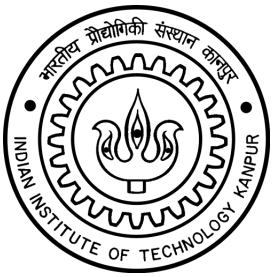
Matrix Completion with Implicit Clustering

Md Enayat Ullah, 12817407

November 4, 2016

Post-graduate Project



Supervisors:

Dr. Purushottam Kar (CSE, IIT Kanpur)
Dr. Debasis Kundu (MTH, IIT Kanpur)
Dr. Prateek Jain (MSR, India)

Contents

List of Figures

1 Introduction

Matrix Completion is a problem wherein given a partially observed matrix, the goal is to fill in the missing entries i.e. to complete the matrix. In all its generality, this is a very ill-posed problem with arbitrary number of solutions. The most common assumption made is that the matrix is low rank, i.e. the rank is a constant which is very small compared to the dimensions of the matrix. This setting is still an NP-Hard problem, and thus computationally intractable. The assumptions which make the problem well-posed is Incoherence (See Incoherence definition) which precisely says that the matrix is not sparse. Moreover, we assume that the observed entries are drawn independently and uniformly. This is important so that the adversary is not given the power to hide the entries which are crucial and reveal only redundant entries.

Matrix Completion finds a direct application in recommendation systems, wherein we have an incomplete user-item ratings matrix, and the goal is to complete them. The matrix is mostly incomplete, as a user rates only a tiny number of items from a plethora of items. The assumption of low rank is motivated here by the conjecture that users behave in a small restricted number of ways (rank). Therefore each of the rating pattern is essentially a linear combination of the permissible ways. This low rank restriction enforces that if two users rating on some items appear similar, their other ratings will be similar as well. This concept is known as Collaborative Filtering, and is an implicit consequence of Low rank Matrix Completion. The Collaborative Filtering aspect of Matrix Completion is used crucially in Recommendation Systems, and was also part of one of winning entries of the infamous Netflix Challenge (Ell08).

In this work, we consider the problem of Matrix Completion, wherein given a few uniformly and independently sampled entries of a matrix, the objective is to recover the complete matrix. The problem becomes well-posed under standard assumptions like low rank and Incoherence. On top of this, we impose more tha usual structural constraints on the matrix. In particular, we assume that the matrix can be partitioned into (vertically) disjoint r rank-k components. Under such a partitioning, the matrix allows a low rank decomposition ($M = UV^T$), and the corresponding low rank matrix V exhibits block sparsity. Thus, the Matrix Completion problem reduces to one resembling Dictionary Learning (with an incomplete matrix) in the under-complete setting, and the sparse-representations being block-sparse (instead of column sparse). Solving the problem not only gives us the estimates of missing entries, it also gives as an implicit grouping or clustering over items. This is because we get r clusters and the items can belong to one (or more) of these clusters as determined by the non-zero entries in the item latent representations.

2 Motivation

Matrix Completion primarily finds its application in Recommendation systems, wherein an incomplete Users-Items Rating matrix is completed under the low-rank assumption.The missing ratings are then calculated as the inner product of the latent user representation and latent item representation. This model, although very simple, is very restrictive, and does not take into account a variety of factors that come into play like different types of items may cater differently to the different categories of users and thus influence the rating pattern.

At a high level, we motivate the modification with the conjecture that in a User-Items rating matrix, the feedback of users towards different items may depend on the category it belongs to. For example: user-movies rating matrix can induce can partition based on different movie categories (genres). Therefore, if two users rate some items(movies) almost identically, and the items belong to the same category, then their ratings for the other items in that category will be similar. We therefore solve the problem of categorizing the items and completing the matrix. The factorization thus will not only give us the user-item ratings missing entries, but will also give us r clusters over items.

3 Related Work

Candes and Tao's seminal work on Compressive sensing showed that it is possible to efficiently reconstruct a signal by finding solutions to under-determined linear systems, majorly by exploiting the sparsity of the signal (CT06). They gave the Restricted Isometry Property, which crucially helped solved related problems as well. The work on recovering vectors(signals) was further was extended to matrices as Matrix Sensing, wherein given linear measurements of an unknown matrix with some known matrices, it is possible to recover the matrix. A special case of Matrix Sensing is Matrix Completion in which the sensing matrices are not random but very sparse. Candes and Tao gave a convex relaxation based method to solve the Matrix Completion problem (CR09).

