NoisyCUR: An algorithm for two-cost budgeted matrix completion

Dong Hu[1], Alex Gittens[1], and Malik Magdon-Ismail[1]

Rensselaer Polytechnic Institute, Troy, NY 12180, USA
{hud3,gittea}@rpi.edu, magdon@cs.rpi.edu

Abstract. Matrix completion is a ubiquitous tool in machine learning and data analysis. Most work in this area has focused on the number of observations necessary to obtain an accurate low-rank approximation. In practice, however, the cost of observations is an important limiting factor, and experimentalists may have on hand multiple modes of observation with differing noise-vs-cost trade-offs. This paper considers matrix completion subject to such constraints: a budget is imposed and the experimentalist's goal is to allocate this budget between two sampling modalities in order to recover an accurate low-rank approximation. Specifically, we consider that it is possible to obtain low noise, high cost observations of individual entries or high noise, low cost observations of entire columns. We introduce a regression-based completion algorithm for this setting and experimentally verify the performance of our approach on both synthetic and real data sets. When the budget is low, our algorithm outperforms standard completion algorithms. When the budget is high, our algorithm has comparable error to standard nuclear norm completion algorithms and requires much less computational effort.

Keywords: Matrix Completion · Low-rank Approximation · Nuclear Norm Minimization.

1 Introduction

Matrix completion (MC) is a powerful and widely used tool in machine learning, finding applications in information retrieval, collaborative filtering, recommendation systems, and computer vision. The goal is to recover a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ from only a few, potentially noisy, observations $\mathbf{y} \in \mathbb{R}^d$, where $d \ll mn$.

In general, the MC problem is ill-posed, as many matrices may give rise to the same set of observations. Typically the inversion problem is made feasible by assuming that the matrix from which the observations were generated is in fact low-rank, $\text{rank}(\boldsymbol{A}) = r \ll \min\{m, n\}$. In this case, the number of degrees of freedom in the matrix is $(n+m)r$, so if the observations are sufficiently diverse, then the inversion process is well-posed.

In the majority of the MC literature, the mapping from the matrix to its observations, although random, is given to the user, and the aim is to design algorithms that minimize the sample complexity under these observation models.

Some works have considered modifications of this paradigm, where the user designs the observation mapping themselves in order to minimize the number of measurements needed [6,21,14].

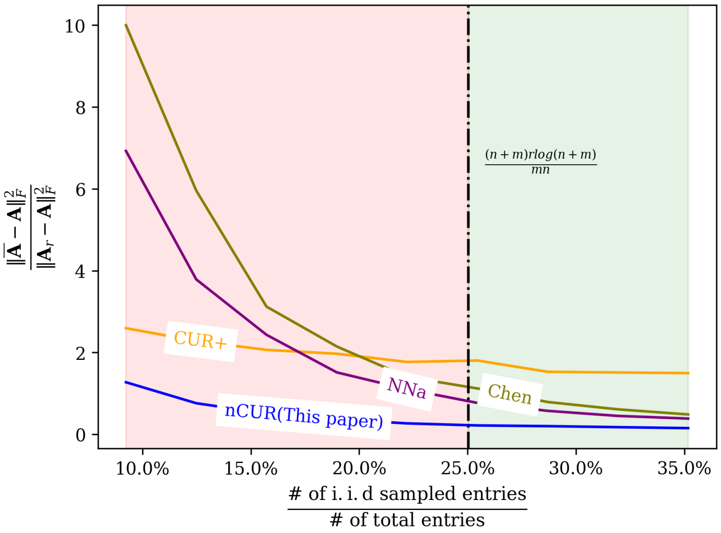

Fig. 1: noisyCUR, a regression-based MC algorithm (CUR+), and two nuclear norm MC algorithms applied to the budgeted MC problem on a synthetic incoherent matrix $\boldsymbol{A} \in \mathbb{R}^{80\times 60}$. See Section 4 for more details on the experimental setup. The performance of each method at each budget level is averaged over 10 runs, with hyper-parameters selected at each value of d by cross-validation. Within the red region the budget is small enough that not enough entries can be sampled for nuclear norm methods to have theoretical support. In this regime, noisyCUR significantly outperforms the baseline algorithms.

This paper considers a budgeted variation of noisy matrix completion with competing observation models: the goal is to allocate a finite budget between the observation models to maximize the accuracy of the recovery. This setup is natural in the context of experimental design: where, given competing options, the natural goal of an experimenter is to spend their budget in a way that maximizes the accuracy of their imputations. Specifically, this work considers a two-cost model where either entire columns can be observed with high per-entry error but low amortized per-entry cost, or entries can be observed with low per-entry error but high per-entry cost. This is a natural model in, for instance, recommender systems applications, where one has a finite budget to allocate between incentivizing users to rate either entire categories of products or individual items. The former corresponds to inexpensive, high-noise measurements of entire columns and the latter corresponds to expensive, low-noise measurements of individual entries.

The noisyCUR algorithm is introduced for this two-cost budgeted MC problem, and guarantees are given on its recovery accuracy. Figure 1 is illustrative of the usefulness of noisyCUR in this setting, compared to standard nuclear norm MC approaches and another regression-based MC algorithm. Even in low-budget settings where nuclear norm matrix completion is not theoretically justifiable, noisyCUR satisfies relative-error approximation guarantees. Empirical comparisons of the performance of noisyCUR to that of nuclear norm MC approaches on a synthetic dataset and on the Jester and MovieLens data sets confirm that noisyCUR can have significantly lower recovery error than standard nuclear norm approaches in the budgeted setting. Additionally, noisyCUR tolerates the presence of coherence in the row-space of the matrix, and is fast as its core computational primitive is ridge regression.

The rest of the paper is organized as follows. In Section 2 we introduce the two-cost model and discuss related works. Section 3 introduces the noisyCUR algorithm and provides performance guarantees; most proofs are deferred to the supplement. Section 4 provides an empirical comparison of the performance of noisyCUR and baseline nuclear norm completion algorithms that demonstrates the superior performance of noisyCUR in the limited budget setting. Section 5 concludes the work.

