

# PCA with Outliers and Missing Data

Sujay Sanghavi

Electrical and Computer Engg.  
University of Texas, Austin

Joint w/ C. Caramanis, Y. Chen, H. Xu

# Outline

---

## PCA and Outliers

- Why SVD fails
- Corrupted features vs. corrupted points

## Our idea + Algorithms

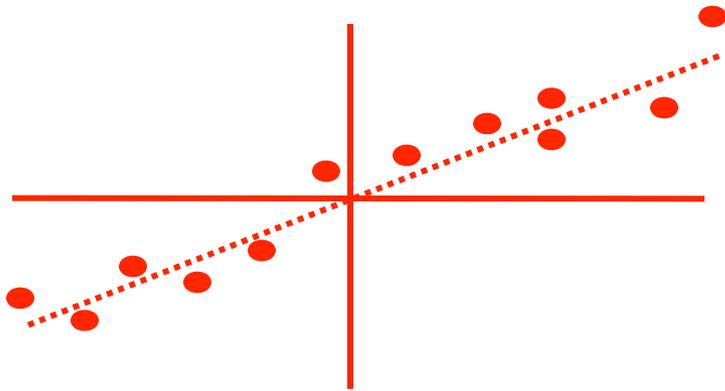
## Results

- Full observation
- Missing Data

**Framework** for Robustness in High Dimensions

# Principal Components Analysis

---



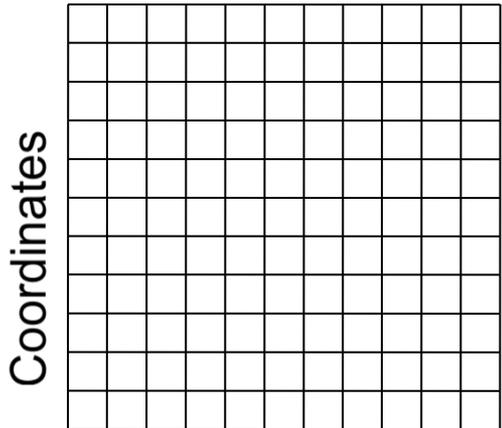
Given points that lie on/near a Lower dimensional subspace, **find this subspace.**

**Classical technique:**

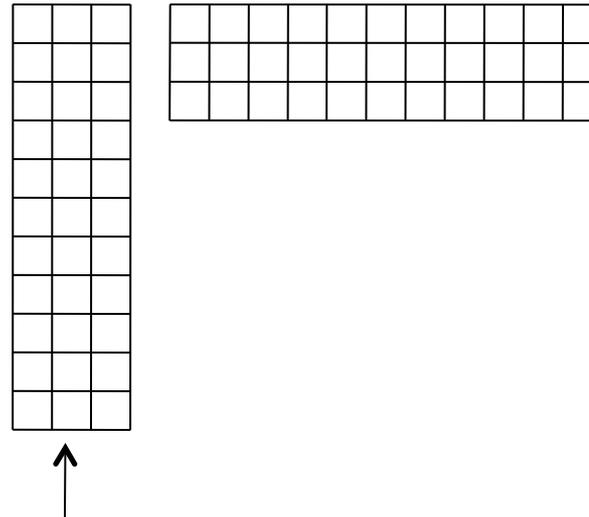
1. Organize points as matrix
2. Take SVD
3. Top singular vectors span space

