Chaz's Theorem: The Return of Hillar

When are sparse sources robustly identifiable?

Abstract

Extension of theorems in HS11 to noisy subsamples of approximately sparse vectors.

Index Terms

bilinear inverse problem, identifiability, dictionary learning, sparse coding, sparse component analysis, 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 yields insight into its structure. Given a dataset $\mathbf{x}_i, \ldots, \mathbf{x}_N \in \mathbb{R}^n$, a simple approach is to assume a linear decomposition:

$$\mathbf{x}_i = A\mathbf{s}_i + \mathbf{n}_i \tag{1}$$

where the unknown matrix $A \in \mathbb{R}^{n\times m}$ and source signals $\mathbf{s}_i \in \mathbb{R}^m$ have some specific properties and the vector $\mathbf{n}_i \in \mathbb{R}^n$ accounts for both noise in the measurements and the degree to which the model fails to accurately capture the structure of the data. Letting $X \in \mathbb{R}^{n\times N}$ and $S \in \mathbb{R}^{m\times N}$ be the matrices with columns $\mathbf{x}_i$ and $\mathbf{s}_i$, respectively, the problem (1) equates to that of approximating the matrix X as the product AS with constraints placed on the individual matrices A and S. For instance, the rows of S can be made to be as statistically independent as possible – this is independent component analysis (ICA). [ref?] Alternatively, given nonnegative X, the matrices A and S can be constrained to be nonnegative as well – this is nonnegative matrix factorization (NMF). [ref?] Sparse component analysis (SCA) [ref?], otherwise known as dictionary learning, sparse coding, or sparse matrix factorization, refers to the problem of solving (1) under the constraint that the columns of S be *sparse*, i.e. they contain at least one zero element. Typically, they should contain as few nonzeros as possible; this emphasizes parsimony in representation by requiring each datum to be some linear combination of only a few elementary atoms from a generating *dictionary*.

An important related question is known as the blind source separation (BSS) problem. Suppose we know *a priori* that a representation such as (1) exists; under what constraints is it unique, or at least approximately so? From the perspective of matrix factorization it is clear that (1) is a special case of bilinear inverse problem, wherein the goal is to identify from some $z \in Z$ a point $(x, y) \in X \times Z$ such that $\mathcal{F}(x, y) \approx z$ given the bilinear mapping $\mathcal{F} : X \times Y \to Z$. In our case we have $\mathcal{F}(A, S) = AS$, from which we can see that if $(A, S) \in \mathbb{R}^{n\times m} \times \mathbb{R}^{m\times N}$ is a solution to the SCA problem, for instance, then so is $(AP^{-1}D^{-1}, PDS)$ for any permutation matrix $P \in \mathbb{R}^{m\times m}$ and invertible diagonal matrix $D \in \mathbb{R}^{m\times m}$. The set of all matrix poducts of permutation and invertible diagonal matrices defines a group, called the *ambiguity transform group*, associated to the sparse matrix factorization problem $X \approx AS$. From this we see that uniqueness can at best be determined only up to the equivalence class of solutions generated by this group action.

Such conditions for *identifiability* of the underlying parameters defining the noiseless SCA problem $\mathbf{x}_i = A\mathbf{s}_i$ were first provided, to our knowledge, by Georgiev et. al. [2005] and Aharon et. al. [2006]. Recently, Hillar and Sommer [2015] proved a more general result, providing deterministic and probabilistic guarantees for the recovery of A and the $\mathbf{s}_i$ (up to permutation and scaling ambiguity). We further generalize their proof to take into account possible noise in the measurement (or inaccuracy of the model) and to require a significantly reduced number of samples. These *robust identifiability* conditions describe when the true model parameters can, in theory, be identified up to noise levels with exact recovery in the noiseless limit. Moreover, we provide conditions under which recovery is theoretically possible given a priori knowledge of only an upper bound on the number of sources. Our results demonstrate that source sparsity is a very reasonable constraint to consider in realistic modeling scenarios – that is, sparse components and the mixing matrix can be uniquely identified (up to permutation and scaling ambiguities and noise levels) from unknown noisy mixtures regardless of which algorithm is used to recover them, provided it produces a representation that adequately encodes the data. The ability to state such a general guarantee is another attractive quality of the relatively simple linear model (1).

