Recovery of Simultaneously Structured Models using Convex Optimization

Maryam Fazel

University of Washington

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

IPAM Workshop, 1/18/2013

Structured models

models with **low-dimensional structure** (low "degrees of freedom"), living in a high-dimensional ambient space

goal: recover/derive such a model from limited observations

applications: signal processing, machine learning, system identification,...

questions: are there suitable convex regularizers? how to quantify their performance?

Typical structured models

- sparse vector (compressed sensing, LASSO,. . .)
- group-sparse vectors (group LASSO)
- low-rank matrix (collaborative filtering, system identification, . . .)
- sparse *plus* low-rank matrix (graphical models with hidden variables, PCA with outliers)
- simultaneously sparse *and* low-rank (phase retrieval)

Typical structured models

- sparse vector (compressed sensing, LASSO,. . .)
- group-sparse vectors (group LASSO)
- low-rank matrix (collaborative filtering, system identification, . . .)
- sparse *plus* low-rank matrix (graphical models with hidden variables, PCA with outliers)
- simultaneously sparse *and* low-rank (phase retrieval)

Recovery of structured models

basic setup: unknown (structured) model $\mathbf{x}_0 \in \mathbf{R}^n$;
we are given observations $\mathcal{G}(\mathbf{x}_0) = \mathbf{y}$ where $\mathcal{G} : \mathbf{R}^n \to \mathbf{R}^m$ is a linear map, $m \ll n$

goal: given $\mathcal{G}$, $\mathbf{y} \in \mathbf{R}^m$ (and structure type), find $\mathbf{x}_0$.

for different structures, much recent research has focused on

- how to find desired model from underdetermined observations?
- how many measurements m suffice? (sample complexity)

for analysis, assume **generic measurements** $\mathcal{G}$: $m \times n$ measurement matrix with **i.i.d. Gaussian** entries.

Example: Sparse vectors and $\|\mathbf{x}\|_1$

generic measurements $\mathcal{G} : \mathbf{R}^n \to \mathbf{R}^m$. $\mathbf{x}_0$ is k-sparse.

non-convex program:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{x}\|_0 \\ \text{subject to} & \mathcal{G}(\mathbf{x}) = \mathcal{G}(\mathbf{x}_0) \end{array}$$

needs $\mathcal{O}(k)$ observations to exactly recover $\mathbf{x}_0$ with high probability (probability goes to 1 exponentially with m)

convex program:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{x}\|_1 \\ \text{subject to} & \mathcal{G}(\mathbf{x}) = \mathcal{G}(\mathbf{x}_0) \end{array}$$

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

some past work: Candes,Romberg,Tao'04; Donoho'04; Tropp'04; Fuchs'04; . . .

Example: Low-rank matrices and $\|\mathbf{X}\|_*$

generic measurements $\mathcal{G} : \mathbf{R}^{n\times n} \to \mathbf{R}^m$. $\mathbf{X}_0$ is rank r.

non-convex program:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\mathbf{X}) \\ \text{subject to} & \mathcal{G}(\mathbf{X}) = \mathcal{G}(\mathbf{X}_0) \end{array}$$

needs $\mathcal{O}(nr)$ observations to exactly recovers $\mathbf{X}_0$ w.h.p

convex program:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{X}\|_* \\ \text{subject to} & \mathcal{G}(\mathbf{X}) = \mathcal{G}(\mathbf{X}_0) \end{array}$$

also needs $\mathcal{O}(nr)$ observations for exact recovery w.h.p.

some past work: Fazel'01; Srebro'04; Recht,Fazel,Parrilo'07; Candes,Recht'08; Candes,Plan'09; Keshavan et al.'09; Negahban et al.'09,. . .

