Polar Deconvolution of Mixed Signals

Zhenan Fan[1] , Halyun Jeong[2] , Babhru Joshi[3] , Michael P. Friedlander[1, 3]

***Abstract*—The signal demixing problem seeks to separate a superposition of multiple signals into its constituent components. This paper studies a two-stage approach that first decompresses and subsequently deconvolves the noisy and undersampled observations of the superposition using two convex programs. Probabilistic error bounds are given on the accuracy with which this process approximates the individual signals. The theory of polar convolution of convex sets and gauge functions plays a central role in the analysis and solution process. If the measurements are random and the noise is bounded, this approach stably recovers low-complexity and mutually incoherent signals, with high probability and with near optimal sample complexity. We develop an efficient algorithm, based on level-set and conditional-gradient methods, that solves the convex optimization problems with sublinear iteration complexity and linear space requirements. Numerical experiments on both real and synthetic data confirm the theory and the efficiency of the approach.**

***Index Terms*—signal demixing, polar convolution, atomic sparsity, convex optimization**

I. Introduction

The signal demixing problem seeks to separate a superposition of signals into its constituent components. In the measurement model we consider, a set of signals $\{x_i^\natural\}_{i=1}^k$ in $\mathbb{R}^n$ are observed through noisy measurements $b \in \mathbb{R}^m$, with $m \leq n$, of the form

$$b = Mx_S^\natural + \eta \quad \text{with} \quad x_S^\natural := \sum_{i=1}^k x_i^\natural. \tag{1}$$

The known linear operator $M : \mathbb{R}^n \to \mathbb{R}^m$ models the acquisition process of the superposition vector $x_S^\natural$. The vector $\eta \in \mathbb{R}^m$ represents noise uncorrelated with the data. This measurement model and its variations are useful for a range of data-science applications, including mixture models [1], [2], blind deconvolution [3], blind source separation [4], and morphological component analysis [5].

A central concern of the demixing problem (1) is to delineate efficient procedures and accompanying conditions that make it possible to recover the constituent signals to within a prescribed accuracy—using the fewest number of measurements m. The recovery of these constituent signals cannot be accomplished without additional information, such as the latent structure in each signal $x_i^\natural$. We build on the general atomic-sparsity framework formalized by Chandrasekaran et al. [6], and assume that each signal $x_i^\natural$ is itself well represented as a superposition of a few atomic signals from a collection $\mathcal{A}_i \subset \mathbb{R}^n$. In other words, the vectors $\{x_i^\natural\}_{i=1}^k$ are paired with atomic sets $\{\mathcal{A}_i\}_{i=1}^k$ that allow the decompositions

$$x_i^\natural = \sum_{a \in \mathcal{A}_i} c_a a, \quad c_a \geq 0 \quad \forall a \in \mathcal{A}_i, \tag{2}$$

where most of the coefficients c_a are zero. This model of atomic sparsity includes a range of important notions of sparsity, such as sparse vectors, which are sparse in the set of canonical vectors, and low-rank matrices, which are sparse in the set of rank-1 matrices with unit spectral norm. Other important generalizations include higher-order tensor decompositions, useful in computer vision [7] and handwritten digit classification [8], and polynomial atomic decomposition [9].

A common approach to recover an atomic signal is to use the gauge function

$$\gamma_{\mathcal{A}}(x) := \inf_{c_a} \left\{ \sum_{a \in \mathcal{A}} c_a \mid x = \sum_{a \in \mathcal{A}} c_a a, \;\; c_a \geq 0 \;\; \forall a \in \mathcal{A} \right\},$$

where $\mathcal{A}$ is the atomic set for x. This gauge function is central to the formulation of convex optimization process that provably leads to solutions that have sparse decompositions in the sense of (2). The properties of gauges and their relationship with atomic sparsity have been well-studied in the literature and are outlined in Chandrasekaran et al. [6] and Fan et al. [10]. The typical approach to the demixing problem is to combine k separate gauge functions, each corresponding to one of the atomic sets $\{\mathcal{A}_i\}_{i=1}^k$, as a weighted sum or similar formulations. We instead combine the k separate gauge functions using a special-purpose convolution operation called polar convolution, that can reflect the additive structure of the superposition, as defined in (1).

A. Polar convolution

For any two atomic sets $\mathcal{A}_1$ and $\mathcal{A}_2$, the polar convolution of the corresponding gauge functions is

$$(\gamma_{\mathcal{A}_1} \diamond \gamma_{\mathcal{A}_2})(x) := \inf_{x_1,\, x_2} \max\{\, \gamma_{\mathcal{A}_1}(x_1),\; \gamma_{\mathcal{A}_2}(x_2) \mid x = x_1 + x_2 \,\}.$$

The resulting function is the gauge to the vector sum $\mathcal{A}_1 + \mathcal{A}_2$, i.e.,

$$\gamma_{\mathcal{A}_1} \diamond \gamma_{\mathcal{A}_2} = \gamma_{\mathcal{A}_1 + \mathcal{A}_2}; \tag{3}$$

see [11, Proposition 6.2]. This convolution operation was first described by Rockafellar [12, Theorem 5.8] as a convexity-preserving operation for general convex functions. Friedlander et al. [11] specialized this operation to the family of gauge functions and describe significant properties that accrue when the convolution is used in that context.

Department of Computer Science, University of British Columbia, Vancouver, BC, Canada

Department of Mathematics, University of California, Los Angeles, California, United States

Department of Mathematics, University of British Columbia, Vancouver, BC, Canada

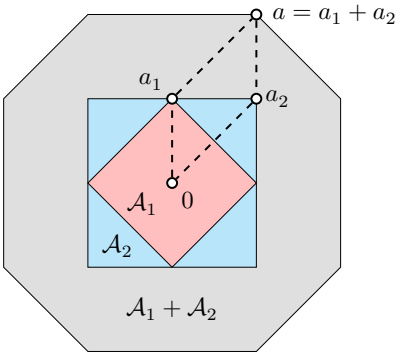


Fig. 1. The sum of two atomic sets. The sum $\mathcal{A}_1 + \mathcal{A}_2$ is the unit level set for the polar convolution $\gamma_{\mathcal{A}_1} \diamond \gamma_{\mathcal{A}_2}$, i.e., $\mathcal{A}_1 + \mathcal{A}_2 = \{\, a \mid \gamma_{\mathcal{A}_1} \diamond \gamma_{\mathcal{A}_2}(a) \leq 1 \,\}$.

The subdifferential properties of polar convolution facilitate our analysis and allow us to build an efficient algorithm that is practical for a range of problems. In particular, the polar convolution decouples under a duality correspondence built around the polarity of convex sets. The polar to a convex set $\mathcal{C} \subset \mathbb{R}^n$,

$$\mathcal{C}^\circ = \{\, y \in \mathbb{R}^n \mid \langle x, y \rangle \leq 1 \text{ for all } x \in \mathcal{C} \,\},$$

contains a dual description of $\mathcal{C}$ in terms of all of its supporting hyperplanes. Under this dual representation,

$$\gamma_{(\mathcal{A}_1 + \mathcal{A}_2)^\circ} = \gamma_{\mathcal{A}_1^\circ} + \gamma_{\mathcal{A}_2^\circ},$$

which implies that the subdifferential decouples as $\partial \gamma_{(\mathcal{A}_1 + \mathcal{A}_2)^\circ} = \partial \gamma_{\mathcal{A}_1^\circ} + \partial \gamma_{\mathcal{A}_2^\circ}$. Thus, a subgradient computation, which is central to all first-order methods for convex optimization, can be implemented using only subdifferential oracles for each of the polar functions $\gamma_{\mathcal{A}_i^\circ}$. We show in Section V how to use this property to implement a version of the conditional gradient method [13], [14] to obtain the polar decomposition using space complexity that scales linearly with the size of the data.

B. Decompression and deconvolution

The principle innovation of our approach to the demixing problem (1) is to decouple the recovery procedure into two stages: an initial *decompression* stage meant to recover the superposition $x_S^\natural$ from the vector of observations b, followed by a *deconvolution* stage that separates the recovered superposition $x_S^\natural$ into its constituent components $\{x_i^\natural\}_{i=1}^k$. We couple the convex theory of polar convolution [11] to the theory of statistical dimension and signal incoherence to derive a recovery procedure and analysis for demixing a compressively sampled mixture to within a prescribed accuracy.

a) Stage 1: Decompression: The initial decompression stage is based on the observation that because each signal $x_i^\natural$ is $\mathcal{A}_i$ sparse, the superposition $x_S^\natural$ must be sparse with respect to the weighted vector sum

$$\mathcal{A}_S := \textstyle\sum_{i=1}^k \lambda_i \mathcal{A}_i \equiv \left\{ \sum_{i=1}^k \lambda_i a_i \;\middle|\; a_i \in \mathcal{A}_i \cup \{0\},\ i \in 1{:}k \right\} \quad (4)$$

of the individual atomic sets $\mathcal{A}_i$. The positive weights λ_i carry information about the relative powers of the individual signals, and serve to equilibrate the gauge values of each signal. Thus, the weights λ_i are defined so that for each $i \in 1{:}k$,

$$\gamma_{\lambda_i \mathcal{A}_i}(x_i^\natural) = \gamma_{\mathcal{A}_1}(x_1^\natural). \quad (5)$$

The initial decompression stage solves the convex optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; \gamma_{\mathcal{A}_S}(x) \;\; \text{subject to} \;\; \|Mx - b\|_2 \leq \alpha, \quad \text{(P1)}$$

where the parameter $\alpha \geq 0$ bounds the acceptable level of misfit between the linear model Mx and the observations b, and correspondingly reflects the anticipated magnitude of the noise η. It follows from (3) that the objective of (P1) is in fact the polar convolution of the individual weighted gauges:

$$\gamma_{\mathcal{A}_S}(x) = \gamma_{\lambda_1 \mathcal{A}_1} \diamond \gamma_{\lambda_2 \mathcal{A}_2} \diamond \cdots \diamond \gamma_{\lambda_k \mathcal{A}_k}(x).$$

Proposition 2 establishes conditions under which the solution x_S^* to (P1) stably approximates the superposition $x_S^\natural$.

b) Stage 2: Deconvolution: The solution x_S^* of the decompression problem (P1) defines the subsequent convex deconvolution problem

$$\underset{x_1, \ldots, x_k}{\text{minimize}} \;\; \max_{i \in 1:k} \gamma_{\lambda_i \mathcal{A}_i}(x_i) \;\; \text{subject to} \;\; \textstyle\sum_{i=1}^k x_i = x_S^* \quad \text{(P2)}$$

to obtain approximations x_i^* to each constituent signal $x_i^\natural$.

In both stages, a variant of the conditional-gradient method provides a computationally and memory efficient algorithm that can be implemented with storage proportional to the number of measurements m [10]. We describe in Section V the details of the method.

C. Prior work

The history of signal demixing can be traced to early work in seismic imaging [15] and morphological component analysis [5], [16], which used 1-norm regularization to separate incoherent signals. More recently, McCoy and Tropp [17], [18] and Oymak and Tropp [19] proposed a unified theoretical framework for signal demixing using modern tools from high-dimensional geometry.

McCoy and Tropp [17] analyzed the recovery guarantees of a convex program that can reconstruct $k = 2$ randomly-rotated signals from a full set of noiseless observations, i.e., $m = n$ and $\|\eta\| = 0$. McCoy and Tropp [18] subsequently extended this framework to demixing $k \geq 2$ randomly-rotated signals from noisy measurements, as modeled by (1). Oymak and Tropp [19] considered a demixing problem similar to (P2) that also incorporates the measurement operator M, and provided guarantees for demixing two unknown vectors from random and noiseless measurements. We build on this line of work by providing explicit recovery error bounds in terms of the complexity of the signal sets and the number of measurements. Our analysis allows for any number of individual signals $k \geq 2$. We postpone to section IV-A a detailed comparison between our results and earlier work.