Data Points

Low Rank Matrix



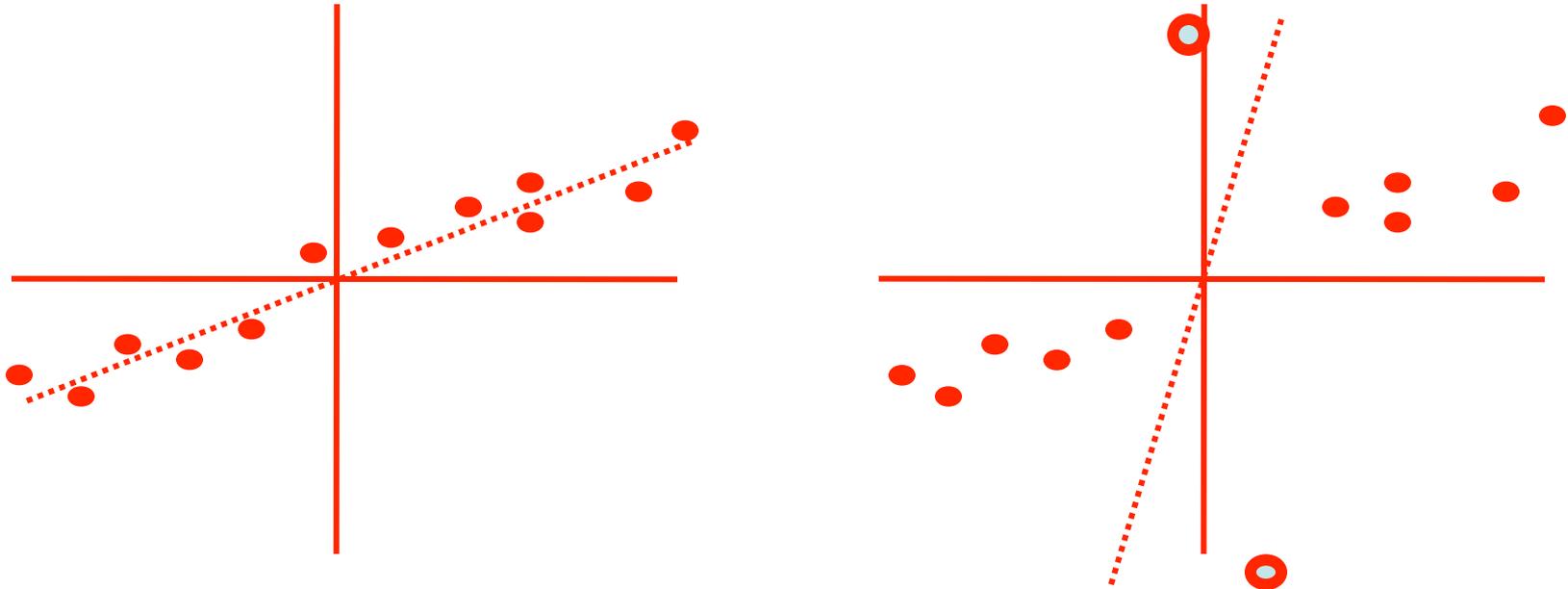
$\approx$



# Fragility

---

Gross errors of even one/few points can completely throw off PCA

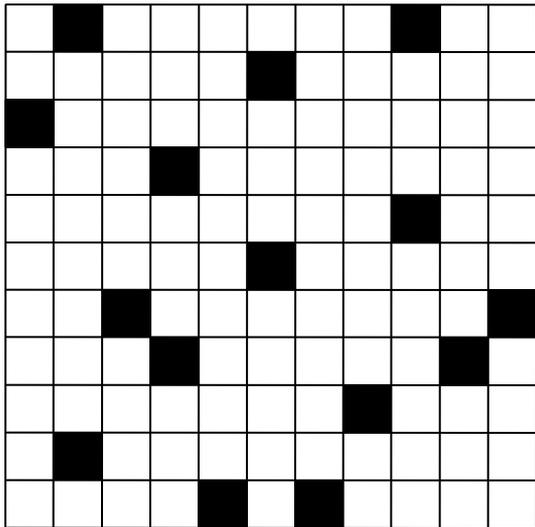


**Reason:** Classical PCA minimizes  $\ell_2$  error, which is susceptible to gross outliers

# Two types of gross errors

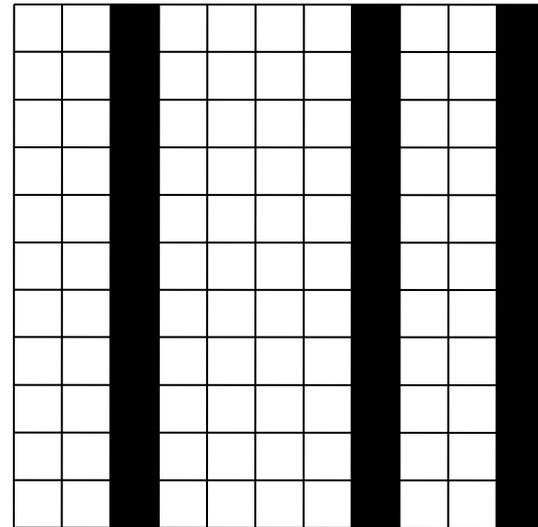
