Tractable Learning of Sparsely Used Dictionaries from Incomplete Samples

Thanh V. Nguyen, Akshay Soni, and Chinmay Hegde*

***Abstract*—In dictionary learning, we seek a collection of atoms that sparsely represent a given set of training samples. While this problem is well-studied, relatively less is known about the more challenging case where the samples are incomplete, i.e., we only observe a fraction of their coordinates. In this paper, we develop and analyze an algorithm to solve this problem, provided that the dictionary satisfies additional low-dimensional structure.**

I. Introduction

In this paper, we consider a variant of the problem of *dictionary learning*, a widely used unsupervised technique for learning compact (sparse) representations of high dimensional data. At its core, the challenge in dictionary learning is to discover a basis (or dictionary) that can sparsely represent a given set of data samples with as little empirical representation error as possible. An important underlying assumption that guides the success of all existing dictionary learning algorithms is the availability of (sufficiently many) data samples that are fully observed. Our focus, on the other hand, is on the special case where the given data points are *only partially observed*, that is, we are given access to only a small fraction of the coordinates of the data samples. Such a "missing data" setting arises naturally in applications such as image-inpainting and demosaicing [1], and hyper-spectral imaging [2].

Earlier works that tackle the incomplete variant of the dictionary learning problem only offer heuristic solutions [2], [3] or involve constructing intractable statistical estimators [4]. Indeed, the recovery of the true dictionary involves analyzing an highly non-convex optimization problem that is, in general, not solvable in polynomial time [5]. To our knowledge, our recent work [6] was the first to give a theoretically sound as well as tractable algorithm to learn the dictionary from incompletely observed samples. However, a key requirement of our approach was the availability of a small hold-out set of *fully* observed samples. In this paper, we circumvent this requirement, provided that the atoms of the unknown dictionary exhibits additional low-dimensional structure.

Our Contributions. Following [7], [8], we assume that each data sample is synthesized from a generative model with an unknown dictionary and a random k-*sparse* coefficient vector (or sparse code). Mathematically, the data samples $Z = [z^{(1)}, z^{(2)}, \ldots, z^{(p)}] \in \mathbb{R}^{n \times p}$ are of the form $Z = A^* X^*$, where $A^* \in \mathbb{R}^{n \times m}$ denotes the dictionary and $X^* \in \mathbb{R}^{m \times p}$ denotes the (column-wise) k-sparse codes. However, we do not have access to the full data samples; instead, each entry of Z is observed independently with probability $\rho \in (0, 1]$. For analysis, we also make the (standard) assumptions that A^* is both *incoherent* (i.e., the columns of A^* are sufficiently close to orthogonal) and *democratic* (i.e., the energy of each atom is well spread). Given a set of such (partially observed) data samples, our goal is to recover the true dictionary A^*. In recent work, [6] we presented a tractable algorithm with convergence results to perform this recovery, *provided* we have a coarse estimate A^0 that is sufficiently close to A^*. In that work, we also presented an algorithm to produce such A^0, albeit by requiring a small hold-out set of *fully* observed samples.

In this paper, we remove this requirement: we present a practical algorithm to estimate A^* from only incomplete samples, provided A^* obeys additional low-dimensional structure.

Techniques. We build upon recent algorithms for dictionary learning — specifically, the framework of [8], which proposes a descent-like algorithm over the dictionary parameters. The descent is achieved by alternating between updating the dictionary estimate and solving for the sparse codes of the data samples given the dictionary. This algorithm provably succeeds so long as the codes are sparse enough, the columns of A^* are incoherent, and we are given sufficiently many samples.

However, a direct application of the above framework to the missing data setting does not work. To resolve this, we leverage a specific property that is commonly assumed in the matrix completion literature: we suppose that the dictionaries are not "spiky" and that the energy of each atom is spread out among its coordinates; specifically, the *sub*-dictionaries formed by randomly sub-selecting rows are still incoherent. We call such dictionaries *democratic*, following the terminology of [9]. Our main contribution in [6] proves that democratic, incoherent dictionaries can be learned via a similar alternating descent scheme if only a small fraction of the data entries are available.

Of course, the above analysis is somewhat local in nature since we are using a descent-style method over a non-convex loss function. In order to get global recovery guarantees, we need to initialize carefully. To achieve this, we leverage known results in matrix completion [10] to prove that a starting estimate for the descent within the basin of attraction of A^* can be constructed in polynomial time, assuming that A^* obeys additional low-dimensional structure. This addresses an open problem in [6], albeit at the cost of the above assumption.

