# Optimally Sparse Representation in General (non-Orthogonal) Dictionaries via $\ell^1$ Minimization

David L. Donoho* and Michael Elad†

**Classification**: Physical Sciences - Engineering

**Manuscript Information**: Number of pages (including this)- 19,
Number of words in the abstract - 246, Number of characters (including spaces) - 41509.

*Corresponding author: Department of Statistics, Room 128, Sequoia Hall, Stanford University, Stanford, CA 94305, USA, Email: donoho@stat.stanford.edu Voice: (650) 723-3350 Fax: (650) 725-8977.

†Department of Computer Science (SCCM), Stanford University, Stanford, CA 94305-9025, USA.

**Abstract**

Given a 'dictionary' $\mathbf{D} = \{\underline{d}_k\}$ of vectors $\underline{d}_k$, we seek to represent a signal $\underline{S}$ as a linear combination $\underline{S} = \sum_k \gamma(k)\underline{d}_k$, with scalar coefficients $\gamma(k)$. In particular, we aim for the *sparsest* representation possible. In general, this requires a combinatorial optimization process. Previous work considered the special case where $\mathbf{D}$ is an overcomplete system consisting of exactly two orthobases, and has shown that, under a condition of mutual incoherence of the two bases, and assuming that $\underline{S}$ has a sufficiently sparse representation, this representation is unique and can be found by solving a *convex* optimization problem: specifically, minimizing the $\ell^1$ norm of the coefficients $\underline{\gamma}$.

In this paper, we obtain parallel results in a more general setting, where the dictionary $\mathbf{D}$ can arise from two or several bases, frames, or even less structured systems. We introduce the *Spark*, a measure of linear dependence in such a system; it is the size of the smallest linearly dependent subset $(\underline{d}_k)$. We show that, when the signal $\underline{S}$ has a representation using less than $\mathrm{Spark}(\mathbf{D})/2$ nonzeros, this representation is necessarily unique. We develop bounds on the Spark that reproduce uniqueness results given in special cases considered earlier. We also show that in the general dictionary case, any sufficiently sparse representation of $\underline{S}$ is also the unique minimal $\ell^1$-representation, using sparsity thresholds related to our Spark bounds.

We sketch three applications: separating linear features from planar ones in 3-D data, noncooperative multiuser encoding, and identification of overcomplete independent components models.

**Keywords:** Sparse Representation, Overcomplete Representation. Matching Pursuit, Basis Pursuit. Convex Optimization, Linear Programming.

# 1 Introduction

Workers throughout engineering and the applied sciences frequently wish to represent data (signals, images) in the most parsimonious terms. In signal analysis specifically, they often consider models proposing that the signal of interest is sparse in some transform domain, such as the wavelet or Fourier domain [18]. However, there is a growing realization that many signals are mixtures of diverse phenomena, and no single transform can be expected to describe them well; instead, we should consider models making sparse combinations of generating elements from several different transforms [3, 4, 18, 25]. Unfortunately, as soon as we start considering general collections of generating elements, the attempt to find sparse solutions enters mostly uncharted territory, and one expects at best to use plausible heuristic methods [1, 5, 16, 23] and certainly to give up hope of rigorous optimality. In this paper, we will develop some rigorous results showing that it can be possible to find optimally sparse representations by efficient techniques in certain cases.

Suppose we are given a dictionary $\mathbf{D}$ of generating elements $\{\underline{d}_k\}_{k=1}^L$, each one a vector in $\mathbf{C}^N$, which we assume normalized: $\underline{d}_k^H \underline{d}_k = 1$. The dictionary $\mathbf{D}$ can be viewed a matrix of size $N \times L$, with generating elements for columns. We do not suppose any fixed relationship between $N$ and $L$. In particular, the dictionary can be overcomplete and contain linearly dependent subsets, and in particular need not be a basis. As examples of such dictionaries, we can mention: wavelet packets and cosine packets dictionaries of Coifman and Meyer [4], which contain $L = N \log(N)$ elements, representing transient harmonic phenomena with a variety of durations and locations; wavelet frames, such as the directional wavelet frames of Ron and Shen [21], which contain $L = CN$ elements for various constants $C > 1$; and the combined Ridgelet/Wavelet systems of Starck, Candès, and Donoho [23, 24]. Faced with such variety, we cannot call individual elements in the dictionary 'basis elements'; we will use the term *atom* instead—see [3, 19] for further elaboration on the atoms/dictionary terminology and other examples.

Given a signal $\underline{S} \in \mathbf{C}^N$, we seek the sparsest coefficient vector $\underline{\gamma}$ in $\mathbf{C}^L$ such that $\mathbf{D}\underline{\gamma} = \underline{S}$. Formally, we aim to solve the optimization problem

$$(P_0) \qquad \text{Minimize} \quad \|\underline{\gamma}\|_0 \qquad \text{subject to} \qquad \underline{S} = \mathbf{D}\underline{\gamma}. \tag{1}$$

Here the $\ell^0$ norm $\|\underline{\gamma}\|_0$ is simply the number of nonzeros in $\underline{\gamma}$. While this problem is easily solved if there happens to be a unique solution $\mathbf{D}\underline{\gamma} = \underline{S}$ among general (not necessarily sparse) vectors $\underline{\gamma}$, such uniqueness does not hold in any of our applications; in fact these mostly involve $L \gg N$, where such uniqueness is of course impossible. In general, solution of $(P_0)$ requires enumerating subsets of the dictionary looking for the smallest subset able to represent the signal; of course, the complexity of such a subset search grows exponentially with $L$.

It is now known that in several interesting dictionaries, highly sparse solutions can be obtained by *convex* optimization; this has been found empirically [2, 3] and theoretically [6, 9, 10]. Consider replacing the $\ell^0$ norm in $(P_0)$ by the $\ell^1$-norm, getting the minimization problem

$$(P_1) \qquad \text{Minimize} \quad \|\underline{\gamma}\|_1 \qquad \text{subject to} \qquad \underline{S} = \mathbf{D}\underline{\gamma}. \tag{2}$$

This can be viewed as a kind of convexification of $(P_0)$. For example, over the set of signals that can be generated with coefficient magnitudes bounded by 1, the $\ell^1$ norm is the largest convex

function less than the $\ell^0$ norm; $(P_1)$ is in a sense the 'closest' convex optimization problem to $(P_0)$.

Convex optimization problems are extremely well-studied [12], and this study has led to a substantial body of algorithmic knowledge and high-quality software. In fact $(P_1)$ can be cast as a linear programming problem and solved by modern interior point methods [3], even for very large $N$ and $L$.

Given the tractability of $(P_1)$ and the general intractibility of $(P_0)$, it is perhaps surprising that for certain dictionaries solutions to $(P_1)$, when sufficiently sparse, are the same as the solutions to $(P_0)$. Indeed, results in [6, 9, 10] show that for certain dictionaries, *if there exists a highly sparse solution* to $(P_0)$ then it is identical to the solution of $(P_1)$; and, if we obtain the solution of $(P_1)$ and observe that if *it happens to be sparse* beyond a certain specific threshold, then we know (without checking) that we have also solved $(P_0)$.

Previous work studying this phenomenon [6, 9, 10] considered dictionaries $\mathbf{D}$ built by concatenating two orthobases $\mathbf{\Phi}$ and $\mathbf{\Psi}$, thus giving that $L = 2N$. For example, one could let $\mathbf{\Phi}$ be the Fourier basis and $\mathbf{\Psi}$ be the Dirac basis, so that we are trying to represent a signal as a combination of spikes and sinusoids. In this dictionary, the latest results in [10] show that if a signal is built from fewer than than $.914\sqrt{N}$ spikes and sinusoids, then this is the sparsest possible representation by spikes and sinusoids—i.e., the solution of $(P_0)$—and it is also necessarily the solution found by $(P_1)$.

