INVITED PAPER

Matrix Completion With Noise

Predictions about the choices of those who may take part in choosing such items as movies for rent can be accurately made with a relatively small number of examples.

By Emmanuel J. Candès and Yaniv Plan

ABSTRACT | On the heels of compressed sensing, a new field has very recently emerged. This field addresses a broad range of problems of significant practical interest, namely, the recovery of a data matrix from what appears to be incomplete, and perhaps even corrupted, information. In its simplest form, the problem is to recover a matrix from a small sample of its entries. It comes up in many areas of science and engineering, including collaborative filtering, machine learning, control, remote sensing, and computer vision, to name a few. This paper surveys the novel literature on matrix completion, which shows that under some suitable conditions, one can recover an unknown low-rank matrix from a nearly minimal set of entries by solving a simple convex optimization problem, namely, nuclear-norm minimization subject to data constraints. Further, this paper introduces novel results showing that matrix completion is provably accurate even when the few observed entries are corrupted with a small amount of noise. A typical result is that one can recover an unknown $n \times n$ matrix of low rank r from just about $nr \log^2 n$ noisy samples with an error that is proportional to the noise level. We present numerical results that complement our quantitative analysis and show that, in practice, nuclear-norm minimization accurately fills in the many missing entries of large low-rank matrices from just a few noisy samples. Some analogies between matrix completion and compressed sensing are discussed throughout.

KEYWORDS | Compressed sensing; duality in optimization; low-rank matrices; matrix completion; nuclear-norm minimization; oracle inequalities; semidefinite programming

Manuscript received March 18, 2009; accepted October 17, 2009. Date of publication April 26, 2010; date of current version May 19, 2010. The work of E. J. Candès was supported by the U.S. Office of Naval Research under grants N00014-09-1-0469 and N00014-08-1-0749 and by the National Science Foundation under a Waterman Award.
The authors are with Department of Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: emmanuel@acm.caltech.edu).

Digital Object Identifier: 10.1109/JPROC.2009.2035722

I. INTRODUCTION

Imagine that we only observe a few samples of a signal. Is it possible to reconstruct this signal exactly or at least accurately? For example, suppose we observe a few entries of a vector $x \in \mathbb{R}^n$, which we can think of as a digital signal or image. Can we recover the large fraction of entries—of pixels, if you will—that we have not seen? In general, everybody would agree that this is impossible. However, if the signal is known to be sparse in the Fourier domain and, by extension, in an incoherent domain, then accurate—and even exact—recovery is possible by ℓ_1 minimization [11]; see also [22] for other algorithms, [17] and [18] for other types of measurements, and [34] for different ideas. This revelation is at the root of the rapidly developing field of compressed sensing and is already changing the way engineers think about data acquisition; hence this Special Issue and others (see [2], for example). Concretely, if a signal has a sparse frequency spectrum and we only have information about a few time or space samples, then one can invoke linear programming to interpolate the signal exactly. One can of course exchange time (or space) and frequency and recover sparse signals from just a few of their Fourier coefficients as well.

Imagine now that we only observe a few entries of a data matrix. Then is it possible to accurately—or even exactly—guess the entries that we have not seen? For example, suppose we observe a few movie ratings from a large data matrix in which rows are users and columns are movies (we can only observe a few ratings because each user is typically rating a few movies as opposed to the tens of thousands of movies which are available). Can we predict the rating a user would hypothetically assign to a movie he/she has not seen? In general, everybody would agree that recovering a data matrix from a subset of its entries is impossible. However, if the unknown matrix is known to have low rank or approximately low rank, then accurate and even exact recovery is possible by nuclear norm minimization [10], [14]. This revelation, which to some extent is inspired by the great body of work in compressed sensing, is the subject of this paper.

0018-9219/$26.00 ©2010 IEEE

From now on, we will refer to the problem of inferring the many missing entries as the *matrix completion* problem. By extension, inferring a matrix from just a few linear functionals will be called the *low-rank matrix recovery* problem. Now just as sparse signal recovery is arguably of paramount importance these days, we do believe that matrix completion and, in general, low-rank matrix recovery is just as important, and will become increasingly studied in years to come. For now, we give a few examples of applications in which these problems do come up.

- *Collaborative filtering*. In a few words, collaborative filtering is the task of making automatic predictions about the interests of a user by collecting taste information from many users [23]. Perhaps the most well-known implementation of collaborating filtering is the Netflix recommendation system alluded to earlier, which seeks to make rating predictions about unseen movies. This is a matrix completion problem in which the unknown full matrix has approximately low rank because only a few factors typically contribute to an individual's tastes or preferences. In the new economy, companies are interested predicting musical preferences (Apple Inc.), literary preferences (Amazon, Barnes and Noble), and many other such things.
- *System identification*. In control, one would like to fit a discrete-time linear time-invariant state-space model

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned}$$

  to a sequence of inputs $u(t) \in \mathbb{R}^m$ and outputs $y(t) \in R^p$, $t = 0, \ldots, N$. The vector $x(t) \in \mathbb{R}^n$ is the state of the system at time t, and n is the order of the system model. From the input/output pair $\{(u(t), y(t)) : t = 0, \ldots N\}$, one would like to recover the dimension of the state vector n (the model order) and the dynamics of the system, i.e., the matrices A, B, C, D, and the initial state $x(0)$. This problem can be cast as a low-rank matrix recovery problem; see [26] and references therein.
- *Global positioning*. Finding the global positioning of points in Euclidean space from a local or partial set of pairwise distances is a problem in geometry that emerges naturally in sensor networks [7], [31], [32]. For example, because of power constraints, sensors may only be able to construct reliable distance estimates from their immediate neighbors. From these estimates, we can form a partially observed distance matrix, and the problem is to infer all the pairwise distances from just a few observed ones so that locations of the sensors can be reliably estimated. This reduces to a matrix completion problem where the unknown matrix is of rank two if the sensors are located in the plane and three if they are located are in space.
- *Remote sensing*. The MUSIC algorithm [30] is frequently used to determine the direction of arrival of incident signals in a coherent radio-frequency environment. In a typical application, incoming signals are being recorded at various sensor locations, and this algorithm operates by extracting the directions of wave arrivals from the covariance matrix obtained by computing the correlations of the signals received at all sensor pairs. In remote sensing applications, one may not be able to estimate or transmit all correlations because of power constraints [35]. In this case, we would like to infer a full covariance matrix from just a few observed partial correlations. This is a matrix completion problem in which the unknown signal covariance matrix has low rank since it is equal to the number of incident waves, which is usually much smaller than the number of sensors.

There are of course many other examples, including the structure-from-motion problem [15], [33] in computer vision, multiclass learning in data analysis [3], [4], and so on.

This paper investigates whether or not one can recover low-rank matrices from fewer entries and, if so, how and how well. In Section II, we will study the noiseless problem in which the observed entries are precisely those of the unknown matrix. Section III examines the more common situation in which the few available entries are corrupted with noise. We complement our study with a few numerical experiments demonstrating the empirical performance of our methods in Section IV and conclude with a short discussion (Section V).

Before we begin, it is best to provide a brief summary of the notations used throughout this paper. We shall use three norms of a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ with singular values $\{\sigma_k\}$. The *spectral norm* is denoted by $\|X\|$ and is the largest singular value. The Euclidean inner product between two matrices is defined by the formula $\langle X, Y \rangle := \mathrm{trace}(X^* Y)$, and the corresponding Euclidean norm is called the *Frobenius norm* and denoted by $\|X\|_F$ (note that this is the ℓ_2 norm of the vector of singular values). The *nuclear norm* is denoted by $\|X\|_* := \sum_k \sigma_k$ and is the sum of singular values (the ℓ_1 norm of the vector $\{\sigma_k\}$). As is standard, $X \succeq Y$ means that $X - Y$ is positive semidefinite.

