Matrix Completion With Data-Dependent Missingness Probabilities

Sohom Bhattacharya and Sourav Chatterjee

***Abstract*—The problem of completing a large matrix with lots of missing entries has received widespread attention in the last couple of decades. Two popular approaches to the matrix completion problem are based on singular value thresholding and nuclear norm minimization. Most of the past works on this subject assume that there is a single number p such that each entry of the matrix is available independently with probability p and missing otherwise. This assumption may not be realistic for many applications. In this work, we replace it with the assumption that the probability that an entry is available is an unknown function f of the entry itself. For example, if the entry is the rating given to a movie by a viewer, then it seems plausible that high value entries have greater probability of being available than low value entries. We propose two new estimators, based on singular value thresholding and nuclear norm minimization, to recover the matrix under this assumption. The estimators involve no tuning parameters, and are shown to be consistent under a low rank assumption. We also provide a consistent estimator of the unknown function f.**

***Index Terms*—Matrix completion, graph limits, missing not at random.**

I. Introduction

Let M be an $m \times n$ matrix, which is only partially observed, possibly with added noise. Given an estimate $\hat{M}$ of M, we define its mean squared error as

$$\mathrm{MSE}(\hat{M}) := \mathbb{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{m}_{ij}-m_{ij})^2\right], \qquad (1)$$

where m_{ij} and $\hat{m}_{ij}$ denote the (i, j)-th entries of M and $\hat{M}$ respectively. Given a sequence of such estimation problems, where M_k and $\hat{M}_k$ denote the parameter and estimator matrices of the k-th problem, we call the sequence of estimators $\hat{M}_k$ consistent if

$$\lim_{k\to\infty} \mathrm{MSE}(\hat{M}_k) = 0.$$

Estimating a large matrix from a few randomly selected (and possibly noisy) entries is a common objective in many statistical problems. The basic assumption in all of the work in this area is that the matrix has either low rank or is approximately of low rank in some suitable sense. Some of the prominent applications of matrix completion include compressed sensing [8]–[12], [19], collaborative filtering [6], [42], multi-class learning [2], [40], dimension reduction [31], [46] and subspace estimation [7]. Theoretical guarantees of matrix completion under various assumptions have been worked out in [1], [3], [4], [13], [17], [18], [24], [26]–[29], [34], [37], [38], [43]. This is only a small sampling of the huge literature on this topic. For a recent survey, see [39].

In many of the above works, it is assumed that the entries are missing uniformly at random. This may not be a realistic assumption in many applications. For example, in the classic problem of movie ratings, if a particular movie gets poor reviews, fewer numbers of viewers are expected to review it and hence the probability of missing entries corresponding to that particular movie would be higher. Work on matrix completion under the ‘missing not at random’ (MNAR) assumption is relatively sparse. Some examples include deterministic missing patterns or missing patterns that depend on the matrix, using spectral gap conditions [5], rigidity theory [44], algebraic geometry [25] and other methods [30], [41], [45]. For random but non-uniform missing patterns, a variety of statistical guarantees for procedures based on nuclear-norm penalization and other ideas are available [9], [16]–[18], [20], [26], [29], [33], [38], [43], [49]. However, these guarantees almost always require a careful choice of the penalty parameter (or some other parameter, such as rank) based on knowledge about the unknown matrix that is unlikely to be available. This is in contrast to the case of uniform missing pattern, where we now have many algorithms that assume no knowledge of the unknown matrix.

In the present work, we assume that the probability of an entry being revealed is a function f of the value of that entry, and the revealed entries are allowed to be noisy. This frequently encountered example of missingness where a variable governs its own missingness is known as *self-masking MNAR* [36]. Under these assumptions, we provide an estimator of the parameter matrix based on a spectral method and prove its consistency under a low rank assumption. We also provide a second estimator based on nuclear norm minimization. This estimator performs significantly better than the spectral estimator in the absence of noise, but may not work well for noisy entries. Moreover, it is computationally expensive for large matrices. Lastly, we give estimates of the function f using both methods, along with theoretical guarantees about it. Some numerical examples are worked out. The main advantage of our estimators is that they do

Manuscript received 8 August 2021; revised 14 February 2022; accepted 14 April 2022. Date of publication 25 April 2022; date of current version 15 September 2022. The work of Sourav Chatterjee was supported in part by NSF under Grant DMS-1855484. *(Corresponding author: Sohom Bhattacharya.)*

The authors are with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: sohomb@stanford.edu; souravc@stanford.edu).

Communicated by C. Suh, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2022.3170244

not involve penalty parameters (or any other user-specified parameters) which have to be carefully chosen to ensure that the theoretical guarantees work out. The cost is that we have asymptotic consistency results rather than finite sample error bounds.

A recent paper that works under the setting of self-masking MNAR, but in the setting of tensor completion, is [47]. In [47], the probabilities of missingness are called 'propensity scores'. The main difference between [47] (and similar papers) and our work is that in [47], it is assumed that the tensor of propensity scores is low-rank, while we make no such assumption. Indeed, one of the main observations in our paper, which we prove using spectral techniques reminiscent of the proof of Szemerédi's lemma in combinatorics, is that the matrix of propensity scores is guaranteed to have an approximately low rank structure under a Lipschitz assumption on f.

A natural extension of our work is to study beyond self-masking MNAR, namely, to consider examples where the process that causes the missingness of an entry depends on multiple entries of the parameter matrix and not only its value itself. Such directions are left for future research.

II. Results

A. The Problem

Let M be an $m \times n$ matrix with all entries in the interval $[-1, 1]$. Let $f : [-1, 1] \to [0, 1]$ be a function. Let X be a noisy version of M, modeled as a matrix with independent entries in $[-1, 1]$, such that $\mathbb{E}(x_{ij}) = m_{ij}$ for each i and j. The (i, j)-th entry of X is revealed with probability $f(m_{ij})$, and remains hidden with probability $1 - f(m_{ij})$, and these events occur independently. Our goal is to estimate M using the observed entries of X.

B. Modified USVT Estimator

Our first proposal is an estimator of M based on singular value thresholding. This is a modification of the Universal Singular Value Thresholding (USVT) estimator of [14]. The estimator is defined as follows:

1) Let Y be the matrix whose (i, j)-th entry is x_{ij} if the (i, j)-th entry of X is revealed, and 0 otherwise.
2) Let $\sum \sigma_i u_i v_i^T$ be the singular value decomposition of Y.
3) Choose a positive number $\eta \in (0, 1)$ and let

$$A = \sum_{i:\sigma_i \geq (2+\eta)\max\{\sqrt{m},\sqrt{n}\}} \sigma_i u_i v_i^T.$$

(In [14], it is recommended that η be chosen to be 0.02. For results concerning the optimal choice of the threshold, see [18].)
4) Truncate the entries of A to force them to belong to the interval $[-1, 1]$. Call the resulting matrix $\hat{Q}$.
5) Let P be the matrix whose (i, j)-th entry is 1 if x_{ij} is revealed, and 0 otherwise.
6) Repeat the above steps for the matrix P instead of Y, to get $\hat{R}$.
7) Define a matrix W as $w_{ij} := \hat{q}_{ij}/\hat{r}_{ij}$ if $\hat{r}_{ij} \neq 0$, and 0 otherwise.
8) Truncate the entries of W to force them to be in $[-1, 1]$. The resulting matrix is our estimator $\hat{M}$.

The idea behind this estimator has some similarity with the one proposed recently by Ma and Chen [33], which is also based on a two-step procedure, first estimating the matrix of missingness probabilities and then using these estimated probabilities to estimate the unknown matrix. The algorithm of Ma and Chen involves a number of user-specified parameters, whereas ours does not, which may be a desirable feature.

Note that if the entries of M and X are known to belong to an interval $[a, b]$ instead of $[-1, 1]$, then subtracting $(a + b)/2$ from each entry of X and dividing by $(b-a)/2$ forces the entries to lie in $[-1, 1]$. Then applying the above procedure, and finally multiplying the end-result by $(b-a)/2$ and adding $(a + b)/2$, we can get the desired estimate of M. The case of unknown a, b is beyond the scope of the paper. Lastly, if $n > m$, one can simply work with the transpose of X to get an estimate for the transpose of M.

C. Modified Candès–Recht Estimator

Our second proposal is an estimator of M based on nuclear norm minimization. This estimator works only in the absence of noise, so we assume that $X = M$. Let $\hat{M}$ be the matrix that minimizes nuclear norm among all matrices that are equal to M at the revealed entries, and have all entries in $[-1, 1]$. (Recall that the nuclear norm of a matrix M, usually denoted by $\|M\|_*$, is the sum of its singular values.) Hence, given a set of observed entries Ω, our estimator is obtained by solving the optimization problem:

$$\hat{M} := \operatorname{argmin}_{Z \in S} \|Z\|_*,$$

where

$$S := \{Z : (Z - M)_{ij} \mathbb{1}_{(i,j)\in\Omega} = 0,\ \|Z\|_\infty \leq 1\}.$$

This is a small modification of the popular Candès–Recht estimator [9], [10], [12], suggested recently in [15]. The original estimator does not have the additional constraint that the entries of $\hat{M}$ have to be in $[-1, 1]$. This extra constraint is not problematic since this is a convex constraint. For example, it can be easily implemented in R by adding an ℓ^∞ constraint using CVXR package [21]. Moreover, from an intuitive point of view, it makes sense to add this constraint since we already know that the entries of the unknown matrix M are in $[-1, 1]$. This estimator is similar to the one proposed by Klopp [26], except that our method does not involve a penalty parameter.

D. Consistency Results

We now state consistency results for the two estimators defined above. Suppose that we have a sequence of matrices $\{M_k\}_{k\geq 1}$, where M_k has order $m_k \times n_k$, and $m_k, n_k \to \infty$ as $k \to \infty$. Let $\{X_k\}_{k\geq 1}$ be a sequence of random matrices with independent entries in $[-1, 1]$ such that $\mathbb{E}(X_k) = M_k$ for each k. In other words, X_k is a noisy version of M_k. Let $\mathcal{M}$ be the union of the sets of entries of all of these matrices. Let $f : \mathcal{M} \to [0, 1]$ be a function such that the noisy version of an entry with true value m is revealed with probability $f(m)$, independently of all else. Note that it is irrelevant how f is defined outside $\mathcal{M}$, which is why we took the domain of f to be this countable set.