---

## Corrupted Features



- Individual entries corrupted

## Outliers

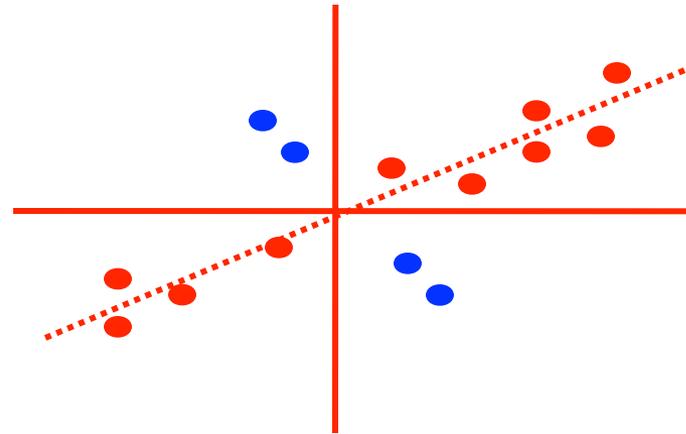
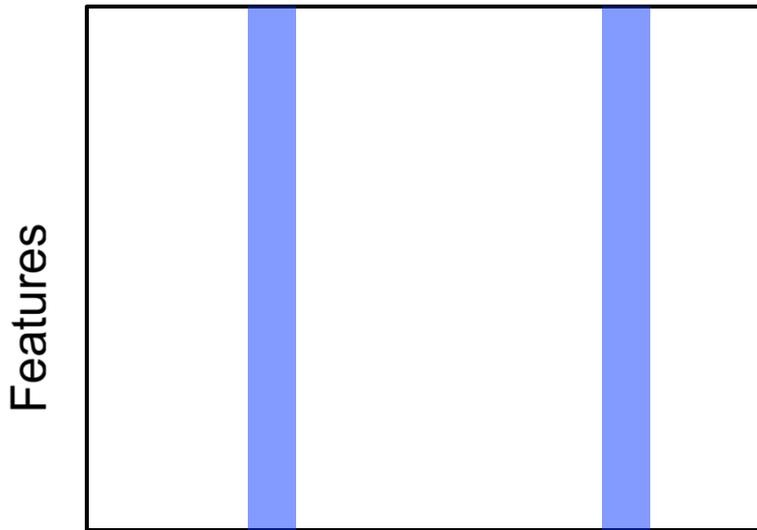


- Entire columns corrupted

.. and **missing data** versions of both

# PCA with Outliers

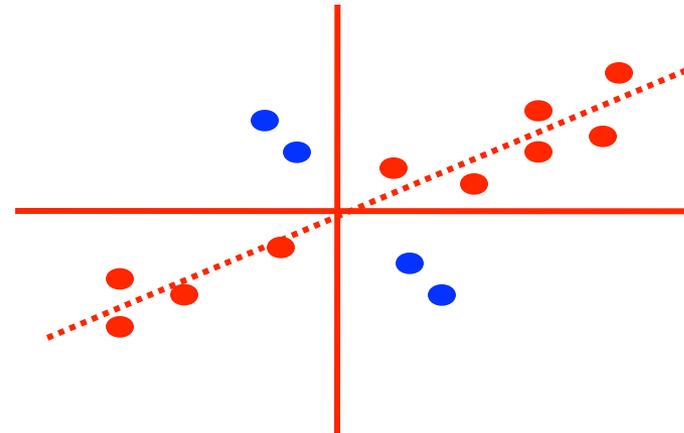
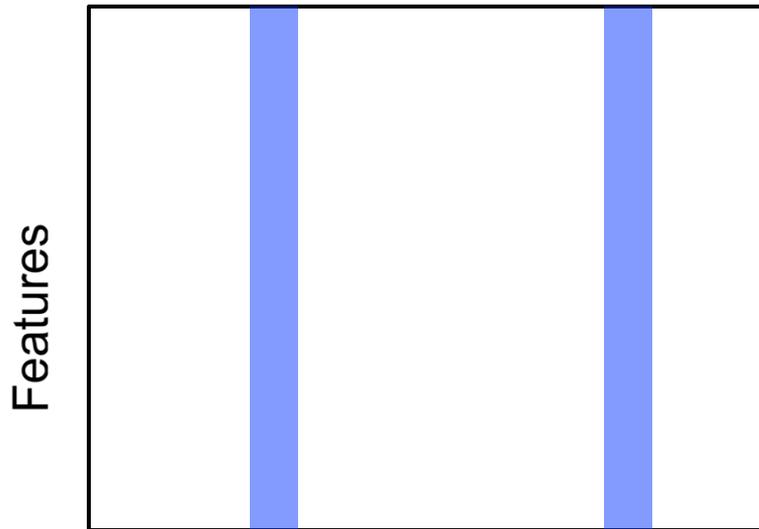
Points



