# Generalization Bounds for Semi-supervised Matrix Completion with Distributional Side Information

**Antoine Ledent[1], Soo Mun Chong[2], Nong Minh Hieu[1]**

[1] Singapore Management University (SMU),
[2] Binance
aledent@smu.edu.sg, mc.s@binance.com, mh.nong.2024@phdcs.smu.edu.sg

## Abstract

We study a matrix completion problem where both the ground truth $R$ matrix and the unknown sampling distribution $P$ over observed entries are low-rank matrices, and *share a common subspace*. We assume that a large amount $M$ of *unlabeled* data drawn from the sampling distribution $P$ is available, together with a small amount $N$ of labeled data drawn from the same distribution and noisy estimates of the corresponding ground truth entries. This setting is inspired by recommender systems scenarios where the unlabeled data corresponds to 'implicit feedback' (consisting in interactions such as purchase, click, etc. ) and the labeled data corresponds to the 'explicit feedback', consisting of interactions where the user has given an explicit rating to the item. Leveraging powerful results from the theory of low-rank subspace recovery, together with classic generalization bounds for matrix completion models, we show error bounds consisting of a sum of two error terms corresponding to sample complexities of $nd$ and $dr$ respectively (ignoring log factors), where $d$ is the rank of $P$ and $r$ is the rank of $M$. In synthetic experiments, we confirm that the true generalization error naturally splits into independent error terms corresponding to the estimations of $P$ and and the ground truth matrix G respectively. In real-life experiments on Douban and MovieLens with most explicit ratings removed, we demonstrate that the method can outperform baselines relying only on the explicit ratings, demonstrating that our assumptions provide a valid toy theoretical setting to study the interaction between explicit and implicit feedbacks in recommender systems.

## Introduction

*Matrix completion* (MC) refers to a broad class of statistical problems where one wishes to recover the entries of an unknown ground truth matrix $\mathrm{G} \in \mathbb{R}^{m\times n}$ based on a set of $N \ll mn$ potentially noisy observations. In short, it is a supervised learning problem where the independent variable is a (row, column) pair where both components can only take a finite set of values $[m]$ or $[n]$. Despite its apparent simplicity, this problem is not only of high practical relevance (in recommender systems (Koren, Bell, and Volinsky 2009; Zhang and Chen 2019), chemical and thermal engineering (Jirasek et al. 2020; Hänsch et al. 2025) and drug discovery (Li et al. 2015), etc.), but also surprisingly challenging and nuanced in its statistical properties. The earliest and most well-known works in the field focused on the *exact recovery* problem: the celebrated works of Candès and Tao (2010); Candès and Recht (2009) showed that minimizing the nuclear norm of the candidate matrix $Z$ subject to $Z_{i,j} = G_{i,j}$ whenever the entry $(i, j)$ is in the set $\Omega$ of observed entries provably recovers the exact ground truth matrix as long as the number of observations taken uniformly at random is larger than $\widetilde{O}(nr)$ where $n$ is the size of the matrix and $r$ is the ground truth rank. However, many different regimes can be considered. For instance, some recent work refine the bounds of Candès and Tao (2010); Candès and Plan (2010) to incorporate a very fine joint dependence on the subgaussianity constant of the noise and the size and rank of the matrix (Chen et al. 2020a), whilst other study excess risk bounds in nonuniform sampling regimes (Foygel et al. 2011; Shamir and Shalev-Shwartz 2011, 2014) or out-of-distribution generalization for certain missingess patterns (Ma and Chen 2019).

A notable set of works consider the so-called 'Inductive Matrix Completion' (IMC) setting (Xu, Jin, and Zhou 2013; Chiang, Dhillon, and Hsieh 2018; Ledent et al. 2021), where one assumes that the learner has access to side information matrices $X \in \mathbb{R}^{m\times d}$ and $Y \in \mathbb{R}^{n\times d}$ with the property that the ground truth matrix G can be represented as $\mathrm{G} = XMY^\top$ for some unknown matrix $M$. The role of the matrices $X, Y$ is to represent some external knowledge about each value of the underlying discrete variable. For instance, in recommender systems, the rows of $X$ and $Y$ would correspond to feature vectors describing the users and items respectively. Similarly, in a drug interaction prediction context, the rows of $X, Y$ would consist of feature vectors containing information about the chemical composition of each drug. Thus, the IMC framework aims to take matrix completion closer to real applications by incorporating available side information, and continues to attract interest in recent years, with many different variants being proposed. For instance, Jalan et al. (2025) study a challenging and biologically-inspired setting where entire rows or columns are missing and study both passive and active sampling regimes, where the user can select whole rows and columns to observe. However, a weakness of existing IMC approaches is the assumption that the side information matrices $X, Y$ are known *a priori*, and the column (resp. row) space of the ground truth matrix is exactly contained in the column space of $X$ (resp. $Y$). In practice, such side information needs to be *estimated* from other forms of training data (social media user graph, molecular drug structure, etc.).

To the best of our knowledge, all theoretical results on matrix completion to date assume that the available samples are all labeled: every observed entry $(i, j) \in [m] \times [n]$ is accompanied by a (possibly noisy) estimate of the ground truth entry $\mathrm{G}_{i,j}$. However, in many real-life scenarios, it is common for a much larger amount of unlabeled samples to be naturally available. In a recommender system, one may observe a large set of implicit interactions $\Omega$, where the presence of the pair $(i, j)$ in $\Omega$ indicates that user $i$ has watched/consumed the item $j$. This type of interaction, often referred to as 'implicit feedback' is distinct from 'explicit' interactions where a *rating* (typically on a scale between 1 and 5 stars) is given by user $i$ to item $j$: such explicit interactions might be much more scarce. Thus, there is a need for a more inclusive learning setting in matrix completion to incorporate semi-supervised settings. Our contributions are as follows:

- We propose a semi-supervised learning paradigm for matrix completion: we assume that the sampling distribution $P \in [0, 1]^{m \times n}$ over entries shares a low-rank subspace with the ground truth matrix G, and that that a large number $M$ of *unlabeled* interactions is provided, together with a much smaller number $N$ of *labeled* interactions.
- Under incoherence, *uniform marginals* and bounded sampling probability assumptions, we leverage powerful results from matrix pertrubation theory to prove generalization bounds which scale like $\widetilde{O}\left(\sqrt{\frac{nd}{M}}\right) + \widetilde{O}\left(\sqrt{\frac{dr}{N}}\right)$, showing that the estimation errors associated to the subspace recovery (from unlabeled samples) and matrix recovery (from labeled samples) can be *disentangled*.
- In synthetic data experiments, we demonstrate that the generalization error can indeed be closely approximated by a sum of terms corresponding to each form of error.
- In RecSys datasets, we demonstrate that a large number of unlabeled samples can substantially improve the performance of explicit feedback prediction methods. This aligns with the conclusions implicitly in recent works (Zhang and Chen 2019; Xia et al. 2022; Ledent et al. 2025), lending legitimacy to our learning paradigm.