Prior Work. The literature on dictionary learning (or sparse coding) is very vast; cf. [1]. Dictionary learning with incompletely observed data, however, is far less well-studied.

*Email: {thanhng, chinmay}@iastate.edu; akson@microsoft.com. T. N. and C. H. are with the Electrical and Computer Engineering Department at Iowa State University. A. S. is with Microsoft. This work was supported in part by the National Science Foundation under grants CCF-1566281 and CCF-1750920, and in part by a Faculty Fellowship from the Black and Veatch Foundation.

Initial attempts in this direction [2] involve Bayesian-style techniques, while more recent attempts have focused on alternating minimization heuristics [3]. However, none of these methods provide rigorous polynomial-time algorithms that provably succeed in recovering the dictionary parameters.

Our setup can also be viewed as an instance of matrix completion, which has been a source of great interest in the machine learning community over the last decade [11], [12]. The typical assumption in such approaches is that the data matrix $Z = A^*X^*$ is low-rank (i.e., A^* typically spans a low-dimensional subspace). This assumption leads to either feasible convex relaxations, or a bilinear form that can be solved approximately via alternating minimization. However, our work differs significantly from conventional matrix completion, since our guarantees are not in terms of estimating the missing entries of Y, but rather obtaining the atoms in A^*.

In the context of matrix-completion, perhaps the most related work to ours is the statistical analysis of matrix-completion under the *sparse-factor model* of [4], which employs a similar generative data model to ours. (Similar sparse-factor models have been studied in the work of [13], but no complexity guarantees are provided.) For this model, [4] propose a non-convex statistical estimator for estimate Z and provide error bounds for this estimator under various noise models. However, they do not discuss an efficient algorithm to realize that estimator. In contrast, we provide a rigorous polynomial time algorithm, together with recovery error bounds. Overall, our work could shed some light on the design of provable algorithms for matrix completion in such more general settings.

II. Preliminaries

Given a vector $x \in \mathbb{R}^m$ and a subset $S \subseteq [m]$, denote $x_S \in \mathbb{R}^m$ as a vector which equals x in indices belonging to S and equals zero elsewhere. Denote by $A_{\Gamma\bullet}$ the submatrix of A with rows not in Γ set to zero. The symbol $\|\cdot\|$ refers to the ℓ_2-norm, unless otherwise specified. We adopt standard big-O notation. The terms "with high probability" (sometimes in abbreviation as w.h.p.) indicates an event with failure probability $O(n^{-\omega(1)})$.

We make two basic assumptions on our dictionary A^*: *incoherence*, and *democracy*. The incoherence property requires the columns of A^* to be approximately orthogonal, and is a canonical property to resolve identifiability issues in dictionary learning and sparse recovery. The *democracy* property shows that the rows of A^* roughly have the same amount of mass. Formally, we have:

Definition 1 (Incoherence)**.** *The matrix A is* incoherent *with parameter μ if the following holds for all columns $i \neq j$:*

$$\frac{|\langle A_{\bullet i}, A_{\bullet j}\rangle|}{\|A_{\bullet i}\|\|A_{\bullet j}\|} \leq \frac{\mu}{\sqrt{n}}.$$

Definition 2 (Democracy)**.** *Suppose A is μ-incoherent. A is further said to be* democratic *if the submatrix $A_{\Gamma\bullet}$ is μ-incoherent for any subset $\Gamma \subset [n]$ of size $\sqrt{n} \leq |\Gamma| \leq n$.*

We seek an algorithm that provides a provably "good" estimate of A^*. For this, we need a suitable measure of "goodness". The following notion of distance records the maximal column-wise difference between any estimate A and A^* in ℓ_2-norm under a suitable permutation and sign flip.

Definition 3 ((δ,κ)-nearness)**.** *The matrix A is said to be δ-close to A^* if $\|\sigma(i)A_{\bullet\pi(i)} - A^*_{\bullet i}\| \leq \delta$ holds for every $i = 1, 2, \ldots, m$ and some permutation $\pi : [m] \to [m]$ and sign flip $\sigma : [m] : \{\pm 1\}$. In addition, if $\|A_{\bullet\pi} - A^*\| \leq \kappa\|A^*\|$ holds, then A is said to be (δ,κ)-near to A^*.*

To keep notation simple, in our convergence theorems below, whenever we discuss nearness, we simply replace the transformations π and σ in the above definition with the identity mapping $\pi(i) = i$ and the positive sign $\sigma(\cdot) = +1$ while keeping in mind that in reality, we are referring to finding one element in the equivalence class of permutations and sign flips.