Apart from the works which use Convex relaxation, Keshavan et al gave a non-convex formulation of Matrix Completion, solved by an Alternating Minimization algorithm (KOM09). Jain et al further showed that it is possible to get to the global minima of this non-convex objective of both Matrix Sensing and Matrix Completion given that we have a good initialization (JNS13). They gave a method to initialize as well as the analysis that the algorithm indeed converges to the global minima. A related problem is Dictionary learning, wherein given a set of vectors, the goal is to learn a basis(dictionary) in which the vectors have sparse representations. We discuss this problem, because our modified Matrix Completion essentially reduces to Dictionary learning. Methods for Dictionary learning mostly comprise of heuristics like K-SVD, Method of Optimal directions(MOD) etc. Recently, there are works which give provable guarantees

with mostly standard set of constraints and assumptions. (SWW12) analyzed this in the noiseless setting and where the Dictionary is a basis. (AAJ+14) and (AGM14) independently gave algorithms and the analysis for the over-complete case. Moreover, (AGM14) provided a neural framework for Dictionary learning which crucially helps the analysis.

4 Problem Formulation

4.1 Matrix Sensing

Matrix Sensing is a problem, in which the goal is recover an unknown matrix and we are only allowed to take a few of its linear measurements with random matrices. Formally put: given A_i's ($\in \mathcal{R}^{m\times n}$) and b_i's ($\in \mathcal{R}$) such that:

$$\langle A_i, X\rangle = b_i \forall\ i\ \ \in [l]$$

Find a low rank matrix X. We define a linear operator $\mathcal{A} : \mathcal{R}^{m\times n} \to \mathcal{R}^l$ which encapsulates all A_i's such that $\mathcal{A}(X) = b$.
The non-convex optimization problem is formulated as:

$$\min_{U,V} \left\|\mathcal{A}(UV^T) - b\right\|_2^2$$
$$\text{where } U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}$$

The problem becomes well-posed when the matrices A_i's are constrained to satisfy the Restricted Isometry Property.

4.2 Generic Matrix Completion

We first look at the generic Matrix Completion problem setting. Matrix Completion can be thought of as a special case of Matrix Sensing, wherein the sensing matrices are all zero expect for one index which is observed.

$$A_i = e_j e_k^T \text{ where } (j,k) \in \text{Observed Set}$$

However, the usual Restricted Isometry property does not quite work here, because the matrix could be very sparse and the chances of observing the few non-zero entries could be very low. We therefore impose another property on the matrix, which is known as Incoherence, which further makes the problem well-posed. Moreover, the observed entries are constrained to be sampled uniformly and independently.

Consider a partially observed matrix $M \in \mathbb{R}^{m\times n}$, and a set Ω such that $\Omega = \{(i,j) : M_{ij} \text{ is observed}\}$. Also, we define a projection of a matrix on a subset

Ω as $P_\Omega(M)$:

$$P_\Omega(M)_{ij} = \begin{cases} M_{ij} & \forall\,(i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

The non-convex way of Matrix Completion essentially solves the following optimization problem:

$$\min_{U,V} \left\| P_\Omega\left(UV^T\right) - P_\Omega(M) \right\|_{\mathbb{F}}^2$$
$$\text{where } U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}$$

Theorem 1. Let $M = U^*\Sigma^*{V^*}^T (n \geq m) \in \mathcal{R}^{m\times n}$ be a rank-r μ-incoherent matrix (See Incoherence definition). Also, let each entry of M be observed uniformly and independently with probability

$$p > C \frac{\left(\frac{\sigma_1^*}{\sigma_r^*}\right)^2 \mu^4 r^{2.5} \log n \log\left(\frac{r\,\|M\|_{\mathcal{F}}}{\epsilon}\right)}{m\delta_{2r}^2}$$

where $\delta_{2r} \leq \frac{\sigma_r^*}{12r\sigma_1^*}$ and $C > 0$ is a global constant. Then with hight probability, for $T = C' \log \frac{\|M\|_{\mathcal{F}}}{\epsilon}$, the outputs $\widehat{U}_T$ and $\widehat{V}_T$ with input $(\Omega, P_\Omega(M))$ satisfy $\left\| M - \widehat{U}\widehat{V}^T \right\|^{\mathcal{F}} \leq \epsilon$.

