Robust Identifiability in Sparse Dictionary Learning

Charles J. Garfinkle, Christopher J. Hillar

Abstract

We study uniqueness in sparse dictionary learning when reconstruction of data is approximate.

Index Terms

bilinear inverse problem, identifiability, dictionary learning, sparse coding, matrix factorization, compressed sensing, combinatorial matrix theory, blind source separation

I. INTRODUCTION

ONE of the fundamental questions in data analysis is how to represent the data in a way that reveals structure. Recently, algorithms have been developed for uncovering *sparse* structure in a given dataset $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^n$ by approximately solving the constrained optimization problem:

$$\min_{A, \mathbf{a}_i} \sum_{i=1}^{N} \|\mathbf{y}_i - A\mathbf{a}_i\|_2 \quad \text{subject to } \|\mathbf{a}_i\|_0 \leq k \tag{1}$$

for unknown $A \in \mathbb{R}^{n \times m}$ and $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ with $k < m$. (It almost goes without saying that $m < N$ as well, lest the solution be completely trivial.) Known variously as dictionary learning, sparse coding, or sparse matrix factorization, this framework emphasizes parsimony in representation by approximating each $\mathbf{y}_i$ as some linear combination of at most k columns of some matrix A, typically with $k \ll m$.

Approximating the optimal solution to (1) has been shown in general to be NP-hard [**?**] **[Razaviyayn 2015, but they did it within an additive error. Tollman 2015 did it for multiplicative error for the dual problem. Should we say this isn't necessarily a big deal since there may be some *particular* instances where NP-hard problems are solvable in poly time?**}. Nonetheless, it is natural to wonder when the solution to (1) is in some sense unique, or at least approximately so. This is a question of fundamental concern in *sparse component analysis* (SCA) [1], wherein one assumes the linear model:

$$\mathbf{y}_i = A\mathbf{a}_i + \eta_i \tag{2}$$

with the vector $\eta_i \in \mathbb{R}^n$ accounting for both "noise" in the data acquisition process and the degree to which the model deviates from reality (e.g. the $\mathbf{a}_i$ may be known to be only approximately k-sparse). The goal in SCA is to recover (approximately) the sparse *sources* $\mathbf{a}_i$ from the observed *mixtures* $\mathbf{y}_i$ by approximately solving (1). **How do sparsity of $\mathbf{a}_i$ and upper bound on $\|\eta_i\|$ tradeoff to make (1) solve (2)?**

Letting Y and X be the matrices with columns $\mathbf{y}_i$ and $\mathbf{a}_i$, respectively, the optimization (1) is seen to minimize the matrix norm $\|Y - AX\|_{2,1}$ subject to sparsity constraints on the columns of X. From this matrix factorization perspective it is clear that SCA is a special case of bilinear inverse problem and thus has an associated *ambiguity transform group*. Clearly, if AX approximates Y then so does $(AD^{-1}P^{-1})(PDX)$ for any permutation matrix $P \in \mathbb{R}^{m \times m}$ and invertible diagonal matrix $D \in \mathbb{R}^{m \times m}$ and moreover this transformation preserves the sparsity of the columns of X. The SCA problem can thus at best (even when $\eta_i = 0$) be solved up to this inherent permutation and scaling ambiguity, which motivates the following definition and problem.

Definition 1. *We say a dataset $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ has a unique k-sparse representation in m dimensions if for some $A \in \mathbb{R}^{n \times m}$ there exist $C, \varepsilon > 0$ such that:*

1) *For some k-sparse $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ we have $\|\mathbf{y}_i - A\mathbf{a}_i\|_2 \leq \varepsilon$ for all $i = 1, \ldots, N$.*
2) *Every alternate $B \in \mathbb{R}^{n \times m}$ and k-sparse $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^m$ for which $\|\mathbf{y}_i - B\mathbf{b}_i\|_2 \leq \varepsilon$ for all $i = 1, \ldots, N$ necessarily satisfy $\|A - BPD\|_1 \leq C\varepsilon$ and $\|\mathbf{b}_i - PD\mathbf{a}_i\|_1 \leq C\varepsilon$ for some permutation matrix $P \in \mathbb{R}^{m \times m}$ and invertible diagonal matrix $D \in \mathbb{R}^{m \times m}$.*

Definition 2. *We say a dataset $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ has a unique k-sparse representation in m dimensions when:*

1) *There exists some $A \in \mathbb{R}^{n \times m}$ and $\varepsilon > 0$ such that for every $\mathbf{y} \in Y$, the ε-ball $\mathcal{B}_2(\mathbf{y}, \varepsilon)$ centered at $\mathbf{y}$ intersects exactly one subspace spanned by k columns of A.*

Dataset $Y = \{A\mathbf{a}_1 + \eta_1, \ldots, A\mathbf{a}_N + \eta_N,\}$ with k-sparse $\mathbf{a}_i \in \mathbb{R}^m$ chosen:	Total sufficient samples N for a unique recovery
in general position with $k\binom{m}{k}$ from each support set in $\mathcal{T}$, independent of A satisfying spark condition (3)	$mk\binom{m}{k}$
$k+1$ chosen "randomly" from each support set in $\mathcal{T}$, given matrix A satisfying spark condition (3)	$m(k+1)$ (with probability one)
$k+1$ chosen "randomly" from each support set in $\mathcal{T}$, independent of $A \notin Z$	$m(k+1)$ (with probability one)
"randomly" from each support set in $\mathcal{T}$, given matrix A satisfying spark condition (3)	$m(k+1)\beta^{-1}$ (with probability $1-\beta$)

TABLE I
THE NUMBER OF SUBSAMPLES SUFFICIENT FOR $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ TO HAVE A UNIQUE SPARSE CODING.

2) *Any matrix $B \in \mathbb{R}^{n \times m}$ for which every such ball $\mathcal{B}_2(\mathbf{y}, \varepsilon)$ intersects a subspace spanned by k columns of B necessarily satisfies $\|A - BPD\|_1 \leq C\varepsilon$ for some permutation matrix $P \in \mathbb{R}^{m \times m}$, invertible diagonal matrix $D \in \mathbb{R}^{m \times m}$ and $C > 0$.*

We will show later that this approximation of A by BPD implies the approximate recovery of original sources in the SCA problem (2).

Problem 1. *Let $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ be generated as $\mathbf{y}_i = A\mathbf{a}_i + \eta_i$ for some matrix $A \in \mathbb{R}^{n \times m}$ and k-sparse $\mathbf{a}_i \in \mathbb{R}^m$ with $\|\eta_i\|_2 \leq \varepsilon$. When does Y have a unique k-sparse representation in m dimensions?*

Conditions for the *identifiability* of the parameters underlying SCA were first investigated for the case $\mathbf{y}_i = A\mathbf{a}_i$ by Georgiev et. al. [1] and subsequently by Aharon et. al. [2]. The most general conditions to date have recently been proven by Hillar and Sommer [cite], who provide deterministic and probabilistic guarantees for the identifiability of the model parameters contained in A and the original sparse codes $\mathbf{a}_i$. We refine their proofs to take into account the possibility of noise contaminating the measurements (or inadequacy of the model) and significantly reduce their theoretically required number of samples. These *robust* identifiability conditions describe when the true model parameters can be uniquely determined in the sense of Definition 2. Moreover, we provide guarantees for unique recovery of A and the $\mathbf{a}_i$ when only an upper bound on the dimensionality of the sources is known. Our conditions are derived from the underlying geometry of the dictionary learning problem, hence they apply regardless of which algorithm is used to solve (1). Table I gives a summary of our results.

Before stating our first theorem, it is instructive to first consider the simpler case where we wish to identify the sources $\mathbf{a}_i$ from the noiseless mixture $\mathbf{y}_i = A\mathbf{a}_i$ when A is already known. Evidently, A must satisfy the following *spark condition*

$$A\mathbf{a}_1 = A\mathbf{a}_2 \implies \mathbf{a}_1 = \mathbf{a}_2 \quad \text{for all } k\text{-sparse } \mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^m \tag{3}$$

for such recovery to be possible. This condition ensures that distinct sparse sources are distinguishable in their measurements. As it turns out, this is also enough to guarantee uniqueness in the case where we don't know A a priori, as our main result states:

Theorem 1. *Given positive integers n, m and $k < m$, there exist $N = mk\binom{m}{k}$ k-sparse vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ with the following property: every matrix $A \in \mathbb{R}^{n \times m}$ satisfying spark condition (3) has associated with it some $\varepsilon > 0$ such that every dataset $Y = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ for which $\|\mathbf{y}_i - A\mathbf{a}_i\|_2 \leq \varepsilon$ for all $i = 1, \ldots, N$ has a unique k-sparse coding in m dimensions.*

In fact, there are many such sets of deterministically produced $\mathbf{a}_i$; we give a parametrized family in Section II. Note that if we restrict our attention to noiseless datasets of the form $\mathbf{y}_i = A\mathbf{a}_i$ then we recover Theorem 1 in [HS11] with an added improvement on N, the number of samples required to guarantee uniqueness. In fact, all of the theorems and corollaries in [HS11] are corollaries of our theorems. Our proofs are generalizations of the basic framework laid out in their paper.

