user biases as entries: $\mathbf{x}_i M \mathbf{y}_j^\top = u_i$ for all i, j. If we set both X and Y as identity matrices, we obtain the specific user-movie match $\mathbf{x}_i M \mathbf{y}_j^\top = z_{i,j}$ for all i, j. In Figure 1, we illustrate an example with $K = L = 3$ which takes into account user and item biases, as well as side information in the form of partitions of users and items into communities (with the movie communities being genres). This representative example is described in more technical detail in Subsection II-D, where we also further explain how equation (2) can be fully incorporated as a particular case of (3). To ensure the uniqueness of the decomposition (3) and thus enable interpretability, we require that the *columns* of $(X^{(1)}, \ldots, X^{(K)})$ and $(Y^{(1)}, \ldots, Y^{(L)})$ form orthonormal bases of their respective spaces. Hence, we refer to our model as OMIC (Orthogonal Inductive Matrix Completion). Note that the choice of matrices $X^{(k)}$s and $Y^{(l)}$s happens at the level of *model selection* (from a preprocessing of available side information or available insight into the likely structure of the data). Thus, the orthogonality assumption is not an assumption on the ground truth matrix, but a restriction of *model choice*. For fixed k, l, the *rows* of $X^{(k)}$ and $Y^{(l)}$ play a similar role to "side information vectors" in traditional IMC, and are not required to be orthogonal. The choice of auxiliary matrices influences both interpretability and accuracy: on the one hand, the model's output will come in a form that is already divided into components corresponding to each pair $(X^{(k)}, Y^{(l)})$. On the other hand, if the ground truth structure matches the choice of prior directions, accuracy will improve. For instance, as we will discuss further below, choosing $X^{(1)} = (\frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}})^\top$ and $X^{(2)}$ to be the complement of $X^{(1)}$ in $\mathbb{R}^m$ corresponds to the assumption that *item biases* are important: this will both (1) *increase accuracy* if item biases indeed play a role in the ground truth; and (2) in any case *direct interpretability* in the direction of the disentanglement of item biases from other low-rank effects in the solution given by the algorithm. If we are given raw side information matrices X, Y, it also makes sense to create an instance of our model where $X^{(1)}$ and $Y^{(1)}$ are orthogonalized versions of X and Y [1]. In that case we obtain an IMC-type model where the solution consists of four terms depending on whether side information was used for users and for items.

Three specific choices of auxiliary matrices $X^{(k)}$s and $Y^{(l)}$s yield especially interesting models with favourable properties in terms optimization and interetability. We develop them in greater detail and we refer to the corresponding models as BOMIC, OMIC+ and BOMIC+.

Our contributions can be summarized as follows:

1) We propose orthogonal inductive matrix completion (OMIC), a class of learning algorithms for matrix completion, which give rise to interpretable solutions, for instance, in terms of the amount of user bias, movie quality, and community effects in a learned matrix.
2) We propose an efficient optimization algorithm. Furthermore, we prove its convergence and give upper bounds on its convergence rate. See Theorems II.1 and II.2.
3) We present in more detail *three especially interesting models*, **BOMIC**, **OMIC+**, and **BOMIC+**, which are particular cases of our framework. For all three cases, we provide a scalable implementation (explained in Algorithms 1 and 3) of our algorithm that allows us to work on large datasets.
4) For the three most relevant models mentioned above, we prove generalisation bounds both in the case of a sampling distribution with uniform marginals, and in the distribution-free case (where we make no assumption on the ground truth distribution). The better the model matches the ground truth, the tighter the bounds.
5) Our empirical analysis shows that OMIC exhibits better performance and flexibility across the whole spectrum of varying quality of side information. On a large array of real data, OMIC surpasses the state-of-the-art in terms of accuracy, with the added benefit of interpretability.

A. Related work

The idea of training user biases, either as a preprocessing step or jointly with a model was frequent in the pre-SoftImpute days [10], [11]. In [18], time-dependent user and item biases were incorporated into a maximum margin matrix factorization framework[2]

[1] And $X^{(2)}$ and $Y^{(2)}$ are chosen to satisfy the orthogonality conditions

[2] A close cousin of nuclear norm minimisation, cf. appendix

with both the biases and the low-rank residuals continuously retrained alternatingly to account for the time variations.

The idea of training a side information term $XMY^\top$ jointly with a residual term Z was expressed in [17], which is the most related paper to ours. We note that only this specific case was studied there, and that no orthogonality/interpretability constraint was imposed, and thus no imputation algorithm was developed (alternating optimization and gradient descent were used instead). Furthermore, the generalization bounds obtained present differences due to the special nature of our side information and auxiliary matrices.

In [19] and [16], the authors study a model equivalent to a single inductive matrix completion term with non-orthogonal side information and prove bounds in the uniform sampling regime. None of these works contain either a sum of IMC terms, cross-term orthogonality constraints such as ours, user/item biases, or any distribution-free generalization bounds.

In [20], the authors introduce an interesting model with similarities (and differences) to both [17] and the present work. First, the matrices X and Y are augmented by column vectors of ones resulting in the matrices $\bar{X} = [X, \mathbf{1}]$ and $\bar{Y} = [Y, \mathbf{1}]$. Predictors then take the form $E = \bar{X}M(\bar{Y})^\top + \Delta$, with nuclear norm regularisation imposed on E and L^1 (or nuclear norm) regularisation on M, and Frobenius norm regularization imposed on Δ, with the constraint that $P_\Omega(E) = R_\Omega$. In [21], the authors solve an explicit rank minimization problem under linear constraints on the matrix (this problem is now commonly referred to as 'Matrix Regression' (see also [22])). In [23], the authors propose a very general optimization framework that encompasses both the matrix regression problem mentioned above and low rank matrix completion, as well one-bit matrix completion.

II. Description of the model and optimization procedure

We always assume that we have an $m \times n$ matrix R whose entries are partially revealed to us. In this section we assume that the entries are observed without noise for notational simplicity[3], whilst in the theoretical results from the next section, we will deal with the more general case of noisy observations. The set of revealed entries is denoted by $\Omega \subset \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\}$, and $\Omega^\perp$ denotes the complement $\{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\} \backslash \Omega$ (i.e., the set of unobserved entries). The projection on the set of matrices with all entries on $\Omega^\perp$ being zero is denoted by P_Ω. $P_{\Omega^\perp}$ is defined similarly. We denote the matrix of observed entries by R_Ω. As explained in the introduction, our model requires choosing some auxiliary matrices $X^{(k)} \in \mathbb{R}^{m \times d_k^{(1)}}$, $Y^{(l)} \in \mathbb{R}^{n \times d_k^{(2)}}$ representing prior knowledge about the problem. The columns of the auxiliary matrices are assumed to form an orthonormal basis of their respective spaces, i.e.

$$\begin{aligned} &\sum_{i=1}^{m} X_{i,j_1}^{(k_1)} X_{i,j_2}^{(k_2)} = \delta_{k_1,k_2}\delta_{j_1,j_2}; \\ &\sum_{i=1}^{n} Y_{i,j_1}^{(l_1)} Y_{i,j_2}^{(l_2)} = \delta_{l_1,l_2}\delta_{j_1,j_2}; \\ &\operatorname{span}_{k,j}(X_{\bullet,j}^{(k)}) = \mathbb{R}^m \qquad \text{and} \\ &\operatorname{span}_{l,j}(Y_{\bullet,j}^{(l)}) = \mathbb{R}^n. \end{aligned} \tag{4}$$

A. Orthogonal inductive matrix completion

The general form of the optimization problem we consider is as follows:

$$\begin{aligned} &\min_M \quad \mathcal{L}(\Omega, M, \Lambda) \qquad \text{with} \\ &\mathcal{L}(\Omega, M, \Lambda) = \sum_{k=1}^{K}\sum_{l=1}^{L} \lambda_{k,l} \|M^{(k,l)}\|_* \\ &+ \frac{1}{2} \sum_{(i,j)\in\Omega} \ell\left[R_{i,j}, \left(\sum_{k=1,l=1}^{K,L} X^{(k)} M^{(k,l)} (Y^{(l)})^\top \right)_{i,j} \right]. \end{aligned} \tag{5}$$

Here the $\lambda_{k,l}$s are non-negative tunable hyper-parameters. Note that the objective is *convex*, but *not strongly convex*. In particular, any stationary point is a solution, but the solution may not be unique (though all solutions correspond to the same (optimal) value of the objective function). In practical implementations such as the specific algorithm and implementation we present below, ℓ is set to the square loss[4]: $\ell(x, x') = |x - x'|^2$.

We note that our orthogonality constraints offer the additional advantage of interpretability: it is easy to check that the spaces $\left\{ X^{(k)} M (Y^{(l)})^\top \middle| M \in \mathbb{R}^{d_1^{(k)} \times d_2^{(l)}} \right\}$ are orthogonal with respect to the Frobenius inner product. Thus, each ground truth matrix R has a unique representation $R = \sum_{k=1}^{K}\sum_{l=1}^{L} X^{(k)} R^{(k,l)} (Y^{(l)})^\top$. We provide the detailed proof of these results in Appendix B (see Proposition B.1). The norms $\left\| X^{(k)} R^{(k,l)} (Y^{(l)})^\top \right\|_{\mathrm{Fr}} = \left\| R^{(k,l)} \right\|_{\mathrm{Fr}}$ can be interpreted as the relative importance

[3] Algorithmic aspects are unchanged

[4] If this loss function is used and the matrix is *fully observed*, the problem becomes strongly convex

of the auxiliary pairs $X^{(k)}, Y^{(l)}$. Furthermore, the tuning of the coefficients $\lambda_{k,l}$ can be assisted by human knowledge. In particular we can dramatically reduce our cross-validation needs by tying many parameters with each other, as well as by setting other parameters corresponding to easy to learn quantities (such as user/community biases) to zero.

In the next three Sections II-B, II-C and II-D below, we present the three most significant instances of our model class, which incorporate user/item biases and/or community side information.

B. First example: jointly trained user/item biases (BOMIC)

One notable example of this setting provides a way to train a low-rank matrix completion model together with user biases, as discussed in the introduction: if we set $X^{(1)} = (\frac{1}{\sqrt{m}}, \dots, \frac{1}{\sqrt{m}})^\top$, $Y^{(1)} = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})^\top$ and set $X^{(2)}$ (resp. $Y^{(2)}$) to be the orthogonal complement of $X^{(1)}$ (resp. $Y^{(1)}$) in $\mathbb{R}^m$ (resp. $\mathbb{R}^n$), then our model (5) is equivalent to optimizing a prediction function $f_{i,j} = c + u_i + m_j + Z_{i,j}$ where Z is constrained to have columns and rows summing to zero, and the regularizer is

$$\lambda_1|c| + \lambda_2\|\mathbf{u}\|_2 + \lambda_3\|\mathbf{m}\|_2 + \lambda_4\|Z\|_*.$$

One would typically set $\lambda_2 = \lambda_3$ and $\lambda_1 = 0$.

Here, the terms $X^{(1)}M^{(1,1)}(Y^{(1)})^\top$ and $X^{(1)}M^{(1,2)}(Y^{(2)})^\top + X^{(2)}M^{(2,1)}(Y^{(1)})^\top$, correspond to a general, matrix-wise bias, and a combination of user/item specific biases respectively. Meanwhile, the term $X^{(2)}M^{(2,2)}(Y^{(2)})^\top$ represents purely bias-free low-rank effects: an entry of $X^{(2)}M^{(2,2)}(Y^{(2)})^\top$ will be large if the item and user are particularly well-fitted to each other, independently of the general behavior of either user or item. This can be especially interesting in terms of interpretability, or if each user must be paired with a single item. We refer to this particular case of our model as BOMIC (Bias-OMIC).

C. Second example: OMIC+

Another highly representative example is the case where we are given side information about the users and items in the form of communities, where each user (resp. item) belongs to exactly one user (resp. item) community. In this situation, we construct the columns of $X^{(1)}$ (resp. $Y^{(1)}$) as normalized indicator functions of the user (resp. item) communities. The columns of $X^{(2)}$ (resp. $Y^{(2)}$) can then be chosen as any orthonormal basis of the orthogonal complement of $X^{(1)}$ (resp. $Y^{(1)}$) in $\mathbb{R}^m$ (resp. $\mathbb{R}^n$). In the case where the user (resp. item) communuties each contain a fixed number of members, the model becomes equivalent to the following optimization problem (up to multiplicatve constants in the regularising parameters):

$$\min_{C,M,U,Z} \mathcal{L} \quad \text{with} \quad \mathcal{L} = \tag{6}$$

$$\sum_{(i,j)\in\Omega} |C_{f(i),g(j)} + M_{i,g(j)} + U_{f(i),j} + Z_{i,j} - R_{i,j}|^2$$

$$+ \lambda_1\|C\|_* + \lambda_{12}\|M\|_* + \lambda_{21}\|U\|_* + \lambda_{22}\|Z\|_*, \tag{7}$$

subject to

$$\sum_{i\in f^{-1}(u)} M_{i,v} = 0 \quad \forall u \leqslant d_1, v \leqslant d_2,$$

$$\sum_{j\in g^{-1}(v)} U_{u,j} = 0 \quad \forall u \leqslant d_1, v \leqslant d_2,$$

$$\sum_{i\in f^{-1}(u)} Z_{i,j} = 0 \quad \forall j \leqslant n,$$

$$\text{and} \sum_{j\in g^{-1}(v)} Z_{i,j} = 0 \quad \forall i \leqslant m. \tag{8}$$

Here, the function $f : \{1, 2, \dots, m\} \to \{1, 2, \dots, d_1^1\}$ (resp. $g : \{1, 2, \dots, n\} \to \{1, 2, \dots, d_2^1\}$) assigns to each user (resp. item) its community. In terms of interpretability, note that in the predictors $C_{f(i),g(j)} + M_{i,g(j)} + U_{f(i),j} + Z_{i,j}$ above, the contribution from C corresponds to effects that only depend on the community of the user and the community of the item, the contribution from M (resp. U) corresponds to 'specific user-item community' (resp. 'specific item-user community') effects, whilst the contribution from Z corresponds to effects that are purely specific to the user and the item (independently of their respective communities).

D. Third example: BOMIC+

We note that several variations of the ideas for the construction of the auxiliary matrices $X^{(k)}$ and $Y^{(l)}$ as above are useful in practice. An important instance is BOMIC+, which combines both of the ideas above by incorporating *both* user and item biases *and* community side information. This is the specific model explained in Figure 1, and is further empirically investigated in the experiments section. Given community side information, we define our auxiliary matrices $X^{(k)}$ and $Y^{(l)}$ as follows:

- $X^{(1)}$ and $Y^{(1)}$ are constructed as in the case of BOMIC (II-B), i.e. $X^{(1)} = (\frac{1}{\sqrt{m}}, \dots, \frac{1}{\sqrt{m}})^\top$, $Y^{(1)} = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})^\top$;

- Let X (resp. Y) denote a matrix whose columns are indicator functions of the user (resp. item) communities. The columns of $X^{(2)}$ (resp. $Y^{(2)}$) form an orthonormal basis of the space $\{v \in \mathrm{span}(X) : \langle v, X^{(1)} \rangle = 0\}$ (resp. $\{v \in \mathrm{span}(Y) : \langle v, Y^{(1)} \rangle = 0\}$), where $\mathrm{span}(X)$ (resp. $\mathrm{span}(Y)$) denotes the span of the columns of X (resp. Y).
- Finally, the columns of $X^{(3)}$ (resp. $Y^{(3)}$) form an orthonormal basis of the orthogonal complement of the columns of $(X^{(1)}, X^{(2)})$ (resp. $(Y^{(1)}, Y^{(2)})$) in $\mathbb{R}^m$ (resp. $\mathbb{R}^n$).

Thus, this model corresponds to equation (2), together with some orthogonality constraints. Indeed, $X^{(1)}M^{(1,1)}(Y^{(1)})^\top$ is a constant and corresponds to c. Furthermore, all the rows of $X^{(1)}M^{(1,3)}(Y^{(3)})^\top$ (resp. columns of $X^{(3)}M^{(3,1)}(Y^{(1)})^\top$) are equal, so that the term $X^{(1)}M^{(1,3)}(Y^{(3)})^\top$ (resp. $X^{(3)}M^{(3,1)}(Y^{(1)})^\top$) corresponds to $\mathbf{u}$ (resp. $\mathbf{m}$). $X^{(2)}M^{(2,2)}(Y^{(2)})^\top$ is the side information term corresponding to $\mathbf{x}_i M \mathbf{y}_j^\top$ in (2). Meanwhile, the remaining terms $X^{(1)}M^{(1,2)}(Y^{(2)})^\top + X^{(2)}M^{(2,1)}(Y^{(1)})^\top + X^{(3)}M^{(3,2)}(Y^{(2)})^\top + X^{(2)}M^{(2,3)}(Y^{(3)})^\top + X^{(3)}M^{(3,3)}(Y^{(3)})^\top$ correspond to the term $Z_{i,j}$ from equation (2), further refined into specific components distinguishing effects involving (1) only the side information of the users but not that of the items (or vice versa), (2) interactions between user bias and item side information (or vice versa) or (3) no side information or biases whatsoever.

E. Optimization algorithm

In this subsection, we propose an iterative imputation algorithm to solve the problem (5) with the square loss. The first step is to develop a method to solve (5) in the case where $\Omega = \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\}$, the so-called "fully known case". The final solution can then be obtained by iteratively using this method.

1) The fully known case: Recall the singular value thresholding operator S_λ from [5] and [24], which is defined by $S_\lambda(Z) = \sum_{i=1}^r (\lambda_i - \lambda)_+ v_i w_i^\top$, where $Z = \sum_{i=1}^r \lambda_i v_i w_i^\top$ is the singular value decomposition (SVD) of Z.

In our case, we now introduce the *Generalized singular value thresholding operator* S_Λ, which, for any set of parameters $\Lambda = (\lambda_{k,l})_{\substack{k \leqslant K \\ l \leqslant L}}$, and given a set of auxiliary matrices $X^{(k)}, Y^{(l)}$ (satisfying the orthogonality conditions (4)), is defined by

$$S_\Lambda(Z) = \sum_{k=1}^K \sum_{l=1}^L X^{(k)} S_{\lambda_{k,l}}(M^{(k,l)})(Y^{(l)})^\top, \quad (9)$$

where $M^{(k,l)} = (X^{(k)})^\top Z (Y^{(l)})$, which ensures $Z = \sum_{k=1}^K \sum_{l=1}^L X^{(k)} M^{(k,l)} (Y^{(l)})^\top$.

Note that $S_\Lambda(Z)$ is well-defined since the spaces $\mathcal{S}_{k,l} = \left\{X^{(k)} M (Y^{(l)})^\top\right\} \subset \mathbb{R}^{m \times n}$ are orthogonal with respect to the Frobenius inner product, and in particular, linearly independent.

Proposition II.1. *The definition in equation* (9) *is equivalent to the following:*

$$S_\Lambda(Z) = \sum_{k=1}^K \sum_{l=1}^L X^{(k)} S_{\lambda_{k,l}} \left((X^{(k)})^\top Z Y^{(l)} \right) (Y^l)^\top. \quad (10)$$

Furthermore, $\tilde{Z} = S_\Lambda(Z)$ *is the solution to the following optimization problem:*

$$\min \frac{1}{2} \|\tilde{Z} - Z\|_{\mathrm{Fr}}^2 + \sum_{k=1}^K \sum_{l=1}^L \lambda_{k,l} \left\| (X^{(k)})^\top Z Y^{(l)} \right\|_*, \quad (11)$$

or equivalently

$$\min \frac{1}{2} \|\tilde{Z} - Z\|_{\mathrm{Fr}}^2 + \sum_{k=1}^K \sum_{l=1}^L \lambda_{k,l} \left\| M^{(k,l)} \right\|_*, \quad (12)$$

$$\text{subject to} \quad \tilde{Z} = \sum_{k=1}^K \sum_{l=1}^L X^{(k)} M^{(k,l)} (Y^{(l)})^\top. \quad (13)$$

2) The OMIC algorithm: For any fixed set of hyperparameters Λ and auxiliary matrices $X^{(k)}, Y^{(l)}$ (for $k \leqslant K, l \leqslant L$), the final solution to the optimization problem (5) is obtained iteratively: at each step, a target matrix is constructed by setting the entries of Ω to the observed values and imputing the values of the previous iteration to the entries of $\Omega^\perp$. We then apply the fully known case (II.1) to this target matrix to reach the next iterate. This algorithm converges for any initial (0th iteration) matrix.

However, if several values of Λ must be calculated, we can use warm starts to improve efficiency. Algorithm 1 below does this in the case where the range of values for Λ is a product $\mathcal{V} = \prod_{k,l} \mathcal{V}_{k,l}$ where the $\mathcal{V}_{k,l}$ are finite sets of candidate values for $\lambda_{k,l}$ (ordered in increasing or decreasing order): initial estimates of $M_{k,l}$ for each value of $\lambda_{k,l}$ are calculated by setting each $\lambda_{k',l'}$ to infinity for $k' \neq k, l' \neq l$ and solving the full problem (5) in this case[5]. Furthermore, each of those sets of $M_{k,l}$ are calculated using warm starts along the sequence of $\lambda_{k,l} \in \mathcal{V}_{k,l}$. For any real

[5] "Setting $\lambda_{k',l'}$ to infinity" amounts to not including the (k', l') term in the sum (3) which defines our predictors. Thus our warm starts are computed by training each term in that sum independently.

number λ, $p_{k,l}(\lambda)$ denotes the set of hyperparameters Λ with $\Lambda_{k,l} = \lambda$ and $\Lambda_{k',l'} = \infty$ otherwise. Note also that this algorithm depends on the auxiliary matrices $X^{(k)}$ and $Y^{(l)}$ through the generalized singular value thresholding operator S_Λ, defined in equation (9) above.

Note that $M^{(k,l)}$ is of dimension $d_1^k \times d_2^l$. Whenever one of d_1^k or d_2^l is small, the computation of SVDs is trivial. For instance, in the specific case of BOMIC, matrices $M^{(k,l)}$ are vectors whenever $k = 1$ or $l = 1$, thus calculating the SVD simply amounts to normalizing a vector. Therefore, although our algorithm theoretically requires KL SVD operations at each iteration, in fact, most of them are trivial. We refer the reader to Appendix D, for a more thorough explanation and an empirical evaluation of the lack of any extra computational burden from the trivial SVDs. Furthermore, in practice, a rank-restricted version of the SVD operation can be employed, together with a suitable warm-start strategy. In Subsection II-F, we develop novel techniques to optimize the computational and memory burden of the algorithm in the most interesting and practically relevant particular cases.

Algorithm 1 OMIC

INPUT: R_Ω, set of regularizing parameters $\mathcal{V} = \prod_{k,l} \mathcal{V}_{k,l}$

OUTPUT: Set of recovered matrices Z^Λ for all $\Lambda \in \mathcal{V}$

for $k \in \{1, 2, \dots, K\}$ **do**
  for $l \in \{1, 2, \dots, L\}$ **do**
    Initialize $Z^{\text{new}} \leftarrow 0$
    for $\lambda \in \mathcal{V}_{k,l}$ **do**
      repeat
        $Z^{\text{old}} \leftarrow Z^{\text{new}}$
        $Z^{\text{new}} \leftarrow S_{p_{k,l}(\lambda)}\left(R_\Omega + P_{\Omega^\perp}(Z^{\text{old}})\right)$
      until converged
      $Z^{k,l,\lambda} \leftarrow Z^{\text{new}}$
    end for
  end for
end for
for $\Lambda \in \mathcal{V}$ **do**
  $Z^{\text{new}} \leftarrow \sum_{k,l=1}^{K,L} Z^{k,l,\lambda_{k,l}}$
  repeat
    $Z^{\text{old}} \leftarrow Z^{\text{new}}$
    $Z^{\text{new}} \leftarrow S_\Lambda\left(R_\Omega + P_{\Omega^\perp}(Z^{\text{old}})\right)$
  until converged
  $Z^\Lambda \leftarrow Z^{\text{new}}$
end for
return Z^Λ for $\Lambda \in \mathcal{V}$

3) Convergence guarantees: Our algorithm enjoys convergence guarantees, which we present here. Here, we fix a Λ and assume that we perform the iterative imputation procedure in the algorithm above starting from $Z^0 = 0$, with (for each $i \geqslant 0$)

$$Z^{i+1} = S_\Lambda\left(P_\Omega(R) + P_{\Omega^\perp}(Z^i)\right). \tag{14}$$

We have the following two results, whose proofs are left to the Appendix.

