Generalization Bounds for Inductive Matrix Completion in Low-Noise Settings

Antoine Ledent[1*], Rodrigo Alves[2], Yunwen Lei[3], Yann Guermeur[4], and Marius Kloft[5]

[1] Singapore Management University (SMU)
[2] Czech Technical University in Prague (CTU)
[3] Hong Kong Baptist University (HKBU)
[4] Centre National de la Recherche Scientique (CNRS)
[5] Technische Universität Kaiserslautern (TUK)
aledent@smu.edu.sg, rodrigo.alves@fit.cvut.cz, yunwen.lei@hotmail.com, yann.guermeur@loria.fr, kloft@cs.uni-kl.de

Abstract

We study inductive matrix completion (matrix completion with side information) under an i.i.d. subgaussian noise assumption at a low noise regime, with uniform sampling of the entries. We obtain for the first time generalization bounds with the following three properties: (1) they scale like the standard deviation of the noise and in particular approach zero in the exact recovery case; (2) even in the presence of noise, they converge to zero when the sample size approaches infinity; and (3) for a fixed dimension of the side information, they only have a logarithmic dependence on the size of the matrix. Differently from many works in approximate recovery, we present results both for bounded Lipschitz losses and for the absolute loss, with the latter relying on Talagrand-type inequalities. The proofs create a bridge between two approaches to the theoretical analysis of matrix completion, since they consist in a combination of techniques from both the exact recovery literature and the approximate recovery literature.

Introduction

Matrix Completion (MC), the problem which consists in predicting the unseen entries of a matrix based on a small number of observations, presents the rare combination of (1) a rich mathematical playground rife with fundamental unsolved problems, and (2) a wealth of unexpected applications in lucrative and meaningful fields, from Recommender Systems (Yao and Kwok 2019; Chen and Li 2017; Aggarwal 2016) to the prediction of drug interaction (Li et al. 2015).

One of the most celebrated algorithms for standard matrix completion is the Softimpute algorithm (Mazumder, Hastie, and Tibshirani 2010), which solves the following optimization problem:

$$\min_{Z\in\mathbb{R}^{m\times n}} \frac{1}{2}\|P_\Omega(Z-R)\|_{\mathrm{Fr}}^2 + \lambda\|Z\|_*, \tag{1}$$

where P_Ω denotes the projection on the set Ω of observed entries, R is the ground truth matrix, $\|\cdot\|_*$ denotes the *nuclear norm* (the sum of the matrix's singular values) and $\|\cdot\|_{\mathrm{Fr}}$ denotes the Frobenius norm. The idea of the algorithm is to encourage *low-rank* solutions in a similar way to how L^1 regularization encourages component sparsity. The parameter λ must be tuned with cross-validation.

Inductive matrix completion (IMC) (Herbster, Pasteris, and Tse 2019; Zhang, Du, and Gu 2018; Menon and Elkan 2011; Chen et al. 2012) is another closely related model which assumes that additional information is available in the form of feature vectors for each user (row) and item (column). It assumes that the side information is summarized in matrices $X \in \mathbb{R}^{m\times a}$ and $Y \in \mathbb{R}^{n\times b}$. IMC then optimizes the following objective function

$$\min_{M\in\mathbb{R}^{a\times b}} \frac{1}{N}\|P_\Omega(XMY^\top - R)\|_{\mathrm{Fr}}^2 + \lambda\|M\|_*. \tag{2}$$

An interesting question is whether one can provide sample complexity guarantees for the optimization problem above. Typically, doing so requires minor modification to the problem for technical convenience. There are several such analogues optimization problems (1) and (2), depending on the type of statistical guarantee expected and the assumptions: in exact recovery (with the assumption of perfectly noiseless observations), the Frobenius norm is replaced by a hard equality constraint, whilst in approximate (noisy) recovery, the nuclear norm regrulariser is replaced by a hard contraint.

More precisely, *exact recovery* results study the following hard version of the optimization problem:

$$\min_{Z\in\mathbb{R}^{m\times n}} \|Z\|_* \quad \text{subject to}$$
$$Z_{i,j} = R_{i,j} \quad \forall (i,j)\in\Omega. \tag{3}$$

In the case of IMC, the equivalent hard version is:

$$\min_{M\in\mathbb{R}^{a\times b}} \|M\|_* \quad \text{subject to}$$
$$[XMY^\top]_{i,j} = R_{i,j} \quad \forall (i,j)\in\Omega. \tag{4}$$

The study of problem (3) is the earliest branch of the related literature: it was shown in a series of papers (Candès and Tao 2010; Candès and Recht 2009; Recht 2011, to name but a few) that if the number of samples is $\geqslant \widetilde{O}(nr)$ (where r is the rank and n is the size of the matrix, i.e. the number of rows or columns, which ever is larger), then it is possible to recover the whole matrix exactly with high probability as long as the entries are sampled uniformly at random. There has also been some more recent interest in the problem (4): it was shown in (Xu, Jin, and Zhou 2013) that *assuming the*

*Corresponding author

side information X, Y is made up of orthonormal columns, exact recovery is possible as long as the number of samples $N = |\Omega|$ satisfies $\widetilde{O}(ar) \leqslant N \leqslant \widetilde{O}(abr)$. Here, the $\widetilde{O}$ notation hides logarithmic factors in all relevant quantities (including the size $m \times n$ of the matrix).