Further, we will also manipulate linear transformations that act on the space $\mathbb{R}^{n_1 \times n_2}$, and we will use calligraphic letters for these operators as in $\mathcal{A}(X)$. In particular, the identity operator on this space will be denoted by $\mathcal{I} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$. We use the same convention as above, and $\mathcal{A} \succeq \mathcal{I}$ means that $\mathcal{A} - \mathcal{I}$ (seen as a big matrix) is positive semidefinite.

We use the usual asymptotic notation, for instance writing $O(M)$ to denote a quantity bounded in magnitude by CM for some absolute constant $C > 0$.

II. EXACT MATRIX COMPLETION

Hereafter, $M \in \mathbb{R}^{n_1 \times n_2}$ is a matrix we would like to know as precisely as possible. However, the only information available about M is a sampled set of entries M_{ij}, $(i,j) \in \Omega$, where Ω is a subset of the complete set of entries $[n_1] \times [n_2]$. (Here and in the sequel, $[n]$ denotes the list $\{1, \ldots, n\}$.) It will be convenient to summarize the information available via $\mathcal{P}_\Omega(M)$, where the sampling operator $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ is defined by

$$[\mathcal{P}_\Omega(X)]_{ij} = \begin{cases} X_{ij}, & (i,j) \in \Omega \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the question is whether it is possible to recover our matrix only from the information $\mathcal{P}_\Omega(M)$. We will assume that the entries are selected at random without replacement as to avoid trivial situations in which a row or a column is unsampled, since matrix completion is clearly impossible in such cases. (If we have no data about a specific user, how can we guess his/her preferences? If we have no distance estimates about a specific sensor, how can we guess its distances to all the sensors?)

Even with the information that the unknown matrix M has low rank, this problem may be severely ill posed. Here is an example that shows why: let x be a vector in $\mathbb{R}^n$ and consider the $n \times n$ rank-1 matrix

$$M = e_1 x^* = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_{n-1} & x_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

where e_1 is the first vector in the canonical basis of $\mathbb{R}^n$. Clearly, this matrix cannot be recovered from a subset of its entries. Even if one sees 95% of the entries sampled at random, then we will miss elements in the first row with very high probability, which makes the recovery of the vector x and, by extension, of M impossible. The analogy in compressed sensing is that one obviously cannot recover a signal assumed to be sparse in the time domain by subsampling in the time domain.

This example shows that one cannot hope to complete the matrix if some of the singular vectors of the matrix are extremely sparse—above, one cannot recover M without sampling all the entries in the first row; see [10] for other related pathological examples. More generally, if a row (or column) has no relationship to the other rows (or columns) in the sense that it is approximately orthogonal, then one would basically need to see all the entries in that row to recover the matrix M. Such informal considerations led the authors of [10] to introduce a geometric incoherence assumption, but for the moment, we will discuss an even simpler notion which forces the singular vectors of M to be spread across all coordinates. To express this condition, recall the singular value decomposition (SVD) of a matrix of rank r

$$M = \sum_{k \in [r]} \sigma_k u_k v_k^* \tag{II.1}$$

in which $\sigma_1, \ldots, \sigma_r \geq 0$ are the singular values and $u_1, \ldots, u_r \in \mathbb{R}^{n_1}$, $v_1, \ldots, v_r \in \mathbb{R}^{n_2}$ are the singular vectors. Our assumption is as follows:

$$\|u_k\|_{\ell_\infty} \leq \sqrt{\mu_B / n_1}, \quad \|v_k\|_{\ell_\infty} \leq \sqrt{\mu_B / n_2} \tag{II.2}$$

for some $\mu_B \geq 1$, where the ℓ_∞ norm is of course defined by $\|x\|_{\ell_\infty} = \max_i |x_i|$. We think of μ_B as being small, e.g., $O(1)$, so that the singular vectors are not too spiky as explained above.

If the singular vectors of M are sufficiently spread, the hope is that there is a unique low-rank matrix that is consistent with the observed entries. If this is the case, one could, in principle, recover the unknown matrix by solving

$$\begin{aligned} &\text{minimize} && \operatorname{rank}(X) \\ &\text{subject to} && \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(M) \end{aligned} \tag{II.3}$$

where $X \in \mathbb{R}^{n_1 \times n_2}$ is the decision variable. Unfortunately, not only is this problem NP-hard but also all known algorithms for exactly solving it are doubly exponential in theory and in practice [16]. This is analogous to the intractability of ℓ_0-minimization in sparse signal recovery.

A popular alternative is the convex relaxation [10], [14], [19], [21], [29]

$$\begin{aligned} &\text{minimize} && \|X\|_* \\ &\text{subject to} && \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(M) \end{aligned} \tag{II.4}$$

(see [6] and [28] for the earlier related trace heuristic). Just as ℓ_1-minimization is the tightest convex relaxation of the combinatorial ℓ_0-minimization problem in the sense that the ℓ_1 ball of $\mathbb{R}^n$ is the convex hull of unit-normed 1-sparse vectors (i.e., vectors with at most one nonzero entry), nuclear-norm minimization is the tightest convex relaxation of the NP-hard rank minimization problem. To be sure, the nuclear ball $\{X \in \mathbb{R}^{n_1 \times n_2} : \|X\|_* \leq 1\}$ is the convex hull of the set of rank-one matrices with spectral norm bounded by one. Moreover, in compressed sensing, ℓ_1 minimization subject to linear equality constraints can

be cast as a linear program (LP) for the ℓ_1 norm has an LP characterization: indeed, for each $x \in \mathbb{R}^n$, $\|x\|_{\ell_1}$ is the optimal value of

$$\begin{aligned} &\text{maximize} && \langle u, x\rangle \\ &\text{subject to} && \|u\|_{\ell_\infty} \le 1 \end{aligned}$$

with decision variable $u \in \mathbb{R}^n$. In the same vein, the nuclear norm of $X \in \mathbb{R}^{n_1 \times n_2}$ has the SDP characterization

$$\begin{aligned} &\text{maximize} && \langle W, X\rangle \\ &\text{subject to} && \|W\| \le 1 \end{aligned} \tag{II.5}$$

with decision variable $W \in \mathbb{R}^{n_1 \times n_2}$. This expresses the fact that the spectral norm is dual to the nuclear norm. The constraint on the spectral norm of W is an SDP constraint since it is equivalent to

$$\begin{bmatrix} I_{n_1} & W \\ W^* & I_{n_2} \end{bmatrix} \succeq 0$$

where I_n is the $n \times n$ identity matrix. Hence, (II.4) is an SDP, which one can express by writing $\|X\|_*$ as the optimal value of the SDP dual to (II.5). We note that specialized algorithms taking advantage of the problem structure have been shown to outperform interior-point methods by several orders of magnitude (see [8] and [27]).

In [14], it is proven that nuclear-norm minimization succeeds nearly as soon as recovery is possible by any method whatsoever.

Theorem 1 [14]: Let $M \in \mathbb{R}^{n_1 \times n_2}$ be a fixed matrix of rank $r = O(1)$ obeying (II.2) and set $n := \max(n_1, n_2)$. Suppose we observe m entries of M with locations sampled uniformly at random. Then there is a positive numerical constant C such that if

$$m \ge C\mu_B^4 n \log^2 n \tag{II.6}$$

then M is the unique solution to (II.4) with probability at least $1-n^{-3}$. In other words, with high probability, nuclear-norm minimization recovers all the entries of M with no error.

