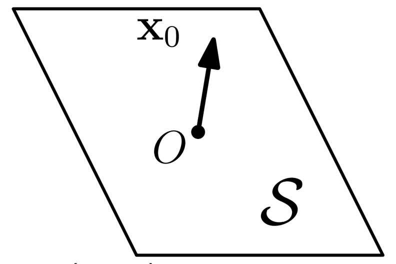
Finding a Sparse Vector in a Subspace: Linear Sparsity Using Alternating Directions

Qing Qu, Ju Sun, and John Wright, Department of Electrical Engineering, Columbia University

Finding a Sparse Vector in a Linear Subspace?

Problem Statement:

- Given a sparse vector $\mathbf{x}_0$ embedded in an n dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^p$, provided any basis of $\mathcal{S}$, can we efficiently recover $\mathbf{x}_0$?



- Equivalently, provided a matrix $\mathbf{A} \in \mathbb{R}^{(p-n)\times p}$ whose row span forms the subspace $\mathcal{S}$, can we solve

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{0}, \ \mathbf{x} \neq \mathbf{0} \quad ? \qquad (1)$$

Motivation:

- In contrast to the standard sparse recovery problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b},$$

convex relaxation works nearly optimally for generic design of $\mathbf{A}$, the computational property of (1) is not nearly as well understood.
- Variants of (1) has been studied in numerical linear algebra, sparse PCA, blind source separation, dictionary learning (DL), spectral estimation and Pony's Problem, and graphical model learning.

Existing Work

- ℓ^1/ℓ^∞ **Recovery** [Spielman et al.] and [Hand et al.]:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \quad \text{s.t.} \quad x_i = 1, \ \mathbf{x} \in \mathcal{S}, \ 1 \le i \le p.$$

- **Semi-Definite Programming (SDP) Relaxation:**

$$\min_{\mathbf{X}} \|\mathbf{X}\|_1, \ \text{s.t.} \ \langle \mathbf{A}^\top\mathbf{A}, \mathbf{X}\rangle = 0, \ \text{tr}[\mathbf{X}] = 1, \ \mathbf{X} \succeq \mathbf{0}.$$

- **Sum-of-Squares (SOS) Relaxation** [Barak et al.]:

Method	Recovery Condition	Computation Complexity
ℓ^1/ℓ^∞	$\theta \in O(1/\sqrt{n})$	$\Omega(p^2)$
SDP	$\theta \in O(1/\sqrt{n})$	$O(p^3)$
SOS	$p \ge \Omega(n^2), \theta \in O(1)$	high order $poly(p)$

Question 1: Is there a practical algorithm that provably recovers a sparse vector with $\theta \gg 1/\sqrt{n}$ from a generic subspace $\mathcal{S}$?

Contributions of this Work

- Proposed a simple ADM algorithm, **addressed the problem** under the PSV model, exact recovery for $\mathbf{x}_0$ to have θp nonzeros, provided $p \ge \Omega(n^4 \log n)$.
- Performs well empirically – succeeds for both the PSV and DL models, with $p \ge \Omega(n \log n)$.

Problem Formulation and Optimality Conditions

- **Planted Sparse Vector (PSV) Model:** A single sparse vector $\mathbf{x}_0$ embedded in an otherwise random subspace:

$$\mathcal{S} = \text{span}\,(\mathbf{x}_0, \mathbf{g}_1, \dots, \mathbf{g}_{n-1}) \subset \mathbb{R}^p,$$

where $\mathbf{g}_k \sim_{\text{i.i.d.}} \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I})$, and $\mathbf{x}_0 \sim_{\text{i.i.d.}} \frac{1}{\sqrt{\theta p}}\text{Ber}(\theta)$.