This talk: Simultaneous structures

- model of interest is known to be structured in *several* ways
- additional structures reduce degrees of freedom

  we hope for recovery with fewer observations

example: matrix is both (block-)sparse and low-rank: $\mathbf{X}_0 \in \mathbf{R}^{n \times n}$

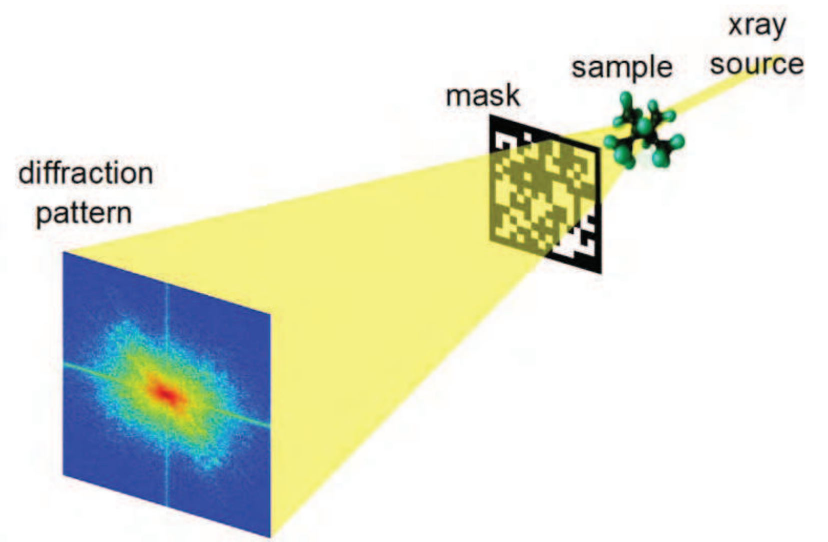
- $\mathrm{rank}(\mathbf{X}_0) = r$ with $r \ll n$
- $\mathbf{X}_0$ supported over a $k \times k$ submatrix

Application: Sparse phase retrieval

phase retrieval: a classic signal processing/optics problem

recover signal $\mathbf{x}_0$ from linear *phaseless* measurements,

$$|\mathbf{a}_i^T \mathbf{x}_0| = b_i, \quad i = 1, \ldots, m$$



reformulate as: find $\mathbf{X} = \mathbf{x}_0 \mathbf{x}_0^T$ s.t. $\langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X} \rangle = b_i^2$

i.e., $\quad \mathbf{X} \succeq 0,\ \mathrm{rank}(\mathbf{X}) = 1,\ \mathcal{A}(\mathbf{X}) = b'$ [Candes,Eldar,Strohmer,Voroninski'11]

in applications, signal $\mathbf{x}_0$ is also often **sparse**. then, $\mathbf{X}$ is rank-1 and (block-)sparse

'combination of norms' recovery program

consider class of convex programs

$$\begin{array}{ll} \underset{\mathbf{x}\in\mathcal{C}}{\text{minimize}} & f(\mathbf{x}) = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(\tau)}) \\ \text{subject to} & \mathcal{G}(\mathbf{x}) = \mathcal{G}(\mathbf{x}_0), \end{array}$$

where $h : \mathbf{R}_+^\tau \to \mathbf{R}_+$ is increasing with respect to the order induced by $\mathbf{R}_+^\tau$.

examples:

$$f(x) = \sum_{i=1}^{\tau} \lambda_i \|x\|_{(i)}$$

where $\lambda_i > 0$ are regularization parameters.

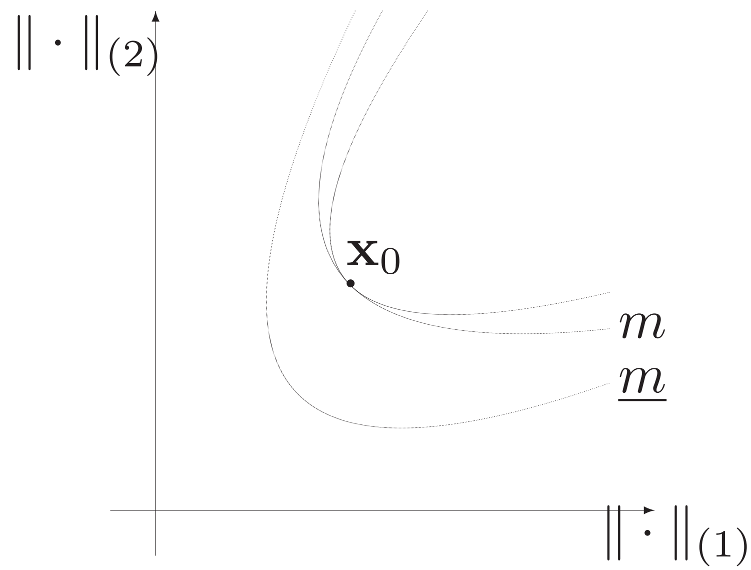
$$f(\mathbf{x}) = \max_{i=1,\ldots,\tau} \frac{1}{\|\mathbf{x}_0\|_{(i)}} \|\mathbf{x}\|_{(i)}$$