As a side remark, one can obtain a probability of success at least $1-n^{-\beta}$ for a given β by taking C in (II.6) of the form $C'\beta$ for some universal constant C'. The probabilistic nature of this result stems from the assumption that the revealed entries of M are sampled from the uniform distribution. Another interpretation is that matrix completion is exact for "most" sampling sets obeying (II.6).

An $n_1 \times n_2$ matrix of rank r depends upon $r(n_1+n_2-r)$ degrees of freedom.[1] When r is small, the number of degrees of freedom is much less than $n_1 n_2$, and this is the reason why subsampling is possible. (In compressed sensing, the number of degrees of freedom corresponds to the sparsity of the signal, i.e., the number of nonzero entries.) What is remarkable here is that exact recovery by nuclear-norm minimization occurs as soon as the sample size exceeds the number of degrees of freedom by a couple of logarithmic factors. Further, observe that if Ω completely misses one of the rows (e.g., one has no rating about one user) or one of the columns (e.g., one has no rating about one movie), then one cannot hope to recover even a matrix of rank 1 of the form $M = xy^*$. Thus one needs to sample every row (and also every column) of the matrix. When Ω is sampled at random, it is well established that one needs at least on the order $O(n \log n)$ for this to happen, as this is the famous coupon collector's problem. Hence, (II.6) misses the information theoretic limit by at most a logarithmic factor.

To obtain similar results for all values of the rank, [14] introduces the *strong incoherence property* with parameter μ stated below.

A) Let P_U (respectively, P_V) be the orthogonal projection onto the singular vectors $u_1, \dots, u_r$ (respectively, $v_1, \dots, v_r$). For all pairs $(a, a') \in [n_1] \times [n_1]$ and $(b, b') \in [n_2] \times [n_2]$

$$\left|\langle e_a, P_U e_{a'}\rangle - \frac{r}{n_1} 1_{a=a'}\right| \le \mu \frac{\sqrt{r}}{n_1}$$
$$\left|\langle e_b, P_V e_{b'}\rangle - \frac{r}{n_2} 1_{b=b'}\right| \le \mu \frac{\sqrt{r}}{n_2}.$$

B) Let E be the "sign matrix" defined by

$$E = \sum_{k \in [r]} u_k v_k^*. \tag{II.7}$$

For all $(a, b) \in [n_1] \times [n_2]$

$$|E_{ab}| \le \mu \frac{\sqrt{r}}{\sqrt{n_1 n_2}}.$$

These conditions do not assume anything about the singular values. As we will see, incoherent matrices with a small value of the strong incoherence parameter μ can be

[1]This can be seen by counting the number of parameters in the singular value decomposition.

recovered from a minimal set of entries. Before we state this result, it is important to note that many model matrices obey the strong incoherence property with a small value of μ.

- Suppose the singular vectors obey (II.2) with $\mu_B = O(1)$ (which informally says that the singular vectors are not spiky). Then with the exception of a very few peculiar matrices, M obeys the strong incoherence property with $\mu = O(\sqrt{\log n})$.[2]
- Assume that the column matrices $[u_1, \ldots, u_r]$ and $[v_1, \ldots, v_r]$ are independent random orthogonal matrices. Then with high probability, M obeys the strong incoherence property with $\mu = O(\sqrt{\log n})$, at least when $r \geq \log n$ so as to avoid small samples effects.

The sampling result below is general, nonasymptotic, and optimal up to a few logarithmic factors.

Theorem 2 [14]: Let $M \in \mathbb{R}^{n_1 \times n_2}$ be a fixed rank-r matrix with strong incoherence parameter μ, and set $n := \max(n_1, n_2)$. Suppose we observe m entries of M with locations sampled uniformly at random. Then there is a numerical constant C such that if

$$m \geq C\mu^2 nr \log^6 n \tag{II.8}$$

then M is the unique solution to (II.4) with probability at least $1-n^{-3}$.

In other words, if a matrix is strongly incoherent and the cardinality of the sampled set is about the number of degrees of freedom times a few logarithmic factors, then nuclear-norm minimization is exact. This improves on an earlier result of Candès and Recht [10], who proved—under slightly different assumptions—that on the order of $n^{6/5} r \log n$ samples were sufficient, at least for values of the rank obeying $r \leq n^{1/5}$.

We would like to point out a result of a broadly similar nature, but with a completely different recovery algorithm and with a somewhat different range of applicability, which was recently established by Keshavan *et al.* [24]. Their conditions are related to the incoherence property introduced in [10] and are also satisfied by a number of reasonable random matrix models. There is, however, another condition that states the singular values of the unknown matrix cannot be too large or too small (the ratio between the top and lowest value must be bounded). This algorithm 1) trims each row and column with too many entries; i.e., replaces the entries in those rows and columns by zero; and 2) computes the SVD of the trimmed matrix, truncates it as to only keep the top r singular values (note that the value of r is needed here), and rescales. The result is that under some suitable conditions discussed above, this recovers a good approximation to the matrix M provided that the number of samples is on the order of nr. The recovery is not perfect, but one can then perform local minimization to achieve exact recovery provided that one has more samples, on the order of $nr \max(\log n, r)$ (the recovery is stable provided that the noise level is small [25]). This work builds upon an earlier spectral technique developed in the literature of computer science [5], which also proves stability, but under stronger conditions.

[2]Specifically, there is a generic random model under which $\mu = O(\sqrt{\log n})$ with very high probability; see [10].

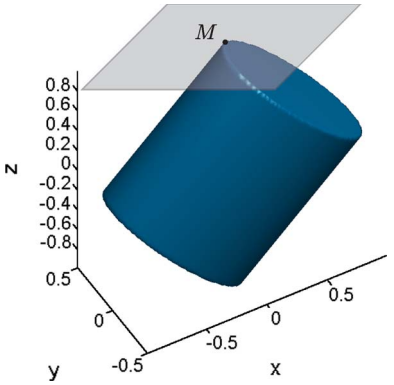

Fig. 1. ***The blue shape (courtesy of Recht) represents the nuclear ball (see the main text); the plane represents the feasible set.***

A. Geometry and Dual Certificates

We cannot possibly rehash the proof of [14, Th. 2] in this paper, or even explain the main technical steps, because of space limitations. We will, however, detail sufficient and almost necessary conditions for the low-rank matrix M to be the unique solution to the SDP (II.4). This will be useful to establish stability results.