**Objective: find identities of outliers**  
(and hence col. space of true matrix)

# Outlier Pursuit - Idea

Points



Standard PCA

$$\min_L \|M - L\|_F$$

$$s.t. \text{rank}(L) = r$$

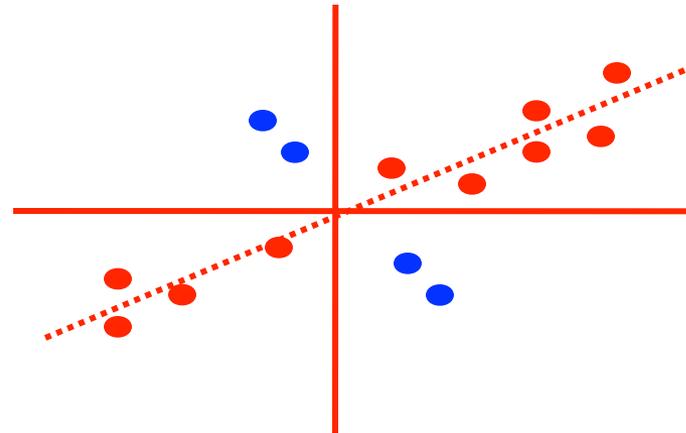
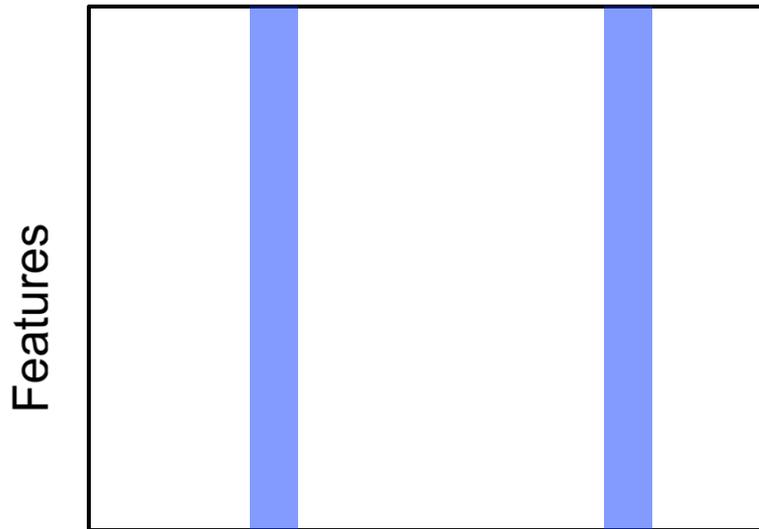
$$\min_{L, C} \|M - L - C\|_F$$

$$s.t. \text{rank}(L) = r$$

$$\text{col}(C) = c$$

# Outlier Pursuit - Method

Points



We propose:

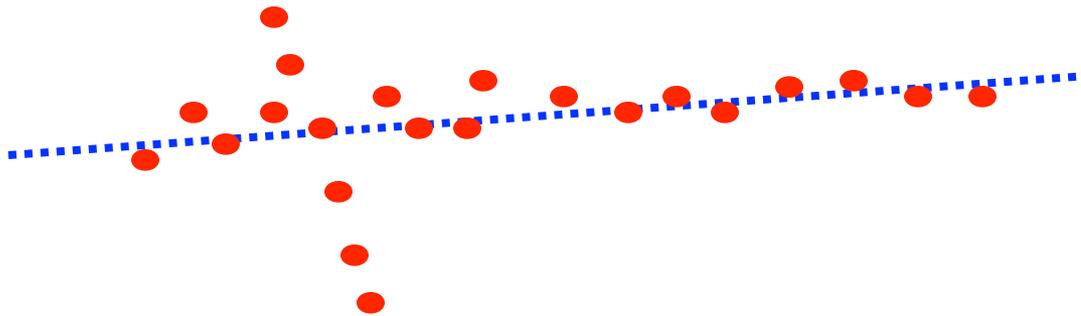
$$\min_{L, C} \|M - L - C\|_F + \lambda_1 \|L\|_* + \lambda_2 \|C\|_{1,2}$$