Theorem II.1. *Consider our general setting with auxiliary matrices satisfying the conditions* (4) *and the operator S_Λ defined accordingly. The sequence Z^i defined in* (14) *converges to a solution Z^∞ of the optimization problem* (5) *with the squared loss. In particular, the loss $\mathcal{L}$ converges to the minimum $\mathcal{L}^*$ of optimization problem* (5)*.*

Theorem II.2. *Let $\mathcal{L}^*$ be the minimum value of the loss $\mathcal{L}$ from problem* (5)*. For every fixed Λ, the sequence Z^i has the following worst-case asymptotic convergence:*

$$\mathcal{L}(Z^i) - \mathcal{L}(Z^\infty) = \mathcal{L}(Z^i) - \mathcal{L}^* \leqslant \frac{2\|Z^0 - Z^\infty\|_{\text{Fr}}^2}{i+1}, \tag{15}$$

where Z^∞ is the limit of the sequence Z^i.

Remark: Note that although the RHS depends on the limit Z^∞, it can be further bounded above as follows

$$\frac{2\|Z^0 - Z^\infty\|_{\text{Fr}}^2}{i+1} \leqslant \frac{2\max_{Z^* \in \mathcal{S}} \|Z^0 - Z^*\|_{\text{Fr}}^2}{i+1}$$

where $\mathcal{S} = \arg\min_Z(\mathcal{L}(Z))$ is the *set* of solutions of Problem (5)[6]. Thus, the rate $1/(i+1)$ holds.

F. A scalable algorithm for BOMIC+

The main computational step at each iteration of the algorithm we propose to solve (5) is the calculation of the solution to an instance of the fully known case, which can be obtained via our *generalized singular value thresholding operator* S_Λ. Observe that the sum of the numbers of columns of all $X^{(k)}$ (resp. $Y^{(l)}$) is m (resp. n). Thus, if m and n are large (which is often the case in RSs contexts) it is infeasible to even store all the auxiliary matrices in memory, let alone perform operations on them directly. The same problem can occur with some of the latent matrices $M^{(k,l)}$. It is easy to see that the largest $M^{(k,l)}$ has at least $m/K \times n/L$ entries which is also very large.

In this subsection, we show how to circumvent this difficulty in the specific case of BOMIC+ where

[6]Note the solution is not necessarily unique, as we discuss in greater detail at the end of Appendix C.

$K = L = 3$ and the auxiliary matrices are defined in Section II-D assuming the side information X, Y consists of indicator functions of communities. For instance, the columns of Y could represent sets of movies belonging to a specific genre, whilst the columns of X could represent countries, genders or professions for users. Our strategy heavily relies on both the "sparse-plus-low-rank" structure present in the target matrices of the "fully known problem" solved at each iteration, as well as the specific structure of community side information.

First, we define some matrices $\widetilde{X}^{(1)}$, $\widetilde{X}^{(2)}$, $\widetilde{X}^{(3)}$, $\widetilde{Y}^{(1)}$, $\widetilde{Y}^{(2)}$, $\widetilde{Y}^{(3)}$ as follows: $\widetilde{X}^{(1)} = X^{(1)}$, $\widetilde{Y}^{(1)} = Y^{(1)}$, $\widetilde{X}^{(2)} = \mathrm{normalize}(X)$, $\widetilde{Y}^{(2)} = \mathrm{normalize}(Y)$, $\widetilde{X}^{(3)} = \mathrm{Id}$ and $\widetilde{Y}^{(3)} = \mathrm{Id}$. Here, normalize denotes the operation of normalizing each column. Note that X (resp. Y) is a sparse matrix composed of the indicator functions of user (resp. item) communities. Therefore, the matrices $\widetilde{X}^{(1)}$, $\widetilde{X}^{(2)}$,$\widetilde{X}^{(3)}$,$\widetilde{Y}^{(1)}$,$\widetilde{Y}^{(2)}$and $\widetilde{Y}^{(3)}$ can easily be stored in memory, and it is easy to multiply them by arbitrary vectors on either side.

To see how these matrices will help us execute BOMIC+, observe first that due to the rotational invariance of SVDs, the operator S_Λ can be rewritten as

$$S_\Lambda(Z) = \sum_{k=1}^{3}\sum_{l=1}^{3} S_{\lambda_{k,l}}\left(X^{(k)}(X^{(k)})^\top Z Y^{(l)}(Y^{(l)})^\top\right). \tag{16}$$

Thus, for a given Z, calculating $S_\Lambda(Z)$ boils down to computing the SVD of the matrices $H^{(k,l)} = X^{(k)}(X^{(k)})^\top Z Y^{(l)}(Y^{(l)})^\top$, which are the projections of Z on the spaces corresponding to each pair of auxiliary matrices. We perform the SVD computation through a rank-restricted alternating least squares algorithm (ALS, see Algorithm 2). This requires an efficient strategy to compute $H^{(k,l)}W_1$ and $W_2H^{(k,l)}$, where W_1 and W_2 are any real conformable matrices.

We now observe that for any orthogonal matrix U, if $W_1 \in \mathbb{R}^{n\times r}$, $UU^\top W_1$ is the projection of W_1 on the span of the columns of U. Crucially, if V is an orthogonal matrix with $\mathrm{span}(V) \subset \mathrm{span}(U)$, then for any $W_1 \in \mathbb{R}^{n\times r}$, $UU^\top W_1 - VV^\top W_1$ is the projection of W_1 on the orthogonal complement of V in U. Applying this to the matrices $Y^{(l)}$ and $\widetilde{Y}^{(l)}$, we obtain, for all $l \leqslant 3$:

$$\begin{aligned} &Y^{(l)}(Y^{(l)})^\top W_1 \\ &= \widetilde{Y}^{(l)}(\widetilde{Y}^{(l)})^\top W_1 \quad -\widetilde{Y}^{(l-1)}(\widetilde{Y}^{(l-1)})^\top W_1. \end{aligned} \tag{17}$$

Similarly, for any $W_2 \in \mathbb{R}^{m\times r}$ and $k \leqslant 3$,

$$\begin{aligned} &X^{(k)}(X^{(k)})^\top W_2 \\ &= \widetilde{X}^{(k)}(\widetilde{X}^{(k)})^\top W_2 \quad -\widetilde{X}^{(k-1)}(\widetilde{X}^{(k-1)})^\top W_2. \end{aligned} \tag{18}$$

Remark II.2. *The operation $W_1 \leftarrow Y^{(3)}(Y^{(3)})^\top W_1$ performed as above can be visualized intuitively: it corresponds to removing from each component of each column $W_1{}_{\bullet,i}$ the average of the components in the same community (in the same column).*

$$\forall i \leqslant m, \quad v_i \leftarrow v_i - \frac{\sum_{j\in c_i} v_j}{\#(c_i)}.$$

With the above techniques in our tool kit, we can now present our algorithm for calculating $H^{(k,l)}W_1$. We will divide the task into three steps.

First, we evaluate $\widetilde{W_1} = Y^{(l)}[(Y^{(l)})^\top W_1]$ using (17).

Next, observe that as a byproduct of the iterative imputation procedure, Z can be decomposed as the sum of a sparse matrix Z_{Sp} and a low-rank matrix Z_{LR} as follows:

$$Z = Z_{Sp} + Z_{LR} = Z_{Sp} + U_{LR}\left[D_{LR}{V_{LR}}^\top\right]. \tag{19}$$

Using this decomposition, it is easy to calculate the quantity $\widetilde{\widetilde{W_1}} = Z\widetilde{W_1}$ as follows:

$$\widetilde{\widetilde{W_1}} = Z\widetilde{W_1} = Z_{Sp}\widetilde{W_1} + U_{LR}D_{LR}({V_{LR}}^\top\widetilde{W_1}). \tag{20}$$

Finally, we calculate $H^{(k,l)}W_1 = \left(X^{(k)}[(X^{(k)})^\top\widetilde{\widetilde{W_1}}]\right)^\top$ using (18). Symmetrically and with the same arguments we can compute $W_2H^{(k,l)}$.

Alg. 2 (based on [25]) describes how to compute $S_\Lambda(Z)$ for a fixed hyperparameter set Λ and Alg. 3 is our fully scalable implementation of Alg. 1 for the BOMIC+ case with community side information. In practice, we can further use warm start strategies such as the one presented in Alg. 1 to speed up convergence.

Remark: The convergence of Alg. 2 follows from that of Alg. 2.1 in [25]: for each combination (k, l) [7], of which there are finitely many, the while loop from lines 11 to 22 essentially runs Algorithm 2.1 from [25] on the component matrix $R^{(k,l)} = (X^{(k)})^\top R Y^{(l)}$, thus the full algorithm converges. The convergence of Alg. (3) then follows from (1) the convergence of Alg. 2 (which is used inside the while loop between lines 3 and 6) together with (2) the convergence of Alg. II-C, which is established in Theorem II.1

Remark: To avoid manipulating or storing large matrices, one must perform the operations in the

[7] leaving aside the extra implementation details required for our specific model

correct order, which is defined by the brackets. This remark applies in particular to Algorithms 2 and 3.

Algorithm 2 (S_Λ)**-ALS**: Generalized restricted truncated singular value thresholding via alternating least squares
INPUT: $Z \in \mathbb{R}^{m\times n}$ decomposed as in (19); thresholding parameters Λ, maximum rank r
OUTPUT: $S_\Lambda(Z)$ represented in storable low-rank format as $(\mathfrak{U}, \mathrm{diag}(\Sigma), \mathfrak{V})$ such that $S_\Lambda(Z) = \mathfrak{U}\,\mathrm{diag}(\Sigma)\mathfrak{V}^\top$

Procedure Projection($E^{(1)}$,$E^{(0)}$,Θ)
return $E^{(1)}[(E^{(1)})^\top\Theta] - E^{(0)}[(E^{(0)})^\top\Theta]$
end Procedure
$\mathfrak{U} \leftarrow \Sigma \leftarrow \mathfrak{V} \leftarrow NULL$
$\widetilde{X}^{(0)} \leftarrow 0, \widetilde{Y}^{(0)} \leftarrow 0$
for k in (1..3) **do**
  for l in (1..3) **do**
    $U \leftarrow$ random orthogonal $m \times r$ matrix
    $D \leftarrow \mathrm{Id}_{r\times r}$
    $A \leftarrow UD$
    while $AB^\top$ not converged **do**
      $\Theta \leftarrow UD(D^2 + \lambda_{k,l}I)^{-1}$
      $\widetilde{\Theta} \leftarrow$ Projection($\widetilde{X}^{(k)}, \widetilde{X}^{(k-1)}, \Theta$)
      $\widetilde{\widetilde{\Theta}} = \widetilde{\Theta}^\top Z_{Sp} + [(\widetilde{\Theta}^\top U_{LR})D_{LR}]V_{LR}{}^\top$
      $\tilde{B} \leftarrow$ Projection($\widetilde{Y}^{(l)}, \widetilde{Y}^{(l-1)}, \widetilde{\widetilde{\Theta}}^\top$)
      Compute the SVD of $\tilde{B}D = \tilde{V}\tilde{D}^2R^\top$ and attribute $V \leftarrow \tilde{V}$, $D \leftarrow \tilde{D}$ and $B \leftarrow VD$
      $\Theta \leftarrow VD(D^2 + \lambda_{k,l}I)^{-1}$
      $\widetilde{\Theta} \leftarrow$ Projection($\widetilde{Y}^{(l)}, \widetilde{Y}^{(l-1)}, \Theta$)
      $\widetilde{\widetilde{\Theta}} = Z_{Sp}\widetilde{\Theta} + U_{LR}[D_{LR}(V_{LR}{}^\top\widetilde{\Theta})]$
      $\tilde{A} \leftarrow$ Projection($\widetilde{X}^{(k)}, \widetilde{X}^{(k-1)}, \widetilde{\widetilde{\Theta}}$)
      Compute the SVD of $\tilde{A}D = \tilde{U}\tilde{D}^2\tilde{R}^\top$ and attribute $U \leftarrow \tilde{U}$, $D \leftarrow \tilde{D}$ and $A \leftarrow UD$
    end while
    $\widetilde{\Theta} \leftarrow$ Projection($\widetilde{Y}^{(l)}, \widetilde{Y}^{(l-1)}, V$)
    $\widetilde{\widetilde{\Theta}} = Z_{Sp}\widetilde{\Theta} + U_{LR}[D_{LR}(V_{LR}{}^\top\widetilde{\Theta})]$
    $M \leftarrow$ Projection($\widetilde{X}^{(k)}, \widetilde{X}^{(k-1)}, \widetilde{\widetilde{\Theta}}$)
    Compute the SVD of M: $M = \bar{U}D_\sigma\bar{R}^\top$.
    $\mathfrak{U} \leftarrow$ CONCAT($\mathfrak{U}, \bar{U}$)
    $\Sigma \leftarrow$ CONCAT($\Sigma, \big((\sigma_1 - \lambda_{k,l})_+, (\sigma_2 - \lambda_{k,l})_+, \ldots, (\sigma_r - \lambda_{k,l})_+\big)$)
    $\mathfrak{V} \leftarrow$ CONCAT($\mathfrak{V}, V\bar{R}$)
  end for
end for
return $\mathfrak{U}; \mathrm{diag}(\Sigma); \mathfrak{V}$

Complexity analysis: Treating K and L as constants, our algorithm can achieve $O(|\Omega|r + (m+n)r^2)$ complexity all three important cases OMIC+, BOMIC and BOMIC+. We note that, as explained above, although in principle the computational burden is multiplied by KL (which can be as much as 9 in the case of BOMIC+), in practice, further refinements can help the algorithm perform at the same speed as SoftImpute, as is hinted at in Subsection II-E2. We refer the reader to Appendix D for more details.

Algorithm 3 Scalable BOMIC+
INPUT R_Ω, regularizing parameters $\Lambda \in \mathbb{R}^{3\times 3}$, maximum rank r
OUTPUT: Recovered matrix Z

$Z_s \leftarrow R_\Omega$
$Z_{LR} \leftarrow \{U_{LR} \leftarrow 0, D_{LR} \leftarrow 0, V_{LR} \leftarrow 0\}$
while Not converged **do**
  $Z_s \leftarrow R_\Omega - P_\Omega(Z_{LR})$
  $Z_{LR} \leftarrow (S_\lambda)$-ALS($Z = \{Z_s, Z_{LR}\}$)
end while
return $Z \leftarrow Z_{LR}$

III. Generalization bounds

In this section, we present some generalization bounds for our model in the three relevant cases BOMIC, OMIC+ and BOMIC+.

A. Distribution-free bounds

We begin with a distribution-free approach, meaning that we do not assume anything about the sampling distribution. In particular, the bounds in this subsection behave worse than the corresponding bounds under uniform sampling assumptions such as the ones in the celebrated works [7], [6].

We first focus on generalization bounds which apply when the side information corresponds to user/item biases and/or community side information (OMIC+).

We will write $C_{k,l}$ for an upper bound on the entries of the ground truth component $X^{(k)}R^{(k,l)}(Y^{(l)})^\top$ where $R^{(k,l)} = (X^{(k)})^\top RY^{(l)}$. Similarly, we will write $r_{k,l}$ for the rank of $R^{(k,l)}$.

Thus, e.g., if $C_{1,2}$ is large, one concludes that in the ground truth matrix, the specific affinities of items in $\{1, 2, \ldots n\}$ to whole communities of users is a significant factor in determining the value of each entry. If $C_{2,2}$ is large, the individual affinities between users and items, independently of their respective communities, is a strong factor.

The following follows from the inequality $\sqrt{x} + \sqrt{y} + \sqrt{z} + \sqrt{t} \leqslant 2\sqrt{x+y+z+t}$ and Theorem E.1 in the Appendix.

Corollary III.1. *Consider the community setting above and assume the user (resp. item) communities are of sizes within a ratio of $O(1)$, as well as that (wlog) $b \geqslant a$, and $m \geqslant n$. For any $\epsilon > 0$, the required number of entries to recover the ground truth matrix within ϵ expected loss (with probability $\geqslant 1 - \delta$) is*

$$O\Bigg((1/\epsilon^2)\Big(C_{1,1}^2 b\sqrt{ar_{1,1}} + C_{1,2}^2 n\sqrt{ar_{1,2}} + C_{2,1}^2 m\sqrt{br_{2,1}} + C_{2,2}^2 m\sqrt{r_{2,2}n} + \log(1/\delta)\Big)\Bigg).$$

Alternatively, using the nuclear norms of the component matrices, we have the following sample complexity bound:

$$O\Bigg((1/\epsilon^2)\Big(\sqrt{b}t_{1,1} + t_{1,2}\sqrt{n} + t_{2,1}\sqrt{m} + t_{2,2}\sqrt{m} + \log(1/\delta)\Big)\Bigg).$$

Note that the above bounds improve with the quality of the side information: the better the ground truth matrix can be approximated by community behaviour, the closer the bound behaves to the bound one would obtain for an $a \times b$ matrix with each user and item being assimilated to its community. Furthermore, the result applies in particular to the BOMIC model from Section II-B by grouping all users (resp. items) into a single community. This yields a bound of the order $(1/\epsilon^2)(C_{1,1}^2 + nC_{1,2}^2 + mC_{2,1} + (\sqrt{n} + \sqrt{m})\sqrt{mnr}C_{2,2}^2 + \log(1/\delta))$ where (e.g.) $C_{2,2}$ is a bound on the bias-free part of the ground truth matrix. Similar bounds hold for the BOMIC+ model II-D (cf. equation (S.70) in the Appendix).

B. Bounds assuming uniform marginals

In this subsection, we present some bounds under the assumption that the sampling distribution has uniform marginals (i.e., whenever an entry is sampled, the probability of choosing an entry in any given row (resp. column) is $1/m$ (resp. $1/n$)). In this case, only the terms containing side information for both users and items present an improvement.

Corollary III.2. *Consider the community setting above (OMIC+) and assume: (1) each user (resp. item) community is of equal size; (2) that (wlog) $b \geqslant a$, and $m \geqslant n$ and (3) the marginals of the sampling distribution are uniform. For any $\epsilon > 0$, (the required number of entries to recover the ground truth matrix within ϵ expected loss is*

$$O\Bigg((1/\epsilon^2)\Big(C_{1,1}^2 br_{1,1}\log(b) + C_{1,2}^2 nr_{1,2}\log(n) + C_{2,1}^2 mr_{2,1}\log(m) + C_{2,2}^2 mr_{2,2}\log(m)\Big)\Bigg).$$

Proof. Follows from Theorem E.3 in Appendix E-B. □

In the supplementary (Section E-D), we experimentally validate the bound on some synthetic data, and observe a good match between the bound and the de facto sample complexities we encounter in practice.

Remark: An interesting observation from the bounds is that the knowledge of the explicit side information vectors (i.e. communities or biases) helps achieve a faster sample complexity (as opposed to the knowledge of the equivalent low-rank constraint) whenever either:

- the first order term $R^{(1,1)}$ is significant; or
- the distribution is arbitrary (doesn't have uniform marginals).

This is explained in detail in Subsection E-C of the appendix. As our synthetic data experiments (cf. Subection IV-C) demonstrate, we can still gain something in practice for moderate-size matrices, even when it comes to cross-terms in the uniform sampling case, but such an improvement cannot be captured at the level of asymptotic results. **Remark:** Note that although the proofs are relatively straightforward, even the generalisation bounds corresponding to a single term do *not* follow from standard results on inductive matrix completion. This is also explained in detail in Section E-C in the appendix.

IV. Experiments

To compare OMIC with the baselines we conducted two experiment strands: synthetic data simulations and real-world applications.

In the first case, we propose a matrix generation procedure that allows us to evaluate how the performance of BOMIC varies in different ground truth regimes: we generated sparsely observed matrices composed of the sum of user and item biases (generic behaviour) and a non-inductive term (specific behaviour). The proportion of each term, the number of observed entries and the distribution of known entries were varied.

In the latter case, we validated our model on five real recommender systems datasets: the Amazon

recommender system dataset, the Douban movie database, the Goodreads spoiler dataset, and two versions of MovieLens. We show that our methods exhibit state-of-the-art performance in all cases.

All the hyperparameter tuning was done through cross-validation. The range of the parameters was adjusted according to each model's needs.

A. Baselines

Our model is a fundamental tool that relies only on the incomplete matrix and some high-level side information and has the benefit of interpretability. We compare our model with other similarly fundamental models such as IMC [12], [17], [13], [16], [19] and Softimpute [5], with the understanding that the basic ideas could be refined and incorporated into more sophisticated recommender systems.

- **SoftImpute (SI)** is a matrix completion method that uses nuclear-norm regularization. This is a standard baseline for non-inductive matrix completion [5].
- **Biased SoftImpute (B-SI)** is a popular approach that consists in first, training user and item biases, and then training the SoftImpute model on the residuals.
- **Inductive Matrix Completion with Noisy Features (IMCNF)** In this model [17] we train a sum of an IMC term and a residual SoftImpute term jointly (via alternating optimization). This model requires side information, and was therefore only considered in real data experiments.

B. Metrics

Let $R \in \mathbb{R}^{m\times n}$ denote the ground truth matrix, $\hat{R}^{(k)}$ the matrix predicted by method k and let $\bar{\Omega}$ be the test set (a subset of entries). Let $\bar{R}^{(k)}$, $\hat{B}_U^{(k)}$ and $\hat{B}_I^{(k)}$ be respectively the zero-order term ($X^{(1)}M^{(1,1)}Y^{(1)}$ in BOMIC), the vector of user biases and the vector of items biases estimated by method k. In the SoftImpute case we need an extra post-processing step to estimate the biases: $\bar{R}^{(SI)} = \sum_{ij}\hat{R}_{ij}^{SI}/mn$, $\hat{B}_{U_i}^{(SI)} = \sum_j(\hat{R}_{ij}^{(SI)} - \bar{R}^{(SI)})/n$ and $\hat{B}_{I_j}^{(SI)} = \sum_i(\hat{R}_{ij}^{(SI)} - \bar{R}^{(SI)})/m$. To assess the methods we used the metrics presented in the list bellow:

- **[RMSE] Root-mean-square error**: $\mathrm{RMSE} = \sqrt{\sum_{i,j\in\bar{\Omega}}(\tilde{R}_{ij} - R_{ij})^2/|\bar{\Omega}|}$
- **[MBD] Matrix bias deviation**: $\mathrm{MBD} = \left|\bar{R} - \bar{R}^{(k)}\right|$
- **[UBD] (resp. IBD): User (resp. item) bias deviation**: $\mathrm{UBD} = \|B_U - \hat{B}_U^{(k)}\|_{\mathrm{Fr}}$ (resp. $\mathrm{IBD} = \|B_I - \hat{B}_I^{(k)}\|_{\mathrm{Fr}}$)
- **[SPC] Spearman correlation**: $\mathrm{SPC} = \rho_S\left(R_{\bar{\Omega}}, \hat{R}_{\bar{\Omega}}^{(k)}\right)$, the Spearman correlation between two vectors composed of the entries of R and $\hat{R}^{(k)}$ on the test set.

Since calculating the metrics MBD,UBD and IBD requires knowledge of all the entries of T, we only calculated them for the synthetic experiments. Note that lower values of RMSE, MBD and UBD and higher values of the Spearman correlation correspond to better performance.

C. Synthetic data simulations

For synthetic data simulations, we evaluated the ability of our model BOMIC to detect and adapt to different regimes in terms of the importance of user and item biases. We constructed two fixed matrices G and S, with the former made up purely of user/item biases, and the latter free of any implicit or explicit user or item biases. Then, we considered combinations $R(\alpha) = \alpha G + (1-\alpha)S$, observed either uniformly (which we describe with $\gamma = 1$) or in a biased way designed to fool a naive bias method into miscalculating the user and item biases ($\gamma = 4$). The proportion of observed entries p_Ω was also varied.

1) Results: For each combination of α, γ and p_Ω we generated 50 different samples of $R(\alpha)$. Given a sampled matrix, we recovered the unknown entries through **BOMIC**, **B-SI** and **SI**. Figure 2 summarizes the results of the performed simulations. We observe that our method consistently outperforms B-SI and SI, in terms of RMSE, UBD and IBD, and performs as well as SI in the MBD case. In addition, OMIC's ability to recover the correct user and item biases (UBD and IBD in Figure 2) is particularly marked in the case of non uniformly sampled entries, as might be expected, in line with Corollary III.1.

D. OMIC as a recommender system

In this subsection, we present our results on real data from the field of recommender systems. We begin by introducing the datasets, then provide our results, and conclude with a practical exploration of the added interpretability benefits of our method.

