Low-Rank Matrix Recovery from Row-and-Column Affine Measurements

Abstract

We propose and study a row-and-column affine measurement scheme for low-rank matrix recovery. Each measurement is a linear combination of elements in one row or one column of a matrix X. This setting arises naturally in applications from different domains. However, current algorithms developed for standard matrix recovery problems do not perform well in our case, hence the need for developing new algorithms and theory for our problem. We propose a simple algorithm for the problem based on Singular Value Decomposition (SVD) and least-squares (LS), which we term SVLS. We prove that (a simplified version of) our algorithm can recover X exactly with the minimum possible number of measurements in the noiseless case. In the general noisy case, we prove performance guarantees on the reconstruction accuracy under the Frobenius norm. In simulations, our row-and-column design and SVLS algorithm show improved speed, and comparable and in some cases better accuracy compared to standard measurements designs and algorithms. Our theoretical and experimental results suggest that the proposed row-and-column affine measurements scheme, together with our recovery algorithm, may provide a powerful framework for affine matrix reconstruction.

1. Introduction

In the low-rank affine matrix recovery problem, an unknown matrix $X \in \mathbb{R}_{n_1 \times n_2}$ with $rank(X) = r$ is measured indirectly via an affine transformation $\mathcal{A} : \mathbb{R}_{n_1 \times n_2} \to \mathbb{R}^d$ and possibly with additive (typically Gaussian) noise $z \in \mathbb{R}^d$. Our goal is to recover X from the vector of noisy measurements $\mathbf{b} = \mathcal{A}(X) + z$. The problem has found numerous applications throughout science and engineering, in different fields such as collaborative filtering(Koren et al., 2009), face recognition (Basri & Jacobs, 2003), quantum state tomography (Gross et al., 2010) and

computational biology (Chi et al., 2013). The problem has been studied mathematically quite extensively in the last few years. Most attention thus far has been given to two particular ensembles of random transformations $\mathcal{A}$: (i) the Matrix Completion (MC) setting, in which each element of $\mathcal{A}(X)$ is a single entry of the matrix where the subset of the observed measurements sampled uniformly at random (Candès & Recht, 2009; Candès & Plan, 2010; Candès & Tao, 2010; Keshavan et al., 2009; 2010; Recht, 2011) (ii) Gaussian-Ensemble (GE) affine-matrix-recovery, in which each element of $\mathcal{A}(X)$ is a weighted sum of all elements of X with i.i.d. Gaussian weights (Candès & Plan, 2011; Recht et al., 2010). Remarkably, although the recovery problem is in general NP-hard, when $r \ll min(n_1, n_2)$ and under certain conditions on the matrix X or the measurements or the measurements operator $\mathcal{A}$ one can recover X from $d \ll n_1 n_2$ measurements with high probability and using efficient algorithms (Candès & Recht, 2009; Recht et al., 2010; Candès & Tao, 2010; Recht, 2011). However, it is desirable to study the problem with other affine transformations $\mathcal{A}$ beyond the two ensembles mentioned above for the following reasons: (i) In some applications we cannot control the measurements operator $\mathcal{A}$, and different models for the measurements may be needed to allow a realistic analysis of the problem (ii) When we can control and design the measurements operator $\mathcal{A}$, other matrices may outperform the two mentioned above with respect to different resources such as number of measurements required, computation time and storage. The main goal of this paper is to present and study a different set of affine transformations, which we term row-and-column affine measurements. This setting may arise naturally in many applications, since it is often natural and possibly cheap to measure a single row or column of a matrix, or a linear combination of a few such rows and columns. For example, (i) In collaborative filtering, we may wish to revoker a users-items preference matrix and have access to only a subset of the users, but can observe their preference scores for *all* items (ii) When recovering a protein-RNA interactions matrix in molecular biology, a single experiment may simultaneously measure the interactions of a specific protein with all RNA molecules (Chu et al., 2011).

In our row and column framework the measurement operator $\mathcal{A}$ is represented by two matrices $A^{(R)}, A^{(C)}$ which multiply X from left and right, respectively. We focus on

two important ensembles of $A^{(R)}, A^{(C)}$: (i) Matrix Completion from single Columns and Rows (RCMC) where we observe single entry measurements in similar to standard matrix completion case, but the measured entries are not scattered randomly along the matrix, but rather we pick at random a few rows and a few columns, and measure *all* entries in these rows and columns. This ensemble is implemented by setting the rows (columns) of $A^{(R)}$ $(A^{(C)})$ as random vectors from the standard basis. (ii) Gaussian Row-and-Column (GRC) measurements. Here each set of measurements is a weighted linear combination of the matrix's rows (or columns) with the weights taken as i.i.d. Gaussians. This ensemble is implemented by setting the entries of $A^{(R)}, A^{(C)}$ as i.i.d. Gaussian random variables.

The measurement operator $\mathcal{A}$ in our model does not satisfy the standard requirements used for GE and MC. As a result, algorithms such as nuclear norm minimization (Recht et al., 2010; Candès & Recht, 2009) fail for our case, and different algorithms and theory are required. However, the specific algebraic structure provided by the row-and-column measurements, allow as to both derive efficient and simple algorithms, as well as to analyze their performance.

1.1. Prior Work

Before giving a detailed derivation and analysis of our design and algorithms, we give here an overview of the existing designs and their properties. We concentrate here on two properties: (i) storage required to represent the measurement operator , and (ii) measurement sparsity, defined as the sum over all measurements of the number of matrix entries participating in each measurement. The latter property may be related to measurement time.