Convex surrogate for  
Rank constraint

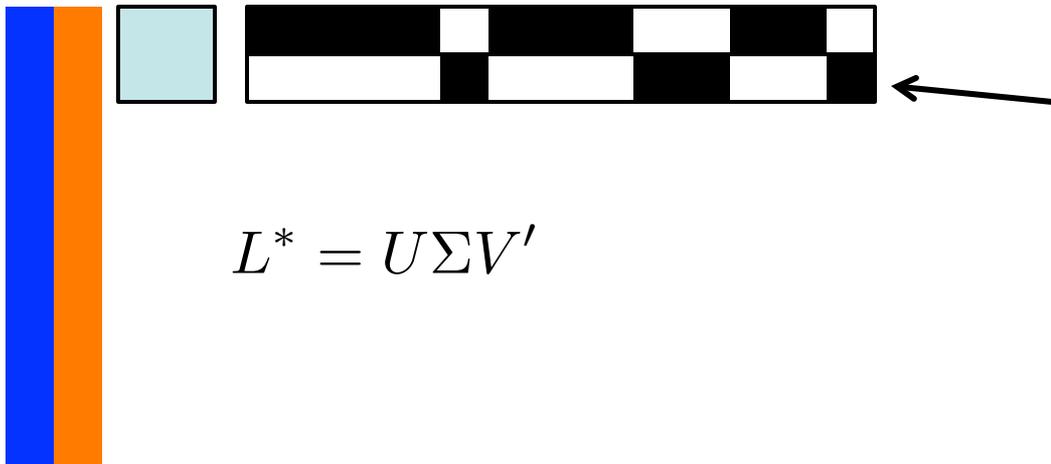
Convex surrogate for  
Column-sparsity

# When does it (not) work ?

---



When certain directions of column space of  $L^*$  poorly represented



This vector has large inner product with some coordinate axes

$$\max_i \|V'e_i\| \quad \text{is large}$$

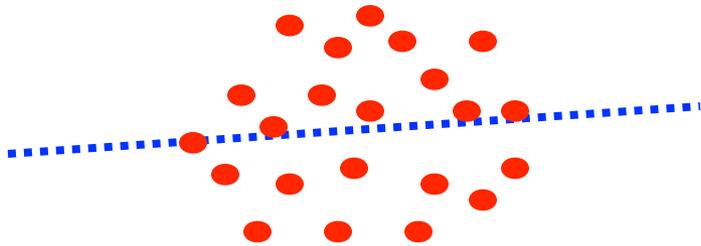
# Results

---

## Assumption:

Columns of true  $L^*$  are **incoherent**:

$$\max_i \|V' e_i\|^2 \leq \frac{\mu r}{n}$$



Note:  $r \leq \mu r \leq n$

First consider: **Noiseless case**

$$\min_{L, C} \|L\|_* + \lambda \|C\|_{1,2}$$

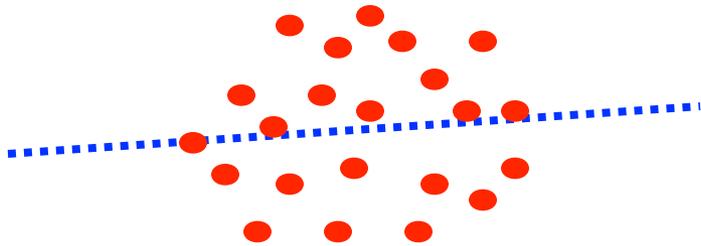
$$s.t. \quad L + C = M$$

# Results

## Assumption:

Columns of true  $L^*$  are **incoherent**:

$$\max_i \|V' e_i\|^2 \leq \frac{\mu r}{n}$$



Note:  $r \leq \mu r \leq n$

## Theorem: (noiseless case)

Our convex program can identify upto a fraction  $\gamma$  of outliers as long as

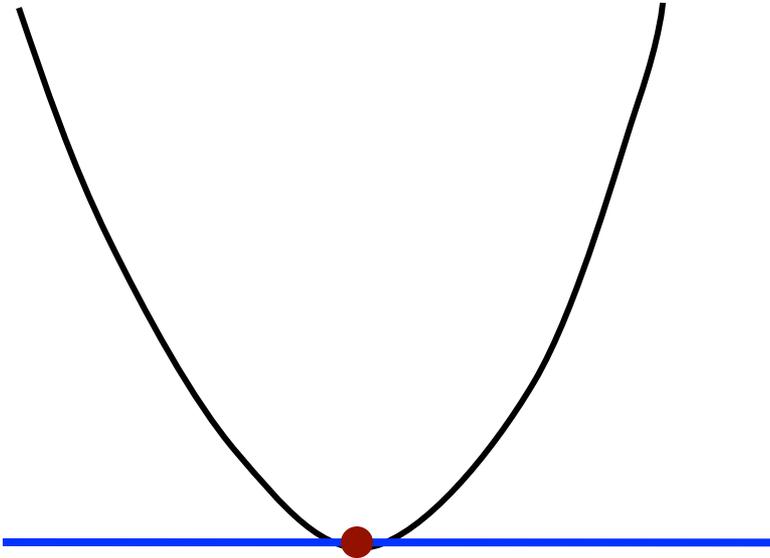
$$\frac{\gamma}{1 - \gamma} \leq \frac{c}{\mu r}$$