2 Problem Statement

2.1 Notation.

Throughout this paper, scalars are denoted by lowercase letters (n), vectors by bolded lowercase letters ($\mathbf{x}$), and matrices by bolded uppercase letters $\boldsymbol{A}$. The spectral, Frobenius, and nuclear norms of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ are written $\|\boldsymbol{A}\|_2$, $\|\boldsymbol{A}\|_F$, and $\|\boldsymbol{A}\|_\star$ respectively, and its singular values are ordered in a decreasing fashion: $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_{\min\{m,n\}}(\boldsymbol{A})$. The smallest nonzero singular value of $\boldsymbol{A}$ is denoted by $\sigma_{\min}(\boldsymbol{A})$. The condition number of $\boldsymbol{A}$ is taken to be the ratio of the largest singular value and the smallest *nonzero* singular value, $\kappa_2(\boldsymbol{A}) = \sigma_1(\boldsymbol{A})/\sigma_{\min}(\boldsymbol{A})$. The orthogonal projection onto the column span of $\boldsymbol{A}$ is denoted by $\boldsymbol{P}_{\boldsymbol{A}}$. Semidefinite inequalities between the positive-semidefinite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are written, e.g., as $\boldsymbol{A} \preceq \boldsymbol{B}$. The standard Euclidean basis vectors are written as $\mathbf{e}_1$, $\mathbf{e}_2$, and so on.

2.2 Problem formulation

Given a limited budget B with which we can pay to noisily observe a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, our goal is to construct a low-rank matrix $\overline{\boldsymbol{A}}$ that approximates $\boldsymbol{A}$ well. There are two modes of observation: very noisy and cheap samples of entire columns of $\boldsymbol{A}$, or much less noisy but expensive samples of individual entries of $\boldsymbol{A}$.

The following parameters quantify this two-cost observation model:

- Each low-noise sample of an individual entry costs p_e.

Algorithm 1 noisyCUR algorithm for completion of a low-rank matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$

```
Require: $d$, the number of column samples; $s$, the number of row samples; $\sigma_e$, the noise level of the column samples; $\sigma_c$, the noise level of the row samples; and $\lambda$, the regularization parameter
Ensure: $\overline{\boldsymbol{A}}$, approximation to $\boldsymbol{A}$
$\tilde{\boldsymbol{C}} \leftarrow \boldsymbol{C} + \boldsymbol{E}_c$, where the $d$ columns of $\boldsymbol{C}$ are sampled uniformly at random with replacement from $\boldsymbol{A}$, and the entries of $\boldsymbol{E}_c$ are i.i.d. $\mathcal{N}(0, \sigma_c^2)$.
$\boldsymbol{U} \leftarrow$ an orthonormal basis for the column span of $\tilde{\boldsymbol{C}}$
$\boldsymbol{\ell}_i \leftarrow \frac{1}{2}\|\mathbf{e}_i^T \boldsymbol{U}\|_2^2 / \|\boldsymbol{U}\|_F^2 + \frac{1}{2m}$ for $i = 1, \ldots, m$
$\boldsymbol{S} \leftarrow$ SamplingMatrix$(\boldsymbol{\ell}, m, s)$, the sketching matrix[1] used to sample $s$ rows of $\boldsymbol{A}$
$\boldsymbol{Y} = \boldsymbol{S}^T \boldsymbol{A} + \boldsymbol{E}_e$, where the entries of $\boldsymbol{E}_e$ are i.i.d. $\mathcal{N}(0, \sigma_e^2)$.
$\boldsymbol{X} \leftarrow \arg\min_{\boldsymbol{Z}} \|\boldsymbol{Y} - \boldsymbol{S}^T \tilde{\boldsymbol{C}} \boldsymbol{Z}\|_F^2 + \lambda \|\boldsymbol{Z}\|_F^2$
$\overline{\boldsymbol{A}} \leftarrow \tilde{\boldsymbol{C}} \boldsymbol{X}$
return $\overline{\boldsymbol{A}}$
```

- Each high-noise sample of a column costs $p_c > 0$. Because columns are cheap to sample, $p_c < mp_e$.
- The low-noise samples are each observed with additive $\mathcal{N}(0, \sigma_e^2)$ noise.
- Each entry of the high-noise column samples is observed with additive $\mathcal{N}(0, \sigma_c^2)$ noise. Because sampling columns is noisier than sampling entries, $\sigma_c^2 > \sigma_e^2$.

2.3 Related Work

To the best of the authors' knowledge, there is no prior work on budgeted low-rank MC. Standard approaches to matrix completion, e.g. [16,12,11] assume that the entries of the matrix are sampled uniformly at random, with or without noise. The most related works in the MC literature concern adaptive sampling models that attempt to identify more informative observations to reduce the sample complexity [6,14,21,13,3]. One of these works is [6], which estimates non-uniform sampling probabilities for the entries in $\boldsymbol{A}$ and then samples according to these to reduce the sample complexity. The work [14] similarly estimates the norms of the columns of $\boldsymbol{A}$ and then samples entries according to these to reduce the sample complexity. The work [21] proposes a regression-based algorithm for noiseless matrix completion that uses randomly sampled columns, rows, and entries of the matrix to form a low-rank approximation.

3 The noisyCUR algorithm

The noisyCUR (nCUR) algorithm, stated in Algorithm 1, is a regression-based MC algorithm. The intuition behind noisyCUR is that, if the columns of $\boldsymbol{A}$ were in general position and $\boldsymbol{A}$ were exactly rank-r, then one could recover $\boldsymbol{A}$

[1] SamplingMatrix$(\boldsymbol{\ell}, m, s)$ returns a matrix $\boldsymbol{S} \in \mathbb{R}^{m \times s}$ such that $\boldsymbol{S}^T \boldsymbol{A}$ samples and rescales, i.i.d. with replacement, s rows of $\boldsymbol{A}$ with probability proportional to their shrinked leverage scores. See the supplement for details.

by sampling exactly r columns of A and sampling r entries from each of the remaining columns, then using regression to infer the unobserved entries of the partially observed columns.

noisyCUR accomodates the presence of noise in the column and entry observations, and the fact that $\boldsymbol{A}$ is not low-rank. It samples d noisy columns from the low-rank matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, with entry-wise error that has variance σ_c^2 and collects these columns in the matrix $\tilde{\boldsymbol{C}} \in \mathbb{R}^{m \times d}$. It then uses $\tilde{\boldsymbol{C}}$ to approximate the span of $\boldsymbol{A}$ and returns an approximation of the form $\overline{\boldsymbol{A}} = \tilde{\boldsymbol{C}}\boldsymbol{X}$.