Approximate recovery results typically study modified problems such as the problem below, for which Equation (1) can be interpreted as a Lagragnian form):

$$\min_{Z} \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \ell(R_{i,j}, Z_{i,j}) \quad \text{subject to} \quad \|Z\|_* \leqslant \mathcal{M}, \tag{5}$$

for some loss function ℓ which is typically assumed to be bounded and Lipschitz, and some constant $\mathcal{M}$ which must be tuned through cross-validation in a way analogous to the tuning of λ in equation (1) in real-life applications. In the case of IMC, the equivalent problem is:

$$\min_{M} \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \ell([XMY^\top]_{i,j}, Z_{i,j}) \quad \text{subject to} \quad \|M\|_* \leqslant \mathcal{M}. \tag{6}$$

Approaching the problem this way allows one to deploy the machinery of Rademacher complexities from traditional statistical learning theory to obtain uniform bounds on the generalization gap of any predictor in the given class. Using such techniques, bounds of $\widetilde{O}\left(\sqrt{\frac{nr}{N}}\right)$ (resp. $\widetilde{O}(a^2 r/\sqrt{N})$, more recently $\widetilde{O}\left(\sqrt{\frac{ar}{N}}\right)$) were shown for approximate recovery MC (resp. IMC) under uniform sampling (MC: see (Shamir and Shalev-Shwartz 2011, 2014), IMC, see (Chiang, Dhillon, and Hsieh 2018; Ledent et al. 2021), cf. also related works). In the distribution-free case, the corresponding rates are $\widetilde{O}\left(\sqrt{\frac{n^{3/2}r^{1/2}}{N}}\right)$ and $\widetilde{O}\left(\sqrt{\frac{a^{3/2}r^{1/2}}{N}}\right)$.

The above rates do not make any assumptions on the noise whatsoever, and depend only on explicit dimensional quantities: they are classified as "uniform convergence" bounds in the classic paradigm of statistical learning theory. In particular, while they do also apply to the noiseless case, they are subsumed by the exact recovery results in this case provided the exact recovery threshold is reached.

Thus the most striking hole in the existing theory is the chasm between exact recovery and approximate recovery in Inductive Matrix Completion: on the one hand, we know that if the entries are observed exactly, solving problem (4) will eventually recover the whole matrix exactly with high probability given enough entries. On the other hand, we know from the approximate recovery literature that regardless of the noise distribution, solving a properly cross-validated version of problem (6) will allow us to approach the Bayes error at speed at least $1/\sqrt{N}$ as we observe more entries. It seems reasonable to expect that in real life, neither of these approaches fully explains the statistical generalization landscape of the problem: we never expect to observe the entries exactly, and the ground truth is probably not exactly low-rank either, but we still do not expect convergence to the Bayes error to be as slow as in the worst case. What would be more reasonable to expect is a sharp decline of the error around a threshold value before which no method can work even if the entries are observed exactly, followed by a slower decline as the model refines its predictions and evens out the noise in the observations. This can be observed practically as well, as can be seen from Figure 1 in the experiments section: the decay of the error as the number of samples increases is neither convex (unlike the functions $1/\sqrt{N}$ and $1/N$), nor completely abrupt (as exact recovery results suggest), which indicates the presence of a threshold phenomenon.

In this paper, we theoretically capture this phenomenon through generalization error bounds for the solutions to problem (2) when the ground truth matrix is observed with some subgaussian noise of subgaussianity constant σ. In addition, our results completely remove the orthogonality assumptions on the side information matrices X, Y which are present in the related work (Xu, Jin, and Zhou 2013), thus improving the state of the art even in the exact recovery case.

In summary, we make the following important contributions:

1. We prove (cf. Theorem 1) that *exact recovery* is possible for IMC (when the entries are observed exactly) with probability $1 - \Delta$ given $\widetilde{O}\left(\mu^5 r^2 (a+b)\sigma_0^{-4} \log\left(\frac{mn}{\Delta}\right)\right)$ samples or more. This is a significant extension of the results in (Xu, Jin, and Zhou 2013) in that we remove most many of their assumptions. In the formula above, μ is a measure of incoherence, and σ_0 denotes the smallest singular value of X or Y assuming they are normalized so that the largest singular value is 1 in each case. This means that after suitable scaling, σ_0 can be replaced by the ratio between the largest and smallest singular values of either X or Y. The presence of this factor underpins one of the main differences between (Xu, Jin, and Zhou 2013) and our work. Indeed, the most limiting assumption in (Xu, Jin, and Zhou 2013) is that the columns of the side information matrices X and Y are *orthonormal*, which is equivalent to assuming that $\sigma_0 = 1$.
2. We experimentally observe the two-phase phenomenon described above via synthetic data experiments.
3. We prove generalization bounds (cf. Theorem 2) which capture this phenomenon in the case of bounded loss functions such as the truncated L^2 loss. Indeed, we show that as long as N exceeds the threshold from the exact recovery result, the expected loss scales as $\widetilde{O}\left(\sigma_0^{-2}\,\sigma\,\mu\frac{\sqrt{a^3 b}}{\sqrt{N}}\log^3(N/\Delta)\right)$, where σ is the subgaussianity constant of the noise. If σ is very small, this implies that before the exact recovery threshold (ERT) is reached, the best available bounds are the uniform convergence bounds (which are vacuous at that regime), whereas as soon as the ERT is crossed, our bounds become valid and already have a small value, which continues to drop further as the number of samples increases. This partially explains the sharp drop in the reconstruction error around the ERT even in the noisy case.
4. Using Talagrand-type inequalities, we further prove a similar result (cf. Theorem 3) which applies to the absolute loss $\ell(x, y) = |x - y|$, despite the fact that it is unbounded.

Note as a side benefit that both of the last two results apply to the Lagrangian formulation of the IMC problem, unlike most of the existing literature on approximate recovery.

Our second result creates a bridge between the approximate recovery literature and the exact recovery literature: as the subgaussianity constant of the noise σ converges to zero, so does the error: the result then reduces to our exact recovery result. Furthermore, our proof techniques also marry both approaches: we rely *both* on the geometry of dual certificates (the tool of choice in the exact recovery literature) *and* Rademacher complexities to reach our result. Beyond our current preliminary results, we believe that the direction we initiate here will prove fertile and that many improved results can be proved, bringing us closer to a complete understanding of the sample complexity landscape of nuclear norm based Inductive Matrix Completion.