Recall that a sequence of estimators $\{\hat{M}_k\}_{k\geq 1}$ is consistent if $\mathrm{MSE}(\hat{M}_k) \to 0$ as $k \to \infty$, where MSE stands for the mean squared error defined in equation (1). We will now prove the consistencies of the two estimators defined above. The crucial assumption will be that the sequence $\{M_k\}_{k\geq 1}$ has *uniformly bounded rank*. This is a version of the frequently occurring *low rank assumption* from the literature. In addition to that, we will need some other technical assumptions. Our first result is the following theorem, which gives a sufficient condition for the consistency of the modified USVT estimator.

Theorem 2.1: In the above setup, suppose that the sequence $\{M_k\}_{k\geq 1}$ has uniformly bounded rank. Let μ_k be the empirical distribution of the entries of M_k. Suppose that for any subsequential weak limit μ of the sequence $\{\mu_k\}_{k\geq 1}$, there is an extension of f to a Lipschitz function from $[-1, 1]$ into $[0, 1]$, also denoted by f, which has no zeros in the support of μ. Then the modified USVT estimator based on $\{X_k\}_{k\geq 1}$ is consistent.

Remark 2.2: The statement of the above Theorem is about asymptotic behavior of MSE. However, in our proofs, we obtain some finite sample error bounds which we have omitted, with the goal of increasing the readability of the result, and also for reducing the stringency of assumptions on f. In fact, the proof shows that if $\|M\|_* \leq q\sqrt{mn}$ for some $q > 0$, and $f \geq \delta$ everywhere for some $\delta > 0$, and f is a Lipschitz function with Lipschitz constant $L > 0$, then for any $\varepsilon > 0$, the MSE can be upper bounded by

$$\frac{12}{\delta^2}\left(c_1 \min\left\{2\sqrt{\frac{r}{m}} + \varepsilon L + \sqrt{2\varepsilon(L+1)},\, 2\right\} + 2c_2 e^{-c_3 n}\right),$$

where r is a constant depending on q and ε, and c_1, c_2, and c_3 are universal constants. Such a bound reveals how the magnitude of error is dependent on the nuclear norm of parameter matrix and the Lipschitz constant of f.

Remark 2.3: Note that in many examples, such as in most recommender systems, the matrix entries can only take values in a fixed finite set. In such examples, there is no loss of generality in the assumption that f has an extension that is Lipschitz and nonzero everywhere on $[-1, 1]$. Also, if f is continuous and nonzero everywhere in $[-1, 1]$, then the condition involving the empirical distribution of the entries is redundant.

The next theorem gives the consistency of the modified Candès–Recht estimator, under the additional assumption that there is no noise.

Theorem 2.4: In the above setup, suppose that the sequence $\{M_k\}_{k\geq 1}$ has uniformly bounded rank, and also suppose that $X_k = M_k$ for each k. Let μ_k be the empirical distribution of the entries of M_k. Suppose that for any subsequential weak limit μ of the sequence $\{\mu_k\}_{k\geq 1}$, there is an extension of f to a measurable function from $[-1, 1]$ into $[0, 1]$, also denoted by f, such that f is nonzero and continuous almost everywhere with respect to μ. Then the modified Candès–Recht estimator is consistent for this problem.

Remark 2.5: We will see in numerical examples that the modified Candès–Recht estimator has superior performance. The advantage of the modified USVT estimator is twofold. First, it can be used when the matrix is very large, where using nuclear norm minimization may become infeasible due to computational cost. Second, in the presence of noise — which is often the case in practice — the modified Candès–Recht estimator may perform badly, as we will see in the simulated and real data examples.

Remark 2.6: Often, in many MNAR examples, identifiability of parameters is an issue (see, e.g., [35]), which corresponds to the notions that there might be two sets of parameter values which yield same observations and hence, the true parameter value cannot be identified. In Theorems 2.1 and 2.4, however, the fact that we are able to approximately recover the true matrix automatically implies that identifiability is not an issue, provided that the low rank assumption holds. (That is, if there are two candidates M_1 and M_2 for the true matrix, and they both have low rank, then our estimate $\hat{M}$ will be close to both M_1 and M_2 with high probability, which means that M_1 must be close to M_2.)

E. Proof Sketch

To prove Theorem 2.1, we first assume that μ_k converges weakly to a limit μ as $k \to \infty$. Let R_k be the matrix obtained by applying f entrywise to M_k and Q_k be entrywise product of M_k and R_k. Let Y_k be the matrix obtained by replacing the unrevealed entries of X_k by zero. Let P_k be the matrix whose (i, j)-th entry is 1 if the (i, j)-th entry of X_k is revealed, and 0 otherwise.

The main step is to show that R_k and Q_k are also approximately low rank matrices, in the sense that $\|R_k\|_* = o(m_k\sqrt{n_k})$ and $\|Q_k\|_* = o(m_k\sqrt{n_k})$. This is proved using a spectral method, similar to the spectral proof of Szemerédi's regularity lemma. The key idea is that a low rank matrix is approximately a block matrix after a suitable permutation of rows and columns, and therefore, applying a Lipschitz function entrywise keeps it close to a block matrix, which, in turn, is approximately low rank. Once this is established, it then follows by the standard results for USVT that if $\hat{Q}_k$ and $\hat{R}_k$ are the estimates of Q_k and R_k obtained by applying the USVT algorithm to Y_k and P_k, then $\hat{Q}_k \approx Q_k$ and $\hat{R}_k \approx R_k$ with high probability (in some appropriate sense).

To prove Theorem 2.4, we first show that one can possibly permute rows and columns in each M_k to get an L^2 limit W. Next we prove there is a measurable function $V : [0, 1]^2 \to [0, 1]$ that is nonzero almost everywhere and P_k converges to V in cut distance almost surely subsequentially. This implies consistency of $\hat{M}_k$ by [15, Theorem 2 and Theorem 3].

F. Estimating f

We will now produce an estimator for the unknown function f that can be used with any consistent estimator. Our procedure is motivated by the nonparametric density estimation methods available in statistics literature. It is interesting to note, if the underlying function f were indeed a constant function, we have observed from simulated examples that our estimator $\hat{f}^b$ is also close to a constant function. Hence, $\hat{f}^b$ can be used to check if the data are MNAR or not. The estimator involves the choice of a tuning parameter b, which is a positive integer, chosen by the user. Given a matrix M with partially revealed entries as in Subsection II-A, and an estimator $\hat{M}$ of M, the estimator $\hat{f}^b$ of f is defined as follows.

1) For $i = 1, \ldots, 2b + 3$, let $c_i := -1 + (i - 2)b^{-1}$. Note that this is a sequence of equally spaced points, starting at $c_1 = -1 - b^{-1}$ and going up to $c_{2b+3} = 1 + b^{-1}$.
2) For each i, choose a_i uniformly at random from the interval $[c_i - (4b)^{-1}, c_i + (4b)^{-1}]$.
3) In the interval $[a_i, a_{i+1}]$, define $\hat{f}^b$ to be the proportion of revealed entries among those entries of M such that the corresponding entry of $\hat{M}$ is in $[a_i, a_{i+1}]$.

Note that the above procedure defines $\hat{f}^b$ on an interval that is slightly larger than $[-1, 1]$, but that should not bother us, because the domain can then be restricted to $[-1, 1]$. The following theorem gives a measure of the performance of $\hat{f}^b$ as an estimate of f.

Theorem 2.7: Suppose that f is Lipschitz, with Lipschitz constant L. Let μ be the empirical distribution of the entries of M and $\theta := \mathrm{MSE}(\hat{M})$. Then

$$\int (\hat{f}^b(x) - f(x))^2 d\mu(x) \le C\theta^{1/3} b^{5/3} + \frac{Cb}{mn} + \frac{CL^2}{b^2},$$

where C is a universal constant.

The above result shows that if b is big, but much smaller than both mn and $\theta^{-1/5}$, then $\hat{f}^b$ is close to f at almost all entries of M. In practice, a good rule of thumb would be to choose b such that b is large, but at the same time, the intervals $[a_l, a_{l+1})$ contain substantial numbers of entries of $\hat{M}$. One can try to choose b optimally using some kind of cross-validation (such as leave-one-out cross-validation), but it may be hard to prove theoretical guarantees for such methods.

Although our method of estimating f has similarities with density estimation methods, the problem is quite different since the entries of the estimated matrix are not independent random variables — in fact, they may have a complicated, or even intractable, dependence structure. One might wonder if traditional nonparametric methods of estimating f can still be applied here under some smoothness constraint. Such questions are left for future investigation.

G. Examples

In this subsection we will see how the two estimators perform in some simulated examples and two real data examples. For real data examples, one should always check whether the matrix is low-rank approximable before applying our methods. Our simulations show taking b of order $\sqrt{n}$ for estimating an $n \times n$ matrix yields good $\hat{f}$, although we do not have a theorem to prove that. Finally, one should also check if data is noisy or not, and should apply spectral estimator when noise is present.