1) Datasets: In this paper, we worked with the following datasets:

- **Amazon** ($R \in \mathbb{R}^{164383\times 101364}$): Amazon is a multinational technology company which mainly focuses on e-commerce. We used the "Electronics" dataset, which we obtained through [26]. Users are Amazon's clients and items are electronic products (e.g. smartphones). The rating

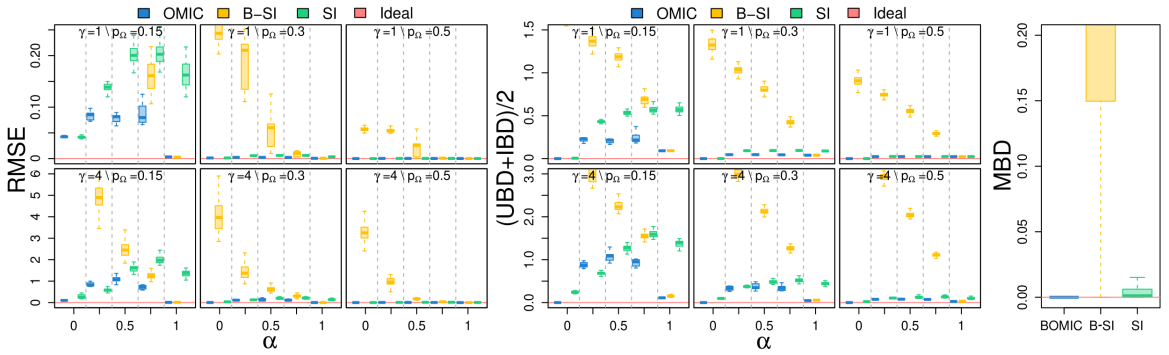

Figure 2: Summary of synthetic data simulations results. The first graph shows the relationship between all combinations of the parameters $(\alpha, \gamma, p_\Omega)$ and the RMSE. The second one shows how $(\alpha, \gamma, p_\Omega)$ influences the correct recovery of user and item biases ((UBD+IBD)/2). Each box plot in the first two graphs is obtained from 50 simulations. The third graph displays the distribution of the MBD over all of the simulations.

range is from 1 to 5 and the entry (i, j) refers to the rating given by client i to product j. Since no systematic side information was provided, we only investigated the performance of BOMIC and SoftImpute for this dataset.

- **Douban** ($R \in \mathbb{R}^{4988\times4903}$): Douban is a social network where users can produce content related to movies, music, and events. The ratings matrix was obtained through [27]. Douban users are members of the social network and Douban items are a subset of popular movies. The rating range is $\{1, 2, \ldots, 5\}$ and the entry (i, j) represents the rating given by user i to movie j. Feature vectors were collected by the authors and can be divided into two distinct parts: general features (e.g year of production, genres and movie duration) and the embedding of the description of the movie given by the pre-trained neural network Bert [28].
- **Goodreads spoiler dataset (GRS)** ($R \in \mathbb{R}^{4199\times3278}$): This dataset was released by [29] and it is available online. Goodreads is a social cataloging website that allows individuals to freely search its database of books, annotations, and reviews. In this case, an entry (i, j) represents the rating of the user i for the book j on a scale from 0 to 5. For each user-book pair, in addition to the rating score, the review text is also available. Each sentence of the review was annotated with respect to whether or not spoilers were present. We generated 89 features such as the length of the review and which percentage of the text contains spoilers.
- **MovieLens**: We consider the MovieLens 1M ($R \in \mathbb{R}^{6040\times3706}$) and MovieLens 20M ($R \in \mathbb{R}^{138493\times27032}$) datasets, which are broadly used and stable benchmark datasets. MovieLens is a non-commercial website for movie recommendations. In MovieLens 1M, we chose movies' genres (resp. age-gender combinations) as item (resp. user) communities.

Train-test setup: For each dataset, we split the set of observed entries (uniformly at random) into a training set (85 %), a validation set (10%) and a test set (5%).

2) Results: Table I summarizes the results of the real-world data experiments. We evaluated the performance of BOMIC, BOMIC+, SI and IMCNF on all datasets above. For BOMIC+, instead of using the side information directly, we performed unsupervised clustering of the corresponding features to translate them into community information, which we then used as the X, Y in the BOMIC+ algorithm in Section II-D. Observe that BOMIC+ has the lowest RMSE on all datasets and largest SPC in two datasets, whilst BOMIC has the best SPC on the MovieLens dataset. It is important to highlight that the standard BOMIC also beat the baselines. One interesting aspect of using BOMIC+ is that the unsupervised clustering step reduces the dimensionality of the side information, which can have a positive regularizing effect.

3) Illustration of OMIC's interpretability: As explained above, one advantage of our method is that it can provide partial explanations for its predictions: each prediction is a sum of terms coming from each of the components of the model. Furthermore, this sum is uniquely determined, since the components of the model live in mutually orthogonal spaces and correspond to distinct intuitive phenomena. For example, if some auxiliary vectors are constructed

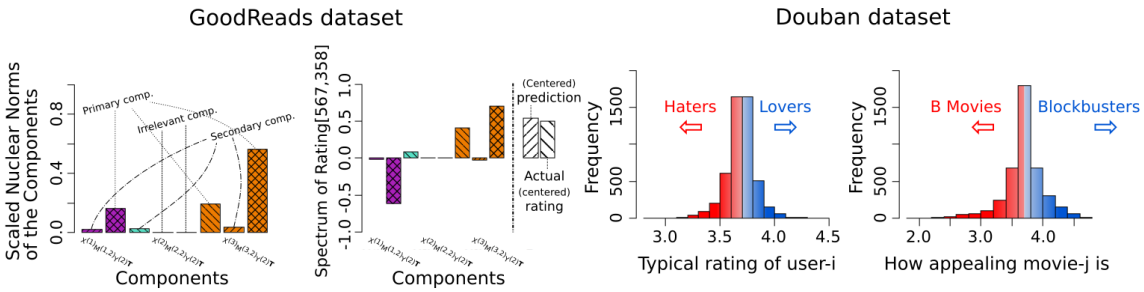

Figure 3: The first two graphs show the relative influence of the BOMIC+ components on the predictions of the whole matrix and one individual entry respectively. The last two graphs show the distribution of the users and item biases obtained by BOMIC on the Douban dataset.

from user community side information, the algorithm can disentangle the users' particular tastes from those of their respective community. In particular, OMIC can discover facts about community-wide behavior.

We illustrate those effects in Figures 3 and 4. On the left of Figure 3, we show the norms of each of the components of the recovered matrix: our recovered matrix takes the form $R = \sum_{k,l\leqslant 3} X^{(k)}\hat{M}^{(k,l)}(Y^{(l)})^\top$ where the $\hat{M}^{(k,l)}$ are obtained as the solution to our optimization problem (5), and each component in $X^{(k)}\hat{M}^{(k,l)}(Y^{(l)})^\top$ in the sum corresponds to an interpretable concept. For instance, $X^{(2)}\hat{M}^{(2,1)}(Y^{(1)})^\top$ correspond to user community biases. The norms of each component can give us an idea of how important each component is globally. Thus we see that over the whole GoodReads dataset, the most important components (excluding the global bias) are: (1) the specific match between the user and the book, (2) user generosity and (3) the quality of each book. These results are intuitively natural and expected.

The second picture presents an explanation for an individual prediction. In other words, we chose one entry of R (say, $R_{i,j}$) and represented the corresponding entry of each of the above mentioned components: for instance, the orange bar to the right of the graph represents the entry $(X^{(3)}\hat{M}^{(3,3)}(Y^{(3)})^\top)_{i,j}$, which corresponds to the same rating. This number represents the part of the rating (i,j) which is attributable to a specific preference of the user for the specific movie (discounting the parts of this preference which are shared by the other members of that user's community or other movies of the same genre).

Here, the book is not generally considered good by the users (cf. large negative component corresponding to the purple bar), but the individual is usually generous (first orange bar), and the specific book and user are a good match for each other (cf. large orange component corresponding to the rightmost bar).

Note that what is interesting here is that our model was specifically trained in a way that treats each of the components as a separate entity, with its own cross-validated hyperparmeter, so that the decomposition along those components is and intertwined with the optimization process (rather than collected as a statistic after applying a standard matrix completion method).

In the last two graphs, we show the distribution of user biases and movie quality in the Douban dataset. The distribution is similar to a normal distribution (squished at the boundaries), allowing us to characterize the users (resp. movies) on a spectrum between haters and lovers (resp. B-movies and blockbusters).

Figure 4 shows bar charts illustrating the affinities between user communities (gender-age combinations) and four movie genres in the MovieLens dataset. Note that these affinity scores are part of our model and can be directly read from the component $X^{(2)}M^{(2,2)}(Y^{(2)})^\top$ in BOMIC+. We observe that BOMIC+ is able to detect noteworthy human behaviour. For instance, female users tend to prefer drama and romance while male users enjoy comedies and thrillers instead. Note also that the biases vary with users' ages: older male users like romance movies more than their younger counterparts.

V. Conclusion

In this paper, we introduced OMIC, a matrix completion framework which relies on orthogonal auxiliary matrices to guide the model in privileged directions corresponding to prior knowledge. This simultaneously allows us to recover interpretable information about the predicted behavior. Our algorithm includes, as particular cases, three models (BOMIC, OMIC+ and BOMIC+) which can train user and item biases (and/or a community component) jointly with a

Table I: Performance comparison of our methods vs baselines on the real datasets

Dataset	P_Ω	RMSE				SPC			
		BOMIC	BOMIC+	SI	IMVNF	BOMIC	BOMIC+	SI	IMVNF
Amazon	0.0001	**1.0406**	-	1.0625	-	**0.4121**	-	0.4110	-
Douban	0.0195	0.7886	**0.7510**	0.8797	0.8034	0.6280	**0.6457**	0.5760	0.6017
GoodReads	0.0331	1.0736	**1.0540**	1.0991	1.0770	0.5113	**0.5120**	0.4857	0.5052
MovieLens1M	0.0446	0.8534	**0.8455**	0.8838	0.8559	**0.6368**	0.6336	0.6289	0.6321
MovieLens20M	0.0051	**0.7803**	-	0.8025	-	**0.6697**	-	0.6521	-

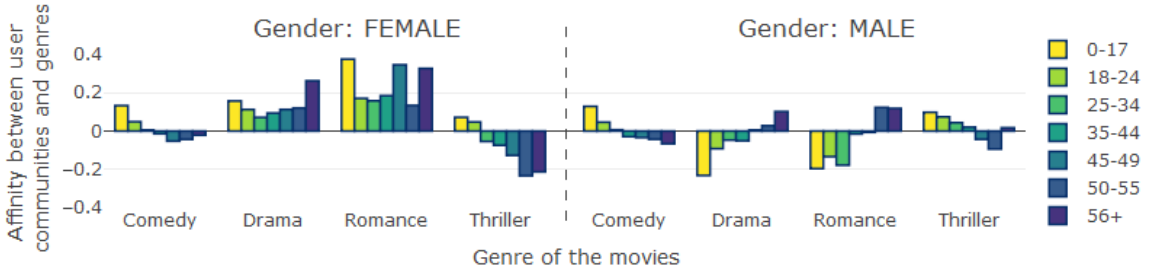

Figure 4: Affinity between user communities (grouped by gender and age) and movie genres for MovieLens. These biases can be directly read from the component $X^{(2)}M^{(2,2)}(Y^{(2)})^\top$ in BOMIC+.

nuclear norm minimization strategy. Finally, synthetic and real-data experiments demonstrate our algorithm's superior ability to adapt to and interpret different qualitative behaviors in the data.

Acknowledgements

We thank the reviewers for their helpful comments. We warmly thank Luís Augusto Weber Mercado for substantial help in designing the figures of the present paper. We also thank Rob Vandermeulen for helpful comments. The authors acknowledge support by the German Research Foundation (DFG) award KL 2698/2-1 and by the Federal Ministry of Science and Education (BMBF) awards 031L0023A, 01IS18051A, and 031B0770E. The simulations were partly executed on the high performance cluster "Elwetritsch II" at the TU Kaiserslautern (TUK) which is part of the "Alliance for High Performance Computing Rheinland-Pfalz"(AHRP). We kindly acknowledge the support of the RHRK especially when using their DGX-2

References

[1] Z. Kang, C. Peng, and Q. Cheng, "Top-n recommender system via matrix completion," 2016.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[3] C.-J. Hsieh, K.-Y. Chiang, and I. S. Dhillon, "Low rank modeling of signed networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12, New York, NY, USA, 2012, p. 507–515.

[4] F. Jirasek, R. A. S. Alves, J. Damay, R. A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft, and H. Hasse, "Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion," *The Journal of Physical Chemistry Letters*, vol. 11, no. 3, pp. 981–985, 2020.

[5] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, p. 2287–2322, Aug. 2010.

[6] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theor.*, vol. 56, no. 5, p. 2053–2080, May 2010.

[7] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, p. 717, 2009.

[8] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 952–960.

[9] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Statist.*, vol. 39, no. 5, pp. 2302–2329, 10 2011.

[10] C. C. Aggarwal, *Recommender Systems: The Textbook*, 1st ed. Springer Publishing Company, Incorporated, 2016.

[11] R. M. Bell, Y. Koren, and C. Volinsky, "The bellkor solution to the netflix prize," 2007.

[12] X. Zhang, S. Du, and Q. Gu, "Fast and sample efficient inductive matrix completion via multi-phase procrustes flow," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 5756–5765.

[13] M. Herbster, S. Pasteris, and L. Tse, "Online matrix completion with side information," *CoRR*, vol. abs/1906.07255, 2019.

[14] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," in *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2011, pp. 437–452.

[15] T. Chen, W. Zhang, Q. lu, K. Chen, Z. Zheng, and Y. Yu, "Svdfeature: A toolkit for feature-based collaborative filtering," *The Journal of Machine Learning Research*, 2012.

[16] P. Jain and I. S. Dhillon, "Provable inductive matrix completion," *CoRR*, vol. abs/1306.0626, 2013.

[17] K.-Y. Chiang, I. S. Dhillon, and C.-J. Hsieh, "Using side information to reliably learn low-rank matrices from missing and corrupted observations," *J. Mach. Learn. Res.*, 2018.

[18] J. Gaillard and J.-M. Renders, "Time-sensitive collaborative filtering through adaptive matrix completion," in *Advances in Information Retrieval*, A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, Eds. Cham: Springer International Publishing, 2015, pp. 327–332.

[19] A. Tasissa and R. Lai, "Low-rank matrix completion in a general non-orthogonal basis," 2018.

[20] J. Lu, G. Liang, J. Sun, and J. Bi, "A sparse interactive model for matrix completion with side information," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/093b60fd0557804c8ba0cbf1453da22f-Paper.pdf

[21] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/32bb90e8976aab5298d5da10fe66f21d-Paper.pdf

[22] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[23] L. Wang, X. Zhang, and Q. Gu, "A Unified Computational and Statistical Framework for Nonconvex Low-rank Matrix Estimation," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. Fort Lauderdale, FL, USA: PMLR, 20–22 Apr 2017, pp. 981–990. [Online]. Available: http://proceedings.mlr.press/v54/wang17b.html

[24] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, 1995.

[25] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank svd via fast alternating least squares," *Journal of Machine Learning Research*, vol. 16, no. 104, pp. 3367–3402, 2015. [Online]. Available: http://jmlr.org/papers/v16/hastie15a.html

[26] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *proceedings of the 25th international conference on world wide web*, 2016, pp. 507–517.

[27] N. Yang, "Douban movie and NetEase music datasets and model code," 2019. [Online]. Available: https://doi.org/10.7910/DVN/JGH1HA

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] M. Wan, R. Misra, N. Nakashole, and J. McAuley, "Fine-grained spoiler detection from large-scale review corpora," *arXiv preprint arXiv:1905.13416*, 2019.

[30] M. Fazel, "Matrix rank minimization with applications," *PhD Thesis*, 2002.

[31] "Learning with matrix factorizations."

[32] G. Watson, "Characterization of the subdifferential of some matrix norms," *Linear Algebra and its Applications*, vol. 170, pp. 33 – 45, 1992.

[33] Y. Nesterov, "Gradient methods for minimizing composite objective function," Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers 2007076, 2007.

[34] O. Shamir and S. Shalev-Shwartz, "Collaborative filtering with the trace norm: Learning, bounding, and transducing," in *Proceedings of the 24th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 19. PMLR, 2011, pp. 661–678.

[35] R. Latała, "Some estimates of norms of random matrices," *Proceedings of the American Mathematical Society*, vol. 133, no. 5, pp. 1273–1282, 2005.

[36] P. Giménez-Febrer, A. Pagès-Zamora, and G. B. Giannakis, "Generalization error bounds for kernel matrix completion and extrapolation," *IEEE Signal Processing Letters*, vol. 27, pp. 326–330, 2020.

[37] R. Foygel, O. Shamir, N. Srebro, and R. R. Salakhutdinov, "Learning with the weighted trace-norm under arbitrary sampling distributions," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 2133–2141.

[38] C. Scott, "Rademacher complexity," *Lecture Notes*, vol. Statistical Learning Theory, 2014.

[39] R. Meir and T. , "Generalization error bounds for bayesian mixture algorithms," *J. Mach. Learn. Res.*, vol. 4, no. null, p. 839–860, Dec. 2003.

[40] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," in *Computational Learning Theory*, D. Helmbold and B. Williamson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 224–240.

[41] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, vol. 12, no. null, p. 3413–3430, Dec. 2011.

[42] R. Vershynin, "High-dimensional probability," 2019. [Online]. Available: https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf

[43] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, p. 389–434, Aug. 2012.

[44] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, p. 150401, Oct 2010.

[45] H. Zhang, L. Cheng, and W. Zhu, "A lower bound guaranteeing exact matrix completion via singular value thresholding algorithm," *Applied and Computational Harmonic Analysis*, vol. 31, no. 3, pp. 454 – 459, 2011.

[46] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," *Proceedings of the Annual ACM Symposium on Theory of Computing*, 12 2012.

[47] C.-J. Hsieh and P. Olsen, "Nuclear norm minimization via active subspace selection," *31st Internat ional Conference on Machine Learning, ICML 2014*, vol. 1, pp. 871–882, 01 2014.

[48] Q. Yao and J. T. Kwok, "Accelerated and inexact soft-impute for large-scale matrix and tensor completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 9, pp. 1665–1679, Sep. 2019.

[49] Y. Yang, M. Pesavento, S. Chatzinotas, and B. Ottersten, "Parallel and hybrid soft-thresholding algorithms with line search for sparse nonlinear regression," 09 2018, pp. 1587–1591.

[50] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank svd via fast alternating least squares," *Journal of Machine Learning Research*, vol. 16, no. 104, pp. 3367–3402, 2015.

[51] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[52] C.-J. Hsieh and P. Olsen, "Nuclear norm minimization via active subspace selection," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 1. Bejing, China: PMLR, 22–24 Jun 2014, pp. 575–583.

[53] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, no. 1, p. 1665–1697, May 2012.

[54] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Completing any low-rank matrix, provably," *Journal of Machine Learning Research*, vol. 16, no. 94, pp. 2999–3034, 2015.

[55] Y. Wan, J. Yi, and L. Zhang, "Matrix completion from non-uniformly sampled entries," 06 2018.

[56] N. Srebro and R. R. Salakhutdinov, "Collaborative filtering in a non-uniform world: Learning with the weighted trace norm," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 2056–2064.

[57] O. Shamir and S. Shalev-Shwartz, "Matrix completion with the trace norm: Learning, bounding, and transducing," *Journal of Machine Learning Research*, vol. 15, pp. 3401–3423, 2014.

[58] R. Li, Y. Dong, Q. Kuang, Y. Wu, Y. Li, M. Zhu, and M. Li, "Inductive matrix completion for predicting adverse drug reactions (adrs) integrating drug–target interactions," *Chemometrics and Intelligent Laboratory Systems*, vol. 144, pp. 71 – 79, 2015.

[59] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene and disease associations," *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 06 2014.

[60] D. Shin, S. Cetintas, K.-C. Lee, and I. S. Dhillon, "Tumblr blog recommendation with boosted inductive matrix completion," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 203–212.

[61] 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*, ser. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 2301–2309.

[62] K. Zhong, P. Jain, and I. S. Dhillon, "Efficient matrix sensing using rank-1 gaussian measurements," in *Algorithmic Learning Theory*, K. Chaudhuri, C. GENTILE, and S. Zilles, Eds. Cham: Springer International Publishing, 2015, pp. 3–18.

[63] K. Zhong, P. Song, Zhao and, and I. S. Dhillon, "Provable non-linear inductive matrix completion," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 11 439–11 449.

[64] S. Xue, W. Qiu, F. Liu, and X. Jin, "Low-rank tensor completion by truncated nuclear norm regularization," *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2600–2605, 2017.

[65] N. Ghadermarzy, Y. Plan, and A. Yilmaz, "Near-optimal sample complexity for convex tensor completion," *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 577–619, 11 2018.

[66] M. Nimishakavi, P. K. Jawanpuria, and B. Mishra, "A dual framework for low-rank tensor completion," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 5484–5495.

[67] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 964–973. [Online]. Available: http://proceedings.mlr.press/v48/tu16.html

[68] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," 08 2008, pp. 426–434.

[69] ——, "Factor in the neighbors: Scalable and accurate collaborative filtering," *ACM Trans. Knowl. Discov. Data*, vol. 4, no. 1, Jan. 2010.

[70] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst, "Matrix Completion on Graphs," *arXiv e-prints*, p. arXiv:1408.1717, Aug. 2014.

[71] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King, "Recommender systems with social regularization," 01 2011, pp. 287–296.

[72] 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*, ser. RecSys '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 135–142.

[73] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: Social recommendation using probabilistic matrix factorization," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 931–940.

[74] W.-J. Li and D.-Y. Yeung, "Relation regularized matrix factorization." 01 2009, pp. 1126–1131.

[75] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, Aug 2011.

[76] G. Guo, J. Zhang, and N. Yorke-Smith, "Trustsvd: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings," 2015.

[77] N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2107–2115.

[78] K. Ahn, K. Lee, H. Cha, and C. Suh, "Binary rating estimation with graph side information," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 4272–4283.

[79] Qiaosheng, Zhang, V. Y. F. Tan, and C. Suh, "Community Detection and Matrix Completion with Two-Sided Graph Side-Information," *arXiv e-prints*, p. arXiv:1912.04099, Dec. 2019.

[80] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 126–135.

[81] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 2013, pp. 252–260.

[82] E. Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018.

[83] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, Jan 2016.

[84] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, Oct 2015, pp. 670–688.

[85] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109 – 137, 1983.

[86] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," 05 2018.

[87] H. Saad and A. Nosratinia, "Exact recovery in community detection with continuous-valued side information," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 332–336, Feb 2019.

[88] J. Yoon, K. Lee, and C. Suh, "On the joint recovery of community structure and community features," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2018, pp. 686–694.

[89] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *2013 IEEE 13th International Conference on Data Mining*, Dec 2013, pp. 1151–1156.

[90] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 3844–3852.

[91] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *CoRR*, vol. abs/1901.00596, 2019.

[92] M. Henaff, J. Bruna, and Y. Lecun, "Deep convolutional networks on graph-structured data," 06 2015.

[93] T. N. Kipf and M. Welling.

[94] A. Micheli, "Neural network for graphs: A contextual constructive approach," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 498–511, March 2009.

[95] G. Liu, Q. Liu, and X. Yuan, "A new theory for matrix completion," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 785–794.

[96] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[97] M. Liu, Y. Luo, Dacheng, C. Xu, and Y. Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," 2015.

[98] M. Z. Alaya and O. Klopp, "Collective matrix completion," *Journal of Machine Learning Research*, vol. 20, no. 148, pp. 1–43, 2019.

[99] S. Gunasekar, M. Yamada, D. Yin, and Y. Chang, "Consistent collective matrix completion under joint low rank structure," 12 2014.

[100] H. Chen and J. Li, "Learning multiple similarities of users and items in recommender systems," in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 811–816.

APPENDIX A
AN ALTERNATIVE FORMULATION OF THE OPTIMIZATION PROBLEM

Instead of a nuclear-norm minimization algorithm, our optimization problem (5) can be equivalently formulated as below.