We are aware of numerous applications where the dictionaries would not be covered by the results just mentioned. These include: cases where the dictionary is overcomplete by more than a factor 2 (i.e. $L > 2N$), and where we have a collection of nonorthogonal elements, for example a frame or other system. In this paper, we consider general dictionaries $\mathbf{D}$ and obtain results paralleling those in [6, 9, 10]. We then sketch applications in the more general dictionaries.

We address two questions for a given dictionary $\mathbf{D}$ and signal $\underline{S}$:

1. **Uniqueness:** Under which conditions is a highly sparse representation necessarily the sparsest possible representation?

2. **Equivalence:** Under which conditions is a highly sparse solution to the $(P_0)$ problem also necessarily the solution to the $(P_1)$ problem?

Our analysis avoids reliance on the orthogonal sub-dictionary structure in [6, 9, 10]; our results are just as strong as those of [6] while working for a much wider range of dictionaries. We base our analysis on the concept of the *Spark* of a matrix—the size of the smallest linearly dependent subset. We show that uniqueness follows whenever the signal $\underline{S}$ is built from fewer than $\mathrm{Spark}(\mathbf{D})/2$ atoms. We develop two bounds on Spark that yield previously known inequalities in the two-orthobasis case but can be used more generally. The simpler bound involves the quantity $M(\mathbf{G})$, the largest off-diagonal entry in the Gram matrix $\mathbf{G} = \mathbf{D}^H\mathbf{D}$. The more general bound involves $\mu_1(\mathbf{G})$, the size of the smallest group of non-diagonal magnitudes arising in a single row or column of $\mathbf{G}$ and having a sum greater than or equal to 1. We have $\mathrm{Spark}(\mathbf{D}) > \mu_1 \geq 1/M$. We also show, in the general dictionary case, that one can determine a threshold on sparsity whereby every sufficiently sparse representation of $\underline{S}$ must be the unique minimal $\ell^1$-representation; our thresholds involve $M(\mathbf{G})$ and a quantity $\mu_{1/2}(\mathbf{G})$ related to $\mu_1$.

We sketch three applications of our results.

- *Separating Lines and Planes.* Suppose we have a data vector $\underline{S}$ representing 3-D volumetric data. It is supposed that the volume contains some number of lines and planes in superposition. Under what conditions can we correctly identify the lines and planes and the densities on each? This is a caricature of a timely problem in extragalactic astronomy [7]. Here, the dictionary to be studied is a collection of nonorthogonal atoms, each one representing a uniform distribution along a digital line or plane in $\{0, \ldots, N-1\}^3$.

- *Robust Multiuser Private Broadcasting.* $J$ different individuals encode information by superposition in a single signal vector $\underline{S}$ that is intended to be received by a single recipient. The encoded vector must look like random noise to anyone but the intended recipient (including the transmitters), it must be decodable perfectly, and the perfect decoding must be robust against corruption of some limited number of entries in the vector. We show that a scheme with these properties can be made by concatenating together $J$ random dictionaries and a Dirac dictionary, with decoding by solving $(P_1)$.

- *Blind Identification of Sources.* Suppose that a random signal $\underline{S}$ is a superposition of independent components taken from an overcomplete dictionary $\mathbf{D}$. Without assuming sparsity of this representation, we cannot in general know which atoms from $\mathbf{D}$ are active in $\underline{S}$ or what there statistics are. However, we show that it is possible, without any sparsity, to determine the activity and higher-order statistics of the sources. In this application, the dictionary is a non-orthogonal collection of $m$-th order tensor products of basis elements, the 'signal' is the $m$-th order cumulant array of $\underline{S}$, and sparsity of the cumulant array emerges automatically for large $m$ even when the original signal is not sparsely represented.

*Note:* After presenting our work publicly we learned of related work about to be published by R. Gribonval and M. Nielsen [14]. Their work initially addresses the same question of obtaining the solution of $(P_0)$ by instead solving $(P_1)$, with results paralleling ours. Later, their work focuses on analysis of more special dictionaries built by concatenating several bases.

This paper is organized as follows: In the next section we briefly recall the main results found in [6, 9, 10] on the Uniqueness and Equivalence Theorems for dictionaries made from two-orthobases. Section 3 then extends the Uniqueness Theorem to arbitrary dictionaries. Section 4 similarly extends the Equivalence Theorem to the general dictionary setting. Applications are discussed in Section 5.

# 2 The Two-Orthobasis Case

Consider a special type of dictionary: the concatenation of two orthobases $\mathbf{\Phi}$ and $\mathbf{\Psi}$, each represented by $N \times N$ unitary matrices; thus $\mathbf{D} = [\mathbf{\Phi}, \mathbf{\Psi}]$. Let $\underline{\phi}_i$ and $\underline{\psi}_j$ $(1 \le i, j \le N)$ denote the elements in the two bases. Following [6] we define the *mutual incoherence* between these two bases by

$$M(\mathbf{\Phi}, \mathbf{\Psi}) = \mathrm{Sup}\{|\langle \underline{\phi}_i, \underline{\psi}_j \rangle|, \quad \forall 1 \le i, j \le N\}.$$

It is easy to show [6, 10] that $1/\sqrt{N} \le M \le 1$; the lower bound is attained by a basis pair such as spikes and sinusoids [6] or spikes and Hadamard-Walsh functions [10]. The upper bound is

attained if at least one of the vectors in $\mathbf{\Phi}$ is also found in $\mathbf{\Psi}$. A condition for uniqueness of sparse solutions to $(P_0)$ can be stated in terms of $M$; the following is proved in [9, 10].

**Theorem 1 - Uniqueness:** *A representation $\underline{S} = \mathbf{D}\underline{\gamma}$ is necessarily the sparsest possible if $\|\underline{\gamma}\|_0 < 1/M$.*

In words, having obtained a sparse representation of a signal, for example by $(P_1)$ or by any other means, if the $\ell^0$ norm of the representation is sufficiently small (below $1/M$), we conclude that this is also the $(P_0)$ solution. In the 'best case' where $1/M = \sqrt{N}$, the sparsity requirement translates into $\|\underline{\gamma}\|_0 < \sqrt{N}$.

A condition for equivalence of the solution to $(P_1)$ with that for $(P_0)$ can also be stated in terms of $M$; see [9, 10].

**Theorem 2 - Equivalence:** *If there exists any sparse representation of $\underline{S}$ satisfying $\|\underline{\gamma}\|_0 < (\sqrt{2} - 0.5)/M$, then this is necessarily the unique solution of $(P_1)$.*

Note the slight gap between the criteria in the above two Theorems. Recently, Feuer and Nemirovsky managed to prove that the bound in the equivalence theorem is sharp, so this gap is unbridgeable [11].

To summarize, if we solve the $(P_1)$ problem and observe that it is sufficiently sparse, we know we have obtained the solution to $(P_0)$ as well. Moreover, if the signal $\underline{S}$ has sparse enough representation to begin with, $(P_1)$ will find it.