The particular value of ε_0 associated to each A satisfying (3) and the values of C and C' alluded to in Definition 2 can be explicitly calculated from certain properties of A and the $\mathbf{a}_i$ which we now define. A simple compactness argument can be made to show that any matrix satisfying (3) must also satisfy

$$\|A\mathbf{a}_1 - A\mathbf{a}_2\|_2 \geq \alpha \|\mathbf{a}_1 - \mathbf{a}_2\|_2 \quad \text{for all } k\text{-sparse } \mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^m \tag{4}$$

for some $\alpha \in (0, 1]$. We can equally well say that $\|A\mathbf{x}\| \geq \alpha\|\mathbf{x}\|$ for all $2k$-sparse $\mathbf{x} \in \mathbb{R}^{n \times m}$, or every $n \times \min(m, 2k)$ submatrix of A has lower bound α. Since A is a real matrix, it also has bounded norm and hence (4) is equivalent to the *restricted isometry property* [3] familiar from work in the field of compressed sensing.

It can be shown (see the Appendix) that (3) implies the existence of a nonzero *Friedrichs angle* $\theta_F \in (0, \frac{\pi}{2}]$ (also known as the minimal Jordan angle) between all pairs of subspaces spanned by the columns of A of dimensionality less than or equal to k. For completeness we state the definition of θ_F here:

Definition 3. *The* Friedrichs *angle $\theta_F(V, W) \in [0, \frac{\pi}{2}]$ between subspaces $V, W \subseteq \mathbb{R}^m$ is the minimal angle formed between unit vectors in $V \cap (V \cap W)^\perp$ and $W \cap (W \cap V)^\perp$, that is*

$$\cos[\theta_F(V, W)] := \max\left\{ \frac{\langle v, w \rangle}{\|v\|\|w\|} : v \in V \cap (V \cap W)^\perp, w \in W \cap (V \cap W)^\perp \right\} \tag{5}$$

We define the following quantity for notational convenience:

Definition 4. *Given $A \in \mathbb{R}^{n \times m}$ and $k < m$, let $\mathcal{T}$ be the set of all $S_i := \{i, \ldots, (i + (k-1)\} \mod m$ for $i = 0, \ldots, m-1$. We define the quantity*

$$\phi_k(A) := \min_{i_1 \neq \ldots \neq i_\ell \in [m]} \left\{ 1 - c(Span\{A_{S_{i_1}}\}, \ldots, Span\{A_{S_{i_\ell}}\}) \right\}, \tag{6}$$

where for any sequence $V_1, \ldots, V_p$ of closed subspaces of $\mathbb{R}^m$,

$$c(V_1, \ldots, V_p) := \left[1 - \prod_{i=1}^{p-1} \left(1 - \cos^2 \left[\theta_F(V_i, \cap_{j=i+1}^{p} V_j) \right] \right) \right]^{1/2} \tag{7}$$

and we set $\phi_1(A) := 1$.

We now state the specifics of Theorem 1 in terms of these quantities.

Remark 1 (Theorem 1 cont'd). *Specifically, the $\mathbf{a}_i$ are constructed so that any k of them are linearly independent. In particular, any k vectors $\mathbf{a}_i$ sharing the same support satisfy*

$$\| \sum_{j=1}^{k} c_j \mathbf{a}_{i_j} \|_2 \geq \sigma \|c\|_2 \quad \text{for all } c = (c_1, \ldots, c_k) \in \mathbb{R}^k. \tag{8}$$

for some $\sigma > 0$. If A has unit norm columns and satisfies spark condition (3) with all $n \times k$ and $n \times \min(m, 2k)$ submatrices of A having lower bounds α_k and α_{2k}, respectively, then for any matrix $B \in \mathbb{R}^{n \times m}$ and k-sparse $\mathbf{b}_i \in \mathbb{R}^m$ for which $\|\mathbf{y}_i - B\mathbf{b}_i\| \leq \varepsilon$ with $\varepsilon < \frac{\alpha^2 \sigma \phi_k(A)}{(2k)^{2/3}}$ we have

$$\|A - BPD\|_1 \leq \frac{2\varepsilon k^{2/3} \sqrt{m}}{\alpha_k \sigma \phi_k(A)} \tag{9}$$

for some permutation matrix $P \in \mathbb{R}^{m \times m}$ and invertible diagonal matrix $D \in \mathbb{R}^{m \times m}$. Moreover,

$$\|\mathbf{b}_i - PD\mathbf{a}_i\|_1 \leq \frac{\varepsilon}{\alpha_{2k} \|D\|_2} \left(1 + \frac{2k^{2/3}}{\alpha_k \sigma \phi_k(A)} \|D^{-1} P^{-1} \mathbf{b}_i\|_1 \right). \tag{10}$$

Our next result exploits randomness to give a different construction requiring fewer samples. Fix a matrix A satisfying spark condition (3) and a small $\beta > 0$. There is a simple random drawing procedure (Definition 6) for generating $N = \beta^{-1} m(k+1)$ k-sparse vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N$ such that $Y = \{A\mathbf{a}_1 + \eta_1, \ldots, A\mathbf{a}_N + \eta_N\}$ has a unique sparse coding (with probability at least $1 - \beta$), provided $\|\eta_i\|$ is bounded by a small enough constant. In particular, sparse matrix factorization of sparsely coded data is typically unique for a number of samples linear in the dimensionality and sparsity level of the sparse codes, provided their supports are distributed enough over those contained in a special support set (specifically, $\mathcal{T}$ from Definition 4).

A subtle difference between these two results is that $\mathbf{a}_1, \ldots, \mathbf{a}_N$ in Theorem 1 are independent of the generation matrix A. The following statement is the best we are able to do in this direction using random methods. With probability one, a random draw of $(k+1)$ k-sparse $\mathbf{a}_i$ from each of the m support sets in $\mathcal{T}$ satisfies the following property. There is a set of matrices $Z \in \mathbb{R}^{n \times m}$ with Lebesgue measure zero such that if $A \notin Z$, then there exists some ε_0 small enough such that $Y = \{A\mathbf{a}_1 + \eta_1, \ldots, A\mathbf{a}_N + \eta_N\}$ for $\|\eta_i\| \leq \varepsilon_0$ has a unique sparse coding.

It is important for us to note how these theorems of ours fit in with others comprising the field of *theoretical dictionary learning*. While a collection of greedy optimization algorithms have been proposed for solving (1), as yet there exist no global convergence guarantees for these techniques. Local identifiability analyses have demonstrated that the global minimum is a local minimum of cost functions of L0 and L1 cost functions, (local) convergence guarantees of greedy and convex algorithms. A most welcome corollary to our theorems is a closed form expression for checking when any dictionary learning algorithm has converged to a global minimum of (1) by asserting whether or not the reconstruction error of the proposed solution is indeed small enough.

It is also informative to elaborate on the relationship of our results to the field of compressive sensing. [ref?] CS theory provides conditions under which it is possible to recover data vectors $\mathbf{x}$ with sparse structure after they have been linearly subsampled as $\mathbf{y} = \Phi \mathbf{a}$ by a known compression matrix Φ. The sparsity usually enforced is that the vectors $\mathbf{x}$ can be expressed as $\mathbf{x} = \Psi \mathbf{a}$ using a known dictionary matrix Ψ and m-dimensional vectors $\mathbf{a}$ with at most $k < m$ nonzero entries. Such vectors $\mathbf{a}$ are called k-*sparse*. A necessary condition for the unique recovery of $\mathbf{a}$ given $\mathbf{y}$ is that the generation matrix $A = \Phi \Psi$ satisfy the spark condition (3). Otherwise, different sparse sources would be indistinguishable in the compressed space. Provided the dimension n of $\mathbf{y}$ satisfies

$$n \geq Ck \log \left(\frac{m}{k} \right), \tag{11}$$

the theory guarantees that with high probability a randomly generated Φ will yield an A satisfying (3). In contrast, the goal of SCA is to recover both the code vectors *and* the generation matrix from measurements. We show that the same uniqueness conditions required by CS also guarantee uniqueness in SCA given enough samples distributed over suitable supports.

The organization of this paper is as follows.in Section II we state our main tool from combinatorial matrix theory and then derive from it Theorem 1. We next provide precise statements in Section C-B of our randomized theorems and then prove them in Section C-C. In Section (??) we describe how to extend the theorems in the previous sections to the case where we have only an upper bound on the number of sparse components. The final section is a discussion. An appendix contains the proof of Lemma 1 as well as some auxiliary proofs.

Have the cool small dimensional example here. Data in $3D$-space coded by planes. $n = 3, k = 2$.

Implication: overcomplete dimensionality of any dataset with respect to ε. If you coded very well some data sparsely then you can do a check and if that check holds up then anyone else who codes the data sparsely too has learned the same dictionary.