It is informative to explain the relationship of this result to the recently emergent field of *compressed sensing* (CS). [ref?] The theory of CS provides techniques 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*:

$$A\mathbf{a}_1 = A\mathbf{a}_2 \implies \mathbf{a}_1 = \mathbf{a}_2 \quad \text{for all } k\text{-sparse } \mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^m. \tag{2}$$

Otherwise, different sparse sources would be indistinguishable in the compressed space. **[But this doesn't necessarily imply that the x are also be indistinguishable...it could be that indistinguishable a's always map to the same x's.]** Provided the dimension n of $\mathbf{y}$ satisfies

$$n \geq Ck \log\left(\frac{m}{k}\right), \tag{3}$$

the theory guarantees that with high probability a randomly generated Φ will yield an A satisfying (2). **[What's the assumption on Ψ?]** 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 data.

It is also important for us to describe how our theorems fit in with the others comprising the field of *theoretical dictionary learning*, many of them only very recently published. (local) identifiability of L0 and L1 cost functions, (local) convergence guarantees of greedy and convex algorithms. Our theorem states conditions when one can be sure one is sitting in a global minimum of an L0 norm problem, and an L1 norm problem when we are in a regime where L1 solves L0.

II. Definitions

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

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

Definition 1: 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 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]

$$\Theta(V, W) := \max\left(\sup_{\substack{v \in V \\ \|v\|=1}} d(v, W), \sup_{\substack{w \in W \\ \|w\|=1}} d(w, V)\right). \tag{4}$$

We note the following useful fact [ref: Morris, Lemma 3.3]:

$$\dim(W) = \dim(V) \implies \sup_{\substack{v \in V \\ \|v\|=1}} d(v, W) = \sup_{\substack{w \in W \\ \|w\|=1}} d(w, V). \tag{5}$$

Definition 2: The *spark* of a matrix $A \in \mathbb{R}^{n \times m}$ is the least number of linearly dependent columns:

$$\text{spark}(A) = \min_{x \neq 0} \|x\|_0 \quad \text{such that } Ax = 0. \tag{6}$$

Definition 3: We say that $A \in \mathbb{R}^{n \times m}$ satisfies the (k, α)-*lower-RIP* when for some $\alpha \in (0, 1]$, [ref: Restricted Isometry Property first introduced in "Decoding by linear programming" by Candes and Tao]

$$\|Aa\|_2 \geq \alpha \|a\|_2 \quad \text{for all } k\text{-sparse } a \in \mathbb{R}^m.$$

Definition 4: The *Friedrichs angle* $\theta_F(V, W) \in [0, \frac{\pi}{2}]$ between subspaces V and W of $\mathbb{R}^m$ is the minimal angle formed between unit vectors in $V \cap (V \cap W)^\perp$ and $W \cap (W \cap V)^\perp$, that is

$$\cos\left[\theta_F(V, W)\right] := \max\left\{\frac{\langle v, w \rangle}{\|v\|\|w\|} : v \in V \cap (V \cap W)^\perp, w \in W \cap (V \cap W)^\perp\right\} \tag{7}$$

In our proof we make use of the following quantity defined for a sequence $V_1, \ldots, V_p$ of closed subspaces of $\mathbb{R}^m$:

$$c(V_1, \ldots, V_p) := 1 - \left[1 - \prod_{i=1}^{p-1} \left(1 - \cos^2\left[\theta_F(V_i, \cap_{j=i+1}^p V_j)\right]\right)\right]^{1/2} \tag{8}$$

Since we will be solely working with subspaces spanned by subsets of columns of a matrix $A \in \mathbb{R}^{n \times m}$, we make the following definition for notational convenience. Given a sequence of supports $S_1, \ldots, S_p \in \binom{[m]}{k}$, let

$$c_A(S_1, \ldots, S_p) := c\left(\text{Span}\{A_{S_1}\}, \ldots, \text{Span}\{A_{S_p}\}\right). \tag{9}$$

