Demixing Sparse Signals from Nonlinear Observations

Mohammadreza Soltani and Chinmay Hegde
Electrical and Computer Engineering Department
Iowa State University

March 30, 2016

Abstract

We study the problem of *demixing* a pair of sparse signals from nonlinear observations of their superposition. Mathematically, we consider the observation model $y = f(Ax)$, where $y \in \mathbb{R}^n$ represents the observations, f is a nonlinear link function, $A \in \mathbb{R}^{m \times n}$ is a measurement operator, and $x \in \mathbb{R}^n$ is the superposition of the signals. Further, we assume that $x = \Phi w + \Psi z$, where Φ, Ψ are *incoherent* bases of $\mathbb{R}^n$, and w and z are s-sparse coefficient vectors. Problems of this nature arise in computer vision and astronomical imaging.

In this paper, we provide a simple and fast algorithm (that we call OneShot) to demix the constituent signals w and z from the observations y. Our algorithm is non-iterative, does not require explicit knowledge of the link function, and is able to recover w and z even in the case where $m \ll n$ (provided the bases are incoherent enough.) Moreover, under certain assumptions, the number of measurements provably required for successful demixing by OneShot is given by $m = O(s \log \frac{n}{s})$, therefore matching the best-known bounds for the (easier) case of linear observations. Finally, we verify the utility of our proposed tools via numerical experiments.

1 Introduction

In several applications in signal processing, data analysis, and statistics, the problem of signal *demixing* assumes a special importance. Here, the goal is to recover a set of signals from their linear superposition. In this context, a common, well-studied observation model is given by:

$$x = \Phi x_1 + \Psi x_2, \tag{1.1}$$

where $x \in \mathbb{R}^m$ represents the observations, $\Phi, \Psi \in \mathbb{R}^{m \times n}$ are bases (or dictionaries), and $x_1, x_2 \in \mathbb{R}^n$ are the coefficients of the constituent signals. Then, the demixing problem refers to recover x_1 and x_2 from the measurements x. This problem has been studied in the context of various applications [1–3]. In image processing and computer vision applications, this problem models tasks such as background-foreground separation [4, 5], while in astronomical imaging applications, this problem represents the separation of astronomical features (stars/galaxies) from background sky phenomena.

At first glance, a natural question is whether such a demixing of signals from their superposition is even possible. Clearly, the answer is no. For instance, suppose that we are given an observation vector x which is modeled as $x = x_1 + x_2$ where x_1, x_2 and $x \in \mathbb{R}^n$. Also assume that x_1 and x_2 both have only one nonzero entry in the first coordinate. Under these assumptions, there is an infinite number of x_1 and x_2 that do satisfy this observation model.

In general, demixing is an ill-posed problem since the number of unknowns, $2n$, is greater than the number of equations, n. This simple example reveals the fact that the recovery of the constituent signals is impossible if these component bases are strongly aligned. In Section 3 we use a quantity known as *incoherence* that captures the degree of alignment of the bases.

However, even if we assume that the constituent signals are incoherent enough, demixing poses a very challenging problem under more general observation models. For instance, assume the measurement model is given by $y = A(\Phi w + \Psi z)$, where $A \in \mathbb{R}^{m \times n}$ denotes a linear measurement operator, and $x, y \in \mathbb{R}^n$. Let us focus on the case where $m \ll n$. Such a measurement model has been the focus of considerable attention in recent advances in signal processing and high-dimensional statistics [6–8]. In this case, it might seem impossible to recover the components x and z since A possesses a nontrivial null space. Once again, this problem becomes ill-posed and additional information about the *structure* of the components is necessary. For example, in the application of separating foreground and background images, the foreground image can be modeled as *sparse* while the background can be assumed to be *low-rank*. Such model assumptions on the constituent signals have been shown to enable successful demixing [9–13].

In this paper, we focus on an *even* more challenging scenario, where the measurements y are *nonlinear* functions of the signal superposition. That is, $y = f(Ax)$ where f is a nonlinear function (sometimes called a *link* function), and x is the target superposition. Nonlinear link functions have long been studied in the statistics literature, and have recently been studied in the context of signal recovery [14–16]. Such nonlinear functions can be used to model numerous phenomena in signal acquisition, including quantization [14] and phaseless imaging [17].

We are interested in the problem of signal demixing in such observation scenarios. Mathematically, we consider the model of $y = f(Ax)$ where $A \in \mathbb{R}^{m \times n}$ is a random matrix and $x = \Phi w + \Psi z$. We exclusively focus on the case where $m \ll n$. As part of our structural assumptions, we suppose that the signal coefficients w and z are s-sparse in the basis Φ and Ψ, respectively (i.e., w and z contain no more than s nonzero entries). Then, the goal is to recover w and z, given m number of measurements, $\{y_1, y_2, \ldots, y_m\}$ and the two bases Φ and Ψ. Furthermore, f may be non-smooth, non-invertible, or even unknown. (The only restriction is that f should be odd.)

In this paper, we provide a simple, fast algorithm (that we call OneShot) to demix the constituent signals w and z from the observations y. Our algorithm is non-iterative, does not require explicit knowledge of the link function, and is able to recover w and z even in the case where $m \ll n$ (provided the bases Φ and Ψ are incoherent enough.)

We support our algorithm with a rigorous theoretical analysis. Our analysis leads us to prove upper bounds for the *sample complexity* (i.e., the number of requisite observations for successful demixing) as well as the estimation error achieved by our proposed algorithm. In particular, under certain conditions, we prove that the sample complexity of OneShot is given by $m = O(s \log \frac{n}{s})$; this matches the best-known bounds for the (easier) case of linear observations. Formally, assume that Φ and Ψ are two incoherent basis with incoherence ε and $x = \Phi w + \Psi z$ where $\|w\|_0 \leq s$, $\|w\|_0 \leq s$ and $x, w, z \in \mathbb{R}^n$. If ε is a small constant (say $\varepsilon \leq 0.25$), then with $m = O(s \log \frac{n}{s})$ observations, we can recover x, as well as the

constituents w, z, up to accuracy $O(\varepsilon)$. See Section 4 for details.

Our technique to prove our desired upper bounds for ONESHOT is based on the approach proposed in [18]. Our main conceptual contribution is to extend this approach for the (more general) nonlinear demixing problem. The theoretical bounds emerging from our approach serve to highlight the role of signal incoherence in algorithm performance. Our technique follows a geometric argument, and leverages the *Gaussian mean width* for the set of sparse vectors. The Gaussian mean width is a geometric measure of complexity of a set of points in a given space, and tight bounds for the Gaussian width of the set of s-sparse vectors in the unit ball in $\mathbb{R}^n$ are available. We use the bounds in deriving our sample complexity results.

Moreover, we provide numerical evidence for the efficiency of our methods. In particular, we compare the performance of ONESHOT with a previous method proposed in [18] that is based on convex optimization. The original idea of this method [18] is not directly related to the nonlinear demixing problem, but it was proposed for nonlinear signal recovery when the underlying signal x belongs to the set of approximately s-sparse signals. Simulation results show that ONESHOT outperforms this convex method significantly in both statistical efficiency as well as running time, and consequently makes it an attractive choice in large-scale nonlinear demixing problems.