Early work on demixing sparse signals implicitly assumed some notion of incoherence between representations of the signals. This concept was made concrete by Donoho and Huo [20], and subsequently Donoho and Elad [21],

who measured the mutual incoherence of finite bases via the maximal inner-products between elements of the sets. Related incoherence definitions appear in compressed sensing [22], [23] and robust PCA [24], [25]. In this paper we adopt McCoy and Tropp's [18] notion of incoherence as the minimal angle between conic representation of the individual signals.

D. Contributions and roadmap

Section II shows that the decompression problem (P1) can stably recover $x_S^\natural$. Proposition 1 characterizes the recovery error in terms of the overall complexity of the signal, provided the measurements are random. This result follows directly from Tropp [26] and a conic decomposition property particular to polar convolution.

Section III shows that the deconvolution problem (P2) can stably approximate each $x_i^\natural$. The bound in the recovery error is given in terms of the error in the initial decompression process and the incoherence between signals as measured by the minimum angle between conic representations of each signal; see Proposition 2. This result requires a general notion of incoherence based on the angle between descent cones, first analyzed by McCoy and Tropp [18].

Section IV shows how a random-rotation model yields a particular level of incoherence with high probability; see Proposition 5. We develop the recovery guarantee under the random-rotation model; see Theorem 1. These results are verified numerically in section VI-A.

Section V outlines an algorithm based on conditional-gradient and level-set methods for computing the decompression and deconvolution process. The worst-case computational complexity of this process is sublinear in the required accuracy. Numerical experiments in Section VI on real and synthetic structured signals verify the efficiency of the approach.

The following blanket assumption holds throughout the paper.

Assumption 1 (Measurement model)**.** *The linear model* (1) *satisfies the following conditions: the linear map* $M : \mathbb{R}^n \to \mathbb{R}^m$ *has i.i.d. standard Gaussian entries; the noise vector* η *satisfies* $\|\eta\|_2 \leq \alpha$ *for some scalar* α*; and the relative signal powers* $\{\lambda_i\}_{i=1}^k$ *satisfy* (5)*.*

Proofs of all theoretical results are given in Section A.

II. Decompressing the superposition

As shown in Section I-B, under the assumption that the individual signals $x_i^\natural$ are $\mathcal{A}_i$ sparse, the aggregate signal $x_S^\natural$ is sparse with respect to the aggregate atomic set $\mathcal{A}_S$. Thus, the decompression of the observations b in (1) is accomplished by minimizing the gauge to $\mathcal{A}_S$ to within the bound on the noise level $\|\eta\| \leq \alpha$, as modeled by the recovery problem (P1). Without noise (i.e, $\alpha = 0$), the aggregate signal $x_S^\natural$ is the unique solution to (P1) when the null space of the measurement operator M has only a trivial intersection with the descent cone

$$\mathcal{D}_S := \mathcal{D}(\mathcal{A}_S, x_S^\natural) = \mathsf{cone}\left\{\, d \in \mathbb{R}^n \mid \gamma_{\mathcal{A}_S}(x_S^\natural + d) \leq \gamma_{\mathcal{A}_S}(x_S^\natural) \,\right\},$$

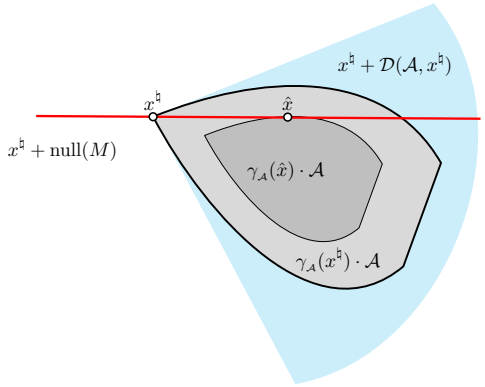


Fig. 2. A non-trivial intersection of $\mathcal{D}(\mathcal{A}, x^\natural)$ and $\mathsf{null}(M)$ is required for successful decompression. The blue shaded region represents the shifted descent cone $x^\natural + \mathcal{D}(\mathcal{A}, x^\natural)$, and red line represents the shifted null space $\mathsf{null}(M) + x^\natural$. If $\mathcal{D}(\mathcal{A}, x^\natural) \cap \mathsf{null}(M) \neq \{0\}$ (as depicted here) then there exists a vector $\hat{x}$ such that $\gamma_{\mathcal{A}}(\hat{x}) < \gamma_{\mathcal{A}}(x^\natural)$ and $M\hat{x} = Mx^\natural$.

where $\mathsf{cone}\,\mathcal{C} := \mathbb{R}_+\mathcal{C}$ is the conic extension of a set $\mathcal{C}$. In other words, $x_S^\natural$ is the unique solution of (P1) if and only if

$$\mathcal{D}_S \cap \mathsf{null}(M) = \{0\}. \tag{6}$$

Figure 2 illustrates the geometry of this optimality condition, and depicts a case in which it doesn't hold.

If the linear operator M is derived from Gaussian measurements, Gordon [27] characterized the probability of the event (6) as a function of the Gaussian width of the descent cone $\mathcal{D}_S$ and the number of measurements m. This result is the basis for recovery guarantees developed by Chandrasekaran et al. [6] and Tropp [26] for a convex formulation similar to (P1).

Intuitively, the number of measurements required for stable recovery of the aggregate $x_S^\natural$ depends on the total complexity of the k constituent $\mathcal{A}_i$-sparse vectors $x_i^\natural$. The complexity is measured in terms of the statistical dimension of each of the descent cones $\mathcal{D}_i$: for any convex cone $\mathcal{D} \subset \mathbb{R}^n$, its statistical dimension is given by

$$\delta(\mathcal{D}) = \mathbb{E}\left[\, \|\, \mathsf{proj}_{\mathcal{D}}(g)\|_2^2 \,\right],$$

where g is standard normal random vector and $\mathsf{proj}_{\mathcal{D}}$ is the orthogonal projection onto the cone $\mathcal{D}$. Tropp [26, Corollary 3.5] established a bound on the recovery error between the solutions of the decompression problem (P1) and the superposition $x_S^\natural$ that depends on the statistical dimension $\delta(\mathcal{D}_S)$ of its descent cone. The following proposition is a restatement of Tropp [26, Corollary 3.5] applied to the decompression problem (P1).

Proposition 1 (Stable decompression of the aggregate)**.** *For any* $t > 0$*, any solution* x^* *of* (P1) *satisfies*

$$\|x^* - x_S^\natural\|_2 \leq 2\alpha \left[\sqrt{m-1} - \sqrt{\delta(\mathcal{D}_S)} - t\right]_+^{-1}$$

with probability at least $1 - \exp(-t^2/2)$*, where* $[\xi]_+ = \max\{0, \xi\}$*.*

The statistical dimension of $\mathcal{D}_S$ is in general challenging to compute. However, we show in section III-A that when all the signals $\{x_i^\natural\}_{i=1}^k$ are incoherent, a reasonable upper bound on $\delta(\mathcal{D}_S)$ can be guaranteed; see Corollary 1.

III. Deconvolving the components

The second stage of our approach is the deconvolution stage which separates the recovered aggregate signal into its constituent components. In order to successfully separate the aggregate $x_S^\natural$ into its components $\{x_i^\natural\}_{i=1}^k$ using the deconvolution problem (P2), additional assumption on dissimilarity between the atomic representations of the individual signals is generally required. For example, it can be challenging to separate the superposition of two sparse signals or two low-rank signals without additional assumptions. We follow McCoy and Tropp [18], and measure the dissimilarity between signal structures—and thus their incoherence—using the angles between corresponding descent cones.

To motivate the incoherence definition, consider the case where there are only $k = 2$ signals $x_1^\natural$ and $x_2^\natural$. If the descent cones $-\mathcal{D}_1$ and $\mathcal{D}_2$ have a nontrivial intersection, then there exists a nonzero direction $d \in -\mathcal{D}_1 \cap \mathcal{D}_2$ such that $\gamma_{\mathcal{A}_1}(x_1^\natural - d) < \gamma_{\mathcal{A}_1}(x_1^\natural)$ and $\gamma_{\mathcal{A}_2}(x_2^\natural + d) < \gamma_{\mathcal{A}_2}(x_2^\natural)$, which contradicts the optimality condition required for $x_1^\natural$ and $x_2^\natural$ to be unique minimizers of (P2). Thus, deconvolution only succeeds if the descent cones have a trivial intersection, which can be characterized using angle between the descent cones. Figure 3 illustrates this geometry.

Obert [28] defined the angle between two cones $\mathcal{K}_1$ and $\mathcal{K}_2$ in $\mathbb{R}^n$ as the minimal angle between vectors in these two cones. It follows that the cosine of the angle between two cones can be expressed as

$$\cos\angle(\mathcal{K}_1, \mathcal{K}_2) = \sup\{\, \langle x, y\rangle \mid x \in \mathcal{K}_1 \cap \mathbb{S}^{n-1},\, y \in \mathcal{K}_2 \cap \mathbb{S}^{n-1} \,\}.$$

For the general case where the number of signals $k \geq 2$, a natural choice for a measure of incoherence between these structured signals is the minimum angle between the descent cone of a signal with respect to the remaining descent cones.

Definition 1. *The pairs $\{(x_i^\natural, \mathcal{A}_i)\}_{i=1}^k$ are β-incoherent with $\beta \in (0, 1]$ if for all $i \in 1{:}k$,*

$$\cos\angle\left(-\mathcal{D}_i, \sum_{j\neq i} \mathcal{D}_j\right) \leq 1 - \beta.$$

We use the incoherence between descent cones to bound the error between the true constituent signals $\{x_i^\natural\}_{i=1}^k$ and the solution set of the deconvolution problem (P2). This bound depends on the accuracy of the approximation x_S^* to the true superposition $x_S^\natural$ and is shown in Proposition 2.

Proposition 2 (Stable deconvolution)**.** *If the pairs $\{(x_i^\natural, \mathcal{A}_i)\}_{i=1}^k$ are β-incoherent for some $\beta \in (0, 1]$, then any set of solutions $\{x_i^*\}_{i=1}^k$ of (P2) satisfies for all $i \in 1{:}k$*

$$\|x_i^* - x_i^\natural\|_2 \leq \|x_S^* - x_S^\natural\|_2/\sqrt{\beta},$$

where x_S^* is any solution of (P1).

In summary, a large angle between negation of a descent cone $-\mathcal{D}_i$ and all the other descent cones—as reflected by a large incoherence constant β—corresponds a small error between each x_i^* and the ground truth $x_i^\natural$.

A. Bound on $\delta(\mathcal{D}_S)$ under incoherence

Proposition 1 gives a stable recovery result for the decompression stage. However, the recovery bound depends on the the statistical dimension of $\mathcal{D}_S$, which is challenging to compute even when the statistical dimension of the individual descent cone $\mathcal{D}_i$ is known. In this section, we show that the incoherence between the structured signals $\{x_i^\natural\}_{i=1}^k$ is sufficient to establish an upper bound for $\delta(\mathcal{D}_S)$. We start with the $k = 2$ case. Proposition 3 shows that if the angle between two cones is bounded, then the statistical dimension of the sum of these two cones is also bounded.

Proposition 3 (Bound on statistical dimension of sum)**.** *Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be two closed convex cones in $\mathbb{R}^n$. If $\cos\angle(-\mathcal{K}_1, \mathcal{K}_2) \leq 1 - \beta$ for some $\beta \in (0, 1]$, then*

$$\sqrt{\delta(\mathcal{K}_1 + \mathcal{K}_2)} \leq \tfrac{1}{\sqrt{\beta}}\left(\sqrt{\delta(\mathcal{K}_1)} + \sqrt{\delta(\mathcal{K}_2)}\right).$$

This result generalizes to an arbitrary number of cones.