Definition 5: Let $A \in \mathbb{R}^{n \times m}$. Let $\mathcal{S}$ be the set of all $S_i := \{i, \ldots, (i + (k-1)\} \mod m$ for $i = 0, \ldots, m-1$. We define the following number associated with A:

$$\phi(A) := \min_{\substack{i_1 \neq \ldots \neq i_\ell \in [m] \\ \ell \in \{k, k+1\}}} c_A(S_{i_1}, \ldots, S_{i_\ell}) \tag{10}$$

Definition 6: We say a set of vectors $\mathbf{a}_1, ..., \mathbf{a}_N$ has an *k-spread* of $\delta > 0$ if for any set of k vectors $\mathbf{a}_{i_1}, ..., \mathbf{a}_{i_k}$, the following property holds:

$$\|\sum_{j=1}^{k} c_j \mathbf{a}_{i_j}\|_2 \geq \delta \|c\|_1 \quad \forall c = (c_1, ..., c_k) \in \mathbb{R}^m. \tag{11}$$

Theorem 1: Fix positive integers n and $k < m$. There exist $N = mk\binom{m}{k}$ k-sparse vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^m$ such that if $Y = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ is a dataset for which $\|\mathbf{y}_i - A\mathbf{a}_i\|_2 \leq \varepsilon$ for all $i \in \{1, \ldots, N\}$ for some $A \in \mathbb{R}^{n \times m}$ satisfying spark$(A) > 2k$ then, provided ε is small enough, there exists some $C > 0$ for which the following holds: any matrix $B \in \mathbb{R}^{n \times m}$ for which $\|\mathbf{y}_i - B\mathbf{b}_i\| \leq \varepsilon$ for some k-sparse $\mathbf{b}_i \in \mathbb{R}^m$ for all $i \in \{1, \ldots, N\}$ is such that

$$\|(A - BPD)\mathbf{e}_i\| \leq C\varepsilon \tag{12}$$

for some permutation matrix $P \in \mathbb{R}^{m \times m}$ and diagonal matrix $D \in \mathbb{R}^{m \times m}$. Specifically, provided $\varepsilon < \frac{\alpha^2 \delta \rho}{2k\sqrt{2}}$, then (12) holds for $C = \frac{2k}{\alpha \delta \rho}$, where $\delta > 0$ is the spread of the $\mathbf{a}_i$ and A has unit norm columns, k-RIP constant $\alpha \in (0, 1]$ and $\phi(A) = \rho$.

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 $\gamma_1, ..., \gamma_N$ 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 claim that there exists some $\delta > 0$ such that for any set of k vectors $\mathbf{a}_{i_1}, ..., \mathbf{a}_{i_k}$, the following property holds:

$$\|\sum_{j=1}^{k} c_j \mathbf{a}_{i_j}\|_2 \geq \delta \|c\|_1 \quad \forall c = (c_1, ..., c_k) \in \mathbb{R}^m. \tag{13}$$

To see why, consider the compact set $\mathcal{C} = \{c : \|c\|_1 = 1\}$ and the continuous map

$$\phi : \mathcal{C} \to \mathbb{R}$$
$$(c_1, ..., c_k) \mapsto \|\sum_{j=1}^{k} c_j \mathbf{a}_{i_j}\|_2.$$

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 $\delta > 0$ such that the neighbourhood $\{x : x < \delta\} \subseteq \mathbb{R} \setminus \phi(\mathcal{C})$. Hence $\phi(c) \geq \delta$ for all $c \in \mathcal{C}$. The property (13) follows by the association $c \mapsto \frac{c}{\|c\|_1}$ 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 (13) for all of them. (We refer the reader to the Appendix for a lower bound on δ 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 $A \in \mathbb{R}^{n \times m}$ with unit norm columns satisfiying the (k, α)-lower-RIP and for which spark$(A) > 2k$ and 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 some support $S' \in \binom{[m]}{k}$. 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 Span$\{A_S\}$. Hence, fixing $\mathbf{z} \in$ 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 $\mathbf{y} = \sum_{i \in \mathcal{I}} c_i \mathbf{y}_i \in$ Span$\{\mathcal{Y}\}$, we have by the triangle inequality that