Recently Cai and Zhang proposed a new design of rank one projection (Cai et al., 2015) where each measurements is of the form $\alpha^T X \beta$ where $\alpha \in \mathbb{R}^{n_1}$ and $\beta \in \mathbb{R}^{n_2}$ have i.i.d standard Gaussian entries, and proved that nuclear norm minimization can recover X with high probability. This is the first model deviating from MC and GE we are aware of. This model is different from our row-and-column model, as each measurement is obtained by multiplying X from both sides, whereas in our model each measurement is obtained by multiplying X from either left or right. Moreover, in our model the measurements are not chosen independently from each other but come in groups of size n_1 or n_2 (corresponding to rows or columns $A^{(R)}, A^{(C)}$). An advantage of rank one projection is that it leads to a significance reduction in measurement storage needed for $\mathcal{A}$ with overall $O(dn_1 + dn_2)$ storage space. However, each measurement is still dense and involve all matrix elements, hence measurement sparsity is $O(dn_1n_2)$. In contrast, our GRC model requires only $O(d)$ storage for $\mathcal{A}$, and every measurements depends only on $O(max(n_1, n_2))$ ele-

ments, leading to a reduced overall time for all measurements $O(dn_1 + dn_2)$. For RCMC, we need only $O(\frac{d\log(n)}{n})$ storage for $\mathcal{A}$, and measurement sparsity $O(d)$.

In the Gaussian Ensemble model, we can look at A in matrix representation $\mathcal{A}(X) = Avec(X)$. If $A \overset{i.i.d.}{\sim} N(0,1)$ one can recover low rank matrix X with $O(rn_1 + rn_2)$ noiseless measurements using nuclear norm minimization (Recht et al., 2010; Candès & Plan, 2011) or other methods such as Singular Value Projection (SVP) (Jain et al., 2010), which is optimal up to constants. Recovery in this model is robust to noise, with only a small increase in measurements. The main disadvantage of this model is that design requires $O(dn_1n_2)$ storage space for r A, which could be problematic for large matrices. Another possible disadvantage of this method is that measurements are dense - each measurement represents a linear combination of all $O(n_1n_2)$ matrix entries, and the time required to calculate $\mathcal{A}(X)$ is in general $O(dn_1n_2)$, which could be problematic for large n_1, n_2.

In the standard matrix completion problem (Candès & Recht, 2009) where we can recover X from single entries in X chosen uniformly at random using nuclear norm minimization (Cai et al., 2010; Toh & Yun, 2010; Candès & Tao, 2010; Ma et al., 2011; Recht, 2011) or using other methods such as SVD and gradient descent (Keshavan et al., 2009; 2010). This model has the lowest storage requirements $O(d)$ and measurement sparsity $O(d)$. However, recovery guarantees for this model are quite weak: it requires some assumptions on X such as incoherent (2) and the number of measurements required for recovery of X is higher compared to the Gaussian Ensemble.

2. Preliminaries and Notations

We denote by $\mathbb{R}_{n_1 \times n_2}$ the space of matrices of size $n_1 \times n_2$, by $\mathcal{O}_{n_1 \times n_2}$ the space of matrices of size $n_1 \times n_2$ with orthonormal columns, and by $\mathcal{M}_{n_1 \times n_2}^{(r)}$ the space of matrices of size $n_1 \times n_2$ and rank $\leqslant r$.

We denote by $||\cdot||_F$, the matrix Frobenius norm, by $||\cdot||_*$ the nuclear norm, and by $||\cdot||_2$ the spectral norm. For a vector, $||\cdot||$ denotes the standard l_2 norm.

For $X \in \mathbb{R}_{n_1 \times n_2}$ we denote by $span(X)$ the subspace of $\mathbb{R}^{n_1}$ spanned by the columns of X and define P_X to be the orthogonal projection into $span(X)$.

For a matrix X we denote by $X_{i\bullet}$ the i-th row, by $X_{\bullet j}$ the j-th column and by X_{ij} the (i,j) element. For two sets of indices I, J, we denote by X_{IJ} the sub-matrix obtained by taking the rows with indices in I and columns with indices in J of X. We denote by $[k]$ the set of indices $1, .., k$. We denote by $vec(X)$ the (column) vector obtained by stacking all the columns of X on top of each other.

We use the notation $X \overset{i.i.d.}{\sim} G$ to denote a random matrix X with i.i.d. entries $X_{ij} \sim G$.

For a matrix $X \in \mathcal{M}_{n_1 \times n_2}^{(r)}$ let $X = U\Sigma V^T$ be the Singular Value Decomposition (SVD) of X where $U \in \mathcal{O}_{n_1 \times r}$, $V \in \mathcal{O}_{r \times n_2}$ and $\Sigma = diag(\sigma_1(X), ..., \sigma_r(X))$ with $\sigma_1(X) \geq \sigma_2(X).. \geq \sigma_r(X) > 0$ the (non-zero) singular values of X (we omit the zero singular values and their corresponding vectors from the decomposition). For a general matrix $X \in \mathbb{R}_{n_1 \times n_2}$ we denote by $X_{(r)}$ the top-r singular value decomposition of X, $X_{(r)} = U_{\bullet[r]}\Sigma_{[r][r]}V_{\bullet[r]}^T$.

Our model assumes two affine transformations applied to X, representing rows and columns, $B^{(C,0)} = XA^{(C)}$ and $B^{(R,0)} = A^{(R)}X$, achieved by multiplications with two matrices $A^{(R)} \in \mathbb{R}_{k^{(R)} \times n_1}$ and $A^{(C)} \in \mathbb{R}_{n_2 \times k^{(C)}}$. We observe noisy observations of these transformations, $B^{(R)}, B^{(C)}$ obtained by applying additive noise:

$$A^{(R)}X + Z^{(R)} = B^{(R)}; \quad XA^{(C)} + Z^{(C)} = B^{(C)} \quad (1)$$

where the total number of measurements is $d = k^{(R)}n_1 + n_2 k^{(C)}$, and $Z^{(R)} \in \mathbb{R}_{n_1 \times k^{(R)}}, Z^{(C)} \in \mathbb{R}_{k^{(C)} \times n_2}$ are two zero-mean noise matrices. Our goal is to recover X from the observed measurements $B^{(C)}$ and $B^{(R)}$. To achieve this goal, we define the loss function

$$\mathcal{F}(X) = ||A^{(R)}X - B^{(R)}||_F^2 + ||XA^{(C)} - B^{(C)}||_F^2 \quad (2)$$

and solve the least squares problem:

$$Minimize\ \mathcal{F}(X)\ s.t.\ X \in \mathcal{M}_{n_1 \times n_2}^{(r)}. \quad (3)$$

If $Z^{(R)}, Z^{(C)} \overset{i.i.d.}{\sim} N(0, \tau^2)$, minimizing the loss function in eq. (2) is equivalent to maximizing the log-likelihood of the data, giving a statistical motivation for the above score. Problem (3) is non-convex due to the non-convex rank constraint $rank(X) \leq r$.