2 Prior Work

The problem of linear demixing has been a significant focus of study in the last several years. In a class of image processing techniques known as *morphological component analysis* (MCA), images are treated as the superposition of a few structured constituents, and the goal is to reliably recover these constituents given knowledge of their structure [1,2]. Typical structural assumptions on the signal constituents include sparsity in a given basis or dictionary. In [19], the authors describe a number of applications of similar demixing techniques in audio processing and hardware diagnosis.

Research in the linear demixing problem has also considered a variety of signal models beyond sparsity-based models. The *robust PCA* problem [9–11] involves inferring a low-rank matrix L and a sparse matrix S, given their sum $M = L + S$. In the machine learning community, the separation of low-rank and sparse matrices is used for latent variable model selection [20] and the robust alignment of multiple occluded images [21]. A different (more general) signal model is the *low-dimensional manifold* model. For instance, in [12, 13], the authors propose a greedy iterative method for demixing signals, arising from a mixture of known manifolds by iterative projections onto the manifolds. We refer to [3] for a comprehensive discussion about the linear demixing problem with various applications.

The linear demixing problem falls under the general category of *inverse problems*, where the number of unknown parameters in an estimation problem far exceeds the number of observations. In signal processing applications, such problems have been widely studied under the moniker of *compressive sensing* [6–8], where the goal is to reconstruct a given signal from a small number of linear measurements. Of particular relevance to us is the 1-*bit compressive sensing* problem. Here, the linear measurements are quantized to a single bit where the measurements are binary (± 1) and only comprise the sign of the signal coefficient (i.e., the amplitude information is completely absorbed by the quantization operator.) Recent

results have shown that if the signal x is sparse enough, this recovery problem can be done efficiently using convex optimization [14–16].

The nonlinear measurement model is also related to classical observation models in statistics. variously known as the *single index model* (SIM), or the *Generalized Linear Model* (GLM). In SIM, the unknown nonlinear function f is assumed to be odd and non-decreasing, and the goal is to estimate f as well as the parameter vector x. For representative algorithmic solutions in the machine learning community, see [22, 23]. Recently, [24] proposed an approach for estimating the coefficients based on a novel second moment estimator. An advantage of their approach is that the unknown link function can be either odd or even.

Our method closely follows, and can be viewed as a generalization of, the method developed in [18]. Here, the authors consider the problem of recovering sparse signals from general nonlinear measurements. Their results implicitly assume that the (possibly unknown) nonlinear link function is odd. Using geometric arguments, they provide a minimax-optimal rate for their bound up to constants. Quantitatively, their results indicate that the required number of measurements for successful nonlinear recovery is proportional to the *effective dimension* of the set of sparse signals; the effective dimension is defined as the square of the *Gaussian mean width*, a geometric notion that captures the intrinsic complexity of a set.

None of the above works have (explicitly) considered the problem of *demixing* signals from nonlinear observations. In this work, we address this problem and provide a very fast algorithm with provable guarantees.

3 Mathematical Setup

In this section, we establish the formal mathematical model and introduce some definitions. Consider the *Gaussian* observation model:

$$y = f(Ax)\,, \tag{3.1}$$

where $A \in \mathbb{R}^{m\times n}$ is a random matrix with i.i.d. standard normal entries, and $x \in \mathbb{R}^n$ is given by sum of two s-sparse vectors in two different basis. That is,

$$x = \Phi w + \Psi z\,, \tag{3.2}$$

where $\Phi, \Psi \in \mathbb{R}^{n\times n}$ are orthonormal bases, and $w, z \in \mathbb{R}^n$ such that $\|w\|_0 \le s_1$, and $\|z\|_0 \le s_2$. Here, f denotes a (possibly unknown) odd link function which is not necessarily smooth, invertible, or continuous. Otherwise mentioned, the symbol $\|\cdot\|$ refers to the ℓ_2-norm.

We define the following quantities:

$$\bar{x} = \frac{\Phi\bar{w} + \Psi\bar{z}}{\|\Phi\bar{w} + \Psi\bar{z}\|} = \alpha(\Phi\bar{w} + \Psi\bar{z}),$$

where:

$$\alpha = \frac{1}{\|\Phi\bar{w} + \Psi\bar{z}\|}, \quad \bar{w} = \frac{w}{\|w\|}, \quad \bar{z} = \frac{z}{\|z\|}.$$

Also, we assume that w and z belong to the following sets:

$$w \in \mathcal{K}_1 = \{\Phi a \mid \|a\|_0 \le s_1\},$$
$$z \in \mathcal{K}_2 = \{\Psi a \mid \|a\|_0 \le s_2\},$$

and we define $K = \{a \mid \|a\|_0 \leq s\}$.

Definition 3.1. *(Linear estimate of x). The linear estimate of x is given by $\hat{x}_{lin} = \frac{1}{m}A^T y = \frac{1}{m}\Sigma_{i=1}^m y_i a_i$ where $a_i \in \mathbb{R}^n$ is the i-th row of A and we have $a_i \sim \mathcal{N}(0, I)$.*

Definition 3.2. *(ε-incoherence). The set of s-sparse vectors in basis Φ and Ψ are said to be ε-incoherent where:*

$$\varepsilon = \sup_{\substack{\|u\|_0 \leq s,\ \|v\|_0 \leq s \\ \|u\|_2 = 1,\ \|v\|_2 = 1}} |\langle \Phi u, \Psi v \rangle|.$$

Definition 3.3. *(Polar norm) For a given $x \in \mathbb{R}^n$ and a subset of $Q \in \mathbb{R}^n$, the polar norm with respect to Q is defined as follows:*

$$\|x\|_{Q^o} = \sup_{u \in Q} \langle x, u \rangle.$$

Furthermore, for a given subset of $Q \in \mathbb{R}^n$, we define $Q_t = (Q - Q) \cap tB_2^n$. Since Q_t is a symmetric set, it can be shown that the polar norm with respect to Q_t defines a semi-norm.

Definition 3.4. *(Local gaussian mean width). For a given set $K \in \mathbb{R}^n$, the local gaussian mean width (or simply, the local mean width) is defined as follows $\forall\ t > 0$:*

$$W_t(K) = \mathbb{E} \sup_{x,y \in K, \|x-y\| \leq t} \langle g, x - y \rangle.$$

4 A Fast Demixing Algorithm

Having defined these quantities, we are ready to present our algorithm and main theoretical results. Recall that we wish to recover components w and z in equation 3.2. Our proposed algorithm, that we call it OneShot, is given in Algorithm 1.