Armed with the above concepts, we now posit a generative model for our observed data. Suppose that we observe p data samples $Y = [y^{(1)}, y^{(2)}, \ldots, y^{(p)}]$ such that each sample is generated according to the rule:

$$y = \mathcal{P}_\Gamma(A^*x^*), \tag{1}$$

where A^* is an unknown, ground truth dictionary; x^* is drawn from a distribution $\mathcal{D}$ specified below; and $\mathcal{P}_\Gamma$ is a uniform sampling operator that retains entries in $\Gamma \subset [n]$ and zeroes out everything else. We emphasize that Γ is independently chosen for each $y^{(i)}$, so more precisely, $y^{(i)} = y^{(i)}_{\Gamma^{(i)}} \in \mathbb{R}^n$. We also make the following assumptions.

Assumption 1. *A^* has size $m \leq Kn$ for a constant $K > 0$, is of rank $r < \min(m, n)$ and democratic with parameter μ. All columns of A^* have unit norms.*

Assumption 2. *A^* has bounded spectral and max norms: $\|A^*\| \leq O(\max(1, \sqrt{m/n}))$, $\|A^*\|_{\max} \leq O(1/\sqrt{n})$.*

Assumption 3. *The code vector x^* is k-sparse random with uniform support S. We assume that the sparsity $k \leq O(\rho\sqrt{m})$ (with ρ defined below) and that the nonzero entries of x^* are pairwise independent sub-Gaussian with variance 1, and bounded below by some known constant C.*

Assumption 4. *Each entry of the sample A^*x^* is independently observed with constant probability $\rho \in (0, 1]$.*

Assumption 5. *Given a set of full samples $Z = A^*X$ whose SVD is $Z = \sum_{i=1}^r \sigma_i u_i v_i^T$, assume that u_i and v_j for $i, j \in [r]$ obey:*

$$\|u_i\|_\infty \leq O(n^{-1/2}), \quad \|v_j\|_\infty \leq O(n^{-1/2}).$$

The first four assumptions should be familiar. The intuition behind Assumption 5 is that Y has entries with magnitude bounded by $O(1/\sqrt{n})$ with high probability. Specifically, we have $Z_{ij} = A^{*\top}_{i\bullet} x^{(j)}$ for the i^{th}-row of A^* and the j^{th}-column of X. According to our aforementioned generative model, the $(i, j) - \text{th}$ entry of Y has mean $\mathbb{E}[Z_{ij}] = 0$ and variance:

$$var(Z_{ij}) = \mathbb{E}[(A^{*\top}_{i\bullet} x^{(j)})^2] = O(k/m)\|A^*_{i\bullet}\|^2 \leq O(k/n).$$

Since each entry of A^* is bounded by $O(1/\sqrt{n})$ (Assumption 2) and A^* has m columns, then with high probability, $|Z_{ij}| \leq$

Algorithm 1 Low-rank Dictionary Learning Algorithm

Input: Y – matrix of p samples with missing entries
Randomly pick $\widetilde{O}(m/\rho)$ samples, denote them as Y_1
Denote the remaining samples as Y_2
$\widehat{Y}_1 = \texttt{EXACT-COMPLETION}(Y_1)$
$A^0 = \texttt{SPECTRAL-INIT}(\widehat{Y}_1, Y_2, \rho, k, m)$
$A = \texttt{DESCENT-ALTMIN}(Y, A^0, \rho, \Theta(m/\rho k))$
Output: A – the recovered dictionary

$O(k/n) \leq O(n^{-1/4})$ because $k \leq O(\sqrt{n})$. Since entries Z are spread out, its singular vectors are expected to be spread out. This assumption allows us to perform matrix completion techniques on a partially sampled Z.

III. Algorithm

In this section, we introduce an algorithm that recovers the ground-truth dictionary A^* from partially observed samples; see Algorithm 1. At a high level, our algorithm mimics the ideas in [6] for learning dictionaries from incomplete samples. The algorithm there involves a spectral initialization (Subroutine 2) followed by a descent-style alternating minimization (Subroutine 3). Given a (coarse) initial estimate A^0, the descent stage attempts to refine it to a much greater accuracy by taking into account only partial samples.