- **Nonconvex ℓ^1/ℓ^2 Minimization Problem:**

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S}, \ \|\mathbf{x}\|_2 = 1.$$

which is equivalent to

$$\min_{\mathbf{q}} \|\mathbf{Y}\mathbf{q}\|_1, \quad \text{s.t.} \quad \|\mathbf{q}\|_2 = 1, \qquad (2)$$

where $\mathbf{Y} \in \mathbb{R}^{p \times n}$ is an arbitrary orthonormal matrix whose columns form a basis of $\mathcal{S}$.

- **Theorem (Global Optimality for ℓ^1/ℓ^2 Recovery):** Suppose $\mathcal{S}$ follows the PSV model, and $\mathbf{q}^\star$ be the optimum to (2), with very high probability, we have $\mathbf{Y}\mathbf{q}^\star = \xi\mathbf{x}_0$ for some $\xi \neq 0$, provided

$$p \ge \Omega(n \log n), \qquad \text{and} \qquad \theta \le \theta_0.$$

Question 2: Can we efficiently solve (2) to global optimality?

Algorithm based on Alternating Direction Method (ADM)

- **Alternating Minimization:** Consider a relaxation of (2):

$$\min_{\mathbf{q},\mathbf{x}} \frac{1}{2}\|\mathbf{Y}\mathbf{q} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \quad \text{s.t.} \quad \|\mathbf{q}\|_2 = 1,$$

minimize the problem by alternating direction:

$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2}\left\|\mathbf{Y}\mathbf{q}^{(k)} - \mathbf{x}\right\|_2^2 + \lambda\|\mathbf{x}\|_1, \qquad (3)$$

$$\mathbf{q}^{(k+1)} = \arg\min_{\mathbf{q}} \frac{1}{2}\left\|\mathbf{Y}\mathbf{q} - \mathbf{x}^{(k+1)}\right\|_2^2 \ \text{s.t.} \ \|\mathbf{q}\|_2 = 1. \qquad (4)$$

Closed form solutions of (3), (4) lead to one ADM iteration

$$\mathbf{q}^{(k+1)} = \frac{\mathbf{Y}^\top \mathcal{S}_\lambda\left[\mathbf{Y}\mathbf{q}^{(k)}\right]}{\left\|\mathbf{Y}^\top \mathcal{S}_\lambda\left[\mathbf{Y}\mathbf{q}^{(k)}\right]\right\|_2}, \qquad (5)$$

where $\mathcal{S}_\lambda[x] = \text{sign}(x)(|x| - \lambda)_+$.

- **Initialization Strategy:** Given $\mathbf{Z} = [\mathbf{x}_0, \mathbf{g}_1, \cdots, \mathbf{g}_{n-1}]$, $x_{0i} \neq 0$,

$$x_{0i} = \Theta\left(1/\sqrt{\theta p}\right), \qquad \mathbf{g}^i \sim \mathcal{N}(\mathbf{0}, 1/p\,\mathbf{I}).$$

Idea: Because $\mathbf{z}^i$ is biased towards the optimizer $\mathbf{q}^\star = \mathbf{e}_1$, use normalized rows of $\mathbf{Z}$ as initializations.
Remark: Analysis shows that it works for the orthogonalized version and invariant to rotations as well.

- **Rounding by Linear Programming (LP):** Let $\mathbf{r} = \bar{\mathbf{q}}$, which is the output of the ADM algorithm,