Corollary 1 (Bound on statistical dimension of sum under incoherence)**.** *If the pairs $\{(x_i^\natural, \mathcal{A}_i)\}_{i=1}^k$ are β-incoherent for some $\beta \in (0, 1]$, then*

$$\sqrt{\delta(\mathcal{D}_S)} \leq \beta^{-\frac{k-1}{2}} \sum_{i=1}^k \sqrt{\delta(\mathcal{D}_i)}.$$

Corollary 1 shows that when the pairs $\{(x_i^\natural, \mathcal{A}_i)\}_{i=1}^k$ are β-incoherent, $\delta(\mathcal{D}_S)$ can be upper bounded in terms of the statistical dimension of individual descent cones.

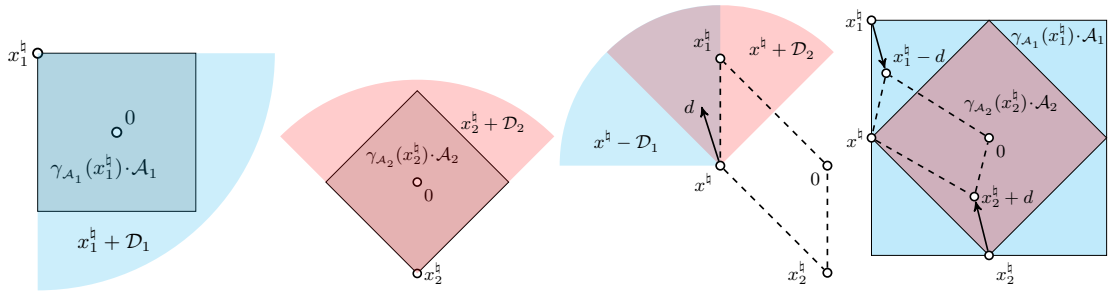
IV. Inducing incoherence through random rotation

Proposition 2 establishes the stability of the deconvolution problem in the case that the unknown signals are β-incoherent, as formalized in Definition 1. However, except in very special cases like randomly rotated signals, it is not feasible to determine the incoherence constant β. We build on McCoy and Tropp's random rotation model [18] to quantify, with high probability, the β-incoherence of k randomly-rotated atomic sparse signals, and present a recovery result for a randomly rotated case.

We first consider a simpler case of two general cones, one of which is randomly rotated. Let $\mathrm{SO}(n)$ denote the special orthogonal group, which consists of all n-by-n orthogonal matrices with unit determinant. The following proposition provides a probabilistic bound on the angle between the two cones in terms of their statistical dimension. This geometric result maybe of intrinsic interest in other contexts.

Proposition 4 (Probabilistic bound under random rotation)**.** *Let Q is drawn uniformly at random from $\mathrm{SO}(n)$. Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be two closed convex cones in $\mathbb{R}^n$. For any $t \geq 0$, we have*

$$\mathbb{P}\left[\cos\angle(\mathcal{K}_1, Q\mathcal{K}_2) \geq \tfrac{3}{\sqrt{n}}\left(\sqrt{\delta(\mathcal{K}_1)} + \sqrt{\delta(\mathcal{K}_2)}\right) + t\right] \leq \exp(-\tfrac{n-2}{8}t^2).$$



(a) $\mathcal{D}_1 \coloneqq \mathcal{D}(\mathcal{A}_1, x_1^\natural)$ (b) $\mathcal{D}_2 \coloneqq \mathcal{D}(\mathcal{A}_2, x_2^\natural)$ (c) $d \in -\mathcal{D}_1 \cap \mathcal{D}_2$ (d) $x_1^\natural - d$ and $x_2^\natural + d$

Fig. 3. The top row depicts two scaled atomic sets $\gamma_{\mathcal{A}_1}(x_i^\natural)\cdot\mathcal{A}_i$ and the corresponding descent cones $x_i^\natural + \mathcal{D}_i$ (shifted to lie at $x_i^\natural$) for $i = 1, 2$. (c) The descent cones shifted to $x^\natural = x_1^\natural + x_2^\natural$, with $\mathcal{D}_1$ negated; the vector d lies in their intersection. (d) The vector d descends on both scaled atomic sets, so that $\gamma_{\mathcal{A}_1}(x_1^\natural - d) < \gamma_{\mathcal{A}_1}(x_1)$ and $\gamma_{\mathcal{A}_2}(x_2^\natural + d) < \gamma_{\mathcal{A}_2}(x_2^\natural)$.

We now assume that the k structured signals $x_i^\natural$ are defined via a random rotations of k underlying structured signals x_i°.

Assumption 2 (Random rotations)**.** *Fix x_i° and $\mathcal{A}_i^\circ$ for $i \in 1{:}k$ such that x_i° is sparse with respect to atomic set $\mathcal{A}_i^\circ$. For each $i \in 1{:}k$, assume*

$$x_i^\natural \coloneqq Q_i x_i^\circ \quad \text{and} \quad \mathcal{A}_i \coloneqq Q_i \mathcal{A}_i^\circ,$$

where the matrices Q_i are drawn uniformly and i.i.d. from $\mathrm{SO}(n)$.

Our next proposition shows that, under mild conditions, randomly rotated structured signals are incoherent with high probability.

Proposition 5. *Suppose that Assumption 2 holds. If $\sum_{i=1}^k \sqrt{\delta(\mathcal{D}_i)} \leq \left(1 - 4^{-\frac{1}{k-1}} - t\right)\sqrt{n}/6$ for some $t > 0$, then the rotated pairs $\{(x_i^\natural, \mathcal{A}_i)\}_{i=1}^k$ are $4^{-\frac{1}{k-1}}$-incoherent with probability at least $1 - k(k-1)\exp(-\frac{n-2}{8}t^2)$.*

Proposition 5 requires $\sum_{i=1}^k \sqrt{\delta(\mathcal{D}_i)}$ to scale as $\sqrt{n}$ and thus controls the total complexity of the k unknown signals. We now state the main theorem and show that randomly rotated vectors can be recovered using the two-stage approach (P1) and (P2).

Theorem 1. *Suppose that Assumptions 1 and 2 hold. For any $t_1, t_2 > 0$, if $\sum_{i=1}^k \sqrt{\delta(\mathcal{D}_i)} \leq \left(1 - 4^{-\frac{1}{k-1}} - t_2\right)\sqrt{n}/6$, then any set of minimizers $\{x_i^*\}_{i=1}^k$ of (P2) satisfies*

$$\|x_i^* - x_i^\natural\|_2 \leq 4\alpha \left[\sqrt{m-1} - c\textstyle\sum_{i=1}^k \sqrt{\delta(\mathcal{D}_i)} - t_1\right]_+^{-1} \quad (7)$$

for all $i \in 1{:}k$ with probability at least $1 - \exp\left(-t_1^2/2\right) - k(k-1)\exp(-\frac{n-2}{8}t_2^2)$ and $c \leq 2$.

The proof follows directly from Proposition 1, Proposition 2, Corollary 1, Proposition 5, and the probability union bound. We verify empirically in section VI-A the tightness of the bound in (7).

A. Comparison of error bound

Here we compare our results to the one provided in [18], which also developed a novel procedure to solve the demixing problem (1). In [18], the authors introduced the constrained optimization problem

$$\begin{aligned} &\underset{x_1,\ldots,x_k}{\text{minimize}} && \left\| M^\dagger \left(M \textstyle\sum_{i=1}^k x_i - b \right) \right\|_2 \\ &\text{subject to} && \gamma(x_i) \leq \gamma_{\mathcal{A}_i}(x_i^\natural),\ \forall i \in 1{:}k, \end{aligned} \quad (8)$$

where $M^\dagger$ is the Moore-Penrose pseudo-inverse of M and showed that if $n \geq m \geq \sum_{i=1}^k \delta(\mathcal{D}_i) + \mathcal{O}(\sqrt{kn})$ and $\{x_i^\natural\}_{i=1}^k$ are randomly rotated as per Assumption 2, then any set of minimizers $\{x_i^*\}_{i=1}^k$ of (8) satisfies with high probability the bound

$$\|x_i^* - x_i^\natural\|_2 \leq C\|M^\dagger \eta\|_2 \quad (9)$$

for all $i \in 1{:}k$ [18, Theorem A]. To our knowledge, this result is the first to show that stable recovery of the constituent signals $\{x_i^\natural\}_{i=1}^k$ is possible with high probability provided the number of measurement grow linearly in k. However, the constant C in the error bound (9) could depend on all of the problem parameters except η. As a comparison to Theorem 1, the error bound in (7) makes explicit the effect of all problem parameters.

V. Deconvolution algorithm

We describe a procedure for obtaining solutions for the decompression (P1) and deconvolution (P2) problems. The procedure first solves the decompression problem (P1) using an algorithm that doesn't store or track an approximation to $x_S^\natural$, which in many contexts may be too large to store or manipulate directly. Instead, the algorithm produces a sequence of iterates $r^{(t)} \coloneqq b - Mx^{(t)}$ that approximate the residual vector corresponding to an implicit approximation $x^{(t)}$ of $x_S^\natural$. The procedure requires only the storage of several vectors of length m, which represents the size of the data b. As we show in section V-C, the solution to the deconvolution problem (P2) is subsequently obtained via an unconstrained linear least-squares problem that uses information implicit in this residual vector. Algorithm 1 summarizes the overall procedure.

A. Level-set method

The loop beginning at Line 2 of Algorithm 1 approximately solves a sequence of problems

$$v(\tau) \coloneqq \min_x \left\{ \tfrac{1}{2}\|Mx - b\|^2 \mid \gamma_{\mathcal{A}_s}(x) \leq \tau \right\}, \quad (10)$$

Algorithm 1 Decompression and deconvolution algorithm

Input: noise level $\alpha > 0$; accuracy $\epsilon > 0$

1 $\tau^0 \leftarrow 0$
2 **for** $t \leftarrow 0, 1, 2, \ldots$ **do** [level-set iterations]
3 $\quad (r^{(t)}, p^{(t)}, \ell^{(t)}) \leftarrow \mathrm{DCG}(\tau^{(t)})$ [solve (10) approximately]
4 $\quad$ **if** $\|r^{(t)}\| > \sqrt{\alpha^2 + \epsilon}$ **then** break [test ϵ-infeasibility]
5 $\quad \tau^{(t+1)} \leftarrow \tau^{(t)} + (\ell^{(t)} - \alpha^2/2)/(\langle p^{(t)}, r^{(t)} \rangle)$ [Newton update]
6 **end**
7 $(x_1, \ldots, x_k) \leftarrow$ solve (13) using $\bar{r} := r^{(t)}$ [solve (P2)]
8 **return** $(x_1, \ldots, x_k)$

parameterized by the scalar τ that defines the level-set constraint. This loop implements the level-set approach [29]–[31], which constructs a monotonically-increasing sequence $\{\tau^{(t)}\}$ that converges to the leftmost root τ_* of the equation $v(\tau) = \frac{1}{2}\alpha^2$, which represents the fidelity constraint of problem (P1). Under modest assumptions satisfied by this problem, the sequence $\tau^{(t)} \to \tau_* = \mathsf{opt}$, the optimal value of (P1). The tail of the resulting sequence of computed solutions to (10) is super-optimal and ϵ-infeasible for (P1), i.e., a solution x satisfies

$$\gamma_{\mathcal{A}_s}(x) \le \mathsf{opt} \quad \text{and} \quad \|Mx - b\| \le \sqrt{\alpha^2 + \epsilon}, \tag{11}$$

where ϵ is a specified optimality tolerance. The level-set algorithm requires $\mathcal{O}(\log(1/\epsilon))$ approximate evaluations of the optimization problem (10) (see Line 3) to achieve this optimality condition. Each approximate evaluation provides a global lower-minorant of v that is used by a Newton-like update to the level-set parameter $\tau^{(t)}$; see line 5.