## Related Works

**Matrix completion in the i.i.d. setting (without side information):** Our results are most closely related to the so-called 'approximate recovery' branch of the literature on matrix completion, which assumes that the observations are drawn i.i.d. from a distribution over entries, and that the performance measure is the (in-distribution) excess risk. In this setting, a sample complexity of $\widetilde{O}\left(nr\right)$ was achieved for empirical risk minimization with a nuclear norm constraint in Foygel et al. (2011) under a uniform marginals assumption, whilst a sample complexity of $\widetilde{O}\left(n^{\frac{3}{2}}\sqrt{r}\right)$ was achieved under arbitrary sampling regimes in Shamir and Shalev-Shwartz (2011, 2014). Before that, similar settings were studied in Neyshabur, Tomioka, and Srebro (2015), which includes the case of explicit rank restriction in a classification setting. When imposing Schatten $p$ quasi norm constraints, recent results show a sample complexity of $\widetilde{O}\left(n^{1+\frac{p}{2}}r^{1-\frac{p}{2}}\right)$ and $\widetilde{O}\left(nr\right)$ in the arbitrary sampling and uniform marginals settings respectively (Ledent and Alves 2024). Very tight bounds in terms of the dependency on the variance of the noise were proved in Chen et al. (2020b) and Chen et al. (2021) for nuclear norm and exactly low-rank cases respectively. However, all of those results depend at least linearly in the size $n$ of the matrix and require all of the samples to be labeled, making the results ineffective in the semi-supervised learning setting we aim to study.

Another closely related branch of the literature studies '**Inductive Matrix Completion**' (with side information), which also studies the in-distribution i.i.d. setting but assumes the presence of known side information matrices $X \in \mathbb{R}^{m \times d}$ and $Y \in \mathbb{R}^{n \times d}$ such that the ground truth matrix is known to be representable in the form $\mathrm{G} = XMY^\top$ for some $M \in \mathbb{R}^{d \times d}$. Similarly to our work, much of the literature on this problem imposes nuclear norm constraints on the core matrix $M$. The problem was initially proposed in Xu, Jin, and Zhou (2013), which studied the exact recovery setting under the uniform sampling regime with nuclear norm minimization, and was later studied under the i.i.d. setting for the first time in Chiang, Dhillon, and Hsieh (2018); Chiang, Hsieh, and Dhillon (2015), where sample complexity bounds of $\widetilde{O}(d^2r)$ are provided. The bounds were improved to $\widetilde{O}\left(d^{\frac{3}{2}}\sqrt{r}\right)$ and $\widetilde{O}\left(dr\right)$ in the arbitrary sampling and uniform marginals cases respecitvely in Ledent et al. (2021), and bounds with a finer dependence on the variance of the noise were proved in Ledent et al. (2023). Whilst those rates are similar to our bounds on the estimation error arising from the labeled data, they do not involve subspace estimation: in fact, we rely on those results as tools to establish our own bounds, but the proofs are completely different. The main difficulty in our work is bounding the estimation of the subspaces $X, Y$ using the unlabeled data and controling the propagation of that error through the downstream inductive matrix completion problem. In particular, our results require more stringent assumptions on the sampling distribution as a result of the increased technical difficulty.

Many of the tools we rely on for the subspace estimation problem come from the matrix **perturbation theory** and its applications to datascience (Chen et al. 2021). Indeed, this book also provides a remarkable array of strong bounds for matrix completion. However, the bounds apply to a different regime: the entries are sampled uniformly at random with Bernouilli sampling (without replacement and without side information) and the emphasis of the results lies in a tighter dependence on the subgaussianity constant of the noise, as well as providing entry-wise estimates. Whilst the book Chen et al. (2021) deals mostly with models involving explicit rank restriction, we note that similar results are also known for nuclear norm minimization (Chen et al. 2020a).

Another nonuniform sampling regime which has gained a lot of attention in recent years is the so called 'Missing Not at Random' (**MNAR**) matrix completion problem (Ma and Chen 2019; Choi and Yuan 2024; Jalan et al. 2025), where the entries are sampled with independent Bernouilli masks whose associated probabilities are obtained by applying a sigmoid function to an unknown low-rank matrix. Thus, this corresponds to an alternative form of nonuniform sampling compared to our i.i.d. setting, and this research direction also involves a low-rank constraint in the sampling distribution. However, there are many notable differences: first, the emphasis is on bounding the (uniform) Frobenius error, making this an *out-of-distribution* problem where the empirical risk is reweighted by inverse propensity scores to compensate for the nonuniformity of the sampling distribution. Second, the 'low-rank' condition on the sampling distribution is $\mathcal{P} = \sigma(\Gamma)$ where $\mathcal{P}$ is the matrix of Bernouillli probabilities (the 'propensity scores') and a nuclear norm constraint is imposed on $\Gamma$. Whilst this is somewhat comparable to a low-rank condition, it is important to note that the choice $\Gamma \simeq 0 \in \mathbb{R}^{m\times n}$ (which is both low-rank and low nuclear norm) leads to a uniform masking probability of 0.5, which is a *dense* observation regime at odds with the sparse observations regime studied in classic matrix completion settings. In addition to the difference in performance measure, this further complicates any comparison betweeen the results in Ma and Chen (2019) and the approximate recovery literature at the common regimes where very few entries are observed. Third, to compensate for the nonuniformity of the sampling distribution despite the uniform performance measure, an inverse multiplicative factor in the minimum sampling probability for any entry is present in Ma and Chen (2019), which further distinguishes its results from our own, which apply even if many entries have zero sampling probability. Whilst the problem setting in Jalan et al. (2025); McGrath et al. (2024) also involves matrices with shared subspaces, they are assumed to come from distinct sources and both matrices are partially observed, instead of corresponding to low-rank sampling distributions. Similarly, the idea of a low-rank distribution is studied in Anandkumar et al. (2014); Vandermeulen and Ledent (2021); Amiridi, Kargas, and Sidiropoulos (2022) and used in a recommender systems setting in Poernomo et al. (2025), but none of these works involve explicit feedback or a shared subspace. Lastly, there is a rich literature on semi-supervised learning in a broader machine learning context (Balcan and Blum 2005; Blum and Mitchell 1998; Bekker and Davis 2020). However, the key techniques such as Contrastive Learning (Lei et al. 2023; Hieu et al. 2025; Hieu and Ledent 2025; Alves and Ledent 2024; Ghanooni et al. 2024, 2025), typically apply to classification problems rather than regression setting, and such methods usually do not apply to discrete input spaces, such as matrix completion, which comes with its own challenges.