These results correspond to the case of dictionaries built from two orthobases. Both Theorems immediately extend in a limited way to non-orthogonal bases $\mathbf{\Phi}$ and $\mathbf{\Psi}$ [6]. We merely replace $M$ by the quantity $\tilde{M}$ in the statement of both results:

$$\tilde{M} = \mathrm{Sup}\left\{ \left| \mathbf{\Phi}^{-1}\mathbf{\Psi} \right|_{i,j}, \left| \mathbf{\Psi}^{-1}\mathbf{\Phi}, \right|_{i,j}, \quad \forall 1 \le i,j \le N \right\}.$$

However, as we may easily have $\tilde{M} \ge 1$, this is of limited value.

# 3 A Uniqueness Theorem for Arbitrary Dictionaries

We return to the setting of the Introduction; the dictionary $\mathbf{D}$ is a matrix of size $N \times L$, with normalized columns $\{\underline{d}_k\}_{k=1}^{L}$. An incoming signal vector $\underline{S}$ is to be represented using the dictionary by $\underline{S} = \mathbf{D}\underline{\gamma}$. Assume that we have two different suitable representations, i.e.

$$\exists \underline{\gamma}_1 \neq \underline{\gamma}_2 \mid \underline{S} = \mathbf{D}\underline{\gamma}_1 = \mathbf{D}\underline{\gamma}_2. \tag{3}$$

Thus the difference of the representation vectors, $\underline{\delta} = \underline{\gamma}_1 - \underline{\gamma}_2$, must be in the null space of the representation: $\mathbf{D}(\underline{\gamma}_1 - \underline{\gamma}_2) = \mathbf{D}\underline{\delta} = 0$. Hence some group of columns from $\mathbf{D}$ must be linearly dependent. To discuss the size of this group we introduce terminology.

**Definition - Spark:** *Given a matrix $\mathbf{A}$ we define $\sigma = Spark(\mathbf{A})$ as the smallest possible number such that there exists a sub-group of $\sigma$ columns from $\mathbf{A}$ that are linearly dependent.*

Clearly, if there are no zero columns in $\mathbf{A}$, then $\sigma \geq 2$, with equality only if two columns from $\mathbf{A}$ are linearly dependent. Note that, although spark and rank are in some ways similar, they are totally different. The rank of a matrix $\mathbf{A}$ is defined as the *maximal* number of columns from $\mathbf{A}$ that are linearly *independent*, and its evaluation is a sequential process requiring $L$ steps. Spark, on the other hand, is the *minimal* number of columns from $\mathbf{A}$ that are linearly *dependent*, and its computation requires a combinatorial process of complexity $2^L$ steps. The matrix $\mathbf{A}$ can be of full rank and yet have $\sigma = 2$. At the other extreme we get that $\sigma \leq \text{Min}\{L, \text{Rank}(\mathbf{A}) + 1\}$.

As an interesting example, let us consider a two-ortho basis case made of spikes and sinusoids. Here $\mathbf{D} = [\mathbf{\Phi}, \mathbf{\Psi}]$, $\mathbf{\Phi} = \mathbf{I}_N$ the identity matrix of size $N \times N$, and $\mathbf{\Psi} = \mathbf{F}_N$ the Discrete Fourier Transform matrix of size $N \times N$. Then, supposing $N$ is a perfect square, using the Poisson Summation formula we can show [6] that there is a group of $2\sqrt{N}$ linearly-dependent vectors in this $\mathbf{D}$, and so $\text{Spark}(\mathbf{D}) \leq 2\sqrt{N}$. As we shall see later, $\text{Spark}(\mathbf{D}) = 2\sqrt{N}$.

We now use the Spark to bound the sparsity of non-unique representations of $\underline{S}$. Let $\sigma = \text{Spark}(\mathbf{D})$. Every non-uniqueness pair $(\underline{\gamma}_1, \underline{\gamma}_2)$ generates $\underline{\delta} = \underline{\gamma}_1 - \underline{\gamma}_2$ in the column null-space; and

$$\mathbf{D}\underline{\delta} = 0 \quad \Longrightarrow \quad \|\underline{\gamma}_1 - \underline{\gamma}_2\|_0 = \|\underline{\delta}\|_0 \geq \sigma. \tag{4}$$

On the other hand the $\ell^0$ pseudo-norm obeys the triangle inequality $\|\underline{\gamma}_1 - \underline{\gamma}_2\|_0 \leq \|\underline{\gamma}_1\|_0 + \|\underline{\gamma}_2\|_0$. Thus, for the two supposed representations of $S$, we must have

$$\|\underline{\gamma}_1\|_0 + \|\underline{\gamma}_2\|_0 \geq \sigma. \tag{5}$$

Formalizing matters:

**Theorem 3** - **Sparsity Bound:** *If a signal $\underline{S}$ has two different representations $\underline{S} = \mathbf{D}\underline{\gamma}_1 = \mathbf{D}\underline{\gamma}_2$, the two representations must have no less than Spark$(\mathbf{D})$ non-zero entries combined.*

From (5), if there exists a representation satisfying $\|\underline{\gamma}_1\|_0 < \sigma/2$, then any other representation $\underline{\gamma}_2$ must satisfy $\|\underline{\gamma}_2\|_0 > \sigma/2$, implying that $\underline{\gamma}_1$ is the sparsest representation. This gives:

**Corollary 4** - **Uniqueness:** *A representation $\underline{S} = \mathbf{D}\underline{\gamma}$ is necessarily the sparsest possible if $\|\underline{\gamma}\|_0 < Spark(\mathbf{D})/2$.*

We note that these results are sharp, essentially by definition.

## 3.1 Lower Bounds on Spark(D)

Let $\mathbf{G} = \mathbf{D}^H\mathbf{D}$ be the Gram matrix of $\mathbf{D}$. As every entry in $\mathbf{G}$ is an inner product of a pair of columns from the dictionary, the diagonal entries are all 1, owing to our normalization of $\mathbf{D}$'s columns; the off-diagonal entries have magnitudes between 0 and 1.

Let $\mathcal{S} = \{k_1, \ldots, k_s\}$ be a subset of indices, and suppose the corresponding columns of $\mathbf{D}$ are linearly independent. Then the corresponding leading minor $G_{\mathcal{S}} = (G_{k_i,k_j} : i, j = 1, ..., s)$ is positive definite [15]. The converse is also true. This proves

**Lemma 5** *$\sigma \leq Spark(\mathbf{D})$ if and only if every $(\sigma - 1) \times (\sigma - 1)$ leading minor of $\mathbf{G}$ is positive definite.*

To apply this criterion, we use the notion of *Diagonal Dominance* [15]: *If a symmetric matrix* $A = (A_{ij})$ *satisfies* $A_{ii} > \sum_{j \neq i} |A_{ij}|$ *for every* $i$*, then* $A$ *is positive definite.* Applying this criterion to all principal minors of the Gram matrix $G$ leads immediately to a lower bound for Spark($G$).

**Theorem 6 - Lower-Bound Using $\mu_1$:** *Given the dictionary* $\mathbf{D}$ *and its Gram matrix* $\mathbf{G} = \mathbf{D}^H\mathbf{D}$*, let* $\mu_1(\mathbf{G})$ *be the smallest integer* $m$ *such that at least one collection of* $m$ *off-diagonal magnitudes arising within a single row or column of* $\mathbf{G}$ *sums at least to* 1*. Then Spark*$(\mathbf{D}) > \mu_1(\mathbf{G})$*.*

Indeed, by hypothesis, every $\mu_1 \times \mu_1$ principal minor has its off-diagonal entries in any single row or column summing to less than 1. Hence every such principal minor is diagonally dominant, and hence every group of $\mu_1$ columns from $\mathbf{D}$ is linearly independent.