B. Dual conditional gradient method

Algorithm 2 `dual-conditional-gradient`(τ). This algorithm solves (P1) without reference to the primal iterate $x^{(t)}$, and instead returns the implied residual $r^{(t)} \equiv b - Mx^{(t)}$.

Input: τ

$r^{(0)} \leftarrow b$; $q^{(0)} \leftarrow 0$
for $t \leftarrow 0, 1, 2, \ldots$ **do**
$\quad p^{(t)} \in \tau\mathcal{F}(M\mathcal{A}_s; r^{(t)})$ [$p^{(t)} \equiv Ma^{(t)}$ with $a^{(t)} \in \tau\mathcal{F}(M\mathcal{A}_s; z^{(t)})$]
$\quad \Delta r^{(t)} \leftarrow p^{(t)} - q^{(t)}$ [$\Delta r^{(t)} \equiv M(a^{(t)} - x^{(t)})$]
$\quad \rho^{(t)} \leftarrow \langle r^{(t)}, \Delta r^{(t)} \rangle$ [optimality gap]
$\quad$ **if** $\rho^{(t)} < \epsilon$ **then** break [break if optimal]
$\quad \theta^{(t)} \leftarrow \min\{1, \rho^{(t)}/\|\Delta r^{(t)}\|_2^2\}$ [exact linesearch on least-squares objective]
$\quad r^{(t+1)} \leftarrow r^{(t)} - \theta^{(t)}\Delta r^{(t)}$ [$r^{(t+1)} \equiv b - Mx^{(t+1)}$]
$\quad q^{(t+1)} \leftarrow q^{(t)} + \theta^{(t)}\Delta r^{(t)}$ [$q^{(t+1)} \equiv Mx^{(t+1)}$]
end
$\ell^{(t)} \leftarrow \frac{1}{2}\|r^{(t)}\|^2 - \rho^{(t)}$ [lower bound on optimal value]
return $r^{(t)}$, $p^{(t)}$, $\ell^{(t)}$

The level-set subproblems are solved approximately using the dual conditional-gradient method described by Algorithm 2. An implementation of this algorithm requires storage for three m-vectors

$$p^{(t)} := Ma^{(t)}, \quad q^{(t)} := Mx^{(t)}, \quad r^{(t)} := b - Mx^{(t)},$$

(A fourth vector $\Delta r^{(t)}$ can be computed at each iteration.) Implicit in these vectors are the iterate $x^{(t)}$ and current atom $a^{(t)} \in \mathcal{A}_s$, which in some situations are prohibitively large to store or manipulate The main computational cost is in Line 3, which uses the residual $r^{(t)}$ to expose an atom in the face

$$\mathcal{F}(M\mathcal{A}_s; r) = \mathrm{conv}\{\, p \in M\mathcal{A}_s \mid \langle p, r \rangle = \sup_{u \in M\mathcal{A}_s} \langle u, r \rangle \,\} \tag{12}$$

of the mapped atomic set $M\mathcal{A}_S \subset \mathbb{R}^m$. Because the exposed faces decompose under set addition, it follows from the expression (4) of $\mathcal{A}_s$ that $\mathcal{F}(M\mathcal{A}_s; r) = \sum_{i=1}^k \mathcal{F}(\lambda_i M\mathcal{A}_i; r)$. Thus, the facial exposure operation on Line 3 can be computed by separately exposing faces on each of the individual mapped atomic sets, which can be implemented in parallel, i.e.,

$$p^{(t)} = \tau \sum_{i=1}^k \lambda_i p_i^{(t)} \quad \text{where} \quad p_i^{(t)} \in \mathcal{F}(M\mathcal{A}_i; r^{(t)}) \ \forall i \in 1{:}k.$$

The conditional-gradient method converges to the required optimality within $\mathcal{O}(1/\epsilon)$ iterations [14]. Combined with the complexity of the level-set method, we thus expect a total worst-case complexity of $\mathcal{O}(\log(1/\epsilon)/\epsilon)$ iterations to satisfy the optimality condition (11).

C. Exposing the signals

Once Algorithm 1 reaches Line 7, the residual vector $r^{(t)}$ contains information about the atoms that are in the support of each of the approximations x_i^* to the signals $x_i^\natural$. It follows from Fan et al. [10, Theorem 7.1] that for all $i \in 1{:}k$,

$$x_i^* \in \mathsf{cone}\,\mathcal{F}(M\mathcal{A}_i; r^*), \quad r^* := b - M\sum_{i=1}^k x_i^*.$$

Thus, a solution of the deconvolution problem (P2) can be recovered by solving

$$\begin{array}{ll} \underset{x_1,\ldots,x_k}{\text{minimize}} & \frac{1}{2}\|M\sum_{i=1}^k x_i - (b - \bar{r})\|^2 \\ \text{subject to} & x_i \in \mathsf{cone}\,\mathcal{F}(M\mathcal{A}_i; \bar{r}), \end{array} \tag{13}$$

which can be implemented as a standard linear least-squares problem over the coefficients of the atoms exposed in each of the atomic sets.

VI. Experiments and novel applications

In section VI-A we empirically verify Theorem 1 through a set of synthetic experiments on recovering multiple randomly-rotated sparse signals from noiseless and noisy measurements. Note that the random rotation guarantees incoherence among the unknown signals $\{x_i^\natural\}_{i=1}^k$. We also empirically show that random rotation is not required for successful recovery of a class of unknown signals with different underlying structures. In section VI-B we separate a sparse signal and sparse-in-frequency signal. In section VI-C we separate the superposition of three signals: a sparse signal, a low-rank matrix, and noise. In section VI-D we separate a multiscale low-rank synthetic image.

We implement the algorithm described in Section V in the Julia language [32] using version 1.6 to recover the unknown signals. All the experiments are conducted on a Linux server with 8 CPUs and 64Gb memory.

A. Stability of Demixing

We provide three experiments that numerically verify the bounds established by Theorem 1 to solve the demixing problem (1). The experiment draws multiple realizations of a random problem specified over a range of parameters k (number of signals), m (number of measurements), n (signal dimension) and s (the sparsity level for each signal). Each signal $x_i^\natural$ in (1) is generated according to Assumption 2, where each vector x_i° is s-sparse with respect to the standard basis. By construction, the atomic sets $i \in 1{:}k$ are defined to be

$$\mathcal{A}_i = Q_i \{ \pm e_1, \ldots, \pm e_n \} \quad \text{where} \quad Q_i \sim \texttt{uniform}(\mathrm{SO}(n)).$$

Amelunxen et al. [33, Proposition 4.5] give an upper bound on the statistical dimension of the descent cone for $(x_i^\natural, \mathcal{A}_i)$, and thus for the descent cone at $(x_i^\circ, \mathcal{A}_i^\circ)$, for s-sparse vectors. We use this bound to approximate the statistical dimension $\delta(\mathcal{D}_i)$ of the descent cone $\mathcal{D}_i$ corresponding to the pair $(x_i^\natural, \mathcal{A}_i)$. We define the maximum absolute error

$$\texttt{maxerr} := \max_{i \in 1:k} \|x_i^* - x_i^\natural\|_2. \tag{14}$$

1) Relation between m and n: We first show a phase portrait for the noiseless case that verifies the relationship between number of measurement m and signal dimension n, as stated in Theorem 1. The number of signals is fixed at $k = 3$ and the sparsity level is fixed at $s = 5$. The phase plot is shown in Figure 4, where the horizontal axis represents the signal dimension $n \in \{50, 65, \ldots, 500\}$ and the vertical axis represents the number of measurements $m \in \{50, 65, \ldots, 500\}$. The colormap indicates the empirical probability of successful demixing over 50 trials, where we say the demixing is successful if $\texttt{maxerr} < 10^{-2}$. The red solid curve and the blue dashed line, respectively, approximate the graphs of the functions

$$\sqrt{m} = \sum_{i=1}^{k} \sqrt{\delta(\mathcal{D}_i)} \quad \text{and} \quad \sqrt{n} = \sum_{i=1}^{k} \sqrt{\delta(\mathcal{D}_i)}.$$

The statistical dimensions of $\mathcal{D}_i$ are approximated using [33, Proposition 4.5], as stated above. The area above the red curve and to the right of the dashed line corresponds to problem parameters with successful recovery and corroborates the bounds stated in Theorem 1.

2) Relation between m and k: We also show a phase portrait for the noiseless case that verifies the relationship between number of measurement m and number of signals k stated in Theorem 1. The signal dimension is fixed at $n = 1000$ and the sparsity level is fixed at $s = 3$. The phase plot is shown in Figure 5, where the horizontal axis represents the number of signals $k \in \{2, 3, \ldots, 10\}$ and the vertical axis represents the number of measurements $m \in \{100, 200, \ldots, 1000\}$. All the other settings are the same as stated in section VI-A1. The red line corresponds to $\sqrt{m} = \sum_{i=1}^{k} \sqrt{\delta(\mathcal{D}_i)}$ and shows that recovery is possible provided the number of measurements scale as k^2, when the complexity of all of unknown signals are the same.

3) Relation between maximal absolute error and noise level: Lastly, we show a plot for the noisy case that verifies the relationship between maximum absolute error $\texttt{maxerr}$ and noise level α stated in Theorem 1. The number of measurement is fixed at $m = 125$, the signal dimension is fixed at $n = 200$, the number of signals is fixed at $k = 3$, and the sparsity level is fixed at $s = 5$. The result is shown in Figure 6, where the horizontal axis represents the noise level $\alpha \in \{0.01, 0.02, \ldots, 2\}$ and the vertical axis represents the maximum absolute error $\texttt{maxerr}$. The blue curve corresponds to the mean of $\texttt{maxerr}$ over 50 trials and the yellow shaded area corresponds to the standard deviation. The figure verifies the linear dependence of the recovery error with the noise level, as stated in Theorem 1.

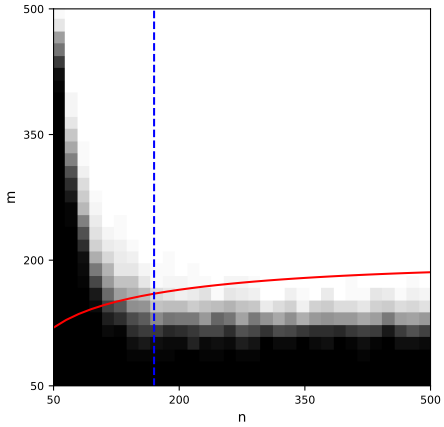


Fig. 4. Phase-transition plots for demixing the sum of randomly-rotated sparse signals $\{x_i^\natural\}_{i=1}^k$ from noiseless measurements b. The horizontal and vertical axes, respectively, represent the signal dimension n and measurement dimension m. The colormap indicates the empirical probability of successful demixing over 50 trials. The red solid curve approximately represents the mapping $\sqrt{m} = \sum_{i=1}^{k} \sqrt{\delta(\mathcal{D}_i)}$ and the blue dashed line approximately represents the position $\sqrt{n} = \sum_{i=1}^{k} \sqrt{\delta(\mathcal{D}_i)}$.

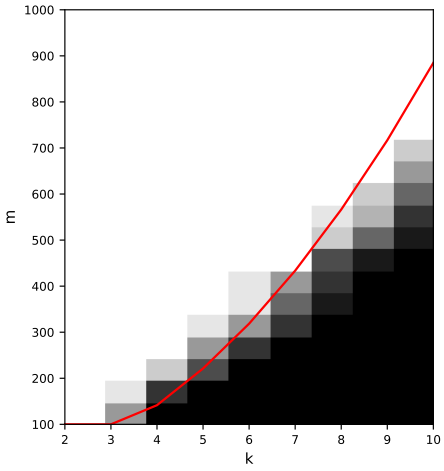


Fig. 5. Phase-transition plots for demixing the sum of randomly-rotated sparse signals $\{x_i^\natural\}_{i=1}^k$ from noiseless measurements b. The horizontal and vertical axes, respectively, represent the number of signals k and measurement dimension m. The colormap indicates the empirical probability of successful demixing over 50 trials. The red solid curve approximately represents the mapping $\sqrt{m} = \sum_{i=1}^{k} \sqrt{\delta(\mathcal{D}_i)}$.

