A Fast Iterative Algorithm for Demixing Sparse Signals from Nonlinear Observations

Mohammadreza Soltani, Chinmay Hegde

Department of Electrical and Computer Engineering
Iowa State University

GlobalSIP 2016, Dec 2016

Demixing problem — Motivation

- **What is demixing and why do we care?**

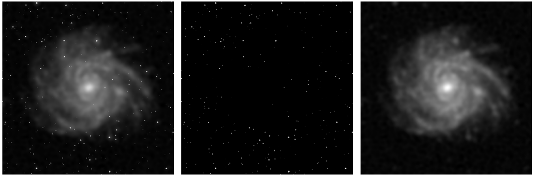
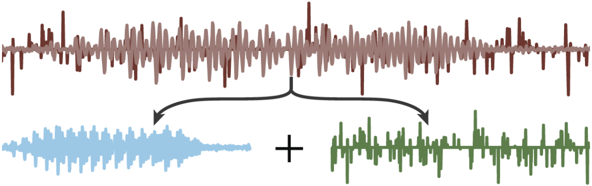


Image credits: NASA

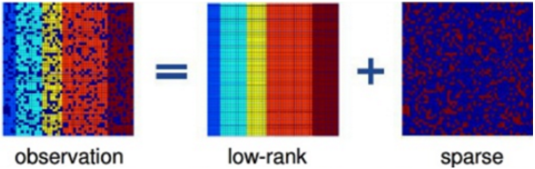
- Another example



Demixing problem — Examples

- *Demixing* problem is special of interest of the numerous applications ranging from
  - signal processing, astronomy, computer vision, and machine learning
- Examples
  1. Morphological Component Analysis (MCA)
  2. Separation of foreground and background in video

  3. Robust PCA

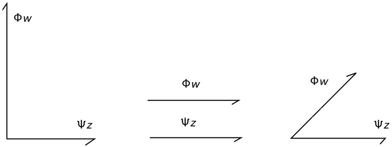


Demixing problem — Definition

- In simple form, demixing involves disentangling two (or more) constituent signals from observations of their linear **superposition**:

$$x = \Phi w + \Psi z$$

  - Φ and Ψ are *incoherent* orthonormal bases of $\mathbb{R}^n$,
  - $w, z \in \mathbb{R}^n$ are the corresponding basis coefficients



- **Goal: reliably recover the constituent signals (equivalently, their basis representations w and z) from the superposition signal x.**

Four problems

$y \in \mathbb{R}^{m \times 1}$, $A \in \mathbb{R}^{m \times n}$, $m \ll n$

- Compressive sensing (Linear inverse problem)

$$y = Ax$$

  where $x \in \mathbb{R}^n$ s.t. $\|x\|_0 \leq s$.

- Linear demixing problem

$$y = A(\Phi w + \Psi z)$$

  where Φ and Ψ are *incoherent* bases in $\mathbb{R}^n$, and $w, z \in \mathbb{R}^n$ are the basis coefficients s.t. $\|w\|_0 \leq s$ and $\|z\|_0 \leq s$.

- Nonlinear signal recovery

$$y = g(Ax)$$

  where g denotes an element wise nonlinear *link* function.

- **Nonlinear demixing problem – Our focus in this talk**

$$y = g(A(\Phi w + \Psi z))$$

Challenges in (nonlinear) demixing Problem

1. Fundamental identifiability issue (Linear demixing)
   - number of unknowns ($2n$) is greater than the number of observations (n).
   - **remedy:** some type of *incoherence* between the constituent signals (or more specifically, between the corresponding bases Φ and Ψ).
2. Limited number of measurements
   - $y = Ax$, x is superposition signal and $A \in \mathbb{R}^{m \times n}$ denotes the measurement operator with $m \ll n$.
   - **remedy:** structural assumptions on the constituent signals.
3. Nonlinear observation model
   - $y = g(Ax) + e$, x is superposition signal and $e \in \mathbb{R}^m$ denotes the additive noise.
   - **remedy:** The subject of this talk.

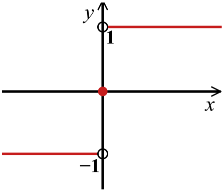
Nonlinear Signal Recovery

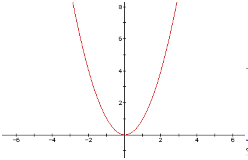
- Formulation with additive noise

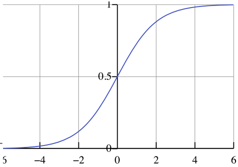
$$y = g(Ax) + e$$

  where g denotes an element wise nonlinear *link* function and $e \in \mathbb{R}^m$ represents the noise.