Algorithm 1 OneShot

Inputs: Basis matrices Φ and Ψ, measurement matrix A, measurements y, sparsity level s.
Outputs: Estimates $\widehat{x} = \Phi\hat{w} + \Psi\widehat{z}$, $\widehat{w} \in \mathcal{K}_1$, $\widehat{z} \in \mathcal{K}_2$

$\widehat{x}_{\text{lin}} \leftarrow \frac{1}{m}A^T y$ {form linear estimation}
$b_1 \leftarrow \Phi^* \widehat{x}_{lin}$ {forming first proxy}
$\widehat{w} \leftarrow \mathcal{P}_s(b_1)$ {Projection on set $\mathcal{K}_1$}
$b_2 \leftarrow \Psi^* \widehat{x}_{lin}$ {forming second proxy}
$\widehat{z} \leftarrow \mathcal{P}_s(b_2)$ {Projection on set $\mathcal{K}_2$}
$\widehat{x} \leftarrow \Phi\widehat{w} + \Psi\widehat{z}$ {Estimating $\widehat{x}$}

In OneShot, we assume that the sparsity levels s_1 and s_2 defined in the sets $\mathcal{K}_1$ and $\mathcal{K}_2$ are equal to each other, *i.e*, $s_1 = s_2 = s$. The algorithm effortlessly extends to the case of unequal sparsity levels. Also, we have used the following projections:

$$\widehat{w} = \mathcal{P}_s(\Phi^* \widehat{x}_{lin}), \quad \hat{z} = \mathcal{P}_s(\Psi^* \hat{x}_{lin}).$$

Here, $\mathcal{P}_s$ denotes projection onto $\mathcal{K}$ and can be implemented by hard thresholding. It is important to note that OneShot is not an iterative algorithm; this fact enables us to achieve a fast running time in large-scale problems.

In our proofs, we use a result from [18] that we state for completeness.

Lemma 4.1. *(Quality of linear estimate). Given the model in Equation 3.1, the linear estimator (Def. 3.1) is an unbiased estimator of $\bar{x}$ (up to constants). That is, $\mathbb{E}(\widehat{x}_{lin}) = \mu\bar{x}$ and the variance of estimator is given by:*

$$\mathbb{E}\|\widehat{x}_{lin} - \mu\bar{x}\|_2^2 = \frac{1}{m}[\sigma^2 + \eta^2(n-1)],$$

where

$$\mu = \mathbb{E}(y_1\langle a_1, \bar{x}\rangle), \sigma^2 = Var(y_1\langle a_1, \bar{x}\rangle), \eta^2 = \mathbb{E}(y_1^2).$$

Theorem 4.2. *(Main theorem.) Let $y \in \mathcal{R}^m$ be given the set of measurements. Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d. standard normal entries. Also, let $\Phi, \Psi \in \mathbb{R}^{n\times n}$ are bases with incoherence ε, as defined in Definition 3.2. If we use* Oneshot *to recover w and z described in equations 3.1 and 3.2, then the error of nonlinear estimation incurred by $\hat{x}$, the output of* Oneshot*, satisfies the following upper bound:*

$$\mathbb{E}\|\widehat{x} - \mu\bar{x}\| \leq \frac{4\sqrt{2}\sigma}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4\sqrt{2}\mu\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C\eta}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}}. \tag{4.1}$$

where $C > 0$ is an absolute constant. The coefficients μ, σ, and η are given in Lemma 4.1. Similar upper bounds can be obtained for estimates of the constituent signals w, z via the triangle inequality.

Proof. See Section 5. □

Corollary 4.3. *(Example quantitative result). If $f(x) = \text{sign}(x)$, then we may substitute*

$$\mu = \sqrt{\frac{2}{\pi}} \approx 0.8, \qquad \sigma^2 = 1 - \frac{2}{\pi} \approx 0.6, \qquad \eta^2 = 1,$$

in the above statement. Hence, the bound in 4.1 becomes:

$$\mathbb{E}\|\widehat{x} - \mu\bar{x}\| \leq \frac{4}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4.53\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}}\,. \tag{4.2}$$

Proof. Using Lemma 4.1, $\mu = \mathbb{E}(y_1\langle a_i, \bar{x}\rangle)$ where $y_1 = \text{sign}(\langle a_i, x\rangle)$. Since $a_i \sim \mathcal{N}(0, I)$ and $\bar{x}$ has unit norm, $\langle a_i, \bar{x}\rangle \sim \mathcal{N}(0, 1)$. Thus, $\mu = \mathbb{E}|g| = \sqrt{\frac{2}{\pi}}$. Moreover, we can write $\sigma^2 = \mathbb{E}(|g|^2) - \mu^2 = 1 - \frac{2}{\pi}$. Here, we have used the fact that $|g|^2$ obeys the $\tilde{\chi}_1^2$ distribution with mean 1. Finally, $\eta^2 = \mathbb{E}(y_1^2) = 1$. □

Remark 4.4. *(Implications). The upper bound on the estimation error in the main theorem shows that to achieve an estimation error of $O(\varepsilon)$, the following number of measurements suffices:*

$$m = \mathcal{O}\left(\frac{s}{\varepsilon^2}\log\frac{n}{s}\right).$$

For constant ε, this matches the sample complexity of demixing in the linear case [13].

The main theorem obtains a bound on the expected value of the estimation error. We can derive a similar upper bound that holds with high probability. In this theorem, we assume that measurements y_i for $i = 1, 2, \ldots, m$ have a *sub-gaussian* distribution. See [25] for a comprehensive discussion about sub-gaussian random variables.

Theorem 4.5. *(High-probability version of main theorem.) Let $y \in \mathcal{R}^m$ be a set of measurements with a sub-gaussian distribution. Assume that $A \in \mathbb{R}^{m\times n}$ is a random matrix with i.i.d standard normal entries. Also, assume that $\Phi, \Psi \in \mathbb{R}^{n\times n}$ are two bases with incoherence ε as in Definition 3.2. Let $0 \le s \le \sqrt{m}$. If we use* ONESHOT *to recover w and z described in* (3.1) *and* (3.2)*, then the estimation error of the output of* ONESHOT *satisfies the following:*

$$\|\widehat{x} - \mu\bar{x}\| \le \frac{2\sqrt{2}\eta s}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4\sqrt{2}\mu\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C\eta}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}} + 4\frac{\eta s}{\sqrt{m}},$$

with probability at least $1 - 4\exp(-\frac{cs^2\eta^4}{\|y_1\|_{\psi_2}^4})$ where $C, c > 0$ are absolute constants. The coefficients μ, σ, and η are given in Lemma 4.1. Here, $\|y_1\|_{\psi_2}$ denotes the ψ_2-norm of the first measurement y_1.

Proof. See Section 5. □

5 Proofs

In this section, we derive the proof of Theorem 4.2. The proof mostly follows the method of [18]. As a precursor, we need the following lemma from geometric functional analysis, restated from [18].

Lemma 5.1. *Assume K is a closed star-shaped set. Then for $u \in K$, and $a \in \mathbb{R}^n$, one has the following result $\forall t > 0$:*

$$\|\mathcal{P}_K(a) - u\|_2 \le \max(t, \frac{2}{t}\|\mathcal{P}_K(a) - u\|_{K_t^o}). \tag{5.1}$$

We also use the following result of [18].

Claim 5.2. *(orthogonal decomposition of a_i) Suppose we decompose the rows of A, a_i, as:*

$$a_i = \langle a_i, \bar{x}\rangle\bar{x} + b_i, \tag{5.2}$$

where $b_i \in \mathbb{R}^n$ is orthogonal to $\bar{x}$. Then we have $b_i \sim \mathcal{N}(0, I_{x^\perp})$ since $a_i \sim \mathcal{N}(0, I)$. Also, $I_{x^\perp} = I - \bar{x}\bar{x}^T$. Moreover, the measurements y_i in equation 3.1 and the orthogonal component b_i are statistically independent.