Our problem is a specialization of the general affine matrix recovery problem (Recht et al., 2010), in which a matrix is measured through a general affine transformation $\mathcal{A}$, with $\mathbf{b} = \mathcal{A}(X) + z$. We consider next and throughout the paper two specific *random ensembles* of measurement matrices $A^{(R)}, A^{(C)}$:

1. **Row and Column Matrix Completion (RCMC):** In this ensemble each row of $A^{(R)}$ and each column of $A^{(C)}$ is a vector of the standard basis e_j for some j - thus each measurement $B_{ij}^{(R)}$ or $B_{ij}^{(C)}$ is obtained from a single entry of X. We define a row-inclusion probability $p^{(R)}$ and column inclusion probability $p^{(C)}$ such that each row (column) of the matrix X will be measured with probability $p^{(R)}(p^{(C)})$. More precisely, we define $r_1, .., r_{n_1}$ i.i.d. Bernoulli variables,

$P(r_i = 1) = p^{(R)}$, and include e_i as a row in $A^{(R)}$ if and only if $r_i = 1$. Similarly, we define $c_1...c_{n_2}$ i.i.d. Bernoulli variables, $P(c_i = 1) = p^{(C)}$, and include e_i as a column in $A^{(C)}$ if and only if $c_i = 1$. The expected number of observed rows (columns) is $k^{(R)} = n_1 p^{(R)} (k^{(C)} = n_2 p^{(C)})$. The model is very close to the possibly more natural model of picking $k^{(R)}$ distinct rows and $k^{(C)}$ distinct columns at random for fixed $k^{(R)}, k^{(C)}$, but allows for easier analysis.

2. **Gaussian Rows and Columns (GRC):** In this ensemble $A^{(R)}, A^{(C)} \overset{i.i.d.}{\sim} N(0,1)$. Each observation $B_{ij}^{(R)}$ or $B_{ij}^{(C)}$ is obtained by a weighted sum of a single row or column of X, with i.i.d. Gaussian weights.

2.1. Comparison to Standard Designs

The rows-and-columns design presented above is distinct from standard measurements ensembles proposed and studied in the literature. It is instructive to compare the GRC ensemble to the Gaussian Ensemble (GE) model (Candès & Plan, 2011), where using the matrix representation $\mathcal{A}(X) = Avec(X)$ with $A \in \mathbb{R}_{d \times n_1 n_2}$, we take $A \overset{i.i.d.}{\sim} N(0,1)$. For the latter, the following r-Restricted Isometry Property (RIP) can be used:

Definition 1. *(r-RIP) Let $\mathcal{A} : \mathbb{R}_{n_1 \times n_2} \to \mathbb{R}^d$ be a linear map. For every integer r with $1 \leq r \leq min(n_1, n_2)$, define the r-Restricted Isometry Constant to be the smallest number ϵ_r such that*

$$(1 - \epsilon_r)||X||_F \leq ||\mathcal{A}(X)||_F \leq (1 + \epsilon_r)||X||_F \quad (4)$$

holds for all matrices X of rank at most r.

The GE model satisfies the r-Restricted Isometry Property (RIP) for $d = O(rn)$ with high probability (Recht et al., 2010). Based on this property it is known that nuclear norm minimization (Recht et al., 2010; Candès & Plan, 2011) and other algorithms such as singular value projection (Jain et al., 2010) can recover X with high probability. Unlike GE, in our GRC model the operator $\mathcal{A}(X)$ doesn't satisfy the RIP, and nuclear norm minimization fails. Instead, $A^{(R)}, A^{(C)}$ preserve matrix Frobenius norm in high probability, which is a weaker property than the RIP which holds for *any* low-rank matrix. (see lemma 7 in Appendix).

We next compare the RCMC model to the standard Matrix Completion model (Candès & Recht, 2009), in which single entries are chosen at random to be observed. Unlike GE, for MC incoherence conditions on X are required in order to recover X (Candès & Recht, 2009):

Definition 2. *(Incoherence). Let U be a subspace of $\mathbb{R}^n$ of dimension r, and P_U be the orthogonal projection on U.*

Then the coherence of U (with respect to the standard basis $\{e_i\}$) is defined as

$$\mu(U) \equiv \frac{n}{r} max_i ||P_U(e_i)||^2. \tag{5}$$

We say that a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ is μ-incoherent if for the SVD $X = U\Sigma V^T$ we have $max(\mu(U), \mu(V)) \leq \mu$.

When X is incoherent, and when known entries are sampled uniformly at random from X, several algorithms (Keshavan et al., 2009; Cai et al., 2010; Jain et al., 2010) succeed to recover X with high probability. In particular, nuclear norm minimization has gained popularity as a solver for the standard MC problem due to it's recovery guarantees, and a convenient representation as a convex optimization problem with availability of many iterative algorithms for the problem. However, nuclear norm minimization fails for the RCMC design, even when the matrix X is incoherent, as shown by the next example:

Example: Take $X \in \mathbb{R}^{n \times n}$ for $\frac{n}{3} \in \mathbb{N}$ with $X_{ij} = 1 \forall (i, j) \in [n] \times [n]$. Thus $||X||_* = n$. Take $k^{(R)} = k^{(C)} = \frac{n}{3}$. One can set all unknown entries to 0.5, giving a matrix X_0 of rank 2 with $\sigma_1(X_0) = \frac{(\sqrt{2}+1)n}{3}$, $\sigma_2(X_0) = \frac{(\sqrt{2}-1)n}{3}$, therefore $||X_0||_* = \frac{n\sqrt{2}}{3} < ||X||_*$ and nuclear norm minimization fails to recover the correct X. In Section 3 we present our SVLS algorithm, which does not rely on nuclear-norm minimization.

3. Algorithms for Recovery of X

In this section we give an efficient algorithm which we call SVLS (Singular Value Least Squares), SVLS very easy to implement, first we give algorithm 1 for the noiseless case and than we expend our algorithm to the general case.