$$\lambda = \frac{3}{7\sqrt{\gamma n}}$$

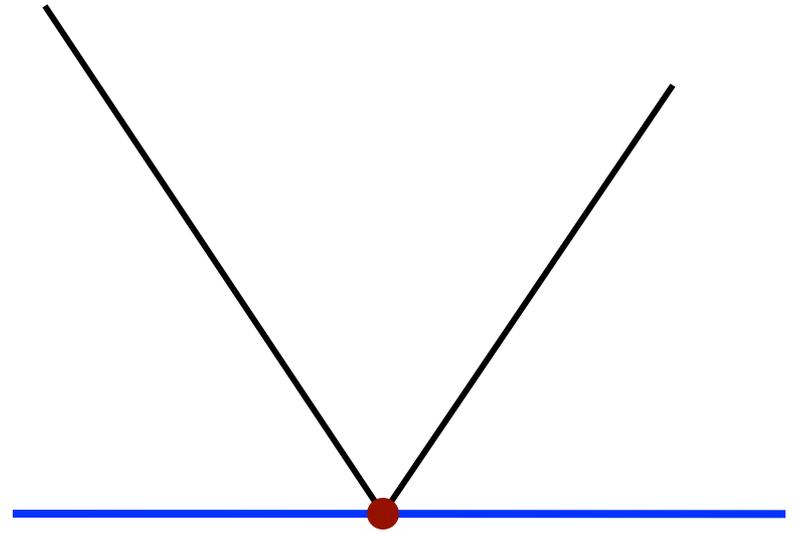
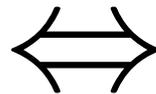
**Outer bound:**  $\gamma > \frac{1}{r+1}$  makes the problem un-identifiable

# Proof Technique

---



A point  $x$  is the optimum of a convex function  $f$



Zero lies in the (sub) gradient  $\partial f(x)$  of  $f$  at  $x$

**Steps:** 1. guess a “nice” point, -- oracle problem

2. show it is the optimum by showing zero is in subgradient

# Proof Technique

---

## Guessing a “nice” optimum

(Note: in “single structure” problems like matrix completion, compressed sensing etc., this is not an issue)

### Oracle Problem:

$$\min_{L, C} \|M - L - C\|_F + \lambda_1 \|L\|_* + \lambda_2 \|C\|_{1,2}$$

$$s.t. \text{ ColSupp}(C) \subset \text{ColSupp}(C^*)$$

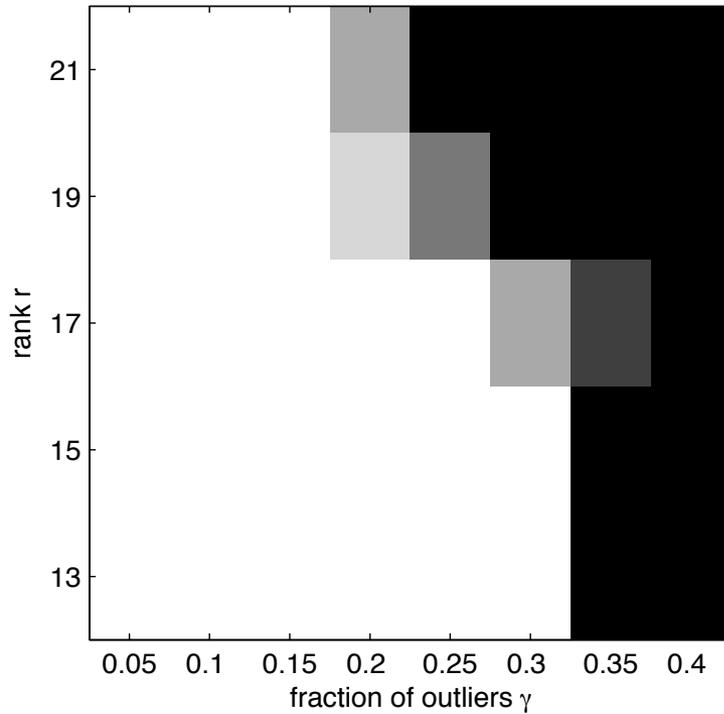
$$\text{ColSpace}(L) \subset \text{ColSpace}(L^*)$$

$(\hat{L}, \hat{C})$  is, by definition, a nice point.