Proof of Theorem 4.2. Observe that the magnitude of the signal x may be completely lost due to the action of the nonlinear measurement function f (such as the sign$(\cdot)$ function). Therefore, our recovered signal $\widehat{x}$ approximates the true signal modulo (possibly unknown)

scaling factor. Indeed, for μ defined in Lemma 4.1, we have:

$$\begin{aligned}\|\widehat{x} - \mu\bar{x}\| &= \|\Phi\widehat{w} + \Psi\widehat{z} - \alpha\mu\Phi\bar{w} - \alpha\mu\Psi\bar{z}\| \\ &\leq \|\Phi\|\|\widehat{w} - \mu\alpha\bar{w}\| + \|\Psi\|\|\widehat{z} - \mu\alpha\bar{z}\| \\ &\leq (t + \frac{2}{t}\|\Phi^*\widehat{x}_{lin} - \mu\alpha\bar{w}\|_{K_t^o}) + (t + \frac{2}{t}\|\Psi^*\widehat{x}_{lin} - \mu\alpha\bar{z}\|_{K_t^o})\,.\end{aligned}$$

The equality comes from the definition of $\bar{x}$. The first inequality results from an application of the triangle inequality and the definition of the operator norm of a matrix, while the second inequality follows from Lemma 5.1.

It suffices to derive a bound on the first term in the above expression (since a similar bound will hold for the second term.) We obtain:

$$\begin{aligned}\|\Phi^*\widehat{x}_{lin} - \mu\alpha\bar{w}\|_{K_t^o} &= \|\Phi^*\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle\bar{x} + y_i b_i) - \mu\alpha\bar{w}\|_{K_t^o} \\ &\leq \|\Phi^*\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle\bar{x}) - \mu\alpha\bar{w}\|_{K_t^o} + \|\Phi^*\frac{1}{m}\Sigma_i y_i b_i\|_{K_t^o} \\ &\leq \underbrace{\|\Phi^*\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle\bar{x}) - \mu\Phi^*\bar{x}\|_{K_t^o}}_{S_1} + \underbrace{\|\mu\alpha\Phi^*\Psi\bar{z}\|_{K_t^o}}_{S_2} + \underbrace{\|\Phi^*\frac{1}{m}\Sigma_i y_i b_i\|_{K_t^o}}_{S_3}\,.\end{aligned} \tag{5.3}$$

The first equality follows from the orthogonal decomposition of a_i, while the second and third inequalities result from the triangle inequality. We first bound S_1 as follows:

$$\begin{aligned}S_1 &= \|\Phi^*\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle\bar{x}) - \mu\Phi^*\bar{x}\|_{K_t^o} \\ &= \|(\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu))\Phi^*\bar{x}\|_{K_t^o} \\ &= |\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|\|\Phi^*\bar{x}\|_{K_t^o}.\end{aligned}$$

Therefore,

$$\mathbb{E}(S_1^2) = \mathbb{E}(|\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|^2\|\Phi^*\bar{x}\|_{K_t^o}^2).$$

Define $\gamma_i \stackrel{\Delta}{=} y_i\langle a_i, \bar{x}\rangle - \mu_i$. Then,

$$\begin{aligned}\mathbb{E}(|\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|^2) &= \mathbb{E}(\frac{1}{m^2}(\Sigma_i\gamma_i)^2) \\ &= \mathbb{E}(\frac{1}{m^2}(\sum_{i=}^{m}\gamma_i^2 + \Sigma_{i\neq j}\gamma_i\gamma_j)) \\ &= \frac{1}{m^2}(\sum_{i=1}^{m}\mathbb{E}\gamma_i^2) = \frac{1}{m}\mathbb{E}\gamma_1^2 \\ &= \frac{\sigma^2}{m}\end{aligned}$$

where σ^2 has been defined in Lemma 4.1. The third and last equalities follow from the fact that the y_i's are independent and identically distributed.

Now, we bound $\|\Phi^*\bar{x}\|_{K_t^o}^2$ as follows::

$$\begin{aligned}
\|\Phi^*\bar{x}\|_{K_t^o} &= \sup_{u\in(K-K)\cap tB_n^2} \langle \Phi^*\bar{x}, u\rangle \\
&= t \sup_{\substack{v_1\in\frac{1}{t}K, v_2\in\frac{1}{t}K \\ \|v_i\|\le 1, i=1,2}} \langle \Phi^*\bar{x}, v_1 - v_2\rangle \\
&\le 2t \sup_{\substack{\|a\|_0\le s \\ \|a\|\le 1}} |\langle \Phi^*\bar{x}, a\rangle| \\
&\le 2t\Big(\sup_{\substack{\|a\|_0\le s \\ \|a\|\le 1}} |\langle \alpha\bar{w}, a\rangle| + \sup_{\substack{\|a\|_0\le s \\ \|a\|\le 1}} |\langle \alpha\Phi^*\Psi\bar{z}, a\rangle|\Big) \\
&\le 2\alpha t(1 + \sup_{\substack{\|a\|_0\le s \\ \|a\|\le 1}} |\langle \alpha\Psi\bar{z}, \Phi a\rangle|) \\
&= 2\alpha t(1+\varepsilon).
\end{aligned}$$

This implies that:

$$\implies \mathbb{E}(S_1^2) \le 4\frac{\alpha^2 t^2\sigma^2}{m}(1+\varepsilon)^2. \tag{5.4}$$

The second inequality follows from (3.2) and the triangle inequality. The last inequality is results from an application of the Cauchy-Schwarz inequality and the definition of ε.

Similarly, we can bound S_2 as follows:

$$\begin{aligned}
\mathbb{E}(S_2) &= \mathbb{E}(\|\mu\alpha\Phi^*\Phi\bar{z}\|_{K_t^o}) \\
&= \mathbb{E}(|\mu\alpha|\|\Phi^*\Phi\bar{z}\|_{K_t^o}) \\
&= |\mu\alpha|\|\Phi^*\Phi\bar{z}\|_{K_t^o} \\
&= |\mu\alpha| \sup_{u\in(K-K)\cap tB_n^2} \langle \Psi\bar{z}, \Phi u\rangle \\
&= |\mu\alpha| t \sup_{\substack{v_1\in\frac{1}{t}K, v_2\in\frac{1}{t}K \\ \|v_i\|\le 1, i=1,2}} \langle \Psi\bar{z}, \Phi(v_1 - v_2)\rangle \\
&\le 2\mu\alpha t\varepsilon.
\end{aligned} \tag{5.5}$$

Finally, we give the bound for S_3. Define $L \stackrel{\Delta}{=} \frac{1}{m}\Sigma_i y_i b_i$. Then, we get:

$$\mathbb{E}(S_3) = \mathbb{E}\|\Phi^*\frac{1}{m}\Sigma_i y_i b_i\|_{K_t^o} = \mathbb{E}\|\Phi^* L\|_{K_t^o}.$$

Our goal is to bound $\mathbb{E}\|\Phi^* L\|_{K_t^o}$. Since y_i and b_i are independent random variables (as per Claim 5.2), we can use the law of conditional covariances and the law of iterated expectation. That is, we first condition on y_i, and then take expectation with respect to b_i.