3.1. Noiseless Case

In the noiseless case we reduce the optimization problem 3 to solving a system of linear equations, and provide a closed-form estimator. We then give conditions under which with high probability, the solution is unique and is equal to the true matrix X. If $rank(A^{(R)}\hat{U}) = r$ one can write the resulting estimator $\hat{X}$ in closed-form as follows:

$$\hat{X} = \hat{U}Y = \hat{U}[\hat{U}^T A^{(R)^T} A^{(R)}\hat{U}]^{-1}\hat{U}^T A^{(R)^T} B^{(R)} \tag{6}$$

Algorithm 1 doesn't treat the row and column measurements symmetrically. We can apply the same algorithm, but changing the role of rows and columns. The resulting closed form solution is then:

$$\hat{X} = B^{(C)}A^{(C)}(\hat{V}^T A^{(C)}A^{(C)^T}\hat{V})^{-1}\hat{V}^T \tag{7}$$

for an orthogonal matrix $\hat{V}$ representing a basis for the rows of X. Since the algorithm uses matrix inversion/Gaussian

---

Algorithm 1

Input $A^{(R)}, A^{(C)}, B^{(R)}, B^{(C)}$ and rank r

1. Compute a basis (of size r) to the row space of $B^{(C)}$ using Gaussian elimination, represented as the columns of a matrix $\hat{U} \in \mathbb{R}_{n_1 \times r}$.

2. Solve the linear system $B^{(R)}_{\bullet j} = A^{(R)}\hat{U}Y_{\bullet j}$ for each $j = 1, .., n_2$ and write the solutions as a matrix $Y = Y_{\bullet 1}...Y_{\bullet n_2}$.

3. Output $\hat{X} = \hat{U}Y$

---

elimination steps, it is crucial that we have exact noiseless measurements. Next, we modify the algorithm to work also for noisy measurements.

3.2. General (Noisy) Case

In the noisy case we seek a matrix X minimizing the loss $\mathcal{F}$ in eq. (2). The minimization problem is non-convex and there is no known algorithm with optimality guarantees. We propose the following algorithm which empirically returns a matrix estimator $\hat{X}$ with a low value of the loss F:

Algorithm 2 SVLS

Input $A^{(R)}, A^{(C)}, B^{(R)}, B^{(C)}$ and rank r

1. compute $B^{(C)} = \hat{U}\hat{\Sigma}\hat{V}^T$, the SVD of $B^{(C)}$, ($\hat{U}$ is a basis for the columns space of $B^{(C)}$)

2. Find the least-squares solution $\hat{Y} = argmin_Y \| B^{(R)} - A^{(R)}\hat{U}Y\|_F$. If $rank(A^{(R)}\hat{U}) = r$ we can write $\hat{Y}$ in closed form as before:

$$\hat{Y} = [\hat{U}^T A^{(R)^T} A^{(R)}\hat{U}]^{-1}\hat{U}^T A^{(R)^T} B^{(R)} \tag{8}$$

3. Return the estimate $\hat{X}^{(R)} = \hat{U}\hat{Y}$.

4. Repeat 1-3 with replacing roles of the columns and the rows and get matrix $\hat{X}^{(C)}$

5. Set $\hat{X} = argmin_{\hat{X}^{(R)}, \hat{X}^{(C)}} \mathcal{F}(X)$

3.2.1. GRADIENT DESCENT

SVLS returns an estimator $\hat{X}$ of X, which may not minimize the loss function in eq. (2). We therefore perform an additional gradient descent stage starting from $\hat{X}$ to achieve an estimator with lower loss (while still a local minima due to the non-convexity of the problem). SVLS can be thus viewed as a fast method for providing a desir-

3.3. Estimation Of Unknown Rank

In real life problems, one doesn't know the true rank of a matrix and should estimate it from data. Our rows-and-columns sampling design is particularly suitable for rank estimation since $rank(B^{(C,0)}) = rank(B^{(R,0)}) = rank(X)$ with high probability when enough rows and columns are sampled. In the noiseless case we can estimate $rank(X)$ by $\hat{r} = rank(B^{(C,0)})$.

For the noisy case we can estimate $rank(X)$ from $B^{(C)}, B^{(R)}$. We used the popular elbow method to estimate $rank(B^{(C)})$ in the following way

$$\hat{r}^{(C)} = argmax_{i \in [k^{(C)}-1]} \left(\frac{\sigma_i(B^{(C)})}{\sigma_{i+1}(B^{(C)})} \right) \quad (9)$$

We compute similarly $\hat{r}^{(R)}$ from $B^{(R)}$ and take the average as our rank estimator, $\hat{r} = round\left(\frac{\hat{r}^{(C)} + \hat{r}^{(R)}}{2} \right)$.

Modern methods for rank estimation from singular values (Gavish & Donoho, 2013) (Gavish & Donoho, 2014) can be similarly applied to $B^{(R)}, B^{(C)}$ and averaged, and may yield more accurate rank estimates.

4. Performance Guarantees

We show guarantees on the accuracy of the estimator $\hat{X}$ returned by SVLS. Our guarantees are probabilistic, with respect to randomizing the design matrices $A^{(R)}, A^{(C)}$. For the noiseless case we give conditions which are close to optimal for exact recovery.

4.1. Noiseless Case

A rank r matrix of size $n_1 \times n_2$ has $r(n_1 + n_2 - r)$ degrees of freedom, and can therefore not be recovered by fewer measurements. Setting $k^{(R)} = k^{(C)} = r$ gives precisely this minimal number of measurements. We next show that this number suffices, with probability 1, to guarantee accurate recovery of X in the GRC model. In the RCMC model we need an additional incoherence assumption on X and a larger number of measurements in order to guarantee accurate recovery with high probability.

Lemma 1. *Let* $X_1, X_2 \in \mathcal{M}_{n_1 \times n_2}^{(r)}$ *and* $A^{(R)} \in \mathbb{R}_{k^{(R)} \times n_1}, A^{(C)} \in \mathbb{R}_{n_2 \times k^{(C)}}$ *such that* $rank(A^{(R)}X_1) = rank(X_1 A^{(C)}) = r$. *If* $A^{(R)}X_1 = A^{(R)}X_2$ *and* $X_1 A^{(C)} = X_2 A^{(C)}$ *then* $X_1 = X_2$.