The optimal coefficient matrix $\boldsymbol{X}$ would require $\boldsymbol{A}$ to be fully observed. The sample complexity is reduced by instead sampling s rows noisily from $\boldsymbol{A}$, with entry-wise error that has variance σ_e^2, to form the matrix $\boldsymbol{Y}$. These rows are then used to estimate the coefficient matrix $\boldsymbol{X}$ by solving a ridge-regression problem; ridge regression is used instead of least-squares regression because in practice it better mitigates the noisiness of the observations. The rows are sampled according to the shrinked leverage scores [15,18] of the rows of $\tilde{\boldsymbol{C}}$.

The cost of observing the entries of $\boldsymbol{A}$ needed to form the approximation $\overline{\boldsymbol{A}}$ is $dp_c + snp_e$, so the number of column observations d and the number of row observations s are selected to use the entire budget, $B = dp_c + snp_e$.

Our main theoretical result is that when $\boldsymbol{A}$ is column-incoherent, has low condition number and is sufficiently dense, noisyCUR can exploit the two sampling modalities to achieve additive error guarantees in budget regimes where the relative error guarantees of nuclear norm completion do not apply.

First we recall the definition of column-incoherence.

Definition 1 *If $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{V} \in \mathbb{R}^{m \times r}$ is an orthonormal basis for the row space of $\boldsymbol{A}$, then the column leverage scores of $\boldsymbol{A}$ are given by*

$$\ell_i = \|\mathbf{e}_i^T \boldsymbol{V}\|_2^2 \quad \text{for } i = 1, \ldots, n.$$

The column coherence of $\boldsymbol{A}$ is the maximum of the column leverage scores of $\boldsymbol{A}$, and $\boldsymbol{A}$ is said to have a β-incoherent column space if its column coherence is smaller than $\beta \frac{r}{n}$.

Theorem 1 *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be a rank-r matrix with β-incoherent column space. Assume that $\boldsymbol{A}$ is dense: there is a $c > 0$ such that at least half* [2] *of the entries of $\boldsymbol{A}$ satisfy $|a_{ij}| \geq c$.*

Fix a precision parameter $\varepsilon \in (0,1)$ and invoke Algorithm 1 with

$$d \geq \max\left\{ \frac{6+2\varepsilon}{3\varepsilon^2}\beta r \log\frac{r}{\delta}, \frac{8(1+\delta)^2}{c^2(1-\varepsilon)\varepsilon} r\kappa_2(\boldsymbol{A})^2\sigma_c^2 \right\} \text{ and } s \geq \frac{6+2\varepsilon}{3\varepsilon^2} 2d \log\frac{d}{\delta}.$$

The returned approximation satisfies

$$\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_F^2 \leq \left(\gamma + \varepsilon + 40\frac{\varepsilon}{1-\varepsilon}\right)\|\boldsymbol{A}\|_F^2 + 12\varepsilon\left(\frac{\sigma_e^2}{\sigma_c^2}\right) d\sigma_r^2(\boldsymbol{A})$$

[2] This fraction can be changed, with corresponding modification to the sample complexities d and s.

with probability at least $0.9 - 2\delta - 2\exp(\frac{-(m-r)\delta^2}{2}) - \exp(\frac{-sn}{32})$. *Here,*

$$\gamma \leq 2\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\left[\frac{\lambda}{(1+\varepsilon)\left(\frac{1}{2}\sqrt{m-r}-\sqrt{d}\right)^2\sigma_c^2+\lambda}\right]^2.$$

The proof of this result is deferred.

Comparison of nuclear norm and noisyCUR approximation guarantees. Theorem 1 implies that, if $\boldsymbol{A}$ has low condition number, is dense and column-incoherent, and the regularization parameter λ is selected to be $o((\sqrt{m-r}-\sqrt{d})^2\sigma_c)$, then a mixed relative-additive bound of the form

$$\begin{aligned}\|\boldsymbol{A}-\overline{\boldsymbol{A}}\|_F^2 &\leq \varepsilon'\|\boldsymbol{A}\|_F^2+\varepsilon''\tilde{O}(r)\sigma_r^2(\boldsymbol{A})\\ &= \varepsilon'\|\boldsymbol{A}\|_F^2+\varepsilon''\tilde{O}(\|\boldsymbol{A}\|_F^2)\end{aligned} \tag{1}$$

holds with high probability for the approximation returned by Algorithm 1, where ε' and ε'' are $o(1)$, when d and s are $\tilde{\Omega}(r)$.

By way of comparison, the current best guarantees for noisy matrix completion using nuclear norm formulations state that if $d = \Omega((n+m)r\log(n+m))$ entries are noisily observed with noise level σ_e^2, then a nuclear norm formulation of the matrix completion problem yields an approximation $\overline{\boldsymbol{A}}$ that satisfies

$$\|\boldsymbol{A}-\overline{\boldsymbol{A}}\|_F^2 = O\left(\frac{\sigma_e^2}{\sigma_r^2(\boldsymbol{A})}\frac{nm}{r}\right)\|\boldsymbol{A}\|_F^2$$

with high probability [7,2]. The conditions imposed to obtain this guarantee [7] are that $\boldsymbol{A}$ has low condition number and that both its row and column spaces are incoherent. If we additionally require that $\boldsymbol{A}$ is dense, so that the assumptions applied to both algorithms are comparable, then $\|\boldsymbol{A}\|_F^2 = \Omega(mn)$ and the guarantee for nuclear norm completion becomes

$$\|\boldsymbol{A}-\overline{\boldsymbol{A}}\|_F^2 = O(\sigma_e^2\kappa_2^2(\boldsymbol{A}))\|\boldsymbol{A}\|_F^2. \tag{2}$$

Comparison of budget requirements for nuclear norm completion and noisy CUR. For nuclear norm completion approaches to assure guarantees of the form (2) it is necessary to obtain $\Omega((n+m)r\log(n+m))$ high precision noisy samples [4], so the budget required is

$$B_{NN} = \Omega((n+m)r\log(n+m)p_e).$$

When B_{NN} exceeds the budget B, there is no theory supporting the use of a mix of more expensive high and cheaper low precision noisy measurements.