Let us first provide some intuition behind the spectral initialization and how we adapt it to the incomplete setting. In essence, the idea is to design a re-weighted covariance matrix whose eigen-spectrum reveals one of the dictionary atoms. The whole procedure is performed iteratively until all m atoms are estimated. The re-weighting are technically based on pairwise correlations between the samples with two fixed samples (says, u, v) from an independent hold-out sampling set. For samples with missing entries, however, such pairwise correlations are not useful because each vector is sparsely observed. Our previous approach [6] overcomes this problem by assuming access to a hold-out sample set that is fully observed. As a result, the spectral initialization with the re-weighting scheme provably succeeds provided a *democratic* ground truth A^* and *sufficiently sparse* codes for the samples.

However, the success of the initialization procedure entirely depends on a *hold-out*, *fully-observed* sample set of size $O(m\,\text{polylog}(n))$. If such a hold-out set is not available, a natural solution is to *estimate* the hold-out set from the available, incompletely observed samples. Our approach aims to circumvent this requirement by *approximately reconstructing* the hold-out set from partial samples alone, borrowing any exact matrix completion algorithm (call it EXACT-COMPLETION). This succeeds provided the underlying ground truth A^* is low-rank, as asserted in Assumption 1.

Our approach fundamentally differs from standard matrix completion, since we only need to complete $O(m\ \text{polylog}(n))$ samples. Consequently, there are computational benefits in performing matrix completion on this smaller set rather than the whole set of samples. Our analysis below suggests that the difference in sample set sizes is $O_\rho(k)$, which can be as high as $n^{\frac{1}{2}-\delta}$ and hence can constitute a significant benefit.

Subroutine 2 $\texttt{SPECTRAL-INIT}(\widehat{Y}_1, Y_2, \rho, k, m)$

Input: $\widehat{Y}_1$ – p_1 approximated full samples (hold-out set)
Y_2 – p_2 samples with missing entries
Set $L = \emptyset$
while $|L| < m$ **do**
  Pick u and v from $\mathcal{P}_1$ at random
  Construct the weighted covariance matrix $\widehat{M}_{u,v}$ using samples $y^{(i)}$ from $\mathcal{P}_2$

$$\widehat{M}_{u,v} \leftarrow \frac{1}{p_2\rho^4}\sum_{i=1}^{p_2}\langle y^{(i)}, u\rangle\langle y^{(i)}, v\rangle y^{(i)}(y^{(i)})^\top$$

  $\delta_1, \delta_2 \leftarrow$ top singular values
  if $\delta_1 \geq \Omega(k/m)$ *and* $\delta_2 < O^*(k/m\log n)$ **then**
    $z \leftarrow$ top singular vector
    if z *is not within distance* $1/\log n$ *of vectors in* L *even with sign flip* **then**
      $L \leftarrow L \cup \{z\}$
    end
  end
end
Output: $A^0 \leftarrow \text{Proj}_{\mathcal{B}}(\tilde{A})$ where $\tilde{A}$ is the matrix whose columns in L and $\mathcal{B} = \{A : \|A\| \leq 2\|A^*\|\}$

Subroutine 3 $\texttt{DESCENT-ALTMIN}(Y, A^0, \rho, \eta)$

Input: Y – p samples with observed entry set $\Gamma^{(i)}$
Initial A^0 that is $(\delta, 2)$-near to A^*
Parameters ρ, η
for $s = 0, 1, \ldots, T$ **do**
  /* Encoding step */
  for $i = 1, 2, \ldots, p$ **do**
    $x^{(i)} \leftarrow \text{threshold}_{C/2}(\frac{1}{\rho}(A^s)^\top y^{(i)})$
  end
  /* Update step */
  $\widehat{g}^s \leftarrow \frac{1}{p}\sum_{i=1}^{p}(\mathcal{P}_{\Gamma^{(i)}}(A^s x^{(i)}) - y^{(i)})\text{sgn}(x^{(i)})^\top$
  $A^{s+1} \leftarrow A^s - \eta\widehat{g}^s$
end
Output: $A \leftarrow A^{(T)}$ as a learned dictionary

IV. Analysis

The remainder of the paper analyzes the above algorithm. Our main theoretical result is stated in Theorem 1, which follows from an appropriate concatenation of the main results of [10], [11], and [6].