Proof. First, $rank(X_2 A^{(C)}) = rank(X_1 A^{(C)}) = r$ and $rank(A^{(R)}X_2) = rank(A^{(R)}X_1) = r$. Since $span(X_1 A^{(C)}), span(X_2 A^{(C)})$ are subspaces of $span(X_1), span(X_2)$ respectively, and

$dim(span(X_2)) = r$ we get $span(X_2) = span(X_2 A^{(C)}) = span(X_1 A^{(C)}) = span(X_1)$, and we define $U \in \mathcal{O}_{n_1 \times r}$ a basis for this subspace.

For X_1, X_2 there are $Y_1, Y_2 \in \mathbb{R}_{r \times n_2}$ such that $X_1 = UY_1, X_2 = UY_2$. Hence $A^{(R)}UY_1 = A^{(R)}UY_2$. Since $rank(A^{(R)}UY_1) = r$ and $U \in \mathcal{O}_{n_1 \times r}$ we get $rank(A^{(R)}U) = r$, hence the matrix $U^T A^{(R)^T} A^{(R)}U$ is invertible, $Y_1 = Y_2$, and $X_1 = UY_1 = UY_2 = X_2$. $\square$

Lemma 2. *Let* $X \in \mathcal{M}_{n_1 \times n_2}^{(r)}$ *such that* $X = U\Sigma V^T$, *and* $A^{(R)} \in \mathbb{R}_{k^{(R)} \times n_1}, A^{(C)} \in \mathbb{R}_{n_2 \times k^{(C)}}$ *such that* $rank(A^{(R)}X) = rank(XA^{(C)}) = r$. *If* $A^{(R)}, A^{(C)}, B^{(R,0)}, B^{(C,0)}$ *and* r *is the input of SVLS then the output* $\hat{X}$ *satisfies*

$$A^{(R)}X = A^{(R)}\hat{X}, \ XA^{(C)} = \hat{X}A^{(C)} \quad (10)$$

Proof. $span(XA^{(C)}) \subseteq span(X)$ and $rank(XA^{(C)}) = rank(XA^{(C)}) = r$, hence $span(XA^{(C)}) = span(X)$ and $\hat{U}$ from stage 1 in SVLS is a basis for $span(X)$. We can write $X = \hat{U}L$ for some matrix $L \in \mathbb{R}_{r \times n_2}$, Since $rank(A^{(R)}\hat{U}L) = rank(\hat{U}) = r$, we have $rank(A^{(R)}\hat{U}) = r$. Thus eq. (8) gives $\hat{X}$ in closed form and we get:

$$A^{(R)}\hat{X} = A^{(R)}\hat{U}[\hat{U}^T A^{(R)^T} A^{(R)}\hat{U}]^{-1}\hat{U}^T A^{(R)^T} B^{(R,0)} =$$
$$A^{(R)}\hat{U}[\hat{U}^T A^{(R)^T} A^{(R)}\hat{U}]^{-1}\hat{U}^T A^{(R)^T} A^{(R)}\hat{U}L =$$
$$A^{(R)}\hat{U}L = A^{(R)}X \quad (11)$$

$$\hat{X}A^{(C)} = \hat{U}[\hat{U}^T A^{(R)^T} A^{(R)}\hat{U}]^{-1}\hat{U}^T A^{(R)^T} A^{(R)}XA^{(C)} =$$
$$\hat{U}[\hat{U}^T A^{(R)^T} A^{(R)}\hat{U}]^{-1}\hat{U}^T A^{(R)^T} A^{(R)}\hat{U}LA^{(C)} =$$
$$\hat{U}LA^{(C)} = XA^{(C)} \quad (12)$$
$\square$

4.1.1. Exact Recovery for GRC

For the GRC model, the above lemma can be used to prove exact recovery of X with the minimal possible number of measurements:

Lemma 3. *Let* $V \in \mathcal{O}_{n \times r}$ *and* $A^{(C)} \in \mathbb{R}_{n \times k}$ *be a random matrix* $A^{(C)} \overset{i.i.d.}{\sim} N(0, \sigma^2)$. *Then* $V^T A^{(C)} \in \mathbb{R}_{r \times k} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

The proof of this lemma is in the appendix (7.1).

Theorem 1. *Let* $\hat{X}$ *be the output of Algorithm SVLS in the GRC model with* $Z^{(C)}$ *and* $Z^{(R)}$ *equal to zero and* $k^{(R)}, k^{(C)} \geq r$. *Then* $P(\hat{X} = X) = 1$.

Proof. Let $U\Sigma V^T$ be the SVD of X, from lemma 3 and since the measure of low rank matrices is zero and $k^{(C)} \geq r$ we get that $rank(V^T A^C) = r$, hence if $B^{(C)} =$

$U\Sigma V^T A^{(C)}$ $rank(B^{(C)}) = rank(U\Sigma V^T A^{(C)}) = r$, In the same way $rank(B^{(R)}) = r$. hence combining lemma 2 with lemma 1 give us the require result $\square$

4.1.2. EXACT RECOVERY FOR RCMC

Unlike in the GRC model we can't promise uniqueness of solution in the RCMC model. We wish to find function such that for any incoherence measure we can find the probability for reconstruct X. We assume the Bernoulli rows and columns model as described in Section 2 and assume for simplicity that $p^{(R)} = p^{(C)} = p$ (and $k^{(R)} = k^{(C)} = k$). We prove that if $U \in \mathcal{O}_{n \times r}$ is orthonormal then with high probability $p^{-1}||U^T A^{(R)^T} A^{(R)} U - pI_r||_2 < 1$. Because U is orthonormal, this is equivalent to

$$p^{-1}||UU^T A^{(R)^T} A^{(R)} UU^T - pUU^T||_2 < 1 \Leftrightarrow$$
$$p^{-1}||P_U P_{A^{(R)^T}} P_U - pP_U||_2 < 1 \quad (13)$$

where $P_U = UU^T, P_{A^{(R)^T}} = A^{(R)^T} A^{(R)}$. We generalize Theorem 4.1 from (Candès & Recht, 2009).