Pareto optimal front

- sets of achievable objective values shrink as the number of measurements grows, always containing $\mathbf{x}_0$

- for $\mathbf{x}_0$ to be recoverable; for any $\underline{m} < m$, $\mathbf{x}_0$ is not on the Pareto optimal front.

- need at least m measurements



Our results

- theoretical analysis of general simultaneous structures
- performance of combined convex penalties, and a **fundamental limitation**
- special case of sparse and low-rank matrix problem
  - performance of convex vs nonconvex penalty, and a **gap**

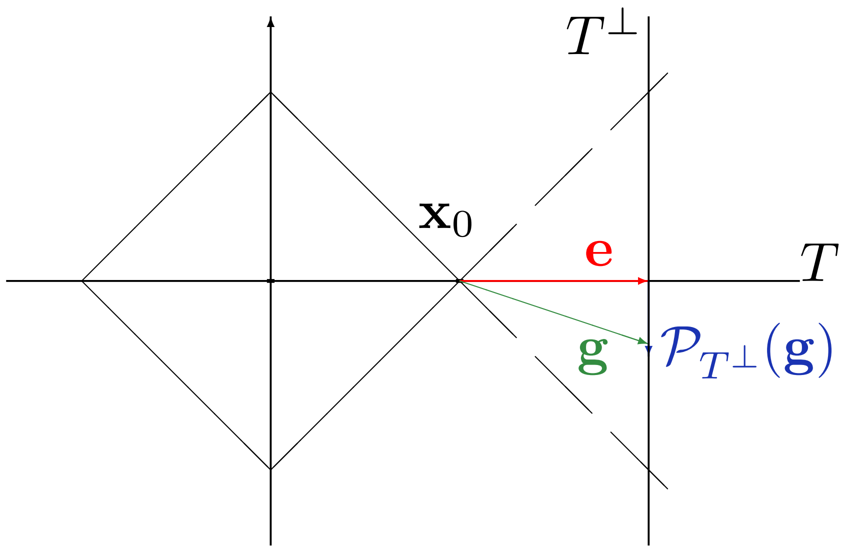
Decomposable norms

Definition. norm $\|\cdot\|$ is decomposable at $\mathbf{x}$ if there exist

$$\text{subspace } T \subset \mathbf{R}^n \text{ (support)}, \quad \text{vector } \mathbf{e} \in T \text{ (sign)}$$

such that $$\partial\|\mathbf{x}\| = \{\mathbf{z} \in \mathbf{R}^n \ : \quad \mathcal{P}_T(\mathbf{z}) = \mathbf{e}, \quad \|\mathcal{P}_{T^\perp}(\mathbf{z})\|_* \leq 1\}$$

and for all $\mathbf{y} \in T^\perp$, $\|\mathbf{y}\| = \sup_{\mathbf{z} \in T^\perp, \|\mathbf{z}\|^* \leq 1} \langle \mathbf{y}, \mathbf{z} \rangle$.