- Nonlinear Signal recovery
  1. 1-bit compressive sensing $\rightarrow$ modeling high speed ADC's
  2. phase retrieval $\rightarrow$ modern astronomical imaging systems
  3. nonlinear matrix completion $\rightarrow$ recommender systems







Nonlinear Demixing Problem — Formulation

- **General formulation**

$$y = g(A(\Phi w + \Psi z)) + e \tag{1}$$

  - Φ and Ψ are *incoherent* bases in $\mathbb{R}^n$
  - $w, z \in \mathbb{R}^n$ are the basis coefficients such that. $\|w\|_0 \leq s$ and $\|z\|_0 \leq s$
  - $e \in \mathbb{R}^m$ denotes the additive noise.
- **The goal is to recover w and z (constituent signals)**

Definition (ε-incoherence)

The orthonormal bases Φ and Ψ are said to be ε-incoherent if:

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

Nonlinear Demixing Problem — First Scenario

- We consider two scenarios:
  1. **In the first scenario,** the model is given by

$$y = g(A(\Phi w + \Psi z))$$

  - g is **unknown** and odd function. It can be non-smooth and non-invertible
  - A with **i.i.d standard normal entries**
  - no additive noise
  - We introduced an algorithm called OneShot [1].
  - **Sample complexity result**
    Suppose we fix $\kappa > 0$ as a small constant, and suppose that the incoherence parameter $\varepsilon = c\kappa$ for some constant c, and that the number of measurements scales as:

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

[1] M. Soltani and C. Hegde, Demixing, Sparse Signals from Nonlinear Observations, Asilomar Conference on Signals, Systems, and Computers, November 2016.

Algorithms — Beyond OneShot

- Disadvantages of OneShot:
  1. sparse components are recovered only up to an arbitrary scale factor
  2. leading to high estimation errors in practice
  3. its sample complexity is inversely dependent on the estimation error
- To solve these problems, we propose a different, iterative algorithm for recovering the signal components
  - **Demixing with Hard Thresholding (DHT)**

Algorithms — Second Scenario

- **In the second scenario,** the model is given by

$$y = g(A(\Phi w + \Psi z)) + e$$

  - g is **known** and its derivative is strictly bounded either within a positive, or within a negative interval
  - A with **independent isotropic rows**
  - A with **independent subgaussian isotropic rows**
  - additive noise is assumed
- By defining $\Gamma = [\Phi \ \Psi]$ and $t = [w; z]$, and $\Theta(x) = \int_{\infty}^{x} g(u)du$, DHT tries to solve the following optimization problem ($F(t) : \mathbb{R}^{2n} \to \mathbb{R}$):

$$\min_{t \in \mathbb{R}^{2n}} \quad F(t) = \frac{1}{m} \sum_{i=1}^{m} \Theta(a_i^T \Gamma t) - y_i a_i^T \Gamma t$$
$$\text{s. t.} \quad \|t\|_0 \leq 2s.$$

Algorithms — Second Scenario

Algorithm 1 Demixing with Hard Thresholding DHT

Inputs: Bases Φ and Ψ, measurement matrix A, link function g, measurements y, sparsity level s, step size η'.
Outputs: Estimates $\widehat{x} = \Phi\widehat{w} + \Psi\widehat{z}$, $\widehat{w}$, $\widehat{z}$
Initialization:
$\left(x^0, w^0, z^0\right) \leftarrow$ ARBITRARY INITIALIZATION
$k \leftarrow 0$
while $k \leq N$ **do**
  $t^k \leftarrow [w^k; z^k]$ {Forming constituent vector}
  $t_1^k \leftarrow \frac{1}{m}\Phi^T A^T(g(Ax^k) - y)$
  $t_2^k \leftarrow \frac{1}{m}\Psi^T A^T(g(Ax^k) - y)$
  $\nabla F^k \leftarrow [t_1^k; t_2^k]$ {Forming gradient}
  $\tilde{t}^k = t^k - \eta' \nabla F^k$ {Gradient update}
  $[w^k; z^k] \leftarrow \mathcal{P}_{2s}\left(\tilde{t}^k\right)$ {Projection}
  $x^k \leftarrow \Phi w^k + \Psi z^k$ {Estimating $\widehat{x}$}
  $k \leftarrow k + 1$
end while
Return: $(\widehat{w}, \widehat{z}) \leftarrow \left(w^N, z^N\right)$