Proof. See (JNS13)

4.3 Modified Matrix Completion

We now impose an additional structural constraint on the matrix M which says that the matrix M can be partitioned into r (vertically) disjoint components, each being of rank k. Note that this is not equivalent to the rank-k constraint on the matrix, because rank k doesn't imply that there exists a partition into k rank 1 sub-matrix (the other way implications holds).

Without loss of generality, we assume that the k partitions are equal sized and are arranged contiguously, and let s be the number of columns in each component, so we have $s = n/r$. Let us denote the components as $M_i \in \mathbb{R}^{m\times n}$ for $i \in [r]$. So we have

$$M_i = U_i V_i^T, \text{ where } U_i \in \mathbb{R}^{m\times r} \text{ and } V_i \in \mathbb{R}^{n\times r}\ \forall\ i \in [r]$$
$$M = \sum_{i=1}^{r} M_i = \sum_{i=1}^{r} U_i V_i^T$$

The following figures present a toy example to graphically illustrate the problem. Figure 1 shows the two($r = 2$) rank k components of M. Figure 2 and 3 shows

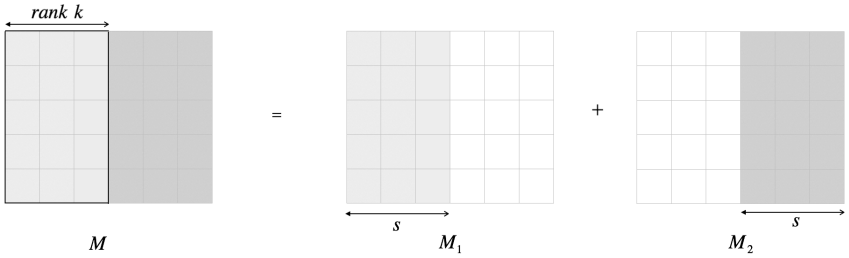


Figure 1: Two rank k components of M

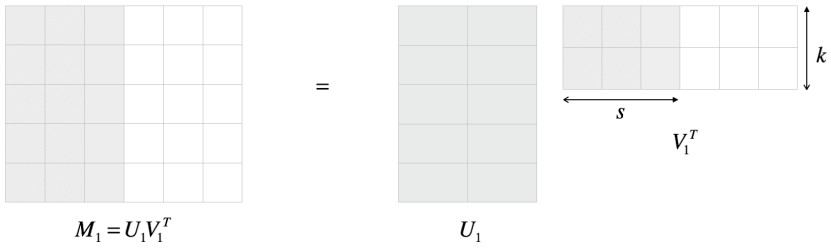


Figure 2: Matrix Factorization of M_1

the low rank decomposition of M_1 and M_2. Figure 4 sums it up to show the low rank decomposition of M and the block sparsity of V is evident here.

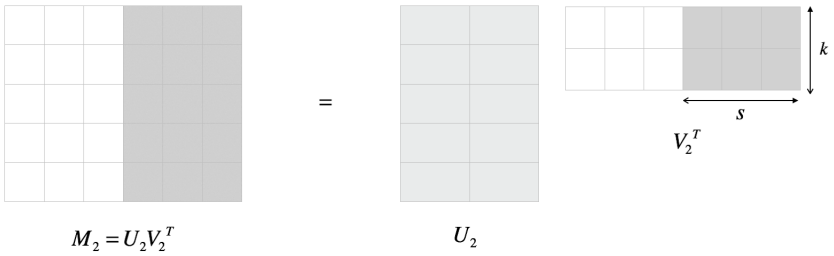


Figure 3: Matrix Factorization of M_2

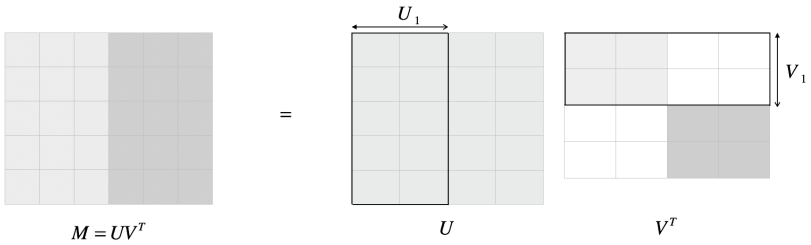


Figure 4: Matrix Completion with r rank-k components

Information theoretically, the matrix M has $(mr + nk)$ bits of information. We therefore conjecture that this case requires $\mathcal{O}((mr + nk)$ entries to recover the matrix, unlike the generic case with requires $\mathcal{O}(m + n)r$ entries ($k \ll r$). Moreover, we also get an implicit cluster over items (V) in this case.