For a second bound on Spark, define $M(\mathbf{G}) = \max_{i \neq j} |\mathbf{G}_{i,j}|$, giving a uniform bound on off-diagonal elements. Note that the earlier quantity $M(\mathbf{\Psi}, \mathbf{\Phi})$ defined in the two-orthobasis case agrees with the new quantity. Now consider the implications of $M = M(\mathbf{G})$ for $\mu_1(\mathbf{G})$. Since we have to sum at least $1/M$ off-diagonal entries of size $M$ to reach a cumulative sum equal or exceeding 1, we always have $\mu_1(G) \geq 1/M(\mathbf{G})$. This proves

**Theorem 7 - Lower Bound Using M:** *If the Gram matrix has all off-diagonal entries* $\leq M$*,*

$$\textit{Spark}(\mathbf{D}) > 1/M. \tag{6}$$

Combined with our earlier observations on Spark, we have: *a representation with fewer than* $(1 + 1/M)/2$ *non-zeros is necessarily maximally sparse.* (Note that the requirement $\|\underline{\gamma}\|_0 < (1 + 1/M)/2$ is actually equivalent to $\|\underline{\gamma}\|_0 \leq 1/(2M)$ because $\|\underline{\gamma}\|_0$ and Spark($\mathbf{D}$) are both integers.) This result was found in [6] for the two-orthobasis case, so we see that the same form of result holds in general, with the extended definition of $M$.

## 3.2 Refinements

In general, Theorems 6 and 7 can contain slack, in two ways.

First, there may be additional dictionary structure to be exploited. For example, return to the special two-orthobasis case of spikes and sinusoids: $\mathbf{D} = [\mathbf{\Phi}, \mathbf{\Psi}]$, $\mathbf{\Phi} = \mathbf{I}_N$ and $\mathbf{\Psi} = \mathbf{F}_N$. Here $M = 1/\sqrt{N}$, and so the theorem says that Spark($\mathbf{D}$) $> \sqrt{N}$. In fact the exact value is $2\sqrt{N}$, so there is room to do better. This improvement follows directly from the uncertainty theorem given in [9, 10], leading to Spark($\mathbf{D}$) $\geq 2/M$. The reasoning is general, and applies to any two-orthobasis dictionary, giving

**Theorem 8 - Improved Lower Bound – Two Orthobasis Case:** *Given a two-orthobasis dictionary with mutual coherence value at most* $M$*,* *Spark*$(\mathbf{D}) \geq 2/M$*.*

This bound, together with Corollary 4, gives precisely Theorem 1 in [9, 10]. Returning again to the spikes/sinusoids example, from $M = 1/\sqrt{N}$ we have Spark($\mathbf{D}$) $\geq 2\sqrt{N}$. On the other hand, if $N$ is a perfect square, the spike train example provides a linearly-dependent subset of $2\sqrt{N}$ atoms, and so in fact $\sigma = 2\sqrt{N}$.

Secondly, Theorem 7 can contain slack simply because the metric information in the Gram matrix is not sufficiently expressive on linear dependence issues. We may construct examples where $\mathrm{Spark}(\mathbf{D}) \gg 1/M(\mathbf{G})$. Our bounds measure diagonal dominance rather than positive definiteness, and this is well-known to introduce a gap in a wide variety of settings [15].

For a specific example, consider a two-basis dictionary made of spikes and *exponentially-decaying* sinusoids [6]. Thus, let $\Phi = I_N$ and, for a fixed $\alpha \in (0,1)$, construct the non-orthogonal basis $\Psi_N^{(\alpha)} = \{\psi_k^\alpha\}$, where $\psi_k^{(\alpha)}(i) = \alpha^i \exp\{\sqrt{-1}\frac{2\pi k}{N} i\}$. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be subsets of indices in each basis. Now suppose this pair of subsets generates a linear dependency in the combined representation, so that, for appropriate $N$-vectors of coefficients $\rho$ and $\omega$,

$$0 = \sum_{k \in \mathcal{S}_1} \rho(k)\delta_k(i) + \sum_{k \in \mathcal{S}_2} \omega(k)\psi_k^{(\alpha)}(i), \qquad \forall i.$$

Then, as $\psi_k^{(\alpha)}(i) = \alpha^i \psi_k^{(0)}(i)$ it is equally true that

$$\alpha^{-i} \cdot 0 = \sum_{k \in \mathcal{S}_1} (\alpha^{-k}\rho(k))\delta_k(i) + \sum_{k \in \mathcal{S}_2} \omega(k)\psi_k^{(0)}(i), \qquad \forall i.$$

Hence, there would have been a linear-dependency in the two-orthobasis case $[I_N, F_N]$, with the same subsets and with coefficients $\tilde{\rho}(k) = \alpha^k \rho(k)$, $\forall k$, and $\tilde{\omega}(k) = \omega(k)$. In short, we see that

$$\mathrm{Spark}([I_N, \Psi_N^{(\alpha)}]) = \mathrm{Spark}([I_N, F_N]).$$

On the other hand, picking $\alpha$ close to zero, we get that the basis elements in $\Psi_N^{(\alpha)}$ decay so quickly that the decaying sinusoids begin to resemble spikes; thus $M(I_N, \Psi_N^{(\alpha)}) \to 1$ as $\alpha \to 0$.

This is a special case of a more general phenomenon, *invariance of the Spark under row and column scaling*, combined with *non-invariance of the Gram matrix under row scaling*. For any diagonal nonsingular matrices $W_1$ ($N \times N$) and $W_2$ ($L \times L$), we have

$$\mathrm{Spark}(W_1 \mathbf{D} W_2) = \mathrm{Spark}(\mathbf{D}),$$

while the Gram matrix $\mathbf{G} = \mathbf{G}(\mathbf{D})$ exhibits no such invariance; even if we force normalization so that $diag(\mathbf{G}) = I_L$, we have only invariance against the action of $W_2$ (column scaling), but not $W_1$ (row scaling). In particular, while the positive-definiteness of any particular minor of $\mathbf{G}(W_1\mathbf{D})$ is invariant under such row scaling, the diagonal dominance is not.

## 3.3 Upper Bounds on Spark

Consider a concrete method for evaluating $\mathrm{Spark}(\mathbf{D})$. We define a sequence of optimization problems, for $k = 1, \ldots, L$:

$$(R_k) \qquad \text{Minimize} \quad \|\underline{\gamma}\|_0 \qquad \text{subject to} \qquad \mathbf{D}\underline{\gamma} = \underline{0} \qquad \gamma_k = 1. \tag{7}$$

Letting $\underline{\gamma}_k^{0,*}$ denote the solution of problem $(R_k)$, we get

$$\mathrm{Spark}(\mathbf{D}) = \mathrm{Min}_{1 \le k \le L} \|\underline{\gamma}_k^{0,*}\|_0. \tag{8}$$

On the other hand, the implied $\ell^0$ optimization is computationally intractable. On the other hand, any sequence of vectors obeying the same constraints furnishes an upper bound. To obtain such a sequence, we replace minimization of the $l_0$ norm by the more tractable $\ell^1$ norm. Define the sequence of optimization problems for $k = 1, 2, ..., L$:

$$(Q_k) \qquad \text{Minimize} \quad \|\underline{\gamma}\|_1 \qquad \text{subject to} \qquad \mathbf{D}\underline{\gamma} = \underline{0} \qquad \gamma_k = 1. \tag{9}$$

(The same sequence of optimization problems was defined in [6], who noted the formal identity with the definition of analytic capacities in harmonic analysis, and called them capacity problems.) The problems can in principle be tackled straightforwardly using Linear Programming solvers. Denote the solution of the $(Q_k)$ problem by $\underline{\gamma}_k^{1,*}$. Then, clearly $\|\underline{\gamma}_k^{0,*}\|_0 \leq \|\underline{\gamma}_k^{1,*}\|_0$, so

$$\text{Spark}(\mathbf{D}) \leq \text{Min}_{1 \leq k \leq L} \|\underline{\gamma}_k^{1,*}\|_0. \tag{10}$$

# 4 An Equivalence Theorem for Arbitrary Dictionaries

We now set conditions under which, if a signal $\underline{S}$ admits a highly sparse representation in the dictionary $\mathbf{D}$, the $\ell^1$ optimization problem $(P_1)$ must necessarily obtain that representation.