By conditioning on y_i, we have $L \sim \mathcal{N}(0, \beta^2 I_{x^\perp})$ where $I_{x^\perp} = I - \bar{x}\bar{x}^T$ is the covariance of vector b_i according to claim 5.2 and $\beta^2 = \frac{1}{m^2}\Sigma_i y_i^2$. Define $g_{x^\perp} \sim \mathcal{N}(0, I_{x^\perp})$. Therefore, $L = \beta g_{x^\perp}$ Putting everything together, we get:

$$\begin{aligned}\mathbb{E}(S_3) &= \mathbb{E}\|\Phi^* L\|_{K_t^o} \\ &= \mathbb{E}\|\Phi^* \beta g_{x^\perp}\|_{K_t^o} \\ &= \beta \mathbb{E}\|\Phi^* g_{x^\perp}\|_{K_t^o}.\end{aligned}$$

We need to extend the support of distribution of $g_{x^\perp}$ and consequently L from $x^\perp$ to $\mathbb{R}^n$. This is done by the following claim in [18]:

Claim 5.3. *Let g_E be a random vector which is distributed as $\mathcal{N}(0, I_E)$. Also, assum that $\Gamma : \mathbb{R}^n \to \mathbb{R}$ is a convex function. Then, for any subspace E of $\mathbb{R}^n$ such that $E \subseteq F$, we have:*

$$\mathbb{E}(\Gamma(g_E)) \leq \mathbb{E}(\Gamma(g_F)).$$

Hence, we can orthogonally decompose $\mathbb{R}^n$ as $\mathbb{R}^n = D \oplus C$ where D is a subspace supporting $x^\perp$ and C is the orthogonal subspace onto it. Thus, $g_{\mathbb{R}^n} = g_D + g_C$ in distribution such that $g_D \sim \mathcal{N}(0, I_D),\ g_C \sim \mathcal{N}(0, I_C)$. Also, $\|.\|_{K_t^o}$ is a convex function since it is a semi-norm. Hence,

$$\begin{aligned}\mathbb{E}_D\|\Phi^* g_D\|_{K_t^o} &= \mathbb{E}_D\|\Phi^* g_D + \mathbb{E}_C(g_C)\|_{K_t^o} \\ &= \mathbb{E}_D\|\mathbb{E}_{C|D}(\Phi^* g_D + g_c)\|_{K_t^o} \\ &\leq \mathbb{E}_D\mathbb{E}_{C|D}\|\Phi^*(g_D + g_C)\|_{K_t^o} \\ &= \mathbb{E}\|\Phi^* g_{\mathbb{R}^n}\|_{K_t^o}.\end{aligned}$$

The first inequality follows from Jensen's inequality, while the second inequality follows from the law of iterated expectation. Therefore, we get:

$$\begin{aligned}\mathbb{E}\|\Phi^* L\|_{K_t^o} &= \mathbb{E}\|\Phi^* \beta g_{x^\perp}\|_{K_t^o} \\ &= \beta \mathbb{E}\|\Phi^* g_{x^\perp}\|_{K_t^o} \\ &\leq \beta \mathbb{E}\|\Phi^* g_{\mathbb{R}^n}\|_{K_t^o} \\ &= \beta \sup_{u \in (K-K) \cap t B_n^2} \langle \Phi^* \beta g_{\mathbb{R}^n}, u \rangle \\ &= \beta W_t(K).\end{aligned}$$

The last equality follows from the fact that $\Phi^* g_{\mathbb{R}^n} \sim \mathcal{N}(0, I)$. The final step is to take an expectation with respect to y_i, giving us a bound on $\mathbb{E}(S_3)$:

$$\begin{aligned}\mathbb{E}(S_3) &= \mathbb{E}\|\Phi^* L\|_{K_t^o} \\ &\leq \mathbb{E}(\beta) W_t(K) \\ &\leq \sqrt{\mathbb{E}(\beta^2)} W_t(K)\,,\end{aligned}$$

where $\beta^2 = \frac{1}{m^2}\sum_{i=1}^m y_i^2$. Hence,

$$\mathbb{E}(S_3) \leq \frac{\eta}{\sqrt{m}} W_t(K)\,. \tag{5.6}$$

Putting together the results from (5.4), (5.5), and (5.6), we have:

$$\begin{aligned}\mathbb{E}(\|\Phi^*\widehat{x}_{lin} - \mu\alpha\bar{w}\|_{K_t}) &\leq \mathbb{E}(S_1) + \mathbb{E}(S_2) + \mathbb{E}(S_3) \\ &\leq \sqrt{\mathbb{E}(S_1)} + \mathbb{E}(S_2) + \mathbb{E}(S_3) \\ &\leq \frac{2\alpha t\sigma}{\sqrt{m}}(1+\varepsilon) + 2\mu\alpha t\varepsilon + \frac{\eta}{\sqrt{m}}W_t(K).\end{aligned}$$

Therefore, we obtain:

$$t + \frac{2}{t}\mathbb{E}(\|\Phi^*\widehat{x}_{lin} - \mu\alpha\bar{w}\|_{K_t}) \leq t + \frac{4\alpha\sigma}{\sqrt{m}}(1+\varepsilon) + 4\mu\alpha\varepsilon + \frac{2\eta}{t\sqrt{m}}W_t(K).$$

Hence,

$$\mathbb{E}(\|\widehat{x} - \mu\bar{x}\|) \leq 2t + \frac{8\alpha\sigma}{\sqrt{m}}(1+\varepsilon) + 8\mu\alpha\varepsilon + \frac{4\eta}{t\sqrt{m}}W_t(K).$$

Moreover, we can bound $\alpha = \frac{1}{\|\Phi\bar{w}+\Psi\bar{z}\|}$ as follows:

$$\begin{aligned}\|\Phi\bar{w} + \Psi\bar{z}\|_2^2 &\geq \|\Phi\bar{w}\|_2^2 + \|\Psi\bar{z}\|_2^2 - 2|\langle\Phi\bar{w}, \Psi z\rangle| \\ &\geq 2 - 2\varepsilon, \quad \text{or,} \\ \alpha &\leq \frac{1}{\sqrt{2}\sqrt{1-\varepsilon}}\end{aligned} \tag{5.7}$$

However, k is a closed star-shaped set, thus $W_t(k) = tW_1(K)$. We can use lemma (2.3) in [15] to bound $W_1(K)$ and let $t \to 0$. Using this, we have the final bound:

$$\mathbb{E}\|\widehat{x} - \mu\bar{x}\| \leq \frac{4\sqrt{2}\sigma}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4\sqrt{2}\mu\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C\eta}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}}\,, \tag{5.8}$$

where $C > 0$ is an absolute constant. This completes the proof of Theorem 4.2. □

As a precursor to the high probability version of the main theorem, we need a few more definitions and preliminary lemmas:

Definition 5.4. *(Sub-Gaussian random variable.) A random variable X is called sub-Gaussian if it satisfies the following:*

$$\mathbb{E}\exp\left(\frac{cX}{\|X\|_{\psi_2}^2}\right) \leq 2,$$

where $c > 0$ is an absolute constant and $\|X\|_{\psi_2}$ denotes the ψ_2-norm which is defined as follows:

$$\|X\|_{\psi_2} = \sup_{p\geq 1}\frac{1}{\sqrt{p}}(\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

Definition 5.5. *(Sub-exponential random variable.) A random variable X is called sub-gaussian if it satisfies the following:*

$$\mathbb{E}\exp\left(\frac{cX}{\|X\|_{\psi_1}}\right) \leq 2,$$

where $c > 0$ is an absolute constant and $\|X\|_{\psi_1}$ denotes the ψ_1-norm which is defined as follows:

$$\|X\|_{\psi_1} = \sup_{p\geq 1} \frac{1}{p}(\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

We should mention that there are other definitions for sub-Gaussian and sub-exponential random variables. Please see [25].