$$\|\mathbf{z} - \mathbf{y}\|_2 = \|\sum_{i \in \mathcal{I}} c_i A\mathbf{a}_i - \sum_{i \in \mathcal{I}} c_i \mathbf{y}_i\|_2 \leq \sum_{i \in \mathcal{I}} \|c_i(A\mathbf{a}_i - \mathbf{y}_i)\|_2 = \sum_{i \in \mathcal{I}} |c_i| \|A\mathbf{a}_i - \mathbf{y}_i\|_2 \leq \varepsilon \sum_{i \in \mathcal{I}} |c_i|. \tag{14}$$

The alternate factorization for the $\mathbf{y}_i$ implies (by a manipulation identical to that of (14)) that for $\mathbf{z}' = \sum_{i \in \mathcal{I}} c_i B\mathbf{b}_i \in$ 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

$$\|\mathbf{z} - \mathbf{z}'\|_2 \leq \|\mathbf{z} - \mathbf{y}\|_2 + \|\mathbf{y} - \mathbf{z}'\|_2 = 2\varepsilon \sum_{i \in \mathcal{I}} |c_i|. \tag{15}$$

Since $\text{supp}(\mathbf{a}_i) = S$ for all $i \in \mathcal{I}$ and A satisfies the (k, α)-lower-RIP, we have

$$\|\mathbf{z}\|_2 = \|\sum_{i\in\mathcal{I}}^{k} c_i A\mathbf{a}_i\|_2 = \|A(\sum_{i\in\mathcal{I}} c_i\mathbf{a}_i)\|_2 \geq \alpha\|\sum_{i\in\mathcal{I}}^{k} c_i\mathbf{a}_i\|_2 \geq \alpha\delta\sum_{i\in\mathcal{I}}^{k}|c_i|, \tag{16}$$

where for the last inequality we have applied the property (13). Combining (15) and (16), we see that for all $\mathbf{z} \in \text{Span}\{A_S\}$ there exists some $\mathbf{z}' \in \text{Span}\{B_{S'}\}$ such that $\|\mathbf{z} - \mathbf{z}'\|_2 \leq \frac{2\varepsilon}{\alpha\delta}\|\mathbf{z}\|_2$. It follows that $d(\mathbf{z}, \text{Span}\{B_{S'}\}) \leq \frac{2\varepsilon}{\alpha\delta}$ for all unit vectors $\mathbf{z} \in \text{Span}\{A_S\}$. Hence,

$$\sup_{\substack{\mathbf{z}\in\text{Span}\{A_S\}\\ \|\mathbf{z}\|=1}} d(\mathbf{z}, \text{Span}\{B_{S'}\}) \leq \frac{2\varepsilon}{\alpha\delta}. \tag{17}$$

Suppose $\varepsilon < \frac{\alpha\delta}{2}$. Then $\dim(\text{Span}\{B_{S'}\}) \geq \dim(\text{Span}\{A_S\}) = k$ by Lemma 4 and the fact that every k columns of A are linearly independent . In fact, since $|S'| = k$, we have $\dim(\text{Span}\{B_{S'}\}) = \dim(\text{Span}\{A_S\})$. Recalling (5), we see the association $S \mapsto S'$ thus defines a map $\pi : \{S_1, \ldots, S_m\} \to \binom{[m']}{k}$ satisfying $\Theta(\text{Span}\{A_S\}, \text{Span}\{\mathcal{B}_{\pi(\mathcal{S})}\}) \leq \frac{2\varepsilon}{\alpha\delta}$.