Related Work

Perturbed Exact Recovery with the Nuclear Norm: For Matrix Completion without side information, bounds which capture the two-phase phenomenon by incorporating a multiplicative factor of the the variance of the noise have been shown: in (Candès and Plan 2010), a bound of order $O\left(\sqrt{\frac{n^3}{N}}\,\sigma+\sigma\right)$ is shown for the L^2 generalisation error of matrix completion with noise of variance σ (Cf. equation III.3 on page 7). The proof relies on a perturbed version of the exact recovery arguments presented in (Candès and Tao 2010). The result considers a different loss function and does not consider side information and the proof is purely based on directly computing various norms without relying on Rademacher complexities. In a recent and very impressive contribution (Chen et al. 2020) provided some bounds in the same setting with a finer multiplicative dependency on the size of the matrix n that matches the order of magnitude of the exact recovery threshold (when expressed in terms of sample complexity). The proof is very involved and contrary to our work, the results do not apply to inductive matrix completion.

Exact Recovery with the Nuclear Norm: In (Recht 2011), extending and simplifying earlier work of (Candès and Tao 2010; Candès and Recht 2009), the author proves that exact recovery is possible for matrix completion with the nuclear norm with $\widetilde{O}(nr)$ entries. The result is extended to the case where side information is present in (Xu, Jin, and Zhou 2013) where it is shown exact recovery is possible with $\widetilde{O}((a+b)r)$ observations, where a, b are the sizes of the side information. However, the result only applies as long as this side information consists of orthonormal columns, significantly reducing the applicability. Other variations of the results exist with improved dependence on certain parameters such as the incoherence constants (Chen 2015).

Perturbed Exact Recovery for other Algorithms in learning settings other than nuclear norm minimization, there is some work with low-noise regimes where the bounds also approach zero as the noise approaches zero (for large enough N). For instance, some work on max norm regularisation has this property (Cai and Zhou 2016). Some results of order $\widetilde{O}\left(\sigma\sqrt{\frac{nr}{N}}\right)$ were also obtained for matrix completion with a special algorithm that *requires explicit rank restriction* (Keshavan, Montanari, and Oh 2009; Wang et al. 2021).

Approximate Recovery Results: There is a wide body of works proving uniform-convergence type generalization bounds for various matrix completion settings. the vast majority are of order $\widetilde{O}(1/\sqrt{N})$, with most bounds differing from each other in their dependence on other quantities such as m, n, r, μ, σ and (in IMC) a, b. For matrix completion, (Shamir and Shalev-Shwartz 2011, 2014) proves bounds of order $\widetilde{O}\left(\sqrt{\frac{n^{3/2}r^{1/2}}{N}}\right)$ in the *distribution-free setting* with replacement, as well as $\widetilde{O}\left(\frac{nr\log(n)}{N}+\sqrt{\frac{\log(1/\delta)}{N}}\right)$ in the transductive setting (i.e. for *uniform* sampling *without* replacement). In the case of inductive matrix completion, rates of $\widetilde{O}\left(\sqrt{\frac{rab}{N}}\right)$ were shown in (Chiang, Dhillon, and Hsieh 2018; Chiang, Hsieh, and Dhillon 2015; Giménez-Febrer, Pagès-Zamora, and Giannakis 2020) in a distribution-free situation, whilst (Ledent et al. 2021) provides rates of order $\widetilde{O}\left(\sqrt{\frac{ra}{N}}\right)$ and $\widetilde{O}\left(\sqrt{\frac{r^{1/2}a^{3/2}}{N}}\right)$ in the uniform sampling and distribution-free cases respectively. Similar rates were implicitly proved in the more algorithmic contribution (Ledent, Alves, and Kloft 2021) under very strict assumptions on the side information X, Y. It is also worth noting that although the component of our result which involves the subgaussianity of the noise is vacuous when the size of the side information approaches that of the matrix, that is also the case of every approximate recovery result for IMC to date except the very recent paper (Ledent et al. 2021), whose results are also uniform convergence bounds. Our bounds are far tighter those in all of those works when the noise is small.

Matrix Sensing: Matrix sensing is a learning setting with some similarities to inductive matrix completion where rank-one measurements $\langle vw^\top, R\rangle$ of an unknown matrix R are taken, and the matrix R is estimated. There are a wide variety of results depending on the assumptions on the matrix and the sampling distribution (Gross et al. 2010; Kueng, Rauhut, and Terstiege 2017; Tanner, Thompson, and Vary 2019; Zhong, Jain, and Dhillon 2015). In most cases, the measurements are sampled i.i.d. from some distribution, which introduces some substantial technical differences to the IMC setting. Often, the underlying measurements need to satisfy the restricted isometry property, which is not directly comparable to the joint incoherence assumptions on the side information matrices made in this paper and in the IMC literature. In addition, most results relate to pure exact recovery rather than a low-noise model such as the one studed here.

Notation and Setting

We assume there is an unknown ground truth matrix $R \in \mathbb{R}^{m\times n}$ that we observe noisily. To draw a sample from the distribution, we first sample an entry $\xi = (\xi_1, \xi_2) = (i, j)$ from the uniform distribution over $[m]\times[n]$. We then observe the quantity $R_{(i,j)} + \zeta_{(i,j)}$ where $\zeta_{(i,j)}$ is the noise, whose distribution can depend on the entry (i, j). The samples are

drawn i.i.d.

We suppose we have a training set of N samples and we write Ω for the set of sampled entries $\xi^1, \xi^2, \ldots, \xi^N$. It is possible to sample the same entry several times (which results in potentially different observations due to the i.i.d. nature of the noise). However, for simplicity of notation we will sometimes write $\sum_{(i,j)\in\Omega} f(R_{(i,j)})$ instead of $\sum_{\xi\in\Omega} f(R_{\xi_1,\xi_2}, \xi)$ as long as no ambiguity is possible. We are given two side information matrices $X \in \mathbb{R}^{m\times a}$ and $Y \in \mathbb{R}^{n\times b}$. Throughout this paper, $\|\cdot\|$ denotes the spectral norm, $\|\cdot\|_{\mathrm{Fr}}$ denotes the Frobenius norm, $\|\cdot\|_*$ denotes the nuclear norm, and for any integer l, $[l] = \{1, 2, \ldots, l\}$.