Lemma 4. *Suppose $A^{(R)}$ as in the rows and columns model with probability p, and $U \in \mathcal{O}_{n \times r}$ with $\mu(U) = \frac{n}{r} max_i ||P_U(e_i)||^2 = \mu$. Then there is a numerical constant C_R such that for all $\beta > 1$, if $C_R\sqrt{\frac{\beta log(n) r \mu}{pn}} < 1$ then:*

$$P\left(p^{-1}||P_U P_{A^{(R)^T}} P_U - pP_U||_2 < C_R\sqrt{\frac{\beta log(n) r \mu}{pn}}\right)$$
$$> 1 - 3n^{-\beta} \quad (14)$$

The proof of lemma 4 is built upon (yet generalizes) the proof of Theorem 4.1 from (Candès & Recht, 2009) and is given in Appendix (7.3).

Theorem 2. *Let $X = U\Sigma V^T$ be the SVD of $X \in \mathbb{R}_{n \times n}$, and $max(\mu(U), \mu(V)) < \mu$. Take $A^{(R)}$ and $A^{(C)}$ as in the RCMC model without noise and probability $p = \frac{k}{n}$. Let $\beta > 1$ such that $C_R\sqrt{\frac{\beta log(n) r \mu}{k}} < 1$ where C_R as in lemma 4 and let $\hat{X}$ is the output of SVLS,. Then $P\left(\hat{X} = X\right) > 1 - 6n^{-\beta}$ provided that $C_R\sqrt{\frac{\beta log(n) r \mu}{k}} < 1$.*

Proof. From lemma 4 we have that with probability $> 1 - 6n^{-\beta}$, $p^{-1}||pI_r - U^T A^{(R)^T} A^{(R)} U||_2 < 1$ and $p^{-1}||pI_r - V^T A^{(C)} A^{(C)^T} V||_2 < 1$. Since the singular values of $pI_r - U^T A^{(R)^T} A^{(R)} U$ are $|p - \sigma_i(U^T A^{(R)^T} A^{(R)} U)|$ for $1 \le i \le r$, we have that

$$p - \sigma_r(U^T A^{(R)^T} A^{(R)} U) \le \sigma_1(pI_r - U^T A^{(R)^T} A^{(R)} U) < p$$
$$\Rightarrow 0 < \sigma_r(U^T A^{(R)^T} A^{(R)} U) \quad (15)$$

and similarly for $V^T A^{(C)} A^{(C)^T} V$. Therefore $rank(A^{(R)} U) = rank(V^T A^{(C)}) = r$ and $rank(A^{(R)} X) = rank(XA^{(C)}) = r$ with probability $> 1 - 6n^{-\beta}$, hence from lemma 2 $A^{(R)} X = A^{(R)} \hat{X}$ $XA^{(C)} = \hat{X} A^{(C)}$ and from lemma (1) $X = \hat{X}$. $\square$

Remark 1. *The combination of row and column measurements is crucial in order to guarantee recovery. If, for example we observe only rows then even with $n - 1$ observed rows and rank $r = 1$ we can only determine the unobserved row up to a constant, and thus cannot recover X uniquely.*

4.2. General (Noisy) Case

In the noisy case we can't ensure to recover the exact X. Instead, we bound the reconstruction error $||X - \hat{X}||_F$ where $\hat{X}$ is the output of Algorithm (2).Here, we give bounds on the error for the GRC model and similar models under some conditions. For simplicity, we show the result for $k^{(R)} = k^{(C)} = k$.

We focus on the high dimensional case $k \le n$, where number of measurements is low. In this case our bound is similar to the bound of the Gaussian Ensemble (GE). In (Candès & Plan, 2011) Candes and Plan proved that with high probability $||X - \hat{X}||_F < C_G\sqrt{\frac{nr\tau^2}{d}}$ for some constant C_G for GE. We next prove an analogous result for our GRC model.

Theorem 3. *Let $A^{(R)}$ and $A^{(C)}$ with $k \ge max(4r, 40)$ be as in the GRC model with noise matrices $Z^{(R)}, Z^{(C)}$. Let $\hat{X}$ be the output of SVLS. Then with probability $> 1 - 5e^{ck}$:*

$$||X - \hat{X}||_F \le c^{(C)}\sqrt{\frac{r}{k}}||Z^{(C)}||_2 + c^{(R)}\sqrt{\frac{r}{k}}||Z^{(R)}||_2 \quad (16)$$

where $c^{(R)}, c^{(C)}, c$ are absolute constants.

Theorem 3 applies for any $Z^{(C)}$ and $Z^{(R)}$. If $k \le n$ and $Z^{(R)}, Z^{(C)} \overset{i.i.d.}{\sim} N(0, \tau^2)$, then from eq. (32) we get $max(||Z^{(R)}||_2, ||Z^{(C)}||_2) \le 4\tau\sqrt{n}$ with probability $1 - e^{-2n}$. We therefore get the next Corollary for i.i.d. Gaussian noise:

Corrolary 1. *Let $A^{(R)}, A^{(C)}$ as in the GRC with $n \ge k \ge max(4r, 40)$, model and $Z^{(R)}, Z^{(C)} \overset{i.i.d.}{\sim} N(0, \tau^2)$. Then with probability $> 1 - 5e^{-ck} - e^{-2n}$:*

$$||X - \hat{X}||_F \le 4\left(c^{(R)} + c^{(C)}\right)\sqrt{\frac{\tau^2 nr}{k}} \quad (17)$$

5. Simulations Results

We studied the performance of our algorithm using simulations. We measured the reconstruction accuracy using the
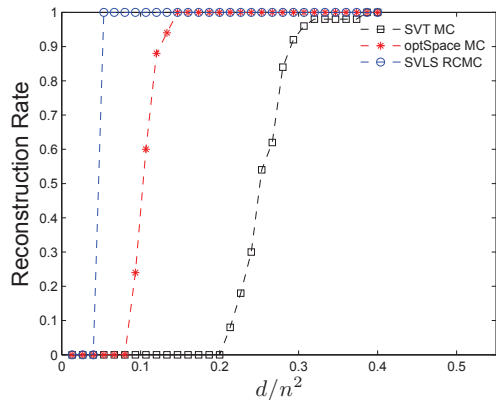
Figure 1. Reconstruction rates for matrices with dimension $n = 150$ and $r = 3$ where d is the number of known entries varied between 0 to 8000. SVT and optSpace are applied to the standard MC design and Algorithm 1 to RCMC. For each d we sampled 50 matrices and calculated the reconstruction rate as discovered in main text.

relative Root-Mean-Squared-Error (RMSE), defined as

$$RRMSE = RRMSE(X, \hat{X}) = ||X - \hat{X}||_F / ||X||_F. \tag{18}$$

For simplicity, we concentrated on square matrices with $n_1 = n_2 = n$ and used an equal number of row and column measurements, $k^{(R)} = k^{(C)} = k$. In all simulations we sampled a random rank-r matrix $X = UV^T$ with $U, V \in \mathbb{R}_{n \times r}$, $U, V \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

In all simulations we assumed that $rank(X)$ is unknown and estimated using the elbow method in eq. (9).