The noisyCUR algorithm allows exactly such a mix: the cost of obtaining the necessary samples is

$$B_{nCUR} = dp_c + snp_e = \tilde{\Omega}(r)p_c + \tilde{\Omega}_r(nr)p_e,$$

where the notation $\tilde{\Omega}_r(\cdot)$ is used to indicate that the omitted logarithmic factors depend only on r. It is evident that

$$B_{nCUR} < B_{NN},$$

so the noisy CUR algorithm is applicable in budget regimes where nuclear norm completion is not.

3.1 Proof of Theorem 1

Theorem 1 is a consequence of two structural results that are established in the supplement.

The first result states that if $\operatorname{rank}(\boldsymbol{C}) = \operatorname{rank}(\boldsymbol{A})$ and the bottom singular value of $\boldsymbol{C}$ is large compared to σ_c, then the span of $\tilde{\boldsymbol{C}}$ will contain a good approximation to $\boldsymbol{A}$.

Lemma 1 *Fix an orthonormal basis $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and consider $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{C} \in \mathbb{R}^{m\times d}$ with factorizations $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{M}$ and $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{W}$, where both $\boldsymbol{M}$ and $\boldsymbol{W}$ have full row rank. Further, let $\tilde{\boldsymbol{C}}$ be a noisy observation of $\boldsymbol{C}$, that is, let $\tilde{\boldsymbol{C}} = \boldsymbol{C} + \boldsymbol{G}$ where the entries of $\boldsymbol{G}$ are i.i.d. $\mathcal{N}(0, \sigma_c^2)$. If $\sigma_{\min}(\boldsymbol{C}) \geq 2(1+\delta)\sigma_c\sqrt{m/\varepsilon}$, then*

$$\|(\boldsymbol{I} - \boldsymbol{P}_{\tilde{\boldsymbol{C}}})\boldsymbol{A}\|_F^2 \leq \varepsilon\|\boldsymbol{A}\|_F^2$$

with probability at least $1 - \exp\left(\frac{-m\delta^2}{2}\right)$.

Recall the definition of a $(1 \pm \varepsilon)$-subspace embedding.

Definition 2 (Subspace embedding [20]) *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and fix $\varepsilon \in (0, 1)$. A matrix $\boldsymbol{S} \in \mathbb{R}^{m\times s}$ is a $(1 \pm \varepsilon)$-subspace embedding for $\boldsymbol{A}$ if*

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|\boldsymbol{S}^T\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2$$

for all vectors $\mathbf{x}$ in the span of $\boldsymbol{A}$, or equivalently, if

$$(1-\varepsilon)\boldsymbol{A}^T\boldsymbol{A} \preceq \boldsymbol{A}^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{A} \preceq (1+\varepsilon)\boldsymbol{A}^T\boldsymbol{A}.$$

Often we will use the shorthand "subspace embedding" for $(1 \pm \varepsilon)$-subspace embedding.

The second structural result is a novel bound on the error of sketching using a subspace embedding to reduce the cost of ridge regression, when the target is noisy.

Corollary 1 *Let $\tilde{\boldsymbol{C}} \in \mathbb{R}^{m\times d}$, where $d \leq m$, and $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{E}$ be matrices, and let $\boldsymbol{S}$ be an $(1 \pm \varepsilon)$-subspace embedding for $\tilde{\boldsymbol{C}}$. If*

$$\boldsymbol{X} = \arg\min_{\boldsymbol{Z}} \|\boldsymbol{S}^T(\tilde{\boldsymbol{A}} - \tilde{\boldsymbol{C}}\boldsymbol{Z})\|_F^2 + \lambda\|\boldsymbol{Z}\|_F^2,$$

then

$$\|\boldsymbol{A}-\tilde{\boldsymbol{C}}\boldsymbol{X}\|_F^2 \leq \|(\boldsymbol{I}-\boldsymbol{P}_{\tilde{\boldsymbol{C}}})\boldsymbol{A}\|_F^2+\gamma\|\boldsymbol{P}_{\tilde{\boldsymbol{C}}}\boldsymbol{A}\|_F^2+\frac{4}{1-\varepsilon}\|\boldsymbol{S}^T\boldsymbol{E}\|_F^2+\frac{4}{1-\varepsilon}\|\boldsymbol{S}^T(\boldsymbol{I}-\boldsymbol{P}_{\tilde{\boldsymbol{C}}})\boldsymbol{A}\|_F^2,$$

where $\gamma = 2\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\left(\frac{\lambda}{(1+\varepsilon)\sigma_d(\tilde{\boldsymbol{C}})^2+\lambda}\right)^2$.

Corollary 1 differs significantly from prior results on the error in sketched ridge regression, e.g. [1,18], in that: (1) it bounds the *reconstruction error* rather than the *ridge regression objective*, and (2) it considers the impact of noise in the target. This result follows from a more general result on sketched noisy proximally regularized least squares problems, stated as Theorem 2 in the supplement.

Together with standard properties of Gaussian noise and subspace embeddings, these two results deliver Theorem 1.

Proof (Proof of Theorem 1). The noisyCUR algorithm first forms the noisy column samples $\tilde{\boldsymbol{C}} = \boldsymbol{C} + \boldsymbol{E}_c$, where $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{M}$. The random matrix $\boldsymbol{M} \in \mathbb{R}^{n\times d}$ selects d columns uniformly at random with replacement from the columns of $\boldsymbol{A}$, and the entries of $\boldsymbol{E}_c \in \mathbb{R}^{m\times d}$ are i.i.d. $\mathcal{N}(0, \sigma_c^2)$. It then solves the sketched regression problem

$$\boldsymbol{X} = \arg\min_{\boldsymbol{Z}} \|\boldsymbol{S}^T(\tilde{\boldsymbol{A}} - \tilde{\boldsymbol{C}}\boldsymbol{Z})\|_F^2 + \lambda\|\boldsymbol{Z}\|_F^2,$$

and returns the approximation $\overline{\boldsymbol{A}} = \tilde{\boldsymbol{C}}\boldsymbol{X}$. Here $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{E}_e$, where $\boldsymbol{E}_e \in \mathbb{R}^{m\times n}$ comprises i.i.d $\mathcal{N}(0, \sigma_e^2)$ entries, and the sketching matrix $\boldsymbol{S} \in \mathbb{R}^{m\times s}$ samples s rows using the shrinked leverage scores of $\tilde{\boldsymbol{C}}$.