We make the following assumptions throughout the paper:

Assumption 1 (Realizability)**.** There exists a matrix $M_* \in \mathbb{R}^{a\times b}$ such that $R = XM_*Y^\top$.

Assumption 2 (Assumptions on the subgaussian noise)**.** We assume the noise is σ subgaussian: $\mathbb{E}(\zeta) = 0$ and $\mathbb{P}(|\zeta| \geqslant t) \leqslant 2\exp(-t^2/(2\sigma^2))$ for all t.

We will write $\bar{X}$ and $\bar{Y}$ for the matrices obtained by normalizing the columns of X, Y and we will write Σ_1, Σ_2 for the diagonal matrices containing the singular values of X, Y. Similarly we will also write $\bar{\bar{X}} = \bar{X}\Sigma_1$ etc.

We also make the following incoherence assumption.

Assumption 3. There exists a constant μ such that the following inequalities hold.

$$\|\bar{X}\|_\infty \leqslant \sqrt{\frac{\mu}{m}}, \qquad \|\bar{Y}\|_\infty \leqslant \sqrt{\frac{\mu}{n}},$$
$$\|A\|_\infty \leqslant \sqrt{\frac{\mu}{a}}, \qquad \|B\|_\infty \leqslant \sqrt{\frac{\mu}{b}}, \tag{7}$$

Here the matrices A, B are from the SVD decomposition of the ground truth core matrix $M_* = ADB^\top$ for some diagonal D.

Note that we do not make the assumption that the matrices X, Y have orthonormal columns (and in particular constant spectrum) as in (Xu, Jin, and Zhou 2013). Therefore, to cope with such extra difficulty (7) is needed in the general non orthogonal case. Whilst that reference simply assumes that the column spaces of X, Y are μ incoherent, our assumption requires that *each individual eigenspace* corresponding to each singular value of X, Y and M *be* μ*-incoherent*. In the supplementary we explain to what extent this slightly stronger assumption is necessary in the non-orthogonal case.

Optimization problem: whether considering inductive matrix completion or matrix completion with the nuclear norm, it is common to assume that the entries are sampled exactly (without noise) and that the algorithm used to recover the ground truth is the following:

$$\arg\min \left(\|M\|_* \text{ s.t. } \forall (i,j) \in \Omega, [XMY^\top]_{i,j} = R_{i,j}\right). \tag{8}$$

This is also the optimization problem we study in the exact recovery portion of our results.

In real situations where there is some noise, some relaxation of the problem is necessary. From an optimization perspective, the most common strategy is to minimize the L^2 loss on the observed entries plus a nuclear norm regularisation term:

$$\min \frac{1}{N} \sum_{\xi\in\Omega} \left|[R_\xi + \zeta_\xi] - XMY^\top\right|^2 + \lambda\|M\|_*, \tag{9}$$

where λ is a regularization parameter. The problem we will consider in this paper is the one defined by equation (9). We will also need to impose the following conditions on λ:

$$\frac{\sigma\,\sigma_0^2}{C\sqrt{aN}} \leqslant \lambda \leqslant \frac{C\,\sigma\,\sigma_0^2}{\sqrt{aN}} \tag{10}$$

for some constant C. It is assumed that λ has been tuned to reach a value which satisfies these conditions.

Main Results

Exact Recovery

We have the following extension of the main theorem in (Xu, Jin, and Zhou 2013):

Theorem 1. *Assume that the entries are observed without noise and that the strong incoherence assumption* (7) *is satisfied for a fixed* μ*. For any* $\Delta > 0$ *as long as*

$$N \geqslant \widetilde{O}\left(\mu^5 r^2 (a+b)\sigma_0^{-4} \log\left(\frac{mn}{\Delta}\right)\right),$$

with probability $\geqslant 1-\Delta$ *we have that any solution* $M_{\min}$ *to the optimization problem below*

$$M_{\min} \in \arg\min \|M\|_* \quad \textit{s.t.}$$
$$\forall (i,j) \in \Omega, \quad [XMY^\top]_{i,j} = R_{i,j}, \tag{11}$$

satisfies

$$XM_{\min}Y^\top = R.$$

Here, as usual, the $\widetilde{O}$ *notation hides further log terms in the quantities* $m, n, \sigma_0^{-1}, \log(\frac{mn}{\Delta})$.

Remark: The above optimization problem can be seen as a limiting case of (9) with $\lambda \to 0$.

Remark: The above theorem has several advantages over the main theorem in (Xu, Jin, and Zhou 2013):

1. It is expressed entirely in terms of a fixed high probability $1-\Delta$ (as opposed to relying on dimensional quantities in the expression for the high probability).
2. It works without assuming that the side information matrices have unit singular values. This is quite a significant improvement as the result in (Xu, Jin, and Zhou 2013) only holds when the side information matrices belong to a given set of measure zero. There is a quadratic dependence on σ_0^{-1} (the inverse of the smallest singular value of either X or Y), which matches the dependence in (Jain and Dhillon 2013) (although that paper works with a completely different optimization problem away from traditional nuclear norm regularization).
3. It holds for any value of N, whereas the result in (Xu, Jin, and Zhou 2013) required $N \leqslant \widetilde{O}(abr)$ and the result in (Recht 2011) (which concerns standard MC without side information) required $N \leqslant mn$.

Approximate Recovery in a Low-Noise Setting

Below we present theorems which provide generalization bounds for the IMC model (2) with the favourable property that they improve when the noise is reduced, and they reduce exactly to the exact recovery result when $\sigma = 0$.

The following theorem provides a generalization bound of order $\widetilde{O}\left(a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\sqrt{\frac{1}{N}}\right)$ for a bounded Lipschitz loss.