5.1. Row-Column Matrix Completion (RCMC)

In the noiseless case we compared our design to the standard matrix completion. We compared the reconstruction rate (probability of exact reconstruction of X as function of the number of measurements d) for the RCMC design with Algorithm SVLS to the reconstruction rate for the standard MC design with the optSpace(Keshavan et al., 2010) and SVT(Cai et al., 2010) algorithms. To allow for numerical errors, for each simulation yielding $X, \hat{X}$ we defined recovery as successful if their $RRMSE$ was lower than 10^{-3}, and for each value of d recorded the percentage of simulations for which recovery was successful . In Figure 5.1 we show results for $n = 150, r = 3$ and $\sigma = 1$. SVLS reconstruct X with optimal number of measurements $d = r(2n - r) = 894$ yielding $\frac{d}{n^2} \approx 0.04$ with probability 1 while MC with optSpace and SVT need roughly 3-fold and 8-fold more measurements, respectively, to guarantee exact recovery.

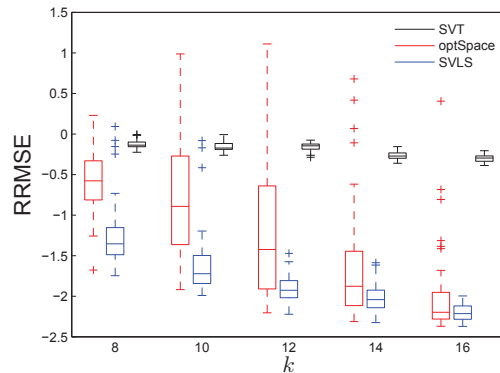The improvement in accuracy is not due to our design or

Figure 2. Box-plots represent the distribution of $RRMSE$ as a function of the number of column and row measurements k over 50 different sampled matrices $X = UV^T$ with $U, V \overset{i.i.d.}{\sim} N(0, 1)$ and $Z^{(R)}, Z^{(C)} \overset{i.i.d.}{\sim} N(0, 0.25^2)$. OptSpace (red) fails to recover X on many instances while SVLS (blue) performs very well on all of them. SVT(green) fails to recover X for all instances. The trimming of dense rows and columns in OptSpace was skipped, since such trimming in our settings may delete all measurement information for low k.

our algorithm alone, but due to their combination. We compared our method to optSpace and SVT for RCMC. We sampled a matrix X with $n = 100$ $r = 3$ and $\sigma = 1$ noise level $\tau^2 = 0.25^2$ and varied the number of row and column measurements k Figure 5.1 shows that while SVLS is very stable even for small k, the optSpace has a lot of outliers and SVT which minimize the nuclear norm achieves poor accuracy. (Remark: The algorithm optSpace has a trimming step which delete dense columns. We omitted this step in the RCMC model since it would delete all the known columns and rows and it's not stable for this type of measurements, but it still get better result than SVT.)

Next, we compared our RCMC to standard MC. We sampled X as before with $U, V \in \mathbb{R}_{1000 \times r}$ with standard Gaussian distribution, different rank and different noise ratio. The observations were corrupted by additive Gaussian noise Z with relative noise $NR \equiv ||Z||_F / ||X||_F$.

For small number of measurements our algorithm gives better results than the standard MC, and is comparable to MC whem the number of measurements is large. Moreover, our algorithm is significantly faster than the other two algorithms in Table 1.

5.2. Gaussian Rows and Columns (GRC)

We tested the performance of the GRC model with $A^{(R)}, A^{(C)} \overset{i.i.d.}{\sim} N(0, \frac{1}{n})$ (for getting independence with n) and with $X = UV^T$ where every entry in U, V has $N(0, \frac{1}{\sqrt{r}})$ entry, we compere our results to the Gaussian

Table 1. RRMSE and time in seconds (in parenthesis) for SVLS applied to RCMC, and optSpace and SVT applied to the standard MC. Results represent average of 5 different random matrices. SVLS is faster than optSpace and SVT by 1 to 3 orders of magnitudes, and shows comparable or better *RRMSE* in all cases.

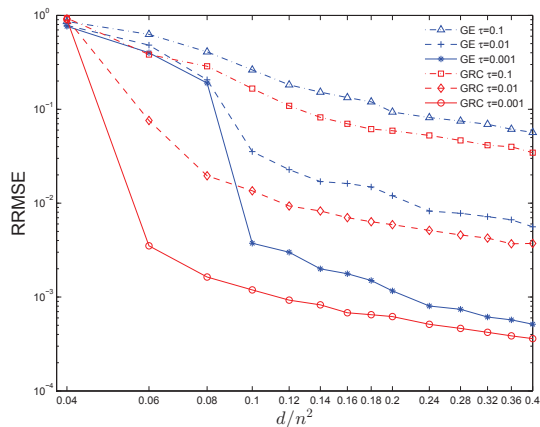| NR	d	r	SVLS	optSpace	SVT
10^{-2}	120156	10	0.063(0.15)	0.005(20.8)	0.0096(18.7)
10^{-1}	120156	10	0.064(0.15)	0.04(21.7)	0.056(11)
1	120156	10	0.64(0.16)	0.49(24.5)	0.52(1)
10^{-2}	59100	20	0.029(0.12)	0.97(25.6)	0.76(4.4)
10^{-1}	59100	20	0.3(0.12)	0.98(40.1)	0.86(6.5)
10^{-1}	391600	50	0.081(0.7)	0.05(1200)	0.069(13)
1	391600	50	0.72(0.6)	0.61(1300)	0.59(5)