Suppose further that $\varepsilon < \frac{\alpha^2\delta\rho}{2k\sqrt{2}}$. Since $\alpha < 1$ and $\rho < 1$, we then indeed have $\varepsilon < \frac{\alpha\delta}{2}$ so that

$$\Theta(\text{Span}\{A_{S_i}\}, \text{Span}\{\mathcal{B}_{\pi(S_i)}\}) \leq \frac{\rho}{k}\Delta \quad \text{for all} \quad i \in [m], \tag{18}$$

where $\Delta = \frac{2k\varepsilon}{\alpha\delta\rho} < \frac{\alpha}{\sqrt{2}}$. 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 $i \in \{1, \ldots, m\}$, $\|(A - BPD)e_i\|_2 \leq C\varepsilon$ for $C = \frac{2k}{\alpha\delta\rho}$. ■

Lemma 1 (Main Lemma): Fix positive integers n and $k < m$. Let $A, B \in \mathbb{R}^{n\times m}$ and suppose that A has unit norm columns, satisfies the (k, α)-lower-RIP, and that spark$(A) > 2k$. Let $\mathcal{S}$ 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 $\pi : \mathcal{S} \to \binom{[m]}{k}$ and some $\Delta < \frac{\alpha}{\sqrt{2}}$ such that

$$\Theta(\text{Span}\{A_{S_i}\}, \text{Span}\{B_{\pi(S_i)}\}) \leq \frac{\rho}{k}\Delta \quad \text{for all} \quad i \in \{0, \ldots, m-1\} \tag{19}$$

where

$$\rho := \min_{\substack{i_1\neq\ldots\neq i_\ell\in[m]\\ \ell\in\{k,k+1\}}} c_A(S_{i_1}, \ldots, S_{i_\ell}) \tag{20}$$

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

$$\|(A - BPD)e_i\|_2 \leq \Delta \quad \text{for all} \quad i \in [m]. \tag{21}$$

Proof of Lemma 1: We assume $k \geq 2$ since the case $k = 1$ is contained in Lemma 3. We begin by proving some useful facts. First, we note that for any $S \in \mathcal{S}$, Lemma 4 applied to (19) implies that $\dim(\text{Span}\{B_{\pi(S)}\}) = k$, i.e. the columns of $B_{\pi(S)}$ are linearly independent for all $S \in \mathcal{S}$. Next, consider any set of ℓ distinct $S_{i_1}, \ldots, S_{i_\ell} \in \mathcal{S}$ for $\ell \in \{k, k+1\}$. Note that condition (19) implies that for all unit vectors $\mathbf{z} \in \text{Span}\{B_{\cap_{j=1}^{\ell}\pi(S_{i_j})}\} \subseteq \cap_{j=1}^{\ell}\text{Span}\{B_{\pi(S_{i_j})}\}$ we have $d(\mathbf{z}, \text{Span}\{A_{S_{i_j}}\}) \leq \frac{\rho}{\ell}\Delta$ for all $j = 1, \ldots, \ell$. It follows by Lemmas 5 and 6 that

$$d\left(\mathbf{z}, \text{Span}\{A_{\cap_{j=1}^{\ell}S_{i_j}}\}\right) \leq \Delta\left(\frac{\rho}{c_A(S_{i_1}, \ldots, S_{i_\ell})}\right) \leq \Delta \tag{22}$$

and by Lemma 4, since $\Delta < 1$, that $\dim(\text{Span}\{B_{\cap_{j=1}^{\ell}\pi(S_{i_j})}\}) \leq \dim(\text{Span}\{A_{\cap_{j=1}^{\ell}S_{i_j}}\})$. Since the columns of $B_{\pi(S)}$ are linearly independent for all $S \in \mathcal{S}$, it follows that

$$|\cap_{j=1}^{\ell}\pi(S_{i_j})| \leq |\cap_{j=1}^{\ell} S_{i_j}|. \tag{23}$$

Given these facts, we are now in a position to construct a map which satisfies the properties required by Lemma 3. Fix $i \in [m]$ and let $h(i) = (i - k + 1) \mod m$. Then $\cap_{j=h(i)}^{i} S_j = \{i\}$. (In general, there may exist combinations of fewer supports S_j with intersection $\{i\}$, e.g. if $m \geq 2k - 1$ then $S_{h(i)} \cap S_i = \{i\}$. For brevity, we consider a construction that is valid for any $m > k$.) By (23) we have that $\cap_{j=h(i)}^{i}\pi(S_j)$ is either empty or it contains a single element. Lemma 2 ensures that the latter case is the only possibility. Thus the association $i \mapsto \cap_{j=h(i)}^{i}\pi(S_j)$ defines a map $\hat{\pi} : [m] \to [m]$. Recalling (5), it follows from (22) that for all unit vectors $\mathbf{z} \in \text{Span}\{A_i\}$ we have $d\left(\mathbf{z}, \text{Span}\{B_{\hat{\pi}(i)}\}\right) \leq \Delta$. Since $\Delta < \frac{\alpha}{\sqrt{2}}$, the result follows by Lemma 3. ■