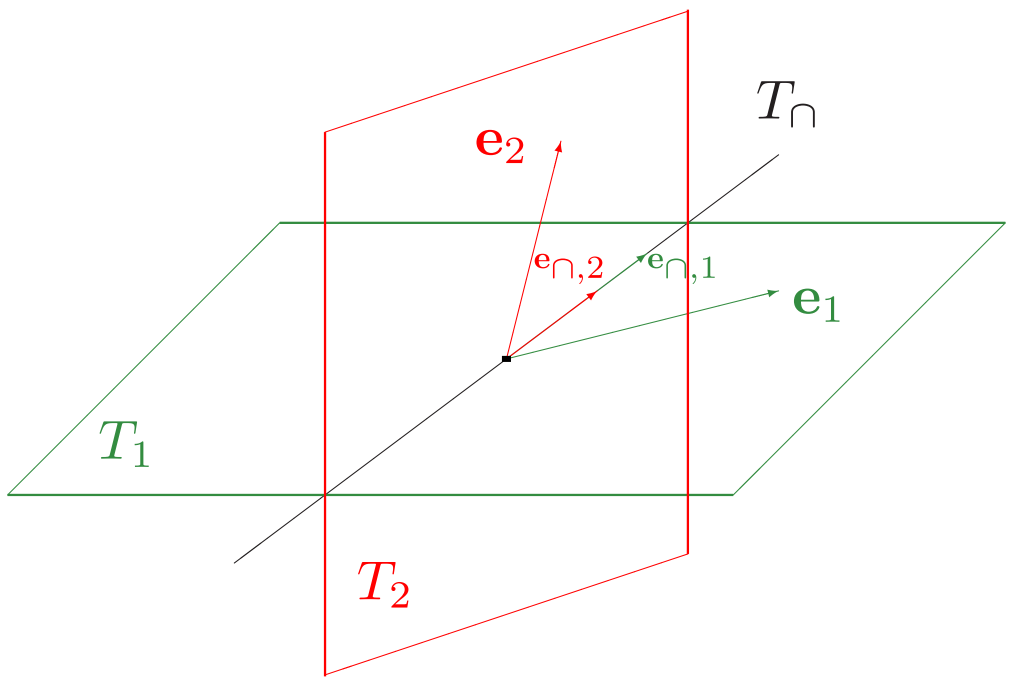


[Candes,Recht'11;Wright et al.'12;Negahban et al.'09]

Simultaneously structured models

suppose norms $\{\|\cdot\|_{(i)}\}_{i=1}^{\tau}$ are decomposable at $\mathbf{x}_0$. $\mathbf{x}_0$ is a *simultaneously structured object* with

- sign vectors $\mathbf{e}_i$, supports T_i, joint support $T_\cap = \bigcap_{i=1}^{\tau} T_i$
- projected signs $\mathbf{e}_{\cap,i} = \mathcal{P}_{T_\cap}(\mathbf{e}_i)$, "angles" $\theta_i = \frac{\|\mathbf{e}_{\cap,i}\|_2}{\|\mathbf{e}_i\|_2}$



Lower bound on measurements for recovery

consider class of convex programs

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(\tau)}) \\ \text{subject to} & \mathcal{G}(\mathbf{x}) = \mathcal{G}(\mathbf{x}_0), \end{array}$$

Theorem 1. program above *fails* to recover $\mathbf{x}_0$ with high probability if

$$m < \frac{n}{81} \inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} \frac{\|\mathcal{P}_{T_\cap}(\mathbf{g})\|_2^2}{\|\mathbf{g}\|_2^2}$$

Main result

Assumption. $\forall\, i \neq j$, let $\langle \mathbf{e}_{\cap,i}, \mathbf{e}_{\cap,j} \rangle \geq 0$.

Theorem 2. Suppose assumption holds. Then program above fails to recover $\mathbf{x}_0$ with high probability, if

$$m < \frac{\kappa\theta^2}{81\tau} \min_i \dim(T_i)$$

note: need measurements on the order of $\min_i \dim(T_i)$, rather than $\dim(T_\cap)$!