TABLE I

COMPARISON TABLE FOR EXAMPLE 2.8

	Modified USVT	Modified Candès–Recht
MSE	0.123	$\sim 10^{-9}$
Run-time	0.31 sec	4.08 min

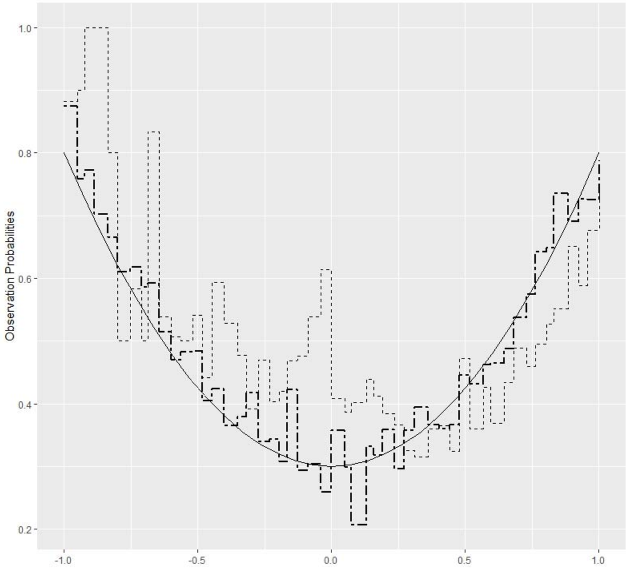

Fig. 1. Estimates of f in Example 2.8. The dashed curve corresponds to modified USVT estimator, the double-dashed curve corresponds to the modified Candès-Recht estimator, and the solid curve is the true f.

Example 2.8: Consider a low rank $n \times n$ matrix M with the entries of M having marginal distribution $Uniform[-1, 1]$. Here, we take $n = 100$ and $\mathrm{rank}(M) = 7$. To generate such a matrix, we define $M_1 = \sum_{i=1}^{6} d_i u_i v_i^T$, where:

- For $i = 1, \ldots, 5$, $d_i = 2^{-i}$, and the components of u_i and v_i are i.i.d. $Bernoulli(1/\sqrt{2})$ random variables.
- $d_6 = 1$, u_6 is a vector of all 1s, and v_6 has i.i.d $Uniform[0, 2^{-5}]$ entries.

It is not difficult to see that the entries of M_1 are i.i.d. $Uniform[0, 1]$ random variables. Multiplying each entry by 2 and subtracting 1, we get M. Then M has rank 7 with probability 1, and the entries of M are uniformly distributed in $[-1, 1]$. We take $f(x) = 0.5x^2 + .3$ to generate missing entries, and do not add noise. To obtain the modified Candès–Recht estimator, we used code from the R package `filling` [48] and imposed the ℓ^∞ constraint using the `CVXR` package [21]. The modified USVT algorithm, being quite straightforward, was coded without the aid of existing packages.

The modified Candès–Recht estimator was able to exactly recover the true M almost all the time, resulting a very small MSE of order 10^{-9}. The modified USVT estimator performed much worse, with an unimpressive MSE of 0.123. The run-time of the modified USVT estimator was much lower than that of the modified Candès–Recht estimator: 0.31 seconds versus 4.08 minutes. We will see in the next example that the performance of the modified USVT estimator becomes better when n is larger, accompanied by a huge gain in run-time over the other estimator. We report both our estimators and their MSEs and run-times in Table I.

Next, for both estimators of M, we estimated f using the method proposed in Section II-F, taking $b = 25$. The estimated

TABLE II
COMPARISON TABLE OF MSE FOR EXAMPLE 2.9

Rank	Modified USVT	Regular USVT
4	0.007	0.060
6	0.012	0.063
8	0.012	0.063
10	0.033	0.062

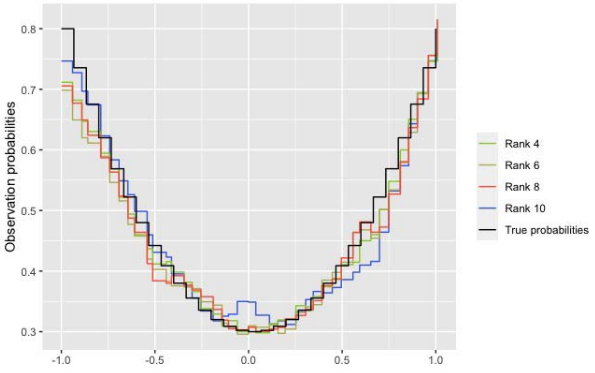

Fig. 2. Estimates of f in Example 2.9.

TABLE III
COMPARISON TABLE FOR EXAMPLE 2.10

	Modified USVT	Modified Candès–Recht
MSE	0.011	$\sim 10^{-9}$
Run-time	0.85 sec	2.51 hrs

$\hat{f}$'s are shown in Figure 1. As expected, the $\hat{f}$ based on the modified Candès–Recht estimator has better performance.

Example 2.9: Here, we want to see how our estimator performs as we vary the rank of underlying parameter matrix. To this end, we take parameter matrix same as previous example, with $n = 500$ and choose rank $r = 4, 6, 8, 10$. We only report result of the modified USVT estimator, the results corresponding to modified Candés- Recht estimator varies similarly. The MSE of the estimator as we vary rank are $0.007, 0.012, 0.012, 0.033$ respectively. The estimated $\hat{f}$ is shown in Figure 2. We also observe our estimator performs better than the vanilla USVT algorithm developed for MCAR. A comparison of MSE of the two estimators has been given below in Table II.

Example 2.10: This is the same as Example 2.8, but with $n = 500$ to show how modified USVT has significant computation time advantage over the modified Candès-Recht estimator. The MSE of the modified USVT estimator is now 0.011, and that of the modified Candès-Recht estimator is of order 10^{-9}. So, with this larger sample size, the modified USVT estimator has reasonably good performance. The time to compute the modified USVT estimator 0.85 seconds, whereas for the modified Candès–Recht estimator, it is 2.51 hours. This shows that even though the latter has much better performance in terms of MSE, it may be more practical to use the former if the matrix is large. We provide the estimators of f in Figure 3, taking $b = 25$. We report the MSEs and run-times for both estimators in Table III.

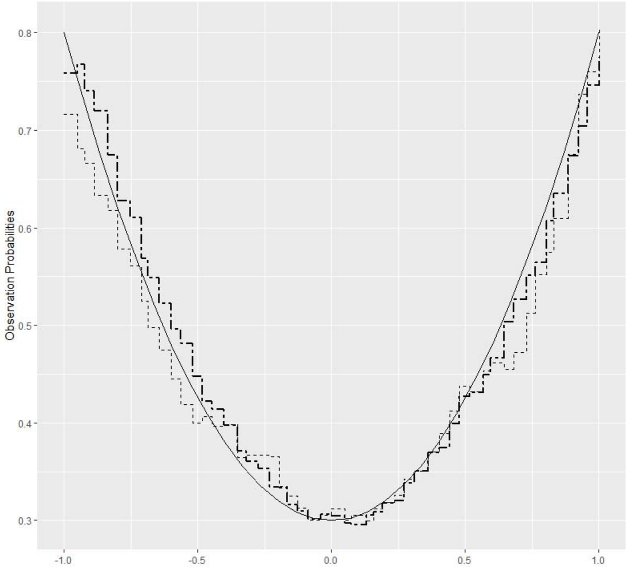

Fig. 3. Estimates of f in Example 2.10. The dashed curve corresponds to modified USVT estimator, the double-dashed curve corresponds to the modified Candès–Recht estimator, and the solid curve is the true f.

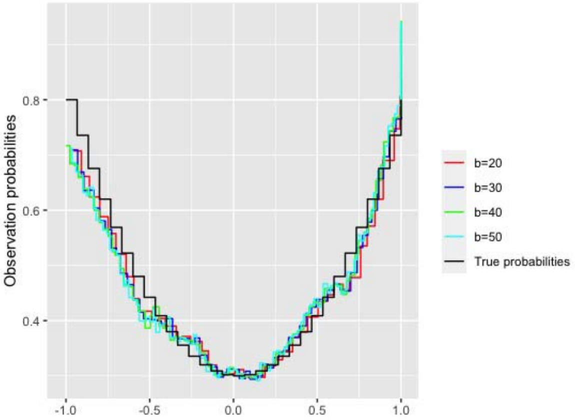

Fig. 4. Estimating $\hat{f}$ under different values of b.

A visual examination shows that both estimators perform well.

Example 2.11: Under the same setup as before, we now show how the change of the parameter b, number of bins, affect the estimate of underlying function f. We choose $b = 20, 30, 40, 50$ and plot the resulting $\hat{f}$ in Figure 4. There does not seem to have much difference in $\hat{f}$ across different values of b.

Example 2.12: We will now show that the modified Candès–Recht estimator performs poorly under presence of noise. Here, we take $n = 100$ and $\text{rank}(M) = 2$, with the marginal distribution of the entries of M being $Uniform[0, 1]$, generated by the same procedure that we used to generate M_1 in Example 2.8. The noisy version of M, namely X, is generated as follows. For each (i, j), generate $x_{ij} = 1$ with probability m_{ij} and $x_{ij} = 0$ with probability $1 - m_{ij}$. Note that $\mathbb{E}(x_{ij}) = m_{ij}$. The entry

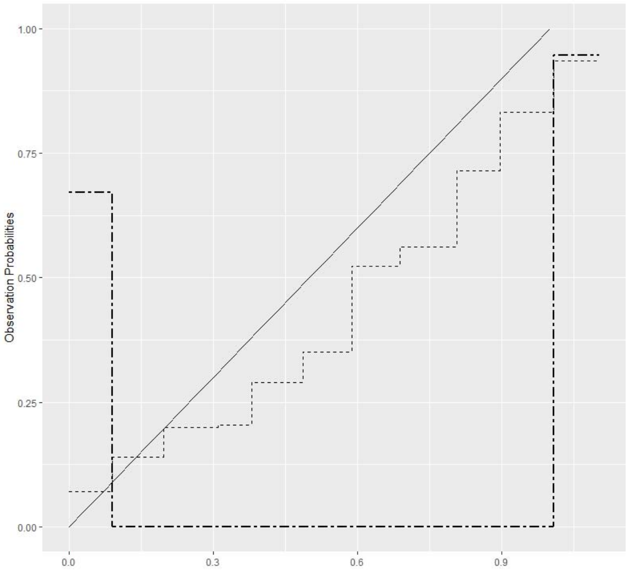

Fig. 5. Estimation of f in Example 2.12. The dashed curve corresponds to modified USVT estimator, the double-dashed curve corresponds to the modified Candès-Recht estimator, and the solid curve is the true f.