Let $\underline{\gamma}_0$ denote the unique sparsest representation of $\underline{S}$, so $\mathbf{D}\underline{\gamma}_0 = \underline{S}$. For any other $\underline{\gamma}_1$ providing a representation (i.e. $\mathbf{D}\underline{\gamma}_1 = \underline{S}$) we therefore have $\|\underline{\gamma}_1\|_0 > \|\underline{\gamma}_0\|_0$. To show that $\underline{\gamma}_0$ solves $(P_1)$, we need to show that $\|\underline{\gamma}_1\|_1 \geq \|\underline{\gamma}_0\|_1$.

Consider the following reformulation of $(P_1)$:

$$\underset{\underline{\gamma}_1}{\text{Minimize}} \quad \|\underline{\gamma}_1\|_1 - \|\underline{\gamma}_0\|_1 \quad \text{subject to} \quad \mathbf{D}\underline{\gamma}_1 = \mathbf{D}\underline{\gamma}_0 = \underline{S}. \tag{11}$$

If the minimum is no less than zero, the $\ell^1$ norm for any alternate representation is at least as large as for the sparsest representation. This in turn means that the $(P_1)$ problem admits $\underline{\gamma}_0$ as a solution. If, also, the minimum is uniquely attained, then the minimizer of $(P_1)$ is $\underline{\gamma}_0$.

As in [6, 10], we develop a simple lower bound. Let $\mathcal{S}$ denote the support of $\underline{\gamma}_0$. Writing the difference in norms in (11) as a difference in sums, and relabeling the summations, gives

$$\begin{aligned} \sum_{k=1}^{L} |\gamma_1(k)| - \sum_{k=1}^{L} |\gamma_0(k)| &= \sum_{\mathcal{S}^c} |\gamma_1(k)| + \sum_{\mathcal{S}} \left( |\gamma_1(k)| - |\gamma_0(k)| \right). \\ &\geq \sum_{\mathcal{S}^c} |\gamma_1(k)| - \sum_{\mathcal{S}} \left( |\gamma_1(k) - \gamma_0(k)| \right) \\ &= \sum_{\mathcal{S}^c} |\delta(k)| - \sum_{\mathcal{S}} |\delta(k)|, \end{aligned} \tag{12}$$

where we introduced the vector $\underline{\delta} = \underline{\gamma}_1 - \underline{\gamma}_0$. Note that the right-hand side of this display is positive only if

$$\sum_{\mathcal{S}} |\delta(k)| < \frac{1}{2} \|\underline{\delta}\|_1,$$

i.e., only if at least 50% of the magnitude in $\delta$ is concentrated in $\mathcal{S}$. This motivates a constrained optimization problem:

$$(C_{\mathcal{S}}) \qquad \text{Maximize} \quad \sum_{\mathcal{S}} |\delta(k)| \quad \text{subject to} \quad \mathbf{D}\underline{\delta} = \underline{0}, \quad \|\underline{\delta}\|_1 = 1. \tag{13}$$

This new problem is closely related to (11) and hence also to $(P_1)$. It has this interpretation: if $val(C_{\mathcal{S}}) < \frac{1}{2}$, then every direction of movement away from $\underline{\gamma}_0$ that remains a representation of $\underline{S}$ (i.e., stays in the nullspace) causes a definite increase in the objective function for (11). Recording this formally,

**Lemma 9 - Less than 50% Concentration Implies Equivalence.** *If $supp(\underline{\gamma}_0) \subset \mathcal{S}$, and $val(C_{\mathcal{S}}) < \frac{1}{2}$, then $\underline{\gamma}_0$ is the unique solution of $(P_1)$ from data $\underline{S} = \mathbf{D}\underline{\gamma}_0$.*

On the other hand, if $val(C_{\mathcal{S}}) \geq \frac{1}{2}$, it can happen that $\underline{\gamma}_0$ is not a solution to $(P_1)$.

**Lemma 10** *If $val(C_{\mathcal{S}}) \geq \frac{1}{2}$, there exists a vector $\underline{\gamma}_0$ supported in $\mathcal{S}$, and a corresponding signal $\underline{S} = \mathbf{D}\underline{\gamma}_0$, such that $\underline{\gamma}_0$ is not a unique minimizer of $(P_1)$.*

**Proof.** Given a solution $\underline{\delta}^*$ to problem $(C_{\mathcal{S}})$ with value $\geq \frac{1}{2}$, pick any $\underline{\gamma}_0$ supported in $\mathcal{S}$ with sign pattern arranged so that $\mathrm{sign}(\delta^*(k)\underline{\gamma}_0(k)) < 0$ for $k \in \mathcal{S}$. Then consider $\underline{\gamma}_1$ of the general form $\underline{\gamma}_1 = \underline{\gamma}_0 + t\underline{\delta}^*$, for small $t > 0$. Equality must hold in (12) for sufficiently small $t$, showing that $\|\underline{\gamma}_1\|_1 < \|\underline{\gamma}_0\|_1$ and so $\underline{\gamma}_0$ is not the unique solution of $(P_1)$ when $\underline{S} = \mathbf{D}\underline{\gamma}_0$.

We now bring the Spark into the picture.

**Lemma 11** *Let $\sigma = Spark(\mathbf{D})$. There is a subset $\mathcal{S}^*$ with $\#(\mathcal{S}^*) \leq \sigma/2 + 1$ and $val(C_{\mathcal{S}^*}) \geq \frac{1}{2}$.*

**Proof.** Let $\mathcal{S}$ index a subset of dictionary atoms exhibiting linear dependence. There is a nonzero vector $\underline{\delta}$ supported in $\mathcal{S}$ with $D\underline{\delta} = 0$. Let $\mathcal{S}^*$ be the subset of $\mathcal{S}$ containing the $\lceil \#(\mathcal{S})/2 \rceil$ largest-magnitude entries in $\underline{\delta}$. Then $\#(\mathcal{S}^*) \leq \#(\mathcal{S})/2 + 1$ while

$$\sum_{\mathcal{S}^*} |\delta(k)| \geq \frac{1}{2}\|\underline{\delta}\|_1.$$

It follows necessarily that $val(C_{\mathcal{S}^*}) > \frac{1}{2}$.

In short, the size of Spark($\mathbf{D}$) controls uniqueness not only in the $\ell^0$ problem, but also in the $\ell^1$ problem. No sparsity bound $\|\underline{\gamma}_0\|_0 < s$ can imply $(P_0)$-$(P_1)$ equivalence in general unless $s \leq \mathrm{Spark}(\mathbf{D})/2$.