Theorem 1. *Suppose* $\mu = O^*(\frac{\sqrt{n}}{k\log^3 n})$, $\frac{1}{\rho} - 1 \leq k \leq O^*(\frac{\rho\sqrt{n}}{\log n})$. *When* $p = \widetilde{O}(\max(m,n)k/\rho^4)$, *then with high probability, Algorithm 1 recovers* A^* *within column-wise* $O(\sqrt{k/n})$ *error. The total running time is* $\widetilde{O}(\rho mn^2 p)$.

In the next lemma, we show that we can construct the hold-out set of size at least $m\text{polylog}(n)$. The best available matrix completion result requires $nr\text{polylog}(n)$ observed entries, which suggests the hold-out set must be at least $r\text{polylog}(n)/\rho$ partial columns. When $r = O(m)$, we require $1/\rho$ factor more partial samples than is necessary to be able to construct the hold-out sampling set.

Lemma 1 (Theorem 1.2, [10]). *Given $p_1 = m\text{polylog}(n)/\rho$ partial samples $Y_1 = [y_1, y_2, \ldots, y_{p_1}] = \mathcal{P}_{\Gamma_1}(A^* X_1)$. With probability $1 - n^{-3}$, nuclear norm minimization recovers all the entries of $Z_1 = A^* X_1$ exactly. The running time is $O(\max(m,n)^3)$. We dub this algorithm as* `EXACT-MATRIX-COMPLETION`.

Proof. We prove this result by construction. The matrix Y_1 has $p_1 = m\text{polylog}(n)/\rho$ columns, then by the uniform sampling, it has $mn\text{polylog}(n)/\rho$ observed entires, which is bigger than $rn\text{polylog}(n)/\rho$ since $m > r$. By Assumption 5, the singular vectors of the original matrix Z_1 have incoherence parameter $\mu_0 = O(\log n)$ with respect to the standard basis (as defined therein). Apply Theorem 1.2, [10], we obtain the result. □

Lemma 2 (Theorem 5, [6]). *Suppose that the available training dataset consists of p_1 fully observed samples, together with p_2 incompletely observed samples according to the sparse factor model. Suppose $\mu = O^*\big(\frac{\sqrt{n}}{k \log^3 n}\big)$, $\frac{1}{\rho} - 1 \le k \le O^*(\frac{\rho\sqrt{n}}{\log n})$. When $p_1 = \widetilde{\Omega}(m)$ and $p_2 = \widetilde{\Omega}(mk/\rho^4)$, then with high probability, Subroutine 2 returns an initial estimate A^0 whose columns share the same support as A^* and is $(\delta, 2)$-near to A^* with $\delta = O^*(1/\log n)$.*

Lemma 3 (Theorem 3, [6]). *Suppose that the initial estimate A^0 is $(\delta, 2)$-near to A^* with $\delta = O^*(1/\log n)$ and the sampling probability satisfies $\rho \ge 1/(k+1)$. If Subroutine 3 is given $p = \widetilde{\Omega}(mk)$ fresh partial samples at each step and uses learning rate $\eta = \Theta(m/\rho k)$, then*

$$\mathbb{E}[\|A^s_{\bullet i} - A^*_{\bullet i}\|^2] \le (1-\tau)^s \|A^0_{\bullet i} - A^*_{\bullet i}\|^2 + O(\sqrt{k/n})$$

for some $0 < \tau < 1/2$ and $s = 1, 2, \ldots, T$. As a corollary, A^s converges geometrically to A^* until column-wise $O(\sqrt{k/n})$ error.

Provided the approximate $\widehat{Y}_1$ of the full samples Z_1, we use them as the hold-out set to perform spectral initialization and obtain a coarse estimate A^0 that is δ-close to the ground truth with closeness $\delta = O^*(1/\log n)$. In order to establish provable guarantees for learning the dictionary A^*, we use the results in Lemma 2 and Lemm 3 obtained from [6].

By way of Lemma 1, we can achieve the exact recovery of Y_1 with near optimal sample complexity at the price of running time. It it important to note that we do not need exact recovery but can tolerate error n^{-1} entrywise. Lemma 4 gives guarantee for such an error.

Let us denote $u = A^*\alpha$ and $v = A^*\alpha'$ sampled from the model without sub-sampling. Consider a sample with missing entries $y = A^*_{\Gamma\bullet} x^*$ under a random subset $\Gamma \subseteq [n]$. Suppose we are given two approximation $\widehat{u}, \widehat{v}$ such that $\widehat{u} = u + \epsilon_u$ and $\widehat{v} = v + \epsilon_v$. Then, denote

$$\beta = \frac{1}{\rho} A^{*T}_{\Gamma\bullet} \widehat{u}, \text{ and } \beta' = \frac{1}{\rho} A^{*T}_{\Gamma\bullet} \widehat{v}$$

representing coarse estimates of α and α' respectively. The following lemma establishes the quality of these estimates (coordinate-wise).