II. DETERMINISTIC UNIQUENESS THEOREM

In what follows, we will use the notation $[m]$ for the set $\{1, ..., m\}$, and $\binom{[m]}{k}$ for the set of subsets of $[m]$ of cardinality k. Also, recall that $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ for real vectors $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$ is the vector space consisting of their $\mathbb{R}$-linear span:

$$\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\} = \left\{ \sum_{i=1}^{\ell} t_i \mathbf{v}_i : t_1, \ldots, t_\ell \in \mathbb{R} \right\}.$$

Finally, for a subset $S \subseteq [m]$ and matrix A with columns $\{A_1, ..., A_m\}$ we define

$$\text{Span}\{A_S\} = \text{Span}\{A_s : s \in S\}.$$

Before proving Theorem 1 in full generality it is illustrative to consider the case when $k = 1$. In this simple case, we only need $N = m$ samples to guarantee uniqueness. Set $\mathbf{a}_i = \mathbf{e}_i$ $(i = 1, \ldots, m)$ to be the standard basis vectors in $\mathbb{R}^m$ and suppose that our data $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ satisfies $\|\mathbf{y}_i - A\mathbf{a}_i\|_2 \leq \varepsilon$ for all $i \in [m]$ for some $A \in \mathbb{R}^{n \times m}$ with unit norm columns satisfying (4) and that $\varepsilon < \frac{\alpha}{2\sqrt{2}}$. Assuming that $\|\mathbf{y}_i - B\mathbf{b}_i\|_2 \leq \varepsilon$ for all $i \in [m]$ for some alternate $B \in \mathbb{R}^{n \times m}$ and 1-sparse $\mathbf{b}_i \in \mathbb{R}^m$, then by the triangle inequality we have $\|A\mathbf{a}_i - B\mathbf{b}_i\| \leq 2\varepsilon$ for all $i \in [m]$. Since the $\mathbf{b}_i$ are 1-sparse, there must exist $c_1, \ldots, c_m \in \mathbb{R}$ such that

$$\|A\mathbf{e}_i - c_i B\mathbf{e}_{\pi(i)}\|_2 \leq 2\varepsilon \quad \text{for all} \quad i \in [m]. \tag{12}$$

for some map $\pi : [m] \to [m]$. Note that if $c_i = 0$ for some i, then $\|A\mathbf{e}_i\| < 1$ (since $\alpha \leq 1$), contradicting the fact that A has unit norm columns. We will now show that π is necessarily injective (and thus defines a permutation). Suppose that $\pi(i) = \pi(j) = \ell$ for some $i \neq j$ and $\ell \in [m]$. Then $\|A\mathbf{e}_i - c_i B\mathbf{e}_\ell\|_2 \leq 2\varepsilon$ and $\|A\mathbf{e}_j - c_j B\mathbf{e}_\ell\|_2 \leq 2\varepsilon$. Summing and scaling these two inequalities by $|c_j|$ and $|c_i|$, respectively, we apply the triangle inequality and then (4) to yield

$$\begin{aligned} (|c_i| + |c_j|)2\varepsilon &\geq |c_j| \|A\mathbf{e}_i - c_i B\mathbf{e}_\ell\|_2 + |c_i| \|c_j B\mathbf{e}_\ell - A\mathbf{e}_j\|_2 \\ &\geq \|c_j A e_i + c_i A e_j\|_2 \\ &\geq \alpha \|c_j e_i + c_i e_j\|_2 \end{aligned}$$

which is in contradiction with the fact that $\|x\|_1 \leq \sqrt{2}\|x\|_2$ for all $x \in \mathbb{R}^2$ and $\varepsilon < \frac{\alpha}{2\sqrt{2}}$. Hence, π is injective. Letting P and D denote the following permutation and diagonal matrices, respectively:

$$P = \left(\mathbf{e}_{\pi(1)} \cdots \mathbf{e}_{\pi(m)} \right), \quad D = \begin{pmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_m \end{pmatrix}, \tag{13}$$

we see that (12) becomes $\|(A - BPD)\mathbf{e}_i\|_2 \leq 2\varepsilon$ for all $i \in [m]$ and it follows that $\|A - BPD\|_1 \leq 2\sqrt{m}\varepsilon$.

Note that the identity $\|(A - BPD)\mathbf{e}_i\|_2 \leq C\varepsilon$ for some $C > 0$ together with (4) already implies in general the recovery result $\|\mathbf{b}_i - PD\mathbf{a}_i\|_1 \leq C'\varepsilon$ for $C' > 0$ for all $i \in [m]$. This follows since for any k-sparse vector $\mathbf{a}_i = \sum_{\ell=1}^{k} c_\ell \mathbf{e}_{i_\ell}$ we have by the triangle inequality that $\|(A - BPD)v\| \leq \sum_{\ell=1}^{k} |c_\ell| \|(A - BPD)\mathbf{e}_{i_\ell}\|$, hence

$$\|(A - BPD)D^{-1}P^{-1}\mathbf{b}_i\| \leq \|D^{-1}P^{-1}\mathbf{b}_i\|_1 C\varepsilon.$$

It then follows by the triangle inequality and (4) that

$$\begin{aligned} \alpha \|D^{-1}P^{-1}\mathbf{b}_i - \mathbf{a}_i\| &\leq \|A(D^{-1}P^{-1}\mathbf{b}_i - \mathbf{a}_i)\| \\ &\leq \|AD^{-1}P^{-1}\mathbf{b}_i - B\mathbf{b}_i\| + \|B\mathbf{b}_i - \mathbf{y}_i\| + \|\mathbf{y}_i - A\mathbf{a}_i\| \\ &\leq \varepsilon \left(2 + C\|D^{-1}P^{-1}\mathbf{b}_i\|_1 \right). \end{aligned}$$

Hence,

$$\|\mathbf{b}_i - PD\mathbf{a}_i\| \leq \frac{\varepsilon}{\alpha \|D\|_2} \left(2 + C\|D^{-1}P^{-1}\mathbf{b}_i\|_1 \right) \tag{14}$$

Unfortunately, the proof for larger k is more challenging since in general it is nontrivial to produce P and D as in (39). For this, we require an additional result in combinatorial linear algebra. We make one more definition before stating the relevant lemma.

Definition 5. *Let V, W be subspaces of $\mathbb{R}^m$ and let $d(v, W) := \inf\{\|v - w\|_2 : w \in W\} = \|v - \Pi_W v\|$ where Π_W is the orthogonal projection operator onto subspace W. The* gap *metric Θ on subspaces of $\mathbb{R}^m$ is [see Theory of Linear Operators in a Hilbert Space p. 69 who cites first reference]*

$$\Theta(V, W) := \max\left(\sup_{\substack{v \in V \\ \|v\| = 1}} d(v, W), \sup_{\substack{w \in W \\ \|w\| = 1}} d(w, V) \right). \tag{15}$$

In our proof we will make use of the following useful facts:

$$\dim(W) = \dim(V) \implies \sup_{\substack{v \in V \\ \|v\| = 1}} d(v, W) = \sup_{\substack{w \in W \\ \|w\| = 1}} d(w, V), \tag{16}$$

a proof of which can be found in [4] (Lemma 3.3) and

$$d(v, W) < \|v\|_2 \quad \forall v \in V \implies \dim(V) \leq \dim(W). \tag{17}$$

A quick proof of (17) is as follows. If $\dim(V) > \dim(W)$ then there exists some $v' \in V \cap W^\perp$ and by Pythagoras' Theorem, $\|v' - w\|_2^2 = \|v'\|_2^2 + \|w\|_2^2 \geq \|v\|_2^2$ for all $w \in W$. This is in contradiction with the LHS of (17), which states that for all $v \in V$ there exists some $w \in W$ such that $\|v - w\|_2 < \|v\|_2$.

We now state our main result from combinatorial matrix theory.

Lemma 1 (Main Lemma). *Fix positive integers n, m and $k < m$. Let $A, B \in \mathbb{R}^{n \times m}$ and suppose that A has unit norm columns, satisfies spark condition (3), and every $n \times \max(2, k)$ submatrix of A has lower bound $\alpha \in (0, 1]$. Let $\mathcal{T}$ be the set of all $S_i := \{i, \ldots, (i + (k-1)\} \mod m$ for $i = 0, \ldots, m - 1$. If there exists a map $\pi : \mathcal{T} \to \binom{[m]}{k}$ and some $\Delta < \frac{\alpha}{\sqrt{2}}$ such that*

$$\Theta(Span\{A_{S_i}\}, Span\{B_{\pi(S_i)}\}) \leq \frac{\phi_k(A)}{k} \Delta \quad for\ all \quad i \in \{0, \ldots, m - 1\} \tag{18}$$

then there exist a permutation matrix $P \in \mathbb{R}^{m \times m}$ and a diagonal matrix $D \in \mathbb{R}^{m \times m}$ such that

$$\|A - BPD\|_1 \leq \sqrt{m}\Delta \quad for\ all \quad i \in [m]. \tag{19}$$

[Is there a better norm to use here?]