Lemma 5.6. *Let x and y be two sub-Gaussian random variables. Then, XY is a sub-exponential random variable.*

Proof. According to the definition of the ψ_2-norm, we have:

$$(\mathbb{E}|XY|^p)^{\frac{1}{p}} \leq p\|X\|_{\psi_2}\|Y\|_{\psi_2}. \tag{5.9}$$

This shows that the random variable XY is sub-exponential according to Definition 5.5. □

Lemma 5.7. *(Gaussian concentration inequality) See [25]. Let $(G_x)_{x\in T}$ be a centered gaussian process indexed by a finite set T. Then $\forall t > 0$:*

$$\mathbb{P}(\sup_{x\in T} G_x \geq \mathbb{E}\sup_{x\in T} G_x + t)) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

where $\sigma^2 = \sup_{x\in T}\mathbb{E}G_x^2 < \infty$.

Lemma 5.8. *(Bernstein-type inequality for random variables) [25]. Let $X_1, X_2, \ldots, X_n$ be independent sub-exponential random variables with zero-mean. Also, assume that $K = \max_i \|X_i\|_{\psi_1}$. Then, for any vector $a \in \mathbb{R}^n$ and every $t \geq 0$, we have:*

$$\mathbb{P}(|\Sigma_i a_i X_i| \geq t) \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right).$$

where $c > 0$ is an absolute constant.

Proof of Theorem 4.5. We follow the proof given in [18]. Let $\beta = \frac{s}{2\sqrt{m}}$ for $0 < s < \sqrt{m}$ where m denotes the number of measurements. In (5.3), we saw that

$$\|\widehat{x} - \mu\bar{x}\| \leq 2(t + \frac{2}{t}(S_1 + S_2 + S_3)) \tag{5.10}$$

We attempt to bound the tail probability separately for each term S_1, S_2, and S_3 and then use a union bound to obtain the desired result.

For S_1, we have:

$$S_1 \leq |\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|\|\Phi^*\bar{x}\|_{K_t^o}.$$

We note that y_i is a sub-gaussian random variable (by assumption) and $\langle a_i, \bar{x}\rangle$ is a standard normal random variable. Hence, by Lemma 5.6, $y_i\langle a_i, \bar{x}\rangle$ is a sub-exponential random variable. Also, $y_i\langle a_i, \bar{x}\rangle$ for $i = 1, 2, \ldots, m$ are independent sub-exponential random variables that can be centered by subtracting of their mean, μ. Now, we can apply Lemma 5.8 on $|\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu)|$. Therefore, We have:

$$\mathbb{P}(|\frac{1}{m}\Sigma_i(y_i\langle a_i, \bar{x}\rangle - \mu))| \geq \eta\beta) \leq 2\exp\left(-\frac{c\beta^2\eta^2 m}{\|y_1\|_{\psi_2}^2}\right).$$

Here, η and μ are as defined in 4.1. Using the bound on $|\Phi^*\bar{x}\|_{k_t^o}$, we have:

$$S_1 \leq \sqrt{2}\eta\beta t\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}, \tag{5.11}$$

with probability at least $1 - 2\exp(-\frac{c\beta^2\eta^2 m}{\|y_1\|_{\psi_2}^2})$ where $c > 0$ is some constant.

For S_2 we have:

$$S_2 \leq \sqrt{2}\mu\alpha t\frac{\varepsilon}{\sqrt{1-\varepsilon}}, \tag{5.12}$$

with probability 1 since S_2 is a deterministic quantity.

For S_3 we have:

$$S_3 \leq \|\Phi^*\frac{1}{m}\Sigma_i y_i b_i\|_{K_t^o}.$$

To obtain a tail bound for S_3, We are using the following:

$$S_3 \leq \frac{1}{m}(\Sigma_i y_i^2)^{1/2}\|\Phi^* g\|_{K_t^o}$$

We need to invoke the Bernstein Inequality (lemaa 5.8) for sub-exponential random variables $(y_i^2 - \eta^2)$ for $i = 1, 2, \ldots, m$ in order to bound $\frac{1}{m}(\Sigma_i y_i^2)^{1/2}$. we have:

$$|\frac{1}{m}|\Sigma_i(y_i^2 - \eta^2)| \leq 3\eta^2$$

with high probability $1 - 2\exp(-\frac{cm\eta^4}{\|y_1\|_{\psi_2}^4})$.

Also, we need to bound $\|\Phi^* g\|$ where $g \sim \mathcal{N}(0, I)$ with high probability. Since Φ is an orthogonal matrix; as a result, $\Phi^* g \sim \mathcal{N}(0, I)$. Hence, we can use the Gaussian concentration inequality to bound $\Phi^* g$ as mentioned in lemma 5.7. Put these pieces together, we have:

$$S_3 \leq \frac{2\eta}{\sqrt{m}}(W_t(K) + t\beta\sqrt{m}), \tag{5.13}$$

with probability at least $1 - 2\exp(-\frac{cm\eta^4}{\|y_1\|_{\psi_2}^4}) - \exp(c\beta^2 m)$. Here, $W_t(K)$ denotes the local mean width for the set $\mathcal{K}_1$ defining in section 3.

Now, using 5.10 to 5.13 together with union bound, we obtain:

$$\|\widehat{x} - \mu\bar{x}\| \leq \frac{2\sqrt{2}\eta s}{\sqrt{m}}\left(\frac{1+\varepsilon}{\sqrt{1-\varepsilon}}\right) + 4\sqrt{2}\mu\left(\frac{\varepsilon}{\sqrt{1-\varepsilon}}\right) + \frac{C\eta}{\sqrt{m}}\sqrt{s\log\frac{2n}{s}} + 4\frac{\eta s}{\sqrt{m}}.$$

with probability at least $1 - 4\exp(-\frac{cs^2\eta^4}{\|y_1\|_{\psi_2}^4})$ where $C, c > 0$ are absolute constants. Here, we have again used the well-known bound on the local mean width of the set of sparse vectors; see lemma (2.3) in [15]. This completes the proof. □

6 Numerical Results

In this section, we provide some representative numerical experiments for our proposed algorithm based on synthetic and real data. We also compare its performance with a LASSO-type technique for demixing. This method, first proposed in [18], was not explicitly developed in the demixing context, but is suitable for our problem. We call this method the *Nonlinear convex demixing LASSO*, or the NLCDLASSO for short. Using our notation from Section 3 and 4, NLCDLASSO solves the following convex problem:

$$\begin{aligned} \min_{z,w} \quad & \|\widehat{x}_{lin} - (\Phi z + \Psi w)\| \\ \text{subject to} \quad & \|w\|_1 \leq \sqrt{s}, \quad \|z\|_1 \leq \sqrt{s}. \end{aligned} \tag{6.1}$$

Here, $\widehat{x}_{lin}$ denotes the linear estimation of x (definition 3.1) and s denotes to the sparsity level of signals w and z in basis Φ and Ψ, respectively. The constraints in problem 6.1 are convex penalties reflecting the knowledge that w and z are s-sparse and have unit ℓ_2-norm. The outputs of this algorithm are the estimates $\widehat{w}$, $\widehat{x}$, and $\widehat{x} = \Phi\widehat{w} + \Psi\widehat{z}$. To solve the optimization problem in 6.1, we have used SPGL1 [26,27]. This solver can handle large scale problems, which is the scenario that we have used in our experimental evaluations.