# Main Results

## Learning Setting and Assumptions

**Matrix Completion in the i.i.d. setting.** We consider a noisy matrix completion problem in the i.i.d. (regression) setting. This setting has been studied in the following papers, however, we reintroduce it with our notation (and in slightly greater formality and generality) for the sake of convenience (Foygel et al. 2011; Shamir and Shalev-Shwartz 2011, 2014; Ledent et al. 2021; Chiang, Dhillon, and Hsieh 2018; Chiang, Hsieh, and Dhillon 2015), as it differs from both the problem of exact matrix recovery (Candès and Tao 2010; Candès and Plan 2010; Chen et al. 2021) or 'missing not at random' matrix completion. The sampling procedure for each sample is i.i.d. according to the following distribution. A full labeled sample $(\xi, \widetilde{G})$ consists in:

- An *entry* $(i, j) = \xi \in [m] \times [n]$ sampled from a categorical probability distribution with Probability Mass Function (PMF) given by $P \in \mathbb{R}^{m\times n}$ over $[m] \times [n]$ (i.e. $\sum_{i\leq m, j\leq n} P_{i,j} = 1$), and
- A *label* $\widetilde{G} \in \mathbb{R}$ drawn from a conditional distribution $\mathcal{G}_{i,j}$ which depends on the entry $\xi = (i, j)$.

The entry is considered as the independent variable and the label is considered as the target variable in the sense of supervised learning. Predictors are functions $\mathrm{Z} \in \mathbb{R}^{m\times n}$ and their performance on a test sample $(\xi, \widetilde{G}, \hat{y})$ is measured via a loss function $\mathrm{l} : ([m] \times [n]) \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+ : (\xi, \widetilde{G}, \hat{y}) \mapsto \mathrm{l}(\xi, \widetilde{G}, \hat{y})$. The population level performance of the predictor Z is the population expected risk $\mathrm{l}(\mathrm{Z}) = \mathbb{E}_{\xi,\widetilde{G}}\left(\mathrm{l}(\xi, \widetilde{G}, \mathrm{Z}_\xi)\right)$. Often, we also consider its empirical analogue $\widehat{\mathbb{E}}_\xi\left(\mathrm{l}(\xi, \widetilde{G}, \mathrm{Z}_{\xi^o{}_\xi})\right) = \frac{1}{N}\sum_{o=1}^{N} \mathrm{l}(\xi_o, \widetilde{G}_o, \mathrm{Z}_{\xi^o})$, where $\xi_1, \ldots, \xi_N \in [m]\times[n]$ denotes a set of empirical samples. The Bayes predictor or *ground truth matrix* $\mathrm{G} \in \mathbb{R}^{m\times n}$ is defined by $\mathrm{G}_{i,j} \in \arg\min_{\hat{y}\in\mathbb{R}} \mathbb{E}_{\xi,\widetilde{G}}\, \mathrm{l}(\xi, \widetilde{G}, \hat{y})$. For instance, if l is the square loss (and in particular, doesn't depend on the entry $(i, j)$) and the observations $\widetilde{G}$ are equal to $R_\xi + \zeta$ for some fixed matrix $R \in \mathbb{R}^{m\times n}$ and some noise $\zeta$ satisfying $\mathbb{E}(\zeta) = 0$, then $R = \mathrm{G}$.

Going further, we consider a **semi-supervised setting** where the learner has access to

- $N$ *labeled* observations $\{(\xi_1^e, \tilde{G}_1), \ldots, (\xi_N^e, \tilde{G}_N)\} := \mathcal{S}_N$, drawn i.i.d. from the distribution above, and
- $M$ i.i.d. *unlabeled* samples $(i_o, j_o)$ (for $o \leq M$) drawn from the distribution $P$ alone, where we do not have access to the label $\widetilde{G}$.

We write $O_M = \frac{1}{M}\sum_{o=1}^{M} 1_{i_o,j_o}$ for the matrix of observed *unlabeled* samples (i.e., the empirical analogue of the PMF $P$), and make the following key assumptions.

**Assumption 1** (Boundedness and Lipschitzness of the Loss Function)**.** *The loss function* l *is uniformly bounded by* $\mathcal{B}$ *and for any value of* $(i, j)$*, is Lipschitz continuous with Lipschitz constant* $\ell$*: For all* $(i, j) \in [m] \times [n]$,

$$|\mathrm{l}((i, j), y, \hat{y}_1)| \leq \mathcal{B} \qquad \textit{and} \qquad (1)$$

$$|\mathrm{l}((i, j), y, \hat{y}_1) - \mathrm{l}((i, j), y, \hat{y}_2)| \leq \ell\, |\hat{y}_1 - \hat{y}_2| . \qquad (2)$$

**Assumption 2** (Existence of Shared Low-rank Subspace)**.** *We assume that the sampling distribution* $P$ *has low-rank* $d$*, i.e.* $P = U^*\Sigma^*[V^*]^\top$ *where* $U^* \in \mathbb{R}^{m\times d}$ *and* $V^* \in \mathbb{R}^{n\times d}$*. We assume that the ground truth matrix can be represented as* $\mathbb{R}^{m\times n} \ni \mathrm{G} = U^*\underline{M}^{*-}[V^*]^\top$ *for some matrix* $\underline{M}^{*-}$*. For convenience and to stick to the scaling used* *in other literature on inductive matrix completion, we introduce the notation* $X^* := \sqrt{\frac{m}{d}}U^*$, $Y^* := \sqrt{\frac{n}{d}}V^*$ *and* $\underline{M}^* := \sqrt{\frac{d^2}{mn}}\underline{M}^{*-}$. *Thus we certainly have*

$$\mathrm{G} = X^*\underline{M}^*[Y^*]^\top. \tag{3}$$

*We also assume that* $\|\underline{M}^*\|_* \leq \mathcal{M}$ *for some constant* $\mathcal{M}$.

This assumption forms the basis of our novel learning setting: to the best of our knowledge, it is the first attempt to formalize the existence of a relationship between implicit feedback (the sampling distribution over ratings) and explicit feedback (the matrix of latent rankings). We choose this assumption because this is the simplest way to assume such a relationship in a matrix completion setting.