Theorem A.1. *The optimization problem* (5) *is equivalent to the following optimization problem:*

$$\begin{aligned}\min \quad & \mathcal{L}\left(R_\Omega, \Lambda, \{U^{(k,l)}, V^{(k,l)}\}_{k\leqslant K, l\leqslant L}\right) \text{ with} \\ & \mathcal{L}(R_\Omega, M, \Lambda) \\ & = \left\| P_\Omega(R) - P_\Omega\left(\sum_{k=1,l=1}^{K,L} U^{(k,l)}(V^{(k,l)})^\top\right)\right\|_{\mathrm{Fr}}^2 \\ & \quad + \sum_{k,l}\lambda_{k,l}\left(\|U^{(k,l)}\|_{\mathrm{Fr}}^2 + \|V^{(k,l)}\|_{\mathrm{Fr}}^2\right),\end{aligned}$$

subject to $(\forall k, l)$*:*

$$\begin{aligned}&\mathrm{span}((U^{(k,l)})_{\bullet,i} : i \leqslant d_1^{(k)}) \subset \mathrm{span}((X^{(k)})_{\bullet,i} : i \leqslant d_1^{(k)}); \\ &\mathrm{span}((V^{(k,l)})_{\bullet,i} : i \leqslant d_2^{(l)}) \subset \mathrm{span}((Y^{(l)})_{\bullet,i} : i \leqslant d_2^{(l)}).\end{aligned}$$

For the proof, we will need the following lemma (lemma 6 from [5], see also [30], [31]):

Lemma A.1. *For any matrix Z, the following holds:*

$$\|Z\|_* = \min_{\substack{U,V;\\ UV^\top = Z}} \|U\|_{\mathrm{Fr}}\|V\|_{\mathrm{Fr}} \tag{S.21}$$

$$= \min_{\substack{U,V;\\ UV^\top = Z}} \frac{1}{2}\left(\|U\|_{\mathrm{Fr}}^2 + \|V\|_{\mathrm{Fr}}^2\right) \tag{S.22}$$

Proof. By Lemma A.1, we have that the optimization problem (5) is equivalent to the following:

$$\begin{aligned}\min \quad & \mathcal{L}\left(R_\Omega, \Lambda, \{U^{(k,l)}, V^{(k,l)}\}_{k\leqslant K, l\leqslant L}\right) \text{ with} \\ & \mathcal{L}(R_\Omega, M, \Lambda) \\ & = \left\| P_\Omega(R) - P_\Omega\left(\sum_{k=1,l=1}^{K,L} X^{(k)}M^{(k,l)}(Y^{(l)})^\top\right)\right\|_{\mathrm{Fr}}^2 \\ & \quad + \sum_{k,l}\lambda_{k,l}\left(\|U^{(k,l)}\|_{\mathrm{Fr}}^2 + \|V^{(k,l)}\|_{\mathrm{Fr}}^2\right), \\ & \text{subject to} \quad M^{(k,l)} = U^{(k,l)}(V^{(k,l)})^\top \quad \forall k, l.\end{aligned} \tag{S.23}$$

Now, note that if for any (k,l), $M^{(k,l)} = U^{(k,l)}(V^{(k,l)})^\top$ and $Z^{(k,l)} = X^{(k)}M^{(k,l)}(Y^{(l)})^\top$, then $Z = (X^{(k)}U^{(k,l)})(Y^{(l)}V^{(k,l)})^\top = \tilde{U}^{(k,l)}(\tilde{V}^{(k,l)})^\top$, where $\tilde{U}^{(k,l)} = (X^{(k)}U^{(k,l)})$ and $\tilde{V}^{(k,l)} = (Y^{(l)}V^{(k,l)})$. Furthermore, for any matrix $A \in \mathbb{R}^{n_1\times n_2}$ and any orthogonal matrix $B \in \mathbb{R}^{n_0\times n_1}$ (resp. $C \in \mathbb{R}^{n_2\times n_3}$),

$$\|A\|_{\mathrm{Fr}} = \|BA\|_{\mathrm{Fr}} = \|AC\|_{\mathrm{Fr}} = \|BAC\|_{\mathrm{Fr}}. \tag{S.24}$$

Thus we have $\|\tilde{U}^{(k,l)}\|_{\mathrm{Fr}} = \|\tilde{X}^\top \tilde{U}^{(k,l)}\|_{\mathrm{Fr}} = \|U^{(k,l)}\|_{\mathrm{Fr}}$, where $\tilde{X}$ is a matrix whose first d_1^k columns are identical to those of $X^{(k)}$, and whose columns form an orthonormal basis of $\mathbb{R}^m$. Similarly, $\|\tilde{V}^{(k,l)}\|_{\mathrm{Fr}} = \|V^{(k,l)}\|_{\mathrm{Fr}}$. Furthermore, conversely, if we can write a matrix Z as $\tilde{U}^{(k,l)}(\tilde{V}^{(k,l)})^\top$ for some $\tilde{U}^{(k,l)}$ and $\tilde{V}^{(k,l)}$ whose columns are in the span of the columns of $X^{(k)}$ and $Y^{(l)}$ respectively, then we can write $Z = (X^{(k)}U^{(k,l)})(Y^{(l)}V^{(k,l)})^\top = \tilde{U}^{(k,l)}(\tilde{V}^{(k,l)})^\top$ where $[U^{(k,l)}]_{i,j} = [(\tilde{X}^{(k)})^\top \tilde{U}^{(k,l)}]_{i,j} \quad \forall i,j \text{ s.t. } i \leqslant d_1^{(k)}$. The theorem follows. □

Appendix B
Proof of uniqueness of decomposition

Proposition B.1. *Let $\mathcal{F}_{k,l} = \{R : \exists M \in \mathbb{R}^{d_k^1 \times d_l^2} : R = X^{(k)} M (Y^{(l)})^\top\}$ denote the KL subspaces corresponding to each pair of auxiliary matrices $(X^{(k)}, Y^{(l)})$. Those vector spaces $\mathcal{F}_{k,l}$ are orthogonal (w.r.t. the Frobenius inner product) and their direct sum is the whole of $\mathbb{R}^{m\times n}$:*

$$\bigoplus \mathcal{F}_{k,l} = \mathbb{R}^{m\times n}. \tag{S.25}$$

In particular, for any $R \in \mathbb{R}^{m\times n}$, there exist a unique collection of matrices $R^{(k,l)} \in \mathcal{F}_{k,l}$ such that $R = \sum_{k,l} R^{(k,l)}$. In fact,

$$R^{(k,l)} = (X^{(k)})^\top R Y^{(l)}. \tag{S.26}$$

Proof. We divide the proof into two parts: the proof that the subspaces are mutually orthogonal, and the proof that their direct sum is the whole of the matrix space.

The subspaces $\mathcal{F}^{k,l}$ are mutually orthogonal. Let $A \in \mathcal{F}^{k,l}$ and $B \in \mathcal{F}^{k',l'}$ where either $k \neq k'$ or $l \neq l$. By definition of the subspaces in question, there exist $M \in \mathbb{R}^{d_k^1 \times d_l^2}$ and $N \in \mathbb{R}^{d_{k'}^1 \times d_{l'}^2}$ such that $A = X^{(k)} M (Y^{(l)})^\top$ and $B = X^{(k')} N (Y^{(l')})^\top$.

We now compute the (Frobenius) inner product between A and B in both cases.

Case 1: $l \neq l'$, in which case by the assumption on $Y^{(1)}, \dots Y^{(l)}$, we have $(Y^{(l)})^\top Y^{(l')} = \mathbf{0} \in \mathbb{R}^{d_l^2 \times d_{l'}^2}$. Then,

$$\begin{aligned}
\langle A, B\rangle &= \mathrm{Tr}\left(X^{(k)} M (Y^{(l)})^\top (X^{(k')} N (Y^{(l')})^\top)^\top\right) \\
&= \mathrm{Tr}\left(X^{(k)} M (Y^{(l)})^\top Y^{(l')} N^\top (X^{(k')})^\top\right) \\
&= \mathrm{Tr}\left(X^{(k)} M \mathbf{0} N^\top (X^{(k')})^\top\right) \\
&= 0
\end{aligned}$$

Case 2: $k \neq k'$

$$\begin{aligned}
\langle A, B\rangle &= \mathrm{Tr}\left((X^{(k')} M (Y^{(l)})^\top)^\top X^{(k)} N (Y^{(l)})^\top\right) \\
&= \mathrm{Tr}\left(Y^{(l)} M^\top (X^{(k')})^\top X^{(k)} M (Y^{(l)})^\top\right) \\
&= \mathrm{Tr}\left(Y^{(l)} M^\top \mathbf{0} M (Y^{(l)})^\top\right) = 0,
\end{aligned}$$

as expected.

The direct sum is the whole of $\mathbb{R}^{m\times n}$: $\bigoplus_{k,l} \mathcal{F}^{k,l} = \mathbb{R}^{m\times n}$.

Let $R \in \mathbb{R}^{m\times n}$. For each column vector $v \in \mathbb{R}^m$ we have immediately $v = \sum_k X^{(k)} (X^{(k)})^\top v$ by our assumption on the X's. Applying this to each column of R, we have $R = \sum_k X^{(k)} (X^{(k)})^\top R$. Similarly, $R = \sum_l R Y^{(l)} (Y^{(l)})^\top$. Plugging the second equation into the first one, we obtain

$$\begin{aligned}
R &= \sum_k X^{(k)} (X^{(k)})^\top R \\
R &= \sum_k X^{(k)} (X^{(k)})^\top \sum_l R (Y^{(l)} (Y^{(l)})^\top) \\
&= \sum_{k,l} X^{(k)} \left[(X^{(k)})^\top R Y^{(l)}\right] (Y^{(l)})^\top \\
&\in \bigoplus_{k,l} \mathcal{F}_{k,l},
\end{aligned}$$

as expected (this also proves equality (S.26)). □

Appendix C
Proof of convergence of our OMIC algorithm

In this section, we prove Theorems II.1 and II.2. The proofs rely mostly on adaptations of the techniques from [5], together with extensive use of the rotational invariance of the Frobenius and nuclear norms, as well as the linear independence of the spaces corresponding to each side information pairs.

Recall the optimization algorithm we propose to solve is the following one (cf. equations (5))

$$\begin{aligned}
&\min \quad \mathcal{L}(R_\Omega, M, \Lambda) \qquad \text{with} \qquad (\text{S.27}) \\
&\mathcal{L}(R_\Omega, M, \Lambda) = \sum_{k,l} \lambda_{k,l} \|M\|_* \\
&+ \frac{1}{2} \left\| P_\Omega(R) - P_\Omega\left(\sum_{k=1,l=1}^{K,L} X^{(k)} M^{(k,l)} (Y^{(l)})^\top\right)\right\|_{\mathrm{Fr}},
\end{aligned}$$

where P_Ω is the projection operator on the set of observed entries: i.e., if an entry is not observed, it is set to zero; if an entry is observed p times, any Frobenius norm counts that entry p times. Here, the output is $(M^{(k,l)})_{k\leqslant K, l\leqslant L}$ and $Z = \sum_{k=1,l=1}^{K,L} X^{(k)} M^{(k,l)} (Y^{(l)})^\top$.

First, let us finish the proof of the fully-known case:

Proof of Proposition II.1. Equation (10) follows from the fact that $M^{(k,l)}$ in the decomposition

is unique and determined by the formula $M^{(k,l)} = (X^{(k)})^\top Z Y^{(l)}$. This itself follows from the orthogonality of the side information matrices after multiplying each side of equation (13) by $(X^{(k)})^\top$ on the left and $Y^{(l)}$ on the right. The equivalence between the next two problems also follows.

As to the fact that $S_\Lambda(Z)$ is the solution to problem (12), let us first note that the case $K = L = 1$ with identity side information is just lemma 1 in [5].

Now, note that

$$\begin{aligned}
&\|\tilde{Z} - Z\|_{\mathrm{Fr}}^2 \\
&= \sum_{k,l} \|X^{(k)} M^{(k,l)} (Y^{(l)})^\top - X^{(k)} \tilde{M}^{(k,l)} (Y^{(l)})^\top\|_{\mathrm{Fr}}^2 \\
&= \sum_{k,l} \|M^{(k,l)} - \tilde{M}^{(k,l)}\|_{\mathrm{Fr}}^2, \qquad \text{(S.28)}
\end{aligned}$$

where at the first equality, we have used the orthogonality of the terms of the sum with respect to the Frobenius inner product, at the second equality, we have used the rotational invariance of the Frobenius norm. Here $\tilde{M}^{(k,l)} = (X^{(k)})^\top Z Y^{(l)}$, so that $Z = \sum_{k,l} X^{(k)} \tilde{M}^{(k,l)} (Y^{(l)})^\top$.

Using this, we can reformulate the problem (12) as follows:

$$\begin{aligned}
\min \quad & \sum_{k,l} \frac{1}{2} \|M^{k,l} - (X^{(k)})^\top Z Y^{(l)}\|_{\mathrm{Fr}}^2 \\
& + \sum_{k=1}^{K} \sum_{l=1}^{L} \lambda_{k,l} \left\| M^{(k,l)} \right\|_*, \qquad \text{(S.29)}
\end{aligned}$$

which can be solved as KL independent optimization problems, with the solution corresponding to index (k,l) being given by $M^{(k,l)} = S_{\lambda_{k,l}}((X^{(k)})^\top Z Y^{(l)})$, by an application of lemma 1 from [5]. The theorem follows.

□

Then, let us dispose with the following straightforward observation:

Lemma C.1. *The generalized singular value thresholding operator S_Λ satisfies, for any two matrices $Z_1, Z_2 \in \mathbb{R}^{m \times n}$,*

$$\|S_\Lambda(Z_1) - S_\Lambda(Z_2)\|_{\mathrm{Fr}} \leqslant \|Z_1 - Z_2\|_{\mathrm{Fr}}, \qquad \text{(S.30)}$$

and in particular, $S_\Lambda(\cdot)$ is a continuous map.

Proof. This follows from the corresponding lemma 3 in [5], together with the definition of the operator S_Λ:

$$\begin{aligned}
&\|S_\Lambda(Z_1) - S_\Lambda(Z_2)\|_{\mathrm{Fr}}^2 \\
&= \Bigg\| \sum_{k=1}^{K} \sum_{l=1}^{L} X^{(k)} S_{\lambda_{k,l}} \left((X^{(k)})^\top Z_1 Y^{(l)} \right) (Y^{(l)})^\top \\
&\quad - \sum_{k=1}^{K} \sum_{l=1}^{L} X^{(k)} S_{\lambda_{k,l}} \left((X^{(k)})^\top Z_2 Y^{(l)} \right) (Y^{(l)})^\top \Bigg\|_{\mathrm{Fr}}^2 \\
&= \left\| \sum_{k=1}^{K} \sum_{l=1}^{L} X^{(k)} S_{\lambda_{k,l}} \left((X^{(k)})^\top (Z_1 - Z_2) Y^{(l)} \right) (Y^{(l)})^\top \right\|_{\mathrm{Fr}}^2 \\
&= \sum_{k=1}^{K} \sum_{l=1}^{L} \left\| S_{\lambda_{k,l}} \left((X^{(k)})^\top (Z_1 - Z_2) Y^{(l)} \right) \right\|_{\mathrm{Fr}}^2 \\
&\leqslant \sum_{k=1}^{K} \sum_{l=1}^{L} \left\| (X^{(k)})^\top (Z_1 - Z_2) Y^{(l)} \right\|_{\mathrm{Fr}}^2 \\
&= \|Z_1 - Z_2\|_{\mathrm{Fr}}^2, \qquad \text{(S.31)}
\end{aligned}$$

where at the fourth line, we have used Lemma 3 from [5].

□

Now, let us define the quantity

$$\begin{aligned}
Q(A|B) = \frac{1}{2} \|P_\Omega(R) + P_{\Omega^\perp}(B) - A\|_{\mathrm{Fr}}^2 \\
+ \sum_{k,l} \lambda_{k,l} \|(X^{(k)})^\top A Y^{(l)}\|_*.
\end{aligned}$$

We have that the loss $\mathcal{L}(Z)$ corresponding to a matrix Z can be written $Q(Z|Z)$. Furthermore, let us define $Z^{i+1} = \arg\min_Z Q(Z|Z^i)$ (since this is an instance of the fully known case, the solution is unique and given by the operator S_Λ above). We now have the following lemma, which shows that the loss decreases monotonically with i:

Lemma C.2. *Define the sequence Z^i by $Z^{i+1} = \arg\min_Z Q(Z, Z^i)$ (with any starting point, for instance $Z^0 = 0$), which is equivalent to definition* (14)*. We have*

$$\mathcal{L}(Z^{i+1}) \leqslant Q(Z^{i+1}|Z^k) \leqslant \mathcal{L}(Z^i). \qquad \text{(S.32)}$$

Proof. The proof is almost the same as the proof of Lemma 2 in [5]. We have

$$\begin{aligned}
&\mathcal{L}(Z^i) \\
&= Q(Z^i|Z^i) \\
&= \frac{1}{2} \|R_\Omega + P_{\Omega^\perp}(Z^i) - Z^i\|_{\mathrm{Fr}}^2 \\
&\qquad + \sum_{k,l} \lambda_{k,l} \|(X^{(k)})^\top Z^i Y^{(l)}\|_* \\
&\geqslant \min_Z \frac{1}{2} \|R_\Omega + P_{\Omega^\perp}(Z^i) - Z\|_{\mathrm{Fr}}^2 \\
&\qquad + \sum_{k,l} \lambda_{k,l} \|(X^{(k)})^\top Z Y^{(l)}\|_*
\end{aligned}$$

$$
\begin{aligned}
&= Q(Z^{i+1}|Z^i) \\
&= \frac{1}{2}\|(R_\Omega - P_\Omega(Z^{i+1}) \\
&\qquad + (P_{\Omega^\perp}(Z^i) - P_{\Omega^\perp}(Z^{i+1}))\|_{\mathrm{Fr}}^2 \\
&+ \sum_{k,l}\lambda_{k,l}\|(X^{(k)})^\top Z^{i+1}Y^{(l)}\|_* \\
&= \frac{1}{2}\left\|(R_\Omega - P_\Omega(Z^{i+1})\right\|_{\mathrm{Fr}}^2 \\
&\qquad + \frac{1}{2}\left\|(P_{\Omega^\perp}(Z^i) - P_{\Omega^\perp}(Z^{i+1}))\right\|_{\mathrm{Fr}}^2 \\
&\qquad + \sum_{k,l}\lambda_{k,l}\|(X^{(k)})^\top Z^{i+1}Y^{(l)}\|_* \\
&\geqslant \frac{1}{2}\left\|(R_\Omega - P_\Omega(Z^{i+1})\right\|_{\mathrm{Fr}}^2 \\
&\qquad + \sum_{k,l}\lambda_{k,l}\|(X^{(k)})^\top Z^{i+1}Y^{(l)}\|_* \\
&= Q(Z^{i+1}, Z^{i+1}) = \mathcal{L}(Z^{i+1}). \qquad \text{(S.33)}
\end{aligned}
$$

□

Next, we have the following lemma:

Lemma C.3. *The sequence $\|Z^i - Z^{i-1}\|_{\mathrm{Fr}}$ is monotone decreasing:*

$$\|Z^i - Z^{i+1}\|_{\mathrm{Fr}} \leqslant \|Z^i - Z^{i-1}\|_{\mathrm{Fr}}. \qquad \text{(S.34)}$$

Furthermore,

$$Z^i - Z^{i+1} \to 0 \quad as \quad i \to \infty. \qquad \text{(S.35)}$$

Proof. We have

$$
\begin{aligned}
&\|Z^i - Z^{i+1}\|_{\mathrm{Fr}}^2 \\
&= \|S_\Lambda\left(P_{\Omega^\perp}(Z^{i-1}) + R_\Omega\right) - S_\Lambda\left(P_{\Omega^\perp}(Z^i) + R_\Omega\right)\|_{\mathrm{Fr}}^2 \\
&\leqslant \|\left(P_{\Omega^\perp}(Z^{i-1}) + R_\Omega\right) - \left(P_{\Omega^\perp}(Z^i) + R_\Omega\right)\|_{\mathrm{Fr}}^2 \\
&\qquad \text{By Lemma C.1} \\
&= \|P_{\Omega^\perp}(Z^{i-1}) - P_{\Omega^\perp}(Z^i)\|_{\mathrm{Fr}}^2 \qquad \text{(S.36)} \\
&\leqslant \|Z^i - Z^{i-1}\|_{\mathrm{Fr}}^2, \qquad \text{(S.37)}
\end{aligned}
$$

which proves the first statement (S.34).

As for the second statement (S.35), it will follow from the following two claims:

Claim 1: $P_\Omega(Z^i - Z^{i+1}) \to 0$. *Claim 2:* $P_{\Omega^\perp}(Z^i - Z^{i+1}) \to 0$.

Proof of Claim 1: Note that by inequality (S.34), the sequence $\|Z^i - Z^{i+1}\|_{\mathrm{Fr}}$ must converge. In particular, $\|Z^i - Z^{i+1}\|_{\mathrm{Fr}} - \|Z^i - Z^{i-1}\|_{\mathrm{Fr}} \to 0$, and by inequalites (S.36) and (S.37),

$$\|P_{\Omega^\perp}(Z^{i-1}) - P_{\Omega^\perp}(Z^i)\|_{\mathrm{Fr}} - \|Z^i - Z^{i-1}\|_{\mathrm{Fr}} \to 0,$$

from which we conclude that

$$\|P_\Omega(Z^i) - P_\Omega(Z^{i+1})\|_{\mathrm{Fr}}^2 \to 0.$$

Claim 1 follows.

Proof of Claim 2: We know by inequality (S.32) that $\mathcal{L}(Z^i)$ must converge, and thus $\mathcal{L}(Z^i) - \mathcal{L}(Z^{i+1}) \to 0$, from which it follows that

$$Q(Z^{i+1}|Z^i) - Q(Z^{i+1}|Z^{i+1}) \to 0. \qquad \text{(S.38)}$$

Now,

$$
\begin{aligned}
&Q(Z^{i+1}|Z^i) - Q(Z^{i+1}|Z^{i+1}) \\
&= \frac{1}{2}\|R_\Omega + P_{\Omega^\perp}(Z^i) - Z^{i+1}\|_{\mathrm{Fr}}^2 \\
&\qquad + \sum_{k,l}\lambda_{k,l}\|(X^{(k)})^\top Z^{i+1}Y^{(l)}\|_* \\
&- \frac{1}{2}\|R_\Omega + P_{\Omega^\perp}(Z^{i+1}) - Z^{i+1}\|_{\mathrm{Fr}}^2 \\
&\qquad - \sum_{k,l}\lambda_{k,l}\|(X^{(k)})^\top Z^{i+1}Y^{(l)}\|_* \\
&= \frac{1}{2}\|P_{\Omega^\perp}(Z^{i+1}) - P_{\Omega^\perp}(Z^i)\|_{\mathrm{Fr}}^2, \qquad \text{(S.39)}
\end{aligned}
$$

which, together with (S.38), implies claim 2. □

The next step is to prove that each limit point of the sequence Z^i is a solution to the optimization problem (5). To prove this, we will need the following lemma:

Lemma C.4. *Let $Z_{n_i} \to Z^\infty$ be a convergent subsequence of Z_i.*

Let $p_{n_i} \in \partial\sum_{k,l}\|(X^{(k)})^\top Z^i Y^{(l)}\|_*$ be a sequence of subgradients of our regularizer $\sum_{k,l}\|M^{(k,l)}\|_*$ evaluated at Z^i. There exists a convergent subsequence of p_{m_i} which converges to some

$$p \in \partial\sum_{k,l}\|(X^{(k)})^\top Z^\infty Y^{(l)}\|_*,$$

a subgradient of our regularizer, evaluated at the limit Z^∞.

Proof. First, recall from [32] and [5] that the set of subgradients of the nuclear norm of a matrix A is given by

$$\partial\|A\|_* = \left\{UV^\top + W, U^\top W = 0 = VW^\top, \|W\|_\sigma \leqslant 1\right\},$$

where $UDV^\top$ is the SVD of the matrix A. Using the chain rule and the fact that the side information matrices $X^{(k)}, Y^{(l)}$ are constant, we can calculate the set of subgradients of our regularizer evaluated at both Z^i and Z^∞ as follows:

$$
\begin{aligned}
&\partial\sum_{k,l}\|(X^{(k)})^\top Z^i Y^{(l)}\|_* \qquad \text{(S.40)} \\
&= \left\{\sum_{k,l}U_{k,l}^i(V_{k,l}^i)^\top + W_{k,l}^i, (U_{k,l}^i)^\top W_{k,l}^i = 0,\right.
\end{aligned}
$$

$$W^i_{k,l}V^i_{k,l} = 0, \|W^i_{k,l}\|_\sigma \leqslant 1\Bigg\}$$

and

$$\partial \sum_{k,l} \|(X^{(k)})^\top Z^i Y^{(l)}\|_* \tag{S.41}$$

$$= \Bigg\{ \sum_{k,l} U_{k,l}V_{k,l}^\top + W_{k,l}, U_{k,l}^\top W_{k,l} = 0,$$

$$W_{k,l}V_{k,l} = 0, \|W_{k,l}\|_\sigma \leqslant 1\Bigg\},$$

where $U_{k,l}D_{k,l}V_{k,l}^\top$ (resp. $U^i_{k,l}D^i_{k,l}(V^i_{k,l})^\top$) is the singular value decomposition of $(X^{(k)})^\top Z^\infty Y^{(l)}$(resp. $(X^{(k)})^\top Z^i Y^{(l)}$).

By compactness, there exists a subsequence m_i of n_i such that W^{m_i} converges to a value W. By continuity of the spectral norm, we also have $\|W\|_* \leqslant 1$. Furthermore, it follows from the convergence of Z^{n_i} (and in particular, of Z^{m_i}) to Z^∞ that $\sum_{k,l} U^{m_i}_{k,l}(V^{m_i}_{k,l})^\top \to \sum_{k,l} U_{k,l}(V_{k,l})^\top$. The result follows.