B. Separation of sparse and sparse-in-frequency signals

We make a direct comparison to the approach described by McCoy et al. [34] and reproduce their experiment on

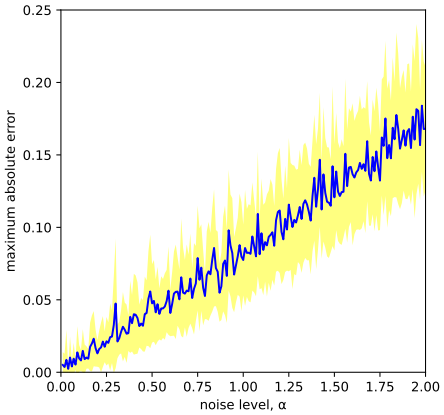


Fig. 6. Error-noise plot for demixing the sum of randomly-rotated sparse signals $\{x_i^\natural\}_{i=1}^k$ from noisy measurements b. The horizontal and vertical axes, respectively, represent the noise level α and the maximum absolute error `maxerr`. The blue curve indicates the relationship between the empirical average of `maxerr` over 50 trials and α, and the yellow shaded area indicated the empirical standard deviation.

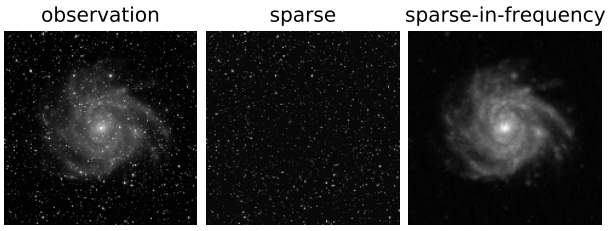


Fig. 7. The star-galaxy separation experiment features two distinct signal components. The image size is 601 × 601 pixels.

separating an astronomical image into a sparse and a sparse-in-frequency signals. An n-vector x is sparse-in-frequency if its discrete cosine transform (DCT) Dx is sparse, where the orthogonal linear map $D : \mathbb{R}^n \to \mathbb{R}^n$ encodes the DCT. Define the observations and corresponding atomic sets

$$b = x_s^\natural + x_d^\natural, \quad \mathcal{A}_s := \{\pm e_1, \ldots, \pm e_n\}, \quad \mathcal{A}_d = D^* \mathcal{A}_s.$$

The star-galaxy image shown in Figure 7 exemplifies this superposition: the stars are well-represented by sparse matrices in $\mathcal{A}_s$, and the galaxy component is well-represented by sinusoidal elements in $\mathcal{A}_d$. The image size is 601×601. The results of the separation are shown in the second two panels of Figure 7.

C. Sparse and low rank matrix decomposition with structured noise

In this next example we decompose an image that contains a sparse foreground, a low-rank background, and structured noise. This is an example of sparse principle component analysis [35]–[38]. Typically, the entry-wise 1-norm and the nuclear norm are used to extract from the matrix each of these qualitatively different structures. Here, we treat the noise as its own signal that also needs to be separated. We consider the observations

$$B = X_s^\natural + X_l^\natural + X_n^\natural,$$

where $X_s^\natural \in \mathbb{R}^{m\times n}$ is sparse, $X_l^\natural \in \mathbb{R}^{m\times n}$ is low-rank matrix, and $X_n^\natural \in \mathbb{R}^{m\times n}$ represents structured noise so that

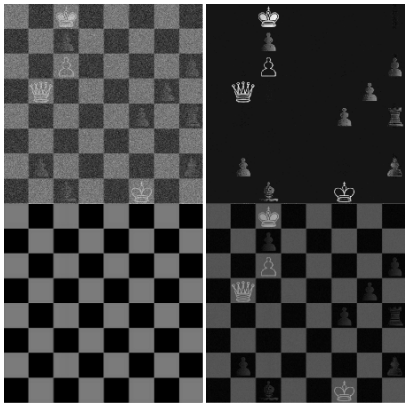


Fig. 8. Noisy chess board in-painting experiment. The image size is 596×596. Northwest: noisy observations; Northeast: recovered sparse component; Southwest: recovered low rank component; Southeast: denoising result.

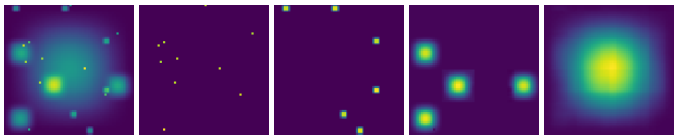


Fig. 9. Multiscale low rank matrix decomposition experiment. The matrix size is 64×64. From left to right: observations; recovered $\mathcal{P}_i$-block-wise low rank component for $i = 1, \ldots, 4$. All the blocks in $\mathcal{P}_i$ have the same size $4^{i-1} \times 4^{i-1}$ for $i = 1, \ldots, 4$.

$PX_n^\natural Q$ is sparse, where P and Q are random orthogonal m-by-m matrices. Based on the atomic framework, we choose the atomic sets for $X_s^\natural$, $X_l^\natural$, and $X_n^\natural$, respective, as

$$\mathcal{A}_s = \{\pm E_{i,j} \mid 1 \le i \le m, 1 \le j \le n\},$$
$$\mathcal{A}_l = \{uv^\intercal \mid u \in \mathbb{R}^m,\ v \in \mathbb{R}^n,\ \|u\|_2 = \|v\|_2 = 1\},$$
$$\mathcal{A}_n = P^\intercal \mathcal{A}_s Q^\intercal,$$

where $E_{i,j}$ is a $m \times n$ matrix with a single nonzero entry (i, j) with value 1.

For the numerical experiment, we consider the noisy chess board in-painting problem. The chess foreground is sparse and the chess board background is low rank. The image size is 596×596. The experiment result is shown in Figure 8.

D. Multiscale low rank matrix decomposition

The multiscale low-rank matrix decomposition problem proposed by Ong and Lustig [39] generalizes the sparse and low-rank matrix decomposition through a block-wise low-rank structure. Let X be an $m \times n$ matrix and $\mathcal{P}$ be a partition of X into multiple blocks. Then X is considered to be block-wise low-rank with respect to $\mathcal{P}$ if all the blocks are low rank. For each block $p \in \mathcal{P}$ with size $m_p \times n_p$, let X_p denote the corresponding part of the matrix X and let $R_p : \mathbb{R}^{m\times n} \to \mathbb{R}^{m_p \times n_p}$ denote the linear operator that can extract X_p from X, namely $R_p(X) = X_p$. The adjoint operator $R_p^* : \mathbb{R}^{m_p \times n_p} \to \mathbb{R}^{m\times n}$ embeds an $m_p \times n_p$ matrix into a $m \times n$ zero matrix. With this operator,

$$X = \sum_{p\in\mathcal{P}} R_p^*(X_p).$$

Each block-wise low-rank signal is represented by a corresponding atomic set. By definition, each block $X_p \in$

$\mathbb{R}^{m_p \times n_p}$ is low rank, and thus X_p is $\mathcal{A}_p$-sparse, where

$$\mathcal{A}_p = \{uv^\intercal \mid u \in \mathbb{R}^{m_p}, v \in \mathbb{R}^{n_p}, \|u\| = \|v\| = 1\}.$$

One and Lustig [39] propose a block-wise nuclear norm and its associated dual norm, respectively, by the functions $\|\cdot\|_{\mathcal{P},1} = \sum_{p\in\mathcal{P}} \|R_p(\cdot)\|_1$, $\|\cdot\|_{\mathcal{P},\infty} = \max_{p\in\mathcal{P}} \|R_p(\cdot)\|_\infty$, where $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are the Schatten 1- and ∞-norms of their matrix arguments. It follows that the block-wise norm $\|\cdot\|_{\mathcal{P},1}$ and dual norm $\|\cdot\|_{\mathcal{P},\infty}$ are the gauge and support functions, respectively, for the atomic set $\mathcal{A}_\mathcal{P} := \bigcup_{p\in\mathcal{P}} R_p^* \mathcal{A}_p$.

We reproduce the synthetic model described by Ong and Lustig, who construct the superposition $B = \sum_{i=1}^k X_i^\natural$, where $X_i^\natural \in \mathbb{R}^{m\times n}$ is block-wise low rank with respect to the multiscale partitions $\{\mathcal{P}_i\}_{i=1}^k$. In our experiment, we set $m = n = 64$, $k = 4$, and for each $i \in 1{:}k$,

$$m_p = n_p = 4^{i-1} \quad \forall p \in \mathcal{P}_i.$$

At the lowest scale $i = 1$, a block-wise low-rank matrix is a scalar, and so 1-sparse matrices are included with the atomic set $\mathcal{A}_{\mathcal{P}_1}$. The solutions of the deconvolution procedure Equation (P2) are shown in Figure 9.

VII. Looking ahead

The random rotation model is a useful mechanism for introducing incoherence among the individual signals. However, even in contexts where it's possible to rotate the signals, it may prove too costly to do so in practice because the rotations need to be applied at each iteration of the algorithm in Line 3. We might then consider other mechanisms for introducing incoherence that are computationally cheaper, and rely instead, for example, on some fast random transform. The literature on demixing abounds with various incoherence notions. We wish to explore what is the relationship between these and definition of β-incoherence that we adopt. Alternative incoherence definitions may prove useful in deriving other mechanisms for inducing incoherence in the signals.

A significant assumption of our analysis is that the parameters λ_i exactly equilibrate the gauge values for each signal; cf. (5). In practice, however, we can only estimate these. It may be possible to analyze how the stability in the recovery of the signals depends on errors that might exist in the ideal parameter choices.

Appendix A
Proofs

This section contains proofs for the mathematical statements in Section III and Section IV. We begin with several technical results needed for analysis, which describe useful properties of descent cones. Some of these results contain their own intrinsic interest.

A. Lemmas

Lemma 1 (Properties of descent cones)**.** *Let $\mathcal{A}$, $\mathcal{A}_1$, $\mathcal{A}_2$ be compact sets in $\mathbb{R}^n$ that contain the origin in their interiors. Fix the vectors x, x_1, x_2. The following properties hold.*

a) A vector d is contained in $\mathcal{D}(\mathcal{A}, x)$ if and only if there is some $\bar{\alpha} > 0$ such that $\gamma_\mathcal{A}(x + \alpha d) \le \gamma_\mathcal{A}(x)$ for all $\alpha \in [0, \bar{\alpha}]$;
b) $\mathcal{D}(\tau\mathcal{A}, x) = \mathcal{D}(\mathcal{A}, x)$ $\forall \tau > 0$;
c) $\mathcal{D}(Q\mathcal{A}, Qx) = Q\mathcal{D}(\mathcal{A}, x)$ if $Q \in \mathrm{SO}(n)$;
d) $\mathcal{D}(\mathcal{A}_1 + \mathcal{A}_2, x_1 + x_2) \subseteq \mathcal{D}(\mathcal{A}_1, x_1) + \mathcal{D}(\mathcal{A}_2, x_2)$ if $\gamma_{\mathcal{A}_1}(x_1) = \gamma_{\mathcal{A}_2}(x_2)$.