**Assumption 3** (Incoherence of the Shared Low-rank Subspaces)**.** *We also write* $\boldsymbol{x}^* = \max_{i\leq m} \|[X^*]_{i,\cdot}\|$ *and* $\boldsymbol{y}^* = \max_{j\leq n} \|[Y^*]_{j,\cdot}\|$, *which can be interpreted as incoherence measures for the row and column subspaces of the ground truth* G *and are treated as* $O(1)$ *constants. We also define the following relevant quantity, which can be interpreted as an overall incoherence constant:*

$$\mathcal{P}^* = \max\left(\sqrt{\frac{n}{m}}\boldsymbol{x}^*, \sqrt{\frac{m}{n}}\boldsymbol{y}^*\right). \tag{4}$$

**Assumption 4** (Approximately Uniform Marginals)**.** *We assume that the marginals of* $P$ *are bounded uniformly as follows, for some constant* $\kappa_1$

$$p_i \leq \frac{\kappa_1}{m} \qquad \text{and} \qquad q_j \leq \frac{\kappa_1}{n}, \tag{5}$$

*where* $p_i := \sum_{j\leq n} P_{i,j}$ *and* $q_j := \sum_{i\leq n} P_{i,j}$ *are the marginals for the* $i$*th row and* $j$*th column respectively.*

This assumption is common in works on approximate recovery in matrix completion. For instance, Foygel et al. (2011); Shamir and Shalev-Shwartz (2011) and Ledent and Alves (2024) use it for some of their stronger results (typically, a non-trivial result is shown without this assumption, and a tighter bound is shown with this assumption). This assumption is weaker than the uniform sampling assumption which is typical in exact recovery results such as those of Candès and Recht (2009); Recht (2011); Chen et al. (2021, 2020a).

**Assumption 5** (Well-conditioning of PMF)**.** *We assume the sampling distribution* $P$ *is well conditioned. More precisely, we rely on the following conditioning number in our bounds:*

$$\kappa_* = \frac{\|P\|}{\Delta_*}, \tag{6}$$

*where* $\Delta_*$ *is the last singular value of* $P$.

**Assumption 6** (Well-conditioning of $X, Y$)**.** *We as assume the following bound on the spectral norm of the ground truth side information matrices* $X, Y$:

$$\|X^*\| \leq \boldsymbol{x}^*\sqrt{\kappa_2\frac{m}{d}} \qquad \text{and} \qquad \|Y^*\| \leq \boldsymbol{x}^*\sqrt{\kappa_2\frac{n}{d}}. \tag{7}$$

**Assumption 7** (Bound on Maximum Sampling Probability)**.** *We assume that the maximum possible entry of the sampling distribution* $P$ *is bounded by a constant* $\Gamma$. *Thus,* $\Gamma$ *is defined as* $\max_{i,j} P_{i,j}mn$ *so that for all* $i, j$:

$$P_{i,j} \leq \frac{\Gamma}{mn}. \tag{8}$$

Assumption 7, which requires a uniform upper bound on the sampling probability for any entry, is the most restrictive of ours. Still, Assumption 4 implies that Assumption 7 is always satisfied with at least a coarse estimate $\Gamma \leq \kappa_1[m+n]$. Relying on this still yields non-trivial results, but at the cost of an assumption of the form $N \geq \frac{m+n}{2}$: one needs at least $O(1)$ *labeled* interactions in each row/column. Whilst this is a significant restriction (because the sample complexity in terms of labeled examples isn't truly independent of the size of the side information $d$), the result is still of interest as this is an absolute threshold rather than a true contribution to the error bound. If $\Gamma$ is constant then the results hold without this caveat. Lastly, we note that, in contrast with the *lower* bound on the sampling probability in Ma and Chen (2019), even Assumption 7 with an *absolute* constant $\Gamma$ covers non-trivial cases and doesn't necessarily imply that the sampling distribution is approximately uniform. Indeed, suppose that the $n$ rows and columns are each divided into $k$ 'groups' or clusters. One can visualize this as a 'check board' where the 1st $n/k$ rows and cols belong to the 1st group, but the example is more general: group memberships can be unknown. Assumption 4 implies that both the ratings $G_{i,j} = \tilde{G}_{g(i),g(j)}$ and the probabilities $P_{i,j} = \tilde{P}_{g(i),g(j)}$ only depend on the (unknown) user and item groups. Assumption 7 only concerns $\tilde{P} \in \mathbb{R}^{k\times k}$ and is independent of $n$. For instance, if $\tilde{P} = \frac{1}{2}I/k + \frac{1}{2}U$ where $U$ is uniform, then $\Gamma = k/2 + 1/2 = [d+1]/2$ and the term $O(\sqrt{\frac{\Gamma nr}{MN}})$ $\leq O(\sqrt{\frac{d}{N}}\sqrt{\frac{nr}{M}})$ in our results below is benign. In this case, the assumptions hold with $\Gamma = O(d)$, $\kappa^*, \kappa_1, \kappa_2, \in O(1)$. In fact, user/item clustered settings are related to the stochastic block model (Abbe 2018; Abbe, Bandeira, and Hall 2016) and have been studied in various works both within (Ledent, Alves, and Kloft 2021; Alves et al. 2020; Alves 2024) and outside (Qiaosheng et al. 2019) matrix completion, achieving strong performance in both cases.

## Model

We assume that the learner is aware of the existence of a shared low-rank subspace and proceeds as follows:

- **Step 1:** first, the unlabeled data is used to estimate the subspaces via singular value decomposition: the matrix $\mathcal{H}$ is constructed, and a singular value decomposition of order $d$ is performed on it, yielding the SVD $\mathcal{H} = U\Sigma V^\top$. Then, the side information matrices $X, Y$ are constructed as $X = \sqrt{\frac{m}{d}}U$ and $Y = \sqrt{\frac{n}{d}}V$.
- **Step 2:** next, after fixing an upper bound constraint $\mathcal{M}$ for the nuclear norm of $\underline{M}$, and the ground truth matrix is estimated via empirical risk minimization using the clas-

sic Inductive Matrix Completion (IMC) algorithm:

$$\underline{M} = \underset{\|\underline{M}\|_* \le \mathcal{M}}{\arg\min}\ \mathrm{l}(X\underline{M}Y^\top) \tag{9}$$

$$= \underset{\|\underline{M}\|_* \le \mathcal{M}}{\arg\min}\ \frac{1}{N}\sum_{o=1}^{N} \mathrm{l}\left((X\underline{M}Y^\top)_{\xi_o}, \widetilde{G}_o\right).$$

We will also use the notation $r$ for the quantity $\frac{\mathcal{M}^2}{d^2}$, which scales as the rank of the matrix $\underline{M}$ (and therefore, G): although this is a real number which depends on a tunable parameter $\mathcal{M}$, in the case of a homogeneous spectrum and $O(1)$ entries, setting $\mathcal{M}$ large enough to ensure that $r$ is $O(\mathrm{rank(G)})$ will guarantee that the ground truth is representable. In the noiseless case, this guarantees that all our bounds also hold for the Population Risk. See Foygel et al. (2011) and Ledent et al. (2021) for more details.