can handle also additional cone constraints on $\mathbf{x}_0$ (affects the constant) quantity $\kappa = \min_i \kappa_i$ where

$$\kappa_i = \frac{n}{\dim(T_i)} \frac{\|\mathbf{e}_i\|_2^2}{L_i^2},$$

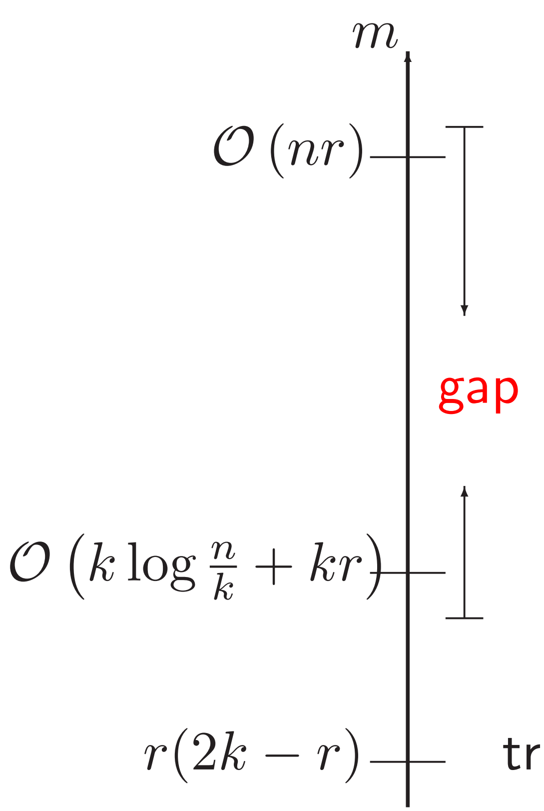
and L is Lipschitz constant of the norm $\quad L = \sup_{\mathbf{z}_1 \neq \mathbf{z}_2} \frac{\|\mathbf{z}_1\| - \|\mathbf{z}_2\|}{\|\mathbf{z}_1 - \mathbf{z}_2\|_2}$

examples. for ℓ_1, $\ell_{1,2}$, and nuclear norm,

$$\kappa_1 = 1, \quad \kappa_{1,2} = 1, \quad 1/2 \leq \kappa_* \leq 1$$

Sparse and low-rank case

a surprising gap. while a nonconvex problem can recover the model from very few measurements (on order of the degrees of freedom), combined convex penalties requires much more measurements.



$\mathbf{X}_0$ is *not* a minimizer of

$$\begin{array}{rl} \text{min.} & \|\mathbf{X}\|_* + \lambda \|\mathbf{X}\|_{1,2} \\ \text{s.t.} & \mathcal{G}(\mathbf{X}) = \mathcal{G}(\mathbf{X}_0) \\ & \mathbf{X} \succeq 0 \end{array}$$

$\mathbf{X}_0$ is the *unique* minimizer of

$$\begin{array}{rl} \text{min.} & \operatorname{rank}(\mathbf{X}) + \lambda \|\mathbf{X}\|_{0,2} + \lambda \|\mathbf{X}^T\|_{0,2} \\ \text{s.t.} & \mathcal{G}(\mathbf{X}) = \mathcal{G}(\mathbf{X}_0) \end{array}$$

true "degrees of freedom"

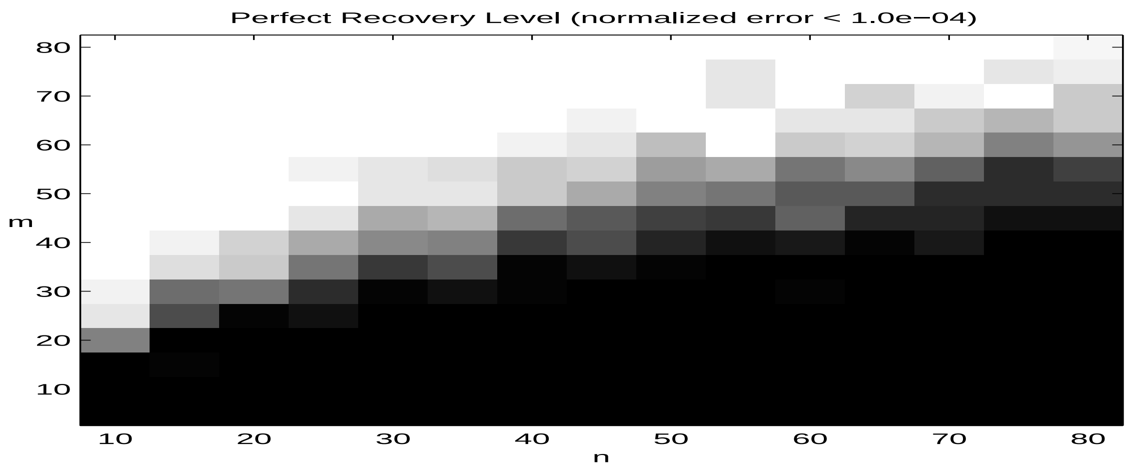
summary of recovery results, for $\mathbf{X} \in \mathbf{R}^{n\times n}$, supported over a $k \times k$ submatrix.