□

Proposition C.5. *Every limit point of the sequence $(Z^i)_{i\in\mathbb{N}}$ defined in* (14) *is a stationary point of the loss function $\mathcal{L}(Z) = \frac{1}{2}\|\tilde{Z} - Z\|^2_{\mathrm{Fr}} + \sum_{k=1}^K \sum_{l=1}^L \left\|(X^{(k)})^\top Z Y^{(l)}\right\|_*$ defined in* (11)*. Hence, it is also a solution to the fixed point equation*

$$Z = S_\Lambda\left(R_\Omega + P_{\Omega^\perp}(Z)\right). \tag{S.42}$$

Proof. Let Z^∞ be such a limit point. There exists a subsequence Z^{n_i} such that $Z^{n_i} \to Z^\infty$.

By Lemma C.3 ,we have $Z^{n_i} - Z^{n_i-1} \to 0$, which by continuity of the operator S_Λ implies that

$$R_\Omega + P_{\Omega^\perp}(Z^{n_i-1}) - Z^{n_i} \to R_\Omega - P_\Omega(Z^\infty). \tag{S.43}$$

Now, note that by definition of Z^i,

$$\begin{aligned}\forall i, \quad 0 &\in \partial Q(Z|Z^{i-1}) \\ &= -(P_\Omega(R) + P_\Omega(Z^{i-1}) - Z^i) \\ &\quad + \partial \sum_{k,l} \|(X^{(k)})^\top Z^i Y^{(l)}\|_*.\end{aligned}$$

Thus, we can choose, for all i, a $p_i \in \partial \sum_{k,l} \|(X^{(k)})^\top Z^i Y^{(l)}\|_*$ such that $p_i - (P_\Omega(R) + P_\Omega(Z^{i-1}) - Z^i) = 0$. Now, by Lemma C.4, there exists a subsequence Z^{m_i} of Z^{n_i} such that $p_{m_i} \to p$ for some

$$p \in \partial \sum_{k,l} \|(X^{(k)})^\top Z^\infty Y^{(l)}\|_*. \tag{S.44}$$

Putting equations (S.43) and (S.44) together, we obtain

$$\begin{aligned}0 = p_{m_i} &- (P_\Omega(R) + P_\Omega(Z^{m_i-1}) - Z^{m_i}) \\ &\to p - R_\Omega - P_\Omega(Z^\infty).\end{aligned} \tag{S.45}$$

Thus, 0 is a subgradient of $\mathcal{L}$ evaluated at Z^∞. The first statement of the Proposition follows.

As for the second statement, note that

$$Z^{m_i} = S_\Lambda\left(R_\Omega + P_{\Omega^\perp}(Z^{m_i-1})\right). \tag{S.46}$$

Furthermore, by Lemma C.3, $Z^{m_i} - Z^{m_i-1} \to 0$, and therefore $Z^{m_i-1} \to Z^\infty$. Thus, using the continuity of the generalized singular value thresholding operator, we obtain by passing to the limits in (S.46):

$$Z^\infty = S_\Lambda\left(R_\Omega + P_{\Omega^\perp}(Z^\infty)\right), \tag{S.47}$$

as expected. □

Proof of Theorem II.1. In Proposition C.5, we have already proved that any limit point of the sequence $(Z^i)_{i\in\mathbb{N}}$ (defined in equation (14)) is a stationary point of the loss function, and therefore a solution to the optimization problem (5). Thus, the only thing left to prove is that the sequence $(Z^i)_{i\in\mathbb{N}}$ converges: indeed, if that is the case, its limit will be its (only) limit point, and will be a solution to problem (5).

Let us first dispense with the following simple observation: by Lemma C.2, for any i, we have

$$\mathcal{L}(Z^i) \leqslant \mathcal{L}(Z^0). \tag{S.48}$$

Since the objective function $\mathcal{L}$ is a continuous function of the matrix Z, the set of matrices Z satisfying equation (S.48) is compact. Thus, by compactness, there exists at least one limit point $\bar{Z}$.

Now, by the continuity of S_Λ and the definition of Z^i, we have, for any i:

$$\begin{aligned}&\|\bar{Z} - Z^i\|^2_{\mathrm{Fr}} \\ &= \left\|S_\Lambda\left(R_\Omega + P_{\Omega^\perp}(\bar{Z})\right) - S_\Lambda\left(R_\Omega + P_{\Omega^\perp}(Z^{i-1})\right)\right\|^2_{\mathrm{Fr}} \\ &\leqslant \left\|\left(R_\Omega + P_{\Omega^\perp}(\bar{Z})\right) - \left(R_\Omega + P_{\Omega^\perp}(Z^{i-1})\right)\right\|^2_{\mathrm{Fr}} \\ &= \|P_{\Omega^\perp}(\bar{Z} - Z^{i-1})\|^2_{\mathrm{Fr}} \leqslant \|\bar{Z} - Z^{i-1}\|^2_{\mathrm{Fr}},\end{aligned} \tag{S.49}$$

where at the first line, we have used Proposition C.5 and the definition of Z^i.

We will now show that the sequence $(Z^i)_{i\in\mathbb{N}}$ actually converges to $\bar{Z}$. To do this, we proceed by contradiction. Assume Z^i doesn't converge towards $\bar{Z}$. By definition of convergence, this implies that there must exist an $\epsilon_* > 0$ such that there exists an infinite subsequence $Z^{I_1}, Z^{I_2}, \ldots$ such that for all i,

$\|Z^{I_i}-\bar{Z}\|_{\mathrm{Fr}} \geqslant \epsilon_*$. Since the subsequence $Z^{I_1}, Z^{I_2}, \ldots$ is contained in the compact set

$$\mathcal{C}_{\epsilon_*} := \left\{Z : \mathcal{L}(Z) \leqslant \mathcal{L}(Z^0) \ \wedge \ \|Z^{I_i} - \bar{Z}\|_{\mathrm{Fr}} \geqslant \epsilon_*\right\},$$

it must have a limit point $\widetilde{Z} \in \mathcal{C}_{\epsilon_*}$ inside that set. In particular, we have

$$\|\widetilde{Z} - \bar{Z}\|_{\mathrm{Fr}} \geqslant \epsilon_* \tag{S.50}$$

Set $\epsilon = \frac{\epsilon_*}{3}$. Since $\bar{Z}$ is a limit point of $(Z^i)_{i\in\mathbb{N}}$, there certainly exists an index k such that

$$\|Z^k - \bar{Z}\|_{\mathrm{Fr}} \leqslant \epsilon. \tag{S.51}$$

Since $\widetilde{Z}$ is also a limit point of $(Z^i)_{i\in\mathbb{N}}$ (specifically, a limit point of the subsequence $(Z^{I_i})_{i\in\mathbb{N}}$), there exists an index l such that $l > k$ and

$$\|Z^l - \widetilde{Z}\|_{\mathrm{Fr}} \leqslant \epsilon. \tag{S.52}$$

On the other hand, since $l > k$ by iteratively applying equation (S.49) $l-k$ times, we obtain:

$$\|\bar{Z} - Z^l\|_{\mathrm{Fr}} \leqslant \|\bar{Z} - Z^k\|_{\mathrm{Fr}} \leqslant \epsilon, \tag{S.53}$$

where the last inequality follows from equation (S.51).

Now, by equations (S.53) and (S.52) and the triangle inequality, we obtain:

$$\|\widetilde{Z} - \bar{Z}\|_{\mathrm{Fr}} \leqslant \|\widetilde{Z} - Z^l\|_{\mathrm{Fr}} + \|Z^l - \bar{Z}\|_{\mathrm{Fr}} \tag{S.54}$$

$$\leqslant \epsilon + \epsilon = 2\epsilon = \frac{2\epsilon_*}{3} < \epsilon_*, \tag{S.55}$$

which is in contradiction with equation (S.50). Thus, we deduce by contradiction that Z^i indeed converges to its only limit point $\bar{Z}$ (which we refer to as Z^∞ in the rest of the appendix). As explained at the beginning of the proof, this, together with Proposition C.5, implies Theorem II.1, as required.

□

We can now proceed with the proof of our Theorem II.2 on the worst-case convergence.

Proof of Theorem II.2. The proof is exactly the same as that of theorem 2 in [5] (and also takes inspiration from [33]), and we reformulate it into our notation here for the sake of completeness only.

For $\theta \in [0,1]$, we write $Z^i(\theta)$ for $(1-\theta)Z^i + \theta Z^\infty$. Note that by convexity of our loss function $\mathcal{L}$, we have $\mathcal{L}(Z^i(\theta)) \leqslant (1-\theta)\mathcal{L}(Z^i) + \theta\mathcal{L}(Z^\infty)$.

Note also that we have

$$\|P_{\Omega^\perp}(Z^i - Z^i(\theta))\|_{\mathrm{Fr}}^2 = \theta^2\|P_{\Omega^\perp}(Z^i - Z^\infty)\|_{\mathrm{Fr}}^2$$
$$\leqslant \theta^2\|Z^i - Z^\infty\|_{\mathrm{Fr}}^2 \leqslant \theta^2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2, \tag{S.56}$$

where we have used Lemmas C.3 and C.1.

Using these facts and the definition in the construction of the sequence Z^i, we can derive the following key inequalities:

$$\begin{aligned}
&\mathcal{L}(Z^{i+1}) \\
&= \min_Z \left[\mathcal{L}(Z) + \frac{1}{2}\|Z - Z^i\|_{\mathrm{Fr}}^2\right] \\
&\leqslant \min_{\theta\in[0,1]} \left[\mathcal{L}(Z^i(\theta)) + \frac{1}{2}\|Z^i(\theta) - Z^i\|_{\mathrm{Fr}}^2\right] \\
&\leqslant \min_{\theta\in[0,1]} \Bigg[\mathcal{L}(Z^i) + \theta(\mathcal{L}(Z^\infty) - \mathcal{L}(Z^i)) \\
&\qquad\qquad + \frac{1}{2}\theta^2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2\Bigg].
\end{aligned} \tag{S.57}$$

The last expression is minimised for $\theta = \theta^i$ where

$$\theta^i = \min\left(\frac{\mathcal{L}(Z^i) - \mathcal{L}(Z^\infty)}{\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2}, 1\right). \tag{S.58}$$

(If $\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2 = 0$, then $Z^i = Z^\infty \quad \forall i$ and there is nothing to prove.)

Recall also that θ^i is a decreasing sequence (cf. Lemma C.2): if $\theta^i \leqslant 1$, then $\theta^j \leqslant 1 \quad \forall j > i$. Suppose $\theta^0 = 1$. Then, plugging this back into equation (S.57), we obtain:

$$\mathcal{L}(Z^1) - \mathcal{L}(Z^\infty) \leqslant \frac{1}{2}\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2, \tag{S.59}$$

and therefore $\theta^1 \leqslant \frac{1}{2}$. Thus, in all cases, $\theta^i < 1 \quad \forall i \geqslant 1$. Note also that if $\theta^0 = 1$, inequality (15) is satisfied (this follows from inequality (S.59)).

Now, for $i \geqslant 1$, we can just use the explicit expression (S.58) for θ, which, plugged back into equation (S.57), gives:

$$\mathcal{L}(Z^{i+1}) - \mathcal{L}(Z^i) \leqslant -\frac{(\mathcal{L}(Z^i) - \mathcal{L}(Z^\infty))^2}{2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2}. \tag{S.60}$$

Now, writing α_i for $(\mathcal{L}(Z^i) - \mathcal{L}(Z^\infty))$ (which is a decreasing sequence, as shown by Lemma C.3) and using the above expression, we obtain

$$\begin{aligned}
\alpha_i &\geqslant \frac{\alpha_i^2}{2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2} + \alpha_{i+1} \\
&\geqslant \frac{\alpha_i\alpha_{i+1}}{2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2} + \alpha_{i+1},
\end{aligned} \tag{S.61}$$

which yields:

$$\alpha_{i+1}^{-1} \geqslant \frac{1}{2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2} + \alpha_i^{-1}. \tag{S.62}$$

Summing both sides for the index running from 1 to $i-1$, we obtain:

$$\alpha_i^{-1} \geqslant \frac{i-1}{2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2} + \alpha_1^{-1}. \tag{S.63}$$

Since $\theta_1 < 1$, by definition of θ_1, we obtain $\frac{\alpha_1}{2\|Z^0-Z^\infty\|_{\mathrm{Fr}}^2} \leqslant \frac{1}{2}$. Plugging this back into equation (S.63), we obtain:

$$\begin{aligned}\alpha_i^{-1} &\geqslant \frac{i-1}{2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2} + \alpha_1^{-1} \\ &\geqslant \frac{i-1}{2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2} + \frac{1}{\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2} \\ &= \frac{i+1}{2\|Z^0 - Z^\infty\|_{\mathrm{Fr}}^2},\end{aligned}$$

which yields inequality (15) after inverting both sides. □

Remark: We use the notation Z^∞ to refer to the limit of the sequence of iterates, instead of referring to 'the solution Z^* of the optimization problem (5)' because the solution is not necessarily unique and Z^∞ may actually depend on the initialization. Indeed, the optimization problem (5) is *convex* but *not strongly convex*. Thus, there can be several solutions, but each of them corresponds to the same value of the objective function. To check that the specific problem (5) can indeed have several equivalent solutions, consider the following example. $X^{(1)} = Y^{(1)} = \mathrm{Id}$ (so that our algorithm coincides with Softimpute), $m = n = 2$, and let the observed entries be $R_{2,1} = R_{1,2} = 1$ (thus, $\Omega = \{(2,1);(1,2)\}$. Here are three equivalent solutions to the optimization problem with regularising parameter λ:

$$A_- = \begin{pmatrix} \lambda-1 & 1-\lambda \\ 1-\lambda & \lambda-1 \end{pmatrix};$$
$$A_+ = \begin{pmatrix} 1-\lambda & 1-\lambda \\ 1-\lambda & 1-\lambda \end{pmatrix}; \qquad \text{and}$$
$$A_0 = \begin{pmatrix} 0 & 1-\lambda \\ 1-\lambda & 0 \end{pmatrix}.$$

In all three cases, the nuclear norm is $2-2\lambda$, and the value of the objective function is $\lambda(2-2\lambda) + \frac{1}{2}(\lambda^2+\lambda^2)$. It is also easy to check that those are actually solutions of (5) because performing any extra iteration of the algorithm yields the same matrix: for A_0, after the imputation step we get the target

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = 1 \times \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} + 1 \times \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix},$$

where the second line is the SVD. It is clear from the SVD that applying the singular value thresholding operator will return the matrix A_0. Thus, the algorithm converges exactly to A_0 in one (zero) iteration(s).

For A_+, note that after the imputation step we get the target

$$\begin{aligned}&\begin{pmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{pmatrix} \\ &= (2-\lambda) \times \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &+ \lambda \times \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix},\end{aligned}$$

and it is clear that after applying the SVT operator we obtain

$$\begin{aligned}&(2-2\lambda) \times \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1-\lambda & 1-\lambda \\ 1-\lambda & 1-\lambda \end{pmatrix} = A_+,\end{aligned}$$

as expected. A similar calculation shows that A_- is also a solution.

APPENDIX D
COMPLEXITY AND RUNTIME ANALYSIS

Runtime analysis of SVT operations in Alg. 1: Whilst Algorithms 1 and 3 theoretically require KL SVD operations, in many instances of our model class (including BOMIC, OMIC+ and BOMIC+) most of the svd calculations are actually *trivial*. Indeed, consider the target matrix $T = P_{\Omega^\top}(Z^{\mathrm{old}}) + R_\Omega$. Note that $M^{(k,l)} := (X^{(k)})^\top T Y^{(l)}$ is a $d_1^{(k)} \times d_2^{(l)}$ matrix. For instance, if $d_1^{(k)} = 1$ or $d_2^{(l)} = 1$ (cf. BOMIC+, $k = 1$ or $l = 1$), this matrix is a vector, making the computation of an SVD unnecessary. In fact, for any combination (k,l) such that $d_1^{(k)} + d_2^{(l)}$ is small, it is easy to compute the small matrix $(X^{(k)})^\top T Y^{(l)}$ and perform its SVD through standard methods.

Figure 5 is a graph which compares the runtimes of SVD calculations in Softimpute and our algorithm 1. For each datapoint, we randomly selected a matrix with the following parameters rank $\in \{5,6,\cdots,10\}; m \in \{100,101,\cdots,1000\}; d_1^{(1)} = d_2^{(1)} \in \{2,3,\cdots,\lceil 0.1m \rceil\}$. More specifically, the users and items were each divided into $d_1^{(1)} = d_2^{(1)}$ communities and the matrices $X^{(1)}, X^{(2)}, X^{(3)}, Y^{(1)}, Y^{(2)}, Y^{(3)}$ were constructed according to the standard procedure for BOMIC+ (see. Subsection II-D). The matrices $M^{(k,l)}$ were chosen with iid Gaussian entries.

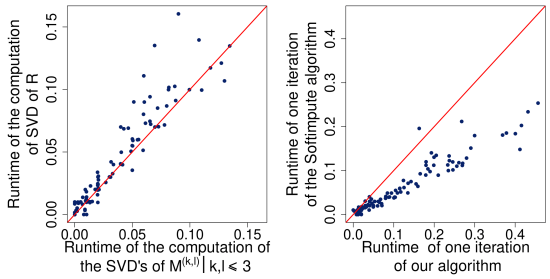

Figure 5: Runtime comparison of main computations required for one iteration of Softimpute and OMIC. The red line is the identity.

We then compared, on the one hand (left part of the figure) the following two operations:

1 Performing the SVD of the full matrix $R = \sum_{k,l} X^{(k)} M^{(k,l)} (Y^{(k)})^\top$,
2 Performing the SVDs of *all nine matrices* $M^{(k,l)}$ for $k, l \leqslant 3$;

and on the other hand (right part of the figure), the following two operations:

1 One full iteration of the Softimpute algorithm, including imputation and singular value thresholding operator.
2 One full iteration of our algorithm, including the imputation and the application of the generalized singular value thresholding operator.

As we can see from the figure, the computational burden of all 9 SVDs required in our algorithm is not significantly bigger than that of the single SVD required in the SoftImpute implementation. Furthermore, although one full iteration of our algorithm is slower than one full iteration of the softimpute algorithm due to extra multiplication steps, this appears to be the case only by a small constant factor.

Formal complexity analysis: We provide an efficient implementation for the special cases BOMIC, OMIC+, and BOMIC+. In those three cases, the number of iterations required at each step of both Algorithm 3 as well as the SVD calculation (Algorithm 2) depend on the many practicalities related to various warm starts applied in both cases. However, it is possible to write down the complexity of performing one iteration.

At each iteration of Algorithm 3, the key step is the SVT operation using Algorithm 2 (the only other operation being an assignment of $O(|\Omega|)$ entries). For each iteration inside Algorithm 2, the complexity can be computed from the following operations which are each required a fixed number of times (here as usual, r is the fixed maximum rank set as a hyperparameter):

- Multiplying each column of a matrix in $\mathbb{R}^{m\times r}$ or $\mathbb{R}^{n\times r}$ by a different constant (e.g. line 12,17). Cost: $O((m+n)r)$.
- Computing the SVD of a $\mathbb{R}^{m\times r}$ or $\mathbb{R}^{n\times r}$ matrix (e.g., lines 16,26). Cost: $O(r^3 + (m+n)r^2)$.
- Performing projections onto the spaces corresponding to $X^{(k)}$ or $Y^{(l)}$ via the procedure from line 1. Cost: $O(m+n)$.
- Multiplying r vectors by the current target (see lines 14,19, 24). Cost: $O(|\Omega|r + (m+n)r^2)$.

Since $r \leqslant m+n$, this yields an overall complexity of $O(KL[|\Omega|r + (m+n)r^2])$. Note that $KL \leqslant 9$, so that the complexity is $O(9[|\Omega|r + (m+n)r^2]) = O(|\Omega|r + (m+n)r^2)$, the same as the SoftImpute algorithm [5].

APPENDIX E
PROOFS OF GENERALIZATION BOUNDS

Notation: In this section, we assume the entries are sampled with i.i.d. noise, so that observations of entry $R_{i,j}$ are of the form $R_{i,j} + \delta_{i,j}$ for $\delta_{i,j} \sim \Delta_{i,j}$ for some noise distribution $\Delta_{i,j}$[8]. Thus, N i.i.d. observations indexed by $\alpha \in \{1, 2, \ldots, N\}$ are denoted by $R_{i_\alpha, j_\alpha} + \delta_\alpha$ where (i_α, j_α) is the α'th i.i.d. choice of entry, and each δ_α is drawn from $\Delta_{i_\alpha, j_\alpha}$ independently. The loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is bounded by a constant B, with Lipschitz constant bounded by L_ℓ[9]. For all $k \leqslant K, l \leqslant L, i \leqslant m, j \leqslant n$, we will write $\mathbf{x}_i^k$ (resp. $\mathbf{y}_j^k$) for the ith row (resp. jth column) of the matrix $X^{(k)}$ (resp. $Y^{(l)}$), $\mathcal{X}^{(k)}$ for $\max_{i=1}^m \|\mathbf{x}_i^k\|_2 = \max_{i=1}^m \|X_{i,\bullet}^{(k)}\|_2$ and $\mathcal{Y}^{(l)}$ for $\max_{i=1}^n \|\mathbf{y}_i^l\|_2$. For a predictor $f : \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\} \to \mathbb{R}$, we will write $\mathcal{R}(f)$ for the expected risk $\mathbb{E}_{(i,j)\sim\mathcal{D}}(\ell(f(i,j), R_{i,j} + \delta_{i,j})$ and $\hat{\mathcal{R}}(f)$ for the empirical risk $(1/N)\sum_{\alpha=1}^N \ell(f(i,j), R_{i_\alpha, j_\alpha} + \delta_\alpha)$. Table (II) summarizes all the notations used in this appendix and in the paper.

A. Proof of bounds in the distribution-free case

First, let us recall the following lemma from [17], [34].

Lemma E.1. *Let ℓ be a loss function bounded by B and with Lipschitz constant bounded by L. Suppose we are given a matrix $R \in \mathbb{R}^{m\times n}$, which is observed*

[8]Furthermore, $R_{i,j}$ and $\Delta_{i,j}$ are defined so that $R_{i,j} = \arg\min_y \mathbb{E}_{\delta_{i,j}}(\ell(R_{i,j} + \delta_{i,j}, y))$.

[9]These conditions are satisfied, for instance, for a hinge loss which could be used to estimate the probability of predicting within a given accuracy, or for the squared loss if one additionally assumes a fixed upper bound on all entries and predictions (which is a reasonable assumption in practice).

with i.i.d. noise, so that observing entry (i,j) results in an output of $R_{i,j}+\delta$ where $\delta \sim \Delta_{i,j}$ where the $\Delta_{i,j}$ are distributions. Let $F_{\mathcal{M}}$ be the set of matrices $\tilde{R}$ with $\|\tilde{R}\|_* \leqslant \mathcal{M}$. Let us write the data-dependent Rademacher complexity for N samples indexed by $\alpha \in \{1,2,\dots,N\}$ by $\mathfrak{R}_N(F_{\mathcal{M}}) = \mathbb{E}\left(\sup_{\tilde{R}\in F_{\mathcal{M}}} \frac{1}{N}\sum_{\alpha=1}^{N} \sigma_\alpha \ell(R_{(i_\alpha,j_\alpha)}+\delta_\alpha,\tilde{R})\right)$, where the σ_α are independent Rademacher random variables, and the (i_α,j_α) are entries sampled independently.

We have the following bound on the expected complexity $\mathfrak{R} = \mathbb{E}_\Omega(\mathfrak{R}_N(F_{\mathcal{M}}))$:

$$\mathfrak{R} = \mathbb{E}_\Omega(\mathfrak{R}_N(F_{\mathcal{M}})) \leqslant \sqrt{\frac{9\mathcal{M}B\mathcal{C}L(\sqrt{d_1}+\sqrt{d_2})}{N}}.$$

Here, $\mathcal{C}$ is the universal constant from [35].

Using this, we can show the following lemma for side information by absorbing the variation between entries corresponding to the same communities into the "noise" of an auxiliary problem to which we apply Lemma E.1.

Lemma E.2. *Let $X \in \mathbb{R}^{m\times A}$ and $Y \in \mathbb{R}^{n\times B}$ be auxiliary matrices whose columns are indicator functions of distinct sets forming partitions $\{c_1,c_2,\dots,c_A\}$ and $\{s_1,\dots,s_B\}$ of $\{1,2,\dots,n\}$ and $\{1,2,\dots,m\}$ respectively. Set $t>0$ and consider the function class $\mathcal{F}_t := \left\{XMY^\top \middle| \|M\|_* \leqslant t\right\}$. The Rademacher complexity $\mathcal{F}_N(\mathcal{F}_t)$ satisfies*

$$\mathcal{R}_N(\mathcal{F}_t) \leqslant \sqrt{\frac{9tC\mathcal{C}L\left[\sqrt{a}+\sqrt{b}\right]}{N}}.$$

In particular, if we consider instead $\mathcal{G}_t := \left\{XMY^\top \middle| \operatorname{rank}(M) \leqslant r\right\}$, we obtain:

$$\mathcal{R}_N(\mathcal{G}_t) \leqslant \sqrt{\frac{9\sqrt{Cabr}B\mathcal{C}L\left[\sqrt{a}+\sqrt{b}\right]}{N}},$$

where C is a bound on the predicted entries.