The recovery is exact if the feasible set is tangent to the nuclear ball at the point M (see Fig. 1), which represents the set of points $(x, y, z) \in \mathbb{R}^3$ such that the 2×2 *symmetric* matrix $\begin{bmatrix} x & y \\ y & z \end{bmatrix}$ has nuclear norm bounded by one. To express this mathematically,[3] standard duality theory asserts that M is a solution to (II.4) if and only if there exists a dual matrix Λ such that $\mathcal{P}_\Omega(\Lambda)$ is a subgradient of the nuclear norm at M, written as

$$\mathcal{P}_\Omega(\Lambda) \in \partial \|M\|_*. \tag{II.9}$$

Recall the SVD (II.1) of M and the "sign matrix" E (II.7). It is well known that $Z \in \partial \|M\|_*$ if and only if Z is of the form

$$Z = E + W \tag{II.10}$$

[3]In general, M minimizes the nuclear norm subject to the linear constraints $\mathcal{A}(X) = b$, $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, if and only if there is $\lambda \in \mathbb{R}^m$ such that $\mathcal{A}^*(\lambda) \in \partial \|M\|_*$.

where

$$P_U W = 0, \quad W P_V = 0, \quad \|W\| \le 1. \tag{II.11}$$

In English, Z is a subgradient if it can be decomposed as the sign matrix plus another matrix with spectral norm bounded by one, whose column (respectively, row) space is orthogonal to the span of $u_1, \ldots, u_r$, (respectively, of $v_1, \ldots, v_r$). Another way to put this is by using notations introduced in [10]. Let T be the linear space spanned by elements of the form $u_k x^*$ and $y v_k^*$, $k \in [r]$, and let $T^\perp$ be the orthogonal complement to T. Note that $T^\perp$ is the set of matrices obeying $P_U W = 0$ and $W P_V = 0$. Then, $Z \in \partial \|M\|_*$ if and only if

$$Z = E + \mathcal{P}_{T^\perp}(Z), \quad \|\mathcal{P}_{T^\perp}(Z)\| \le 1.$$

This motivates the following definition.

Definition 3 (Dual Certificate): We say that Λ is a dual certificate if Λ is supported on Ω ($\Lambda = \mathcal{P}_\Omega(\Lambda)$), $\mathcal{P}_T(\Lambda) = E$, and $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1$.

Before continuing, we would like to pause to observe the relationship with ℓ_1 minimization. The point $x^\star \in \mathbb{R}^n$ is solution to

$$\begin{aligned} &\text{minimize} && \|x\|_{\ell_1} \\ &\text{subject to} && Ax = b \end{aligned} \tag{II.12}$$

with $A \in \mathbb{R}^{m \times n}$ if and only if there exists $\lambda \in \mathbb{R}^m$ such that $A^* \lambda \in \partial \|x^\star\|_{\ell_1}$. Note that if $S^\star$ is the support of $x^\star$, $z \in \partial \|x^\star\|_{\ell_1}$ is equivalent to

$$z = e + w, \quad e = \begin{cases} \operatorname{sgn}(x_i^\star), & i \in S^\star \\ 0, & i \notin S^* \end{cases}$$

and

$$w_i = 0 \text{ for all } i \in S, \quad \|w\|_{\ell_\infty} \le 1.$$

Hence, there is a clear analogy and one can think of T defined above as playing the role of the support set in the sparse recovery problem.

With this in place, we shall make use of the following lemma from [10].

Lemma 4 [10]: Suppose there exists a dual certificate Λ and consider any H obeying $\mathcal{P}_\Omega(H) = 0$. Then

$$\|M + H\|_* \ge \|M\|_* + (1 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H)\|_*.$$

Proof: For any $Z \in \partial \|M\|_*$, we have

$$\|M + H\|_* \ge \|M\|_* + \langle Z, H \rangle.$$

With $\Lambda = E + \mathcal{P}_{T^\perp}(\Lambda)$ and $Z = E + \mathcal{P}_{T^\perp}(Z)$, we have

$$\begin{aligned} \|M + H\|_* &\ge \|M\|_* + \langle \Lambda, H \rangle + \langle \mathcal{P}_{T^\perp}(Z - \Lambda), H \rangle \\ &= \|M\|_* + \langle Z - \Lambda, \mathcal{P}_{T^\perp}(H) \rangle \end{aligned}$$

since $\mathcal{P}_\Omega(H) = 0$. Now we use the fact that the nuclear and spectral norms are dual to one another. In particular, there exists $\|\bar{Z}\| \le 1$ such that $\langle \bar{Z}, \mathcal{P}_{T^\perp}(H) \rangle = \|\mathcal{P}_{T^\perp}(H)\|_*$. Now pick Z such that $\mathcal{P}_{T^\perp}(Z) = \mathcal{P}_{T^\perp}(\bar{Z})$ so that $\langle Z, \mathcal{P}_{T^\perp}(H) \rangle = \|\mathcal{P}_{T^\perp}(H)\|_*$, and note that $|\langle \Lambda, \mathcal{P}_{T^\perp}(H) \rangle| = |\langle \mathcal{P}_{T^\perp}(\Lambda), \mathcal{P}_{T^\perp}(H) \rangle| \le \|\mathcal{P}_{T^\perp}(\Lambda)\| \|\mathcal{P}_{T^\perp}(H)\|_*$. Therefore

$$\|M + H\|_* \ge \|M\|_* + (1 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H)\|_*$$

which concludes the proof. ■

A consequence of this lemma is the sufficient conditions below.

Lemma 5 [10]: Suppose there exists a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| < 1$ and that the restriction $\mathcal{P}_\Omega|_T : T \to \mathcal{P}_\Omega(\mathbb{R}^{n \times n})$ of the (sampling) operator $\mathcal{P}_\Omega$ restricted to T is injective. Then M is the unique solution to the convex program (II.4).

Proof: Consider any feasible perturbation $M + H$ obeying $\mathcal{P}_\Omega(H) = 0$. Then by assumption, Lemma 4 gives

$$\|M + H\|_* > \|M\|_*$$

unless $\mathcal{P}_{T^\perp}(H) = 0$. Assume then that $\mathcal{P}_{T^\perp}(H) = 0$; that is to say, $H \in T$. Then $\mathcal{P}_\Omega(H) = 0$ implies that $H = 0$ by the injectivity assumption. The conclusion is that M is the unique minimizer since any nontrivial perturbation increases the nuclear norm. ■

The methods for proving that matrix completion by nuclear-norm minimization is exact consist in constructing a dual certificate.

Theorem 6 [14]: Under the assumptions of either Theorem 1 or Theorem 2, there exists a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$. In addition, if $p = m/(n_1 n_2)$ is the fraction of observed entries, the operator $\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T : T \to T$ is one-to-one and obeys

$$\frac{p}{2}\mathcal{I} \preceq \mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T \preceq \frac{3p}{2}\mathcal{I} \tag{II.13}$$

where $\mathcal{I} : T \to T$ is the identity operator.

The second part, namely, (II.13), shows that the mapping $\mathcal{P}_\Omega : T \to \mathbb{R}^{n_1 \times n_2}$ is injective. Hence, the sufficient conditions of Lemma 5 are verified, and the recovery is exact. What is interesting is that the existence of a dual certificate together with the near-isometry (II.13)—in fact, the lower bound—is sufficient to establish the robustness of matrix completion vis-à-vis noise.

III. STABLE MATRIX COMPLETION

In any real-world application, one will only observe a few entries corrupted at least by a small amount of noise. In the Netflix problem, users' ratings are uncertain. In the system identification problem, one cannot determine the locations $y(t)$ with infinite precision. In the global positioning problem, local distances are imperfect. Lastly, in the remote sensing problem, the signal covariance matrix is always modeled as being corrupted by the covariance of noise signals. Hence, to be broadly applicable, we need to develop results that guarantee that reasonably accurate matrix completion is possible from noisy sampled entries. This section develops novel results showing that this is, indeed, the case.

