Recovery and denoising with simultaneous structures

Maryam Fazel
University of Washington

Joint work with:
Samet Oymak (Berkeley), Amin Jalali (UW),
Babak Hassibi (Caltech), Yonina Eldar (Technion)

SPARS 2015, Cambridge University, July 7, 2015
Outline

• simultaneous structures: where and why?
• review: single structure case
  compressed sensing, low-rank recovery, . . .
• fundamental limitation of combining convex penalties, for
  – arbitrary norms
  – a variety of measurements, beyond Gaussian
• similar result for the problem of ‘denoising’
• what next?
Low-dimensional structures

classic examples:

• sparse vectors (e.g., compressed sensing) \( \ell_1 \) norm

• group-sparse vectors (group LASSO) \( \ell_{1,2} \) norm

• low-rank matrices (collaborative filtering, phase retrieval, \ldots) nuclear (trace) norm

• sparse \textit{plus} low-rank matrices, \( \mathbf{X} = \mathbf{L} + \mathbf{S} \) (PCA with outliers, graphical models with hidden variables) \( \ell_1 \) plus nuclear norm
Low-dimensional structures

multiple, simultaneous structures

• simultaneously sparse and low-rank matrices (sparse phase retrieval, sparse PCA, quadratic compressed sensing, . . . ) \( \ell_1 \) and nuclear norms

• tensors with low Tucker rank
  nuclear norms of unfolded matrices

• simultaneously sparse and piece-wise constant vectors (e.g., ‘fused lasso’)
  \( \ell_1 \) norm and total-variation norm, \( \| x \|_{TV} = \sum_{i=1}^{n-1} |x_{i+1} - x_i| \)
Sparse and low-rank matrices: an application

phase retrieval, a classic signal processing/optics problem

recover signal $x_0$ from linear *phaseless* measurements,

$$|a_i^T x_0| = b_i, \quad i = 1, \ldots, m$$

reformulate as: find $X = x_0x_0^T$ s.t. $\langle a_i a_i^T, X \rangle = b_i^2$

i.e., $X \succeq 0$, $\text{rank}(X) = 1$, $A(X) = b'$

[Candès, Eldar, Strohmer, Voroninski’11]

signal $x_0$ can also be *sparse*. then, $X$ is rank-1 and (block-)sparse.
other applications (for sparse and low-rank matrices):

- sparse PCA [d’Aspremont et al’08, . . . ]
  - find approximate eigenvectors of $X$ that are sparse, e.g., $X \approx xx^T$ with $x$ $k$-sparse

- cluster detection [Richard, Savalle, Vayatis’12]
  - ideal cluster adjacency matrix is low-rank & sparse
Recovery of structured models

unknown structured model $x_0 \in \mathbb{R}^n$

• recovery from **compressed measurements**: $A(x_0) = y$
  linear $A : \mathbb{R}^n \to \mathbb{R}^m$, $m \ll n$. can write as $Ax = y$ with $A \in \mathbb{R}^{m \times n}$

• **denoising**: $A$ is identity; $y = x_0 + z$, noise $z$ is i.i.d

• LASSO: $y = A(x_0) + z$

goal: given $A$ and $y \in \mathbb{R}^m$, find $x_0$.

• how many measurements $m$ suffice? (sample complexity)
• how does mean-squared error behave with noise level?
**Example: Sparse vectors and $\|x\|_1$**

$A : \mathbb{R}^n \rightarrow \mathbb{R}^m$, suppose $A$ is Gaussian. $x_0$ is $k$-sparse.

**non-convex program:**

\[
\begin{align*}
& \text{minimize} \quad \|x\|_0 \\
& \text{subject to} \quad A(x) = A(x_0)
\end{align*}
\]

needs $O(k)$ observations to exactly recover $x_0$ with high probability$^\ast$

**convex program:**

\[
\begin{align*}
& \text{minimize} \quad \|x\|_1 \\
& \text{subject to} \quad A(x) = A(x_0)
\end{align*}
\]

needs $O(k \log n)$ observations for exact recovery w.h.p.

$^\ast$ means: there exists constant $c$ s.t. $x_0$ is found with probability $> 1 - \exp(-cm)$

[Candes,Romberg,Tao’04; Donoho’04; Tropp’04; Fuchs’04; . . . ]
Example: Low-rank matrices and $\|X\|_*$

$\mathcal{A} : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^m$, suppose $\mathcal{A}$ is Gaussian. $X_0$ is rank $r$.