Proof. This follows from Lemma E.1 applied to the modified problem where for $u \leqslant A, v \leqslant B$, the observations $R_{u,v}$ are distributed according to the distribution of $R_{i,j}+\delta_{i,j}$ where (i,j) is drawn from $\mathcal{D}$ conditioned on $i \in c_u, j \in c_v$ (note that Lemma E.1 allows for the observations of $R_{i,j}$ to be perturbed by random variables with distributions $\Delta_{i,j}$ conditioned on (i,j), the differences between the values of the ground truth matrix at different pairs (i,j) where the communities of i and j are fixed can be absorbed into this perturbation). □

Proposition E.3. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix and let $v \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$ be two vectors. We have*

$$\|vw^\top \odot A\|_* \leqslant \max_{i,j}|v_i||w_j|\|A\|_* \tag{S.64}$$

where $\odot$ denotes the Hadamard (entry wise) product.

Proof. By an equivalent formulation of the nuclear norm A.1 (see. also [31])

$$\begin{aligned}
&\|vw^\top \odot A\|_* \\
&= \min\left(\|B\|_{\mathrm{Fr}}\|C\|_{\mathrm{Fr}} \middle| \right. \\
&\qquad \left. BC^\top = vw^\top \odot A\right) \\
&\leqslant \min\left(\|\operatorname{diag}(v)B\|_{\mathrm{Fr}}\|\operatorname{diag}(w)C\|_{\mathrm{Fr}} \middle| \right. \\
&\qquad \left. \operatorname{diag}(v)B(\operatorname{diag}(w)C)^\top = vw^\top \odot A\right) \\
&= \min\left(\|\operatorname{diag}(v)B\|_{\mathrm{Fr}}\|\operatorname{diag}(w)C\|_{\mathrm{Fr}} \middle| \right. \\
&\qquad \left. (vw^\top)\odot(BC^\top) = vw^\top \odot A\right) \\
&\leqslant \min\left(\|\operatorname{diag}(v)B\|_{\mathrm{Fr}}\|\operatorname{diag}(w)C\|_{\mathrm{Fr}} \middle| \right. \\
&\qquad \left. BC^\top = A\right) \\
&\leqslant \min\left(\max_i|v_i|\|B\|_{\mathrm{Fr}}\max_j|w_j|\|\operatorname{diag}(w)C\|_{\mathrm{Fr}} \middle| \right. \\
&\qquad \left. BC^\top = A\right) \\
&= \max_{i,j}|v_i||w_j|\|A\|_*,
\end{aligned}$$

as expected. □

We are now in a position to present the main result.

Theorem E.1. *Consider the OMIC+ setting from Section II-C (in particular, $K=L=2$). Let $\mathcal{K}$ denote the maximum ratio between the sizes of any two user or item communities. Choose some $\mathcal{M}_{k,l}$ and $C_{k,l}$ such that $\|R^{(k,l)}\|_* \leqslant \mathcal{M}_{k,l}$ and $\max_{i,j}|R^{(k,l)}_{i,j}| \leqslant C_{k,l}$ for all (k,l). Let $\hat{f}$ be the solution to the optimization problem*

$$\begin{aligned}
&\min \quad \hat{\mathcal{R}}(f) \text{ s.t. } \forall k,l, \\
&f = \sum_{k,l}(X^{(k)}M^{(k,l)}(Y^{(l)})^\top);\ \|M^{(k,l)}\|_* \leqslant \mathcal{M}_{k,l}; \\
&\text{and } \|X^{(k)}M^{(k,l)}(Y^{(l)})^\top\|_\infty \leqslant C_{k,l} \quad \forall k,l,j,i
\end{aligned} \tag{S.65}$$

With probability $\geqslant 1-\delta$ over the draw of the training set, the solution to the optimization problem (S.65) *satisfies*

$$\begin{aligned}
\mathcal{R}(\hat{f}) \leqslant\ & 2L_\ell\sqrt{\frac{9\mathcal{C}}{N}}\left(\frac{1}{\sqrt{c}}\sqrt{C_{1,1}\mathcal{M}_{1,1}}\left[\sqrt{a}+\sqrt{b}\right]^{1/2}\right. \\
&+ \frac{1}{\sqrt[4]{c}}\sqrt{C_{1,2}\mathcal{M}_{1,2}}\left[\sqrt{a}+\sqrt{n}\right]^{1/2} \\
&\left.+ \frac{1}{\sqrt[4]{c}}\sqrt{C_{2,1}\mathcal{M}_{2,1}}\left[\sqrt{m}+\sqrt{b}\right]^{1/2}\right.
\end{aligned}$$

$$+\sqrt{C_{2,2}\mathcal{M}_{2,2}}\left[\sqrt{m}+\sqrt{n}\right]^{1/2}\Bigg)$$
$$+2B\sqrt{\frac{\log(1/\delta)}{2M}}+\mathcal{E}. \tag{S.66}$$

Expressed in terms of matrix ranks instead, we obtain:

$$\begin{aligned}\mathcal{R}(\hat{f}) \leqslant 2L_\ell\sqrt{\frac{9\mathcal{C}}{N}}\Bigg(&\sqrt{\mathcal{K}}C_{1,1}\sqrt[4]{abr_{1,1}}\left[\sqrt{a}+\sqrt{b}\right]^{1/2}\\&+\sqrt[4]{\mathcal{K}}C_{1,2}\sqrt[4]{anr_{1,2}}\left[\sqrt{a}+\sqrt{n}\right]^{1/2}\\&+\sqrt[4]{\mathcal{K}}C_{2,1}\sqrt[4]{mbr_{2,1}}\left[\sqrt{m}+\sqrt{b}\right]^{1/2}\\&+C_{2,2}\sqrt[4]{mnr_{2,2}}\left[\sqrt{m}+\sqrt{n}\right]^{1/2}\Bigg)\\&+2B\sqrt{\frac{\log(1/\delta)}{2M}}+\mathcal{E}.\end{aligned} \tag{S.67}$$

(Here $\mathcal{C}$ is an absolute constant and c is the number of elements in the smallest community.)

Proof. For convenience we prove a slightly more general result where c_1 (resp. c_2) is the size of the smallest community of users (resp. items). Let X' and Y' denote matrices whose columns are the (non-normalised) indicator functions of the communities. By Lemma E.1, the Rademacher complexity of the function class $= \{X'M(Y')^\top \| \|M\|_* \leqslant \mathcal{M}_{1,1} \wedge \|X'M(Y')^\top\|_\infty \leqslant C_{1,1}\}$ is bounded by

$$\sqrt{C_{1,1}\frac{9\mathcal{M}_{1,1}\mathcal{C}(\sqrt{a}+\sqrt{b})}{N}}.$$

Now observe that by Lemma E.3, the function class

$$\begin{aligned}\mathcal{F}_{1,1} := \{&XM(Y)^\top \| \|M\|_* \leqslant \mathcal{M}_{1,1}\\&\wedge \|X^{(1)}M(Y^{(1)})^\top\|_\infty \leqslant C_{1,1}\}\end{aligned}$$

satisfies

$$\begin{aligned}\mathcal{F}_{1,1} \subset \{&X'M'(Y')^\top \| \|M'\|_* \leqslant \mathcal{M}_{1,1}c_1^{-1/2}c_2^{-1/2}\\&\wedge \|X'M'(Y')^\top\|_\infty \leqslant C_{1,1}\},\end{aligned}$$

where c_1 (resp. c_2) is the size of the smallest community of users (resp. items). It follows that

$$\mathfrak{R}(\mathcal{F}_{1,1}) \leqslant \frac{1}{\sqrt[4]{c_1c_2}}\sqrt{C_{1,1}\frac{9\mathcal{M}_{1,1}\mathcal{C}(\sqrt{a}+\sqrt{b})}{N}}.$$

By the same argument applied to the two situations where each user or item is a single community, we obtain the following results for $\mathcal{F}_{1,2} := \{X^{(1)}M(Y^{(2)})^\top \| \|M\|_* \leqslant \mathcal{M}_{1,2} \wedge \|X^{(1)}M(Y^{(2)})^\top\|_\infty \leqslant C_{1,2}\}$, $\mathcal{F}_{2,1} := \{X^2M(Y^1)^\top \| \|M\|_* \leqslant \mathcal{M}_{2,1} \wedge \|X^{(2)}M(Y^{(1)})^\top\|_\infty \leqslant C_{2,1}\}$ and $\mathcal{F}_{2,2} := \{X^{(2)}M(Y^{(2)})^\top \| \|M\|_* \leqslant \mathcal{M}_{2,2} \wedge \|X^{(2)}M(Y^2)^\top\|_\infty \leqslant C_{2,2}\}$:

$$\mathfrak{R}(\mathcal{F}_{1,2}) \leqslant$$
$$\mathfrak{R}(\widetilde{\mathcal{F}_{1,2}}) \leqslant \frac{1}{\sqrt[4]{c_1}}\sqrt{C_{1,2}\frac{9\mathcal{M}_{1,2}\mathcal{C}(\sqrt{a}+\sqrt{n})}{N}};$$

$$\mathfrak{R}(\mathcal{F}_{2,1}) \leqslant$$
$$\mathfrak{R}(\widetilde{\mathcal{F}_{2,1}}) \leqslant \frac{1}{\sqrt[4]{c_2}}\sqrt{C_{2,1}\frac{9\mathcal{M}_{2,1}\mathcal{C}(\sqrt{m}+\sqrt{b})}{N}};$$

and

$$\mathfrak{R}(\mathcal{F}_{2,2}) \leqslant$$
$$\mathfrak{R}(\widetilde{\mathcal{F}_{2,2}}) \leqslant \sqrt{C_{2,2}\frac{9\mathcal{M}_{2,2}\mathcal{C}(\sqrt{m}+\sqrt{n})}{N}};$$

after noting that $\mathcal{F}_{1,2} \subset \widetilde{\mathcal{F}}_{1,2} := \{X^{(1)}M(\tilde{Y}^{(2)})^\top \| \|M\|_* \leqslant \mathcal{M}_{1,2} \wedge \|X^{(1)}M(\tilde{Y}^{(2)})^\top\|_\infty \leqslant C_{1,2}\} = \{X^{(1)}M \| \|M\|_* \leqslant \mathcal{M}_{1,2} \wedge \|X^{(1)}M\|_\infty \leqslant C_{1,2}\}$; $\mathcal{F}_{2,1} \subset \widetilde{\mathcal{F}}_{2,1} := \{\tilde{X}^{(2)}M(Y^{(2)})^\top \| \|M\|_* \leqslant \mathcal{M}_{2,1} \wedge \|\tilde{X}^{(2)}M(Y^{(2)})^\top\|_\infty \leqslant C_{2,1}\} = \{M(Y^2)^\top \| \|M\|_* \leqslant \mathcal{M}_{2,1} \wedge \|M(Y^{(2)})^\top\|_\infty \leqslant C_{2,1}\}$; and $\mathcal{F}_{2,2} \subset \widetilde{\mathcal{F}}_{2,2} := \{M \| \|M\|_* \leqslant \mathcal{M}_{2,2} \wedge \|M\|_\infty \leqslant C_{2,2}\}$. Here $\tilde{X}^{(1)} = X^{(1)}$ is a matrix whose columns are the (non normalised) indicator functions of the communities, and $\tilde{X}^{(2)}$ and $\tilde{Y}^{(2)}$ are identity matrices.

By the subadditivity of Rademacher complexity, Talagrand's lemma and the classic Rademacher theorem then immediately yield the first result:

$$\begin{aligned}\mathcal{R}(\hat{f}) &\leqslant 2B\sqrt{\frac{\log(1/\delta)}{2M}}+\mathcal{E}+\\&\quad 2L_\ell\left[\mathfrak{R}(\mathcal{F}_{1,1})+\mathfrak{R}(\mathcal{F}_{1,2})+\mathfrak{R}(\mathcal{F}_{2,1})+\mathfrak{R}(\mathcal{F}_{2,2})\right]\\&\leqslant 2L_\ell\sqrt{\frac{9\mathcal{C}}{N}}\Bigg(\frac{1}{\sqrt{c}}\sqrt{C_{1,1}\mathcal{M}_{1,1}}\left[\sqrt{a}+\sqrt{b}\right]^{1/2}\\&\quad+\frac{1}{\sqrt[4]{c}}\sqrt{C_{1,2}\mathcal{M}_{1,2}}\left[\sqrt{a}+\sqrt{n}\right]^{1/2}\\&\quad+\frac{1}{\sqrt[4]{c}}\sqrt{C_{2,1}\mathcal{M}_{2,1}}\left[\sqrt{m}+\sqrt{b}\right]^{1/2}\\&\quad+\sqrt{C_{2,2}\mathcal{M}_{2,2}}\left[\sqrt{m}+\sqrt{n}\right]^{1/2}\Bigg)\\&\quad+2B\sqrt{\frac{\log(1/\delta)}{2M}}+\mathcal{E}.\end{aligned} \tag{S.68}$$

Regarding the second result, note that since the entries of $M_{1,1}$ are bounded by $C_{1,1}\bar{c}$ where $\bar{c}$ is the

size of the largest community, we have $\|M_{1,1}\| \leqslant \bar{c}C_{1,1}\sqrt{abr_{1,1}}$, $\|M_{1,2}\| \leqslant \sqrt{\bar{c}}C_{1,2}\sqrt{anr_{1,2}}$, $\|M_{2,1}\| \leqslant \sqrt{\bar{c}}C_{2,1}\sqrt{mbr_{2,1}}$ and $\|M_{2,2}\|_* \leqslant C_{2,2}\sqrt{mnr_{2,2}}$. Plugging this back into the first result yields the second result, as expected.

□

If we set the number of user and item communities to one, we obtain the BOMIC model. Furthermore, results can also be similarly extended to the BOMIC+ model, obtaining the full theorem below.

Theorem E.2. *For the BOMIC algorithm from Subsection II-B, with probability $\geqslant 1-\delta$ over the draw of the training set, we have*

$$\begin{aligned}\mathcal{R}(\hat{f}) \leqslant\ & 2L_\ell\sqrt{\frac{9\mathcal{C}}{N}}\Big(C_{1,1} + C_{1,2}\sqrt[4]{n}\left[1+\sqrt{n}\right]^{1/2} \\ &+ C_{2,1}\sqrt[4]{m}\left[\sqrt{m}+1\right]^{1/2} \\ &+ C_{2,2}\sqrt[4]{mnr_{2,2}}\left[\sqrt{m}+\sqrt{n}\right]^{1/2}\Big) \\ &+ 2B\sqrt{\frac{\log(1/\delta)}{2M}} + \mathcal{E}. \end{aligned} \tag{S.69}$$

For the BOMIC+ algorithm from Subsection II-B, with probability $\geqslant 1-\delta$ over the draw of the training set, we have

$$\begin{aligned}\mathcal{R}(\hat{f}) \leqslant\ & 2L_\ell\sqrt{\frac{9\mathcal{C}}{N}}\Big(C_{1,1} \\ &+ C_{1,2}\sqrt[4]{b}\left[1+\sqrt{b}\right]^{1/2} + C_{2,1}\sqrt[4]{a}\left[1+\sqrt{a}\right]^{1/2} \\ &+ C_{2,2}\sqrt[4]{abr_{2,2}}\left[\sqrt{a}+\sqrt{b}\right]^{1/2} \\ &+ C_{1,3}\sqrt[4]{n}\left[\sqrt{n}+1\right]^{1/2} + C_{3,1}\sqrt[4]{m}\left[\sqrt{m}+1\right]^{1/2} \\ &+ C_{2,3}\sqrt[4]{anr_{2,3}}\left[\sqrt{a}+\sqrt{n}\right]^{1/2} \\ &+ C_{3,2}\sqrt[4]{bmr_{3,2}}\left[\sqrt{m}+\sqrt{b}\right]^{1/2} \\ &+ C_{3,3}\sqrt[4]{mnr_{3,3}}\left[\sqrt{m}+\sqrt{n}\right]^{1/2}\Big) \\ &+ 2B\sqrt{\frac{\log(1/\delta)}{2M}} + \mathcal{E}. \end{aligned} \tag{S.70}$$

B. Bounds under the assumption of uniform marginals

In this section, we prove bounds assuming that the sampling distribution has the probability that each row and each column each have equal probability of being sampled.

Proposition E.4. *Let $\mathcal{F}_\mathcal{M}$ be the class of functions $R \in \mathbb{R}^{m\times n} : \|R\|_* \leqslant \mathcal{M}$, where as usual wlog $m \geqslant n$ and we also assume $m \geqslant 3$. Further assume that the sampling distribution has uniform marginals. We have the following bound on the expected Rademacher complexity of $\mathcal{F}$:*

$$\mathbb{E}(\mathfrak{R}_N(\mathcal{F})) \leqslant 20(\mathcal{M}/\sqrt{n})\sqrt{\frac{\log(m)}{N}}. \tag{S.71}$$

Proof. Writing $x_1, \ldots, x_N$ for the iid samples from $\{1,2,\ldots,m\} \times \{1,2,\ldots,n\}$, $\sigma_1,\ldots,\sigma_N$ for the Rademacher variables and Σ for the matrix with $\Sigma_{i,j} = \sum_k \sigma_k 1_{x_k=(i,j)}$, we have by the duality of the trace norm

$$\begin{aligned}\mathbb{E}(\mathfrak{R}_N(\mathcal{F})) &= \mathbb{E}_{x,\sigma}\frac{1}{N}\sup_{R\in\mathcal{F}}(\langle R,\Sigma\rangle) \\ &\leqslant \frac{\mathcal{M}}{N}\mathbb{E}(\|\Sigma\|). \end{aligned} \tag{S.72}$$

Note that $\sigma = \sum_{k=1}^N X_k$ where $X_k = \sigma_k 1_{x_k}$. Note also that the $\mathbb{E}(X_k X_k^\top)$ (resp. $\mathbb{E}(X_k^\top X_k)$) is a diagonal matrix whose ith entry is the sampling probability of the ith row (resp. column). Thus, by our uniform marginal assumption, we have using the notation from proposition F.3 $\rho_k = \sqrt{1/n}$ and $\sigma = \sqrt{N/n}$.

Thus, by the Bernstein inequality in expectation (proposition F.3), we can continue, assuming without loss of generality that $m \geqslant 2$ and $N \geqslant m\log(m)$:

$$\begin{aligned}&\mathbb{E}(\mathfrak{R}_N(\mathcal{F})) \\ &\leqslant \frac{\mathcal{M}}{N}\Big[\sqrt{8/3}\sigma(1+\sqrt{\log(m+n)}) \\ &\qquad + \frac{8}{3}(1+\log(m+n))\Big] \\ &\leqslant \frac{\mathcal{M}}{N}\Big[\sqrt{8/3}\sigma(1+\sqrt{2\log(m)}) \\ &\qquad + \frac{8}{3}(1+2\log(m))\Big] \\ &\leqslant \frac{\mathcal{M}}{N}\Big[\sqrt{8/3}\sigma\sqrt{\log(m)}(1/\sqrt{\log(2)}+\sqrt{2}) \\ &\qquad + \frac{8}{3}(1/\sqrt{\log(2)}+2)\sigma\sqrt{\log(m)}\Big] \\ &\leqslant 20\sigma(\mathcal{M}/N)\sqrt{\log(m)} \\ &\leqslant \frac{20(\mathcal{M}/\sqrt{n})\sqrt{\log(m)}}{\sqrt{N}}, \end{aligned} \tag{S.73}$$

as expected. □

Theorem E.3. *Consider the community side information setting of Section II-C (OMIC+, in particular, $K = L = 2$). Let $\mathcal{K}$ denote the maximum ratio between the sizes of any two user or item communities. Assume that the sampling distribution has uniform marginals. Choose some $\mathcal{M}_{k,l}$ such that*

$\|R^{(k,l)}\|_* \leqslant \mathcal{M}_{k,l}$ *for all* (k,l). *Let* $\hat{f}$ *be the solution to the optimization problem*

$$\begin{aligned} &\min \quad \hat{\mathcal{R}}(f) \text{ s.t. } \forall k,l, \\ &f = \sum_{k,l} (X^{(k)} M^{(k,l)} (Y^{(l)})^\top);\ \|M^{(k,l)}\|_* \leqslant \mathcal{M}_{k,l}. \end{aligned} \tag{S.74}$$

In expectation over the draw of the training set, the solution to the optimization problem (S.74) *satisfies*

$$\begin{aligned} &\mathbb{E}(\mathcal{R}(\hat{f})) \\ &\leqslant \frac{40 L_\ell}{\sqrt{N}} \left[\frac{1}{c} \frac{\mathcal{M}_{1,1}}{\sqrt{a}} \sqrt{\log(b)} + \frac{1}{\sqrt{c}} \frac{\mathcal{M}_{1,2}}{\sqrt{a}} \sqrt{\log(m)} \right. \\ &\qquad \left. + \frac{1}{\sqrt{c}} \frac{\mathcal{M}_{2,1}}{\sqrt{b}} \sqrt{\log(n)} + \frac{\mathcal{M}_{2,1}}{\sqrt{n}} \sqrt{\log(m)} \right] + \mathcal{E}. \end{aligned} \tag{S.75}$$

Expressed in terms of the ranks and maximum sizes of the entries $r_{k,l}, C_{k,l}$, *chosen to satisfy the feasibility condition* $\operatorname{rank}(R_{k,l}) \leqslant r_{k,l}$ *and* $\|R^{(k,l)}\|_\infty \leqslant C_{k,l}$, *we have*

$$\begin{aligned} &\mathbb{E}(\mathcal{R}(\hat{f})) \\ &\leqslant \frac{40 L_\ell}{\sqrt{N}} \left[\mathcal{K} C_{1,1} \sqrt{b r_{1,1} \log(b)} \right. \\ &\qquad + \sqrt{\mathcal{K}} C_{1,2} \sqrt{m r_{1,2} \log(m)} \\ &\qquad + \sqrt{\mathcal{K}} C_{2,1} \sqrt{n r_{2,1} \log(n)} \\ &\qquad \left. + C_{2,2} \sqrt{m r_{2,2} \log(m)} \right] + \mathcal{E}. \end{aligned} \tag{S.76}$$

Proof. The proof is similar to that of Theorem E.1.

Let $\ell \circ \mathcal{F}$ be the loss function class associated with the first problem.

By the Talagrand Lemma and the subadditivity of Rademacher complexity, we have

$$\begin{aligned} &\mathfrak{R}(\ell \circ \mathcal{F}) \\ &\leqslant 2 L_\ell \left[\mathfrak{R}(\mathcal{F}_{1,1}) + \mathfrak{R}(\mathcal{F}_{1,2}) + \mathfrak{R}(\mathcal{F}_{2,1}) + \mathfrak{R}(\mathcal{F}_{2,2}) \right], \end{aligned} \tag{S.77}$$

where $\mathcal{F}_{k,l} := \{X M (Y)^\top | \|M\|_* \leqslant \mathcal{M}_{k,l}\}$.

Next, note, similarly to the proof of Theorem E.1, that

$$\begin{aligned} &\mathcal{F}_{1,1} \subset \widetilde{\mathcal{F}_{1,1}} := \\ &\{\widetilde{X}^1 M' (\widetilde{Y}^1)^\top | \|M'\|_* \leqslant \mathcal{M}_{1,1} c_1^{-1/2} c_2^{-1/2}\}, \end{aligned}$$

and thus, by Proposition E.4,

$$\mathfrak{R}(\mathcal{F}_{1,1}) \leqslant \mathfrak{R}(\widetilde{\mathcal{F}_{1,1}}) \leqslant \frac{20}{\sqrt{N}\sqrt{c_1 c_2}} \frac{\mathcal{M}_{1,1}}{\sqrt{a}} \sqrt{\log(b)}, \tag{S.78}$$

with similar results for the other terms.

Plugging this into equation (S.77) yields the first result. The final result then follows after noting that $\|M_{1,1}\|_* \leqslant \sqrt{c_1 c_2} C_{1,1} \sqrt{abr_{1,1}}$, $\|M_{1,2}\|_* \leqslant \sqrt{c_1} C_{1,2} \sqrt{anr_{1,2}}$, $\|M_{2,1}\|_* \leqslant \sqrt{c_2} C_{2,1} \sqrt{mbr_{1,1}}$, and $\|M_{2,2}\|_* \leqslant C_{2,2} \sqrt{mnr_{1,1}}$.