Figure 3. Relative error *RRMSE* as function of d number of measurements ,where we take $X \in \mathcal{M}^{(2)}_{100 \times 100}$, d is varied from 400 to 4000 and for different noise levels: $\tau = 0.1, 0.01$ and 0.001. For every point we simulated 5 random matrices and computed the average relative error.

ensemble model (GE) for getting same scale of measurements we normalize $\mathcal{A}(X)$ with n. In Figure 3 we take $n = 100$ and $r = 2$ and change $d = 2nk$ the number of measurements , (where $A^{(R)} \in \mathbb{R}_{k \times n}$ and $A^{(C)} \in \mathbb{R}_{n \times k}$), in addition we added Gaussian noise $Z^{(R)}$ and $Z^{(C)}$ with different level of noise τ. Even for small k our method give good results. The error decays at a rate of $\sqrt{k}$ for the GE model we use algorithm APGL (Toh & Yun, 2010) for nuclear norm minimization.

In the next tests we ran SVLS on different noise levels. We take $n = 1000$ and $k = 100$ with different rank level every entry in $Z^{(C)}, Z^{(R)} \overset{i.i.d.}{\sim} N(0, \tau^2)$ and different values of τ. Results are shown in Figure 4. The change in the relative error $RRMSE$ is linear in τ while the rate depends on r.

We next examined the behaviour of the $RRMSE$ when $n \to \infty$ and when $n, k, r \to \infty$ together. Results (shown in Appendix Section 7.5) show that when properly scaled, the $RRMSE$ error is not sensitive to the value of n and other parameters, in agreement with Theorem 3.
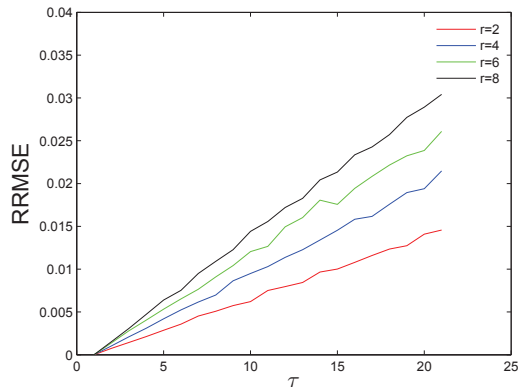

Figure 4. For $X \in \mathbb{R}_{1000 \times 1000}$ of different ranks , we plot the $RRMSE$ as a function of τ varied from 0 to 0.1. For each curve we fitted a linear regression line, with fitted slopes $0.145, 0.208, 0.25, 0.3$ for $r = 2, 4, 6, 8$, respectively. The slope is roughly proportional to $\sqrt{r}$ in concordance with the error bound in Theorem 3 but further investigation of the relation using extensive simulations is required in order to evaluate the dependency of the recovery error in r in a more precise manner.

6. Discussion

We introduced a new measurements ensemble for low rank matrix recovery where every measurements is an affine combination of a row or column of X. We focused on two models: matrix completion from single columns and rows (RCMC) and matrix recovery from Gaussian combination of columns and rows (GRC). We proposed a fast algorithm for this ensemble. For the RCMC model we proved that in the noiseless case our method recovers X with high probability and Simulation results show that the RCMC model outperform the standard approach for matrix completion in both speed and accuracy for models with small noise.

For the GRC model we proves that our method return X with optimal number of measurements in the noiseless case and gave am upper bounds on the error for the noisy case. For RCMC, our simulations show that the RCMC design may achieve comparable or favorable results, compared to the standard MC design, especially for low noise level. Proving recovery guarantees for this RCMC model is an interesting challenge.

Our proposed measurement schemes is not restricted to recovery of low-rank matrices. One can employ this measurement scheme and recover X by minimizing other matrix norms. This direction can lead to new algorithms that may improve matrix recovery for real datasets.

References

Basri, Ronen and Jacobs, David W. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2):218–233, 2003.

Cai, Jian-Feng, Candès, Emmanuel J, and Shen, Zuowei. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Cai, T Tony, Zhang, Anru, et al. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

Candès, Emmanuel J and Plan, Yaniv. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Candès, Emmanuel J and Plan, Yaniv. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

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

Candès, Emmanuel J and Romberg, Justin. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

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

Chi, Eric C, Zhou, Hua, Chen, Gary K, Del Vecchyo, Diego Ortega, and Lange, Kenneth. Genotype imputation via matrix completion. *Genome research*, 23(3):509–518, 2013.

Chu, Ci, Qu, Kun, Zhong, Franklin L, Artandi, Steven E, and Chang, Howard Y. Genomic maps of long noncoding rna occupancy reveal principles of rna-chromatin interactions. *Molecular cell*, 44(4):667–678, 2011.

Dasgupta, Sanjoy and Gupta, Anupam. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Gavish, Matan and Donoho, David L. The optimal hard threshold for singular values is 4/sqrt(3). *arXiv preprint*, 2013.

Gavish, Matan and Donoho, David L. Optimal shrinkage of singular values. *Stanford University Statistics Department technical report 2014-08,*, 2014.

Gross, David, Liu, Yi-Kai, Flammia, Steven T, Becker, Stephen, and Eisert, Jens. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

Jain, Prateek, Meka, Raghu, and Dhillon, Inderjit S. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pp. 937–945, 2010.

Keshavan, Raghunandan H, Montanari, Andrea, and Oh, Sewoong. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems*, pp. 952–960, 2009.

Keshavan, Raghunandan H, Montanari, Andrea, and Oh, Sewoong. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

Koren, Yehuda, Bell, Robert, and Volinsky, Chris. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Ma, Shiqian, Goldfarb, Donald, and Chen, Lifeng. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

Recht, Benjamin. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

Recht, Benjamin, Fazel, Maryam, and Parrilo, Pablo A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Shalev-Shwartz, Shai and Ben-David, Shai. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Szarek, Stanislaw J. Condition numbers of random matrices. *Journal of Complexity*, 7(2):131–149, 1991.

Talagrand, Michel. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

Toh, Kim-Chuan and Yun, Sangwoon. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.