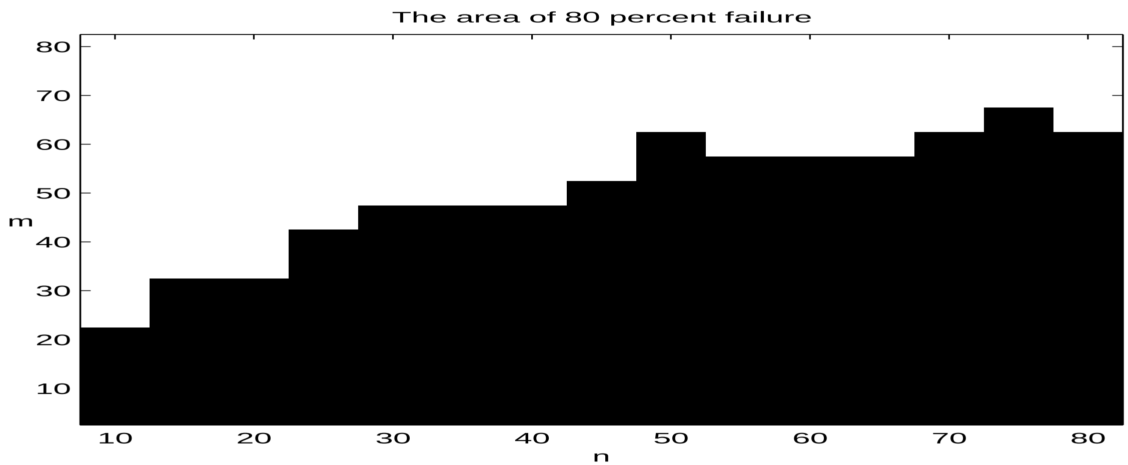
nonconvex approaches are optimal up to a logarithmic factor, while convex approaches perform poorly.

Setting	Nonconvex sufficient m	Convex required m
General model	$O(\max\{rk, k\log\frac{n}{k}\})$	$\Omega(rn)$
PSD, arbitrary rank	$O(\max\{rk, k\log\frac{n}{k}\})$	$\Omega(rn)$
PSD, rank 1	$O(k\log\frac{n}{k})$	$\Omega(\min\{k^2, n\})$

Numerical experiments

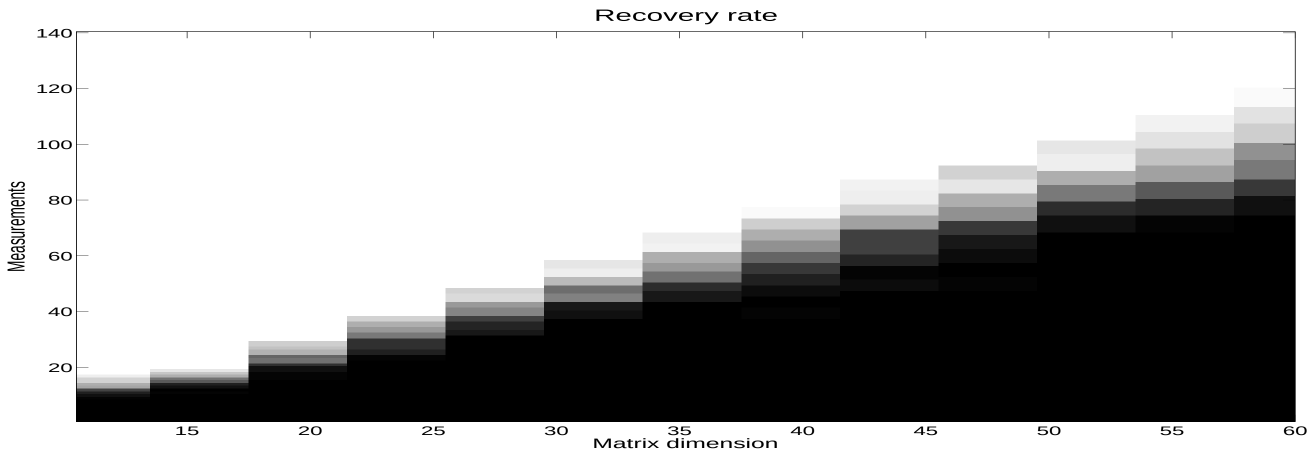
grayscale shows probability of success over 25 runs for each case. recovery using $f(\mathbf{X}) = \mathrm{Tr}(\mathbf{X}) + \lambda \|\mathbf{X}\|_1$. $\mathbf{X}_0$ is PSD, rank 1, $k = 8$. n ranges up to 80.