Lemma 2: Let $\mathcal{S}$ be the set of all $S_i := \{i, \ldots, (i + (k-1)\} \mod m$ for $i = 0, \ldots, m-1$ and suppose there exists a map $\pi : \mathcal{S} \to \binom{[\mathbb{Z}/m\mathbb{Z}]}{k}$ such that for $k' \in \{k, k+1\}$,

$$|\cap_{\ell=1}^{k'}\pi(S_{i_\ell})| \leq |\cap_{\ell=1}^{k'} S_{i_\ell}| \tag{24}$$

for any set of distinct $i_1, \ldots, i_{k'} \in [m]$. Then $|\pi(S_v) \cap \cdots \cap \pi(S_{v+(k-1)})| = 1$ for all $v \in \mathbb{Z}/m\mathbb{Z}$.

Proof of Lemma 2: Consider the set $T_m = \{(i,j) : j \in \mathbb{Z}/m\mathbb{Z}, i \in \pi(S_j)\}$, which has km elements. By the pigeon-hole principle, there is some $p \in \mathbb{Z}/m\mathbb{Z}$ and k distinct $j_1, \ldots, j_k$ such that $\{(p,j_1), \ldots, (p,j_k)\} \subseteq T_m$. Hence, $p \in \pi(S_{j_1}) \cap \cdots \cap \pi(S_{j_k})$ and by (24) we must have $S_{j_1} \cap \cdots \cap S_{j_k} \neq \emptyset$. In particular, $j_1, \ldots, j_k$ must be consecutive modulo $\mathbb{Z}/m\mathbb{Z}$, i.e. there exists some $v \in \mathbb{Z}/m\mathbb{Z}$ such that $\{i_1, \ldots, i_k\} = \{v-(k-1), \ldots, v\}$ and $S_{j_1} \cap \cdots \cap S_{j_k} = \{v\}$. Hence $\pi(S_{j_1}) \cap \cdots \cap \pi(S_{j_k}) = \{p\}$ by (24). Furthermore, we can be sure there exists no additional $j^* \in \mathbb{Z}/m\mathbb{Z}$, $j^* \neq j_1 \neq \ldots \neq j_k$ such that $p \in \pi(S_{j^*})$, since we would then have $\pi(S_{j^*}) \cap \pi(S_{j_1}) \cap \cdots \cap \pi(S_{j_k}) \neq \emptyset$ and (24) would imply $S_{j^*} \cap S_{j_1} \cap \cdots \cap S_{j_k} \neq \emptyset$, which is impossible given $\mathcal{S}$. Now, let $T_{m-1} \subset T_m$ be the set of elements of T_m not having p as a first coordinate. By the preceding argument, we have $|T_{m-1}| = mk - k$ and the proof follows by iterating this procedure. ■

Lemma 3: Fix positive integers n and $m \leq m'$. Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times m'}$ have unit norm columns and suppose that A satsifies the $(2,\alpha)$-lower-RIP. If there exists a map $\pi : \{1, \ldots, m\} \to \{1, \ldots, m'\}$ and some $\Delta < \frac{\alpha}{\sqrt{2}}$ such that

$$d\left(Ae_i, \text{Span}\{B\mathbf{e}_{\pi(i)}\}\right) \leq \Delta \quad \text{for all} \quad i \in [m] \tag{25}$$

then there exist a partial permutation matrix $P \in \mathbb{R}^{m' \times m'}$ and diagonal matrix $D \in \mathbb{R}^{m' \times m}$ such that $\|(A - BPD)\mathbf{e}_i\|_2 \leq \Delta$ for all $i \in \{1, \ldots, m\}$.