**Remark:** In applications, the minimization problem from eq. (12) is replaced by a Lagrangian form, which can be solved with gradient methods in Pytorch. The equivalence between the regularizer term $\left[\|A\|_{\mathrm{Fr}}^2 + \|B\|_{\mathrm{Fr}}^2\right]$ and the nuclear norm of $\underline{M}$ is a consequence of the classic Lemma 6 in (Mazumder, Hastie, and Tibshirani 2010). The precise algorithm can be found in Alg. 1.

$$\textbf{minimize}\ \frac{1}{N}\sum_{o=1}^{N} \mathrm{l}\left((XAB^\top Y^\top)_{\xi_o}, \widetilde{G}_o\right) \tag{10}$$

$$+\lambda_{\mathcal{M}}\left[\|A\|_{\mathrm{Fr}}^2 + \|B\|_{\mathrm{Fr}}^2\right] \tag{11}$$

Algorithm 1: DAMC (Distributionally Aware Matrix Completion)

**Require:** Observed unlabeled data $\mathcal{S}_M = \{\xi_1, \ldots, \xi_M\}$, Observed labeled data $\mathcal{S}_N = \{(\xi_1^e, \tilde{G}_1), \ldots, (\xi_N^e, \tilde{G}_N)\}$, parameters $d$ (size of the side information), and upper bound constraint $\mathcal{M}$ on the nuclear norm of the core matrix.

**Output:** Predictions $X\underline{M}Y \in \mathbb{R}^{m\times n}$

1: Construct the matrix $\mathcal{H} = \frac{1}{M}\sum_{o=1}^{M} \xi_o$.

2: Compute the truncated SVD $\mathcal{H} = U\Sigma V^\top$ (up to rank $d$) and define $X = \sqrt{\frac{m}{d}}U$ and $Y = \sqrt{\frac{n}{d}}V$.

3: Solve the optimization problem:

$$\underline{M} = \underset{\|\underline{M}\|_* \le \mathcal{M}}{\arg\min}\ \mathrm{l}(X\underline{M}Y^\top)$$

$$= \underset{\|\underline{M}\|_* \le \mathcal{M}}{\arg\min}\ \frac{1}{N}\sum_{o=1}^{N} \mathrm{l}\left((X\underline{M}Y^\top)_{\xi_o}, \widetilde{G}_o\right) \tag{12}$$

4: **Return** Core matrix $\underline{M}$; side information matrices $X$, $Y$, and matrix of predictions $\widehat{Z} = X\underline{M}Y^\top$.

## Main Results

**Theorem 1.** *Instate Assumptions 1, 2, 4, 5, 6,and 7, then:*

$$M \ge 470\log\left(\frac{4[m+n]}{\delta}\right)\kappa_*^2\mathcal{P}^{*2}[m+n]. \tag{13}$$

*With probability greater than $1-\delta$ over the draw of both the implicit and explicit feedbacks, the following generalization bound holds simultaneously over any predictor $X\underline{M}Y^\top \in \mathbb{R}^{m\times n}$ for $\underline{M} \in \mathbb{R}^{d\times d}$ such that $\|\underline{M}\| \le \mathcal{M}$*

$$\mathrm{l}\left(X\underline{M}Y^\top\right) - \widehat{\mathrm{l}}\left(X\underline{M}Y^\top\right) \le \tag{14}$$

$$\frac{2\,\mathcal{B}\log(6/\delta)}{\sqrt{N}} + 16\,\ell\,\mathcal{P}^{*2}\sqrt{\kappa_1\kappa_2}\log(2de)\sqrt{\frac{dr}{N}} +$$

$$75\mathcal{P}^*\ell\kappa_*\kappa_1\log\left(\frac{12[m+n]}{\delta}\right)\sqrt{\frac{[m+n]r}{M}} +$$

$$25\mathcal{P}^*\ell\kappa_*\log\left(\frac{12[m+n]}{\delta}\right)\sqrt{\frac{[m+n]\Gamma r}{MN}}$$

*where as usual,* $\mathrm{l}(Z) := \mathbb{E}\left(\mathrm{l}(Z_\xi, \widetilde{G})\right)$ *and* $\widehat{\mathrm{l}}(Z) := \frac{1}{N}\sum_{o=1}^{N}\mathrm{l}(Z_{\xi_o^e}, \widetilde{G}_o)$.

We also have the following immediate consequence in terms of excess risk.

**Corollary 1.** *Let $\widehat{Z}$ be the matrix output by algorithm 1, under assumptions 1-7, we have the following excess risk bound w.p. $\ge 1-\delta$ : $\mathbb{E}(\mathrm{l}(\widehat{Z}, \widetilde{G})) - \widehat{\mathbb{E}}(\mathrm{l}(\widehat{Z}, \widetilde{G})) \le$*

$$O\left[[\ell+\mathcal{B}]\kappa_*\mathcal{P}^*\kappa_1\log\left(\frac{[m+n]}{\delta}\right)\sqrt{\frac{[m+n]r}{M}}\right] +$$

$$O\left[[\ell+\mathcal{B}]\kappa_*\mathcal{P}^*\kappa_1\log\left(\frac{[m+n]}{\delta}\right)\sqrt{\frac{\Gamma[m+n]r}{MN}}\right] +$$

$$O\left[\ell\,\mathcal{P}^{*2}\sqrt{\kappa_1\kappa_2}\log(2de)\sqrt{\frac{dr}{N}}\right].$$

If only Assumptions 1-6 hold, then one can replace $\Gamma$ (from Assumption 7) by the cruder estimate $\Gamma = \kappa_1[m+n]$ which can be obtained from assumption 4 instead. This allows us to replace Assumption 7 by

$$N \ge \frac{m+n}{2}, \quad \text{so that we have} \tag{15}$$

$$\mathbb{E}(\mathrm{l}(\widehat{Z}, \widetilde{G})) - \widehat{\mathbb{E}}(\mathrm{l}(\widehat{Z}, \widetilde{G})) \le \tag{16}$$

$$\widetilde{O}\left[[\ell+\mathcal{B}]\kappa_*\mathcal{P}^*\kappa_1^{3/2}\sqrt{\frac{[m+n]r}{M}} + \ell\,\mathcal{P}^{*2}\sqrt{\kappa_1\kappa_2}\sqrt{\frac{dr}{N}}\right],$$

where the notation $\widetilde{O}$ hides polylogarithmic factors in $m, n, \delta$. Although the condition (15) implies the number of labeled samples $N$ must pass a threshold which is linear in the size of the full matrix, this is a fixed threshold which doesn't depend on the desired error level. In addition, this is only a worst-case scenario: one can alternatively assume that Assumption 7 holds with $\Gamma \simeq d$, also allowing the absorption of the higher order term $\sqrt{\frac{\Gamma[m+n]r}{MN}}$ into the others.