By [18, Appendix A.1.1], $\boldsymbol{S}$ is a subspace embedding for $\tilde{\boldsymbol{C}}$ with failure probability at most δ when s is as specified. Thus Corollary 1 applies and gives that

$$\begin{aligned}\|\boldsymbol{A} - \tilde{\boldsymbol{C}}\boldsymbol{X}\|_F^2 \leq \|(\boldsymbol{I} - \boldsymbol{P}_{\tilde{\boldsymbol{C}}})\boldsymbol{A}\|_F^2 &+ \gamma'\|\boldsymbol{P}_{\tilde{\boldsymbol{C}}}\boldsymbol{A}\|_F^2 \\ &+ \frac{4}{1-\varepsilon}\|\boldsymbol{S}^T\boldsymbol{E}\|_F^2 + \frac{4}{1-\varepsilon}\|\boldsymbol{S}^T(\boldsymbol{I} - \boldsymbol{P}_{\tilde{\boldsymbol{C}}})\boldsymbol{A}\|_F^2 \\ &= T_1 + T_2 + T_3 + T_4,\end{aligned}$$

where $\gamma' = 2\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\left(\frac{\lambda}{(1+\varepsilon)\sigma_d(\tilde{\boldsymbol{C}})^2+\lambda}\right)^2$. We now bound the four terms T_1, T_2, T_3, and T_4.

To bound T_1, note that by [19, Lemma 13], the matrix $\sqrt{\frac{n}{d}}\boldsymbol{M}$ is a subspace embedding for $\boldsymbol{A}^T$ with failure probability at most δ when d is as specified. This gives the semidefinite inequality $\frac{n}{d}\boldsymbol{C}\boldsymbol{C}^T = \frac{n}{d}\boldsymbol{A}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{A}^T \succeq (1-\varepsilon)\boldsymbol{A}\boldsymbol{A}^T$, which in turn gives that

$$\begin{aligned}\sigma_r^2(\boldsymbol{C}) \geq (1-\varepsilon)\frac{d}{n}\sigma_r^2(\boldsymbol{A}) &\geq \frac{8(1+\delta)^2}{c^2\varepsilon}\frac{r}{n}\|\boldsymbol{A}\|_2^2\sigma_c^2 \\ &\geq \frac{8(1+\delta)^2}{c^2\varepsilon n}\|\boldsymbol{A}\|_F^2\sigma_c^2 \geq 4(1+\delta)^2\frac{m}{\varepsilon}\sigma_c^2.\end{aligned}$$

The second inequality holds because

$$d \geq \frac{8(1+\delta)^2}{c^2(1-\varepsilon)\varepsilon} r\kappa_2(\boldsymbol{A})^2\sigma_c^2 \quad \text{implies} \quad \sigma_r^2(\boldsymbol{A}) \geq \frac{8(1+\delta)^2}{c^2(1-\varepsilon)\varepsilon}\frac{r}{d}\|\boldsymbol{A}_2\|^2\sigma_c^2 \tag{3}$$

The third inequality holds because $r\|\boldsymbol{A}\|_2^2$ is an overestimate of $\|\boldsymbol{A}_F\|_2^2$. The final inequality holds because the denseness of $\boldsymbol{A}$ implies that $\|\boldsymbol{A}\|_F^2 \geq \frac{1}{2}c^2mn$.

Note also that the span of $\boldsymbol{C} = \boldsymbol{AM}$ is contained in that of $\boldsymbol{A}$, and since $\frac{n}{d}\boldsymbol{CC}^T \succeq (1-\varepsilon)\boldsymbol{AA}^T$, in fact $\boldsymbol{C}$ and $\boldsymbol{A}$ have the same rank and therefore span the same space. Thus the necessary conditions to apply Lemma 1 are satisfied, and as a result, we find that

$$T_1 \leq \varepsilon\|\boldsymbol{A}\|_F^2$$

with failure probability at most $\exp(-\frac{m\delta^2}{2})$.

Next we bound T_2. Observe that $\|\boldsymbol{P}_{\tilde{\boldsymbol{C}}}\boldsymbol{A}\|_F^2 \leq \|\boldsymbol{A}\|_F^2$. Further, by Lemma 3 in the supplement,

$$\sigma_d(\tilde{\boldsymbol{C}}) \geq \left(\frac{1}{2}\sqrt{m-r} - \sqrt{d}\right)^2 \sigma_c^2$$

with failure probability at most $\exp(\frac{-(m-r)\delta^2}{2})$. This allows us to conclude that

$$T_2 \leq \gamma\|\boldsymbol{A}\|_F^2,$$

where γ is as specified in the statement of this theorem.

To bound T_3, we write

$$\begin{aligned} T_3 &= \frac{4}{1-\varepsilon}\|\boldsymbol{S}^T\boldsymbol{P}_{\boldsymbol{S}}\boldsymbol{E}\|_F^2 \leq \frac{4}{1-\varepsilon}\|\boldsymbol{S}\|_2^2\|\boldsymbol{P}_{\boldsymbol{S}}\boldsymbol{E}\|_F^2 \\ &\leq \frac{8}{1-\varepsilon}\frac{m}{s}\|\boldsymbol{Q}^T\boldsymbol{E}\|_F^2, \end{aligned}$$

where $\boldsymbol{Q}$ is an orthonormal basis for the span of $\boldsymbol{S}$. The last inequality holds because [18, Appendix A.1.2] shows that $\|\boldsymbol{S}\|_2^2 \leq 2\frac{m}{s}$ always. Finally, note that $\boldsymbol{Q}$ has at most s columns, so in the worst case $\boldsymbol{Q}^T\boldsymbol{E}$ comprises sn i.i.d. $\mathcal{N}(0, \sigma_e^2)$ entries. A standard concentration bound for χ^2 random variables with sn degrees of freedom [17, Example 2.11] guarantees that

$$\|\boldsymbol{Q}^T\boldsymbol{E}\|_F^2 \leq \frac{3}{2}sn\sigma_e^2$$

with failure probability at most $\exp(\frac{-sn}{32})$. We conclude that, with the same failure probability,

$$T_3 \leq \frac{12}{1-\varepsilon}mn\sigma_e^2.$$

Now recall (3), which implies that

$$\begin{aligned} \varepsilon(1-\varepsilon)d\sigma_r^2(\boldsymbol{A}) &\geq \frac{8(1+\delta)^2}{c^2}r\|\boldsymbol{A}_2\|^2\sigma_c^2 \geq \frac{8(1+\delta)^2}{c^2}\|\boldsymbol{A}\|_F^2\sigma_c^2 \\ &\geq 4(1+\delta)^2mn\sigma_c^2 \geq mn\sigma_c^2. \end{aligned}$$