Proof of Lemma 3: We will show that π is injective and thus defines a permutation when its codomain is restricted to its image. First, note that (25) implies that there exist $c_1, \ldots, c_m \in \mathbb{R}$ such that

$$\|Ae_i - c_i Be_{\pi(i)}\| \leq \Delta \quad \text{for all} \quad i \in [m]. \tag{26}$$

Now 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 $\|A\mathbf{e}_j - c_j B\mathbf{e}_\ell\|_2 \leq \Delta$. (Note that $c_i \neq 0$ and $c_j \neq 0$ since A has unit norm columns and $\alpha \leq 1$.) Summing and scaling these two inequalities by c_j and c_i, respectively, we apply the triangle inequality and the $(2,\alpha)$-lower-RIP on A to yield

$$\begin{aligned} \alpha\|c_j e_i + c_i e_j\|_2 &\leq \|c_j Ae_i + c_i Ae_j\|_2 && (27)\\ &\leq |c_j|\|A\mathbf{e}_i - c_i B\mathbf{e}_\ell\|_2 + |c_i|\|c_j B\mathbf{e}_\ell - A\mathbf{e}_j\|_2 && (28)\\ &\leq (|c_i| + |c_j|)\Delta && (29)\end{aligned}$$

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 and the matrix $P \in \mathbb{R}^{m' \times m'}$ whose i-th column is $e_{\pi(i)}$ for all $1 \leq i \leq m$ and $\mathbf{0}$ for all $m < i \leq m'$ is a partial permutation matrix. Letting $D \in \mathbb{R}^{m' \times m}$ be the diagonal matrix with diagonal elements $c_1, ..., c_m$, (26) becomes $||(A - BPD)\mathbf{e}_i|| \leq \Delta$ for all $i \in [m]$. ■

Lemma 4: Let V, W be closed subspaces of $\mathbb{R}^m$. If $d(v, W) < \|v\|_2$ for all $v \in V$ then $\dim(V) \leq \dim(W)$.

Proof of Lemma 4: The condition implies that for all $v \in V$ there exists some $w \in W$ such that

$$\|v - w\|_2 < \|v\|_2. \tag{30}$$

If $\dim(V) > \dim(W)$ then there exists some $v' \in V \cap W^\perp$. By Pythagoras' Theorem, $\|v' - w\|_2^2 = \|v'\|_2^2 + \|w\|_2^2 \geq \|v\|_2^2$ for all $w \in W$, which is in contradiction with (30). ■.

Lemma 5: Let $M \in \mathbb{R}^{n \times m}$. If every $2k$ columns of M are linearly independent, then for any $\mathcal{S} \subseteq \binom{[m]}{k}$,

$$y \in \text{Span}\{M_{\cap \mathcal{S}}\} \iff y \in \bigcap_{S \in \mathcal{S}} \text{Span}\{M_S\}. \tag{31}$$

Proof of Lemma 5: The forward direction is trivial; we prove the reverse direction by induction. Enumerate $\mathcal{S} = (S_1, \ldots, S_{|\mathcal{S}|})$ and let $y \in \bigcap_i \text{Span}\{M_{S_i}\}$. Then for all $S_i \in \mathcal{S}$ 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 $\cap_{i=1}^{\ell-1} S_i$, $\ell \leq |\mathcal{S}|$ 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. $\text{supp}(x) \subseteq \cap_{i=1}^{\ell} S_i$. ■

Lemma 6: For $p \geq 2$ let $V_1, \ldots, V_p$ be closed linear subspaces of $\mathbb{R}^m$, let $V = \cap_{i=1}^p V_i$. Then

$$\|x - \Pi_V x\| \leq \frac{1}{c(V_1, \ldots, V_p)} \sum_{i=1}^{p} \|x - \Pi_{V_i} x\| \quad \text{for all} \quad x \in \mathbb{R}^m. \tag{32}$$