Proof.

a) See [17, Proposition 2.5];
b) It follows from the fact that a gauge function is positive homogenous.
c) Because $\gamma_{Q\mathcal{A}} = \gamma_\mathcal{A}(Q^*\cdot)$,

$$\begin{aligned}\mathcal{D}(Q\mathcal{A}, Qx) &= \mathsf{cone}\,\{\, d \mid \gamma_{Q\mathcal{A}}(Qx + d) \le \gamma_{Q\mathcal{A}}(Qx)\,\} \\ &= \mathsf{cone}\,\{\, d \mid \gamma_\mathcal{A}(x + Q^*d) \le \gamma_\mathcal{A}(x)\,\} \\ &= Q\mathcal{D}(\mathcal{A}, x).\end{aligned}$$

d) For every $d \in \mathcal{D}(\mathcal{A}_1 + \mathcal{A}_2, x_1 + x_2)$, by Lemma 1(a), there exists $\alpha > 0$ such that

$$\gamma_{\mathcal{A}_1+\mathcal{A}_2}(x_1 + x_2 + \alpha d) \le \gamma_{\mathcal{A}_1+\mathcal{A}_2}(x_1 + x_2).$$

Then there exists d_1, d_2 such that $d_1 + d_2 = \alpha d$ and $\max\{\gamma_{\mathcal{A}_1}(x_1 + d_1), \gamma_{\mathcal{A}_2}(x_2 + d_2)\} \le \gamma_{\mathcal{A}_1+\mathcal{A}_2}(x_1+x_2)$. By the fact that $\gamma_{\mathcal{A}_1+\mathcal{A}_2}(x_1 + x_2) \le \max\{\gamma_{\mathcal{A}_1}(x_1), \gamma_{\mathcal{A}_2}(x_2)\}$ and the assumption $\gamma_{\mathcal{A}_1}(x_1) = \gamma_{\mathcal{A}_2}(x_2)$, it follows that $d_i \in \mathcal{D}(\mathcal{A}_i, x_i)$, which implies $\alpha d = d_1 + d_2 \in \mathcal{D}(\mathcal{A}_1, x_1) + \mathcal{D}(\mathcal{A}_2, x_2)$. Thus $d \in \mathcal{D}(\mathcal{A}_1, x_1) + \mathcal{D}(\mathcal{A}_2, x_2)$.

□

The Gaussian width of a set $T \subset \mathbb{R}^n$ is defined as

$$\omega(T) = \mathbb{E}_g \sup\{\, \langle g, y\rangle \mid y \in T\,\},$$

where the expectation is taken with respect to the standard Gaussian $\texttt{normal}(0, I_n)$. The following lemma summarizes the main properties that we use regarding the relationship between the conic summaries δ and ω.

Lemma 2 (Properties of conic statistical summaries)**.** *Let $\mathcal{K}$ be a closed and convex cones in $\mathbb{R}^n$ and let $Q \in \mathrm{SO}(n)$. Then the following properties hold.*

a) $\delta(Q\mathcal{K}) = \delta(\mathcal{K})$;
b) $\delta(\mathcal{K}) = \mathbb{E}_g\left[\sup\{\, \langle g, y\rangle \mid y \in \mathcal{K} \cap \mathbb{B}^n\,\}^2\right]$;
c) $\omega(\mathcal{K} \cap \mathbb{B}^n)^2 \le \delta(\mathcal{K})$.

Proof. See [33, Proposition 3.1(6) and Proposition 3.1(5)], respectively, for (a) and (b).

c) Indeed, we know that,

$$\begin{aligned}\omega(\mathcal{K} \cap \mathbb{B}^n)^2 &= [\mathbb{E}_g \sup\{\, \langle g, y\rangle \mid y \in \mathcal{K} \cap \mathbb{B}^n\,\}]^2 \\ &\le \mathbb{E}_g\left[\sup\{\, \langle g, y\rangle \mid y \in \mathcal{K} \cap \mathbb{B}^n\,\}^2\right] \\ &= \delta(\mathcal{K}),\end{aligned}$$

where the first equality follows from the definition of gaussian with, the first inequality follows from the fact that $\mathbb{E}(X)^2 \le \mathbb{E}(X)^2$ for any random variable X, and the last equality follows from Lemma 2(b).

□

Our next lemma shows that if the angle between two cones is bounded, then the norms of individual vectors are bounded by the norm of their sum.

Lemma 3. *Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be two closed convex cones in $\mathbb{R}^n$. If $\cos\angle(-\mathcal{K}_1,\mathcal{K}_2)\le 1-\beta$ for some $\beta\in(0,1]$, then for any $u\in\mathcal{K}_1$ and $v\in\mathcal{K}_2$,*

$$\max\{\|u\|,\|v\|\}\le\frac{1}{\sqrt{\beta}}\|u+v\|.$$

Proof. By expanding the norm square of $u+v$ we can get that

$$\begin{aligned}\|u+v\|^2&=\|u\|^2+\|v\|^2-2\langle -u,v\rangle\\&=\|u\|^2+\|v\|^2-2\cos(\angle(-u,v))\|u\|\|v\|\\&\ge\|u\|^2+\|v\|^2-2(1-\beta)\|u\|\|v\|\\&=\beta(\|u\|^2+\|v\|^2)+(1-\beta)(\|u\|-\|v\|)^2\\&\ge\beta\max\{\|u\|^2,\|v\|^2\},\end{aligned}$$

where the first inequality follows from the definition of the cosine of the angle between two cones. □

Our next lemma is a technical lemma for the expectation.

Lemma 4. *Let X and Y be nonnegative random variables, then we have*

$$\mathbb{E}[(X+Y)^2]\le\left(\sqrt{\mathbb{E}[X^2]}+\sqrt{\mathbb{E}[Y^2]}\right)^2.$$

Proof. By expanding the right hand side, we can get

$$\begin{aligned}\left(\sqrt{\mathbb{E}[X^2]}+\sqrt{\mathbb{E}[Y^2]}\right)^2&=\mathbb{E}[X^2]+\mathbb{E}[Y^2]+2\sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}\\&\ge\mathbb{E}[X^2]+\mathbb{E}[Y^2]+2\,\mathbb{E}[XY]\\&=\mathbb{E}[(X+Y)^2],\end{aligned}$$

where the inequality follows from the Cauchy–Schwarz inequality. □

B. Proof for Proposition 2

For each $i\in 1{:}k$, let $\epsilon_i:=x_i^*-x_i^\natural$ and $\epsilon_{-i}:=\sum_{j\neq i}\epsilon_j$. By the definition of descent cone, $\epsilon_i\in\mathcal{D}(\mathcal{A}_i,x_i^\natural)$. Because $\{(x_i^\natural,\mathcal{A}_i)\}_{i=1}^k$ are β-incoherent for some $\beta\in(0,1]$, by Definition 1,

$$\cos\angle(-\epsilon_i,\epsilon_{-i})\le 1-\beta.$$

By Lemma 3, it follows that

$$\|\epsilon_i+\epsilon_{-i}\|\ge\sqrt{\beta}\|\epsilon_i\|.$$

The desired result follows.

C. Proof for Proposition 3

In this proof, we define $\mathcal{K}_{i,\beta}=\mathcal{K}_i\cap\frac{1}{\sqrt{\beta}}\mathbb{B}^n$ and $f_{i,\beta}(g)=\sup\{\langle g,u\rangle\mid u\in\mathcal{K}_{i,\beta}\}$ for $i=1,2$. By Lemma 2(b), we know that the statistical dimension can be expressed as

$$\begin{aligned}&\delta(\mathcal{K}_1+\mathcal{K}_2)=\mathbb{E}_g\left[\sup\{\langle g,y\rangle\mid y\in(\mathcal{K}_1+\mathcal{K}_2)\cap\mathbb{B}^n\}^2\right]\\&=\mathbb{E}_g\left[\sup\{\langle g,u+v\rangle\mid u\in\mathcal{K}_1,v\in\mathcal{K}_2,\|u+v\|\le 1\}^2\right]\\&\le\mathbb{E}_g\left[\sup\{\langle g,u+v\rangle\mid u\in\mathcal{K}_{1,\beta},v\in\mathcal{K}_{2,\beta}\}^2\right]\\&=\mathbb{E}_g\left[(\sup\{\langle g,u\rangle\mid u\in\mathcal{K}_{1,\beta}\}+\sup\{\langle g,v\rangle\mid v\in\mathcal{K}_{2,\beta}\})^2\right]\\&\le\left(\sqrt{\mathbb{E}_g[f_{1,\beta}(g)^2]}+\sqrt{\mathbb{E}_g[f_{2,\beta}(g)^2]}\right)^2\\&=\tfrac{1}{\beta}\left(\sqrt{\delta(\mathcal{K}_1)}+\sqrt{\delta(\mathcal{K}_2)}\right)^2\end{aligned}$$

where the first inequality follows from Lemma 3 and the fact that the supremum is always nonnegative, and the second inequality follows from Lemma 4.

D. Proof for Corollary 1

Throught this proof, for all $i\in 1:k$, we define $\mathcal{D}_i=\mathcal{D}(\mathcal{A}_i,x_i^\natural)$, $\delta_i=\delta(\mathcal{D}_i)$ and $\delta_{1:i}=\delta(\sum_{j=1}^i\mathcal{D}_i)$. By Assumption 1 and Lemma 1(d), we know that $\mathcal{D}_s\subseteq\sum_{i=1}^k\mathcal{D}_i$, and it follows that $\delta(\mathcal{D}_s)\le\delta_{1:k}$. So we only need to give an upper bound for $\delta_{1:k}$. Since $\cos\angle\left(-\mathcal{D}_k,\sum_{i=1}^{k-1}\mathcal{D}_i\right)\le 1-\beta$, by Proposition 3, it follows that

$$\sqrt{\delta_{1:k}}\le\beta^{-\frac{1}{2}}\left(\sqrt{\delta_{1:(k-1)}}+\sqrt{\delta_k}\right).\tag{15}$$

Since $\sum_{i=1}^{k-2}\mathcal{D}_i\subseteq\sum_{j\neq(k-1)}\mathcal{D}_j$, it follows that $\cos\angle\left(-\mathcal{D}_{k-1},\sum_{i=1}^{k-2}\mathcal{D}_i\right)\le 1-\beta$. By Proposition 3, we have

$$\sqrt{\delta_{1:(k-1)}}\le\beta^{-\frac{1}{2}}\left(\sqrt{\delta_{1:(k-2)}}+\sqrt{\delta_{k-1}}\right).\tag{16}$$

Combining (15) and (16), we know that

$$\sqrt{\delta_{1:k}}\le\beta^{-\frac{2}{2}}\left(\sqrt{\delta_{1:(k-2)}}+\sqrt{\delta_{k-1}}+\sqrt{\delta_k}\right).$$

Repeating this process, we can conclude that

$$\sqrt{\delta_{1:k}}\le\beta^{-\frac{k-1}{2}}\sum_{i=1}^k\sqrt{\delta_i}.$$

E. Proof for Proposition 4

Throught this proof, we define the following notations:

- $\overline{\mathcal{K}}_i:=\mathcal{K}_i\cap\mathbb{S}^{n-1}$ for $i=1,2$;
- $\widehat{\mathcal{K}}_i:=\mathcal{K}_i\cap\mathbb{B}^n$ for $i=1,2$;
- $f(W\in\mathbb{R}^{n\times n})=\sup\{\langle x,Wy\rangle\mid x\in\overline{\mathcal{K}}_1,y\in\overline{\mathcal{K}}_2\}$;
- $\hat{f}(W\in\mathbb{R}^{n\times n})=\sup\{\langle x,Wy\rangle\mid x\in\widehat{\mathcal{K}}_1,y\in\widehat{\mathcal{K}}_2\}$;
- $\mathcal{O}_n=\{Q\in\mathbb{R}^{n\times n}:Q^TQ=I_n\}$;
- $\mathcal{SO}_{n,+}=\{Q\in\mathcal{O}_n:\det(Q)=1\}$;
- $\mathcal{SO}_{n,-}=\{Q\in\mathcal{O}_n:\det(Q)=-1\}$.

Our proof consists of three steps.