It follows from the last two displays that

$$T_3 \leq 12\varepsilon \left(\frac{\sigma_e^2}{\sigma_c^2} \right) d\sigma_r^2(\boldsymbol{A}).$$

The bound for T_4 is an application of Markov's inequality. In particular, it is readily verifiable that $\mathbb{E}[\boldsymbol{S}\boldsymbol{S}^T] = \boldsymbol{I}$, which implies that

$$\mathbb{E}T_4 = \frac{4}{1-\varepsilon} \|(\boldsymbol{I} - \boldsymbol{P}_{\tilde{\boldsymbol{C}}})\boldsymbol{A}\|_F^2 = \frac{4}{1-\varepsilon} T_1 \leq \frac{4\varepsilon}{1-\varepsilon} \|\boldsymbol{A}\|_F^2.$$

The final inequality comes from the bound $T_1 \leq \varepsilon \|\boldsymbol{A}\|_F^2$ that was shown earlier. Thus, by Markov's inequality,

$$T_4 \leq \frac{40\varepsilon}{1-\varepsilon} \|\boldsymbol{A}\|_F^2$$

with failure probability at most 0.1.

Collating the bounds for T_1 through T_4 and their corresponding failure probabilities gives the claimed result.

4 Empirical Evaluation

In this section we investigate the performance of the noisyCUR method on a small-scale synthetic data set and on the Jester and MovieLens data sets. We compare with the performance of three nuclear norm-based algorithms in a low and a high-budget regime.

4.1 Experimental setup

Four parameters are manipulated to control the experiment setup:

1. The budget, taken to be of the size $B = c_0 m r p_e$ for some constant positive integer c_0. This choice ensures that the $O((n+m)r\log(n+m))$ high precision samples needed for nuclear norm completion methods cannot be obtained.
2. The ratio of the cost of sampling a column to that of individually sampling each entry in that column, $\alpha = \frac{p_c}{m p_e}$. For all three experiments, we set $\alpha = 0.2$.
3. The entry sampling noise level σ_e^2.
4. The column sampling noise level σ_c^2.

Based on the signal-to-noise ratio between the matrix and the noise level of the noisiest observation model, σ_c^2, we classify an experiment as being high noise or low noise. The entry-wise signal-to-noise ratio is given by

$$SNR = \frac{\|\boldsymbol{A}\|_F^2}{mn\sigma_c^2}.$$

High SNR experiments are said to be low noise, while those with low SNR are said to be high noise.

4.2 Methodology: noisyCUR and the baselines

We compare to three nuclear norm-based MC algorithms, as nuclear norm-based approaches are the most widely used and theoretically investigated algorithms for low-rank MC. We additionally compare to the CUR+ algorithm of [21] as it is, similarly to noisyCUR, a regression-based MC algorithm.

To explain the baselines, we introduce some notation. Given a set of indices Ω, the operator $\mathcal{P}_\Omega : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ returns a matrix whose values are the same as those of the input matrix on the indices in Ω, and zero on any indices not in Ω. The set Ω_s below comprises the indices of entries of $\boldsymbol{A}$ sampled with high accuracy, while Ω_c comprises the indices of entries of $\boldsymbol{A}$ sampled using the low accuracy column observation model.

(nCUR) Given the settings of the two-cost model, the noisyCUR algorithm is employed by selecting a value for d, the number of noisy column samples; the remaining budget then determines s, the number of rows that are sampled with high precision. Cross-validation is used to select the regularization parameter λ.

(CUR+) The CUR+ algorithm is designed for noiseless matrix completion [21]; it is adapted in a straightforward manner to our setting. Now d is the number of noisy row and column samples, and $d/2$ columns and $d/2$ rows are sampled uniformly with replacement from $\boldsymbol{A}$ and noisily observed to form column and row matrices $\boldsymbol{C}$ and $\boldsymbol{R}$. The remaining budget is used to sample entries to form Ω_e and $\boldsymbol{A}_{\text{obs}}$, the partially observed matrix which contains the observed noisy entry samples and is zero elsewhere. The CUR+ algorithm then returns the low-rank approximation $\overline{\boldsymbol{A}} = \boldsymbol{CUR}$, where $\boldsymbol{U}$ is obtained by solving

$$\boldsymbol{U} = \arg\min \|\mathcal{P}_{\Omega_e}(\boldsymbol{A}_{\text{obs}} - \boldsymbol{CUR})\|_F^2.$$

(NNa) The first of the nuclear norm baselines is the formulation introduced in [5], which forms the approximation

$$\begin{aligned} \overline{\boldsymbol{A}} = & \arg\min\nolimits_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_\star \\ & \text{s.t. } \|\mathcal{P}_{\Omega_e}(\boldsymbol{Z} - \boldsymbol{A}_{\text{obs}})\|_F \leq \delta, \quad (i,j) \in \Omega_e. \end{aligned} \tag{4}$$

All of the budget is spend on sampling entries to form Ω_e and $\boldsymbol{A}_{\text{obs}}$, the partially observed matrix which contains the observed noisy entry samples and is zero elsewhere. Thus the performance of this model is a constant independent of d in the figures. The hyperparameter δ is selected through cross-validation. This baseline is referred to as NNa (nuclear norm minimization for all entries) in the figures.

(NNs) The second nuclear norm baseline is a nuclear norm formulation that penalizes the two forms of observations separately, forming the approximation

$$\begin{aligned} \overline{\boldsymbol{A}} = & \arg\min\nolimits_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_\star \\ & \text{s.t. } \|\mathcal{P}_{\Omega_c}(\boldsymbol{Z} - \boldsymbol{A}_{\text{obs}})\|_F^2 \leq C_1 d m \sigma_c^2 \\ & \quad \|\mathcal{P}_{\Omega_e}(\boldsymbol{Z} - \boldsymbol{A}_{\text{obs}})\|_F^2 \leq C_2 f \sigma_e^2 \end{aligned} \tag{5}$$

where C_1 and C_2 are parameters, and again $\boldsymbol{A}_{\text{obs}}$ is the partially observed matrix which contains the observed noisy column samples and entry samples and is zero elsewhere. As with the noisyCUR method, given a value of d, the remaining budget is spent on sampling s rows with high precision. The hyperparameters C_1 and C_2 are selected through cross-validation. This baseline is referred to as NNs (nuclear norm split minimization) in the figures.