Our noisy model assumes that we observe

$$Y_{ij} = M_{ij} + Z_{ij}, \quad (i,j) \in \Omega \tag{III.1}$$

where $\{Z_{ij} : (i,j) \in \Omega\}$ is a noise term that may be stochastic or deterministic (adversarial). Another way to express this model is as

$$\mathcal{P}_\Omega(Y) = \mathcal{P}_\Omega(M) + \mathcal{P}_\Omega(Z)$$

where Z is an $n \times n$ matrix with entries Z_{ij} for $(i,j) \in \Omega$ (note that the values of Z outside of Ω are irrelevant). All we assume is that $\|\mathcal{P}_\Omega(Z)\|_F \le \delta$ for some $\delta > 0$. For example, if $\{Z_{ij}\}$ is a white noise sequence with standard deviation σ, then $\delta^2 \le (m + \sqrt{8m})\sigma^2$ with high probability, say. To recover the unknown matrix, we propose solving the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & \|\mathcal{P}_\Omega(X - Y)\|_F \le \delta. \end{aligned} \tag{III.2}$$

Among all matrices consistent with the data, find the one with minimum nuclear norm. This is also an SDP, and let $\hat{M}$ be the solution to this problem.

Our main result is that this reconstruction is accurate.

Theorem 7: With the notations of Theorem 6, suppose there exists a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$ and that $\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T \succeq (p/2)\mathcal{I}$ (both these conditions are true with very large probability under the assumptions of the noiseless recovery Theorems 1 and 2). Then $\hat{M}$ obeys

$$\|M - \hat{M}\|_F \le 4\sqrt{\frac{C_p \min(n_1, n_2)}{p}}\delta + 2\delta \tag{III.3}$$

with $C_p = 2 + p$.

For small values of p (recall this is the fraction of observed entries), the error is of course at most just about $4\sqrt{2\min(n_1, n_2)/p}\delta$. As we will see from the proof, there is nothing special about 1/2 in the condition $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$. All we need is that there is a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le a$ for some $a < 1$ [the value of a only influences the numerical constant in (III.3)]. Further, when Z is random, (III.3) holds on the event $\|\mathcal{P}_\Omega(Z)\|_F \le \delta$.

Roughly speaking, our theorem states the following: *when perfect noiseless recovery occurs, then matrix completion is stable vis-à-vis perturbations*. To be sure, the error is proportional to the noise level δ; when the noise level is small, the error is small. Moreover, improving conditions under which noiseless recovery occurs has automatic consequences for the more realistic recovery from noisy samples.

A significant novelty here is that there is just no equivalent of this result in the compressed sensing or statistical literature, for our matrix completion problem does not obey the *restricted isometry property* (RIP) [12]. For matrices, the RIP would assume that the sampling operator obeys

$$(1 - \delta)\|X\|_F^2 \le \frac{1}{p}\|\mathcal{P}_\Omega(X)\|_F^2 \le (1 + \delta)\|X\|_F^2 \tag{III.4}$$

for all matrices X with sufficiently small rank and $\delta < 1$ sufficiently small [29]. However, the RIP does not hold here. To see why, let the sampled set Ω be arbitrarily chosen and fix $(i,j) \notin \Omega$. Then the rank-1 matrix $e_i e_j^*$ whose (i,j)th entry is one, and vanishes everywhere else, obeys $\mathcal{P}_\Omega(e_i e_j^*) = 0$. Clearly, this violates (III.4).

It is nevertheless instructive to compare (III.3) with the bound one would achieve if the RIP (III.4) were true. In this case, [20] would give

$$\|\hat{M} - M\|_F \le C_0 p^{-1/2}\delta$$

for some numerical constant C_0—that is, an estimate that would be better by a factor proportional to $1/\sqrt{\min(n_1, n_2)}$. It would be interesting to know whether or not estimates,

which are as good as what is achievable under the RIP, hold for the RIPless matrix completion problem. We will return to such comparisons later (Section III-B).

We close this section by emphasizing that our methods are also applicable to sparse signal recovery problems in which the RIP does not hold (the authors are currently writing a paper describing these results).

A. Proof of Theorem 7

We use the notation of the previous section and begin the proof by observing two elementary properties. The first is that since M is feasible for (III.2), we have the *cone constraint*

$$\|\hat{M}\|_* \le \|M\|_*. \tag{III.5}$$

The second is that the triangle inequality implies the *tube constraint*

$$\left\|\mathcal{P}_\Omega(\hat{M}-M)\right\|_F \le \left\|\mathcal{P}_\Omega(\hat{M}-Y)\right\|_F + \left\|\mathcal{P}_\Omega(Y-M)\right\|_F \le 2\delta \tag{III.6}$$

since M is feasible. We will see that under our hypotheses, (III.5) and (III.6) imply that $\hat{M}$ is close to M. Set $\hat{M} = M + H$ and put $H_\Omega := \mathcal{P}_\Omega(H)$, $H_{\Omega^c} := \mathcal{P}_{\Omega^c}(H)$ for short. We need to bound $\|H\|_F^2 = \|H_\Omega\|_F^2 + \|H_{\Omega^c}\|_F^2$, and since (III.6) gives $\|H_\Omega\|_F \le 2\delta$, it suffices to bound $\|H_{\Omega^c}\|_F$. Note that by the Pythagorean identity, we have

$$\|H_{\Omega^c}\|_F^2 = \|\mathcal{P}_T(H_{\Omega^c})\|_F^2 + \|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F^2 \tag{III.7}$$

and it is thus sufficient to bound each term in the right-hand side.

We start with the second term. Let Λ be a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$; we have

$$\|M + H\|_* \ge \|M + H_{\Omega^c}\|_* - \|H_\Omega\|_*$$

and

$$\|M + H_{\Omega^c}\|_* \ge \|M\|_* + [1 - \|\mathcal{P}_{T^\perp}(\Lambda)\|]\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_*.$$

The second inequality follows from Lemma 4. Therefore, with $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$, the cone constraint gives

$$\|M\|_* \ge \|M\|_* + \frac{1}{2}\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_* - \|H_\Omega\|_*$$

or, equivalently

$$\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_* \le 2\|H_\Omega\|_*.$$

Since the nuclear norm dominates the Frobenius norm $\|\mathcal{P}_{T^\perp}(H_\Omega c)\|_F \le \|\mathcal{P}_{T^\perp}(H_\Omega c)\|_*$, we have

$$\begin{aligned}\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F &\le 2\|H_\Omega\|_* \\ &\le 2\sqrt{n}\|H_\Omega\|_F \le 4\sqrt{n}\delta\end{aligned} \tag{III.8}$$

where the second inequality follows from the Cauchy–Schwarz inequality and the last from (III.6).

To develop a bound on $\|\mathcal{P}_T(H_\Omega c)\|_F$, observe that the assumption $\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T \succeq (p/2)\mathcal{I}$ together with $\mathcal{P}_T^2 = \mathcal{P}_T$, $\mathcal{P}_\Omega^2 = \mathcal{P}_\Omega$, gives

$$\begin{aligned}\|\mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c})\|_F^2 &= \langle \mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c}), \mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c})\rangle \\ &= \langle \mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c}), \mathcal{P}_T(H_{\Omega^c})\rangle \\ &\ge \frac{p}{2}\|\mathcal{P}_T(H_{\Omega^c})\|_F^2.\end{aligned}$$

But since $\mathcal{P}_\Omega(H_\Omega c) = 0 = \mathcal{P}_\Omega\mathcal{P}_T(H_\Omega c) + \mathcal{P}_\Omega\mathcal{P}_{T^\perp}(H_\Omega c)$, we have

$$\begin{aligned}\|\mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c})\|_F &= \|\mathcal{P}_\Omega\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F \\ &\le \|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F.\end{aligned}$$

Hence, the last two inequalities give

$$\begin{aligned}\|\mathcal{P}_T(H_{\Omega^c})\|_F^2 &\le \frac{2}{p}\|\mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c})\|_F^2 \\ &\le \frac{2}{p}\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F^2.\end{aligned} \tag{III.9}$$

As a consequence of this and (III.7), we have

$$\|H_{\Omega^c}\|_F^2 \le \left(\frac{2}{p} + 1\right)\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F^2.$$

The theorem then follows from this inequality together with (III.8).