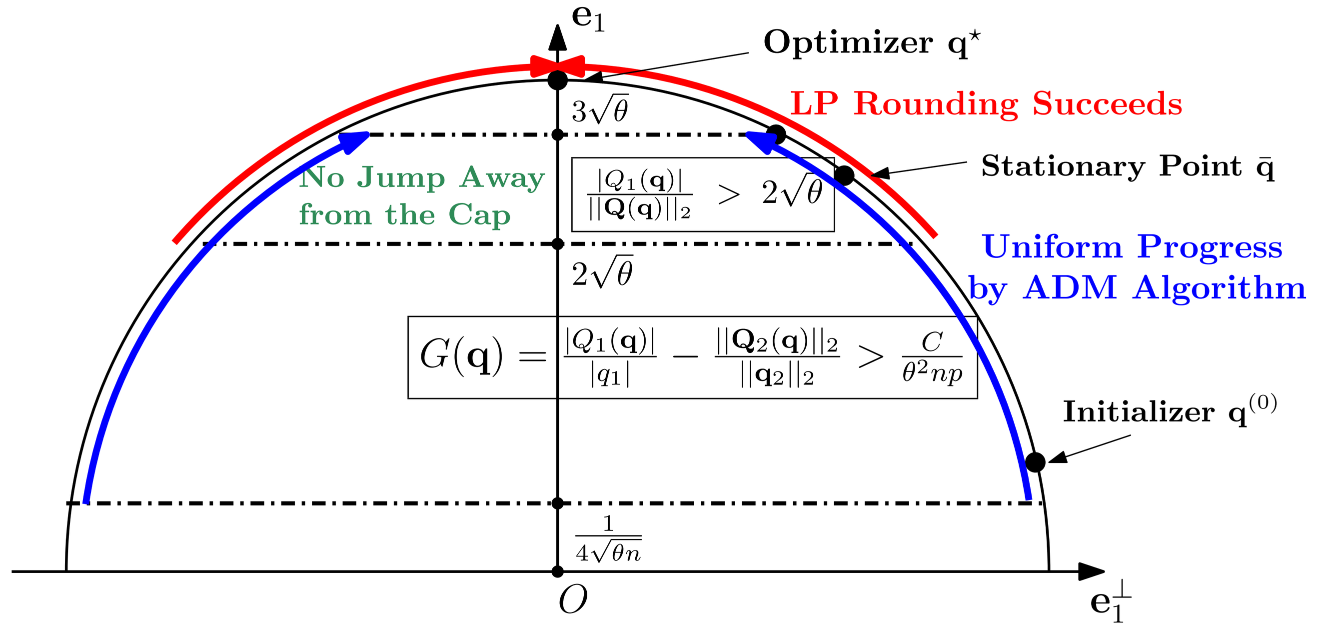
$$\min_{\mathbf{q}} \|\mathbf{Y}\mathbf{q}\|_1, \quad \text{s.t.} \quad \langle \mathbf{r}, \mathbf{q}\rangle = 1. \qquad (6)$$

Theorem (Exact Recovery for the ADM Algorithm, PSV)

- *Apply the ADM algorithm (5) with $\lambda = 1/\sqrt{p}$, using all rows of $\mathbf{Y}$ as initializations for $\mathbf{q}^{(0)}$ to produce $\bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_p$. Solve the LP rounding (6) with $\mathbf{r} = \bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_p$, to produce $\widehat{\mathbf{q}}_1, \dots, \widehat{\mathbf{q}}_p$.*
- *Set $i^\star \in \arg\min_i \left\|\mathbf{Y}\widehat{\mathbf{q}}_i\right\|_0$, with very high probability, $\mathbf{Y}\widehat{\mathbf{q}}_{i^\star} = \gamma\mathbf{x}_0$ for some $\gamma \neq 0$, provided*

$$p > \Omega(n^4 \log n), \qquad \text{and} \qquad \theta \le \theta_0.$$

A Sketch of Analysis



Under the PSV model, let $\mathbf{q} = \left[q_1, \mathbf{q}_2^\top\right]^\top$, $\mathbf{G} = [\mathbf{g}_1, \cdots, \mathbf{g}_{n-1}]$, assume the orthonormal matrix

$$\mathbf{Y} = \left[\frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \mid \mathcal{P}_{\mathbf{x}_0^\perp}\mathbf{G}\left(\mathbf{G}^\top\mathcal{P}_{\mathbf{x}_0^\perp}\mathbf{G}\right)^{-1/2}\right].$$