We defer the proof of this lemma to the Appendix.

Proof of Theorem 1. First, we produce a set of $N = mk\binom{m}{k}$ vectors in $\mathbb{R}^k$ in general linear position (i.e. any set of k of them are linearly independent). Specifically, let $\gamma_1, \ldots, \gamma_N$ be any distinct numbers. Then the columns of the $k \times N$ matrix $V = (\sigma_j^i)_{i,j=1}^{k,N}$ are in general linear position (since the σ_j are distinct, any $k \times k$ "Vandermonde" sub-determinant is nonzero). Next, form the k-sparse vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ with supports $S_j = \{j, \ldots, (j + k - 1) \mod m\}$ for $j \in [m]$ (partitioning the a_i evenly among these supports, i.e. for each support S_j there are $k\binom{m}{k}$ vectors a_i with that support) by setting the nonzero values of vector $\mathbf{a}_i$ to be those contained in the ith column of V. Note that by construction every k vectors a_i are linearly independent.

We will show how the existence of these $\mathbf{a}_i$ proves the theorem. First, we will show that they satisfy property (8). Consider the compact set $\mathcal{C} = \{c \in \mathbb{R}^k : \|c\|_2 = 1\}$ and the continuous map

$$\phi : \mathcal{C} \to \mathbb{R}$$

$$(c_1, \ldots, c_k) \mapsto \|\sum_{j=1}^k c_j \mathbf{a}_{i_j}\|_2.$$

By general linear position of the $\mathbf{a}_i$, we know that $0 \notin \phi(\mathcal{C})$. Since $\mathcal{C}$ is compact, we have by continuity of ϕ that $\phi(\mathcal{C})$ is also compact; hence it is closed and bounded. Therefore 0 can't be a limit point of $\phi(\mathcal{C})$ and there must be some $\sigma > 0$ such that the neighbourhood $\{x : x < \sigma\} \subseteq \mathbb{R} \setminus \phi(\mathcal{C})$. Hence $\phi(c) \geq \sigma$ for all $c \in \mathcal{C}$. The property (8) follows by the association $c \mapsto \frac{c}{\|c\|_2}$ and the fact that there are only finitely many subsets of k vectors $\mathbf{a}_i$ (actually, for our purposes we need only consider those subsets of k vectors $\mathbf{a}_i$ having the same support), hence there is some minimal δ satisfying (8) for all of them. (We refer the reader to the Appendix for a lower bound on σ given as a function of k and the sequence $\gamma_1, \ldots, \gamma_N$ used to generate the a_i.)

Now suppose that $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ is a dataset for which for all $i \in \{1, \ldots, N\}$ we have $\|\mathbf{y}_i - A\mathbf{a}_i\| \leq \varepsilon$ for some $\varepsilon > 0$ and $A \in \mathbb{R}^{n \times m}$ with all $n \times k$ submatrices of A having lower bound α. Suppose further that for some alternate $B \in \mathbb{R}^{n \times m}$

there exist k-sparse $\mathbf{b}_i \in \mathbb{R}^m$ for which $\|\mathbf{y}_i - B\mathbf{b}_i\| \leq \varepsilon$ for all $i \in \{1, \ldots, N\}$. Since there are $k\binom{m}{k}$ vectors $\mathbf{a}_i$ with a given support S, the pigeon-hole principle implies that there are at least k vectors $\mathbf{y}_i$ such that $\|\mathbf{y}_i - A\mathbf{a}_i\| \leq \varepsilon$ for these $\mathbf{a}_i$ and also $\|\mathbf{y}_i - B\mathbf{b}_i\| \leq \varepsilon$ for $\mathbf{b}_i$ all sharing the same support $S' \in \binom{[m]}{k}$. **[Do we really need this many samples per support? Consider the information gained by noting that we have vectors distributed over multiple supports and that A satisfies the spark condition.]** Let $\mathcal{Y}$ be a set of k such vectors $\mathbf{y}_i$ which we will index by $\mathcal{I}$, i.e. $\mathcal{Y} = \{\mathbf{y}_i : i \in \mathcal{I}\}$.

It follows from the general linear position of the $\mathbf{a}_i$ and the fact that every k columns of A are linearly independent that the set $\{A\mathbf{a}_i : i \in \mathcal{I}\}$ is a basis for $\mathrm{Span}\{A_S\}$. Hence, fixing $\mathbf{z} \in \mathrm{Span}\{A_S\}$, there exists a unique set of $c_i \in \mathbb{R}$ (for notational convenience we index these c_i with $\mathcal{I}$ as well) such that $\mathbf{z} = \sum_{i\in\mathcal{I}} c_i A\mathbf{a}_i$. Letting $\mathbf{y} = \sum_{i\in\mathcal{I}} c_i \mathbf{y}_i \in \mathrm{Span}\{\mathcal{Y}\}$, we have by the triangle inequality that

$$\|\mathbf{z} - \mathbf{y}\|_2 = \|\sum_{i\in\mathcal{I}} c_i A\mathbf{a}_i - \sum_{i\in\mathcal{I}} c_i \mathbf{y}_i\|_2 \leq \sum_{i\in\mathcal{I}} \|c_i(A\mathbf{a}_i - \mathbf{y}_i)\|_2 = \sum_{i\in\mathcal{I}} |c_i| \|A\mathbf{a}_i - \mathbf{y}_i\|_2 \leq \varepsilon \sum_{i\in\mathcal{I}} |c_i|. \tag{20}$$

The alternate factorization for the $\mathbf{y}_i$ implies (by a manipulation identical to that of (20)) that for $\mathbf{z}' = \sum_{i\in\mathcal{I}} c_i B\mathbf{b}_i \in \mathrm{Span}\{B_{S'}\}$ we have $\|\mathbf{y} - \mathbf{z}'\|_2 \leq \varepsilon \sum_{i\in\mathcal{I}} |c_i|$ as well. It follows again by the triangle inequality that

$$\|\mathbf{z} - \mathbf{z}'\|_2 \leq \|\mathbf{z} - \mathbf{y}\|_2 + \|\mathbf{y} - \mathbf{z}'\|_2 = 2\varepsilon \sum_{i\in\mathcal{I}} |c_i|. \tag{21}$$

Since $\mathrm{supp}(\mathbf{a}_i) = S$ for all $i \in \mathcal{I}$ and A has k-sparse restricted lower bound α, we have

$$\|\mathbf{z}\|_2 = \|\sum_{i\in\mathcal{I}}^{k} c_i A\mathbf{a}_i\|_2 = \|A(\sum_{i\in\mathcal{I}}^{k} c_i \mathbf{a}_i)\|_2 \geq \alpha \|\sum_{i\in\mathcal{I}}^{k} c_i \mathbf{a}_i\|_2 \geq \frac{\alpha\sigma}{\sqrt{k}} \sum_{i\in\mathcal{I}}^{k} |c_i|, \tag{22}$$

where for the last inequality we have applied the property (8), noting that $\sqrt{k}\|c\|_2 \geq \|c\|_1$ for all $c \in \mathbb{R}^k$. Combining (21) and (22), we see that for all $\mathbf{z} \in \mathrm{Span}\{A_S\}$ there exists some $\mathbf{z}' \in \mathrm{Span}\{B_{S'}\}$ such that $\|\mathbf{z} - \mathbf{z}'\|_2 \leq \frac{2\varepsilon\sqrt{k}}{\alpha\sigma}\|\mathbf{z}\|_2$. It follows that $d(\mathbf{z}, \mathrm{Span}\{B_{S'}\}) \leq \frac{2\varepsilon\sqrt{k}}{\alpha\sigma}$ for all unit vectors $\mathbf{z} \in \mathrm{Span}\{A_S\}$. Hence,

$$\sup_{\substack{\mathbf{z}\in\mathrm{Span}\{A_S\} \\ \|\mathbf{z}\|=1}} d(\mathbf{z}, \mathrm{Span}\{B_{S'}\}) \leq \frac{2\varepsilon\sqrt{k}}{\alpha\sigma}. \tag{23}$$

If $\varepsilon < \frac{\alpha\sigma}{2\sqrt{k}}$, we have that $\dim(\mathrm{Span}\{B_{S'}\}) \geq \dim(\mathrm{Span}\{A_S\}) = k$ by (17) and the fact that every k columns of A are linearly independent . In fact, since $|S'| = k$, we have $\dim(\mathrm{Span}\{B_{S'}\}) = \dim(\mathrm{Span}\{A_S\})$. Recalling (16), we see the association $S \mapsto S'$ thus defines a map $\pi : \{S_1, \ldots, S_m\} \to \binom{[m]}{k}$ satisfying $\Theta(\mathrm{Span}\{A_S\}, \mathrm{Span}\{\mathcal{B}_{\pi(S)}\}) \leq \frac{2\varepsilon\sqrt{k}}{\alpha\delta}$.