Lemma 4. *Suppose $\|\epsilon_u\| \le O(n^{-1/4})$. With high probability over the randomness in u and Γ, we have:*

(a) $|\beta_i - \alpha_i| \le \frac{\mu k \log n}{\sqrt{n}} + 2\sqrt{\frac{1}{\rho n^{1/2}}}$ for each $i = 1, 2, \ldots, m$; and

(b) $\|\beta\| \le \frac{\sqrt{k}\log n}{\rho} + \frac{1}{\rho n^{1/4}}$.

Proof. By definition of β, we have

$$|\beta_i - \alpha_i| = \Big|\frac{1}{\rho} A^{*T}_{\Gamma,i} u - \alpha_i + \frac{1}{\rho} A^{*T}_{\Gamma,i} \epsilon_u\Big| \tag{2}$$

By Lemma 2, [6], we have

$$\Big|\frac{1}{\rho} A^{*T}_{\Gamma,i} u - \alpha_i\Big| \le \frac{\mu k \log n}{\sqrt{n}} + \sqrt{\frac{1-\rho}{\rho n^{1/2}}}.$$

The latter term is bounded by

$$\Big|\frac{1}{\rho} A^{*T}_{\Gamma,i} \epsilon_u\Big| \le \frac{1}{\rho} \|A^*_{\Gamma,i}\| \|\epsilon_u\| \le \frac{1}{\sqrt{\rho}} \|\epsilon_u\| \le \sqrt{\frac{1}{\rho n^{1/2}}}$$

since $\|A^*_{\Gamma,i}\|^2 \le \rho + o(\rho)$ and $\|\epsilon_u\| \le n^{-1/2}$.

Combining these two bounds, we get

$$|\beta_i - \alpha_i| \le \frac{\mu k \log n}{\sqrt{n}} + 2\sqrt{\frac{1}{\rho n^{1/2}}},$$

w.h.p., which is the first part of the claim.

The second part is easily bounded as follows:

$$\begin{aligned} \|\beta\| &= \frac{1}{\rho} \|A^{*T}_{\Gamma\bullet}(u + \epsilon_u)\| \\ &\le \frac{1}{\rho} \|A^*_{\Gamma\bullet}\| \big(\|A^*_{\bullet U}\| \|\alpha_U\| + \|\epsilon_u\|\big), \end{aligned}$$

for $U = \text{supp}(\alpha)$. Then, using $\|\alpha_U\| \le \sqrt{k}\log n$ w.h.p., $\|A^*\| \le O(1)$ and $\|\epsilon_u\| \le O(n^{-1/4})$, then $\|\beta\| \le \sqrt{k}\log n/\rho + \frac{1}{\rho n^{1/4}}$. □

References

[1] R. Rubinstein, A. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proc. IEEE*, 98(6):1045–1057, 2010.

[2] Z. Xing, M. Zhou, A. Castrodad, G. Sapiro, and L. Carin. Dictionary learning for noisy and incomplete hyperspectral images. *SIAM J. Imaging Sciences*, 5(1):33–56, 2012.

[3] V. Naumova and K. Schnass. Dictionary learning from incomplete data. *arXiv preprint arXiv:1701.03655*, 2017.

[4] A. Soni, S. Jain, J. Haupt, and S. Gonella. Noisy matrix completion under sparse factor models. *IEEE Trans. Info. Theory*, June 2016.

[5] P. Loh and M. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *NeurIPS*, 2011.

[6] T. Nguyen, A. Soni, and C. Hegde. On learning sparsely used dictionaries from incomplete samples. In *ICML*, 2018.

[7] S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conf. Learning Theory*, 2014.

[8] S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on Learning Theory*, 2015.

[9] M. Davenport, J. Laska, P. Boufounos, and R. Baraniuk. A simple proof that random matrices are democratic. *arXiv preprint arXiv:0911.0736*, 2009.

[10] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Info. Theory*, 2010.

[11] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[12] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Info. Theory*, 56(6):2980–2998, 2010.

[13] A. Lan, A. Waters, C. Studer, and R. Baraniuk. Sparse factor analysis for learning and content analytics. *J. Machine Learning Research*, 15(1):1959–2008, 2014.