5 Mathematical Preliminaries

5.1 Restricted Isometry Property

Definition 1. A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy *k-RIP* with RIP constant δ_k if $\forall\ X \in \mathbb{R}^n$ with $\|X\|_0 \leq k$, the following hold:

$$(1 - \delta_k) \|X\|_2 \leq \|AX\|_2 \leq (1 + \delta_k) \|X\|_2$$

Random matrices with iid entries and with sufficiently many measurements $\left(\Theta\left(\frac{kn \log n}{\delta_k^2}\right)\right)$ satisfy the RIP property with a high probability.

5.2 Incoherence

Definition 2. A matrix $M \in \mathcal{R}^{m \times n}$ is incoherent with parameter μ if

$$\left\|U^i\right\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{m}}\ \forall\ i \in [m]$$

$$\left\|V^j\right\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{n}} = \frac{\mu}{\sqrt{s}}\ \forall\ j \in [n]$$

where $M = U\Sigma V$ is the SVD of M and U^i and V^j are the i^{th} and j^{th} row of U and V respectively.

6 Algorithm

We use Alternating Minimization to solve this problem. In order to show convergence to the global optima, we first need a good initialization which is *close* to the true solution. We do a thin SVD to get an initialization for U (say U^0). However, the same would not have worked for V, because SVD gives a dense matrix for V, but since we already know that there's a richer inherent structure in the matrix M, we want to exploit it to give us better representations (block sparse diagonal matrix) for V. Fortunately, we don't need to initialize both U and V as we use alternating minimization to solve for the other.

The sparsity constraint in V reduces this to the problem of Dictionary Learning. In particular, this is essentially a Dictionary Learning problem in an undercomplete (basis) setting, and the sparse representations here are block sparse as opposed to the usual column sparse. We therefore solve for V employing Sparse Recovery techniques in the alternating minimization paradigm. Algorithm 1 gives the Alternating Minimization based algorithm for Matrix Completion.

Algorithm 1: Alternating Minimization for Matrix Completion

1: Initialize $U^0 = \text{LSVD}\left(P_\Omega\left(M\right), r\right)$
2: **for** $t = 1, 2, \ldots$ **do**
3: $\quad \widehat{V}^{t+1} = \arg\min_V \left\| P_\Omega\left(\widehat{U}^t V^T - M\right)\right\|_{\mathbb{F}}^2$ s.t V is block-sparse
$\quad$ //Sparse Recovery
4: $\quad \widehat{U}^{t+1} = \arg\min_U \left\| P_\Omega\left(U\left(\widehat{V}^{t+1}\right)^T - M\right)\right\|_{\mathbb{F}}^2$
$\quad$ //Least-squared Estimator
5: **end for**

7 Rank-1 Case

For simplicity, we look at the case in which the rank of each component is 1, i.e $k = 1$. We ignore the matrix of singular values Σ for now. We have:

$$M = UV^T$$

where each row column of V contains s non-zero entries. We assume that the columns of U are orthogonal. Columns of V are orthogonal by construction. We have the usual incoherence constraint of M.

The sample complexity conjecture is that we need $\mathcal{O}(mr+n)$ samples to recover the matrix. To prove the same, we will show that following:

$$\left\| \frac{1}{p^2} P_\Omega(M) P_\Omega(M)^T - MM^T \right\|_2 \leq \delta$$

when $p = \mathcal{O}\left(\frac{mr+n}{mn}\right)$, where p is the probability with with each entry of the matrix is observed (uniformly and independently), and U^1 is the first left singular vector of $P_\Omega(M)$. Algorithm 2 gives the complete algorithm for matrix completion with rank-1 components.