**non-convex program:**

$$\begin{align*}
\text{minimize} & \quad \text{rank}(X) \\
\text{subject to} & \quad \mathcal{A}(X) = \mathcal{A}(X_0)
\end{align*}$$

needs $O(nr)$ observations to exactly recover $X_0$ w.h.p.

**convex program:**

$$\begin{align*}
\text{minimize} & \quad \|X\|_* \\
\text{subject to} & \quad \mathcal{A}(X) = \mathcal{A}(X_0)
\end{align*}$$

also needs $O(nr)$ observations for exact recovery w.h.p.

[Recht,Fazel,Parrilo'07; Candes,Recht’08; Candes,Plan’09; Negahban,Wainwright’09,. . . ]
also true for other classic examples:

- sparse vectors (e.g., compressed sensing) \( \ell_1 \) norm
- group-sparse vectors (group LASSO) \( \ell_{1,2} \) norm
- low-rank matrices (collaborative filtering, phase retrieval, . . . ) nuclear (trace) norm
- sparse plus low-rank matrices, \( \mathbf{X} = \mathbf{L} + \mathbf{S} \) (compressive PCA, . . . ) \( \ell_1 \) plus nuclear norm
Simultaneously structured $x_0$

- object $x_0$ has several structures, each with a structure-promoting norm
- additional structures reduce degrees of freedom

consider class of convex programs

$$\begin{align*}
\text{minimize} \quad & f(x) = h(\|x\|_{(1)}, \ldots, \|x\|_{(S)}) \\
\text{subject to} \quad & A(x) = A(x_0)
\end{align*}$$

where $h : \mathbb{R}_+^S \rightarrow \mathbb{R}_+$ is convex and non-decreasing in each argument

examples:

$$\begin{align*}
f(x) &= \sum_{i=1}^{S} \lambda_i \|x\|_{(i)}, \quad f(x) = \max_{i=1,\ldots,S} \alpha_i \|x\|_{(i)} \\
\lambda_i, \alpha_i > 0 \text{ are parameters}
\end{align*}$$
Pareto optimal front

pick $m$. consider set of norm values achieved by $\{x \mid A(x) = A(x_0)\}$ and fill the upper-right points to get the Pareto optimal set for each $m$. observe

- if we have $m_1 < m$ measurements, $x_0$ doesn’t correspond to Pareto optimal front
  - cannot be recovered by minimizing any combination of norms

- need at least $m$ measurements for $x_0$ to be recoverable
Some results

• a limitation for combining convex penalties: simpler proof

• holds true for a variety of measurements $\mathcal{A}$:
  – Gaussian iid entries
  – independent subgaussian rows
  – sampling operator
    (e.g., sampled rows of identity as in ‘completion’ problems, or sampled rows of Fourier matrix)
  – quadratic (or rank-1) measurements: $\langle a_i a_i^T, X \rangle = b_i^2$

• special case: sparse and low-rank matrix

[Oymak et al. '12,'15 ]
Recovery failure: sufficient condition

suppose $x_0$ has structures $i = 1, \ldots, S$. when does program

$$\text{minimize } f(x) = h(\|x\|_{(1)}, \ldots, \|x\|_{(S)}) \quad \text{subject to } Ax = Ax_0$$

fail to give $x_0$ as its solution?

**Theorem.** if

$$\inf_{g \in \partial f(x_0)} |\bar{g}^T \bar{x}_0| > \frac{\|A\bar{x}_0\|_2}{\sigma_{\min}(A)},$$

where $\bar{x}_0 = \frac{x_0}{\|x_0\|_2}$, $\bar{g} = \frac{g}{\|g\|_2}$, then $x_0$ is not a minimizer and recovery fails.

($\partial f(x_0)$ is the set of subgradients of $f$ at $x_0$)
Recovery failure: sufficient condition

**Theorem.** If

$$\inf_{g \in \partial f(x_0)} |\bar{g}^T \bar{x}_0| > \frac{\|A\bar{x}_0\|_2}{\sigma_{\min}(A)},$$

then $x_0$ is **not** a minimizer and recovery fails.

- LHS depends only on $f$ and $\bar{x}_0$
- cannot be made too small, as subgradients are ‘aligned’ with $\bar{x}_0$ (we bound this with a geometric quantity)
- RHS depends only on $A$ and $\bar{x}_0$
- for many random ensembles, RHS $\gtrsim \sqrt{\frac{m}{n}}$
Recovery failure: sufficient condition

**Theorem.** If

\[
\inf_{g \in \partial f(x_0)} |\bar{g}^T \bar{x}_0| > \frac{\|A\bar{x}_0\|_2}{\sigma_{\min}(A)},
\]

then \(x_0\) is **not** a minimizer and recovery fails.