## 4.1 Equivalence Using $M(\mathbf{G})$

**Theorem 12** *Suppose that the dictionary $\mathbf{D}$ has a Gram matrix $\mathbf{G}$ with off-diagonal entries bounded by $M$. If $\underline{S} = D\underline{\gamma}_0$ where $\|\underline{\gamma}_0\|_0 < (1 + 1/M)/2$, then $\underline{\gamma}_0$ is the unique solution of $(P_1)$.*

We prove this by showing that if $\mathbf{D}\underline{\delta} = 0$, then for any set $\mathcal{S}$ of indices,

$$\sum_{\mathcal{S}} |\delta(k)| \leq \frac{M\#(\mathcal{S})}{M+1} \cdot \|\underline{\delta}\|_1. \tag{14}$$

Then because $\#(\mathcal{S}) < (1 + 1/M)/2$, it is impossible to achieve 50% concentration on $\mathcal{S}$, and so Lemma 9 gives our result.

To get (14), we recall the sequence of capacity problems $(Q_k)$ introduced earlier. By their definition, if $\underline{\delta}$ obeys $\mathbf{D}\underline{\delta} = 0$,

$$|\delta(k)| \le val(Q_k)^{-1} \|\underline{\delta}\|_1.$$

It follows (see also [6]) that for any set $\mathcal{S}$,

$$\sum_{k \in \mathcal{S}} |\delta(k)| \le \left( \sum_{k \in \mathcal{S}} val(Q_k)^{-1} \right) \cdot \|\underline{\delta}\|_1.$$

Result (14) and hence Theorem 12 follow immediately on substituting the following bound for the capacities:

**Lemma 13** *Suppose the Gram matrix has off-diagonal entries bounded by $M$. Then*

$$val(Q_k) \ge (1/M + 1), \qquad k = 1, \ldots, L.$$

To prove the Lemma, we remark that as $D\underline{\delta} = 0$ then also $G\underline{\delta} = D^H D\underline{\delta} = 0$. In particular $(G\underline{\delta})_k = 0$, where $k$ is the specific index mentioned in the statement of the lemma. Now as $\delta_k = 1$ by definition of $(Q_k)$, and as $G_{k,k} = 1$ by normalization, in order that $(G\underline{\delta})_k = 0$ we must have

$$1 = G_{k,k}\delta_k = -\sum_{j \ne k} G_{k,j}\delta_j.$$

Now

$$1 = |\sum_{j \ne k} G_{k,j}\delta_j| \;\; \le \;\; \sum_{j \ne k} |G_{k,j}||\delta_j| \le M \sum_{j \ne k} |\delta_j| = M(\|\underline{\delta}\|_1 - 1).$$

Hence $\|\underline{\delta}\|_1 \ge 1/M + 1$, and the Lemma follows.

## 4.2 Equivalence Using $\mu_{1/2}(\mathbf{G})$

Recall the quantity $\mu_1$ defined in our analysis of the uniqueness problem in Section 3. We now define an analogous concept relevant to equivalence.

**Definition 14** *For $\mathbf{G}$ a symmetric matrix, $\mu_{1/2}(\mathbf{G})$ denotes the smallest number $m$ such that some collection of $m$ off-diagonal magnitudes arising in a single row or column of $\mathbf{G}$ sums at least to $\frac{1}{2}$.*

We remark that $\mu_{1/2}(\mathbf{G}) \le \frac{1}{2}\mu_1(\mathbf{G}) < \mu_1(\mathbf{G})$. Using this notion, we can show:

**Theorem 15** *Consider a dictionary $\mathbf{D}$ with Gram matrix $\mathbf{G}$. If $\underline{S} = D\underline{\gamma}_0$ where $\|\underline{\gamma}_0\|_0 < \mu_{1/2}(\mathbf{G})$, then $\underline{\gamma}_0$ is the unique solution of $(P_1)$.*

Indeed, let $\mathcal{S}$ be a set of size $\#(\mathcal{S}) < \mu_{1/2}(\mathbf{G})$. We show that any vector in the nullspace of $\mathbf{D}$ exhibits less than 50% concentration to $\mathcal{S}$, so that

$$\sum_{\mathcal{S}} |\delta(k)| < \frac{1}{2}\|\underline{\delta}\|_1, \qquad \forall \underline{\delta} : \mathbf{D}\underline{\delta} = 0. \tag{15}$$

The Theorem then follows from Lemma 9.

Note again that $\mathbf{D}\underline{\delta} = 0$ implies $\mathbf{G}\underline{\delta} = 0$. In particular, $(\mathbf{G} - I)\underline{\delta} = -\underline{\delta}$. Let $m = \#\mathcal{S}$ and let $H = (H_{i,k})$ denote the $m \times L$ matrix formed from the rows of $\mathbf{G} - I$ corresponding to indices in $\mathcal{S}$. Then $(\mathbf{G} - I)\underline{\delta} = -\underline{\delta}$ implies

$$\sum_{\mathcal{S}} |\delta(k)| = \|H\underline{\delta}\|_1. \tag{16}$$

As is well known, $H$, viewed as a matrix mapping $\ell_L^1$ into $\ell_m^1$, has its norm controlled by its maximum $\ell^1$ column sum:

$$\|H\underline{x}\|_1 \le |||H|||_{(1,1)} \cdot \|\underline{x}\|_1 \qquad \forall \underline{x} \in R^L,$$

where

$$|||H|||_{(1,1)} = \max_k \sum_{i=1}^{m} |H_{i,k}|.$$

By the assumption that $\mu_{1/2}(\mathbf{G}) > \#\mathcal{S}$,

$$\max_j \sum_{k \in \mathcal{S}} |G_{k,j} - 1_{\{k=j\}}| < \frac{1}{2}.$$

It follows that $|||H|||_{(1,1)} < \frac{1}{2}$. Together with (16) this yields (15) and hence also the Theorem.

# 5 Stylized Applications

We now sketch a few application scenarios in which our results may be of interest.

## 5.1 Separating Points, Lines, and Planes

A timely problem in extragalactic astronomy is to analyze 3-D volumetric data and quantitatively identify and separate components concentrated on 'filaments' and 'sheets' from those scattered uniformly through 3-D space; see [7] for related references, and [23] for some examples of empirical success in performing such separations. It seems of interest to have a theoretical perspective assuring us that, at least in an idealized setting, such separation is possible in principle.

For sake of brevity, we work with specific *algebraic* notions of digital line, digital point, and digital plane. Let $p$ be a prime (e.g. 127, 257) and consider a 'vector' to be represented by a $p \times p \times p$ array $S(i_1, i_2, i_3) = S(\underline{i})$. As representers of digital points we consider the Kronecker sequences $Point_{\underline{k}}(\underline{i}) = 1_{\{k_1=i_1, k_2=i_2, k_3=i_3\}}$ as $\underline{k} = (k_1, k_2, k_3)$ ranges over all triples with $0 \le k_i < p$. We consider a digital plane $Plane_p(\underline{j}, k_1, k_2)$ to be the collection of all triples $(i_1, i_2, i_3)$ obeying $i_{j_3} = k_1 i_{j_1} + k_2 i_{j_2} \bmod p$, and a digital line $Line_p(\underline{i}_o, \underline{k})$ to be the collection of all triples $\underline{i} = \underline{i}_0 + \ell \underline{k} \bmod p,\ \ell = 0, \ldots, p$.