In all cases, treating $\kappa_1, \kappa_2, \kappa_*, p^*, \ell, \mathcal{B}$ as constants and under either assumption 7 with $\Gamma = O(d)$ constant or condition (15), we see that the error scales as

$$\widetilde{O}\left(\sqrt{\frac{[m+n]r}{M}} + \sqrt{\frac{dr}{N}}\right).$$

Here, the first term corresponds to the error in the estimation of the shared low-rank subspaces with the unlabeled data, and the second one corresponds to the error in estimating the ground truth matrix assuming a perfect knowledge of the side information matrices $X, Y$. Thus, accurate recovery can be performed as long as we have at least $M = \widetilde{O}([m+n]r)$ unlabeled samples and $N = \widetilde{O}(dr)$ *labeled* samples. Thus, the result shows that the errors corresponding to the estimation of the common subspace and the estimation of the ground truth matrix based on the subspace information only combine additively. In particular, this implies that successful recovery of the ground truth matrix (in terms of in-distribution excess risk) is possible with only a very small number of labeled samples, as long as a larger number of unlabeled samples is available. This conclusion is of interest in the field of recommender systems, where the labeled interactions correspond to 'explicit feedback' and the unlabeled interactions correspond to 'implicit feedback'.

This observation is in sharp contrast to the sample complexities which can be obtained via a direct application of existing MC results (ignoring the unlabeled samples): if we were to apply the state-of-the-art results for matrix completion in the i.i.d. setting with uniform marginals without the use of the side information matrices $X, Y$, one would obtain a bound of $\widetilde{O}\left(\sqrt{\frac{[m+n]r}{N}}\right)$ (cf. Foygel et al. (2011)): the number of labeled samples would need to be as high as $\widetilde{O}([m+n]r)$. In contrast, we only require $\widetilde{O}(dr)$ labeled interactions. A remarkable property of this result is that the bound is meaningful even if the average number of labeled samples per row/column is vanishingly small, as long as there are $\widetilde{O}(r)$ *unlabeled* samples in each row or column.

**Remarks on the value of $d$ in Assumption 2 and Assumptions 5 and 6:** Well-conditioning assumptions such as Assumption 2 and Assumptions 5 and 6 are common in matrix perturbation theory (Chen et al. 2021). We note that true rank $r$ of the matrix G can be much smaller than the dimension $d$ of the shared low-rank subspace. Indeed, the core matrix $\underline{M}^*$ can be low-rank. Thus, Assumption 2, 5 and 6 can be summarized as follows: (1) the sampling distribution $P$ is well approximated by a rank $d$ matrix whose row and column spaces include those of the ground truth and (2) the row and column spaces of the ground truth G are (possibly strict) subspaces of those of the sampling distribution $P$. Cf. Remark 1 in the Appendix for more details.

## Experiments

We perform both synthetic and real data experiments to validate the pertinance of our bounds. Our key claims are:

- **Claim 1** The errors stemming from the subspace estimation (with the unlabeled samples) and (inductive) matrix completion components only combine additively.
- **Claim 2** In recommender systems datasets, unlabeled interaction data (often referred to as 'implicit feedback') contains relevant information to estimate the row and column subspaces of the ground truth matrix of labeled interaction data (containing the ratings from 1 to 5 given by each user to the interacted items).

We validate Claim 1 with synthetic data experiments and Claim 2 with real data respectively.

### Synthetic Data Experiments

To demonstrate the decomposition of the error into independent terms corresponding to the estimation errors of the labeled and unlabeled data respectively, we generated $\mathrm{G}, P \in \mathbb{R}^{200\times 200}$ satisfying our assumptions with $d = r = 4$. The detailed experimental setup can be found in the Appendix. For a broad range of values of both $N$ and $M$ ($M \in \{10000, 20000, \ldots, 100000\}$ and $N \in \{50, 100, 150, \ldots, 1000\}$), we evaluate the average generalization error over 30 independent runs. This range is selected because $N = 100, M = 100000$ results in perfect recovery up to a high decimal point. We compared two quantities:

1 The generalization gap (test error − training error), and

2 A disentangled estimate of the generalization error calculated as follows:

$$\text{Disentangled Estimate}(M, N) = \text{GAP}(M, 1000) + \text{GAP}(100000, N), \tag{17}$$

where 1000 an 100000 are the maximum possible values for $N$ and $M$ respectively.

Thus, the two terms in equation (17) can be interpreted as corresponding to the errors introduced from the estimation of the subspace and the ground truth matrix respectively. The results are presented below in Figure 1.

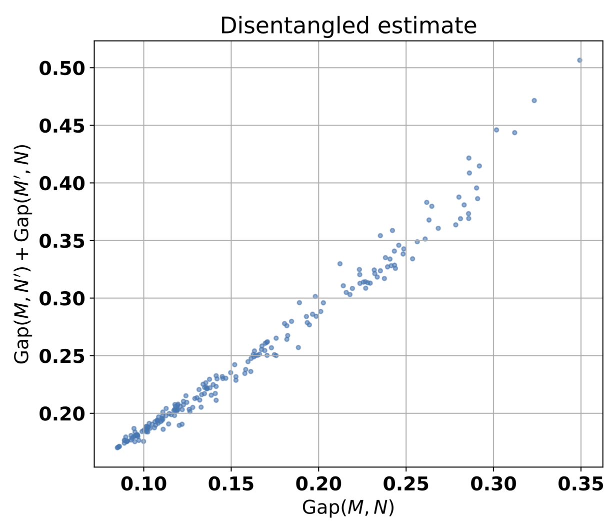


Figure 1: Comparison of generalization error (x-axis) and the corresponding disentangled estimate (y axis) in the synthetic dataset. Each point in the scatter plot corresponds to one configuration $(M, N)$, with the results averaged over 30 independent runs.

We observe a strong correlation between the error and its disentangled estimate. This suggests the two forms of errors indeed combine additively without strong interactive effects.

| Dataset | Method | 0.0 | 0.05 | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|
| ML-100K | userKNN | 1.0123 | 1.0120 | 1.0151 | 1.0231 | 1.0380 | 1.0706 | 1.1675 | 1.2462 |
| | IGMC | 0.9281 | 0.9238 | 0.9321 | 0.9604 | 0.9824 | 1.0268 | 1.1165 | 1.1482 |
| | Soft Impute | 0.9179 | 0.9324 | 0.9373 | 0.9487 | 0.9616 | 1.0380 | 1.4190 | 1.8230 |
| | DAMC | **0.9068** | **0.9165** | **0.9241** | **0.9354** | **0.9364** | **0.9621** | **1.0060** | **1.0460** |
| Douban | userKNN | 0.7946 | 0.7948 | 0.7973 | 0.8048 | 0.8288 | 0.8848 | 0.9639 | 0.9838 |
| | IGMC | 0.7437 | 0.7480 | 0.7579 | 0.7665 | 0.8010 | 0.8256 | 0.8713 | **0.8462** |
| | Soft Impute | 0.7383 | 0.7443 | 0.7427 | 0.7621 | 0.8289 | 1.0145 | 1.8471 | 3.1966 |
| | DAMC | **0.7178** | **0.7195** | **0.7205** | **0.7259** | **0.7382** | **0.7618** | **0.8323** | 0.8577 |
| Yelp | userKNN | 1.0955 | 1.1020 | 1.1060 | 1.1219 | 1.1496 | 1.2015 | 1.2230 | 1.2356 |
| | IGMC | 1.0707 | 1.0955 | 1.0759 | 1.1150 | 1.0899 | 1.1492 | 1.2028 | **1.0859** |
| | Soft Impute | 1.3888 | 1.4010 | 1.4126 | 1.7608 | 1.8428 | 2.0162 | 3.2690 | 3.4140 |
| | DAMC | **1.0320** | **1.0650** | **1.0470** | **1.0283** | **1.0580** | **1.0750** | **1.1290** | 1.1390 |

Table 1: Performance in terms of Root Mean Squared Error (RMSE) across three datasets for varying values of $p$. Lower is better. The best results in each setting are presented in **boldface**.