Some definitions — Second Scenario

Definition (Subgaussian random variable)

A random variable X is called subgaussian if it satisfies the following:

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

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 (Isotropic random vectors)

A random vector-valued variable $v \in \mathbb{R}^n$ is said to be isotropic if $\mathbb{E}vv^T = I_{n \times n}$.

Some definitions — Second Scenario

Definition (Cross-coherence)

The cross-coherence parameter between the measurement matrix A and the dictionary $\Gamma = [\Phi \ \Psi]$ is defined as follows:

$$\vartheta = \max_{i,j} \frac{a_i^T \Gamma_j}{\|a_i\|_2},$$

where a_i and Γ_j denote the i^{th} row of A and the j^{th} column of Γ.

Definition (*Restricted Strong Convexity/Smoothness*)

A loss function f satisfies (RSC/RSS) if:

$$m_{4s} \leq \|\nabla_\xi^2 f(t)\| \leq M_{4s}, \quad t \in \mathbb{R}^{2n},$$

where $\xi = \mathrm{supp}(t_1) \cup \mathrm{supp}(t_2)$, for all $\|t_i\|_0 \leq 2s$ and $i = 1, 2$. Also, m_{4s} and M_{4s} are (respectively) called the RSC and RSS constants.

Algorithms — Second Scenario

Theorem (Performance of DHT)

Consider the model $y = g(A(\Phi w + \Psi z)) + e$*. Suppose that the corresponding objective function* F *satisfies the RSS/RSC properties with constants* M_{6s} *and* m_{6s} *on the set* J *with* $\|J\|_0 \leq 6s$ *such that* $1 \leq \frac{M_{6s}}{m_{6s}} \leq \frac{2}{\sqrt{3}}$ *. Choose a step size parameter* η' *with* $\frac{0.5}{M_{6s}} < \eta' < \frac{1.5}{m_{6s}}$*. Then,* DHT *outputs a sequence of estimates* (w^k, z^k) *such that the estimation error of the true constituent vector,* $t^* = [w^*; z^*]$ *satisfies the following upper bound (in expectation) for any* $k \geq 1$*:*

$$\|t^{k+1} - t^*\|_2 \leq (2q)^k \|t^0 - t^*\|_2 + C\tau\sqrt{\frac{s}{m}}, \tag{2}$$

where $q = 2\sqrt{1 + {\eta'}^2 M_{6s}^2 - 2\eta' m_{6s}}$ *and* $C > 0$ *is a constant that depends on the step size* η' *and the convergence rate* q*.*

Algorithms — Second Scenario

Theorem (Sample complexity when the rows of A are isotropic)

Suppose that the rows of A are independent isotropic random vectors. In order to achieve the requisite RSS/RSC properties of Theorem of DHT*, the number of samples needs to scale as:* $m = \mathcal{O}(s \log n \log^2 s \log(s \log n))$*, provided that the bases* Φ *and* Ψ *are incoherent enough.*

Theorem (Sample complexity when the elements of A are subgaussian)

Assume that all assumptions and definitions in Theorem of DHT *holds except that the rows of matrix A are independent subgaussian isotropic random vectors. Then, in order to achieve the requisite RSS/RSC properties of Theorem* DHT*, the number of samples needs to scale as:* $m = \mathcal{O}\left(s \log \frac{n}{s}\right)$ *, provided that the bases* Φ *and* Ψ *are incoherent enough.*

Algorithms — Second Scenario

Proof sketch

- Assuming the defined objective function satisfies RSC/RSS.
- Establishing linear convergence of DHT in expectation using *Khintchine inequality*

$$\|t^{k+1} - t^*\|_2 \leq (2q)^k \|t^0 - t^*\|_2 + C\tau\sqrt{\frac{s}{m}}$$

- Verifying the objective function satisfies RSC/RSS in two cases:
  1. A with **independent isotropic rows**
     - using *Uniform Rudelson's inequality* and *Uniform symmetrization*
  2. A with **independent subgaussian isotropic rows**
     - using *D-RIP* argument

Please see the following for more details:
"M. Soltani and C. Hegde, Fast Algorithms for Demixing Sparse Signals from Nonlinear Observations, arXiv:1608.01234."