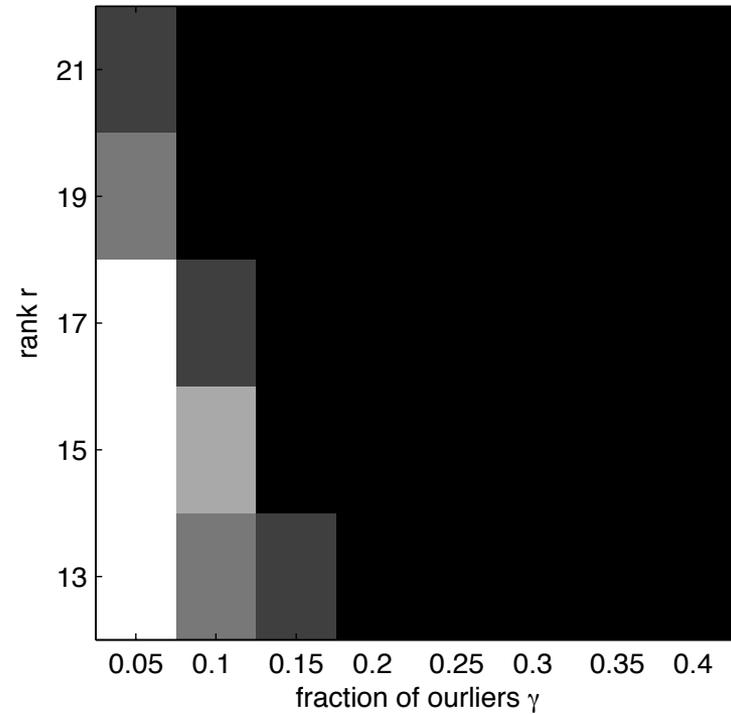
**Rest of proof:** showing it is the optimum of original program, under our assumption.

# Performance

---



L + C formulation



L + S formulation  
( from [Chandrasekaran et. al.],  
[Candes,et. al.] )

# Another view...

---

**Mean** is solution of

$$\min_x \sum_i (x_i - x)^2$$

**Fragile:** Can be easily skewed by one / few points

**Median** is solution of

$$\min_x \sum_i |x_i - x|$$

**Robust:** skewing requires Error in constant fraction of pts

**Standard PCA** of M is solution of

$$\sum_j \|M_j - L_j\|^2$$

$$\text{rank}(L) \leq r$$

**Our method** is (convex rel. of)

$$\sum_j \|M_j - L_j\|$$

$$\text{rank}(L) \leq r$$

# Collaborative Filtering w/ Adversaries

Address: <http://www.amazon.com/exec/obidos/ASIN/1590520556/102-0262843-0526515>

Amazon Exclusive!! Order a Segway now! It's only at Amazon

amazon.com

VIEW CART | WISH LIST | YOUR ACCOUNT | HELP

WELCOME YOUR STORE BOOKS APPAREL & ACCESSORIES ELECTRONICS TOYS & GAMES KITCHEN & HOUSEWARES SOFTWARE SEE MORE STORES Your Gold

SEARCH BROWSE SUBJECTS BESTSELLERS MAGAZINES CORPORATE ACCOUNTS E-BOOKS & DOCS BARGAIN BOOKS USED BOOKS

Now getting jeans is as simple as shopping at Amazon!

Introducing Apparel at Amazon.com. 400+ Brands • One Cart • Shop the Amazon way

**SEARCH**

Books

**BOOK INFORMATION**

[buying info](#)  
[editorial reviews](#)  
[table of contents](#)

**RATE THIS ITEM**

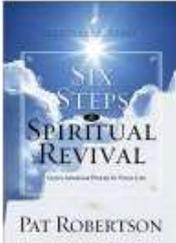
I dislike it I love it!

1 2 3 4 5

[Edit your ratings](#)

**Favorite Magazines!**

**Six Steps to Spiritual Revival**  
by [Pat Robertson](#)



List Price: \$9.99  
Price: **\$9.99** & eligible for **FREE Super Saver Shipping** on orders over \$25. [See details.](#)

Availability: Usually ships in 24 hours

[Used & new](#) from \$7.09

Edition: Hardcover

[see larger photo](#)

[See more product details](#)

**READY TO BUY?**

**or**  
[Sign in](#) to turn on 1-Click ordering.

**MORE BUYING CHOICES**

[Used & new](#) from \$7.09

Have one to sell? [Sell your h](#)

[Win Your Wishes](#)