Theorem 2. *Let ℓ be an L_ℓ-Lipschitz loss function bounded by B_ℓ. Assume that condition* (10) *on λ holds. For any $\Delta > 0$, with probability $1-\Delta$ as long as*

$$N \geqslant \widetilde{O}\left(\mu^5 r^2 (a+b)\sigma_0^{-4}\log\left(\frac{mn}{\Delta}\right)\right),$$

we have the following bound on the performance of the solution $\widehat{R}$ to the optimization problem (2)*:*

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}(\ell(\widehat{R}_{(i,j)}, [R+\zeta]_{(i,j)})) \leqslant \tag{12}$$
$$O\left(a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\mathrm{L}_\ell\log^3\left(\frac{Nmn}{\Delta}\right)\sqrt{\frac{1}{N}} + B_\ell\frac{\log(\frac{1}{\Delta})}{N}\right),$$

where $\mathcal{U}$ stands for the uniform distribution on the entries $[m]\times[n]$.

Next, our proof techniques also allow us to prove results which apply to the absolute value loss, despite the fact that it is unbounded. Indeed, a bound of order $\sqrt{N}$ on the nuclear norm of the difference between the solution and the ground truth is a byproduct of the approximations we perform before applying Rademacher arguments. It can also be used to provide a bound on the *effective* value of B_ℓ, still yielding an overall rate of $1/\sqrt{N}$ thanks to the fact that the last term in equation (12) has the strong decay $1/N$. This is a result of our use of the more fine-grained, talagrand-type results from (Bartlett, Bousquet, and Mendelson 2005) and would not have been possible if we had used standard results on Rademacher complexities such as (Bartlett and Mendelson 2001).

Theorem 3. *Assume that condition* (10) *on λ holds. For any $\Delta > 0$, with probability $1-\Delta$ as long as*

$$N \geqslant \widetilde{O}\left(\mu^5 r^2 (a+b)\sigma_0^{-4}\log\left(\frac{mn}{\Delta}\right)\right),$$

we have

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}\left|\widehat{R}_{(i,j)} - [R+\zeta]_{(i,j)}\right| \leqslant$$
$$O\left(a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\mathrm{L}_\ell\log^3\left(\frac{Nmn}{\Delta}\right)\sqrt{\frac{1}{N}}\right). \tag{13}$$

Here, $\mathcal{U}$ stands for the uniform distribution on the entries $[m]\times[n]$.

Proof Strategy

The main ideas of our proof are (1) to redefine a norm on $\mathbb{R}^{m\times n}$ matrices that captures the effect of the side information matrices, and (2) to combine proof techniques from both the approximate recovery literature and the exact recovery literature: we perturb the analysis from the exact recovery literature to obtain a bound on the discrepancy between the ground truth and the recovered matrix, and then bootstrap the argument by exploiting the i.i.d. nature of the noise and results from traditional complexity analysis to yield a generalization bound.

In this informal description, we sometimes write formulae with such as $P_\Omega(\widehat{R}-R)$, denoting the projection of $\widehat{R}-R$ onto the set of matrices whose non zero entries are in Ω, which requires assuming that each entry was sampled only once. However, this assumption is made purely for simplicity of exposition and it is not made or needed in the formal proofs in the supplementary.

Background on Existing Techniques

The main strategy of the proof of the exact recovery results in both (Xu, Jin, and Zhou 2013) and (Recht 2011), which goes back to earlier work (Candès and Tao 2010; Candès and Recht 2009; Candès and Plan 2010) is to use the duality between the nuclear norm and the spectral norm to study the behavior of the nuclear norm around the ground truth.

It is easiest to explain the strategy in the case of standard matrix completion (as in Recht 2011; Candès and Plan 2010 etc.). For a given matrix R with singular value decomposition $EDF^\top$, if the columns and rows of W are orthogonal to those of R and it satisfies $\|W\| \leqslant 1$, the matrix $\mathcal{Y} := EF^\top + W$ is a *subgradient to the nuclear norm* at R, and a solution to the maximization problem

$$\max_{\mathcal{Y}} \langle \mathcal{Y}, R\rangle \quad \text{subject to}$$
$$\|\mathcal{Y}\| \leqslant 1.$$

The subgradients as above allow us to understand the local behavior of the nuclear norm around the ground truth, and one of the most important observations in the early exact recovery analysis of matrix completion is that exact recovery is guaranteed if there exists such a subgradient *whose non zero entries are all in the set of observed entries* and whose spectral norm is < 1. A subgradient with this property is referred to as a *dual certificate*. Indeed, we have the following result from (Candès and Plan 2010):

Lemma 4. *If there exists a dual certificate $\mathcal{Y}$, then for any Z with $Z_{i,j} = 0 \quad \forall (i,j) \in \Omega$ we have*

$$\|R+Z\|_* \geqslant \|R\|_* - (1 - P_{T^\top}(\mathcal{Y}))\|P_{T^\top}(Z)\|_*. \tag{14}$$

In particular, R is the unique solution to the optimization problem (8)*. Here $P_T(Z) = ZP_F + P_E Z - P_E Z P_F$ where P_E and P_F are the projection operators onto the column and row spaces of the ground truth respectively.*

The high-level intuition behind such a result is that if the set of "observable" matrices whose entries are constrained to lie in the set of observed entries is big enough to contain suitable subgradients, then it is big enough to make the solution to (8) unique.

Whilst most of the early works in the field (Candès and Tao 2010; Candès and Recht 2009) work with sampling without replacement and rely on complex combinatorial arguments

to prove the existence of a dual certificate, the breakthrough in the work of (Recht 2011) is to sample with replacement (simplifying the concentration arguments) and to show that the existence of an *approximate* dual certificate is also enough to guarantee uniqueness. More precisely, let $Z \in \mathbb{R}^{\Omega^\top}$ be a matrix with zeros in all entries outside Ω, and let $U, U^\top$ be the canonical subgradients of R and $P_T(Z)$ respectively. Assume there is an approximate dual certificate $\mathcal{Y}$ with the property that $\|U - P_T(\mathcal{Y})\|_{\mathrm{Fr}}$ is very small and $P_{T^\top}(\mathcal{Y}) < 1/2$, then we have

$$\begin{aligned}
&\|R + Z\|_* \\
&\geqslant \left\langle U + U^\top, R + Z \right\rangle \\
&= \|R\|_* + \left\langle U + U^\top, Z \right\rangle \\
&= \|R\|_* + \langle U - P_T(\mathcal{Y}), P_T(Z) \rangle \\
&\qquad + \left\langle U^\top - P_{T^\top}(\mathcal{Y}), P_{T^\top}(Z) \right\rangle \\
&\geqslant \|R\|_* - \|U - P_T(\mathcal{Y})\|_{\mathrm{Fr}} \|P_T(Z)\|_{\mathrm{Fr}} \\
&\qquad + \|P_{T^\top}(Z)\|_* \left(1 - \|P_T(\mathcal{Y})\|\right). \qquad (15)
\end{aligned}$$

As long as $\|P_T(\mathcal{Y})\| < 1$, $\|U - P_T(\mathcal{Y})\|_{\mathrm{Fr}}$ is small enough and $\|P_T(Z)\|_*$ is not too large in relation to $\|P_{T^\top}(Z)\|_*$, the solution will thus be unique.