B. Comparison With an Oracle

We would like to return to discussing the best possible accuracy one could ever hope for. For simplicity, assume that $n_1 = n_2 = n$, and suppose that we have an oracle

informing us about T. In many ways, going back to the discussion from Section II-A, this is analogous to giving away the support of the signal in compressed sensing [13]. With this precious information, we would know that M lives in a linear space of dimension $2nr - r^2$ and would probably solve the problem by the method of least squares

$$\begin{aligned} &\text{minimize} && \|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(Y)\|_F \\ &\text{subject to} && X \in T. \end{aligned} \tag{III.10}$$

That is, we would find the matrix in T that best fits the data in a least squares sense. Let $\mathcal{A} : T \to \Omega$ (we abuse notations and let Ω be the range of $\mathcal{P}_\Omega$) defined by $\mathcal{A} := \mathcal{P}_\Omega \mathcal{P}_T$. Then assuming that the operator $\mathcal{A}^*\mathcal{A} = \mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T$ mapping T onto T is invertible (which is the case under the hypotheses of Theorem 7), the least squares solution is given by

$$\begin{aligned} M^{\text{Oracle}} &:= (\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Y) \\ &= M + (\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z). \end{aligned} \tag{III.11}$$

Hence

$$\|M^{\text{Oracle}} - M\|_F = \left\|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z)\right\|_F.$$

Let Z' be the minimal (normalized) eigenvector of $\mathcal{A}^*\mathcal{A}$ with minimum eigenvalue $\lambda_{\min}$, and set $Z = \delta\lambda_{\min}^{-1/2}\mathcal{A}(Z')$ (note that by definition $\mathcal{P}_\Omega(Z) = Z$ since Z is in the range of $\mathcal{A}$).[4] By construction, $\|Z\|_F = \delta$ and

$$\left\|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z)\right\|_F = \lambda_{\min}^{-1/2}\delta \gtrsim p^{-1/2}\delta$$

since by assumption, all the eigenvalues of $\mathcal{A}^*\mathcal{A} = \mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T$ lie in the interval $[p/2, 3p/2]$. The matrix Z defined above also maximizes $\|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z)\|_F$ among all matrices bounded by δ, and so the oracle achieves

$$\|M^{\text{Oracle}} - M\|_F \approx p^{-1/2}\delta \tag{III.12}$$

with adversarial noise. Consequently, our analysis loses a $\sqrt{n}$ factor vis-à-vis an optimal bound that is achievable via the help of an oracle.

The diligent reader may argue that the least squares solution above may not be of rank r (it is at most of rank $2r$) and may thus argue that this is not the strongest possible oracle. However, as explained below, if the oracle gave T and r, then the best fit in T of rank r would not do much better than (III.12). In fact, there is an elegant way to understand the significance of this oracle, which we now present. Consider a stronger oracle that reveals the row space of the unknown matrix M (and thus the rank of the matrix). Then we would know that the unknown matrix is of the form

$$M = M_C R^*$$

where M_C is an $n \times r$ matrix and R is an $n \times r$ matrix whose columns form an orthobasis for the row space (which we can build since the oracle gave us perfect information). We would then fit the nr unknown entries by the method of least squares and find $X \in \mathbb{R}^{n\times r}$ minimizing

$$\|\mathcal{P}_\Omega(XR^*) - \mathcal{P}_\Omega(Y)\|_F.$$

Using our previous notations, the oracle gives away $T_0 \subset T$, where T_0 is the span of elements of the form yv_k^*, $k \in [r]$, and is more precise. If $\mathcal{A}_0 : T_0 \to \Omega$ is defined by $\mathcal{A}_0 := \mathcal{P}_\Omega\mathcal{P}_{T_0}$, then the least squares solution is now

$$\left(\mathcal{A}_0^*\mathcal{A}_0\right)^{-1}\mathcal{A}_0^*(Y).$$

Because all the eigenvalues of $\mathcal{A}_0^*\mathcal{A}_0$ belong to $[\lambda_{\min}(\mathcal{A}^*\mathcal{A}), \lambda_{\max}(\mathcal{A}^*\mathcal{A})]$, the previous analysis applies, and this stronger oracle would also achieve an error of size about $p^{-1/2}\delta$. In conclusion, when all we know is $\|\mathcal{P}_\Omega(Z)\|_F \le \delta$, one cannot hope for a root-mean-squared (rms) error better than $p^{-1/2}\delta$.

Note that when the noise is stochastic, e.g., when Z_{ij} is white noise with standard deviation σ, the oracle gives an error bound that is adaptive and is smaller as the rank gets smaller. Indeed, $\mathbb{E}\|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z)\|_F^2$ is equal to

$$\sigma^2\text{trace}\left((\mathcal{A}^*\mathcal{A})^{-1}\right) \approx \frac{2nr - r^2}{p}\sigma^2 \approx \frac{2nr}{p}\sigma^2 \tag{III.13}$$

since all the $2nr - r^2$ eigenvalues of $(\mathcal{A}^*\mathcal{A})^{-1}$ are just about equal to p^{-1}. When $nr \ll m$, this is better than (III.12).

IV. NUMERICAL EXPERIMENTS

We have seen that matrix completion is stable amid noise. To emphasize the practical nature of this result, a series of numerical matrix completion experiments were run with noisy data. To be precise, for several values of the dimension n (our first experiments concern $n \times n$ matrices), the rank r, and the fraction of observed entries $p = m/n^2$, the following

[4] To clarify, Z' is itself a matrix, but it may be useful to think of it as vector with $n_1 n_2$ entries.

Table 1 RMS Error ($\|\hat{M} - M\|_F/n$) as a Function of n When Subsampling 20% of an $n \times n$ Matrix of Rank Two. Each RMS Error Is Averaged Over 20 Experiments

n	100	200	500	1000
RMS error	.99	.61	.34	.24

numerical simulations were repeated 20 times, and the errors averaged. A rank-r matrix M is created as the product of two rectangular matrices $M = M_L M_R^*$, where the entries of $M_L, M_R \in \mathbb{R}^{n\times r}$ are independent identically distributed (i.i.d.) $N(0, \sigma_n^2 := 20/\sqrt{n})$.[5] The sampled set Ω is picked uniformly at random among all sets with m entries. The observations $\mathcal{P}_\Omega(Y)$ are corrupted by noise as in (III.1), where $\{Z_{ij}\}$ is i.i.d. $N(0, \sigma^2)$; here, we take $\sigma = 1$. Lastly, $\hat{M}$ is recovered as the solution to (IV.1) below.

For a peek at the results, consider Table 1. The rms error defined as $\|\hat{M} - M\|_F/n$ measures the rms error per entry. From the table, one can see that even though each entry is corrupted by noise with variance one, when M is a 1000 by 1000 matrix, the rms error per entry is 0.24. To see the significance of this, suppose one had the chance to see *all* the entries of the noisy matrix $Y = M + Z$. Naively accepting Y as an estimate of M would lead to an expected MS error of $\mathbb{E}\|Y - M\|_F^2/n^2 = \mathbb{E}\|Z\|_F^2/n^2 = 1$, whereas the MS error achieved from only viewing 20% of the entries is $\|\hat{M} - M\|_F^2/n^2 = .24^2 = .0576$ when solving the SDP (IV.1). Not only are we guessing accurately the entries we have not seen but also we "denoise" those we have seen.