If $\varepsilon < \frac{\alpha^2\sigma\phi_k(A)}{(2k)^{2/3}}$, we then indeed have $\varepsilon < \frac{\alpha\sigma}{2\sqrt{k}}$ (since $\alpha \leq 1$, $\phi_k(A) \leq 1$ and $k \geq 1$) so that

$$\Theta(\mathrm{Span}\{A_{S_i}\}, \mathrm{Span}\{\mathcal{B}_{\pi(S_i)}\}) \leq \frac{\phi_k(A)}{k}\Delta \quad \text{for all} \quad i = 0, \ldots, m-1, \tag{24}$$

where $\Delta = \frac{2k^{2/3}\varepsilon}{\alpha\sigma\phi_k(A)} < \frac{\alpha}{\sqrt{2}}$. Moreover, it follows by Lemma 1 that there exists a permutation matrix $P \in \mathbb{R}^{m\times m}$ and a diagonal matrix $D \in \mathbb{R}^{m\times m}$ such that for all $i \in \{1, \ldots, m\}$, $\|A - BPD\|_1 \leq C\varepsilon$ for $C = \frac{2k^{2/3}\sqrt{m}}{\alpha\sigma\phi_k(A)}$. $\square$

Corollary 1.

Proof. $\square$

III. DIFFERENT CODING DIMENSIONS

In this section we address the issue of recovering $A \in \mathbb{R}^{n\times m}$ and the $\mathbf{a}_i \in \mathbb{R}^m$ when only an upper bound m' on the number m of sparse sources is known.

Theorem 2. *Given positive integers n, m, m' and k with $k < m \leq m'$, there exist $N = mk\binom{m'}{k}$ k-sparse vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ with the following property: if $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ is a dataset for which for some $A \in \mathbb{R}^{n\times m}$ satisfying spark condition (3) we have $\|\mathbf{y}_i - A\mathbf{a}_i\|_2 \leq \varepsilon$ for all $i = 1, \ldots, N$, then any other matrix $A' \in \mathbb{R}^{n\times m'}$ also satisfying spark condition (3) and for which there exist k-sparse $\mathbf{a}'_1, \ldots, \mathbf{a}'_N \in \mathbb{R}^{m'}$ such that $\|\mathbf{y}_i - A\mathbf{a}'_i\|_2 \leq \varepsilon$ for all $i = 1, \ldots, N$ necessarily satisfies $\|A - A'PD\| \leq C\varepsilon$ for some $C > 0$, permutation matrix $P \in \mathbb{R}^{m'\times m'}$ and diagonal matrix $D \in \mathbb{R}^{m'\times m}$, provided ε is small enough. Moreover, this implies $\|\mathbf{b}_i - PD\mathbf{a}_i\|_1 \leq C'\varepsilon$ for some $C' > 0$.*

Specifically, the $\mathbf{a}_i$ are constructed so that any k of them are linearly independent. In particular, any set of k vectors $\mathbf{a}_i$ that shate the same support satisfy (8) for some $\sigma > 0$. Letting $\Phi = \min(\phi_k(A), \phi_k(B))$, if all $n \times \max(2, k)$ and $n \times 2k$ submatrices of A have lower bound α_k and α_{2k}, respectively, then provided $\varepsilon < (2k)^{-2/3}\alpha_k^2\sigma\Phi$ we have

$$\|A - BPD\|_1 \leq \frac{2\varepsilon k^{2/3}\sqrt{m}}{\alpha_k\sigma\Phi} \tag{25}$$

and

$$\|\mathbf{b}_i - PD\mathbf{a}_i\|_1 \le \frac{\varepsilon}{\alpha_{2k}\|D\|_2} \left(1 + \frac{2k^{2/3}}{\alpha_k\sigma\phi_k(A)}\|D^T P^T \mathbf{b}_i\|_1\right). \tag{26}$$

The proof of this adaptation of Theorem 2 is very similar to the proof of Theorem 1, the only difference being that now we establish a map $\pi : [m] \to [m']$ satisfying the requirements of Lemma 2, the statement of which follows, by pigeonholing $k\binom{m'}{k}$ vectors with respect to $[m']$ holes.

Lemma 2 (Main Lemma for $m \le m'$). *Fix positive integers n, m, m' and k with $k < m \le m'$. Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times m'}$ with having unit norm columns. Suppose both A and B satisfy spark condition (3) with $\alpha \in (0, 1]$ as a lower bound for every one of their $n \times \max(2, k)$ submatrices. Let $\mathcal{T}$ be the set of all $S_i := \{i, \ldots, (i + (k-1)\} \mod m$ for $i = 0, \ldots, m - 1$. If there exists a map $\pi : \mathcal{T} \to \binom{[m]}{k}$ and some $\Delta < \frac{\alpha}{\sqrt{2}}$ such that*

$$\Theta(Span\{A_{S_i}\}, Span\{B_{\pi(S_i)}\}) \le \frac{\min(\phi_k(A), \phi_k(B))}{k}\Delta \quad for \ all \quad i \in \{0, \ldots, m-1\} \tag{27}$$

then there exist a permutation matrix $P \in \mathbb{R}^{m' \times m'}$ and a diagonal matrix $D \in \mathbb{R}^{m' \times m}$ such that

$$\|A - BPD\|_1 \le \sqrt{m}\Delta \quad for \ all \quad i \in [m]. \tag{28}$$

We again defer the proof of this lemma to the Appendix.

IV. DISCUSSION

The theory of CS informs also informs another practical consequence of our result. Since our derived sample complexity is independent of the ambient dimension of the data, n, given a lower bound on the sparsity of the latent variables $\mathbf{a}_i$ we can generate a random matrix to compress the data to a dimension in the regime of (11) before applying a dictionary learning. This could significantly reduce the computational cost of dictionary learning when the sparsity is high by reducing the number of parameters required to define each dictionary element. Such improvements are crucial to scaling up dictionary learning to larger datasets. **[Is this actually a significant computational boost..?]**

Uniqueness in data analysis.

Identifying uniqueness: the bispectrum

APPENDIX A
COMBINATORIAL MATRIX THEORY

In this section, we prove Lemma 1, which was a main ingredient in our proof of Theorem 1. We suspect that there is an appropriate generalization to matroids. First, we prove some auxiliary lemmas.

Lemma 3. *Let $M \in \mathbb{R}^{n \times m}$. If every $2k$ columns of M are linearly independent, then for any $\mathcal{T} \subseteq \binom{[m]}{k}$,*

$$y \in Span\{M_{\cap \mathcal{T}}\} \iff y \in \bigcap_{S \in \mathcal{T}} Span\{M_S\}. \tag{29}$$

Proof. The forward direction is trivial; we prove the reverse direction by induction. Enumerate $\mathcal{T} = (S_1, \ldots, S_{|\mathcal{T}|})$ and let $y \in \bigcap_i Span\{M_{S_i}\}$. Then for all $S_i \in \mathcal{T}$ there exists some x_i with support contained in S_i such that $y = Mx_i$. In particular, $y = Mx_1$ for x_1 with support in S_1. Now suppose there exists some x with support contained in $\cap_{i=1}^{\ell-1}S_i$, $\ell \le |\mathcal{T}|$ such that $y = Mx$. Then $y = Mx = Mx_\ell$, implying $x = x_\ell$ (since every $2k$ columns of M are linearly independent). Hence the support of x is also contained in S_ℓ, i.e. $\text{supp}(x) \subseteq \cap_{i=1}^{\ell}S_i$. $\qquad\square$

Lemma 4. *Fix $k \ge 2$. Let $V_1, \ldots, V_k$ be closed subspaces of $\mathbb{R}^m$, let $V = \cap_{i=1}^k V_i$ and let $c_i := \cos\left[\theta_F\left(V_i, \cap_{j=i+1}^k V_k\right)\right]$ for $i = 1, \ldots, k$. Then for each $x \in \mathbb{R}^m$,*

$$\|x - \Pi_V x\| \le \frac{1}{1-c}\sum_{i=1}^k \|x - \Pi_{V_i}x\|, \tag{30}$$

where

$$c := \left[1 - \prod_{i=1}^k(1 - c_i)\right]^{1/2}. \tag{31}$$

Proof. Fix $x \in \mathbb{R}^m$ and $k \ge 2$. We begin with a proof by induction to show that

$$\|x - \Pi_V x\| \le \sum_{i=1}^k \|x - \Pi_{V_i}x\| + \|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}x - \Pi_V x\|. \tag{32}$$