6.1 Synthetic data

We precisely describe the setup of our simulations for synthetic data. First, we generate w and z belonging to $\mathbb{R}^n$ with $n = 2^{20}$, whose support is randomly generated among all supports in $\mathbb{R}^n$ with s nonzero elements. According to the discussion in the Introduction, for successful recover we require that the constituent signals to be incoherent enough. To achieve this, we consider that the signal w to be s-sparse in the Haar wavelet basis, and z to be s-sparse in the noiselet basis [28]. For the measurement operator A, we choose a partial DFT matrix. Such matrices are known to have similar recovery performance as random Gaussian matrices, but enable fast numerical operations [29].

For our nonlinear link function, we set $f(x) = \text{sign(x)}$. Due to this measurement model, the scale (amplitude) of the underlying signal is irrevocably lost. To measure recovery performance in the absence of scale information, we use the criterion of *Cosine Similarity* between x and $\widehat{x}$ to compare the performance of different methods. More precisely, suppose

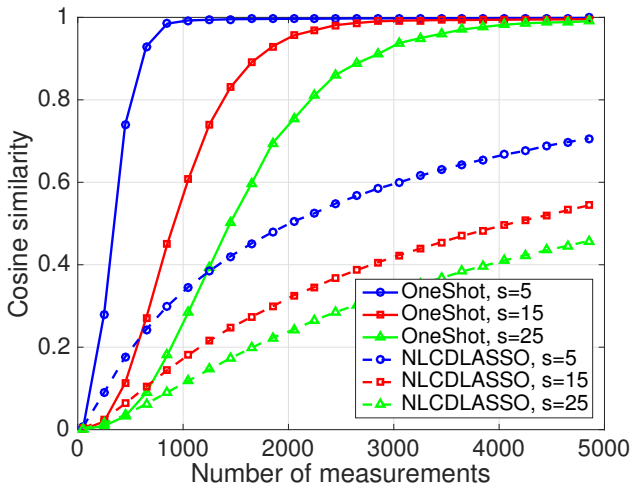


Figure 1: Performance of ONESHOT and NLCDLASSO according to the COSINE SIMILARITY for different choices of sparsity level s and for different number of measurements m.

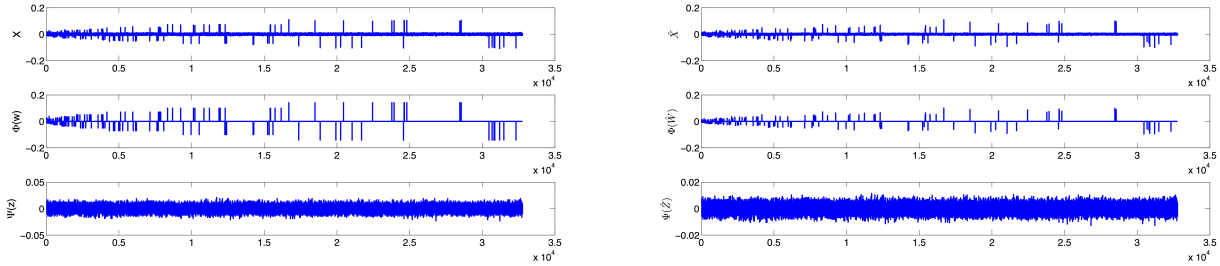
ONESHOT (or NLCDLASSO) outputs $\widehat{w}$ and $\widehat{z}$, such that $\widehat{x} = \Phi\widehat{w} + \Psi\widehat{z}$. Then, the Cosine Similarity criterion is defined as follows:

$$\cos(x, \widehat{x}) = \frac{x^T \hat{x}}{\|x\|\|\widehat{x}\|}.$$

Figure 1 illustrates the performance of ONESHOT and NLCDLASSO according to the Cosine Similarity for different choices of sparsity level s. The horizontal axis denotes an increasing number of measurements. Each data point in the plot is obtained by conducting a Monte Carlo experiment in which a new random measurement matrix A is generated, and averaging over 100 trials. As we can see from the plot, the performance of NLCDLASSO is worse than ONESHOT for any fixed choice of m and s. Even for extreme situations ($m = 4850$ and $s = 5$), ONESHOT outperforms NLCDLASSO significantly.

Figure 2 shows the true signal and its components (x, Φw and Φz) as well as the reconstructed signals using ONESHOT. In this experiment, we have used a standard normal random matrix as the measurement matrix A, and a test signal of length 2^{15}. The sparsity parameter s is set to 15, and the number of measurements is set to 5000. From visual inspection, we can observe that both true and reconstructed signals as well as true constituent signals and their estimations are close to each other (up to a scaling factor).

Finally, we contrast the running time of both algorithms, illustrated in Figure 3. In this experiment, we have measured the wall-clock running time of the two algorithms, by varying signal size x from $n = 2^{10}$ to $n = 2^{20}$. Here, we set the number of measurements to $m = 500$ and the number of Monte Carlo trials to 1000. The horizontal axis in the plot is in log scale. As we can see, ONESHOT is 12 times faster than NLCDLASSO when the size of signal equals to 2^{20} which makes ONESHOT very efficient even for large scale nonlinear demixing problems.



(a) (b)

Figure 2: (a) Ground truth x, Φw, and Φz. (b) ONESHOT estimates $\hat{x}$, $\Phi\hat{w}$, and $\Phi\hat{z}$.

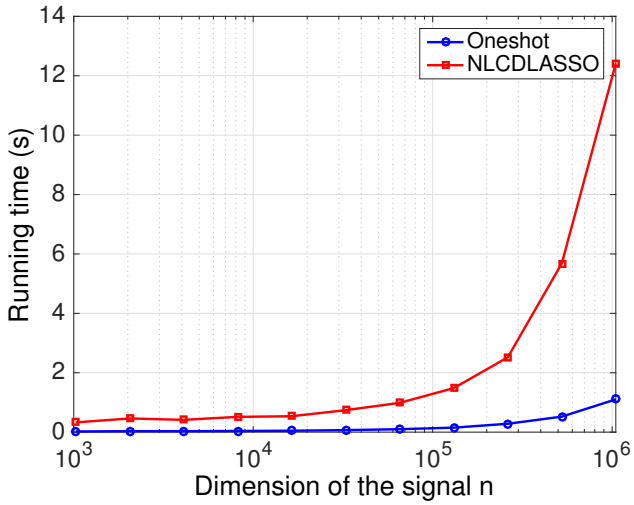


Figure 3: Comparison of running times of ONESHOT with NLCDLASSO.

6.2 Real data

In this section, we provide representative results on real-world 2-dimensional data using ONESHOT and NLCDLASSO.