In (Xu, Jin, and Zhou 2013) these ideas are extended to the case where side information matrices X, Y *with orthonormal columns* is provided. The key here is that with this assumption on the columns, $\|XMY^\top\|_* = \|M\|_*$ for any matrix M, so that most of the arguments above still hold with minor modification, even after replacing the projection operator P_T by its inductive analogue $P_T(Z) = P_X Z P_F + P_E Z P_B - P_E Z P_F$.

Removing the Homogeneity Assumption: Proof Strategy

In our case, where X, Y are arbitrary (they can without loss of generality be assumed to have orthogonal columns, though not necessarily of norm 1), it is no longer true that $\|XMY^\top\|_* = \|M\|_*$ for any M. To tackle this issue, we define a norm $\|Z\|_{\mathcal{I},*}$ on the set of matrices $\mathbb{R}^{m \times n}$ which equals the minimum possible nuclear norm of a matrix M such that $XMY^\top = Z$:

$$\|Z\|_{\mathcal{I},*} = \min \left(\|M\|_* \quad : XMY^\top = Z \right). \qquad (16)$$

A key observation is that both this norm and *its dual* can be computed easily. Indeed, it is easy to see that $\|Z\|_{\mathcal{I},*} = \Sigma_1^2 X^\top Z Y \Sigma_2^2$ where Σ_1, Σ_2 are matrices containing the singular values of X, Y. Furthermore, we also show in the supplementary that in fact the dual norm $\|\cdot\|_{\mathcal{I},\sigma}$ is simply the spectral norm of the matrix $X^\top R Y$. These modifications mean that during the proof, we must manipulate 5 different norms ($\|\cdot\|, \|\cdot\|_*, \|\cdot\|_{\mathcal{I},\sigma}, \|\cdot\|_{\mathcal{I},*}$ and $\|\cdot\|_{\mathrm{Fr}}$), sometimes incurring factors of the smallest singular value σ_0 of X, Y.

We note that removing the homogeneity assumption has consequences in the proofs, including the need for a stronger incoherence assumption.

Fast Decay in Low-Noise Settings: Proof Strategy

In addition, we need to account for the noise, thus instead of perturbing the matrix R only by a matrix Z with $P_\Omega(Z) = 0$, we also perturb it by a matrix H with $P_{\Omega^\top}(H) = 0$ corresponding to the difference between the recovered matrix and the ground truth on the observed entries. Thus our recovered matrix, the solution to algorithm (2), $\widehat{R}$, can be written $\widehat{R} = R + H + Z$.

Our next step is to perform a perturbed version of the calculation in equation (15) taking into account the difference $H = P_\Omega(\widehat{R} - R)$. This is the calculation performed in the proof of Lemma C.1. As previously we write U for a subgradient of $\|R\|_{\mathcal{I},*}$ and $U^\top$ for a subgradient of $\|P_T(Z)\|_{\mathcal{I},*}$. We start by expressing $\|\widehat{R}\|_{\mathcal{I},*}$ as $\left\langle R + H + Z, U + U^\top \right\rangle$ and after some calculations we obtain the following conclusion:

$$\begin{aligned}
\|R\|_{\mathcal{I},*} &\geqslant \|\widehat{R}\|_{\mathcal{I},*} \\
&\geqslant \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} + \frac{1}{4}\|P_{T^\top}(Z)\|_{\mathcal{I},*}, \qquad (17)
\end{aligned}$$

which holds as long as several concentration phenomena occur (which will happen with high probability as long as N is large enough).

Our next step is to bound $\|H\|_{\mathcal{I},*}$. With high probability, the noisily observed entries of R on Ω (the $R_\xi + \zeta_\xi$) are close to the actual entries R, which in turn implies that the entries of H will not be too large (see the beginning of the proof of Theorem D.2).

This yields a bound of order $\widetilde{O}(\sqrt{N}\nu)$ for $\|H\|_{\mathcal{I},*}$, and then via equation (17), on $\|P_T(Z)\|_*$. Together with further modifications, this eventually yields a bound on the nuclear norm of $Z + H = \widehat{R} - R$. This means that our perturbed version of the exact recovery results places the recovered matrix $\widehat{R}$ inside of the smaller function class of matrices within a bounded spectral norm of the ground truth matrix. At this point, we can leverage classical results on the Rademacher complexity of the function class of matrices with bounded nuclear norm (see Lemma 5 below for the inductive version we use in practice) to further bound the generalization gap. Several further steps are needed to process the final result into an elegant formula that holds for any value of N. The details are in the supplementary material.

Lemma 5 (Chiang, Dhillon, and Hsieh 2018)**.** *The function class* $\left\{XMY^\top : \|M\|_* \leqslant \mathcal{M}\right\}$ *satisfies*

$$\mathfrak{R}(\mathcal{F}_\mathcal{M}) \leqslant \mathbf{x}\mathbf{y}\mathcal{M}\sqrt{\frac{1}{N}}, \qquad (18)$$

where $\mathbf{x} := \|X^\top\|_{2,\infty}$ *and* $\mathbf{y} := \|Y^\top\|_{2,\infty}$.