(Chen) The final nuclear norm baseline is an adaptation of the two-phase sampling method of [6]. This method spends a portion of the budget to uniformly at random sample entries to estimate leverage scores, then uses the rest of the budget to sample entries according to the leverage score and reconstructs from the union of these samples using the same optimization as NNa. The performance is therefore independent of d. This baseline is referred to as Chen in the figures.

The details of cross-validation of the parameters for the nuclear norm methods are omitted because there are many relevant hyperparameters: in addition to the constraint parameters in the optimization formulations, there are important hyperparameters associated with the ADMM solvers used (e.g., the Lagrangian penalty parameters).

4.3 Synthetic Dataset

Figure 2 compares the performance of the baseline methods and noisyCUR on an incoherent synthetic data set $\boldsymbol{A} \in \mathbb{R}^{80 \times 60}$ generated by sampling a matrix with i.i.d. $\mathcal{N}(5, 1)$ entries and taking its best rank four approximation. For each value of d, the regularization parameter λ of noisyCUR is selected via cross-validation from 500 logarithmically spaced points in the interval $(10^{-4}, 10)$.

4.4 Jester

Figure 3 compares the performance of the baseline methods and noisyCUR on a subset of the Jester dataset of [9]. Specifically the data set was constructed by extracting the submatrix comprising the 7200 users who rated all 100 jokes. For each value of d, the regularization parameter λ of noisyCUR is selected via cross-validation from 200 points logarithmically spaced in the interval $(10, 10^5)$.

4.5 Movielens-100K

Figure 4 compares the performance of the baseline methods and noisyCUR on the Movielens-100K dataset of [10]. The original 1682×943 data matrix is quite sparse, so the iterative SVD algorithm of [8] is used to complete it to a low-rank matrix before applying noisyCUR and the three baseline algorithms. For each value of d, the parameter λ of noisyCUR is cross-validated among 200 points linearly spaced in the interval $[1, 200]$.

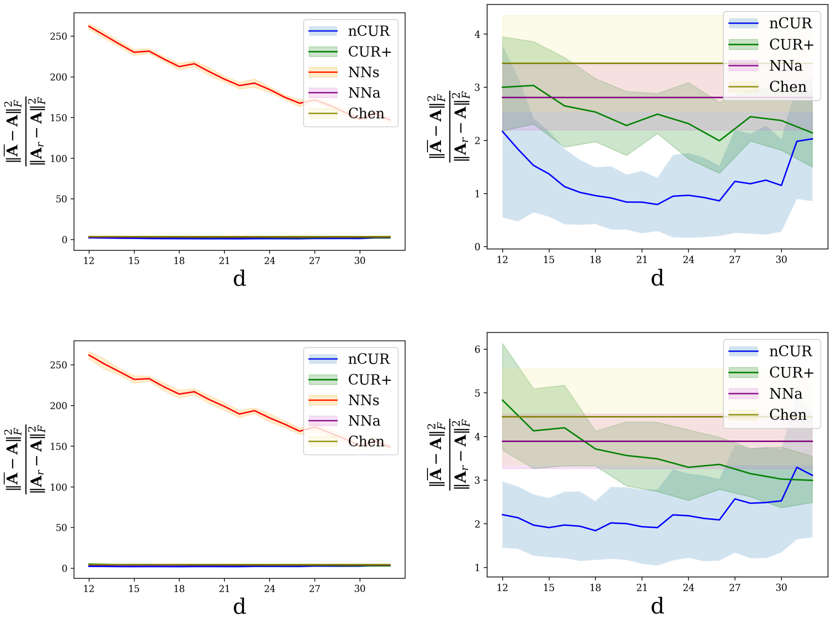

Fig. 2: Performance on the synthetic dataset. $\beta = 15\%$ for all plots. The noise level is low in the top plots: $\sigma_e^2 = 0.01, \sigma_c^2 = 0.05$. The upper left plot shows all methods, while the upper right plot removes the NNs method to facilitate comparison of the better performing methods. In the bottom two plots, the noise level is higher: $\sigma_e^2 = 0.04, \sigma_c^2 = 0.2$. Similarly, the bottom left plot shows all methods, while the bottom right removes the NNs method. Each point in the plots is the average of 100 runs.

4.6 Observations

In both the lower and higher noise regimes on all the datasets, noisyCUR exhibits superior performance when compared to the nuclear norm baselines, and in all but one of the experiments, noisyCUR outperforms CUR+, the regression-based baseline. Also, noisyCUR produces a v-shaped error curve, where the optimal approximation error is achieved at the bottom of the v. This convex shape of the performance of noisyCUR with respect to d suggests that there may be a single optimal d given a dataset and the parameters of the two-cost model.

We note a practical consideration that arose in the empirical evaluations: the noisyCUR method is much faster than the nuclear norm algorithms, as they invoke an iterative solver (ADMM was used in these experiments) that computes SVDs of large matrices during each iteration, while noisyCUR solves a single ridge regression problem.

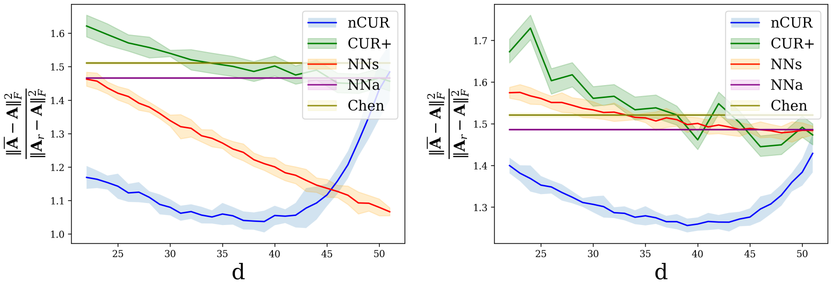

Fig. 3: Performance on the Jester data set. $\beta = 11\%$ for all plots. The noise level is low in the top plots: $\sigma_e^2 = 0.04, \sigma_c^2 = 2$. In the bottom two plots, the noise level is higher: $\sigma_e^2 = 0.25$, $\sigma_c^2 = 12.5$. As in Figure 2, the plots to the left contain the NNs baseline while those to the right do not. Each point in the plots is the average over 10 runs.