For the 2-dimensional scenario, we start with a 256×256 test image. First, we obtain its 2D Haar wavelet decomposition and retain the $s = 500$ largest coefficients, denoted by the s-sparse vector w. Then, we reconstruct the image based on these largest coefficients, denoted by $\widehat{x} = \Phi w$. Similar to the synthetic case, we generate a noise component in our superposition model based on 500 noiselet coefficients z. In addition, we consider a parameter which controls the strength of the noiselet component contributing to the superposition model. We set this parameter to 0.1. Therefore, our test image x is given by $x = \Phi w + 0.1\Psi z$.

Figure 4 illustrates both the true and the reconstructed images x and $\widehat{x}$ as well as both the true and the reconstructed wavelet-sparse component, denoting by Φw and $\Psi\widehat{w}$, respectively. The number of measurements is set to 35000. In this simulation, for measurement matrix A, we have used a partial (subsampled) DFT matrix due to the difficulties in large-scale computations with Gaussian measurement matrices. However, the main theoretical

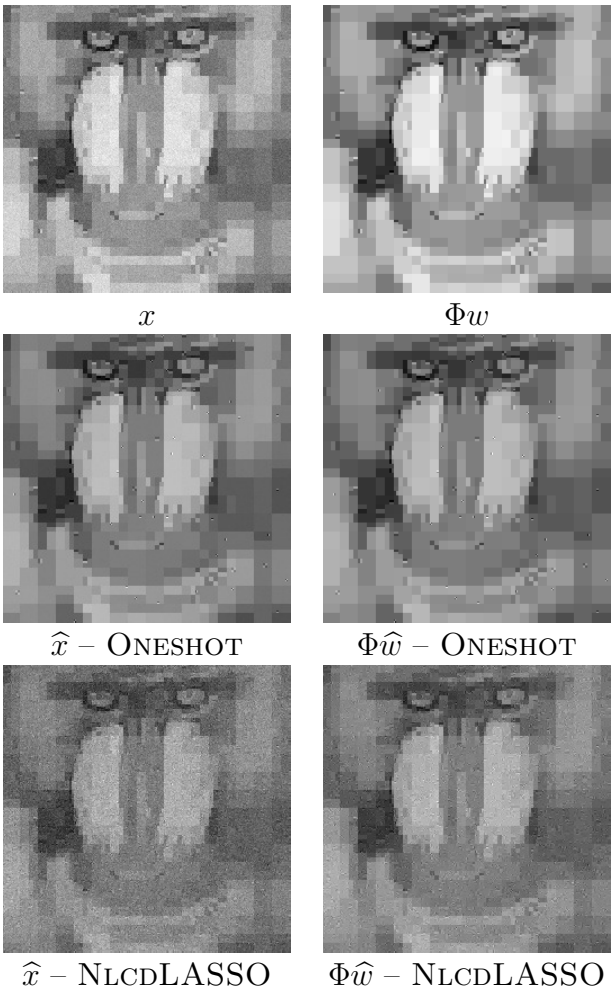


Figure 4: Comparison of ONESHOT and NLCDLASSO for demixing problem in real 2-dimensional case.

results obtained in this paper are applicable to random Gaussian measurement matrices. Hence, our theoretical results do not directly apply in this, and we defer a thorough investigation of random DFT measurement matrices for the nonlinear demixing problem as future research. Nevertheless, the results are promising, and from visual inspection we see that the reconstructed image, $\widehat{x}$, using ONESHOT is better than the reconstructed image by NLCDLASSO. Quantitatively, we also calculate Peak signal-to-noise-ratio (PSNR) of the reconstructed images using both algorithms relative to the test image, x. We have:

$$PSNR_{\text{ONESHOT}} = 19.8335\,dB \qquad PSNR_{\text{NLCDLASSO}} = 17.9092\,dB$$

therefore illustrating the superior performance of ONESHOT compared to NLCDLASSO.

7 Conclusions

We proposed a simple (and fast) algorithm for demixing sparse signals from nonlinear observations. We call this algorithm ONESHOT since it is non-iterative. The performance of ONESHOT is characterized via upper bounds on the statistical error that depend on the (in)coherence between the sparse signal constituents. An important implication of our bound is that under the assumption of sufficient incoherence, the number of measurements required for reliable recovery of the constituent signals is given by $m = \mathcal{O}(s \log \frac{2n}{s})$ where s is the sparsity level of the constituent signals and n is the ambient dimension. We anticipate that the problem of demixing signals from nonlinear observations can be used in numerous different practical applications. As future work, we intend to extend our methods to more general signal models (including rank-sparsity models), as well as robust algorithms for more general measurement models.

References

[1] M. Elad, J-L. Starck, P. Querre, and D. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Appl. Comput. Harmonic Analysis*, 19(3):340–358, 2005.

[2] J. Bobin, J.-L. Starck, J. Fadili, Y. Moudden, and D. Donoho. Morphological component analysis: An adaptive thresholding strategy. *IEEE Trans. Image Proc.*, 16(11):2675–2681, 2007.

[3] M. B. McCoy and J. A. Tropp. Sharp recovery bounds for convex demixing, with applications. *Foundations of Computational Mathematics*, 14(3):503–567, 2014.

[4] H. Guo, C. Qiu, and N. Vaswani. An online algorithm for separating sparse and low-dimensional signal sequences from their sum. *IEEE Trans. Sig. Proc.*, 62(16):4284–4297, 2014.

[5] C. Qiu, N. Vaswani, B. Lois, and L. Hogben. Recursive robust pca or recursive sparse recovery in large but structured noise. *IEEE Trans. Inform. Theory*, 60(8):5007–5039, 2014.

[6] E. Candès. Compressive sampling. In *Proc. Int. Congress of Math.*, Madrid, Spain, Aug. 2006.

[7] D. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[8] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Springer.

[9] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[10] V. Chandrasekaran, S. Sanghavi, P. A Parrilo, and A. S. Willsky. Sparse and low-rank matrix decompositions. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 962–967. IEEE, 2009.

[11] V. Chandrasekaran, S. Sanghavi, P. A Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[12] C. Hegde and R. Baraniuk. SPIN : Iterative signal recovery on incoherent manifolds. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, July 2012.

[13] C. Hegde and R. Baraniuk. Signal recovery on incoherent manifolds. *IEEE Trans. Inform. Theory*, 58(12):7204–7214, Dec. 2012.

[14] P. Boufounos and R. Baraniuk. 1-bit compressive sensing. In *Int. Conf. Info. Sciences and Systems (CISS)*, pages 16–21. IEEE, 2008.

[15] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on*, 59(1):482–494, 2013.

[16] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

[17] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[18] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[19] C. Studer, P. Kuppinger, G. Pope, and H. Bölcskei. Recovery of sparsely corrupted signals. *Information Theory, IEEE Transactions on*, 58(5):3115–3130, 2012.

[20] V. Chandrasekaran, P. A Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1610–1613. IEEE, 2010.

[21] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2233–2246, 2012.

[22] A. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

[23] S. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

[24] X. Yi, Z. Wang, C. Caramanis, and H. Liu. Optimal linear estimation under unknown nonlinear transform. In *Advances in Neural Information Processing Systems*, pages 1549–1557, 2015.

[25] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[26] E. van den Berg and M. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comp.*, 31(2):890–912, 2008.

[27] E. van den Berg and M. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

[28] R. Coifman, F. Geshwind, and Y. Meyer. Noiselets. *Appl. Comput. Harmonic Analysis*, 10(1):27–44, 2001.

[29] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.