Algorithm 2: Alternating Minimization for Matrix Completion: Rank-1

Initialize $U^0 = \text{LSVD}\left(P_\Omega\left(M\right), r\right)$
for $t = 1, 2, \ldots$ **do**
  for $i = 1, 2, \ldots r$ **do**
    $\widehat{V}^{t+1} = \mathbf{0}$
    $x_i = \dfrac{\sum_{j:(i,j)\in\Omega} M_{ij} U_j^i}{\sum_{j:(i,j)\in\Omega} U_j^i U_j^i}$
    $\widehat{i} = \min_i \sum_{j:(i,j)\in\Omega} \left(M_{ij} - x_i U_j^i\right)^2$
    $\widehat{V}_i^{t+1}(\ \widehat{i}\ ) = x_{\widehat{i}}$
  end for
  $\widehat{V}^{t+1}$ =normalizeColumns($\widehat{V}^{t+1}$)
  for $i = 1, 2, r$ **do**
    $\widehat{U}_i^{t+1} = \left(\sum_{i:(i,j)\in\Omega} \left(\widehat{V}_i^t\right)^T \widehat{V}_i^t\right) \left(\sum_{i:(i,j)\in\Omega} M_{ij} \widehat{V}_i^t\right)$
  end for
  $\widehat{U}^{t+1}$ =GramSchmidt($\widehat{U}^{t+1}$)
end for

8 Experiments

We generate synthetic datasets wherein the vectors U_i and V_i are drawn from multivariate normal distribition and orthonormalized. The details of the generative model is given below:

$$
\begin{aligned}
&A_i \sim \mathcal{N}\left(\mathbf{0}, \mathcal{I}_m\right) \forall\ i \in\ [r] \\
&U = \text{GramSchmidtOrthonormalization}(A) \\
&B_i \sim \mathcal{N}\left(\mathbf{0}, \mathcal{I}_s\right) \forall\ i \in\ [r] \\
&V_i = [\underbrace{0 \ldots 0}_{\text{(i-1)s}}\ B_i\ \underbrace{0 \ldots 0}_{\text{(r-i-1)s}}]^T\ i \in\ [r] \\
&V = \text{NormalizeColumns}(V) \\
&M = UV^T
\end{aligned}
$$

Generating U and V as above gives us a good enough incoherence which ensures that the algorithm works well. We then sample observed entries set $Omega$ uniformly and independently, wherein the size of the set taken as:

$$|\Omega| = c(mr + n)$$

The ideal value of c according to our hypothesis is 1, because this$(mr+n)$ is essentially the number bits present in such a matrix, and information theoretically

these number bits should be sufficient to recover the complete matrix. However, the number of entries needed to convergence to a global minima strongly depends on the the sampled observed entries, and may not always reveal crucial information which suffices to complete the matrix. For example: if we do not observe any entry of a particular row or column of M, we cannot recover the true entries of that column or row. In the experiments, we try out different values of (m,n), r and c, and observe the Frobenius norm error, defined as:

$$\text{Error} = \left\| M - \widehat{U}\widehat{V}^T \right\|_{\mathcal{F}}$$

We have three plots below.In each of them, two parameters (out of (m,n), r and c) are fixed, and the third is varied. Moreover, in the same plot, we do this for different sets of the fixed parameters.