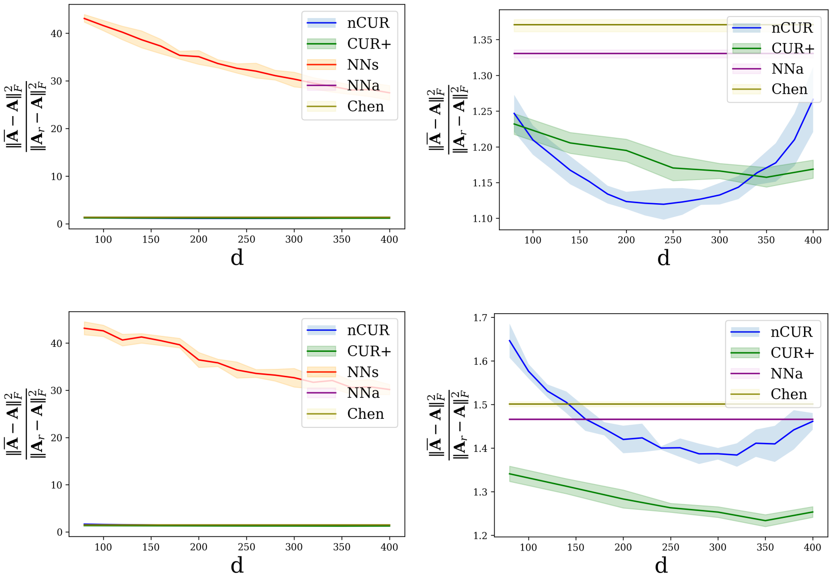

Fig. 4: Performance on the MovieLens-100K data set. $\beta = 10.6\%$ for all plots. The noise level is low in the upper two plots: $\sigma_e^2 = 0.003, \sigma_c^2 = 0.06$. In the bottom two plots, the noise level is higher: $\sigma_e^2 = 0.0225, \sigma_c^2 = 0.65$. As in Figure 2, the plots to the left show the NNs baseline while the plots to the right do not. Each point in the plots is the average of 10 runs.

5 Conclusion

This paper introduced the noisyCUR algorithm for solving a budgeted matrix completion problem where the two observation modalities consist of high-accuracy expensive entry sampling and low-accuracy inexpensive column sampling. Recovery guarantees were proven for noisyCUR; these hold even in low-budget regimes where standard nuclear norm completion approaches have no recovery guarantees. noisyCUR is fast, as the main computation involved is a ridge regression. Empirically, it was shown that noisyCUR has lower reconstruction error than standard nuclear norm completion baselines in the low-budget setting. It is an interesting open problem to determine optimal or near-optimal values for the number of column and row samples (d and s), given the parameters of the two-cost model (B, p_e, p_c, σ_e, and σ_c). Implementations of noisyCUR and the baseline methods used in the experimental evaluations are available at `https://github.com/jurohd/nCUR`, along with code for replicating the experiments.

References

1. Avron, H., Clarkson, K.L., Woodruff, D.P.: Sharper Bounds for Regularized Data Fitting. In: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017). vol. 81, pp. 27:1–27:22. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik (2017)
2. Balcan, M.F., Liang, Z., Song, Y., Woodruff, D.P., Zhang, H.: Non-convex matrix completion and related problems via strong duality. Journal of Machine Learning Research **20**(102), 1–56 (2019)
3. Balcan, M.F., Zhang, H.: Noise-tolerant life-long matrix completion via adaptive sampling. In: Advances in Neural Information Processing Systems 29, pp. 2955–2963. Curran Associates, Inc. (2016)
4. Candés, E.J., Tao, T.: The power of convex relaxation: Near-optimal matrix completion. IEEE Transactions on Information Theory **56**(5), 2053–2080 (2010)
5. Candes, E.J., Plan, Y.: Matrix completion with noise. Proceedings of the IEEE **98**(6), 925–936 (2010)
6. Chen, Y., Bhojanapalli, S., Sanghavi, S., Ward, R.: Coherent matrix completion. Proceedings of The 31st International Conference on Machine Learning (ICML) pp. 674–682 (2014)
7. Chen, Y., Chi, Y., Fan, J., Ma, C., Yan, Y.: Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. arXiv preprint, arXiv:1902.07698 (2019)
8. Cho, K., Reyhani, N.: An iterative algorithm for singular value decomposition on noisy incomplete matrices. In: The 2012 International Joint Conference on Neural Networks (IJCNN). pp. 1–6 (2012)
9. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: An iterative algorithm for singular value decomposition on noisy incomplete matrices. Information Retrieval **6**(2), 133–151 (2001)
10. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5 (2015)

11. Hastie, T., Mazumder, R., Lee, J.D., Zadeh, R.: Matrix completion and low-rank SVD via fast alternating least squares. Journal of Machine Learning Research **16**(1), 3367–3402 (2015)
12. Keshavan, R., Montanari, A., Oh, S.: Matrix completion from noisy entries. In: Advances in Neural Information Processing Systems. pp. 952–960 (2009)
13. Krishnamurthy, A., Singh, A.: Low-rank matrix and tensor completion via adaptive sampling. In: Advances in Neural Information Processing Systems 26, pp. 836–844. Curran Associates, Inc. (2013)
14. Krishnamurthy, A., Singh, A.R.: On the power of adaptivity in matrix completion and approximation. arXiv preprint arXiv:1407.3619 (2014)
15. Ma, P., Mahoney, M.W., Yu, B.: A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research **16**(1), 861–911 (2015)
16. Recht, B.: A simpler approach to matrix completion. Journal of Machine Learning Research **12**(12) (2011)
17. Wainwright, M.J.: High-dimensional statistics: a non-asymptotic viewpoint. Cambridge University Press (2019)
18. Wang, S., Gittens, A., Mahoney, M.W.: Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. Journal of Machine Learning Research **18**(218), 1–50 (2018)
19. Wang, S., Zhang, Z., Zhang, T.: Towards more efficient SPSD matrix approximation and CUR matrix decomposition. Journal of Machine Learning Research **17**(1), 7329–7377 (2016)
20. Woodruff, D.P.: Sketching as a tool for numerical linear algebra. Foundations and Trends® in Theoretical Computer Science **10**(1–2), 1–157 (2014)
21. Xu, M., Jin, R., Zhou, Z.H.: CUR algorithm for partially observed matrices. In: Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 1412–1421. PMLR (2015)