## Real Data Experiments

We also perform real data experiments on three popular datasets to evaluate whether the sampling distribution over observed user-item interactions (often referred to as 'implicit feedback') contains information which can be used to improve performance at the prediction of *ratings* on a scale from 1 to 5 (often referred to as the 'explicit feedback'). We performed experiments on three well-known datasets: Douban (Zhu et al. 2020, 2019), Yelp (Zhang, Zhao, and LeCun 2015) and MovieLens 100K (Harper and Konstan 2015). As a result of our hypothesis, which concerns semi-supervised matrix completion, our training setup is somewhat different from standard benchmarks: instead of relying on a training set of labeled interactions, we remove a fraction $p$ of the labels associated to interactions in the training set. In other words, a proportion $(1 - p)$ of the observations in the training set contain the observed rating, while the remaining interactions are provided merely in the form of user-item interaction pairs $(i, j)$ with no rating information. Due to the nonlinearity of real-world data, we tested a slight modification of DAMC where the singular value decomposition of the empirical distribution is replaced by a nonlinear autoencoder. However, the inductive matrix completion components (line 3 of Algorithm 1) was kept unchanged. We compare the method to various classic baselines for explicit feedback prediction: UserKNN (Herlocker et al. 1999), Softimpute (Mazumder, Hastie, and Tibshirani 2010), and IGMC (Zhang and Chen 2019). We emphasize that our aim is mostly to demonstrate the validity of our learning paradigm, rather than to provide a state-of-the-art recommender systems model: we show that a method which relies on the unsupervised information can provide better predictions on the unseen labeled test data, compared to purely supervised methods which rely only on the fully observed labeled entries (cf. Table 1).

We observe that DAMC, which relies on the unsupervised information in the pure implicit feedback, significantly outperforms classic methods which rely only on the labeled observations in most situations. This indicates that a relationship between the sampling distribution and the ground truth matrix indeed exists, lending legitimacy to our theoretical learning setting. In particular, DAMC significantly outperform its counterpart Softimpute (which is the same algorithm without using side information) for all nonzero values of $p$. Most notably, we observe that for large values of $p$ such as $p = 0.90$ and $p = 0.95$, many of the baseline models relying only on explicit ratings are not able to perform much better than random, whilst the semi-supervised DAMC can still achieve consistently good performance.

## Conclusion

We introduced a new matrix completion learning setting where the sampling distribution and the ground truth matrix are both low-rank and *share common row and column subspaces*. This setting is inspired by the recommender systems application, where unlabeled interactions ('implicit feedback') are usually much more abundant than labeled interactions ('explicit feedback'). Assuming access to a larger amount $M$ of unlabeled samples and a smaller number $N$ of labeled samples, we show generalization error bounds of the form $\widetilde{O}\left(\sqrt{\frac{[m+n]r}{M}} + \sqrt{\frac{dr}{N}} + \sqrt{\frac{\Gamma[m+n]r}{MN}}\right)$. When either $\Gamma$ (the ratio between the maximum and average sampling probability) is $O(d)$ or $N \geq \frac{m+n}{2}$, the higher-order term $\sqrt{\frac{\Gamma[m+n]r}{MN}}$ vanishes, demonstrating a disentanglement between two sources of error: the estimation of the shared low-dimensional subspaces relying on the unlabeled samples, and the estimation of the ground truth matrix. In particular, our results demonstrate the ground truth matrix can be recovered accurately even with a vanishingly small number of labeled interactions per row/column. On real data, we show that unlabeled samples can dramatically improve the performance of explicit feedback prediction methods, lending validity to our assumptions. In future work, it would be interesting to distill the real-data results into a true SVD and to tackle the difficulty of removing Assumption 7 or the uniform marginals assumption by imposing suitable modifications on the algorithm, or to attempt to derive *optimistic bounds* with a fast decay rate in $N$.

## Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2025-08-066T). Antoine Ledent and Soo Mun Chong's research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

## References

Abbe, E. 2018. Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, 18(177): 1–86.

Abbe, E.; Bandeira, A. S.; and Hall, G. 2016. Exact Recovery in the Stochastic Block Model. *IEEE Transactions on Information Theory*, 62(1): 471–487.

Alves, R. 2024. Regionalization-Based Collaborative Filtering: Harnessing Geographical Information in Recommenders. *ACM Transactions on Spatial Algorithms and Systems*, 10(2): 1–23.

Alves, R.; and Ledent, A. 2024. Context-aware representation: Jointly learning item features and selection from triplets. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 6492–6502.

Alves, R.; Ledent, A.; Assunção, R.; and Kloft, M. 2020. An Empirical Study of the Discreteness Prior in Low-Rank Matrix Completion. *Proceedings of Machine Learning Research (PMLR): NeurIPS 2020 Workshop on the Preregistration Experiment: An Alternative Publication Model For Machine Learning Research*.

Amiridi, M.; Kargas, N.; and Sidiropoulos, N. D. 2022. Low-rank characteristic tensor density estimation part I: Foundations. *IEEE Transactions on Signal Processing*, 70: 2654–2668.

Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S. M.; and Telgarsky, M. 2014. Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research*, 15: 2773–2832.

Balcan, M.-F.; and Blum, A. 2005. A PAC-style model for learning from labeled and unlabeled data. In *International Conference on Computational Learning Theory*, 111–126. Springer.

Bekker, J.; and Davis, J. 2020. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4): 719–760.

Blum, A.; and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, 92–100.

Candès, E. J.; and Recht, B. 2009. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6): 717.

Candès, E. J.; and Tao, T. 2010. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Trans. Inf. Theor.*, 56(5): 2053–2080.

Candès, E.; and Plan, Y. 2010. Matrix Completion With Noise. *Proceedings of the IEEE*, 98: 925 – 936.

Chen, Y.; Chi, Y.; Fan, J.; and Ma, C. 2021. *Spectral Methods for Data Science: A Statistical Perspective*.

Chen, Y.; Chi, Y.; Fan, J.; Ma, C.; and Yan, Y. 2020a. Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization.

Chen, Y.; Chi, Y.; Fan, J.; Ma, C.; and Yan, Y. 2020b. Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization. *SIAM Journal on Optimization*, 30(4): 3098–3121.