□

C. In-depth discussion of bounds

In the case of OMIC+ and for a fixed tolerance threshold ϵ, our distribution-free sample complexity bounds scale as

$$\begin{aligned} &\mathcal{K} C_{1,1}^2 b \sqrt{a r_{1,1}} + \sqrt{\mathcal{K}} C_{1,2}^2 n \sqrt{a r_{1,2}} \\ &+ \sqrt{\mathcal{K}} C_{2,1}^2 m \sqrt{b r_{2,1}} + C_{2,2}^2 m \sqrt{r_{2,2} n}, \end{aligned}$$

whilst in the case of uniform marginals they scale like

$$\begin{aligned} &\mathcal{K}^2 C_{1,1}^2 b r_{1,1} \log(b) + \mathcal{K} C_{1,2}^2 m r_{1,2} \log(m) \\ &+ \mathcal{K} C_{2,1}^2 n r_{2,1} \log(n) + C_{2,2}^2 m r_{2,2} \log(m). \end{aligned}$$

In both cases, if $C_{1,1}$ is much larger than $C_{1,2}, C_{2,1}$ and $C_{2,2}$, the bound behaves similarly to the situation where the users and items are identified with their category. Whilst this is not very surprising, the bound does show that this remains true for small but non zero values of $C_{1,2}, C_{2,1}$ and $C_{2,2}$: the model can effectively learn a combination of community behaviour and user behaviour with no further difficulty than if it were learning both problems independently. Note also conversely that if a, b are very small (as in the particular case BOMIC where $a = b = 1$), the first three terms are very small and the bound essentially tells us that the model is about as hard as learning the low-rank residual alone.

Note that the bounds show that prior knowledge of the community structure helps more than knowledge of the generic low-rank structure. Indeed, consider a typical situation where the communities are of equal and small size, $C_{1,2} = C_{2,1} = 0$, $a = b$, $m = n$ and $r_{1,1} = a$, and assume that the absolute values of the maximum and minimum entries of $R^{(2,2)}$ are of the same order so that the maximum absolute value of an entry of R is $\simeq C_{1,1} + C_{2,2}$. With knowledge of the community structure, our model requires in the distribution-free case $O(C_{1,1}^2 a^2 + C_{2,2}^2 m^{3/2} \sqrt{r_{2,2}})$ entries. If we were to apply a generic low rank method instead, the number of required entries would then be $O((C_{2,2} + C_{1,1})^2 m \sqrt{(r_{2,2} + a)m})$. More generally, we see that ignoring the change in the $C_{k,l}$'s, absorbing ground truth community component

of order r into the generic low-rank term results in an increase of at least $O([\sqrt{r_{2,2}+r}-\sqrt{r_{2,2}}]m^{3/2})$ in the number of required entries, whilst the sample complexity only grows by $O(a^{3/2}[\sqrt{r_{1,1}+r}-\sqrt{r_{1,1}}])$ instead when the community side information is duly considered. Similar results hold for the case of uniform marginals, where absorbing a community component of rank r into the generic low rank component costs $O(rm\log(m))$ in sample complexity, compared to $O(ra\log(a))$ when the side information is duly considered.

Admittedly, absorbing cross terms ($R^{(2,1)}$ or $R^{(1,2)}$) into the generic low-rank component does not cause the bound for uniform marginals to change at the asymptotic level. On the other hand, the *distribution-free bounds* still show a significant advantage in exploiting the community side information even when it comes to cross terms. Indeed, consider a similar situation to above with $a=b=1$ (BOMIC) and $C_{2,1}=C_{1,1}=0$. Absorbing the cross term $R^{(1,2)}$ into the generic low rank component results in a sample complexity bound of order $O((C_{2,2}+C_{1,2})^2\sqrt{r_{2,2}+1}n^{3/2})$. Ignoring again the effects on the $C_{k,l}$ for simplicity, this represents an increase of at least $O([\sqrt{r_{2,2}+1}-\sqrt{r_{2,2}}]n^{3/2})$ to the fourth term, compared to a contribution of order $O(n)$ when the side information is taken into account. This means that (at least for small values of $r_{2,2}$), the knowledge of the specific singular vector (i.e. the singular vector $(1/\sqrt{n})_{i\leqslant n}$) helps our model more than the simple knowledge of the equivalent restriction in the rank.

Remarks on proof techniques and comparison to IMC bounds

The bounds admittedly follow reasonably straightforwardly from similar techniques as those used for standard matrix completion, applied to auxiliary problems corresponding to each of the four terms $(X^{(1)}M^{(1,1)}(Y^1)^\top$, $X^{(1)}M^{(1,2)}(Y^2)^\top$, $X^{(2)}M^{(2,1)}(Y^1)^\top$ and $X^{(2)}M^{(2,2)}(Y^2)^\top)$ independently and then merged via the subadditivity of Rademacher complexity. Indeed, in the distribution-free case state-of-the-art bounds take the form (cf. [34], [17], [36] etc.)

$$O\left(\sqrt{\frac{\mathcal{M}(\sqrt{n}+\sqrt{m})}{N}}\right) \tag{S.79}$$

where $\mathcal{M}$ is a bound on the nuclear norm and in the case of uniform marginals, bounds of the form

$$O\left(\sqrt{\frac{(\mathcal{M}^2/n)\log(m)}{N}}\right) \tag{S.80}$$

were proved in [37].

However, we note that the state-of-the-art bounds for *inductive matrix completion* (whose predictors take the form $XMY^\top$ for some fixed side information X, with nuclear norm minimization at work on M), applied to any of the first three terms in question do **not** yield bounds as tight as ours.

Indeed, there is no suitable equivalent to (S.79) or (S.80) in the inductive case, and the state-of-the-art bounds for inductive matrix completion (cf. [36], [16], etc.), actually scale like $O\left(\mathbf{x}\mathbf{y}\mathcal{M}\sqrt{\frac{1}{N}}\right)$ (or rd_1d_2 where d_1 and d_2 are the dimensions of the side information) instead, where $\mathbf{x}$ (resp. $\mathbf{y}$) stands for the maximum norm of a row of X (resp. Y).

In the representative case $C_{1,2}=C_{2,1}=C_{2,2}=0$, our distribution-free bound (S.67) takes the form

$$\begin{aligned} &O\left(\sqrt{\frac{C_{1,1}\mathcal{M}_{1,1}\left[\sqrt{a}+\sqrt{b}\right]}{cN}}\right) \\ &\simeq O\left(C_{1,1}\sqrt{\frac{\mathcal{K}\sqrt{abr_{1,1}}\left[\sqrt{a}+\sqrt{b}\right]}{N}}\right), \end{aligned} \tag{S.81}$$

whereas the bound (S.79) takes the form

$$\frac{1}{c\sqrt{N}}\mathcal{M}_{1,1}\simeq\frac{\sqrt{r_{1,1}}\sqrt{ab\bar{c}}}{\sqrt{N}c}=\frac{\sqrt{r_{1,1}}\sqrt{ab}\mathcal{K}}{\sqrt{N}}. \tag{S.82}$$

In terms of sample complexity, this corresponds to a required number of entries of the order of $abr_{1,1}$, in line with the results of [16]. This sample complexity bound makes an excellent use of the side information in the sense that the bound is *independent of the size of the matrix*, but further refining the dependence on the *dimensions of the side information*, which is relevant in our case, was not one of the aims of that paper. In the case where the side information is composed of identity functions, this result is clearly vacuous, contrary to our own, which instead then scales as the state-of-the-art bounds for matrix completion (from which it was derived).

D. Experimental verification of the bounds under uniform marginals

In this section, we aim to experimentally validate our generalisation bounds. We focus on the case of uniform marginals for the following reasons: even without side information, for the case of traditional matrix completion for a square $n\times n$ matrix, it is not clear in what sense the bound $O(n^{3/2}\sqrt{r})$ is tight.

Indeed, it certainly is tight for a full rank matrix, but this is not very informative, since in that case, it simply says that the required number of entries is of the same order as the number of entries in the matrix. The case of a lower rank constraint is less clear. This makes it difficult to design a sampling regime and a way to evaluate the bound.

To verify and experimentally explore the behaviour of our bound which applies to the case of uniform marginals, we construct some ground truth matrices with $C_{1,2} = C_{2,1} = 0$, $C_{1,1} = 1$, $a = b$, $m = n$. Set $r_{1,1} = a$ and we vary the values of a, $C = C_{2,2}$ and $r_{2,2}$.

Ground truth matrix generation In each case, the matrix $A := R^{(1,1)} = \frac{aM^{(1,1)}}{m}$ is constructed with iid Rademacher entries, i.e., the entries of the community×community component of the matrix are iid Rademacher variables. For each rank $r = r_{2,2}$, we construct a matrix B as follows: generate r iid column vectors V (resp. W) in $\mathbb{R}^{m-a}$ (resp. $\mathbb{R}^{n-b}$) whose entries are $N(0, 1/(m-a))$ (resp. $N(0, 1/(n-b))$). We then form the matrix $\tilde{B} := X^{(2)}VW^\top(Y^{(2)})^\top$, and finally obtain the matrix B by dividing by the maximum entry in absolute value: $B = \tilde{B}/\{\max_{i,j}|\tilde{B}|\}$. The final matrix is constructed as $A + CB$.

Sampling regime: we evaluate two sampling regimes with uniform marginals:

- Uniform sampling
- Checkerboard sampling: uniformly sample from entries (i, j) with $i = j \mod 2$.

Training procedure: we train the bias-free part of OMIC+ via our algorithm. To isolate variables and for simplicity, $\lambda_{1,2}$ and $\lambda_{2,1}$ are both set to ∞. To evaluate the performance of our algorithm on many sparsity regimes, we pick an ordering $O = \{O_1, \ldots, O_{(mn)!}\}$ (or $O = \{O_1, \ldots, O_{(mn/4)!}\}$ if in the other sampling regime) of all entries (here $O_1 \subset \ldots \subset O_{(mn)!} = \{1, 2, \ldots m\} \times \{1, 2, \ldots, n\}$ are increasing subsets of the set of entries. We train our algorithm for Ω taking each value in O, using the previous value as a warm start to dramatically reduce computation For simplicity, hyperparameters $\lambda_{1,1}$ and $\lambda_{2,2}$ are determined through cross validation *on a fixed, reasonably performing sparsity regime* once and for all.

We set a tolerance threshold $\epsilon = 0.1$ based on convergence analysis, and for each configuration of a, r, C, we compute the minimum N_ϵ such that the algorithm achieves RMSE $\leqslant \epsilon$ for the set O_{N_ϵ}.

Figure 6 is a graph of N_ϵ versus the bound in III.2 for various values of a, C, r, n. We set $\epsilon = 0.1$, $a = b$. We explored the following values: $a \in \{2, 3, \cdots, 8\}$ $r \in \{2, 3, \cdots, 8\}$ $C \in [0.5, 2]$ $m \in \{100, 101, \cdots, 400\}$. For each datapoint, we choose a random combination of the above parameters, generate a random matrix accordingly, and compare the N_ϵ to our bound. We observe a very good match between the bound and the observed de facto sample complexity.

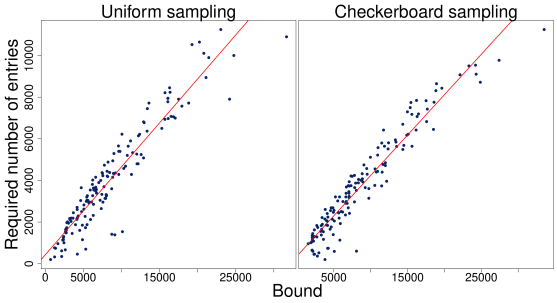

Figure 6: Comparison of our bound in III.2 and the observed required number of entries to reach 0.1 validation RMSE. The red lines are obtained through linear regression.

APPENDIX F
MISCELLANEOUS LEMMAS

Recall the definition of the Rademacher complexity of a function class $\mathcal{F}$:

Definition F.1. *Let $\mathcal{F}$ be a class of real-valued functions with range X. Let also $S = (x_1, x_2, \ldots, x_n) \subset X$ be n samples from the domain of the functions in $\mathcal{F}$. The empirical Rademacher complexity $\mathfrak{R}_S(\mathcal{F})$ of $\mathcal{F}$ with respect to $x_1, x_2, \ldots, x_n$ is defined by*

$$\mathfrak{R}_S(\mathcal{F}) := \mathbb{E}_\delta \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \delta_i f(\mathbf{x}_i), \tag{S.83}$$

where $\delta = (\delta_1, \delta_2, \ldots, \delta_n) \in \{\pm 1\}^n$ is a set of n i.i.d. Rademacher random variables (which take values 1 or -1 with probability 0.5 each).

Recall the following classic theorem [38], [39], [40]:

Theorem F.1. *Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables taking values in a set $\mathcal{Z}$, and let $a < b$. Consider a set of functions $\mathcal{F} \in [a, b]^{\mathcal{Z}}$. $\forall \delta > 0$, we have with probability $\geqslant 1 - \delta$ over the draw of the sample S that*

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(f(Z)) \tag{S.84}$$
$$\leqslant \frac{1}{n}\sum_{i=1}^{n} f(z_i) + 2\mathbb{E}(\mathfrak{R}_S(\mathcal{F})) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

With a simple extra integration argument, we obtain the following version in expectation:

Theorem F.2. *Let $Z, Z_1, \dots, Z_n$ be i.i.d. random variables taking values in a set $\mathcal{Z}$, and let $a < b$. Consider a set of functions $\mathcal{F} \in [a,b]^{\mathcal{Z}}$. $\forall \delta > 0$, we have in expectation over the draw of the sample S that*

$$\inf_{f\in\mathcal{F}} \left(\mathbb{E}(f(Z)) - \frac{1}{n}\sum_{i=1}^n f(z_i) \right) \leqslant 2\mathbb{E}(\mathfrak{R}_S(\mathcal{F})) + 3(b-a)\sqrt{\frac{1}{n}}. \tag{S.85}$$

Proof. Let $X = \inf_{f\in\mathcal{F}} \left(\mathbb{E}(f(Z)) - \frac{1}{n}\sum_{i=1}^n f(z_i) \right) - 2\mathbb{E}(\mathfrak{R}_S(\mathcal{F}))$. Let us also write $\phi(\delta)$ for $(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}$. By (S.84), we have

$$\mathbb{P}\left(X \geqslant \phi(\delta)\right) \leqslant \delta \tag{S.86}$$

For all $i \geqslant 1$, let us write A_i for the event $\{X \leqslant \phi(\delta_i)\}$ where $\delta_i = 2^{-i}$. Let us also write $\widetilde{A}_1 := A_1$ and for $i \geqslant 2$, $\widetilde{A}_i := A_i \backslash A_{i-1}$. We have, for $i \geqslant 2$, $\mathbb{P}(\widetilde{A}_i) \leqslant \mathbb{P}(A_{i-1}^c) \leqslant \delta_{i-1}$, and for $i = 1$, $\mathbb{P}(\widetilde{A}_1) \leqslant 1 = \delta_{i-1}$.

Thus,

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{i=1}^\infty \mathbb{E}(X|\widetilde{A}_i)\mathbb{P}(\widetilde{A}_i) \\
&\leqslant \sum_{i=1}^\infty \mathbb{E}(X|\widetilde{A}_i)\delta_{i-1} \\
&= \sum_{i=1}^\infty \phi(\delta_i)\frac{1}{2^{i-1}} \\
&= \sum_{i=1}^\infty (b-a)\sqrt{\frac{1+i}{2n}}\frac{1}{2^{i-1}} \\
&\leqslant (b-a)\sqrt{\frac{2}{n}}\sum_{i=1}^\infty 2^{-i/2} \\
&\leqslant (b-a)\sqrt{\frac{2}{n}}\frac{1/\sqrt{2}}{1-1/\sqrt{2}} \\
&= (b-a)\sqrt{\frac{2}{n}}\frac{1}{\sqrt{2}-1} \\
&\leqslant 5(b-a)\sqrt{\frac{1}{n}},
\end{aligned}$$

as expected. □

Proposition F.2 (Cf. [41])**.** *Let $X_1, \dots, X_S$ be independent, zero mean random matrices of dimension $m \times n$. For all k, assume $\|X_k\| \leqslant M$ almost surely, and denote $\rho_k^2 = \max(\|\mathbb{E}(X_k X_k^\top)\|, \|\mathbb{E}(X_k^\top X_k)\|)$. For any $\tau > 0$,*

$$\begin{aligned}
&\mathbb{P}\left(\left\| \sum_{k=1}^S X_k \right\| \geqslant \tau \right) \\
&\leqslant (m+n)\exp\left(-\frac{\tau^2/2}{\sum_{k=1}^S \rho_k^2 + M\tau/3} \right).
\end{aligned} \tag{S.87}$$

Proposition F.3. *Under the assumptions of Proposition F.2, writing $\sigma^2 = \sum_{k=1}^S \rho_k^2$, we have*

$$\begin{aligned}
\mathbb{E}\left(\left\| \sum_{k=1}^S X_k \right\| \right) &\leqslant \sqrt{8/3}\sigma(1+\sqrt{\log(m+n)}) \\
&\quad + \frac{8M}{3}(1+\log(m+n)).
\end{aligned} \tag{S.88}$$

Proof. The result in O notation is an exercise from [42], and a similar result is also mentioned in both [37] and [43].

For completeness and to get the exact constants, we include a proof as follows.

Let $Y = \left\| \sum_{k=1}^S X_k \right\|$. By Proposition F.2, splitting into two cases depending on whether $\tau M \leqslant \sigma^2$ or $\tau M \geqslant \sigma^2$ we have

$$\begin{aligned}
\mathbb{P}(Y \geqslant \tau) &\leqslant \min\left(1, (m+n)\exp\left[-\frac{3\tau^2}{8\sigma^2}\right]\right) \\
&\quad + \min\left(1, (m+n)\exp\left[-\frac{3\tau}{8M}\right]\right).
\end{aligned} \tag{S.89}$$

Now note that writing κ for $\log(m+n)8M/3$, we have

$$\begin{aligned}
&\int_0^\infty 1 \wedge (m+n)\exp\left(-\frac{3\tau}{8M}\right) d\tau \\
&\leqslant \int_0^\kappa 1 \wedge (m+n)\exp\left(-\frac{3\tau}{8M}\right) d\tau \\
&\qquad + \int_\kappa^\infty (m+n)\exp\left(-\frac{3\tau}{8M}\right) d\tau \\
&\leqslant \kappa + \left[\frac{-8M}{3}(m+n)\exp\left(-\frac{3\tau}{8M}\right)\right]_\kappa^\infty \\
&= \kappa + \frac{8M(m+n)}{3}\exp\left(-\frac{3\kappa}{8M}\right) \\
&= \kappa + \frac{8M(m+n)}{3} = \frac{8M}{3}(1+\log(m+n)).
\end{aligned} \tag{S.90}$$

We also have, writing ψ for $\sigma\sqrt{\log(m+n)8/3}$,

$$\int_0^\infty 1 \wedge (m+n)\exp\left(-\frac{3\tau^2}{8\sigma^2}\right) d\tau$$

$$
\begin{aligned}
&\leqslant \int_0^{\psi} 1 d\tau + \int_{\psi}^{\infty} (m+n) \exp\left(-\frac{3\tau^2}{8\sigma^2}\right) d\tau \\
&\leqslant \psi + \int_{\psi}^{\infty} \exp\left(-\frac{3(\tau^2-\psi^2)}{8\sigma^2}\right) d\tau \\
&\leqslant \psi + \int_{\psi}^{\infty} \exp\left(-\frac{3(\tau-\psi)^2}{8\sigma^2}\right) d\tau \\
&\leqslant \psi + \sigma\sqrt{2\pi/3} \\
&= \sigma\left[\sqrt{\log(m+n)8/3} + \sqrt{2\pi/3}\right] \\
&\leqslant \sqrt{8/3}\sigma(1+\sqrt{\log(m+n)}). \qquad \text{(S.91)}
\end{aligned}
$$

Plugging inequalities (S.90) and (S.91) into equation (S.89), we obtain:

$$
\begin{aligned}
\mathbb{E}(Y) &\leqslant \int_0^{\infty} \mathbb{P}(Y \geqslant \tau) d\tau \\
&\leqslant \sqrt{8/3}\sigma(1+\sqrt{\log(m+n)}) \qquad \text{(S.92)} \\
&\quad + \frac{8M}{3}(1+\log(m+n)),
\end{aligned}
$$

as expected.

□

APPENDIX G
DETAILS OF THE MATRIX GENERATION PROCEDURE FOR THE SYNTHETIC DATA EXPERIMENTS

Our generation procedure can be described as follows: let $\tilde{a}$ be the vector with components $\tilde{a}_i = i - \frac{m+1}{2}$ and let $\tilde{b}$ be the vector with components $\tilde{a}_j = i - \frac{n+1}{2}$. Let $a = \frac{\tilde{a}}{\|\tilde{a}\|}$ and $b = \frac{\tilde{b}}{\|\tilde{b}\|}$. We also write $v_1 \in \mathbb{R}^m$ for $(\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}})^\top$ and $v_2 \in \mathbb{R}^n$ for $(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^\top$ Then we define $G = \frac{1}{2}av_2^\top + \frac{1}{2}v_1b^\top$ and $S \in \mathbb{R}^{m\times n}$ where $S_{i,j} = (1/mn)$, if $(i,j) \in \{1,\cdots,m/2\} \times \{1,\cdots,n/2\} \cup \{m/2+1,\cdots,m\} \times \{n/2+1,\cdots,n\}$, and $S_{i,j} = -(1/mn)$ otherwise. Therefore, we can generate a matrix $R \in \mathbb{R}^{m\times n}$ as

$$
R(\alpha) = \alpha c G + (1-\alpha) c S, \qquad \text{(S.93)}
$$

where $\alpha \in [0,1]$ is a parameter that controls the relative intensities of the user/item biases and the non-inductive component, and c is a scaling constant. Note that G is composed of the sum of two terms. The first term is a matrix with all rows being equal, whilst the second term's columns are all equal. Thus G is made up of user and item biases. On the other hand, the S matrix can be divided in four blocks of equal sizes. The top left and bottom right blocks entries have a constant value of $(1/mn)$. The remaining block has entries with the value $-(1/mn)$.

To perform the experiments we needed to select the parameters m, n, c and α. We chose[10] $m = n = c = 100$. The parameter $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ was empirically selected in such a way that the expected intensity of the biases' component varied. Note that in the extremes of the α interval the generated matrix is just composed of one of the components.

To determine the number of observed entries and the sampling distribution, we considered two extra parameters: the percentage of observed entries p_Ω and a parameter $\gamma \in \mathbb{N}$ that manages the sparsity distribution. Given a fixed p_Ω we randomly selected $\gamma(p_\Omega mn/(\gamma+1))$ entries in the first $m/2$ rows and $(p_\Omega mn/(\gamma+1))$ in the remaining ones. The parameter p_Ω was varied in $\{0.15, 0.30, 0.50\}$ and the parameter γ as varied in the range $\{1, 4\}$ ($\gamma = 1$ indicates uniformly sampled observations).

Note that for SoftImpute we need an extra post-processing step to estimate the biases. In this case, we calculate the matrix bias as the average of the SI-predicted matrix entries. After subtracting the SI matrix bias, we calculated the users and the items bias by averaging the columns and the rows, respectively.

APPENDIX H
IN-DEPTH LITERATURE REVIEW

A major step signaling the beginning of the construction of a formal theory of matrix completion was the introduction of the SoftImpute algorithm [5], which uses the nuclear norm as a regularizer. Around the same time, the field witnessed a series of breakthroughs in the study of how many entries are required to recover a low-rank matrix exactly [6], [7] or approximately from noisy entries [8], [9]. Those works assume that the entries of the matrix are sampled uniformly. A simpler and more complete approach to the same results was provided in both [41] and [44]. The conclusion of the works on exact recovery is that if the entries are sampled uniformly, it is possible to recover the matrix with high probability assuming $O(\mu rn \log(n)^2)$ entries are observed, where n is the dimension of the matrix, and μ is some notion of *coherence*, which is $O(1)$ if the singular vectors have roughly equal components.

Of course, there is a huge branch of literature focusing on the optimization aspect of matrix completion [45], [46], [47], [48], [49], [50], [51], [52].

In [11], user biases were trained jointly with other methods, including methods taking time dependence

[10]For a small number of incoherent eigenvectors, which is the situation in the case described here, the choice $c = \sqrt{mn}$ ensure entries of size close to one.

into account, but no nuclear-norm regularization was used.

Other works [53], [54], [55], [56] have focused on the case of non uniformly sampled entries. The general form of the results obtained is (similarly to the uniform case) that $O(rn\log(n)^2)$ observed entries are sufficient. However, these results come at the cost of either making strong explicit assumptions on the distributions, sometimes with relevant constants showing up in the bounds, or strong modifications of the algorithm.