Proof. Follows directly from Theorem 1 in (Kakade, Sridharan, and Tewari 2009), together with the duality between the nuclear and spectral norms (Fazel, Hindi, and Boyd 2001). Cf. also (Chiang, Dhillon, and Hsieh 2018). □

Experiments

In this paper, we have posited that an accurate understanding of the sample complexity landscape of inductive matrix completion requires treating the noise component differently

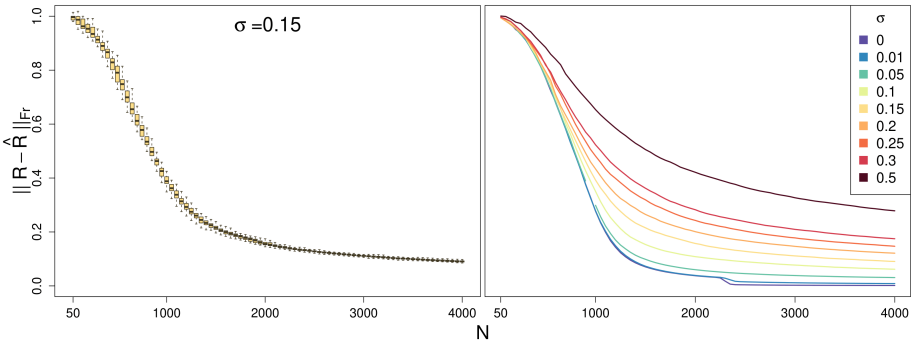

Figure 1: $\|\widehat{R} - R\|_{\mathrm{Fr}}$ as a function of N, σ

from the ground truth entries for the purposes of complexity. In this section we present the experiments we ran to confirm that a two-phase phenomenon as suggested by our bounds does in fact occur in practice.

We considered random matrices of size 100×100 and of rank 10[1], and created random orthonormal side information of rank 40, ensuring that the singular vectors of the ground truth matrix are in the span of the relevant side information, but with the orientation being otherwise uniformly random. The ground truth matrices were normalized to have Frobenius norm 100, and we then added i.i.d. $N(0, \sigma^2)$ gaussian noise to each observation. We performed classic inductive matrix completion (with the square loss) on the resulting training set, cross-validating the parameter λ on a validation set, and evaluated the RMSE distance between the resulting trained matrix and the ground truth. We performed this whole procedure for a wide range of different values for the number of samples N. For each value of N we perform the procedure on 40 different random matrix and side information.

The results are presented in Figure 1 below. The graph on the left contains box plots for our simulation with $\sigma = 0.15$ whilst the graph on the right presents our results, averaged over the 40 simulations, for several values of σ.

As can be observed in the figure, the graph of the error as a function of N is not convex, despite the fact that traditional approximate recovery bounds $\widetilde{O}(1/\sqrt{N})$ are convex. Instead, the graph looks like a sigmoid: we can clearly observe a thresholding phenomenon where the performance is very poor initially, but very quickly improves past a minimum number of entries. Furthermore, as can be expected, after the threshold is crossed the error decreases slowly (at least as $\widetilde{O}(\frac{\sigma}{\sqrt{N}})$ as per the bounds in Theorem 2 above), confirming that inductive matrix completion in low-noise settings exhibits a two-phase phenomenon matching our theoretical results. Furthermore, the fact that the post threshold error curve scales as σ is also apparent from the graphs.

[1]To generate such a random matrix, we generate matrices $U, V \in \mathbb{R}^{100\times 10}$ with i.i.d. gaussian entries, then we form the matrix $UV^\top$ and we normalize it to have Frobenius norm 100.

Conclusion and Future Directions

In this paper, we have studied *Inductive Matrix Completion* with nuclear norm regularisation in low-noise regimes. Our first contribution is an exact recovery result which generalizes the existing ones to the case where the side information is no longer assumed to be orthonormal, and to an arbitrary sampling regime (previously, the number of samples was required to be bounded *above* by $\widetilde{O}(abr)$). Our second contribution consists in generalization bounds composed of two components: (1) the requirement that the number of samples should exceed a given threshold and (2) a term of order $\widetilde{O}(\sigma\, \sigma_0^{-2} a^{3/2}\sqrt{b}\log^3(N/\Delta)\sqrt{\frac{1}{N}})$ (ignoring incoherence constants and other constant quantities), which is directly proportional to the subgaussianity constant σ of the noise. In particular, the result forms a bridge between exact recovery results and approximate recovery results: at the regimes where exact recovery is possible, the error converges to zero when the noise converges to zero.

We believe our result and proof strategy open the door to a new and unexplored direction of research. Possible future directions include improving the dependence on N from $1/\sqrt{N}$ to $\frac{1}{N}$, extending the results to non-trivially non uniform distributions or providing analogues of our results for other low-rank learning problems such as density estimation (Song et al. 2014; Vandermeulen and Ledent 2021; Kargas and Sidiropoulos 2019; Anandkumar et al. 2014; Vandermeulen 2020; Amiridi, Kargas, and Sidiropoulos 2020, 2021) or more complex recommender systems models that involve implicit feedback or graph/cluster information (Zhang and Chen 2020; Alves et al. 2020; Wu et al. 2021; Steck 2019; Vančura et al. 2022; Lin et al. 2022; Shen et al. 2021). Improving the dependence on a, b to match the scaling of the ERT is also a very ambitious and interesting aim.

Acknowledgements

Rodrigo Alves thanks Recombee for supporting his research. Marius Kloft acknowledges support by the Carl-Zeiss Foundation, the DFG awards KL 2698/2-1, KL 2698/5-1, KL 2698/6-1, and KL 2698/7-1, and the BMBF awards 01|S18051A, 03|B0770E, and 01|S21010C.

References

Aggarwal, C. C. 2016. *Recommender Systems: The Textbook*. Springer Publishing Company, Incorporated, 1st edition. ISBN 3319296574.

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 Pre-registration Experiment: An Alternative Publication Model For Machine Learning Research*.

Amiridi, M.; Kargas, N.; and Sidiropoulos, N. D. 2020. Low-rank Characteristic Tensor Density Estimation Part I: Foundations. *arXiv e-prints*, arXiv:2008.12315.