By using properties of arithmetic mod $p$, it is not hard to show the following:

- Any digital plane contains $p^2$ points.
- Any digital line contains $p$ points.
- Any two digital lines are disjoint, or intersect in a single point.
- Any two digital planes are parallel, or intersect in a digital line.
- Any digital line intersects a digital plane not at all, in a single point, or along the whole extent of the digital line.

Suppose now that we construct a dictionary $\mathbf{D}$ consisting of indicators of points, of digital lines and of digital planes. These should all be normalized to unit $l^2$ norm. Let $\mathbf{G}$ denote the Gram matrix. We make the observation that

$$M(\mathbf{G}) = 1/\sqrt{p}.$$

Indeed, if $\underline{d}_1$ and $\underline{d}_2$ are each normalized indicator functions corresponding to subsets $I_1$, $I_2 \subset \mathbf{Z}_p^3$, then

$$\langle \underline{d}_1, \underline{d}_2 \rangle = \frac{|I_1 \cap I_2|}{|I_1|^{1/2} |I_2|^{1/2}}.$$

Using this fact and the above incidence estimates we get the following:

$$\langle Plane_1, Plane_2 \rangle \le \frac{p}{\sqrt{p^2} \cdot \sqrt{p^2}} = 1/p,$$

while

$$\langle Plane, Line \rangle \leq \frac{p}{\sqrt{p^2} \cdot \sqrt{p}} = 1/\sqrt{p},$$

and

$$\langle Line_1, Line_2 \rangle \leq \frac{1}{\sqrt{p} \cdot \sqrt{p}} = 1/p,$$

and

$$\langle Point, Line \rangle \leq \frac{1}{1 \cdot \sqrt{p}} = 1/\sqrt{p},$$

where for example by *Plane* in the above display we mean the indicator function of a digital plane normalized to have $\ell^2$-norm one, and similarly for *Line* and *Point*. We obtain immediately:

**Theorem 16** *If the 3-D array $S$ can be written as a superposition of fewer than $\sqrt{p}$ digital planes, lines, and points, this is the unique sparsest decomposition. If it can be written as a superposition of fewer than $\sqrt{p}/2$ digital planes, lines, and points, this is the unique minimal $\ell^1$ decomposition.*

This shows that there is at least some possibility to separate 3-D data rigorously into unique geometric components. It is in accord with recent experiments in [23] successfully separating data into a relatively few such components.

## 5.2 Robust Multiuser Private Broadcasting

Consider now a stylized problem in private digital communication. Suppose that $J$ individuals are to communicate privately and without coordination over a broadcast channel with an intended recipient. The $j$-th individual is to create a signal $\underline{S}_j$ that is superposed with all the other individuals' signals, producing $\underline{S} = \sum_{j=1}^{J} \underline{S}_j$. The intended recipient gets a copy of $\underline{S}$. The signals $\underline{S}_j$ are encoded in such a way that, to everyone but the intended recipient, encoders included, $\underline{S}$ looks like white Gaussian noise; participating individuals cannot understand each other's messages, but nevertheless the recipient is able to separate $\underline{S}$ unambiguously into its underlying components $\underline{S}_i$ and decipher each such apparent noise signal into meaningful data. Finally, it is desired that all this remain true even if the recipient gets a defective copy of $\underline{S}$ that may be badly corrupted in a few entries.

Here is a way to approach this problem. The intended recipient arranges things in advance so that each of the $J$ individuals has a private 'codebook' $\mathbf{C}_j$ containing $K$ different $N$-vectors whose entries are generated by sampling independent Gaussian white noises. The codebook is known to the $j$-th encoder and to the recipient. An overall dictionary is created by concatenating these together with an identity matrix, producing

$$\mathbf{D} = [I_N, \ \mathbf{C}_1, \ \ \ldots \ , \ \mathbf{C}_J].$$

The $j$-th individual creates a signal by selecting a random subset of $I_j$ indices in $\mathbf{C}_j$ and generating

$$\underline{S}_j = \sum_{i \in I_j} \alpha_i^j \underline{c}_{j,i}.$$

The positions of the indices $i \in I_j$ and coefficients $\alpha_i^j$ are the meaningful data to be conveyed. Now each $\underline{c}_{j,i}$ is the realization of a Gaussian white noise; hence each $\underline{S}_j$ resembles such a white noise. Individuals don't know each other's codebooks, so they are unable to decipher each other's meaningful data (in a sense that can be made precise; compare [22]).

The recipient can extract all the meaningful data simply by performing minimum $\ell^1$ atomic decomposition in dictionary $\mathbf{D}$. This will provide a list of coefficients associated with spikes and with the components associated with each codebook. Provided the participants encode a sparse enough set of coefficients, the scheme works perfectly to separate errors and the components of the individual message.

To see this, note that by using properties of extreme values of Gaussian random vectors (compare [6]) we can show that the normalized dictionary obeys

$$M(\mathbf{D}) \leq \frac{\sqrt{2\log(K \cdot J + N)}}{\sqrt{N}}(1 + \epsilon_N),$$

where $\epsilon_N$ is a random variable that tends to zero in probability as $N \to \infty$. Hence we have:

**Theorem 17** *There is a capacity threshold $C(\mathbf{D})$ such that, if each participant transmits an $\underline{S}_j$ using at most $\#(I_j) \leq C(\mathbf{D})$ coefficients and if there are fewer than $C(\mathbf{D})$ errors in the recipient's copy of $\hat{\underline{S}}$, then the minimum $\ell^1$ decomposition of $\hat{\underline{S}}$ precisely recovers the indices $I_j$ and the coefficients $\alpha_i^j$. We have*

$$(J+1)C(\mathbf{D}) \geq \frac{1}{2M}.$$

An attractive feature of the scheme is that the logarithmic dependence on $K$ means that a considerable 'oversupply' of codebooks and codebook sizes does not dramatically impact performance of the method. Also, the system can be initialized so that a certain number of new individuals can be added to the scheme in the future without coordination with others.

## 5.3 Identification of Overcomplete Blind Sources

Suppose that a statistically independent set of random variables $\underline{X} = (X_i)$ generates an observed data vector according to

$$\underline{Y} = \mathbf{A}\underline{X} \tag{17}$$

where $\underline{Y}$ is the vector of observed data and $\mathbf{A}$ represents the coupling of independent 'sources' to observable 'sensors'. Such models are often called independent components models [17]. We are interested in the case where $\mathbf{A}$ is known, but is overcomplete, in the sense that $L \gg N$ (many more sources than sensors, a realistic assumption). We are not supposing that the vector $\underline{X}$ is highly sparse; hence there is no hope in general to recover uniquely the components $\underline{X}$ generating a single realization $\underline{Y}$. We ask instead whether it is possible to recover statistical properties of $\underline{X}$; in particular higher-order cumulants.