Chiang, K.-Y.; Dhillon, I. S.; and Hsieh, C.-J. 2018. Using Side Information to Reliably Learn Low-Rank Matrices from Missing and Corrupted Observations. *J. Mach. Learn. Res.*

Chiang, K.-Y.; Hsieh, C.-J.; and Dhillon, I. S. 2015. Matrix Completion with Noisy Side Information. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Choi, J.; and Yuan, M. 2024. Matrix completion when missing is not at random and its applications in causal panel data models. *Journal of the American Statistical Association*, 1–15.

Foygel, R.; Shamir, O.; Srebro, N.; and Salakhutdinov, R. R. 2011. Learning with the weighted trace-norm under arbitrary sampling distributions. *Advances in neural information processing systems*, 24.

Ghanooni, N.; Mustafa, W.; Lei, Y.; Lin, A. W.; and Kloft, M. 2024. Generalization Bounds with Logarithmic Negative-Sample Dependence for Adversarial Contrastive Learning. *Transactions on Machine Learning Research*.

Ghanooni, N.; Mustafa, W.; Wagner, D.; Fellenz, S.; Lin, A. W.; and Kloft, M. 2025. Mitigating Spurious Features in Contrastive Learning with Spectral Regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hänsch, S.; Sajdoková, A.; Rabau, A.; Rybář, V.; Alves, R.; and Kordík, P. 2025. Data-driven closure model selection for multiphase CFD via matrix completion. *AI Thermal Fluids*, 100019.

Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4): 1–19.

Herlocker, J. L.; Konstan, J. A.; Borchers, A.; and Riedl, J. 1999. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 230–237. ACM.

Hieu, N. M.; and Ledent, A. 2025. Generalization Analysis for Supervised Contrastive Representation Learning under Non-IID Settings. In *Forty-second International Conference on Machine Learning*.

Hieu, N. M.; Ledent, A.; Lei, Y.; and Ku, C. Y. 2025. Generalization analysis for deep contrastive representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17186–17194.

Jalan, A.; Jedra, Y.; Mazumdar, A.; Mukherjee, S. S.; and Sarkar, P. 2025. Optimal Transfer Learning for Missing Not-at-Random Matrix Completion. *arXiv preprint arXiv:2503.00174*.

Jirasek, F.; Alves, R. A. S.; Damay, J.; Vandermeulen, R. A.; Bamler, R.; Bortz, M.; Mandt, S.; Kloft, M.; and Hasse, H. 2020. Machine Learning in Thermodynamics: Prediction of Activity Coefficients by Matrix Completion. *The Journal of Physical Chemistry Letters*, 11(3): 981–985.

Koren, Y.; Bell, R. M.; and Volinsky, C. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer*, 42(8): 30–37.

Ledent, A.; and Alves, R. 2024. Generalization Analysis of Deep Non-linear Matrix Completion. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 26290–26360. PMLR.

Ledent, A.; Alves, R.; and Kloft, M. 2021. Orthogonal Inductive Matrix Completion. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.

Ledent, A.; Alves, R.; Lei, Y.; Guermeur, Y.; and Kloft, M. 2023. Generalization bounds for inductive matrix completion in low-noise settings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8447–8455.

Ledent, A.; Alves, R.; Lei, Y.; and Kloft, M. 2021. Fine-grained Generalization Analysis of Inductive Matrix Completion. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 25540–25552. Curran Associates, Inc.

Ledent, A.; Kasalickỳ, P.; Alves, R.; and Lauw, H. W. 2025. Conv4Rec: A 1-by-1 Convolutional Autoencoder for User Profiling Through Joint Analysis of Implicit and Explicit Feedback. *IEEE Transactions on Neural Networks and Learning Systems*.

Lei, Y.; Yang, T.; Ying, Y.; and Zhou, D.-X. 2023. Generalization analysis for contrastive representation learning. In *International Conference on Machine Learning*, 19200–19227. PMLR.

Li, R.; Dong, Y.; Kuang, Q.; Wu, Y.; Li, Y.; Zhu, M.; and Li, M. 2015. Inductive matrix completion for predicting adverse drug reactions (ADRs) integrating drug–target interactions. *Chemometrics and Intelligent Laboratory Systems*, 144: 71 – 79.

Ma, W.; and Chen, G. H. 2019. Missing Not at Random in Matrix Completion: The Effectiveness of Estimating Missingness Probabilities Under a Low Nuclear Norm Assumption. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mazumder, R.; Hastie, T.; and Tibshirani, R. 2010. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J. Mach. Learn. Res.*, 11: 2287–2322.

McGrath, S.; Zhu, C.; Guo, M.; and Duan, R. 2024. LEARNER: A Transfer Learning Method for Low-Rank Matrix Estimation. arXiv:2412.20605.

Neyshabur, B.; Tomioka, R.; and Srebro, N. 2015. Norm-Based Capacity Control in Neural Networks. In Grunwald, P.; Hazan, E.; and Kale, S., eds., *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, 1376–1401. Paris, France: PMLR.

Poernomo, J.; Tan, N. G. L.; Alves, R.; and Ledent, A. 2025. Probabilistic Modeling, Learnability and Uncertainty Estimation for Interaction Prediction in Movie Rating Datasets. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 1261–1266.

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

Recht, B. 2011. A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research*, 12: 3413–3430.

Shamir, O.; and Shalev-Shwartz, S. 2011. Collaborative Filtering with the Trace Norm: Learning, Bounding, and Transducing. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, 661–678. PMLR.

Shamir, O.; and Shalev-Shwartz, S. 2014. Matrix Completion with the Trace Norm: Learning, Bounding, and Transducing. *Journal of Machine Learning Research*, 15: 3401–3423.

Vandermeulen, R. A.; and Ledent, A. 2021. Beyond smoothness: Incorporating low-rank analysis into nonparametric density estimation. *Advances in Neural Information Processing Systems*, 34: 12180–12193.

Xia, L.; Huang, C.; Xu, Y.; Zhao, J.; Yin, D.; and Huang, J. 2022. Hypergraph contrastive collaborative filtering. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, 70–79.

Xu, M.; Jin, R.; and Zhou, Z.-H. 2013. Speedup Matrix Completion with Side Information: Application to Multi-Label Learning. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, 2301–2309. Red Hook, NY, USA: Curran Associates Inc.

Zhang, M.; and Chen, Y. 2019. Inductive matrix completion based on graph neural networks. *arXiv preprint arXiv:1904.12058*.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhu, F.; Chen, C.; Wang, Y.; Liu, G.; and Zheng, X. 2019. DTCDR: A framework for dual-target cross-domain recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1533–1542.

Zhu, F.; Wang, Y.; Chen, C.; Liu, G.; and Zheng, X. 2020. A Graphical and Attentional Framework for Dual-Target Cross-Domain Recommendation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 3001–3008.