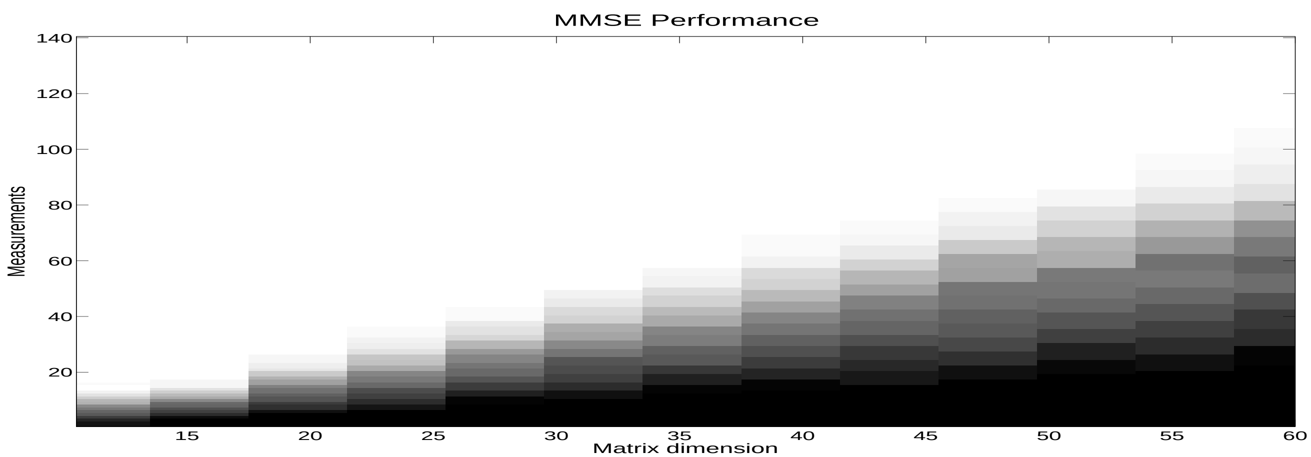




Numerical experiments

grayscale shows probability of success over 25 runs for each case. recovery using $f(\mathbf{X}) = \mathrm{Tr}(X) + \lambda \|\mathbf{X}\|_{1,2}$ with PSD constraint. $\mathbf{X}_0$ is PSD, rank 1, $k = 7$, n ranges up to 60.





Summary

- regularizers for recovery of a model known to have several structures simultaneously
- result: combined convex penalty requires many more generic measurements than degrees of freedom
- contrast with card vs ℓ_1, rank vs $\| \cdot \|_*$, . . .

Future work

- recovery error and phase transition
- can we directly define atoms and take convex hulls to find better norm in some cases?
- partial relaxation
- other applications
- other measurement models, e.g., phase retrieval measurements $\langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X} \rangle = b_i'$

Reference

- "Simultaneously Structured Models with Application to Sparse and Low-rank Matrices",

  Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C. Eldar, Babak Hassibi. arXiv:1212.3753, Dec 2012.