In order to stably recover M from a fraction of noisy entries, the following regularized nuclear-norm minimization problem was solved using the FPC algorithm from [27]:

$$\text{minimize} \quad \frac{1}{2}\|\mathcal{P}_\Omega(X - Y)\|_F^2 + \mu\|X\|_*. \qquad \text{(IV.1)}$$

It is a standard duality result that (IV.1) is equivalent to (III.2) for some value of μ, and thus one could use (IV.1) to solve (III.2) by searching for the value of $\mu(\delta)$ giving $\|\mathcal{P}_\Omega(\hat{M} - Y)\|_F = \delta$ (assuming $\|\mathcal{P}_\Omega(Y)\|_F > \delta$). We use (IV.1) because it works well in practice and because the FPC algorithm solves (IV.1) nicely and accurately. We also remark that a variation on our stability proof could also give a stable error bound when using the SDP (IV.1).

It is vital to choose a suitable value of μ, which we do with the following heuristic argument: first, simplifying to the case when Ω is the set of all elements of the matrix, note that the solution of (IV.1) is equal to Y but with singular values shifted towards zero by μ (soft-thresholding), as can be seen from the optimality conditions of Section II by means of subgradients; or see [9]. When Ω is not the entire set, the solution is no longer exactly a soft-thresholded version of Y but, experimentally, it is generally close. Thus, we want to pick μ large enough to threshold away the noise (keep the variance low) and small enough not to overshrink the original matrix (keep the bias low). To this end, μ is set

[5]The value of σ_n is rather arbitrary. Here, it is set so that the singular values of M are quite larger than the singular values of $\mathcal{P}_\Omega(Z)$ so that M can be distinguished from the null matrix. Having said that, note that for large n and small r, the entries of M are much smaller than those of the noise, and thus the signal appears to be completely buried in noise.

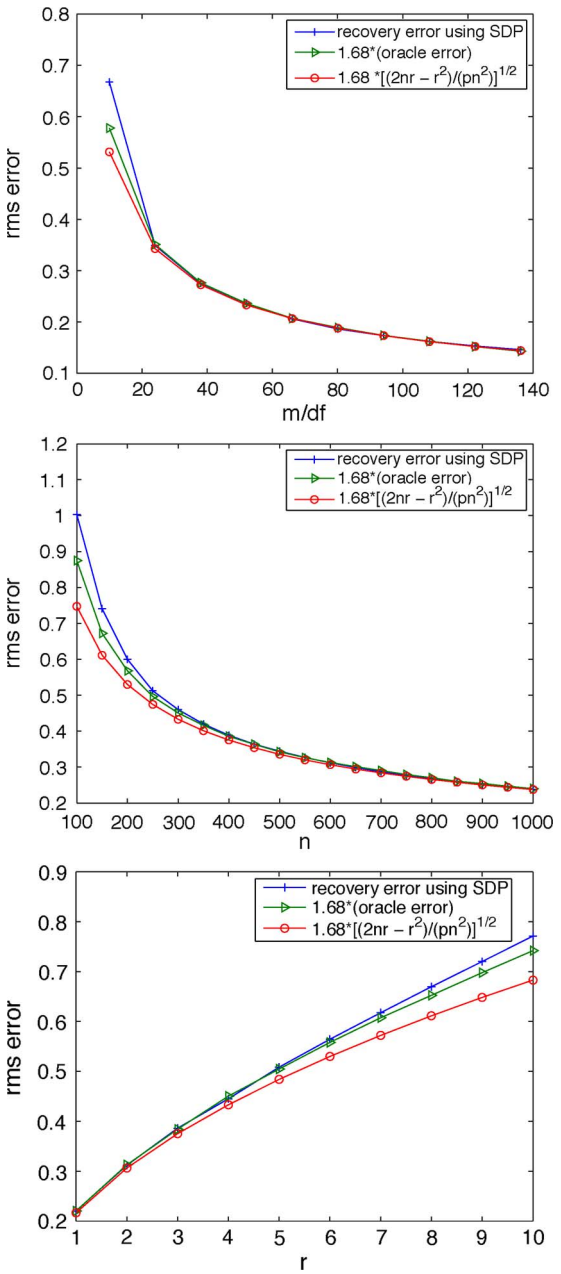

Fig. 2. *Comparison among the recovery error, the oracle error times 1.68, and the estimated oracle error times 1.68. Each point on the plot corresponds to an average over 20 trials. (Top) In this experiment, $n = 600$, $r = 2$, and p varies. The x-axis is the number of measurements per degree of freedom (df). Middle: n varies whereas $r = 2$, $p = 0.2$. (Bottom) $n = 600$, r varies, and $p = 0.2$.*

to be the smallest possible value such that if $M = 0$ and $Y = Z$. Then it is likely that the minimizer of (IV.1) satisfies $\hat{M} = 0$. It can be seen that the solution to (IV.1) is $\hat{M} = 0$ if $\|\mathcal{P}_\Omega(Y)\| \leq \mu$ (once again, check the subgradient or [9]). Then the question is: what is $\|\mathcal{P}_\Omega(Z)\|$? If we make a nonessential change in the way Ω is sampled, then the answer follows from random matrix theory. Rather than picking Ω uniformly at random, choose Ω by selecting each entry with probability p, independently of the others. With this modification, each entry of $\mathcal{P}_\Omega(Z)$ is i.i.d. with variance $p\sigma^2$. Then if $Z \in \mathbb{R}^{n\times n}$, it is known that $n^{-1/2}\|\mathcal{P}_\Omega(Z)\| \to \sqrt{2p}\sigma$, almost surely as $n \to \infty$. Thus we pick $\mu = \sqrt{2np}\sigma$, where $p = m/n^2$. In practice, this value of μ seems to work very well for square matrices. For $n_1 \times n_2$ matrices, based on the same considerations, the proposal is $\mu = (\sqrt{n_1} + \sqrt{n_2})\sqrt{p}\sigma$ with $p = m/(n_1 n_2)$.

In order to interpret our numerical results, they are compared to those achieved by the oracle; see Section III-B. To this end, Fig. 2 plots three curves for varying values of n, p, and r: 1) the rms error introduced above, 2) the rms error achievable when the oracle reveals T and the problem is solved using least squares, and 3) the estimated oracle root expected MS error derived in Section III-B, i.e., $\sqrt{\mathrm{df}/[n^2 p]} = \sqrt{\mathrm{df}/m}$, where $\mathrm{df} = r(2n - r)$. In our experiments, as n and m/df increased, with $r = 2$, the rms error of the nuclear norm problem appeared to be fit very well by $1.68\sqrt{\mathrm{df}/m}$. Thus, to compare the oracle error to the actual recovered error, we plotted the oracle errors times 1.68. We also note that in our experiments, the rms error was never greater than $2.25\sqrt{\mathrm{df}/m}$.

No one can predict the weather. We conclude the numerical section with a real-world example. We retrieved from the Web site [1] a 366×1472 matrix whose entries are daily average temperatures at 1472 different weather stations throughout the world in 2008. Checking its SVD reveals that this is an approximately low-rank matrix as expected. In fact, letting M be the temperature matrix and calling M_2 the matrix created by truncating the SVD after the top two singular values gives $\|M_2\|_F/\|M\|_F = .9927$.