- LHS depends only on \(f(\cdot)\) and \(\bar{x}_0\)
- cannot be made too small, as subgradients are ‘aligned’ with \(\bar{x}_0\)
- RHS depends only on \(A\) and \(\bar{x}_0\)
- for many random ensembles, RHS \(\gtrsim \sqrt{\frac{m}{n}}\)
Recovery failure: sufficient condition

**Theorem.** If

$$\inf_{g \in \partial f(x_0)} |\bar{g}^T \bar{x}_0| > \frac{||A\bar{x}_0||_2}{\sigma_{\min}(A)}$$

then $x_0$ is not a minimizer and recovery fails.

- LHS depends only on $f$ and $\bar{x}_0$
- cannot be made too small, as subgradients are ‘aligned’ with $\bar{x}_0$
- RHS depends only on $A$ and $\bar{x}_0$
- for many random ensembles, RHS $\approx \sqrt{\frac{m}{n}}$
Bound the left-hand side

def.: correlation between $x_0$ and set $S$ (largest angle)

$$\rho(x, S) = \inf_{0 \neq s \in S} |\bar{x}^T \bar{s}|$$

if set $S$ is subdiff. of norm $i$:

$$\rho(x_0, \partial \|x_0\|(i)) = \frac{\|\bar{x}_0\|(i)}{\sup_{g \in \partial \|x_0\|(i)} \|g\|_2} \geq \frac{\|\bar{x}_0\|(i)}{L_i} := \kappa_i$$

where $L_i$ is the norm’s Lipschitz constant. now lower bound the LHS,

$$\inf_{g \in \partial f(x_0)} |\bar{g}^T \bar{x}_0| \geq \kappa_{\text{min}} = \min_i \kappa_i.$$ 

(see also [Mu,Huang,Wright,Goldfarb'13])
proof:

- from convex analysis: any subgradient of $f = h(\|x\|_{(1)}, \ldots, \|x\|_{(S)})$ can be written as $g = \sum_i w_i g_i$ with $w_i \geq 0$, where $g_i$ is a subgradient of $\|x\|_{(i)}$

- $g^T \bar{x}_0 = \sum_i w_i \|\bar{x}_0\|_{(i)}$ (since $g_i^T \bar{x}_0 = \|\bar{x}_0\|_{(i)}$

- $\|g\|_2 \leq \sum_i w_i \|g_i\|_2 \leq \sum_i w_i L_i$, so

$$\inf_{g \in \partial f(x_0)} |g^T \bar{x}_0| \geq \frac{\sum_i w_i \|\bar{x}_0\|_{(i)}}{\sum_i w_i L_i} \geq \min_i \frac{w_i \|\bar{x}_0\|_{(i)}}{w_i L_i} = \kappa_{\min}.$$
random vector $\mathbf{x} \in \mathbb{R}^n$ is subgaussian, if marginals $\mathbf{x}^T \mathbf{v}$ are subgaussian random variables for all $\mathbf{v} \in \mathbb{R}^n$

**Lemma [subgaussian measurements]** if $\mathbf{A}$ has

- i.i.d zero-mean isotropic subgaussian rows, or
- i.i.d zero-mean, unit-variance subgaussian entries

there exists constant $c_1$ such that whenever $m \leq c_1 n$, w.h.p. we have

$$\frac{\| \mathbf{A} \mathbf{x}_0 \|_2^2}{\sigma_{\min}^2(\mathbf{A})} \leq \frac{2m}{n}$$

[see review by Vershynin,’14]
Bound the left-hand side

**Lemma [sampling]** sample rows of $\mathbf{A}$ uniformly from the identity matrix, discard duplicate rows. Then w.h.p.,

$$\frac{\|\mathbf{A}\mathbf{x}_0\|_2^2}{\sigma_{\text{min}}^2(\mathbf{A})} \leq \frac{2m}{n}$$

several other measurements give similar bounds, e.g., $\langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X} \rangle = b_i^2$ (also studied in [Li, Voroninski ’12])
Recovery failure

putting bounds together:

**Theorem.** $x_0$ will not be a minimizer of the recovery program w.h.p., if

$$m \leq c \, n\kappa_{\min}^2$$

for all measurement types mentioned.