Since $\Pi_V x \in V_k$ for all $x \in \mathbb{R}^m$, we have by the triangle inequality that

$$\|x - \Pi_V x\| = \|x - \Pi_{V_k} x\| + \|\Pi_{V_k}(I - \Pi_{V_{k-1}})x\| + \|\Pi_{V_k}\Pi_{V_{k-1}}x - \Pi_V x\|$$

$$\leq \sum_{i=k-1}^{k} \|x - \Pi_{V_i} x\| + \|\Pi_{V_k}\Pi_{V_{k-1}}x - \Pi_V x\|,$$

where we have used the fact that $\|\Pi_{V_k}\| \leq 1$. If $k = 2$ then we are done; otherwise, suppose now that for $1 \leq \ell < k - 1$ we have

$$\|x - \Pi_V x\| \leq \sum_{i=k-\ell}^{k} \|x - \Pi_{V_i} x\| + \|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_{k-\ell}}x - \Pi_V x\|$$

Applying the triangle inequality to the second term, we have

$$\|x - \Pi_V x\| = \sum_{i=k-\ell}^{k} \|x - \Pi_{V_i} x\| + \|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_{k-\ell}}(I - \Pi_{V_{k-(\ell+1)}})x\| + \|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_{k-(\ell+1)}}x - \Pi_V x\|$$

$$\leq \sum_{i=k-(\ell+1)}^{k} \|x - \Pi_{V_i} x\| + \|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_{k-(\ell+1)}}x - \Pi_V x\|,$$

again using $\|\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_{k-\ell}}\| \leq 1$. This completes the proof of (32) by induction. Making use of the following result by [5] (Theorem 9.33):

$$\|\left(\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1}\right)x - \Pi_V x\| \leq c\|x\| \quad \text{for all} \quad x \in \mathbb{R}^m, \tag{33}$$

having recalled the definition of c in (31), and noting that

$$\|(\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1})(x - \Pi_V x) - \Pi_V(x - \Pi_V x)\| = \|(\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1})x - \Pi_V x\|, \tag{34}$$

since $\Pi_{V_\ell}\Pi_V = \Pi_V$ for all $\ell = 1, \ldots, k$ and $\Pi_V^2 = \Pi_V$, we have by (33) and (34) that

$$\|(\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1})x - \Pi_V x\| = \|(\Pi_{V_k}\Pi_{V_{k-1}}\cdots\Pi_{V_1})(x - \Pi_V x) - \Pi_V(x - \Pi_V x)\|$$

$$\leq c\|x - \Pi_V x\|.$$

Substituting the LHS into (32), we get

$$\|x - \Pi_V x\| \leq \sum_{i=1}^{k} \|x - \Pi_{V_i} x\| + c\|x - \Pi_V x\|,$$

from which the result follows by solving for $\|x - \Pi_V x\|$. $\qquad \square$

Lemma 5. *Fix positive integers $k < m$ and let $\mathcal{T} = \{S_0, \ldots, S_{m-1}\}$, where for $i = 0, \ldots m - 1$,*

$$S_i = \{i, i+1, \ldots, i+(k-1)\} \bmod m.$$

Suppose there exists a map $\pi : \mathcal{T} \to \binom{\mathbb{Z}/m\mathbb{Z}}{k}$ such that for all $\mathcal{F} \in \binom{\mathcal{T}}{k}$,

$$\bigcap_{S \in \mathcal{F}} S = \emptyset \implies \bigcap_{S \in \mathcal{F}} \pi(S) = \emptyset. \tag{35}$$

Then $\pi(S_i) \cap \cdots \cap \pi(S_{i+(k-1)}) \neq \emptyset$ for all $i \in \mathbb{Z}/m\mathbb{Z}$.

Proof. Consider the set $Q_m = \{(i, j) : i \in \mathbb{Z}/m\mathbb{Z}, j \in \pi(S_i)\}$, which has mk elements. By the pigeon-hole principle, there is some $p \in \mathbb{Z}/m\mathbb{Z}$ and at least k distinct $i_1, \ldots, i_k$ such that $\{(i_1, p), \ldots, (i_k, p)\} \subseteq Q_m$. Hence, $p \in \pi(S_{i_1}) \cap \cdots \cap \pi(S_{i_k})$ and by (35) there must be some $v \in \mathbb{Z}/m\mathbb{Z}$ such that $v \in S_{i_1} \cap \cdots \cap S_{i_k}$. This is only possible (given $\mathcal{T}$) if $i_1, \ldots, i_k$ are consecutive modulo $\mathbb{Z}/m\mathbb{Z}$, i.e. $\{i_1, \ldots, i_k\} = \{v - (k-1), \ldots, v\}$.

We now claim that there exists no additional $i \notin \{v - (k-1), \ldots, v\}$ such that $p \in \pi(S_i)$. To see why, note that if there were such an i then we would have $p \in \pi(S_i) \cap \pi(S_{v-(k-1)}) \cap \cdots \cap \pi(S_v)$ and (35) would imply that every k-element subset of $\{i\} \cup \{v - (k-1), \ldots, v\}$ is a set of consecutive integers. This is only possible if $m = k + 1$; but then there can't have been $k + 1$ distinct elements of $\binom{\mathbb{Z}/m\mathbb{Z}}{k}$ all containing p since there are only $\binom{m-1}{m-2} = m - 1 = k$ distinct elements of $\binom{\mathbb{Z}/m\mathbb{Z}}{m-1}$ which contain p. Thus, letting $Q_{m-1} \subset Q_m$ be the set of elements of Q_m not having p as a second coordinate, we have $|Q_{m-1}| = (m-1)k$ and the proof follows by iterating these arguments. $\qquad \square$

Proof of Lemma 1 (Main Lemma). We assume $k \geq 2$ since the case $k = 1$ was proven at the beginning of Section II. We begin by proving some useful facts. Note that by (18), for any $S \in \mathcal{T}$ we have for all unit vectors $u \in \text{Span}\{A_S\}$

that $d(u, \text{Span}\{B_{\pi(S)}\}) \leq \frac{\phi_k(A)}{k}\Delta$. Note also that since $\phi_k(A) \in [0, 1]$, we have $\frac{\phi_k(A)}{k}\Delta < 1$; hence by (17) we have $\dim(\text{Span}\{B_{\pi(S)}\}) \geq \dim(\text{Span}\{A_S\}) = k$. Since $|\pi(S)| = k$ we in fact have $\dim(\text{Span}\{B_{\pi(S)}\}) = k$, i.e. the columns of $B_{\pi(S)}$ are linearly independent for any $S \in \mathcal{T}$.

Next, fix $\mathcal{I} = \{i_1, \ldots, i_k\} \in \binom{[m]}{k}$. Condition (18) implies that for all unit vectors $\mathbf{z} \in \text{Span}\{B_{\cap_{i \in \mathcal{I}} \pi(S_i)}\} \subseteq \cap_{i \in \mathcal{I}}\text{Span}\{B_{\pi(S_i)}\}$ we have $d(\mathbf{z}, \text{Span}\{A_{S_i}\}) \leq \frac{\phi_k(A)}{k}\Delta$ for all $i \in \mathcal{I}$. It follows by Lemmas 3 and 4 that

$$d\left(\mathbf{z}, \text{Span}\{A_{\cap_{i \in \mathcal{I}} S_i}\}\right) \leq \Delta \left(\frac{\phi_k(A)}{1 - c(\text{Span}\{A_{S_{i_1}}\}, \ldots, \text{Span}\{A_{S_{i_\ell}}\})} \right) \leq \Delta. \tag{36}$$

By (17) we have that $\dim(\text{Span}\{B_{\cap_{i \in \mathcal{I}} \pi(S_i)}\}) \leq \dim(\text{Span}\{A_{\cap_{i \in \mathcal{I}} S_i}\})$. It follows by the linear independence of the columns of $B_{\pi(S)}$ that

$$|\cap_{i \in \mathcal{I}} \pi(S_i)| \leq |\cap_{i \in \mathcal{I}} S_i| \quad \text{for all} \quad \mathcal{I} \in \binom{[m]}{k}. \tag{37}$$

Now, fix $i \in [m]$ and let $\mathcal{I} = \{i - (k-1), \ldots, i\}$. Then $\cap_{i \in \mathcal{I}} S_i = \{i\}$. (In general, there may exist combinations of fewer supports S_i with intersection $\{i\}$, e.g. if $m \geq 2k - 1$ then $S_{i-(k-1)} \cap S_i = \{i\}$. For brevity, we consider a construction that is valid for any $m > k$.) By (37) we have that $\cap_{i \in \mathcal{I}} \pi(S_i)$ is either empty or it contains a single element. Lemma 5 ensures that the latter case is the only possibility. Thus the association $i \mapsto \cap_{i \in \mathcal{I}} \pi(S_i)$ defines a map $\hat{\pi} : [m] \to [m]$. Recalling (16), it follows from (36) that for all unit vectors $\mathbf{z} \in \text{Span}\{A_i\}$ we have $d\left(\mathbf{z}, \text{Span}\{B_{\hat{\pi}(i)}\}\right) \leq \Delta$ as well, i.e. for every canonical basis vector $\mathbf{e}_i \in \mathbb{R}^m$, there exists some $c_i \in \mathbb{R}$ such that $\|A\mathbf{e}_i - c_i B\mathbf{e}_{\hat{\pi}(i)}\| \leq \Delta$ where $\Delta < \frac{\alpha}{\sqrt{2}}$. This is exactly the supposition in (12) and the result follows from the subsequent arguments of Section II. $\square$

APPENDIX B
DIFFERENT CODING DIMENSIONS

In this section we prove Lemma 2 from Section III.