The case of non-uniform entries with absolutely no assumption on the sampling distribution is an interesting one that commands a completely different approach. It was studied in [34], [57]. The most related work to ours is [17], where the authors study, and provide generalization bounds for a model composed of a sum of an IMC term and a standard SoftImpute model. This model is a particular case of ours. Note we require to adapt proofs to obtain bounds with a tighter dependence on the dimensions of both left and right side information for the bounds to be non trivial in case of user biases. Furthermore, no notion of interpretability or orthogonality was presented in [17].

Inductive matrix completion [12], [13], [14], [15] is the problem of solving matrix completion with some side information: given some features $X \in \mathbb{R}^{m\times d_1}$ and $Y \in \mathbb{R}^{n\times d_2}$, it tries to find a low-rank matrix M such that $R = XMY^\top$ approximates the observed matrix well. It has found many successful applications in recent years [58], [59], [60]. Theoretical guarantees were provided in [46], [61], [62]. Note that in the basic model, successful IMC requires that the columns of X (resp. Y) span the left (resp. right) singular vectors of the SVD of the ground truth matrix (this case is often referred to as "perfect" side information). In [17] the extended model $R = XMY^\top + N$, with nuclear-norm regularization applied to both M and N was proposed. Recently, progress was made in the direction of matrix completion with side information with the need to extract features jointly [63].

The idea of nuclear-norm minimization was also extended to tensors in various ways [64], [65], [66] as there is no unique approach which provides all the benefits the nuclear norm enjoys in the matrix case.

A. Models with similarities to ours

In this section, we explain the particulars of some recently proposed models which may be understood as using combinations of side information and generic low rank constraints, or other variations of parts of the ideas we propose. We note that in each case, substantial differences remain between the works in question and the present paper.

In [20], the authors introduce a model with similarities (and differences) to both [17] and the present work: assuming one is given side information matrices X and Y, the authors present the following model: first, the matrices X and Y are augmented by a columns of ones resulting in the matrices $\bar{X} = [X, \mathbf{1}]$ and $\bar{Y} = [Y, \mathbf{1}]$. Predictors then take the form $E = \bar{X}M(\bar{Y})^\top + \Delta$, with nuclear norm regularisation imposed on E and L^1 (or nuclear norm) regularisation on M, and Frobenius norm regularization imposed on Δ, with the constraint that $P_\Omega(E) = R_\Omega$ where R_Ω denotes the observed entries. Thus, similarly to [17], the authors allow hyperparameters to decide how much to trust the side information by employing nuclear norm regularisation on *both* the M inside the IMC term *and* nuclear norm regularisation on the whole predictor. However, the strategy remains different from [17], which applies nuclear norm regularisation to a residual term instead. The authors prove generalisation bounds in the uniform sampling regime which scale as the bounds in [61] in the case where the corresponding realizability assumptions are satisfied and scale as the state-of-the-art bounds in [34], [17] in the case where the side information is not good and the model must rely purely on the generic nuclear norm regularisation. A similarity between that model and ours is the augmentation of the matrices X and Y by a vector of ones (which also modifies the corresponding realizability assumptions). As such the proposed predictors involve a combination of lower order terms (which depend on one of the side information terms but not the others: $xv^\top + yw^\top$ where x, y are the side information vectors and u, w are trainable parameters) and higher order terms ($xMy^\top$ where M is trainable). Thus if the side information is categorical, the model predictors will be similar to the predictors in our model OMIC+. However, the extra interpretability and optimization benefits we reap from the cross-term orthogonality constraints are absent in that work. Further, the optimization problem remains different and the generalization bounds do not explicitly exploit community structure. Contrary to our models BOMIC and BOMIC+ (but similarly to our model OMIC+, to which it is more closely related), there are also actually no user or item biases in that work (unless the matrices X and Y include identity matrices as submatrices), although there is a matrix-wise bias.

In [21], the authors solve an explicit rank minimiz-

ation problem under linear constraints on the matrix (this problem is now commonly referred to as 'Matrix Regression' (see also [22])). In that formulation, instead of observing entries of the matrix, iid measurements of the form $v^\top M w$ are made where w, v are Gaussian vectors. As such, the problem is related to both inductive matrix completion (because the vs and ws play a similar role as side information vectors in IMC) and classic matrix completion (because different observations never correspond to the same v or w and the dimensions of v and w are the same as the ground truth matrix). In this context and with an explicit low-rank assumption the authors show a sample complexity bound of $O(r^3 n \log(n))$, which is in line with bounds for classic matrix completion under a uniform sampling assumption. A similar problem with nuclear norm regularisation was studied in [62], together with its link to IMC. As explained in our Theory section, none of the bounds in either of those works give tight bounds for the community setting. Later in [67], an ingenious new algorithm was proposed, reducing the sample complexity to $O((m+n)r)$.

In [23], the authors propose a very general optimization framework that encompasses both the matrix regression problem mentioned above and low rank matrix completion, as well as one-bit matrix completion.

B. Matrix completion with graph side information

In [68], [69], the authors propose a model based on user biases combined with neighborhood-based models. In [70], [71], [72], [73], [74], [75], the authors construct various low-rank matrix completion problems with regularizers inspired from the graph side information (typically, the feature vectors of adjacent nodes in the graph are regularised to be close to each other). In [76], the authors ingeniously combine this idea with user biases. Notably, in [77], some generalization guarantees were provided for such regularization strategies.

C. Community discovery

In [78], the authors propose a probabilistic model to solve binary matrix completion with graph side information based on the assumption that the users form communities: the assumption is that each user's rating is a noisy measurement of the preference of the cluster. The clusters are recovered from the graph information via the Stochastic Block Model (SBM), and the cluster preferences are then recovered from the observed data. Generalisation bounds and an asymptotic analysis are provided for this model. In [79] a similar model with further twists such as the existence of atypical users and items is introduced, and a thorough and impressive complexity and generalization analysis is performed.

In [80], the authors consider the problem of simultaneously clustering users and items in an efficient way based on *a single fully observed matrix*. In [81], another similar matrix factorization approach was provided to cluster users based on side information.

Those works rely heavily on the more general problem of community discovery, which is concerned with recovering "groups" or clusters of users given some side information such as a graph of interaction [82], [83], [84], [85], [86], [87], [88], [89]. These models typically rely on the assumption that the graph we observe is generated under the Stochastic Block Model, i.e., each edge in the graph is present or absent with a given probability that depends (only) on the cluster assignments of the two relevant nodes. We refer the reader to [82] for a survey.

We note that the above approaches are crucially different from ours in that *they do not allow for non random user-specific behaviour within each cluster*: in all of these works (except [79]), the behaviour of users/items is an independent noisy measurement of cluster behaviour, whilst in our model, users exhibit their own behaviour on top of the cluster-specific behaviour. In particular, in the works above, there is no difference between predicting the matrix and predicting the clusters, whilst in our setting, we usually assume the clusters are given and recover the matrix from them. In that respect, our setting is more similar to the regularization-based techniques [70], [71], [72], [73], [74], [75], but our method is different. The paper [79] is to the best of our knowledge the only work that incorporates item specific behavior in a community detection context. They do so in a discrete fashion with the concept of "atypical" movies and users, whilst our approach is a continuous one, which includes the possibility of representing any matrix (at a regularization cost).

We note in passing that a different approach to extracting community information from graphs is offered by graph neural networks [90], [91], [92], [93], [94].

D. On some variants of the matrix completion problem

In [95], the authors study a different problem, closely related to matrix completion, which assumes the existence of an exact dictionary, then introduce

a much weaker condition than uniform sampling and incoherence, and show that typical optimization algorithms can recover the matrix. Before that, in [96], similar results were shown under different assumptions. In [22], the author gives recovery guarantees for the more general problem of linear matrix equations. In [97], the authors study a multi-view model for image data where a common low-rank representation of several different views of the data points is constructed, to be later fed to a matrix completion algorithm.

Recently, [98] and [99] studied a form of transfer learning problem, where the same users rank different media such as movies, music, series etc. In [100], the authors perform *non negative matrix completion* with multiple sources of side information through a regularization-based approach different form the IMC setting.

Very recently, progress was made in the direction of establishing theoretical guarantees for matrix completion with "side information" where features are extracted jointly through a shallow network [63]. This opens up an interesting new avenue of research for matrix completion methods such as ours, which could perhaps be further combined with these techniques in the future. We note that this preliminary work [63] only deals with the case of fixed rank constraints.

REFERENCES

[1] Z. Kang, C. Peng, and Q. Cheng, "Top-n recommender system via matrix completion," 2016.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[3] C.-J. Hsieh, K.-Y. Chiang, and I. S. Dhillon, "Low rank modeling of signed networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12, New York, NY, USA, 2012, p. 507–515.

[4] F. Jirasek, R. A. S. Alves, J. Damay, R. A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft, and H. Hasse, "Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion," *The Journal of Physical Chemistry Letters*, vol. 11, no. 3, pp. 981–985, 2020.

[5] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, p. 2287–2322, Aug. 2010.

[6] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theor.*, vol. 56, no. 5, p. 2053–2080, May 2010.

[7] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, p. 717, 2009.

[8] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 952–960.

[9] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Statist.*, vol. 39, no. 5, pp. 2302–2329, 10 2011.

[10] C. C. Aggarwal, *Recommender Systems: The Textbook*, 1st ed. Springer Publishing Company, Incorporated, 2016.

[11] R. M. Bell, Y. Koren, and C. Volinsky, "The bellkor solution to the netflix prize," 2007.

[12] X. Zhang, S. Du, and Q. Gu, "Fast and sample efficient inductive matrix completion via multi-phase procrustes flow," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 5756–5765.

[13] M. Herbster, S. Pasteris, and L. Tse, "Online matrix completion with side information," *CoRR*, vol. abs/1906.07255, 2019.

[14] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," in *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2011, pp. 437–452.

[15] T. Chen, W. Zhang, Q. lu, K. Chen, Z. Zheng, and Y. Yu, "Svdfeature: A toolkit for feature-based collaborative filtering," *The Journal of Machine Learning Research*, 2012.

[16] P. Jain and I. S. Dhillon, "Provable inductive matrix completion," *CoRR*, vol. abs/1306.0626, 2013.

[17] K.-Y. Chiang, I. S. Dhillon, and C.-J. Hsieh, "Using side information to reliably learn low-rank matrices from missing and corrupted observations," *J. Mach. Learn. Res.*, 2018.

[18] J. Gaillard and J.-M. Renders, "Time-sensitive collaborative filtering through adaptive matrix completion," in *Advances in Information Retrieval*, A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, Eds. Cham: Springer International Publishing, 2015, pp. 327–332.

[19] A. Tasissa and R. Lai, "Low-rank matrix completion in a general non-orthogonal basis," 2018.

Notation	Meaning
$\mathcal{D}$	Sampling distribution over entries
N	Sample size
$\{(i_1, j_1), (i_2, j_2), \ldots, (i_N, j_N)\}$	Set of observed entries
P_Ω	Projection on the set of observed entries
$(\{(i_\alpha, j_\alpha) : \alpha \leqslant N\} = \Omega)$	
$P_{\Omega^\top}$	Projection on complement of set of observed entries
$X^{(k)} \in \mathbb{R}^{m \times d_1^{(k)}}$ $(k \leqslant K)$	k'th left side information matrix
$Y^{(l)} \in \mathbb{R}^{n \times d_2^{(l)}}$ $(l \leqslant L)$	l'th right side information matrix
$d_1^{(k)}$	Dimension of k'th left side information
$d_2^{(l)}$	Dimension of l'th right side information
$M^{(k,l)} \in \mathbb{R}^{d_1^{(k)} \times d_2^{(l)}}$	Matrices to be trained in our model $f_{i,j} = \sum_{k,l=1}^{K,L} X^{(k)} M^{(k,l)} (Y^{(l)})^\top$
$\mathcal{M}_{k,l}$	Upper bound on $\|M^{(k,l)}\|_*$
R_Ω	Matrix of observed entries
$\|\cdot\|_*$	Nuclear norm
$\|\cdot\|_\sigma$	Spectral Norm
$f : \{1, \ldots, m\} \times \{1, \ldots, n\} \to \mathbb{R}$	Prediction function
f_M	Prediction function $f_{i,j} = \sum_{k,l=1}^{K,L} X^{(k)} M^{(k,l)} (Y^{(l)})^\top$ $(M = (M_1, \ldots, M_{K,L}))$
$\mathcal{R}(f)$	Expected risk $\mathbb{E}_{(i,j)\sim\mathcal{D}}(\ell(f(i,j), R_{i,j} + \delta_{i,j}))$
$\hat{\mathcal{R}}(f)$	Empirical risk $\frac{\sum_{(i,j)\in\Omega} \ell(f(i,j), R_{i,j}+\delta_{i,j}))}{\#(\Omega)} = \sum_{\alpha=1}^{N} \frac{\ell(f_{i_\alpha, j_\alpha}, R_{i_\alpha, j_\alpha} + \delta_{i,j})}{N}$
$\hat{f}$	Solution to optimisation problem (S.65)
ℓ	Loss function
B	Upper bound on value of the loss function ℓ
L_ℓ	Upper bound on the Lipschitz constant of the loss function ℓ
$R \in \mathbb{R}^{m \times n}$	Ground truth matrix (can be observed with noise)
$\delta_{i,j}, \delta_\alpha$ (if $(i_\alpha, j_\alpha) = (i, j)$)	Sample from noise distribution of entry $R_{i,j}$, (follows distribution $\Delta_{i,j}$ independently at each draw)
$R^{(k,l)} = (X^{(k)})^\top R Y^{(l)}$	Core matrix in (k, l) component of ground truth matrix
$X^{(k)} R^{(k,l)} (Y^{(l)})^\top$	(k, l) component of ground truth matrix
$C_{k,l}$	Upper bound on entries of $X^{(k)} R^{(k,l)} (Y^{(l)})^\top$ (or R)
$r_{k,l}$	Upper bound on rank of $X^{(k)} R^{(k,l)} (Y^{(l)})^\top$ (equivalently $R^{(k,l)}$)
C	Constant upper bound on the entries of the ground truth matrix R
$\mathbf{x}_i^k$	ith row of the matrix $X^{(k)})$
$\mathbf{y}_i^k$	ith row of the matrix $Y^{(k)})$
$\mathcal{X}_k$	$\max_{i=1}^m \|\mathbf{x}_i^k\|_2 = \max_{i=1}^m \|X_{i,\bullet}^{(k)}\|_2$
$\mathcal{Y}_l$	$\max_{i=1}^n \|\mathbf{y}_i^l\|_2 = \max_{i=1}^m \|Y_{i,\bullet}^{(l)}\|_2$
$\mathcal{E}$	Optimal risk $\mathbb{E}_{(i,j)\sim\mathcal{D}}(\ell(R_{i,j} + \delta_{i,j}, R_{i,j}))$
$\tilde{X}^{(k)}$	Matrix whose columns come from completing $X^{(k)}$ into an orthonormal basis
$\tilde{Y}^{(l)}$	Matrix whose columns come from completing $Y^{(l)}$ into an orthonormal basis
S_λ	Singular value thresholding operator from earlier work
$\Lambda = (\lambda_{k,l})_{k\leqslant K; l\leqslant L}$	A set of hyperparameters
S_Λ	Generalised SVTO defined in (9)

Table II: Table of notations for quick reference

[20] J. Lu, G. Liang, J. Sun, and J. Bi, "A sparse interactive model for matrix completion with side information," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/093b60fd0557804c8ba0cbf1453da22f-Paper.pdf

[21] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/32bb90e8976aab5298d5da10fe66f21d-Paper.pdf

[22] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[23] L. Wang, X. Zhang, and Q. Gu, "A Unified Computational and Statistical Framework for Nonconvex Low-rank Matrix Estimation," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. Fort Lauderdale, FL, USA: PMLR, 20–22 Apr 2017, pp. 981–990. [Online]. Available: http://proceedings.mlr.press/v54/wang17b.html

[24] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, 1995.

[25] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank svd via fast alternating least squares," *Journal of Machine Learning Research*,

vol. 16, no. 104, pp. 3367–3402, 2015. [Online]. Available: http://jmlr.org/papers/v16/hastie15a.html

[26] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *proceedings of the 25th international conference on world wide web*, 2016, pp. 507–517.

[27] N. Yang, "Douban movie and NetEase music datasets and model code," 2019. [Online]. Available: https://doi.org/10.7910/DVN/JGH1HA

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] M. Wan, R. Misra, N. Nakashole, and J. McAuley, "Fine-grained spoiler detection from large-scale review corpora," *arXiv preprint arXiv:1905.13416*, 2019.

[30] M. Fazel, "Matrix rank minimization with applications," *PhD Thesis*, 2002.

[31] "Learning with matrix factorizations."

[32] G. Watson, "Characterization of the subdifferential of some matrix norms," *Linear Algebra and its Applications*, vol. 170, pp. 33 – 45, 1992.

[33] Y. Nesterov, "Gradient methods for minimizing composite objective function," Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers 2007076, 2007.

[34] O. Shamir and S. Shalev-Shwartz, "Collaborative filtering with the trace norm: Learning, bounding, and transducing," in *Proceedings of the 24th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 19. PMLR, 2011, pp. 661–678.

[35] R. Latała, "Some estimates of norms of random matrices," *Proceedings of the American Mathematical Society*, vol. 133, no. 5, pp. 1273–1282, 2005.

[36] P. Giménez-Febrer, A. Pagès-Zamora, and G. B. Giannakis, "Generalization error bounds for kernel matrix completion and extrapolation," *IEEE Signal Processing Letters*, vol. 27, pp. 326–330, 2020.

[37] R. Foygel, O. Shamir, N. Srebro, and R. R. Salakhutdinov, "Learning with the weighted trace-norm under arbitrary sampling distributions," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 2133–2141.

[38] C. Scott, "Rademacher complexity," *Lecture Notes*, vol. Statistical Learning Theory, 2014.

[39] R. Meir and T. , "Generalization error bounds for bayesian mixture algorithms," *J. Mach. Learn. Res.*, vol. 4, no. null, p. 839–860, Dec. 2003.

[40] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," in *Computational Learning Theory*, D. Helmbold and B. Williamson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 224–240.

[41] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, vol. 12, no. null, p. 3413–3430, Dec. 2011.

[42] R. Vershynin, "High-dimensional probability," 2019. [Online]. Available: https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf

[43] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, p. 389–434, Aug. 2012.

[44] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, p. 150401, Oct 2010.

[45] H. Zhang, L. Cheng, and W. Zhu, "A lower bound guaranteeing exact matrix completion via singular value thresholding algorithm," *Applied and Computational Harmonic Analysis*, vol. 31, no. 3, pp. 454 – 459, 2011.

[46] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," *Proceedings of the Annual ACM Symposium on Theory of Computing*, 12 2012.

[47] C.-J. Hsieh and P. Olsen, "Nuclear norm minimization via active subspace selection," *31st Internat ional Conference on Machine Learning, ICML 2014*, vol. 1, pp. 871–882, 01 2014.

[48] Q. Yao and J. T. Kwok, "Accelerated and inexact soft-impute for large-scale matrix and tensor completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 9, pp. 1665–1679, Sep. 2019.

[49] Y. Yang, M. Pesavento, S. Chatzinotas, and B. Ottersten, "Parallel and hybrid soft-thresholding algorithms with line search for sparse nonlinear regression," 09 2018, pp. 1587–1591.

[50] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank svd via fast alternating least squares," *Journal of Machine Learning Research*, vol. 16, no. 104, pp. 3367–3402, 2015.

[51] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[52] C.-J. Hsieh and P. Olsen, "Nuclear norm minimization via active subspace selection," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 1. Bejing, China: PMLR, 22–24 Jun 2014, pp. 575–583.

[53] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, no. 1, p. 1665–1697, May 2012.

[54] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Completing any low-rank matrix, provably," *Journal of Machine Learning Research*, vol. 16, no. 94, pp. 2999–3034, 2015.

[55] Y. Wan, J. Yi, and L. Zhang, "Matrix completion from non-uniformly sampled entries," 06 2018.

[56] N. Srebro and R. R. Salakhutdinov, "Collaborative filtering in a non-uniform world: Learning with the weighted trace norm," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 2056–2064.

[57] O. Shamir and S. Shalev-Shwartz, "Matrix completion with the trace norm: Learning, bounding, and transducing," *Journal of Machine Learning Research*, vol. 15, pp. 3401–3423, 2014.

[58] R. Li, Y. Dong, Q. Kuang, Y. Wu, Y. Li, M. Zhu, and M. Li, "Inductive matrix completion for predicting adverse drug reactions (adrs) integrating drug–target interactions," *Chemometrics and Intelligent Laboratory Systems*, vol. 144, pp. 71 – 79, 2015.

[59] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene and disease associations," *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 06 2014.

[60] D. Shin, S. Cetintas, K.-C. Lee, and I. S. Dhillon, "Tumblr blog recommendation with boosted inductive matrix completion," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 203–212.

[61] 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*, ser. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 2301–2309.

[62] K. Zhong, P. Jain, and I. S. Dhillon, "Efficient matrix sensing using rank-1 gaussian measurements," in *Algorithmic Learning Theory*, K. Chaudhuri, C. GENTILE, and S. Zilles, Eds. Cham: Springer International Publishing, 2015, pp. 3–18.

[63] K. Zhong, P. Song, Zhao and, and I. S. Dhillon, "Provable non-linear inductive matrix completion," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 11 439–11 449.

[64] S. Xue, W. Qiu, F. Liu, and X. Jin, "Low-rank tensor completion by truncated nuclear norm regularization," *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2600–2605, 2017.

[65] N. Ghadermarzy, Y. Plan, and A. Yilmaz, "Near-optimal sample complexity for convex tensor completion," *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 577–619, 11 2018.

[66] M. Nimishakavi, P. K. Jawanpuria, and B. Mishra, "A dual framework for low-rank tensor completion," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 5484–5495.

[67] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 964–973. [Online]. Available: http://proceedings.mlr.press/v48/tu16.html

[68] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," 08 2008, pp. 426–434.

[69] ——, "Factor in the neighbors: Scalable and accurate collaborative filtering," *ACM Trans. Knowl. Discov. Data*, vol. 4, no. 1, Jan. 2010.

[70] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst, "Matrix Completion on Graphs," *arXiv e-prints*, p. arXiv:1408.1717, Aug. 2014.

[71] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King, "Recommender systems with social regularization," 01 2011, pp. 287–296.

[72] 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*, ser. RecSys '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 135–142.

[73] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: Social recommendation using probabilistic matrix factorization," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 931–940.

[74] W.-J. Li and D.-Y. Yeung, "Relation regularized matrix factorization." 01 2009, pp. 1126–1131.

[75] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, Aug 2011.

[76] G. Guo, J. Zhang, and N. Yorke-Smith, "Trustsvd: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings," 2015.

[77] N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2107–2115.

[78] K. Ahn, K. Lee, H. Cha, and C. Suh, "Binary rating estimation with graph side information," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 4272–4283.

[79] Qiaosheng, Zhang, V. Y. F. Tan, and C. Suh, "Community Detection and Matrix Completion with Two-Sided Graph Side-Information," *arXiv e-prints*, p. arXiv:1912.04099, Dec. 2019.

[80] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 126–135.

[81] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 2013, pp. 252–260.

[82] E. Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018.

[83] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, Jan 2016.

[84] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, Oct 2015, pp. 670–688.

[85] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109 – 137, 1983.

[86] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," 05 2018.

[87] H. Saad and A. Nosratinia, "Exact recovery in community detection with continuous-valued side information," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 332–336, Feb 2019.

[88] J. Yoon, K. Lee, and C. Suh, "On the joint recovery of community structure and community features," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2018, pp. 686–694.

[89] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *2013 IEEE 13th International Conference on Data Mining*, Dec 2013, pp. 1151–1156.

[90] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 3844–3852.

[91] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *CoRR*, vol. abs/1901.00596, 2019.

[92] M. Henaff, J. Bruna, and Y. Lecun, "Deep convolutional networks on graph-structured data," 06 2015.

[93] T. N. Kipf and M. Welling.

[94] A. Micheli, "Neural network for graphs: A contextual constructive approach," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 498–511, March 2009.

[95] G. Liu, Q. Liu, and X. Yuan, "A new theory for matrix completion," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 785–794.

[96] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization,"

Proceedings of the National Academy of Sciences, vol. 100, no. 5, pp. 2197–2202, 2003.

[97] M. Liu, Y. Luo, Dacheng, C. Xu, and Y. Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," 2015.

[98] M. Z. Alaya and O. Klopp, "Collective matrix completion," *Journal of Machine Learning Research*, vol. 20, no. 148, pp. 1–43, 2019.

[99] S. Gunasekar, M. Yamada, D. Yin, and Y. Chang, "Consistent collective matrix completion under joint low rank structure," 12 2014.

[100] H. Chen and J. Li, "Learning multiple similarities of users and items in recommender systems," in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 811–816.