Define a random process over $\mathbf{q} \in \mathbb{S}^{n-1}$:

$$\mathbf{Q}(\mathbf{q}) = \frac{1}{p}\sum_{k=1}^{p} \mathbf{y}^k \mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}^k\right] = \left[Q_1(\mathbf{q}), \mathbf{Q}_2^\top(\mathbf{q})\right]^\top$$

- **Good initialization:** One of initializers $\mathbf{q}_i^{(0)} = \mathbf{y}^i$, w.h.p.,

$$\left|\left\langle \mathbf{q}_i^{(0)}, \mathbf{e}_1 \right\rangle\right| \ge 1/(4\sqrt{\theta n})$$

- **Uniform progress away from the equator:** Because

$$\left\langle \frac{\mathbf{Q}(\mathbf{q})}{\|\mathbf{Q}(\mathbf{q})\|_2}, \mathbf{e}_1 \right\rangle > \langle \mathbf{q}, \mathbf{e}_1 \rangle \ \Leftrightarrow \ \frac{|Q_1(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2} > 0,$$

we show for any $\mathbf{q} \in \mathbb{S}^{n-1}$ with $\frac{1}{4\sqrt{\theta n}} \le |q_1| \le 3\sqrt{\theta}$, w.h.p.,

$$G(\mathbf{q}) = \frac{|Q_1(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2} > \frac{C}{\theta^2 np}. \qquad (7)$$

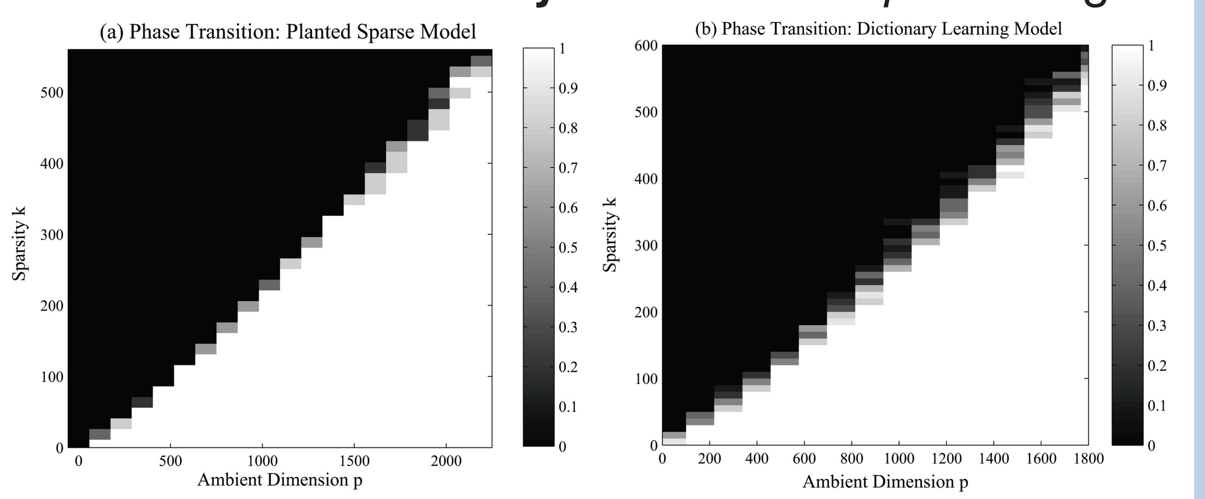
- **No jumps away from the cap:** For all $\mathbf{q}$ with $|q_1| > 3\sqrt{\theta}$,

$$|Q_1(\mathbf{q})| / \|\mathbf{Q}(\mathbf{q})\|_2 > 2\sqrt{\theta}. \qquad (8)$$

- **Location of the stationary point:** Steps above implies if the ADM algorithm starts from a point $\mathbf{q}^{(0)}$ with $\left|q_1^{(0)}\right| > \frac{1}{4\sqrt{\theta n}}$, it will converge to a stationary point $\bar{\mathbf{q}}$ such that $|\bar{q}_1| > 2\sqrt{\theta}$.
- **LP rounding succeeds:** Solving (6) with $\mathbf{r} = \bar{\mathbf{q}}$, w.h.p., will output a solution $\mathbf{q}^\star = \mathbf{e}_1$.