Proof. We begin with the case $k = 1$ and then show how the case for $k > 1$ reduces to this. Since $\phi_1(A) = \phi_1(B) := 1$, equation (27) asserts that for all canonical basis vectors $\mathbf{e}_i \in \mathbb{R}^m$ there is some $c_i \in \mathbb{R}$ such that

$$\|A\mathbf{e}_i - c_i B\mathbf{e}_{\pi(i)}\|_2 \leq \Delta \quad \text{for all} \quad i \in [m]. \tag{38}$$

Note that if $c_i = 0$ for some i, then $\|A\mathbf{e}_i\| < 1$ (since $\alpha \leq 1$), contradicting the fact that A has unit norm columns. We will now show that π is necessarily injective (and thus defines a permutation). Suppose that $\pi(i) = \pi(j) = \ell$ for some $i \neq j$ and $\ell \in [m']$. Then $\|A\mathbf{e}_i - c_i B\mathbf{e}_\ell\|_2 \leq \Delta$ and $\|A\mathbf{e}_j - c_j B\mathbf{e}_\ell\|_2 \leq \Delta$. Summing and scaling these two inequalities by $|c_j|$ and $|c_i|$, respectively, we apply the triangle inequality and then (4) to yield

$$\begin{aligned}(|c_i| + |c_j|)2\varepsilon &\geq |c_j|\|A\mathbf{e}_i - c_i B\mathbf{e}_\ell\|_2 + |c_i|\|c_j B\mathbf{e}_\ell - A\mathbf{e}_j\|_2 \\ &\geq \|c_j A\mathbf{e}_i + c_i A\mathbf{e}_j\|_2 \\ &\geq \alpha\|c_j \mathbf{e}_i + c_i \mathbf{e}_j\|_2\end{aligned}$$

which is in contradiction with the fact that $\|x\|_1 \leq \sqrt{2}\|x\|_2$ for all $x \in \mathbb{R}^2$ and $\Delta < \frac{\alpha}{\sqrt{2}}$. Hence, π is injective. Letting P and D denote the following permutation and diagonal matrices, respectively:

$$P = \left(\mathbf{e}_{\pi(1)} \cdots \mathbf{e}_{\pi(m)}, \mathbf{1}, \cdots, \mathbf{1}\right), \quad D = \begin{pmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_m \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}, \tag{39}$$

we see that (38) becomes $\|(A - BPD)\mathbf{e}_i\|_2 \leq \Delta$ for all $i \in [m]$ and it follows that $\|A - BPD\|_1 \leq \sqrt{m}\Delta$.

Now suppose $k \geq 2$ and let $\mathcal{F} = \{S_{i_1}, \ldots, S_{i_k}\} \in \binom{\mathcal{T}}{k}$. Condition (18) implies that for all unit vectors $\mathbf{z} \in \text{Span}\{B_{\cap_{S \in \mathcal{F}} \pi(S)}\} \subseteq \cap_{S \in \mathcal{F}}\text{Span}\{B_{\pi(S)}\}$ we have $d(\mathbf{z}, \text{Span}\{A_S\}) \leq \frac{\phi_k(A)}{k}\Delta$ for all $S \in \mathcal{F}$. It follows by Lemmas 3 and 4 that

$$d\left(\mathbf{z}, \text{Span}\{A_{\cap_{i \in \mathcal{I}} S_i}\}\right) \leq \Delta \left(\frac{\phi_k(A)}{1 - c(\text{Span}\{A_{S_{i_1}}\}, \ldots, \text{Span}\{A_{S_{i_\ell}}\})} \right) \leq \Delta. \tag{40}$$

By (17) we have that $\dim(\mathrm{Span}\{B_{\cap_{S\in\mathcal{F}}\pi(S)}\}) \leq \dim(\mathrm{Span}\{A_{\cap_{S\in\mathcal{F}}\S}\})$. Likewise, for all unit vectors $\mathbf{z} \in \mathrm{Span}\{A_{\cap_{S\in\mathcal{F}}S}\} \subseteq \cap_{S\in\mathcal{F}}\mathrm{Span}\{A_S\}$ we have $d(\mathbf{z}, \mathrm{Span}\{B_{\pi(S)}\}) \leq \frac{\phi_k(B)}{k}\Delta$ for all $S \in \mathcal{F}$. It follows by Lemmas 3 and 4 that

$$d\left(\mathbf{z}, \mathrm{Span}\{B_{\bigcap_{S\in\mathcal{F}}\pi(S)}\}\right) \leq \Delta\left(\frac{\phi_k(B)}{1 - c(\mathrm{Span}\{B_{\pi(S_{i_1})}\}, \dots, \mathrm{Span}\{B_{\pi(S_{i_\ell})}\})}\right) \leq \Delta. \tag{41}$$

By (17) we have that $\dim(\mathrm{Span}\{A_{\cap_{S\in\mathcal{F}}S}\}) \leq \dim(\mathrm{Span}\{B_{\cap_{S\in\mathcal{F}}\pi(S)}\})$. Hence, $\dim(\mathrm{Span}\{A_{\cap_{S\in\mathcal{F}}S}\}) = \dim(\mathrm{Span}\{B_{\cap_{S\in\mathcal{F}}\pi(S)}\})$ and since every k columns of A and every k columns of B are linearly independent, we have

$$|\cap_{S\in\mathcal{F}}\pi(S)| = |\cap_{S\in\mathcal{F}}S| \quad \text{for all} \quad \mathcal{F} \in \binom{\mathcal{T}}{k}. \tag{42}$$

Now, fix $i \in [m]$ and let $\mathcal{F} \subseteq \mathcal{T}$ be any collection of sets for which $\cap_{S\in\mathcal{F}}S = \{i\}$ e.g. $\mathcal{F} = \{S_{i-(k-1)}, \dots, S_i\}$. (In general, there may exist combinations of fewer supports S_i with intersection $\{i\}$, e.g. if $m \geq 2k - 1$ then $S_{i-(k-1)} \cap S_i = \{i\}$. For brevity, we consider a construction that is valid for any $m > k$.) Then $\cap_{S\in\mathcal{F}}S = \{i\}$ and the association $i \mapsto \cap_{S\in\mathcal{F}}\pi(S)$ defines a map $\hat{\pi} : [m] \to [m]$. It follows from (42) that for every canonical basis vector $\mathbf{e}_i \in \mathbb{R}^m$, there exists some $c_i \in \mathbb{R}$ such that $\|A\mathbf{e}_i - c_iB\mathbf{e}_{\hat{\pi}(i)}\| \leq \Delta$. This is exactly the supposition in (38). $\square$

APPENDIX C
CALCULATING C

In this section we derive the constant C for a particular generating dictionary $A \in \mathbb{R}^{n\times m}$ and $\mathbf{a}_1, \dots, \mathbf{a}_N \in \mathbb{R}^m$.

Lemma 6. *Let $\gamma_1 < \dots < \gamma_N$ be any distinct numbers such that $\gamma_{i+1} = \gamma_i + \delta$ and form the $k \times N$ Vandermonde matrix $V = (\gamma_j^i)_{i,j=1}^{k,N}$. Then for all $S \in \binom{[N]}{k}$,*

$$\|V_S x\|_2 > \rho\|x\|_1 \quad \text{where} \quad \rho = \frac{\delta^k}{\sqrt{k}}\left(\frac{k-1}{k}\right)^{\frac{k-1}{2}}\prod_{i=1}^{k}(\gamma_1 + (i-1)\delta) \quad \text{for all} \quad x \in \mathbb{R}^k \tag{43}$$

Proof. The determinant of the Vandermonde matrix is

$$\det(V) = \prod_{1\leq j\leq k}\gamma_j \prod_{1\leq i\leq j\leq k}(\gamma_j - \gamma_i) \geq \delta^k\prod_{i=1}^{k}(\gamma_1 + (i-1)\delta). \tag{44}$$

Since the γ_i are distinct, the determinant of any $k \times k$ submatrix of V is nonzero; hence V_S is nonsingular for all $S \in \binom{[N]}{k}$. Suppose $x \in \mathbb{R}^k$. Then $\|x\|_2 = \|V_S^{-1}V_S x\|_2 \leq \|V_S^{-1}\|\|V_S x\|_2$, implying $\|V_S x\|_2 \geq \|V_S^{-1}\|^{-1}\|x\|_2 \geq \frac{1}{\sqrt{k}}\|V_S\|_2^{-1}\|x\|_1$. For the Euclidean norm we have $\|V_S^{-1}\|_2^{-1} = \sigma_{\min}(V_S)$, where $\sigma_{\min}$ is the smallest singular value of V_S. A lower bound for the smallest singular value of a nonsingular matrix $M \in \mathbb{R}^{k\times k}$ is given in [Hong and Pan]:

$$\sigma_{\min}(M) > \left(\frac{k-1}{k}\right)^{\frac{k-1}{2}}|\det M| \tag{45}$$

and the result follows. $\square$

ACKNOWLEDGMENT

REFERENCES

[1] A. C. P Georgiev, F Theis, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16(4), pp. 992–996, 2005.
[2] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.
[3] E. J. Candes and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.
[4] I. D. Morris, "A rapidly-converging lower bound for the joint spectral radius via multiplicative ergodic theory," *Advances in Mathematics*, vol. 225, no. 6, pp. 3425–3445, 2010.
[5] F. R. Deutsch, *Best approximation in inner product spaces*. Springer Science & Business Media, 2012.