We first tested whether the incoherence assumptions described above were satisfied. Since M_2 contained almost all of the energy in M, we measured μ_B in terms of the singular vectors of M_2 and found $\mu_B = 3.83$, which we considered to be small.

To test the performance of our matrix completion algorithm, we subsampled 30% of M and then recovered an estimate $\hat{M}$ using (IV.1). Note that this is a much different problem than those proposed earlier in this section. Here, we attempt to recover a matrix that is not exactly low rank, but only approximately. The solution gives a relative error of $\|\hat{M} - M\|_F/\|M\|_F = .166$. For comparison,[6] exact knowledge of the best rank-2 approximation achieves $\|M_2 - M\|_F/\|M\|_F = .121$. Here μ has been selected to give a good cross-validated error and is about 535.

V. DISCUSSION

This paper reviewed and developed some new results about matrix completion. By and large, low-rank matrix recovery is a field in complete infancy and abounding with interesting and open questions. If the recent avalanche of results in compressed sensing is any indication, it is likely that this field will experience tremendous growth in the next few years.

At an information-theoretic level, one would like to know whether one can recover low-rank matrices from a few general linear functionals, i.e., from $\mathcal{A}(M) = b$, where $\mathcal{A}$ is a linear map from $\mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$. In this direction, we would like to single out the original result of Recht *et al.* [29], who showed—by leveraging the techniques and proofs from the compressed sensing literature—that if each measurement is of the form $\langle A_k, X\rangle$, where A_k is an independent array of i.i.d. Gaussian variables (a la compressed sensing), then the nuclear norm heuristics recovers rank-r matrices from on the order of $nr \log n$ such randomized measurements.

At a computational level, one would like to have available a suite of efficient algorithms for minimizing the nuclear norm under convex constraints and, in general, for finding low-rank matrices obeying convex constraints. Algorithms with impressive performance in some situations have already been proposed [9], [27], but the computational challenges of solving problems with millions if not billions of unknowns obviously still require much research. ■

Acknowledgment

E. J. Candès would like to thank T. Tao and S. Becker for some very helpful discussions.

[6]The number two is somewhat arbitrary here, although we picked it because there is a large dropoff in the size of the singular values after the second. If, for example, M_{10} is the best rank-10 approximation, then $\|M_{10} - M\|_F/\|M\|_F = .081$.

REFERENCES

[1] National Climatic Data Center. [Online]. Available: http://www.ncdc.noaa.gov/oa/ncdc.html

[2] *IEEE Signal Processing Mag. (Special Issue on Sensing, Sampling, and Compression)*, vol. 25, Mar. 2008.

[3] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *Proc. 24th Int. Conf. Machine Learn.*, 2007.

[4] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," *Neural Inf. Process. Syst.*, 2007.

[5] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia, "Spectral analysis of data," in *Proc. 33rd Annu. ACM Symp. Theory Comput.*, New York, 2001, pp. 619–626.

[6] C. Beck and R. D'Andrea, "Computational study and comparisons of LFT reducibility methods," in *Proc. Amer. Contr. Conf.*, 1998.

[7] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, "Semidefinite programming based algorithms for sensor network localization," *ACM Trans. Sens. Netw.*, vol. 2, no. 2, pp. 188–220, 2006.

[8] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," submitted for publication.

[9] J.-F. Cai, E. J. Candès, and Z. Shen. (2008). A singular value thresholding algorithm for matrix completion, Tech. Rep. [Online]. Available: http://arxiv.org/abs/0810.3286

[10] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, to be published.

[11] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[12] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[13] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when *p* is much larger than *n*," *Ann. Statist.*, vol. 35, 2007.

[14] E. J. Candès and T. Tao. (2009). The power of convex relaxation: Near-optimal matrix completion, Tech. Rep., submitted for publication. [Online]. Available: http://arxiv.org/abs/0903.1476

[15] P. Chen and D. Suter, "Recovering the missing components in a large noisy low-rank matrix: Application to SFM source," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 8, pp. 1051–1063, 2004.

[16] A. L. Chistov and D. Y. Grigoriev, "Complexity of quantifier elimination in the theory of algebraically closed fields," in *Proc. 11th Symp. Math. Found. Comput. Sci.*, vol. 176, *Lecture Notes in Computer Science*, 1984, pp. 17–31.

[17] D. Donoho, "For most large underdetermined systems of linear equations, the minimal L1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, Jun. 2006.

[18] D. L. Donoho and J. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, to be published.

[19] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2002.

[20] M. Fazel, E. Candès, B. Recht, and P. Parrilo, "Compressed sensing and robust recovery of low rank matrices," in *Proc. 2008 42nd Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, Oct. 2008.

[21] M. Fazel, H. Hindi, and S. Boyd, "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices," in *Proc. Amer. Contr. Conf.*, Jun. 2003.

[22] A. Gilbert, S. Muthukrishnan, and M. Strauss, "Improved time bounds for near-optimal sparse Fourier representation," in *Proc. Wavelets XI SPIE Opt. Photon.*, San Diego, CA, 2005.

[23] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, pp. 61–70, 1992.

[24] R. Keshavan, S. Oh, and A. Montanari, "Matrix completion from a few entries," in *Proc. ISIT'09*, 2009, submitted for publication.

[25] R. H. Keshavan, A. Montanari, and S. Oh. (2009). Matrix completion from noisy entries. [Online]. Available: http://arxiv.org/abs/0906.2027

[26] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," submitted for publication.

[27] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and Bregman iterative methods for matrix rank minimization," Tech. Rep., 2008.

[28] M. Mesbahi and G. P. Papavassilopoulos, "On the rank minimization problem over a positive semidefinite linear matrix inequality," *IEEE Trans. Autom. Control*, vol. 42, no. 2, pp. 239–243, 1997.

[29] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization," submitted for publication.

[30] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[31] A. Singer, "A remark on global positioning from local distances," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 28, pp. 9507–9511, 2008.

[32] A. Singer and M. Cucuringu, "Uniqueness of low-rank matrix completion by rigidity theory," submitted for publication.

[33] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. Computer. Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[34] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1417–1428, 2002.

[35] C. C. Weng and P. P. Vaidyanathan, "Matrix completion for DOA estimation," Unpublished.

ABOUT THE AUTHORS

Emmanuel J. Candès received the B.Sc. degree from the Ecole Polytechnique, France, in 1993 and the Ph.D. degree in statistics from Stanford University, Stanford, CA, in 1998.

He is a Professor of Mathematics and of Statistics at Stanford University, and is the Ronald and Maxine Linde Professor of Applied and Computational Mathematics at the California Institute of Technology, Pasadena (on leave). His areas of interest are in mathematical signal processing, information theory, scientific computing, and mathematical optimization.

Dr. Candès was awarded the James H. Wilkinson Prize in Numerical Analysis and Scientific Computing by SIAM in 2005. He received the U.S. National Science Foundation's highest honor in 2006: the Alan T. Waterman Award. He received the 2008 Information Theory Society Paper Award, and is a recipient of the George Polya Prize awarded by SIAM.

Yaniv Plan received the B.A. degree from the University of California, Berkeley, in 2004, where he double majored in applied math and physics and received honors in applied math. Currently, he is working towards the Ph.D. degree at the Applied and Computational Math Department, California Institute of Technology, Pasadena, working under the supervision of Dr. Candès.

His areas of interest are in probability, statistics, compressive sensing, sparse reconstruction, and low-rank matrix recovery.