Experimental Results

- **Phase Transition on Synthetic Data:** $p = 5n \log n$.



- **Exploratory Experiments on Faces:**

Figure: Four sparse vectors extracted by the ADM algorithm for one person in the Yale B database under different illuminations.

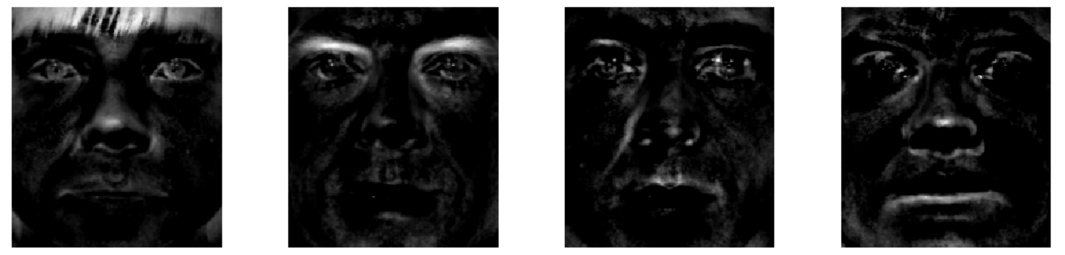
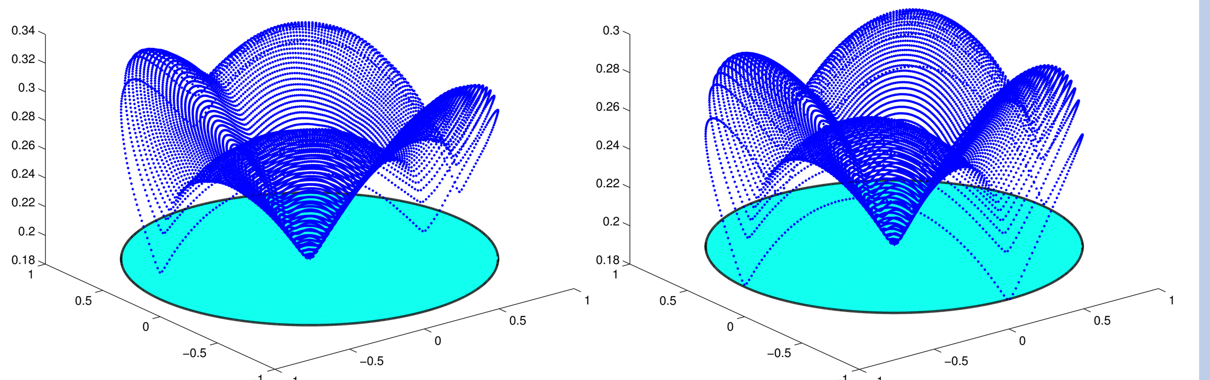


Figure: Four sparse vectors extracted by the ADM algorithm for 10 different persons in the Yale B database under normal illuminations.

Discussions

- **More Application Ideas?**
- **Intriguing Experiments on Dictionary Learning:**



Efficient algorithms can also achieve **linear sparsity** regime for the **squared dictionary learning** under the Bernoulli-Gaussian model!

- **Generalization:** Can we develop general tools for

$$\min_{\mathbf{w}} \frac{1}{p}\sum_{k=1}^{p} f_k(\mathbf{w}), \qquad \text{s.t.} \quad \mathbf{w} \in \mathcal{M} \quad ?$$

$f_k(\mathbf{w})$: nonconvex function, $\mathcal{M}$: smooth manifold.

- **Nonconvex Problems as a Whole:** Phase retrieval, matrix/tensor completion, robust PCA, blind deconvolution, etc.

Email: <qq2105, js4038, jw2966>@columbia.edu