Charles J. Garfinkle

Christopher J. Hillar

SCRAP PAPER

A. Introduction

We also require the data to satisfy certain properties. Consider the problem where we wish to identify the mixing matrix A from the mixtures $\mathbf{y}_i$ when the sources $\mathbf{a}_i$ are known. In this case, a necessary condition for uniqueness of A given $\mathbf{y}_i = A\mathbf{a}_i + \eta_i$ (even when $\eta_i = 0$) is:

$$A^{(1)}_{i,:}(\mathbf{a}_1 \cdots \mathbf{a}_N) = A^{(2)}_{i,:}(\mathbf{a}_1 \cdots \mathbf{a}_N) \text{ for all } i \in [n] \implies A^{(1)} = A^{(2)} \quad \text{for all } A^{(1)}, A^{(2)} \in \mathbb{R}^{n \times m}. \tag{46}$$

Otherwise, blah blah blah (figure this out). Trying to introduce the "spread" of the data as a necessary condition.

B. Statements of Probabilistic Theorems

We next give precise statements of our probabilistic versions of Theorem 1, all of which rely on the following construction.

Definition 6 (Random Sparse Vectors). *Given the support set for its k nonzero entries, a random draw of $\mathbf{a}$ is the k-sparse vector with support entries chosen uniformly from the interval $[0,1] \subset \mathbb{R}$, independently. When a support set is not specified, a random draw is a choice of one support set uniformly from all of them in $\binom{[m]}{k}$ and then a random draw.*

Definition 7 (Special Random Sparse Vectors). *Given the support set for its k nonzero entries, a random draw of $\mathbf{a}$ is the k-sparse vector with support entries chosen uniformly from the interval $[0,1] \subset \mathbb{R}$, independently. When a support set is not specified, a random draw is a choice of one support set uniformly from all m of them in $\mathcal{T}$ and then a random draw.*

Theorem 3. *Fix positive integers $k < m$ and n, and a generation matrix $A \in \mathbb{R}^{n \times m}$ satisfying spark condition (3). If $(k+1)$ k-sparse $\mathbf{a}_i$ are randomly drawn from each support set in $\mathcal{T}$, then $Y = \{A\mathbf{a}_1 + \eta_1, \ldots, A\mathbf{a}_N + \eta_N\}$ for $\|\eta_i\| \leq \varepsilon$ has a unique sparse coding with probability one, provided ε is small enough.*

The following is a direct application of this result, partially answering one of our questions from the introduction.

Corollary 2. *Suppose m, n, and k satisfy inequality (11). With probability one, a random [insert footnote explaining 'random'] $n \times m$ generation matrix A satisfies (3). Fixing such an A, for ε small enough we have with probability one that a dataset $Y = \{A\mathbf{a}_1 + \eta_1, ..., A\mathbf{a}_N + \eta_1\}$ for $\|\eta_i\| \leq \varepsilon$ generated from $N = (k+1)m$ k-sparse samples $\mathbf{a}_i$ uniquely prescribes A and these sparse vectors $\mathbf{a}_i$ up to an approximate fixed permutation and scaling.*

We now state our third theorem. Note that an *algebraic set* is a solution to a finite set of polynomial equations.

Theorem 4. *Fix positive integers $k < m$ and n. If $(k+1)$ k-sparse $\mathbf{a}_i$ are randomly drawn from each support set in $\mathcal{T}$, then with probability one the following holds. There is an algebraic set $Z \subset \mathbb{R}^{n \times m}$ of Lebesgue measure zero with the following property: if $A \notin Z$ then there exists some ε small enough such that $Y = \{A\mathbf{a}_1 + \eta_1, ..., A\mathbf{a}_N + \eta_1\}$ for $\|\eta_i\| \leq \varepsilon$ has a unique sparse coding.*

Corollary 3. *Suppose m, n, and k obey inequality (11). With probability one, a random draw of $N = (k+1)m$ k-sparse samples $\mathbf{a}_i$ with supports in $\mathcal{T}$ satisfies: for almost every matrix A there is some ε small enough that gives $Y = \{A\mathbf{a}_1 + \eta_1, ..., A\mathbf{a}_N + \eta_1\}$ for $\|\eta_i\| \leq \varepsilon$ a unique sparse coding.*

C. Proofs of Probabilistic Theorems

Definition 6: can I bound the probability of drawing from one of the $m!$ special support sets without explicitly making that restriction?

We first generalize Lemma 3 in [?] to the noisy case:

Lemma 7. *Fix $A \in \mathbb{R}^{n \times m}$ satisfying (4). With probability one, if $(k+1)$ k-sparse vectors $\mathbf{a}_i \in \mathbb{R}^{n \times m}$ are such that $d(A\mathbf{a}_i, V) \leq (??)$ for some k-dimensional subspace $V \subset \mathbb{R}^m$ then all of the $(k+1)$ vectors $\mathbf{a}_i$ have the same supports.*

Proof. We need only show that the $k+1$ vectors $A\mathbf{a}_i$ are linearly dependent; the rest follows by Lemma 3 from [?]. Let these $k+1$ vectors $\mathbf{a}_i$ be indexed by $\mathcal{I}$ and let $W = \text{Span}\{A\mathbf{a}_{i \in \mathcal{I}}\}$. Then for all $w \in W$ we can write $w = \sum_{i \in \mathcal{I}} c_i A\mathbf{a}_i$ for some set of $c_i \in \mathbb{R}$. Letting $v = \sum_{i \in)} c_i v_i$, it follows that

$$\|w - v\| = \|\sum_{i \in \mathcal{I}} c_i A\mathbf{a}_i - \sum_{i \in \mathcal{I}} c_i v_i\| \leq \sum_{i \in \mathcal{I}} |c_i| \|A\mathbf{a}_i - v_i\| \leq 2\varepsilon \sum_{i \in \mathcal{I}} |c_i|$$

Need RHS less thatn $\|w\|$ to prove that $\dim(W) \leq \dim(V) = k$... $\qquad \square$

Proof of Theorem 3 ($\varepsilon = 0$). Consider any alternate factorization $\mathbf{y}_i = B\mathbf{b}_i$ for Y. Given a support set $S \in \binom{[m]}{k}$, let $J(S) = \{j : \text{supp}(\mathbf{b}_j) \subseteq S\}$ and note that those $\mathbf{y}$ indexed by $J(S)$ span at most a k-dimensional space. By Lemma 3 in [?], either $|J(S)| \leq k$ or with probability one all $\mathbf{a}_j$ with $j \in J(S)$ have the same support S'. Since there are only $(k+1)$ vectors $\mathbf{a}_i$

with a given support, the latter case actually implies (with probability one) that $|J(S)| = k + 1$. If $N = (k+1)\binom{m}{k}$, though, then with probability one we would reach a contradiction if $|J(S)| \leq k$ for any $S \in \binom{[m]}{k}$; hence with probability one we have $|J(S)| = k + 1$. Can we reach the same conclusion when $N = m(k+1)$? (Perhaps by creating 'virtual data' spanning all supports by those data points with supports in $\mathcal{T}$)?

$\square$

Proof of Theorem 3. Let $\mathbf{y}_i = A\mathbf{a}_i + \eta_i$ for all $i \in \{1, \ldots, N\}$ and suppose there is some $B \in \mathbb{R}^{n \times m}$ and k-sparse $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^m$ such that $\|\mathbf{y}_i - B\mathbf{b}_i\| \leq \varepsilon$ for all $i \in \{1, \ldots, N\}$. Then by the triangle inequality we have $\|A\mathbf{a}_i - B\mathbf{b}_i\| \leq 2\varepsilon$ for all $i \in \{1, \ldots, N\}$. Given a support set $S \in \binom{[m]}{k}$, let $J(S) = \{j : \operatorname{supp}(\mathbf{b}_j) \subseteq S\}$ and note that for all $j \in J(S)$ there exists some $v \in \operatorname{Span}\{B_S\}$ such that $\|A\mathbf{a}_i - v\| \leq 2\varepsilon$. By Lemma (??), with probability one, either $|J(S)| \leq k$ for all $\mathbf{a}_j$ with $j \in J(S)$ have the same support. $\square$