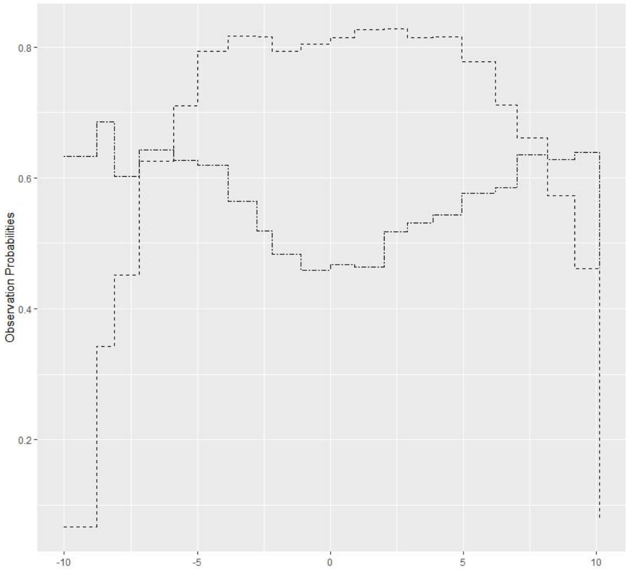

Fig. 6. Estimation of f in Example 2.13. The dashed curve corresponds to the modified USVT estimator and the double-dashed curve corresponds to the modified Candès-Recht estimator.

x_{ij} is revealed with probability m_{ij}, and remains hidden with probability $1 - m_{ij}$ (that is, we took $f(x) = x$). For $n = 100$, the MSE of modified USVT estimator turned out to be 0.017, much better than the MSE of the modified Candès–Recht estimator, which was 0.112. The estimates of f based on the two methods, with $b = 10$, are depicted in Figure 5. The estimate based on the modified USVT method is reasonably good, even with n as small as 100 in this example. The estimate based on the modified Candès–Recht estimator, however, is completely off: It estimates f to be large near 0 and 1 and zero everywhere in between. This is because the observed entries consist solely of zeros and ones, and $\hat{M}$ coincides with X at the observed values. So the estimation procedure for $\hat{f}$ deduces, incorrectly, that there is no chance of observing an entry if its non-noisy value is strictly between 0 and 1.

Example 2.13: We now consider a real data example. In real data, it is not possible to compare the performance of $\hat{f}$ with the 'true f', because we do not know what the true f is (or if our model is actually valid). Still, if $\hat{f}$ turns out to be substantially different than a constant function, it validates the viewpoint that entries are not missing uniformly at random. We consider the well-known Jester data [22], which consists of 100 jokes rated by $73{,}421$ users. The ratings are continuous values between -10 and 10, entered by the users by clicking on an on-screen 'funniness' bar. Not every user rates every joke, so there are many missing entries. Due to the prohibitively large run-time of the modified Candès–Recht estimator, we first took a submatrix consisting of all 100 jokes but a random sample of 300 users. Approximately 45% of the values were missing in this submatrix. The estimates of f based on the two methods (with $b = 10$) are shown in Figure 6.

Interestingly, the two estimates are very different. We posit that this is due to the presence of noise in the observed matrix, which messes up the modified Candès–Recht estimator. Indeed, the continuous nature of the ratings makes it very unlikely that the observed matrix is without noise. This is further validated by Figure 7, where we plot the percentage of the modified Candès-Recht estimator matrix $\hat{M}$ that is captured by its rank-k approximation, $k = 1, 2, \ldots, 100$. (The percentage is simply the sum of squares of the top k singular values divided by the sum of squares of all singular values.) This figure shows that to even get within 80% of $\hat{M}$, we need to consider a rank-25 approximation. Thus, $\hat{M}$ is not of low rank, even approximately. This invalidates the low rank assumption of the Candès–Recht procedure, and allows us to conjecture that the $\hat{f}$ given by the modified USVT estimator is a better reflection of the true f, assuming that the model is correct.

Example 2.14: We continue with the Jester data example. Assuming that the $\hat{f}$ given by the modified USVT estimator reflects the true state of affairs, we ran the modified USVT method on the whole dataset. The estimated f, with $b = 70$, is shown in Figure 8. The inverted U-shape is mysterious. It is not clear to us what may have led to this, if it is indeed close to the true f, because we do not know what caused entries to be missing in this dataset.

Example 2.15: For our final example, we consider the Film Trust dataset of movie ratings [23]. This dataset consists of ratings given by 1508 users to 2071 movies, with many missing entries. The user ratings range in the set $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$. This dataset is much sparser than the Jester data; only 35497 ratings are available, which is about 1.13 percent of the total number of possible ratings. Due to the large size of the dataset, we implemented only the modified USVT algorithm. We assume that each user has a 'true' rating for each movie, and the observed rating, if any, is a noisy version of the true rating. The observation probability is then

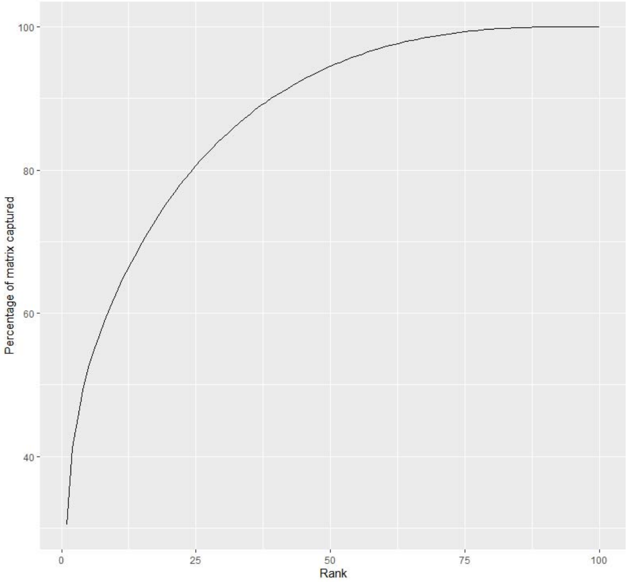

Fig. 7. Let $\hat{M}$ be the modified Candès–Recht estimate of M in Example 2.13. This graph shows that percentage of $\hat{M}$ that is captured by its rank-k approximation, $k = 1, 2, \dots, 100$.

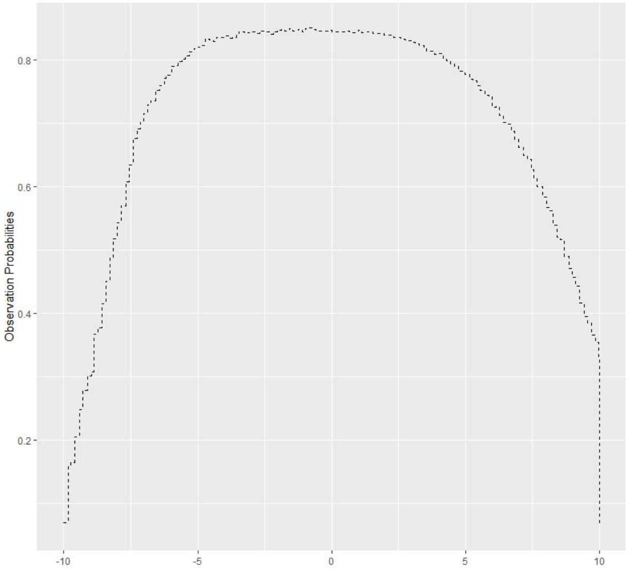

Fig. 8. Estimation of f in Example 2.14 using the modified USVT estimator.

a function f of the true rating. The estimate of f, with $b = 30$, is plotted in Figure 9. As expected, a high rating increases the chance of the rating being available; however, there is a dip towards the end of the curve which we do not know how to explain. One possible explanation is that very highly rated movies are often classics that not many people watch and rate because they have already watched those movies before.

III. Proof of Theorem 2.4

For an $m \times n$ matrix A, define

$$\|A\|_2 := \left(\frac{1}{mn}\sum_{i,j} a_{ij}^2\right)^{1/2}.$$

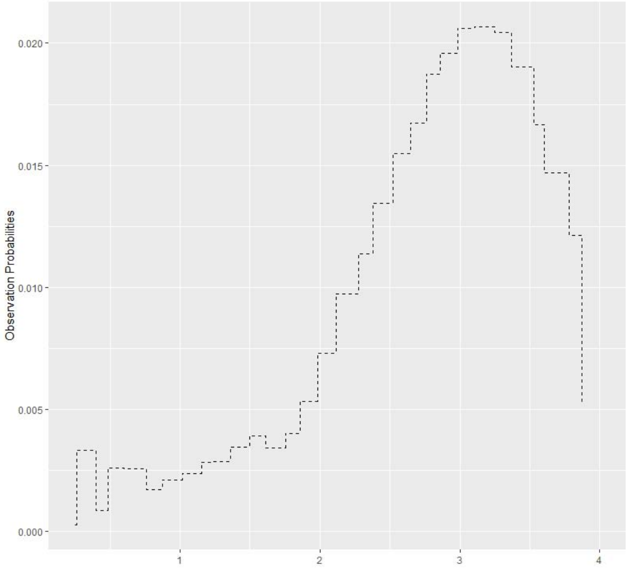

Fig. 9. Estimation of f in Example 2.15 using the modified USVT estimator.

Note that $\|A\|_2^2$ is the sum of squares of the singular values of A, divided by mn. Given the matrix A, we will also denote by A the function $A : [0,1]^2 \to [-1,1]$ which equals a_{ij} in the rectangle $(\frac{i-1}{m}, \frac{i}{m}) \times (\frac{j-1}{n}, \frac{j}{n})$ for each $1 \le i \le m$ and $1 \le j \le n$. On the boundaries of the rectangles, we define the function A is to be zero. Note that with this convention, $\|A\|_2$ equals the L^2 norm of the function A, which will also be denoted by $\|A\|_2$.

For each k, let S_k denote the group of all permutations of $\{1, \dots, k\}$. Given an $m \times n$ matrix A and a measurable map $W : [0,1]^2 \to [-1,1]$, we define

$$d_2(A, B) := \min_{\pi \in S_m,\, \tau \in S_n} \|A^{\pi,\tau} - W\|_2, \tag{2}$$

where $A^{\pi,\tau}$ is the matrix whose (i,j)-th entry is $a_{\pi(i)\tau(j)}$. The first key step in the proof of Theorem 2.4 is the following lemma.