Sample complexity

Table: *Summary of our contributions, and comparison with existing methods for the concrete case where Φ is the identity and Ψ is the DCT basis. Here, s denotes the sparsity level of the components, n denotes the ambient dimension, m denotes the number of samples, and κ denotes estimation error.*

Algorithms	Sample complexity	Running time	Measurements	Link function
LASSO[1]	$\mathcal{O}(\frac{s}{\kappa^2} \log \frac{n}{s})$	poly(n)	Gaussian	unknown
OneShot	$\mathcal{O}(\frac{s}{\kappa^2} \log \frac{n}{s})$	$\mathcal{O}(mn)$	Gaussian	unknown
DHT	$\mathcal{O}(s \text{ polylog } n)$	$\mathcal{O}(mn \log \frac{1}{\kappa})$	Isotropic rows	known
DHT	$\mathcal{O}(s \log \frac{n}{s})$	$\mathcal{O}(mn \log \frac{1}{\kappa})$	Subgaussian	known

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

Experimental Results

We compare OneShot and DHT with two other algorithms:

1. *Nonlinear convex demixing with LASSO* or NlcdLASSO. This algorithm solves the following optimization problem:

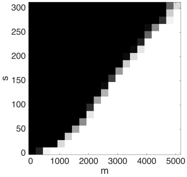
$$\begin{aligned} \min_{z,w} \quad & \|\widehat{x}_{\text{lin}} - (\Phi z + \Psi w)\|_2 \\ \text{subject to} \quad & \|w\|_1 \leq \sqrt{s}, \quad \|z\|_1 \leq \sqrt{s}. \end{aligned} \tag{3}$$

2. *Demixing with Soft Thresholding* or DST. This algorithm solves the following optimization problem:

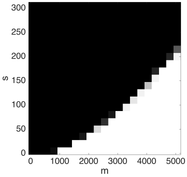
$$\min_{t} \ \frac{1}{m} \sum_{i=1}^{m} \Theta(a_i^T \Gamma t) - y_i a_i^T \Gamma t + \beta \|t\|_1, \tag{4}$$

Experimental Results — Synthetic data

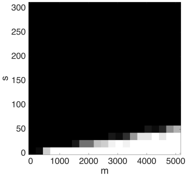
- Second Scenario — Link function, g is known



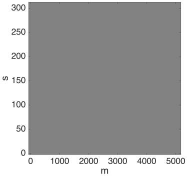
(a) DHT



(b) DST



(c) OneShot

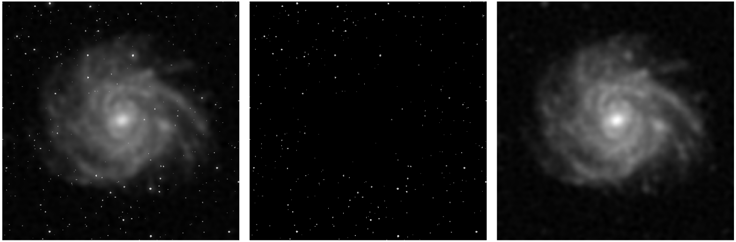


(d) NlcdLASSO

- Phase transition plots with cosine similarity as the criterion. Link function is defined as $g(x) = 2x + sin(x)$.

Experimental Results — Real data

- Second Scenario — Link function, g is known



(a) Original x (b) $\Phi(\widehat{w})$ $\Psi(\widehat{z})$

Image credits: NASA and Convexity in Source Separation

- Parameters:
  $n = 512 \times 512, s = 1000, m = 15000, g(x) = \frac{1}{2}\frac{1-e^{-x}}{1+e^{-x}}$.

Conclusion

- Considering the problem of demixing sparse signals from their nonlinear measurements
- Specifically, studying the more challenging scenario with a limited number of nonlinear measurements
- As our contribution:
    1. proposing a fast algorithm for recovery of the constituent signals
    2. supporting the proposed algorithm with the rigorous theoretical analysis
    3. deriving nearly-tight upper bounds on their sample complexity
    4. verifying experimentally the superiority of the proposed algorithms compared to existing convex demixing methods both on synthetic and real data