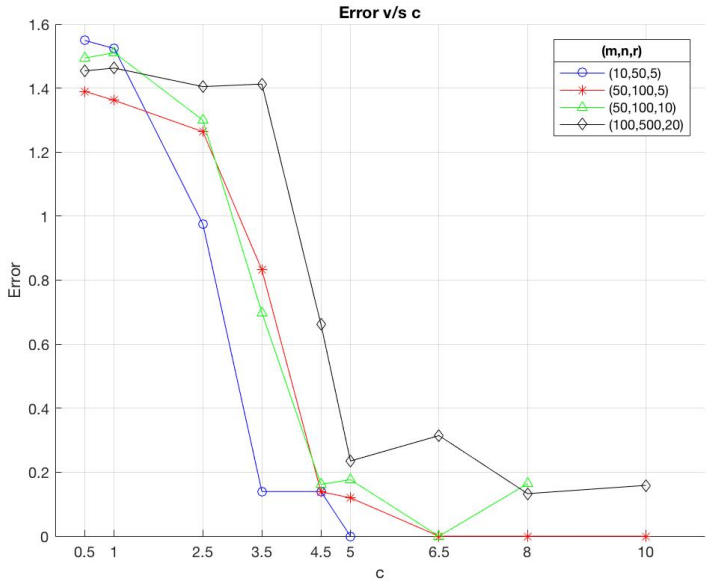


Figure 5: Error vs. varying c

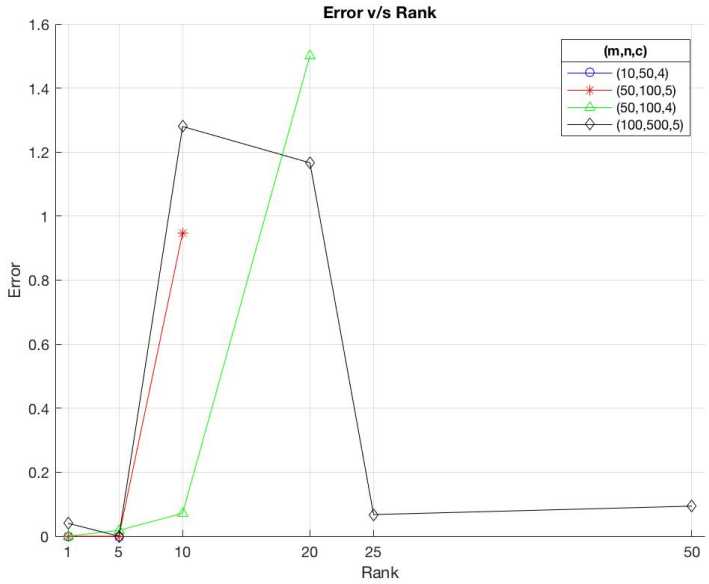


Figure 6: Error vs. varying rank

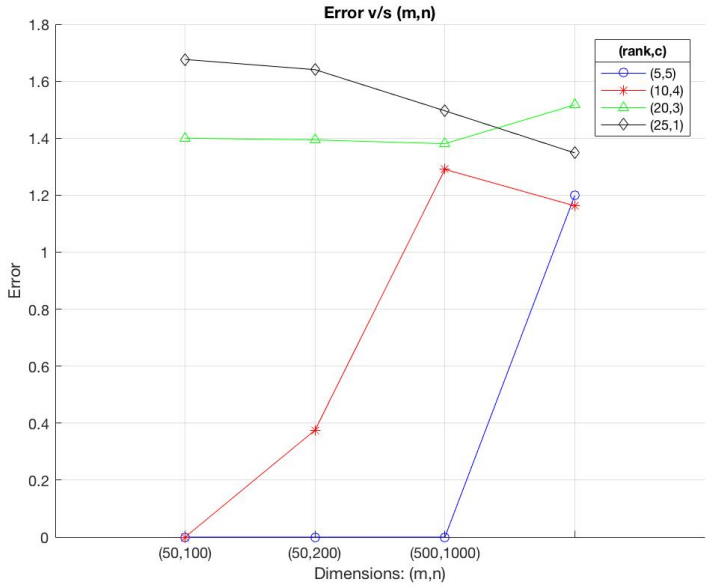


Figure 7: Error vs. varying dimensions of matrix

9 Conclusion

In this work, we looked at the Matrix Completion problem, with subtle modifications. We impose sparsity on the latent Items matrix (V). This in turn gives us an implicit way to cluster items, based on user-pattern ratings. The problem reduces to one resembling Dictionary Learning, which we solve via Alternating Minimization using sparse recovery techniques. The experiments validate our conjecture regarding the number of samples required to recover the matrix. What we need to do now is extend it to the case where we have components of rank-k, and do a theoretical analysis establishing sample complexity and convergence guarantees. The other thing is to relax the equal-sized components constraint. A further extension is to impose sparsity on the users matrix as well, which would in turn give a clustering over the users and items both.

References

[AAJ+14] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *COLT*, pages 123–137, 2014.

[AGM14] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *COLT*, pages 779–806, 2014.

[CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[CT06] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[Ell08] Jordan Ellenberg. The netflix challenge. *WIRED-SAN FRANCISCO-*, 16(3):114, 2008.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[KOM09] Raghunandan H Keshavan, Sewoong Oh, and Andrea Montanari. Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pages 324–328. IEEE, 2009.

[SWW12] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *COLT*, pages 37–1, 2012.