**Examples:**

<table>
<thead>
<tr>
<th>model</th>
<th>$f(\cdot)$</th>
<th>$L$</th>
<th>$|\bar{x}_0| \leq$</th>
<th>$m$ at least</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$ sparse vector</td>
<td>$|\cdot|_1$</td>
<td>$\sqrt{n}$</td>
<td>$\sqrt{k}$</td>
<td>$k$</td>
</tr>
<tr>
<td>$k$ column-sparse matrix</td>
<td>$|\cdot|_{1,2}$</td>
<td>$\sqrt{d}$</td>
<td>$\sqrt{k}$</td>
<td>$kd$</td>
</tr>
<tr>
<td>rank $r$ matrix</td>
<td>$|\cdot|_*$</td>
<td>$\sqrt{d}$</td>
<td>$\sqrt{r}$</td>
<td>$rd$</td>
</tr>
<tr>
<td>sparse &amp; Low-rank matrix</td>
<td>$h(|\cdot|_*, |\cdot|_1)$</td>
<td>$-$</td>
<td>$-$</td>
<td>$\min{k^2, rd}$</td>
</tr>
</tbody>
</table>

last three lines: $d \times d$ matrix with $k \times k$ nonzero block, rank $r$, and $n = d^2$
Sparse and low-rank case

A gap. A nonconvex problem can recover the model from few measurements (on order of the degrees of freedom), while combined convex penalties require much more measurements (suppose $\mathcal{A}$ is Gaussian).

\[
\mathcal{O}(nr) \quad m \\
\mathcal{O}(k \log \frac{n}{k} + kr) \quad \text{gap} \\
r(2k - r) \quad \text{“degrees of freedom”}
\]

\[
\begin{align*}
X_0 \text{ is not a minimizer of} & \quad \min_{X} \|X\|_* + \lambda \|X\|_{1,2} \\
\text{s.t.} & \quad \mathcal{A}(X) = \mathcal{A}(X_0) \\
& \quad X \succeq 0 \\
\end{align*}
\]

\[
\begin{align*}
X_0 \text{ is the unique minimizer of} & \quad \min_{X} \text{rank}(X) + \lambda \|X\|_{0,2} + \lambda \|X^T\|_{0,2} \\
\text{s.t.} & \quad \mathcal{A}(X) = \mathcal{A}(X_0) \\
\end{align*}
\]
Numerical experiments

grayscale shows probability of success over 100 runs for each case. recovery using
\[ f(X) = \text{Tr}(X) + \lambda \|X\|_1. \]
\( X_0 \) is PSD, rank 1, \( k = 8 \).

Figure 1: \( \lambda = 0.2 \) (left) and \( \lambda = 0.35 \) (right).
A related problem: Denoising

denosing this bottleneck also appears in another problem:

Suppose \( x_0 \) has \( S \) structures; estimate \( x_0 \in \mathbb{R}^n \) given \( y = x_0 + z \), where \( z \sim \mathcal{N}(0, \sigma^2 I) \). Use:

\[
\hat{x} = \arg \min_x \{ \|y - x\|_2^2 + \sum_{i=1}^{S} \lambda_i \|x\|_{(i)} \}
\]

(aka proximal operator of function \( \sum_{i=1}^{S} \lambda_i \|x\|_{(i)} \))

**Def.** The **MSE risk** of above estimator at \( x_0 \) is

\[
\eta(\lambda_1, \ldots, \lambda_S) = \max_{\gamma > 0} \mathbb{E}[\|\hat{x}(\gamma x_0 + z) - \gamma x_0\|_2^2]
\]

How low can MSE risk get?

Can show: performance is order-wise the same as using the best single norm

Similar statements for the case \( y = A(x) + z \)
Summary

• simultaneously structured models: weighted sum of norms is often used in applications, lacked performance theory

• result: combined convex penalty displays a fundamental gap, both for recovery sample complexity and denoising error

• lower bound holds for various measurements, e.g., sampling (matrix or tensor completion), quadratic measurements (phase retrieval, sparsePCA)

• tight upper bounds on $m$ can be obtained for the Gaussian case, for lin comb of norms with $\lambda_i$’s tuned optimally; differs from lower bounds by a log factor [Oymak et al, 2015]
Discussion: what next?

is the situation all gloomy...?

- find better penalty/regularizer:
  - can we directly define atoms and take convex hulls to find the atomic norm? [Chandrasekaran et al’10] seems intractable for sparse and low-rank case, but may help in other problems
  - convex relaxation hierarchies for the atomic norm
  - some improvements (though not orderwise) on a case-by-case basis:
    * tensors with low Tucker rank [Mu, Huang, Wright, Goldfarb ’13]
    * a relaxation for sparse and low-rank [Richard, ’]

- find more suitable measurements schemes (e.g., sequential measurements [Bahmani, Romberg ’15])
References