Amiridi, M.; Kargas, N.; and Sidiropoulos, N. D. 2021. Low-rank Characteristic Tensor Density Estimation Part II: Compression and Latent Density Estimation. *arXiv e-prints*, arXiv:2106.10591.

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.

Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local Rademacher complexities. *The Annals of Statistics*, 33(4): 1497 – 1537.

Bartlett, P. L.; and Mendelson, S. 2001. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. In Helmbold, D.; and Williamson, B., eds., *Computational Learning Theory*, 224–240. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-44581-4.

Cai, T. T.; and Zhou, W.-X. 2016. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1): 1493 – 1525.

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, H.; and Li, J. 2017. Learning Multiple Similarities of Users and Items in Recommender Systems. In *2017 IEEE International Conference on Data Mining (ICDM)*, 811–816.

Chen, T.; Zhang, W.; lu, Q.; Chen, K.; Zheng, Z.; and Yu, Y. 2012. SVDFeature: A Toolkit for Feature-based Collaborative Filtering. *The Journal of Machine Learning Research*.

Chen, Y. 2015. Incoherence-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, 61(5): 2909–2923.

Chen, Y.; Chi, Y.; Fan, J.; Ma, C.; and Yan, Y. 2020. 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.

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

Giménez-Febrer, P.; Pagès-Zamora, A.; and Giannakis, G. B. 2020. Generalization Error Bounds for Kernel Matrix Completion and Extrapolation. *IEEE Signal Processing Letters*, 27: 326–330.

Gross, D.; Liu, Y.-K.; Flammia, S. T.; Becker, S.; and Eisert, J. 2010. Quantum State Tomography via Compressed Sensing. *Phys. Rev. Lett.*, 105: 150401.

Herbster, M.; Pasteris, S.; and Tse, L. 2019. Online Matrix Completion with Side Information. *CoRR*, abs/1906.07255.

Jain, P.; and Dhillon, I. S. 2013. Provable Inductive Matrix Completion. *CoRR*, abs/1306.0626.

Kakade, S. M.; Sridharan, K.; and Tewari, A. 2009. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 21*, 793–800. Curran Associates, Inc.

Kargas, N.; and Sidiropoulos, N. D. 2019. Learning Mixtures of Smooth Product Distributions: Identifiability and Algorithm. In Chaudhuri, K.; and Sugiyama, M., eds., *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, 388–396. PMLR.

Keshavan, R.; Montanari, A.; and Oh, S. 2009. Matrix Completion from Noisy Entries. In Bengio, Y.; Schuurmans, D.; Lafferty, J. D.; Williams, C. K. I.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 22*, 952–960. Curran Associates, Inc.

Kueng, R.; Rauhut, H.; and Terstiege, U. 2017. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1): 88–116.

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.; 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.

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.

Lin, W.-Y.; Liu, S.; Ren, C.; Cheung, N.-M.; Li, H.; and Matsushita, Y. 2022. Shell Theory: A Statistical Model of Reality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6438–6453.

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

Menon, A. K.; and Elkan, C. 2011. Link Prediction via Matrix Factorization. In *Machine Learning and Knowledge Discovery in Databases*, 437–452. Springer Berlin Heidelberg.

Recht, B. 2011. A Simpler Approach to Matrix Completion. *J. Mach. Learn. Res.*, 12(null): 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.

Shen, W.; Zhang, C.; Tian, Y.; Zeng, L.; He, X.; Dou, W.; and Xu, X. 2021. Inductive Matrix Completion Using Graph Autoencoder. *CoRR*, abs/2108.11124.

Song, L.; Anandkumar, A.; Dai, B.; and Xie, B. 2014. Nonparametric Estimation of Multi-View Latent Variable Models. In Xing, E. P.; and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 640–648. Bejing, China: PMLR.

Steck, H. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference*, WWW '19, 3251–3257. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Tanner, J.; Thompson, A.; and Vary, S. 2019. Matrix Rigidity and the Ill-Posedness of Robust PCA and Matrix Completion. *SIAM Journal on Mathematics of Data Science*, 1(3): 537–554.

Vančura, V.; Alves, R.; Kasalickỳ, P.; and Kordík, P. 2022. Scalable Linear Shallow Autoencoder for Collaborative Filtering. In *Proceedings of the 16th ACM Conference on Recommender Systems*, 604–609.

Vandermeulen, R. A. 2020. Improving nonparametric density estimation with tensor decompositions. *arXiv preprint arXiv:2010.02425*.

Vandermeulen, R. A.; and Ledent, A. 2021. Beyond Smoothness: Incorporating Low-Rank Analysis into Nonparametric Density Estimation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 12180–12193. Curran Associates, Inc.

Wang, J.; Wong, R. K. W.; Mao, X.; and Chan, K. C. G. 2021. Matrix Completion with Model-free Weighting. arXiv:2106.05850.

Wu, Q.; Zhang, H.; Gao, X.; Yan, J.; and Zha, H. 2021. Towards open-world recommendation: An inductive model-based collaborative filtering approach. In *International Conference on Machine Learning*, 11329–11339. PMLR.

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.

Yao, Q.; and Kwok, J. T. 2019. Accelerated and Inexact Soft-Impute for Large-Scale Matrix and Tensor Completion. *IEEE Transactions on Knowledge and Data Engineering*, 31(9): 1665–1679.

Zhang, M.; and Chen, Y. 2020. Inductive Matrix Completion Based on Graph Neural Networks. In *International Conference on Learning Representations*.

Zhang, X.; Du, S.; and Gu, Q. 2018. Fast and Sample Efficient Inductive Matrix Completion via Multi-Phase Procrustes Flow. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 5756–5765. Stockholmsmässan, Stockholm Sweden: PMLR.

Zhong, K.; Jain, P.; and Dhillon, I. S. 2015. Efficient Matrix Sensing Using Rank-1 Gaussian Measurements. In Chaudhuri, K.; GENTILE, C.; and Zilles, S., eds., *Algorithmic Learning Theory*, 3–18. Cham: Springer International Publishing. ISBN 978-3-319-24486-0.