Lemma 3.1: Suppose that for each k, we have a matrix M_k of order $m_k \times n_k$ with entries in $[-1,1]$, where $m_k, n_k \to \infty$ as $k \to \infty$. Suppose that this sequence has uniformly bounded rank. Then there exists a subsequence M_{k_l} and a measurable map $W : [0,1]^2 \to [-1,1]$ such that $d_2(M_{k_l}, W) \to 0$ as $l \to \infty$.

We will now prove Lemma 3.1. The proof closely follows the proof of [15, Theorem 1]. Let m and n be two positive integers. Let $\mathcal{P}$ be a partition of $\{1, \dots, m\}$ and let $\mathcal{Q}$ be a partition of $\{1, \dots, n\}$. The pair $(\mathcal{P}, \mathcal{Q})$ defines a block structure for $m \times n$ matrices in the natural way: Two pairs of indices (i,j) and (i',j') belong to the same block if and only if i and i' belong to the same member of $\mathcal{P}$ and j and j' belong to the same member of $\mathcal{Q}$.

If A is an $m \times n$ matrix, let $A^{\mathcal{P},\mathcal{Q}}$ be the 'block averaged' version of A, obtained by replacing the entries in each block (in the block structure defined by $(\mathcal{P}, \mathcal{Q})$) by the average value in that block. It is easy to see that

$$\|A^{\mathcal{P},\mathcal{Q}}\|_2 \le \|A\|_2. \tag{3}$$

We need the following lemma.

Lemma 3.2: For any $m \times n$ matrix A with entries in $[-1, 1]$, and $\operatorname{rank}(A) \le r$, there is a sequence of partitions $\{\mathcal{P}_j\}_{j\ge1}$ of $\{1, \dots, m\}$ and a sequence of partitions $\{\mathcal{Q}_j\}_{j\ge1}$ of $\{1, \dots, n\}$ such that for each j,

1) $\mathcal{P}_{j+1}$ is a refinement of $\mathcal{P}_j$ and $\mathcal{Q}_{j+1}$ is a refinement of $\mathcal{Q}_j$,
2) $|\mathcal{P}_j|$ and $|\mathcal{Q}_j|$ are bounded by $(2^{j+2}j)^{j^2}$, and
3) $\|A - A^{\mathcal{P}_j,\mathcal{Q}_j}\|_2 \le 2\sqrt{r}/j + 6j^3 2^{-j}$.

Proof: Let $A = \sum_{i=1}^r \sigma_i u_i v_i^T$ be the singular value decomposition of A, where $\sigma_1 \ge \cdots \ge \sigma_r$, and some of the σ_i's are zero if the rank is strictly less than r. Take any $j \ge 1$. Let l be the largest number such that $\sigma_l > \sqrt{mn}/j$. If there is no such l, let $l = 0$. Let

$$A_1 := \sum_{i=1}^{l} \sigma_i u_i v_i^T.$$

We define $\mathcal{P}_j$, $\mathcal{Q}_j$, and $\tilde{A}_1$ as in the proof of [15, Lemma 4], as follows. For $1 \le i \le l$ and $1 \le a \le m$, let u_{ia} denote the a^{th} component of u_i. Let $\tilde{u}_{ia}^{(j)}$ be the largest integer multiple of $2^{-j}m^{-1/2}$ that is $\le u_{ia}$. Let $\tilde{u}_i^{(j)}$ be the vector whose a^{th} component is $\tilde{u}_{ia}^{(j)}$. Similarly, for $1 \le b \le n$, let $\tilde{v}_{ib}^{(j)}$ be the largest integer multiple of $2^{-j}n^{-1/2}$ that is $\le v_{ib}$. Finally, define $\tilde{A} = \sum_{j=1}^{l} \sigma_i \tilde{u}_i \tilde{v}_i^\top$. This matrix $\tilde{A}$ is used as a block-approximation of A_1. As shown in [15], this sequence of partitions satisfy property (1) and (2) in the statement of the lemma. Now, using the properties of the $\|\cdot\|_2$ norms noted earlier, and the facts that $l \le r$ and $\sigma_{l+1} \le \sqrt{mn}/j$, we have

$$\|A - A_1\|_2 = \left(\frac{1}{mn}\sum_{i=l+1}^{r} \sigma_i^2\right)^{1/2} \le \left(\frac{r\sigma_{l+1}^2}{mn}\right)^{1/2} \le \frac{\sqrt{r}}{j}.$$

Again, as in the proof of [15, Lemma 4], we obtain

$$\|A_1 - \tilde{A}_1\|_2 \le 3j^3 2^{-j}.$$

Combining, we get

$$\|A - \tilde{A}_1\|_2 \le \sqrt{r}/j + 3j^3 2^{-j}.$$

Now note that $\tilde{A}_1$ is constant within the blocks defined by the pair $(\mathcal{P}_j, \mathcal{Q}_j)$. Thus, by (3),

$$\begin{aligned}\|A - A^{\mathcal{P}_j,\mathcal{Q}_j}\|_2 &\le \|A - \tilde{A}_1\|_2 + \|\tilde{A}_1 - A^{\mathcal{P}_j,\mathcal{Q}_j}\|_2 \\ &\le \|A - \tilde{A}_1\|_2 + \|\tilde{A}_1^{\mathcal{P}_j,\mathcal{Q}_j} - A^{\mathcal{P}_j,\mathcal{Q}_j}\|_2 \\ &\le 2\|A - \tilde{A}_1\|_2.\end{aligned}$$

This completes the proof. □

We are now ready to prove Lemma 3.1.

Proof of Lemma 3.1: Let r be a uniform upper bound on the rank of M_k. Lemma 3.2 tells us that for each k and j, we can find a partition $\mathcal{P}_{k,j}$ of $\{1, \dots, m_k\}$ and a partition $\mathcal{Q}_{k,j}$ of $\{1, \dots, n_k\}$ such that

1) $\mathcal{P}_{k,j+1}$ is a refinement of $\mathcal{P}_{k,j}$ and $\mathcal{Q}_{k,j+1}$ is a refinement of $\mathcal{Q}_{k,j}$,
2) $|\mathcal{P}_{k,j}|$ and $|\mathcal{Q}_{k,j}|$ are bounded by $(2^{j+2}j)^{j^2}$, and
3) $\|M_k - M_k^{\mathcal{P}_{k,j},\mathcal{Q}_{k,j}}\|_2 \le 2\sqrt{r}/j + 6j^3 2^{-j}$.

To reduce notation, let us denote $M_k^{\mathcal{P}_{k,j},\mathcal{Q}_{k,j}}$ by $M_{k,j}$. Following the proof of [15, Theorem 1] and passing to a subsequence if necessary, we get that for every j, there exists a measurable function $W_j : [0,1]^2 \to [-1,1]$ such that $M_{k,j}^{\pi_k,\tau_k} \to W_j$ in L^2 as $k \to \infty$, where π_k and τ_k are permutations that depend only on k (and not on j). Without loss of generality, let us assume π_k and τ_k are identity permutations for each k.

By construction, the block structure for W_{j+1} is a refinement of the block structure for W_j. Also by construction, the value of W_j in one of its blocks is the average value of W_{j+1} within that block. From this, by a standard martingale argument (for example, as in the proof of [32, Theorem 9.23]) it follows that W_j converges pointwise almost everywhere to a function W as $j \to \infty$. In particular, $W_j \to W$ in L^2. We claim that $M_k \to W$ in L^2 as $k \to \infty$. To show this, take any $\varepsilon > 0$. Find j so large that $\|W - W_j\|_2 \le \varepsilon$ and $2\sqrt{r}/j + 6j^3 2^{-j} \le \varepsilon$. Then for any k,

$$\begin{aligned}&\|W - M_k\|_2 \\ &\le \|W - W_j\|_2 + \|W_j - M_{k,j}\|_2 + \|M_{k,j} - M_k\|_2 \\ &\le \varepsilon + \|W_j - M_{k,j}\|_2 + 2\sqrt{r}/j + 6j^3 2^{-j} \\ &\le 2\varepsilon + \|W_j - M_{k,j}\|_2.\end{aligned}$$

Since $M_{k,j} \to W_j$ in L^2 as $k \to \infty$ and ε is arbitrary, this completes the proof. □

Henceforth, let us work in the setting of Theorem 2.4. For each k, let P_k be the random binary matrix whose (i,j)-the entry is 1 if the (i,j)-th entry of M_k is revealed, and 0 otherwise. Then note that as functions on $[0,1]^2$, $\mathbb{E}(P_k) = f \circ M_k$, where $\mathbb{E}(P_k)$ denotes the matrix of expected values of the entries of P_k.

Recall the *cut norm* on the set of $m \times n$ matrices, as defined in [15]:

$$\|A\|_\square := \frac{1}{mn}\max\{|x^T A y| : x \in \mathbb{R}^m,\, y \in \mathbb{R}^n,\\ \|x\|_\infty \le 1,\, \|y\|_\infty \le 1\},$$

where $\|x\|_\infty$ denotes the ℓ^∞ norm of a vector x. If A is an $m \times n$ matrix and $W : [0,1]^2 \to \mathbb{R}$ is a measurable function, we define $d_\square(A, W)$ to be $\|A - B\|_\square$, where B is the $m \times n$ matrix whose (i,j)-th entry is the average value of W in the rectangle $(\frac{i-1}{m}, \frac{i}{m}) \times (\frac{j-1}{n}, \frac{j}{n})$.

The following lemma shows that P_k and $\mathbb{E}(P_k)$ are close in cut norm.

Lemma 3.3: As $k \to \infty$, $\|P_k - \mathbb{E}(P_k)\|_\square \to 0$ in probability.