The order 1 and 2 cumulants (mean and variance) are well-known; the order 3 and 4 cumulants (essentially, skewness and kurtosis) are also well-known; for general definitions see [17] and other standard references, such as McCullagh [20]. In general, the $m$-th order joint *cumulant tensor* of a

random variable $\underline{U} \in \mathbf{C}^N$ is an $m$-way array $Cum^m_{\underline{U}}(i_1, ...., i_m)$ with $1 \leq i_m \leq N$. If $\underline{Y}$ is generated in terms of the independent components model (17) then we have an additive decomposition:

$$Cum^m_{\underline{Y}}(\underline{i}) = \sum_k Cum^m_{X_k} \cdot (\underline{a}_k \otimes \underline{a}_k \otimes \ldots \underline{a}_k)\,(\underline{i}),$$

where $Cum^m_{X_k}$ is the $m$-th order cumulant of the scalar random variable $X_k$. Here $(\underline{a}_k \otimes \underline{a}_k \otimes \ldots \underline{a}_k)$ denotes an $m$-way array built from exterior powers of columns of $\mathbf{A}$:

$$(\underline{a}_k \otimes \underline{a}_k \otimes \ldots \underline{a}_k)\,(i_1, \ldots, i_m) = \underline{a}_k(i_1) \cdot \underline{a}_k(i_2) \cdot \ldots \underline{a}_k(i_m).$$

Define then a dictionary $\mathbf{D}^{(m)}$ for $m$-way arrays with $k$-th atom the $m$-the exterior power of $\underline{a}_k$:

$$\underline{d}_k = (\underline{a}_k \otimes \underline{a}_k \otimes \ldots \underline{a}_k)\,, \qquad k = 1, \ldots, L.$$

Now let $\underline{S}^{(m)}$ denote the $m$-order Cumulant array $Cum^m_{\underline{Y}}$. Then

$$\underline{S}^{(m)} = \mathbf{D}^{(m)} \underline{\gamma}^{(m)}, \tag{18}$$

where $\mathbf{D}^{(m)}$ is the dictionary and $\underline{\gamma}^{(m)}$ is the vector of $m$-order scalar cumulants of the independent components random variable $\underline{X}$. If we can uniquely solve this system, then we have a way of learning cumulants of the (unobservable) $\underline{X}$ from cumulants of the (observable) $\underline{Y}$. Abusing notation so that $M(\mathbf{D}) \equiv M(\mathbf{D}^H\mathbf{D})$ etc., we have

$$M(\mathbf{D}^{(m)}) = M(\mathbf{D})^m.$$

In short, if $M(\mathbf{D}) < 1$ then $M(\mathbf{D}^{(m)})$ can be very small for large enough $m$. Hence, even if the $M$-value for determining a not-very-sparse $\underline{X}$ from $\underline{Y}$ is unfavorable, the corresponding $M$-value for determining $Cum^m_{\underline{X}}$ from $Cum^m_{\underline{Y}}$ can be very favorable, for $m$ large enough. Suppose for example that $\mathbf{A}$ is an $N$ by $L$ system with $M(\mathbf{A}) = 1/\sqrt{N}$, but $\underline{X}$ has (say) $N/3$ active components. This is not very sparse; on typical realizations $\|X\|_0 = N/3 \gg \sqrt{N} = 1/M$. In short, unique recovery of $\underline{X}$ based on sparsity will not hold. However, the second-order cumulant array (the covariance) $Cum^2_{\underline{X}}$ still has $N/3$ active components, while $M(\mathbf{D}^{(2)}) = M(\mathbf{D}^{(2)})^2 = 1/N$; as $N/3 < 1/2M$, the cumulant array $Cum^2_{\underline{Y}}$ is sparsely represented and $Cum^2_{\underline{X}}$ is recovered uniquely by $\ell^1$ optimization. Generalizing, we have

**Theorem 18** *Consider any coupling matrix* $\mathbf{A}$ *with* $M < 1$. *For all sufficiently large* $m \geq m^*(L, M)$, *the minimum* $\ell^1$ *solution of the cumulant decomposition problem (18) is unique (and correct!).*

**Acknowledgements:** Partial support from NSF DMS 00-77261, NSF ANI-008584 (ITR), NSF DMS 01-40698 (FRG), DARPA ACMP, DMS 98–72890 (KDI), and AFOSR MURI 95–P49620–96–1–0028.

# References

[1] A.P. Berg & W.B. Mikhael, A survey of mixed transform techniques for speech and image coding, In *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems*, Volume 4, pages 106–109, 1999.

[2] S.S. Chen, *Basis Pursuit*, PhD thesis, Stanford University, November 1995.

[3] S.S. Chen, D.L. Donoho & M.A. Saunders, Atomic decomposition by basis pursuit, *SIAM Review*, Volume 43, number 1, pages 129–59, January 2001.

[4] R. Coifman, Y. Meyer & M.V. Wickerhauser, Adapted waveform analysis, wavelet-packets and applications, In *ICIAM 1991, Proceedings of the Second International Conference on Industrial and Applied Mathematics*, pages 41–50, SIAM, Philadelphia, 1992.

[5] V.E. DeBrunner, L.X. Chen & H.J. Li, Lapped multiple basis algorithms for still image compression without blocking effect, *IEEE Trans. Image Proc.*, Volume 6, pages 1316–1322, September 1997.

[6] D.L. Donoho & X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Transactions on Information Theory*, November 2001, volume 47, number 7, pages 2845–62.

[7] D.L. Donoho, O. Levi, J.L. Starck, & V. Martinez, Multiscale geometric analysis for 3D galaxy catalogs, *Proceeding of SPIE on Astronomical Telescopes and Instrumentations*, Hawaii, August 2002.

[8] D.L. Donoho & P.B. Stark, Uncertainty principles and signal recovery, *SIAM J. on Applied Mathematics*, Volume 49/3, pages 906–931, June 1989.

[9] M. Elad & A.M. Bruckstein, On sparse representations, *International Conference on Image Processing (ICIP)*, Tsaloniky, Greece, November 2001.

[10] M. Elad & A.M. Bruckstein, A generalized uncertainty principle and sparse representation in pairs of $\Re^N$ bases, *IEEE Transactions on Information Theory*, Volume 48, pages 2558–2567, September 2002.

[11] A. Feuer & A. Nemirovsky, On sparse representations in pairs of bases, Accepted to the *IEEE Transactions on Information Theory* in November 2002.

[12] P.E. Gill, W. Murray & M.H. Wright, *Numerical Linear Algebra and Optimization*, Addison-Wesley, 1991.

[13] G.H. Golub & C.F. Van Loan, *Matrix Computations*, Third edition, The John Hopkins University Press, 1996.

[14] R. Gribonval & M. Nielsen, Sparse representations in unions of bases, submitted to the *IEEE Transactions on Information Theory* in November 2002.

[15] R.A. Horn & C.R. Johnson, *Matrix Analysis*, Addison-Wesley, Redwood City, 1991.

[16] X. Huo, *Sparse Image Representation via Combined Transforms*, PhD thesis, Stanford University, 1999.

[17] A. Hyvarinen & J. Karhunen & E. Oja, *Independent Components Analysis.* John Wiley & Sons, 2001.

[18] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, Second Edition, 1998.

[19] S. Mallat & Z. Zhang, Matching pursuit with time-frequency dictionaries, *IEEE Transactions on Signal Processing*, Volume 41, number 12, pages 3397–3415, December 1993.

[20] P. McCullagh, *Tensor Methods in Statistics*, Chapman and Hall, London, 1987.

[21] A. Ron & X. Shen, Compactly supported tight affine spline frames in $L_2(R^d)$, *Math. Comp.* Volume 65(216), pages 1513–1530, February 1996.

[22] N.J.A. Sloane, Encrypting by random rotations, In *Cryptography*, volume 149 of *Lecture Notes in Computer Science*, pages 71–128. Springer, 1983.

[23] J.L. Starck, D.L. Donoho & E.J. Candès, Curvelet transform for astronomical image processing, Accepted to *Astron. Astrophys.* in September 2002.

[24] J.L. Starck, E.J. Candès & D.L. Donoho, Curvelet Transform for Image Denoising, *IEEE Transactions on Signal Processing*, Volume 11, no. 6, pages 670–84, June 2002.

[25] M.V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, Wellesley, 1994.