First step: show that both f and $\hat{f}$ are convex and 1-Lipschitz functions. First, we show that both f and $\hat{f}$ are convex. For any $W_1,W_2\in\mathbb{R}^{n\times n}$ and any $t\in[0,1]$,

$$\begin{aligned}&f(tW_1+(1-t)W_2)\\&=\sup\{\langle x,(tW_1+(1-t)W_2)y\rangle\mid x\in\overline{\mathcal{K}}_1,y\in\overline{\mathcal{K}}_2\}\\&=\sup\{\langle x,tW_1y\rangle+\langle x,(1-t)W_2y\rangle\mid x\in\overline{\mathcal{K}}_1,y\in\overline{\mathcal{K}}_2\}\\&\le tf(W_1)+(1-t)f(W_2).\end{aligned}$$

So f is convex. The same reason holds for $\hat{f}$, and thus $\hat{f}$ is also convex. Next, by [40, Lemma 2.6], in order to show that both f and $\hat{f}$ are 1-Lipschitz, we only need to show that the norm of any subgradient of f or $\hat{f}$ is bounded by 1. By [41, Theorem D.4.4.2], we know that for any $W\in\mathbb{R}^{n\times n}$,

$$\begin{aligned}\partial f(W)&=\operatorname{conv}\{xy^T\mid x\in\overline{\mathcal{K}}_1,y\in\overline{\mathcal{K}}_2,\langle x,Wy\rangle=f(W)\},\\\partial\hat{f}(W)&=\operatorname{conv}\{xy^T\mid x\in\widehat{\mathcal{K}}_1,y\in\widehat{\mathcal{K}}_2,\langle x,Wy\rangle=f(W)\}.\end{aligned}$$

Since $\|x\|\le 1$ and $\|y\|\le 1$, it is easy to verify that for any $W\in\mathbb{R}^{n\times n}$ and for any $Z\in\partial f(W)\cup\partial\hat{f}(W)$,

$$\|Z\|_F\le 1,$$

where $\|\cdot\|_F$ is the Frobenius norm. Therefore, we can conclude that both f and $\hat{f}$ are 1-Lipschitz functions.

Second step: bound $\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}[f(Q)]$. First, we give the bound on $\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{O}_n)}\left[\hat{f}(Q)\right]$. From the first step, we know that $\hat{f}$ is convex. Then by the comparison principle developed by Tropp; see [42, Theorem 5 and Lemma 8], we can conclude that

$$\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{O}_n)}\left[\hat{f}(Q)\right] \le \tfrac{1.5}{\sqrt{n}}\,\mathbb{E}_{G\sim\mathcal{N}(0,I_n)}\left[\hat{f}(G)\right], \quad (17)$$

Next, we give the bound on $\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}\left[\hat{f}(Q)\right]$. By expanding the uniform distribution over $\mathcal{O}_n$, we can get

$$\begin{aligned}&\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{O}_n)}\left[\hat{f}(Q)\right]\\ &= \tfrac{1}{2}\,\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}\left[\hat{f}(Q)\right] + \tfrac{1}{2}\,\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,-})}\left[\hat{f}(Q)\right]\\ &\ge \tfrac{1}{2}\,\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}\left[\hat{f}(Q)\right],\end{aligned}$$

where the inequality follows from the fact that $\hat{f}$ is non-negative. Combine this result with (17), we can conclude that

$$\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}\left[\hat{f}(Q)\right] \le \tfrac{3}{\sqrt{n}}\,\mathbb{E}_{G\sim\mathcal{N}(0,I_n)}\left[\hat{f}(G)\right]. \quad (18)$$

Then, by the Gaussian Chevet's inequality; see [43, Exercise 8.7.4], we know that

$$\mathbb{E}_{G\sim\mathcal{N}(0,I_n)}\left[\hat{f}(G)\right] \le \omega(\widehat{\mathcal{K}}_1)+\omega(\widehat{\mathcal{K}}_2) \le \sqrt{\delta(\mathcal{K}_1)}+\sqrt{\delta(\mathcal{K}_2)}, \quad (19)$$

where the second inequality follows from Lemma 2(c). Combine (18) and (19), we can get

$$\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}\left[\hat{f}(Q)\right] \le \tfrac{3}{\sqrt{n}}\left(\sqrt{\delta(\mathcal{K}_1)}+\sqrt{\delta(\mathcal{K}_2)}\right). \quad (20)$$

Finally, by the fact that $f\le\hat{f}$ and (20), we can conclude that

$$\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}[f(Q)] \le \tfrac{3}{\sqrt{n}}\left(\sqrt{\delta(\mathcal{K}_1)}+\sqrt{\delta(\mathcal{K}_2)}\right). \quad (21)$$

Third step: concentration bound for $f(Q)$. From step 1, we know that f is 1-Lipschitz. For clearness, we denote $\mathbb{P}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}$ and $\mathbb{E}_{Q\sim\texttt{uniform}(\mathcal{SO}_{n,+})}$ as $\mathbb{P}_Q$ and $\mathbb{E}_Q$. By the concentration bounds of Lipschitz functions over the special orthogonal group develop by Meckes; see [44, Theorem 5.5 and Theorem 5.16], we can get that for every $t\ge 0$,

$$\mathbb{P}_Q\left[f(Q)\ge\mathbb{E}_Q[f(Q)]+t\right]\le\exp(-\tfrac{n-2}{8}t^2).$$

Note that a similar result can be obtained from [43, Theorem 5.2.7]. Combining with (21), we can conclude that for every $t\ge 0$,

$$\mathbb{P}_Q\left[f(Q)\ge\tfrac{3}{\sqrt{n}}\left(\sqrt{\delta(\mathcal{K}_1)}+\sqrt{\delta(\mathcal{K}_2)}\right)+t\right]\le\exp(-\tfrac{n-2}{8}t^2).$$

F. Lemmas needed for the proof of Proposition 5

In this section, we present two lemmas that are needed for the proof of Proposition 5. These two lemmas provide probabilistic bound on the statistical dimension of sum of randomly rotated cones.

The next lemma provides a probabilistic bound on the statistical dimension of the sum of two cones.

Lemma 5 (Probabilistic bound on statistical dimension under random rotation)**.** *Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be two closed convex cones in $\mathbb{R}^n$. Then*

$$\begin{aligned}\mathbb{P}\left[\sqrt{\delta(\mathcal{K}_1+Q\mathcal{K}_2)} \le \tfrac{1}{\sqrt{\beta(t)}}\left(\sqrt{\delta(\mathcal{K}_1)}+\sqrt{\delta(\mathcal{K}_2)}\right)\right]\\ \ge 1-\exp(-\tfrac{n-2}{8}t^2)\\ \text{with}\quad \beta(t)=1-\tfrac{3}{\sqrt{n}}\left(\sqrt{\delta(\mathcal{K}_1)}+\sqrt{\delta(\mathcal{K}_2)}\right)-t,\end{aligned}$$

where Q is drawn uniformly at random from $\mathrm{SO}(n)$.

Proof. By Proposition 3 and Proposition 4. □

Our next lemma extends Lemma 5 to arbitrary number of cones.

Lemma 6. *Let $\mathcal{K}_1,\dots,\mathcal{K}_p$ be closed convex cones in $\mathbb{R}^n$ and let $Q_1,\dots,Q_p$ be i.i.d. matrices uniformly drawn from $\mathrm{SO}(n)$. If $\sum_{i=1}^p\sqrt{\delta(\mathcal{K}_i)}\le\left(1-4^{-\frac{1}{p-1}}-t\right)\sqrt{n}/6$ for some $t>0$, then*

$$\mathbb{P}\left[\sqrt{\delta\left(\textstyle\sum_{i=1}^p Q_i\mathcal{K}_i\right)}\le 2\textstyle\sum_{i=1}^p\sqrt{\delta(\mathcal{K}_i)}\right]\ge 1-(p-1)\exp(-\tfrac{n-2}{8}t^2).$$

Proof. Throughout this proof, we define the following notations:

- $\delta_i=\delta(\mathcal{K}_i)$, for all $i\in 1:p$;
- $\delta_{1:i}=\delta\left(\sum_{j=1}^i Q_j\mathcal{K}_j\right)$, for all $i\in 1:p$;
- For each $i\in 2:p$, define the event

$$E_i(t)=\left\{\sqrt{\delta_{1:i}}\le\tfrac{1}{\sqrt{\beta_i(t)}}\left(\sqrt{\delta_{1:(i-1)}}+\sqrt{\delta_i}\right)\right\}$$
$$\text{with}\quad \beta_i(t)=1-\tfrac{3}{\sqrt{n}}\left(\sqrt{\delta_{1:(i-1)}}+\sqrt{\delta_i}\right)-t.$$

Our proof consists of three steps.

Step 1: bound the probability of $E_2(t)\wedge\cdots\wedge E_p(t)$. Denote the indicator random variable for $E_i(t)$ by $\mathbb{1}_{E_i(t)}$, which evaluates to 1 if $E_i(t)$ occurs and otherwise evaluates to 0. Then for each $i\in 2:p$, we have

$$\begin{aligned}\mathbb{P}(E_i(t)) &= \mathbb{E}(\mathbb{1}_{E_i(t)})\\ &= \mathbb{E}_{\{Q_j\}_{j=1}^{i-1}}\left[\mathbb{E}\left(\mathbb{1}_{E_i(t)}\mid\{Q_j\}_{j=1}^{i-1}\right)\right]\\ &\ge \mathbb{E}_{\{Q_j\}_{j=1}^{i-1}}\left[1-\exp(-\tfrac{n-2}{8}t^2)\right]\\ &= 1-\exp(-\tfrac{n-2}{8}t^2),\end{aligned}$$

where the inequality follows from Lemma 5 and the assumption that Q_i are all independent. Extending the bound on $\mathbb{P}(E_i(t))$ to all $i\in 2{:}p$ via the union bound, we have

$$\mathbb{P}(E_2(t)\wedge\cdots\wedge E_p(t))\ge 1-(p-1)\exp(-\tfrac{n-2}{8}t^2).$$

Step 2: show that $E_2(t)\wedge\cdots\wedge E_p(t)$ implies bound on $\sqrt{\delta_{1:p}}\le\frac{1}{\sqrt{\beta_2(t)\dots\beta_p(t)}}\sum_{i=1}^p\sqrt{\delta_i}$. Indeed, we have

$$\begin{aligned}\sqrt{\delta_{1:p}} &\le \tfrac{1}{\sqrt{\beta_p(t)}}\left(\sqrt{\delta_{1:(p-1)}}+\sqrt{\delta_p}\right)\\ &\le \tfrac{1}{\sqrt{\beta_p(t)}}\left(\tfrac{1}{\sqrt{\beta_{p-1}(t)}}\left(\sqrt{\delta_{1:(p-2)}}+\sqrt{\delta_{p-1}}\right)+\sqrt{\delta_p}\right)\\ &\le \tfrac{1}{\sqrt{\beta_p(t)\beta_{p-1}(t)}}\left(\sqrt{\delta_{1:(p-2)}}+\sqrt{\delta_{p-1}}+\sqrt{\delta_p}\right)\\ &\vdots\\ &\le \tfrac{1}{\sqrt{\beta_2(t)\dots\beta_i(t)}}\sum_{j=1}^i\sqrt{\delta_j}.\end{aligned}$$

Step 3: show that $E_2(t)\wedge\cdots\wedge E_p(t)$ and the assumption $\sum_{i=1}^p\sqrt{\delta(\mathcal{K}_i)}\le\left(1-4^{-\frac{1}{p-1}}-t\right)\sqrt{n}/6$ implies that

$\beta_i(t) \geq 4^{-\frac{1}{k-1}}$ **for** $i \in 2:p$**.** We prove this by induction on i. First we show that $\beta_2(t) \geq 4^{-\frac{1}{k-1}}$. Indeed, we have

$$\begin{aligned}\beta_2(t) &= 1 - \tfrac{3}{\sqrt{n}}\left(\sqrt{\delta_1} + \sqrt{\delta_2}\right) - t \\ &\geq 1 - \tfrac{3}{\sqrt{n}} \frac{\left(1-4^{-\frac{1}{k-1}}-t\right)\sqrt{n}}{6} - t \geq 4^{-\frac{1}{k-1}}.\end{aligned}$$