Proof of Lemma 6: We will prove by induction on p that

$$\|x - \Pi_V x\| = \sum_{i=1}^{p} \|x - \Pi_{V_i} x\| + \|\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_p} x - \Pi_V x\| \tag{33}$$

and show that (32) follows from this. Fix $x \in \mathbb{R}^m$ and suppose $p = 2$. Then since $\Pi_V x \in V_1$ for all $x \in \mathbb{R}^m$, we have by the triangle inequality that

$$\begin{aligned} \|x - \Pi_V x\| &= \|x - \Pi_{V_1} x\| + \|\Pi_{V_1} x - \Pi_{V_1}\Pi_{V_2} x\| + \|\Pi_{V_1}\Pi_{V_2} x - \Pi_V x\| && (34)\\ &\leq \|x - \Pi_{V_1} x\| + \|x - \Pi_{V_2} x\| + \|\Pi_{V_1}\Pi_{V_2} x - \Pi_V x\|, && (35)\end{aligned}$$

where we have used the fact that $\|\Pi\| \leq 1$ for any projection operator Π. Suppose now that

$$\|x - \Pi_V x\| = \sum_{i=1}^{p-1} \|x - \Pi_{V_i} x\| + \|\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_{p-1}}x - \Pi_V x\| \tag{36}$$

Applying the triangle inequality, we have

$$\|x - \Pi_V x\| = \sum_{i=1}^{p-1} \|x - \Pi_{V_i} x\| + \|\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_{p-1}}x - \Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_p}x\| + \|\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_p}x - \Pi_V x\| \tag{37}$$

$$\leq \sum_{i=1}^{p} \|x - \Pi_{V_i} x\| + \|\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_p}x - \Pi_V x\| \tag{38}$$

which proves (33). We now make use of the following result by [Deutsch, "Best Approximation in Inner Product Spaces, Theorem 9.33 "]:

$$\|\left(\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_\ell}\right)x - \Pi_V x\| \leq (1 - c(V_1,\ldots,V_p))\,\|x\| \quad \text{for all} \quad x \in \mathbb{R}^m \tag{39}$$

$$\begin{aligned}\|(\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_k})(x - \Pi_V x) - \Pi_V(x - \Pi_V x)\| &= \|(\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_k})x - (\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_k})\Pi_V x - \Pi_V x + \Pi_V^2 x\| \\ &= \|(\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_\ell})x - \Pi_V x\|, \end{aligned} \tag{40}$$

We then have by (39) and (40) that

$$\begin{aligned}\|(\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_\ell})x - \Pi_V x\| &= \|(\Pi_{V_1}\Pi_{V_2}\cdots\Pi_{V_\ell})(x - \Pi_V x) - \Pi_V(x - \Pi_V x)\| \\ &\leq (1 - c(V_1,\ldots,V_p))\|x - \Pi_V x\|\end{aligned}$$

It follows from this and (33) that

$$\|x - \Pi_V x\| \leq \sum_{i=1}^{p} \|x - \Pi_{V_i} x\| + (1 - c(V_1,\ldots,V_p))\|x - \Pi_V x\|$$

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

III. Appendix

Lemma 7: Let $\gamma_1 < ... < \gamma_N$ 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 $S \in \binom{[N]}{k}$,

$$\|V_S x\|_2 > \rho\|x\|_1 \quad \text{where} \quad \rho = \frac{\delta^k}{\sqrt{k}}\left(\frac{k-1}{k}\right)^{\frac{k-1}{2}} \prod_{i=1}^{k}(\gamma_1 + (i-1)\delta) \quad \text{for all} \quad x \in \mathbb{R}^k \tag{41}$$

Proof of Lemma 7: The determinant of the Vandermonde matrix is

$$\det(V) = \prod_{1\leq j\leq k} \gamma_j \prod_{1\leq i\leq j\leq k} (\gamma_j - \gamma_i) \geq \delta^k \prod_{i=1}^{k}(\gamma_1 + (i-1)\delta). \tag{42}$$

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]:

$$\sigma_{\min}(M) > \left(\frac{k-1}{k}\right)^{\frac{k-1}{2}} |\det M| \tag{43}$$

and the result follows. ■