Proof: It is easy to see from the definition of cut norm that for an $m \times n$ matrix A,

$$\|A\|_\square \le \frac{\|A\|_{op}}{\sqrt{mn}},$$

where $\|A\|_{op}$ is the ℓ^2 operator norm of A. Now take any $t > 0$. Using [14, Theorem 3.4], $\mathbb{P}(\|P_k - \mathbb{E}(P_k)\| \ge 3\sqrt{n_k}) \le C_1 e^{-C_2 n_k}$ for some positive universal constants C_1 and C_2. Hence, for k large enough,

$$\begin{aligned}\mathbb{P}(\|P_k - \mathbb{E}(P_k)\|_\square \ge t) &\le \mathbb{P}(\|P_k - \mathbb{E}(P_k)\|_{op} \ge t\sqrt{m_k n_k}) \\ &\le \mathbb{P}(\|P_k - \mathbb{E}(P_k)\|_{op} \ge 3\sqrt{n_k}) \\ &\le C_1 e^{-C_2 n_k}.\end{aligned}$$

This shows that $\|P_k - \mathbb{E}(P_k)\|_\square \to 0$ in probability as $k \to \infty$. □

Next, we relate the limiting empirical distribution of the entries of M_k with the L^2 limit of M_k as a function on $[0,1]^2$. In the following, λ denotes Lebesgue measure on $[0,1]^2$.

Lemma 3.4: Suppose that $M_k \to W$ in L^2 as a sequence of functions on $[0,1]^2$. Then μ_k converges weakly to $\mu = \lambda \circ W^{-1}$.

Proof: Take any bounded continuous function $g : [-1,1] \to \mathbb{R}$. It is not difficult to see that

$$\int g d\mu_k = \iint g(M_k(x,y))dxdy.$$

Since $M_k \to W$ in L^2 and g is bounded and continuous, we get

$$\lim_{k\to\infty} \iint g(M_k(x,y))dxdy = \iint g(W(x,y))dxdy.$$

But the right side is the integral of g with respect to the measure $\lambda \circ W^{-1}$. This completes the proof. □

The purpose of the next lemma is to investigate the convergence of $f \circ M_k$ under the hypotheses of Theorem 2.4.

Lemma 3.5: Suppose that $M_k \to W$ in L^2 as a sequence of functions on $[0,1]^2$. Let $\mu := \lambda \circ W^{-1}$. Suppose that $g : [-1,1] \to [0,1]$ is a measurable function which is continuous almost everywhere with respect to μ. Then $g \circ M_k \to g \circ W$ in L^2.

Proof: Since $M_k \to W$ in L^2, any subsequence has a further subsequence along which $M_k(x,y) \to W(x,y)$ for λ-a.e. (x,y). By assumption, g is continuous at $W(x,y)$ for λ-a.e. (x,y). Combining these two observations, we get that for any subsequence, there is a further subsequence along with $g \circ M_k(x,y) \to g \circ W(x,y)$ for λ-a.e. (x,y). Since g, M_k and W are all taking values in $[0,1]$, this implies that $g \circ M_k \to g \circ W$ in L^2 along this subsequence. This completes the proof. □

As a consequence of the above lemmas, we obtain the following result.

Lemma 3.6: Suppose that $M_k \to W$ in L^2 as a sequence of functions on $[0,1]^2$. Then, under the hypotheses of Theorem 2.4, there is a measurable function $V : [0,1]^2 \to [0,1]$ that is nonzero almost everywhere and $d_\square(P_k, V) \to 0$ in probability as $k \to \infty$.

Proof: By Lemma 3.3, it suffices to show that $d_\square(\mathbb{E}(P_k), V) \to 0$ for some V as in the statement of the lemma. By Lemma 3.4, μ_k converges weakly to $\mu = \lambda \circ W^{-1}$. By the hypotheses of Theorem 2.4, there is a measurable extension of f to $[-1,1]$, also denoted by f, which is nonzero and continuous μ-a.e. As noted earlier, $\mathbb{E}(P_k) = f \circ M_k$. Therefore, by Lemma 3.5, $\mathbb{E}(P_k) \to f \circ W$ in L^2. It is not hard to see that this implies that $d_\square(\mathbb{E}(P_k), f \circ W) \to 0$. But $f \circ W$ is nonzero λ-a.e. Thus, we can take $V = f \circ W$. □

We are now ready to prove Theorem 2.4.

Proof: Suppose that $\hat{M}_k$ is not a consistent sequence of estimators. Then, passing to a subsequence if necessary, we may assume that

$$\inf_{k\geq 1} \mathbb{E}\|\hat{M}_k - M_k\|_2^2 > 0. \tag{4}$$

Note that this condition continues to hold true if we pass to further subsequences and permute rows and columns in each M_k, which we will do shortly. Passing to a further subsequence, and permuting rows and columns in each M_k if necessary, we use Lemma 3.1 to get an L^2 limit W of M_k as $k \to \infty$. Then, by Lemma 3.6, there is a measurable function $V : [0,1]^2 \to [0,1]$ that is nonzero almost everywhere and $d_\square(P_k, V) \to 0$ in probability as $k \to \infty$. Again passing to a subsequence, we get that $d_\square(P_k, V) \to 0$ almost surely. But this implies, by [15, Theorem 2 and Theorem 3], that $\|\hat{M}_k - M_k\|_2 \to 0$ almost surely. Since the entries of M_k and $\hat{M}_k$ are in $[-1,1]$ for all k, this contradicts (4). □

IV. Proof of Theorem 2.1

Without loss of generality, suppose that $m_k \leq n_k$ for each k. (Otherwise, we can just transpose the matrices.) Let r be a uniform upper bound on the rank of M_k. Let R_k be the matrix obtained by applying f entrywise to M_k. Let Q_k be the entrywise (i.e., Hadamard) product of M_k and R_k. Let Y_k be the matrix obtained by replacing the unrevealed entries of X_k by zero. Let P_k be the matrix whose (i,j)-th entry is 1 if the (i,j)-th entry of X_k is revealed, and 0 otherwise. Note that $\mathbb{E}(Y_k) = Q_k$ and $\mathbb{E}(P_k) = R_k$. Note also that the entries of Y_k and P_k are all in $[-1,1]$.

First, let us assume that μ_k converges weakly to a limit μ as $k \to \infty$. Then by the hypotheses of Theorem 2.1, f has an extension to a Lipschitz function on $[-1,1]$, also called f, which has no zeros in the support of μ. Let us fix such an extension, and let L denote its Lipschitz constant.

Lemma 4.1: As $k \to \infty$, $\|R_k\|_* = o(m_k\sqrt{n_k})$.

Proof: Fix $\varepsilon > 0$. It is an easy consequence of the Cauchy–Schwarz inequality that for any k,

$$\|M_k\|_* \leq \|M_k\|_2\sqrt{\operatorname{rank}(M_k)m_kn_k} \leq \sqrt{rm_kn_k}.$$

By [15, Lemma 2], this implies that there is a block matrix B_k with at most b blocks, where b depends only on ε and r, and entries in $[-1,1]$, such that $\|M_k - B_k\|_2 \leq \varepsilon$. Let D_k be obtained by applying f to B_k entrywise. Then by the Lipschitz property of f, we get

$$\|R_k - D_k\|_2 \leq \varepsilon L.$$

Note that just like B_k, D_k has at most b blocks. In particular, $\operatorname{rank}(D_k) \leq b$. Therefore again by the Cauchy–Schwarz inequality,

$$\begin{aligned}\|R_k\|_* &\leq \|R_k - D_k\|_* + \|D_k\|_* \\ &\leq \|R_k - D_k\|_2\sqrt{\operatorname{rank}(R_k - D_k)m_kn_k} \\ &\quad + \|D_k\|_2\sqrt{\operatorname{rank}(D_k)m_kn_k} \\ &\leq \varepsilon L m_k\sqrt{n_k} + \sqrt{bm_kn_k}.\end{aligned}$$

Thus,

$$\limsup_{k\to\infty} \frac{\|R_k\|_*}{m_k\sqrt{n_k}} \leq \varepsilon L.$$

Since this holds for arbitrary $\varepsilon > 0$, this completes the proof. □

Lemma 4.2: As $k \to \infty$, $\|Q_k\|_* = o(m_k\sqrt{n_k})$.

Proof: Let B_k, b, and D_k be as in Lemma 4.2. Let E_k be the Hadamard product of B_k and D_k, and F_k be the Hadamard product of B_k and R_k. Then E_k also has b blocks. Moreover, since the entries of all these matrices are in $[-1,1]$, it is not hard to see that

$$\begin{aligned}\|Q_k - E_k\|_2 &\le \|Q_k - F_k\|_2 + \|F_k - E_k\|_2\\ &\le \|M_k - B_k\|_2 + \|R_k - D_k\|_2\\ &\le (L+1)\varepsilon.\end{aligned}$$

The rest of the proof is the same as the proof of Lemma 4.1, with R_k replaced by Q_k and D_k replaced by E_k. □

As a consequence of the above lemmas, we obtain the following result.

Lemma 4.3: Let $\hat{Q}_k$ and $\hat{R}_k$ be the estimates of Q_k and R_k obtained by applying the USVT algorithm to Y_k and P_k. Then $\mathbb{E}\|\hat{Q}_k - Q_k\|_2^2 \to 0$ and $\mathbb{E}\|\hat{R}_k - R_k\|_2^2 \to 0$ as $k \to \infty$.

Proof: This is a direct consequence of Lemmas 4.1 and 4.2 and the consistency of the USVT estimator [14, Theorem 1.1]. □

Let us now prove Theorem 2.1 under the simplifying assumption under which we are currently working. Let $\hat{M}_k$ denote the modified USVT estimator. Let m_{kij} denote the (i,j)-th element of M_k, $\hat{m}_{kij}$ denote the (i,j)-th element of $\hat{M}_k$, etc.

Since f is nonzero and continuous on the support of μ, and the support is a compact set, there exists $\delta > 0$ such that $f > \delta$ everywhere on the support of μ. In particular, $\mu(\{x : f(x) \le \delta\}) = 0$. Since $\mu_k \to \mu$ weakly, and $\{x : f(x) \le \delta\}$ is a closed set due to the continuity of f, we get

$$\limsup_{k\to\infty} \mu_k(\{x : f(x) \le \delta\}) \le \mu(\{x : f(x) \le \delta\}) = 0.$$

In other words, if we let $I_k := \{(i,j) : r_{kij} \le \delta\}$, then $|I_k| = o(m_k n_k)$ as $k \to \infty$.