Next for any $i \in 3:k$, we assume that $\beta_j(t) \geq 4^{-\frac{1}{k-1}}$ for all $2 \leq j \leq (i-1)$, then we have

$$\begin{aligned}\beta_i(t) &= 1 - \tfrac{3}{\sqrt{n}}\left(\sqrt{\delta_{1:(i-1)}} + \sqrt{\delta_i}\right) - t \\ &\geq 1 - \tfrac{3}{\sqrt{n}} \frac{1}{\sqrt{\beta_2(t)\ldots\beta_{i-1}(t)}} \sum_{j=1}^{i} \sqrt{\delta_j} - t \\ &\geq 1 - \tfrac{3}{\sqrt{n}} 2^{\frac{i-2}{k-1}} \frac{\left(1-4^{-\frac{1}{k-1}}-t\right)\sqrt{n}}{6} - t \geq 4^{-\frac{1}{k-1}}.\end{aligned}$$

Finally, combining all three steps, we can conclude that $\mathbb{P}\left[\sqrt{\delta\left(\sum_{i=1}^{p} Q_i\mathcal{K}_i\right)} \leq 2\sum_{i=1}^{p}\sqrt{\delta(\mathcal{K}_i)}\right] \geq 1-(p-1)\exp(-\frac{n-2}{8}t^2)$. □

G. Proof for Proposition 5

Throughout this proof, we define the following notations for all $i \in 1:k$:

- $\mathcal{D}_i = \mathcal{D}(\mathcal{A}_i, x_i^\natural)$;
- $\hat{\mathcal{D}}_i = \mathcal{D}(\hat{\mathcal{A}}_i, \hat{x}_i^\natural)$;
- $\delta_i = \delta(\mathcal{D}_i)$;
- $\delta_{1:i} = \delta\left(\sum_{j=1}^{i}\mathcal{D}_i\right)$;
- $\delta_{-i} = \delta\left(\sum_{j\neq i}\mathcal{D}_j\right)$.

By Lemma 1(c), for all $i \in 1:k$, we have

$$\mathcal{D}_i = \mathcal{D}(Q_i\hat{\mathcal{A}}_i, Q_i\hat{x}_i^\natural) = Q_i\hat{\mathcal{D}}_i.$$

Then it follows from Lemma 2(a) that

$$\delta(\hat{\mathcal{D}}_i) = \delta(Q_i^T\mathcal{D}_i) = \delta_i.$$

For all $i \in 1{:}k$, define

- $\hat{\mathcal{D}}_i \coloneqq \mathcal{D}(\hat{\mathcal{A}}_i, \hat{x}_i^\natural)$;
- $\delta_i = \delta(\mathcal{D}_i)$;
- $\delta_{-i} = \delta\left(\sum_{j\neq i}\mathcal{D}_j\right)$

First, fix $t > 0$, for each $i \in 1{:}k$, define the event

$$E_i(t) = \left\{\cos\angle\left(-\mathcal{D}_i, \sum_{j\neq i}\mathcal{D}_j\right) \leq \tfrac{3}{\sqrt{n}}\left(\sqrt{\delta_i} + \sqrt{\delta_{-i}}\right) + t\right\}.$$

Denote the indicator random variable for $E_i(t)$ by $\mathbb{1}_{E_i(t)}$, which evaluates to 1 if $E_i(t)$ occurs and otherwise evaluates to 0. Then, the following chain of inequalities gives the upper bound for the probability of the event $E_i(t)$:

$$\begin{aligned}\mathbb{P}(E_i(t)) &= \mathbb{E}(\mathbb{1}_{E_i(t)}) \\ &= \mathbb{E}_{\{Q_j\}_{j\neq i}}\mathbb{E}\left[\mathbb{1}_{E_i(t)} \;\middle|\; Q_j\ \forall j \neq i\right] \\ &\geq \mathbb{E}_{\{Q_j\}_{j\neq i}}[1-\exp(-\tfrac{n-2}{8}t^2)] \\ &= 1-\exp(-\tfrac{n-2}{8}t^2),\end{aligned} \tag{22}$$

where the inequality follows from Proposition 4.

Next, by Lemma 6, we know that

$$\mathbb{P}\left[\sqrt{\delta_{-i}} \leq 2\sum_{j\neq i}\sqrt{\delta_j}\right] \geq 1-(k-2)\exp(-\tfrac{n-2}{8}t^2). \tag{23}$$

Thirdly, for each $i \in 1{:}k$, define the event,

$$\hat{E}_i(t) = \left\{\cos\angle\left(-\mathcal{D}_i, \sum_{j\neq i}\mathcal{D}_j\right) \leq \tfrac{6}{\sqrt{n}}\sum_{i=1}^{k}\sqrt{\delta_i} + t\right\}.$$

By combining Equation (22) and Equation (23) together, we can conclude that

$$\mathbb{P}(\hat{E}_i(t)) \geq 1-(k-1)\exp(-\tfrac{n-2}{8}t^2).$$

Extend the bound on $\mathbb{P}(\hat{E}_i(t))$ to all $i \in 1{:}k$ via the union bound:

$$\mathbb{P}(\hat{E}_1 \wedge \cdots \wedge \hat{E}_k) \geq 1-k(k-1)\exp(-\tfrac{n-2}{8}t^2).$$

Finally, by our assumption that $\sum_{i=1}^{k}\sqrt{\delta(\mathcal{D}_i)} \leq \left(1-4^{-\frac{1}{k-1}}-t\right)\sqrt{n}/6$, it follows that

$$\tfrac{6}{\sqrt{n}}\sum_{i=1}^{k}\sqrt{\delta_i} + t \leq 1-4^{-\frac{1}{k-1}}.$$

Therefore, we can conclude that the rotated pairs $\{(x_i^\natural, \mathcal{A}_i)\}_{i=1}^{k}$ are $4^{-\frac{1}{k-1}}$-incoherent with probability at least $1-k(k-1)\exp(-\frac{n-2}{8}t^2)$.

Acknowledgment

The authors would like to thank...

References

[1] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with dirichlet prior," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 33–36.

[2] A. Quirós and S. P. Wilson, "Dependent gaussian mixture models for source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 239, 2012.

[3] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inform. Theory*, vol. 60, no. 3, pp. 1711–1732, 2013.

[4] T.-H. Chan, W.-K. Ma, C.-Y. Chi, and Y. Wang, "A convex analysis framework for blind separation of non-negative sources," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 5120–5134, 2008.

[5] J. Bobin, J.-L. Starck, J. M. Fadili, Y. Moudden, and D. L. Donoho, "Morphological component analysis: An adaptive thresholding strategy," *IEEE transactions on image processing*, vol. 16, no. 11, pp. 2675–2681, 2007.

[6] V. Chandrasekaran, B. Recht, P. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[7] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proceedings of the 7th European Conference on Computer Vision-Part I*, ser. ECCV '02. Berlin, Heidelberg: Springer-Verlag, 2002, p. 447–460.

[8] B. Savas and L. Eldén, "Handwritten digit classification using higher order singular value decomposition," *Pattern Recognition*, vol. 40, p. 993–1003, 2007.

[9] D. Carando and S. Lassalle, "Atomic decompositions for tensor products and polynomial spaces," *Journal of mathematical analysis and applications*, vol. 347, no. 1, pp. 243–254, 2008.

[10] Z. Fan, H. Jeong, Y. Sun, and M. P. Friedlander, "Atomic decomposition via polar alignment: The geometry of structured optimization," *Foundations and Trends in Optimization*, vol. 3, no. 4, pp. 280–366, 2020. [Online]. Available: http://dx.doi.org/10.1561/2400000028

[11] M. P. Friedlander, I. Macêdo, and T. K. Pong, "Polar convolution," *SIAM J. Optim.*, vol. 29, no. 4, pp. 1366–1391, 2019.

[12] R. T. Rockafellar, *Convex Analysis*. Princeton: Princeton University Press, 1970.

[13] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics (NRL)*, vol. 3, no. 1-2, pp. 95–110, 1956.

[14] M. Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization." in *ICML (1)*, 2013, pp. 427–435.

[15] J. F. Claerbout and F. Muir, "Robust modeling with erratic data," *Geophysics*, vol. 38, no. 5, pp. 826–844, 1973.

[16] J.-L. Starck, Y. Moudden, J. Bobin, M. Elad, and D. Donoho, "Morphological component analysis," in *Wavelets XI*, vol. 5914. International Society for Optics and Photonics, 2005, p. 59140Q.

[17] M. B. McCoy and J. A. Tropp, "Sharp recovery bounds for convex demixing, with applications," *Found. Comput. Math.*, vol. 14, no. 3, pp. 503–567, 2014.

[18] ——, "The achievable performance of convex demixing," *arXiv preprint arXiv:1309.7478*, 2013.

[19] S. Oymak and J. A. Tropp, "Universality laws for randomized dimension reduction, with applications," *IMA Inform. Inference*, vol. 7, no. 3, pp. 337–446, 2017.

[20] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 2845–2862, November 2001. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=959265

[21] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, 2003. [Online]. Available: http://www.pnas.org/cgi/content/abstract/100/5/2197

[22] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[23] A. Maleki, "Coherence analysis of iterative thresholding algorithms," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2009, pp. 236–243.

[24] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. Assoc. Comput. Mach.*, vol. 58, no. 3, p. 11, 2011.

[25] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *IMA Inform. Inference*, vol. 2, no. 1, pp. 32–68, 2013.

[26] J. A. Tropp, "Convex recovery of a structured signal from independent random linear measurements," in *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[27] Y. Gordon, "On Milman's inequality and random subspaces which escape through a mesh in R^n," in *Geometric Aspects of Functional Analysis*. Springer, 1988, pp. 84–106.

[28] D. G. Obert, "The angle between two cones," *Linear Algebra and its Applications*, vol. 144, pp. 63–70, 1991.

[29] E. van den Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM J. Sci. Comput.*, vol. 31, no. 2, pp. 890–912, 2008. [Online]. Available: http://link.aip.org/link/?SCE/31/890

[30] ——, "Sparse optimization with least-squares constraints," *SIAM J. Optim.*, vol. 21, no. 4, pp. 1201–1229, 2011.

[31] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and S. Roy, "Level-set methods for convex optimization," *Math. Program., Ser. B*, vol. 174, no. 1-2, pp. 359–390, December 2018.

[32] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," November 2014.

[33] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[34] M. B. McCoy, V. Cevher, Q. T. Dinh, A. Asaei, and L. Baldassarre, "Convexity in source separation: Models, geometry, and algorithms," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 87–95, 2014.

[35] M. Fazel and J. Goodman, "Approximations for partially coherent optical imaging systems," *Technical Report*, 1998.

[36] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *American Control Conference*, Arlington, 2001.

[37] Y. Pati and T. Kailath, "Phase-shifting masks for microlithography: automated design and mask requirements," *JOSA A*, vol. 11, no. 9, pp. 2438–2452, 1994.

[38] L. G. Valiant, "Graph-theoretic arguments in low-level complexity," in *International Symposium on Mathematical Foundations of Computer Science*. Springer, 1977, pp. 162–176.

[39] F. Ong and M. Lustig, "Beyond low rank+ sparse: Multiscale low rank matrix decomposition," *IEEE journal of selected topics in signal processing*, vol. 10, no. 4, pp. 672–687, 2016.

[40] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[41] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. New York, NY: Springer, 2001.

[42] J. A. Tropp, "A comparison principle for functions of a uniformly random subspace," *Probability Theory and Related Fields*, vol. 153, no. 3, pp. 759–769, 2012.

[43] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.

[44] E. S. Meckes, *The random matrix theory of the classical compact groups*. Cambridge University Press, 2019, vol. 218.