Let $J_k := \{(i,j) : \hat{r}_{kij} \le \delta/2\}$. Then

$$\begin{aligned}|J_k| &\le |I_k| + |\{(i,j) : |\hat{r}_{kij} - r_{kij}| > \delta/2\}|\\ &\le |I_k| + \frac{4}{\delta^2}\sum_{i,j}(\hat{r}_{kij} - r_{kij})^2\\ &= |I_k| + \frac{4 m_k n_k}{\delta^2}\|\hat{R} - R\|_2^2.\end{aligned}$$

By Lemma 4.3 and the fact that $|I_k| = o(m_k n_k)$, this shows that $|J_k| = o_P(m_k n_k)$ as $k \to \infty$ (meaning that $|J_k|/(m_k n_k) \to 0$ in probability as $k \to \infty$).

Now take $(i,j) \notin I_k \cup J_k$. Then

$$\begin{aligned}\left|\frac{\hat{q}_{kij}}{\hat{r}_{kij}} - m_{kij}\right| &= \left|\frac{\hat{q}_{kij}}{\hat{r}_{kij}} - \frac{q_{kij}}{r_{kij}}\right|\\ &\le \frac{|\hat{q}_{kij} - q_{kij}|}{\hat{r}_{kij}} + \frac{|q_{kij}||\hat{r}_{kij} - r_{kij}|}{\hat{r}_{kij} r_{kij}}\\ &\le \frac{2}{\delta}|\hat{q}_{kij} - q_{kij}| + \frac{2}{\delta^2}|\hat{r}_{kij} - r_{kij}|.\end{aligned}$$

Since $\hat{m}_{kij}$ is obtained by truncating $\hat{q}_{kij}/\hat{r}_{kij}$, the above upper bound also holds for $|\hat{m}_{kij} - m_{kij}|$ when $(i,j) \notin I_k \cup J_k$. But $|\hat{m}_{kij} - m_{kij}| \le 2$ for any (i,j). Thus,

$$\begin{aligned}&\sum_{i,j}(\hat{m}_{kij} - m_{kij})^2\\ &\le 4|I_k \cup J_k| + \sum_{(i,j)}\left(\frac{2}{\delta}|\hat{q}_{kij} - q_{kij}| + \frac{2}{\delta^2}|\hat{r}_{kij} - r_{kij}|\right)^2\\ &\le 4|I_k \cup J_k| + \frac{8}{\delta^2}\sum_{i,j}(\hat{q}_{kij} - q_{kij})^2\\ &\quad + \frac{8}{\delta^4}\sum_{i,j}(\hat{r}_{kij} - r_{kij})^2.\end{aligned}$$

By Lemma 4.3 and our previous deduction that $|I_k \cup J_k| = o_P(m_k n_k)$, the above inequality shows that $\|\hat{M}_k - M_k\|_2 \to 0$ in probability as $k \to \infty$. Since this is a uniformly bounded sequence of random variables, this proves the consistency of $\hat{M}_k$. This proves Theorem 2.1 under the simplifying assumption that μ_k converges weakly to some μ as $k \to \infty$. We are now ready to prove Theorem 2.1 in full generality.

Proof of Theorem 2.1 Let $\hat{M}_k$ be the modified USVT estimator of M_k. Suppose that $\{\hat{M}_k\}_{k\ge 1}$ is not a consistent sequence of estimators. Passing to a subsequence if necessary, we may assume that

$$\inf_{k\ge 1}\mathbb{E}\|\hat{M}_k - M_k\|_2^2 > 0. \tag{5}$$

Note that this will continue to hold true if we pass to further subsequences. Passing to a further subsequence, we may assume that μ_k converges weakly to some μ. But then we already know that (5) is violated. This completes the proof of the theorem. □

V. Proof of Theorem 2.7

In this proof, C will denote any universal constant, whose value may change from line to line. Let $[x,y)$ be a subinterval of $[-2,2]$. Let $p_{ij} = 1$ if the (i,j)-th entry of M is revealed and 0 otherwise. Let

$$\begin{aligned}S_{x,y} &:= \{(i,j) : m_{ij} \in [x,y)\},\\ T_{x,y} &:= \{(i,j) : \hat{m}_{ij} \in [x,y)\},\end{aligned}$$

and let

$$\hat{f}_{x,y} := \frac{1}{|T_{x,y}|}\sum_{(i,j)\in T_{x,y}} p_{ij},\ \ g_{x,y} := \frac{1}{|S_{x,y}|}\sum_{(i,j)\in S_{x,y}} p_{ij}$$

where the right sides are declared to be zero if the corresponding sums are empty. Note that $\hat{f}_{x,y}$ and $g_{x,y}$ are always in $[0,1]$. Take some $\delta < (y-x)/2$, to be chosen later. Let

$$\mu_{x,y} := \frac{1}{mn}|\{(i,j) : m_{ij} \in [a-\delta, a+\delta] \cup [b-\delta, b+\delta]\}|. \tag{6}$$

Take any $(i,j) \in T_{x,y} \setminus S_{x,y}$. There are two cases. First suppose that $m_{ij} \notin [x-\delta, y+\delta]$. Since $(i,j) \in T_{x,y}$, we have $\hat{m}_{ij} \in [x,y)$, and hence in this case, $|\hat{m}_{ij} - m_{ij}| > \delta$. By Markov's inequality, the number of such (i,j) is bounded above by

$$\frac{1}{\delta^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(m_{ij} - \hat{m}_{ij})^2. \tag{7}$$

The second case is that $m_{ij} \in [x-\delta, x) \cup [y, y+\delta]$. By the definition of $\mu_{x,y}$, the number of such (i,j) is at most $mn\mu_{x,y}$. Combining, we get that

$$|T_{x,y} \setminus S_{x,y}| \le \frac{1}{\delta^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(m_{ij} - \hat{m}_{ij})^2 + mn\mu_{x,y}.$$

Now take any $(i,j) \in S_{x,y} \setminus T_{x,y}$. Then, again, there are two cases. First, suppose that $m_{ij} \in [x+\delta, y-\delta]$. Since $\hat{m}_{ij} \notin [x,y)$, in this case we have that $|\hat{m}_{ij} - m_{ij}| > \delta$. Thus, by Markov's inequality, the number of such (i,j) is bounded above by (7). The other case is $m_{ij} \in [x, x+\delta) \cup (y-\delta, y)$. As before, the number of such (i,j) is bounded above by $mn\mu_{x,y}$. Combining these two observations, we get that

$$\mathbb{E}|T_{x,y}\Delta S_{x,y}| \le \frac{2mn\theta}{\delta^2} + 2mn\mu_{x,y}. \tag{8}$$

We will now work under the assumption that $S_{x,y} \neq \emptyset$. The final estimate will be valid even if $S_{x,y} = \emptyset$. First, note that

$$\mathrm{Var}(g_{x,y}) = \frac{1}{|S_{x,y}|^2}\sum_{(i,j)\in S_{x,y}} \mathrm{Var}(Y_{ij}) \le \frac{1}{4|S_{x,y}|}.$$

Let $f_{x,y} := \mathbb{E}(g_{x,y})$. Then the above bound can be written as

$$|S_{x,y}|\mathbb{E}[(g_{x,y} - f_{x,y})^2] \le \frac{1}{4}. \tag{9}$$

Clearly, the above bound holds even if $S_{x,y} = \emptyset$. Next, note that

$$\begin{aligned}|g_{x,y} - \hat{f}_{x,y}| &\le \frac{1}{|S_{x,y}|}\left|\sum_{(i,j)\in S_{x,y}} Y_{ij} - \sum_{(i,j)\in T_{x,y}} Y_{ij}\right| \\ &\quad + \left|\frac{1}{|S_{x,y}|} - \frac{1}{|T_{x,y}|}\right| \sum_{(i,j)\in T_{x,y}} Y_{ij} \\ &\le \frac{1}{|S_{x,y}|}\sum_{(i,j)\in T_{x,y}\Delta S_{x,y}} Y_{ij} + \frac{||T_{x,y}| - |S_{x,y}||}{|S_{x,y}|} \\ &\le \frac{2|T_{x,y}\Delta S_{x,y}|}{|S_{x,y}|}.\end{aligned}$$

This shows, by (8) and the fact that $\hat{f}_{x,y}$ and $g_{x,y}$ are both in $[0,1]$, that

$$\begin{aligned}|S_{x,y}|\mathbb{E}[(\hat{f}_{x,y} - g_{x,y})^2] &\le |S_{x,y}|\mathbb{E}|\hat{f}_{x,y} - g_{x,y}| \\ &\le 2\mathbb{E}|T_{x,y}\Delta S_{x,y}| \\ &\le \frac{4mn\theta}{\delta^2} + 4mn\mu_{x,y}. \end{aligned} \tag{10}$$

Again, this bound holds even if $S_{x,y} = \emptyset$. Combining (9) and (10), we get

$$|S_{x,y}|\mathbb{E}[(\hat{f}_{x,y} - f_{x,y})^2] \le \frac{8mn\theta}{\delta^2} + 8mn\mu_{x,y} + \frac{1}{4}. \tag{11}$$

Using the notation (6), we see that for any $1 \le l \le b+2$,

$$\begin{aligned}&\mathbb{E}(\mu_{a_l,a_{l+1}}) \\ &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[\mathbb{P}(|m_{ij} - a_l| \le \delta) + \mathbb{P}(|m_{ij} - a_{l+1}| \le \delta)] \\ &\le 8b\delta.\end{aligned}$$

Applying (11) to the interval $[a_l, a_{l+1})$, taking expectation over the randomness of the a_l's and applying the above inequality, and then summing over l, we get

$$\begin{aligned}&\frac{1}{mn}\sum_{l=1}^{b+2}|S_{a_l,a_{l+1}}|\mathbb{E}[(\hat{f}_{a_l,a_{l+1}} - f_{a_l,a_{l+1}})^2] \\ &\quad\le \frac{C\theta b}{\delta^2} + Cb^2\delta + \frac{Cb}{mn}.\end{aligned}$$

Choosing $\delta = (\theta/b)^{1/3}$ gives

$$\begin{aligned}&\frac{1}{mn}\sum_{l=1}^{b+2}|S_{a_l,a_{l+1}}|\mathbb{E}[(\hat{f}_{a_l,a_{l+1}} - f_{a_l,a_{l+1}})^2] \\ &\quad\le C\theta^{1/3}b^{5/3} + \frac{Cb}{mn}.\end{aligned}$$

For $x \in [a_l, a_{l+1})$, let

$$\tilde{f}(x) := \frac{1}{|S_{a_l,a_{l+1}}|}\sum_{(i,j)\in S_{a_l,a_l+1}} f(m_{ij})$$

Then note that for any $(i,j) \in S_{a_l,a_{l+1}}$,

$$|\tilde{f}(m_{ij}) - f(m_{ij})| \le \frac{CL}{b}.$$

Since

$$\begin{aligned}&\frac{1}{mn}\sum_{l=1}^{b+2}\mathbb{E}[|S_{a_l,a_{l+1}}|(\hat{f}_{a_l,a_{l+1}} - f_{a_l,a_{l+1}})^2] \\ &\quad= \frac{1}{mn}\sum_{i,j}\mathbb{E}[(\hat{f}(m_{ij}) - \tilde{f}(m_{ij}))^2],\end{aligned}$$

this completes the proof of the theorem.

Acknowledgment

Sohom Bhattacharya would like to thank Debangan Dey, Samriddha Lahiry, Samyak Rajanala, and Subhabrata Sen for helpful discussions. The authors thank the two anonymous referees for their useful suggestions.

References

[1] D. Achlioptas and F. Mcsherry, "Fast computation of low-rank matrix approximations," *J. ACM*, vol. 54, no. 2, p. 9, Apr. 2007.

[2] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, Dec. 2008.

[3] M. Azadkia, "Adaptive estimation of noise variance and matrix estimation via USVT algorithm," 2018, *arXiv:1801.10015*.

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

[5] S. Bhojanapalli and P. Jain, "Universal matrix completion," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2014, pp. 1881–1889.

[6] D. Billsus and M. J. Pazzani, "Learning collaborative information filters," in *Proc. 15th Int. Conf. Mach. Learn.*, vol. 98, 1998, pp. 46–54.

[7] C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen, "Subspace estimation from unbalanced and incomplete data matrices: $\mathrm{L}_{2,\infty}$ statistical guarantees," *Ann. Statist.*, vol. 49, no. 2, pp. 944–967, 2021.

[8] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[9] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[10] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[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, Feb. 2006.
[12] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.
[13] A. Carpentier, O. Klopp, and M. Löffler, "Constructing confidence sets for the matrix completion problem," in *Proc. Conf. Int. Soc. Non-Parametric Statist.* Cham, Switzerland: Springer, 2016, pp. 103–118.
[14] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *Ann. Statist.*, vol. 43, no. 1, pp. 177–214, 2015.
[15] S. Chatterjee, "A deterministic theory of low rank matrix completion," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 8046–8055, Dec. 2020.
[16] J. Cho, D. Kim, and K. Rohe, "Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise," *Statistica Sinica*, vol. 27, no. 4, pp. 1921–1948, 2017.
[17] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Inference*, vol. 3, no. 3, pp. 189–223, 2014.
[18] D. Donoho and M. Gavish, "Minimax risk of matrix denoising by singular value thresholding," *Ann. Statist.*, vol. 42, no. 6, pp. 2413–2440, 2014.
[19] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[20] R. Foygel, R. Salakhutdinov, O. Shamir, and N. Srebro, "Learning with the weighted trace-norm under arbitrary sampling distributions," 2011, *arXiv:1106.4251*.
[21] A. Fu, B. Narasimhan, and S. Boyd, "CVXR: An R package for disciplined convex optimization," *J. Stat. Softw.*, vol. 94, no. 14, pp. 1–34, 2020, doi: 10.18637/jss.v094.i14.
[22] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: A constant time collaborative filtering algorithm," *Inf. Retr.*, vol. 4, no. 2, pp. 133–151, Jul. 2001.
[23] G. Guo, J. Zhang, and N. Yorke-Smith, "A novel Bayesian similarity measure for recommender systems," in *Proc. IJCAI*, vol. 13, 2013, pp. 2619–2625.
[24] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, Jul. 2010.
[25] F. J. Király, L. Theran, and R. Tomioka, "The algebraic combinatorial approach for low-rank matrix completion," *Mach. Learn.*, vol. 16, pp. 1391–1436, Jan. 2015.
[26] O. Klopp, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.
[27] O. Klopp, "Matrix completion by singular value thresholding: Sharp bounds," *Electron. J. Statist.*, vol. 9, no. 2, pp. 2348–2369, Jan. 2015.
[28] V. Koltchinskii *et al.*, "Von Neumann entropy penalization and low-rank matrix estimation," *Ann. Statist.*, vol. 39, no. 6, pp. 2936–2973, 2011.
[29] V. Koltchinskii, K. Lounici, and A. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Statist.*, vol. 39, no. 5, pp. 2302–2329, 2011.
[30] T. Lee and A. Shraibman, "Matrix completion from any given set of observations," in *Proc. NIPS*, 2013, pp. 1781–1787.
[31] N. Linial, E. London, and Y. Rabinovich, "The of graphsgeometry and some of its algorithmic applications," *Combinatorica*, vol. 15, no. 2, pp. 215–245, 1995.
[32] L. Lovász, "*Large Networks and Graph Limits*. Providence, RI, USA: American Mathematical, 2012.
[33] W. Ma and G. H. Chen, "Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption," 2019, *arXiv:1910.12774*.
[34] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, Jan. 2010.
[35] W. Miao, P. Ding, and Z. Geng, "Identifiability of normal and normal mixture models with nonignorable missing data," *J. Amer. Stat. Assoc.*, vol. 111, no. 516, pp. 1673–1683, Oct. 2016.
[36] K. Mohan, F. Thoemmes, and J. Pearl, "Estimation with incomplete data: The linear case," in *Proc. Int. Conf. Artif. Intell. Org.*, 2018, pp. 5082–5088.
[37] A. Montanari, F. Ruan, and J. Yan, "Adapting to unknown noise distribution in matrix denoising," 2018, *arXiv:1810.02954*.
[38] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *Ann. Statist.*, vol. 39, no. 2, pp. 1069–1097, Apr. 2011.
[39] L. T. Nguyen, J. Kim, and B. Shim, "Low-rank matrix completion: A contemporary survey," *IEEE Access*, vol. 7, pp. 94215–94237, 2019.
[40] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statist. Comput.*, vol. 20, no. 2, pp. 231–252, Apr. 2010.
[41] D. Pimentel-Alarcón, N. Boston, and R. D. Nowak, "A characterization of deterministic sampling patterns for low-rank matrix completion," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 623–636, Jun. 2016.
[42] J. D. M. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 713–719.
[43] A. Rohde and A. Tsybakov, "Estimation of high-dimensional low-rank matrices," *Ann. Statist.*, vol. 39, no. 2, pp. 887–930, 2011.
[44] A. Singer and M. Cucuringu, "Uniqueness of low-rank matrix completion by rigidity theory," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 4, pp. 1621–1641, Jan. 2010.
[45] A. Sportisse, C. Boyer, and J. Josse, "Imputation and low-rank estimation with missing not at random data," *Statist. Comput.*, vol. 30, no. 6, pp. 1629–1643, 2020.
[46] K. Q. Weinberger and L. K. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 77–90, 2006.
[47] C. Yang, L. Ding, Z. Wu, and M. Udell, "TenIPS: Inverse propensity sampling for tensor completion," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3160–3168.
[48] Kisung You. (2020) *Filling: Matrix Completion, Imputation, and Inpainting Methods*. [Online]. Available: https://CRAN.R-project.org/package=filling
[49] Z. Zhu, T. Wang, and R. J. Samworth, "High-dimensional principal component analysis with heterogeneous missingness," 2019, *arXiv:1906.12125*.

Sohom Bhattacharya received the Bachelor of Statistics and Master of Statistics degrees from the Indian Statistical Institute, Kolkata, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Department of Statistics, Stanford University.

He will be joining the Department of Operations Research and Financial Engineering, Princeton University, as a Post-Doctoral Scholar, in Fall 2022. His areas of interests are mathematical statistics, machine learning, and probability theory.

Sourav Chatterjee received the Bachelor of Statistics and Master of Statistics degrees from the Indian Statistical Institute, Kolkata, in 2000 and 2002, respectively, and the Ph.D. degree in statistics from Stanford University in 2005.

He joined UC Berkeley as a Neyman Assistant Professor of statistics in 2005 and started as a Tenure-Track Assistant Professor with the Statistics Department in 2006. In 2009, he worked as an Associate Professor of mathematics with the Courant Institute of Mathematical Sciences, New York University. Since 2013, he has been a Professor of mathematics and statistics with Stanford University, Stanford, CA, USA. His areas of interests are probability theory, statistics, and mathematical physics.

Dr. Chatterjee was awarded a Sloan Research Fellowship in mathematics in 2007, the 2008 Tweedie New Researcher Award from the Institute of Mathematical Statistics, the 2010 Rollo Davidson Prize for work in probability theory, the 2012 Doeblin Prize from the Bernoulli Society, the 2012 Young Researcher Award from the International Indian Statistical Association, and the 2013 Line and Michel Loève Prize in probability from UC Berkeley. He gave a Medallion Lecture at the Institute of Mathematical Statistics in 2012 and was an Invited Speaker at the International Congress of Mathematicians in 2014. He was elected as a fellow of the Institute of Mathematical Statistics in 2018 and received the Infosys Prize in mathematical sciences in 2020.