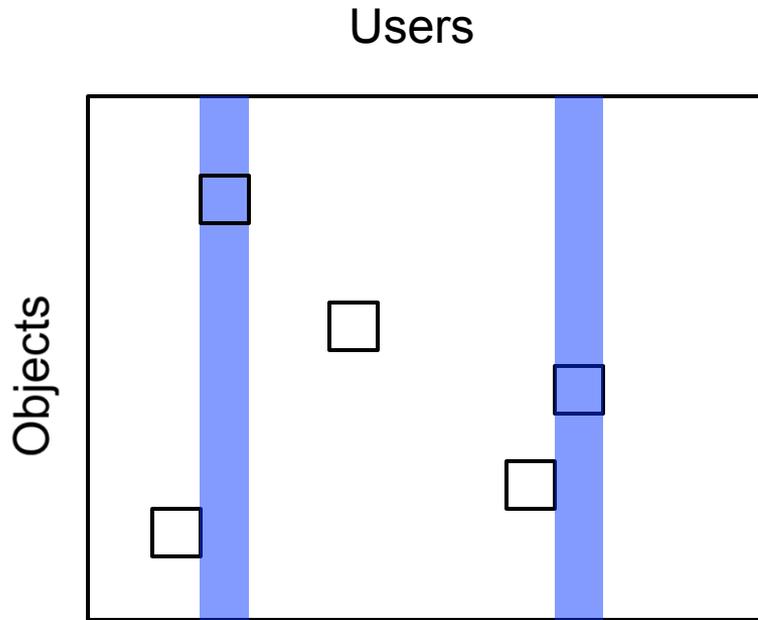
Add any item to your Wish List and you could **win** **up to \$350** of your select items. [Learn more.](#)  
Don't have one? We'll set one up for you.

**Customers who shopped for this item also shopped for these items:**

- [The End of the Age](#) by Pat Robertson
- [The Ultimate Guide to Sex for Men](#) by Bill Brent
- [Esther's Gift](#) by Jan Karon

# Collaborative Filtering w/ Adversaries

---



Low-rank matrix that

- Is partially observed
- Has some corrupted columns

**== outliers with missing data !**

## Our setting:

- Good users == random sampling of incoherent matrix (as in matrix completion)
- Manipulators == completely arbitrary sampling, values

# Outlier Pursuit with Missing Data

---

$$\begin{aligned} \min \quad & \|L\|_* + \gamma \|C\|_{1,2} \\ \text{s.t.} \quad & l_{ij} + c_{ij} = m_{ij} \quad \text{for observed } (i, j) \end{aligned}$$

Now: need **row space to be incoherent** as well

- since we are doing matrix completion and manipulator identification

# Our Result

## Theorem:

Convex program optimum  $(\hat{L}, \hat{C})$  is such that  $\hat{L}$  has the correct column space and the support of  $\hat{C}$  is exactly the set of manipulators, whp, provided  $n \geq p$

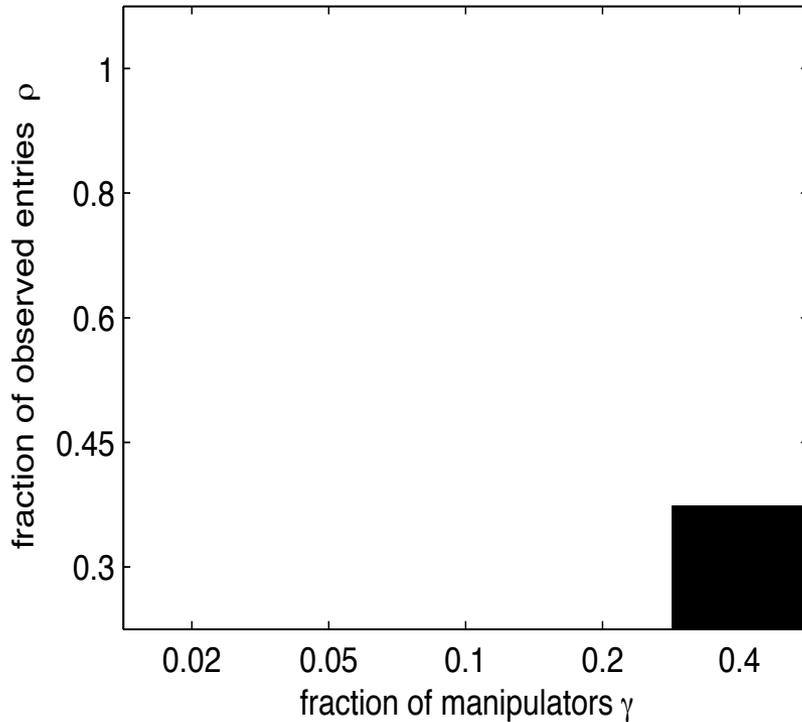
$$\text{Sampling density } \rho \geq c_1 \frac{\mu^2 r^2 \log^3(4n)}{p}$$

$$\text{Fraction of users that are manipulators } \frac{\eta}{1 - \eta} \leq c_2 \frac{\rho^2}{\left(1 + \frac{\mu r}{\rho \sqrt{p}}\right) \mu^2 r^2 \log^6(4n)}$$

*Note: no assumptions on manipulators*

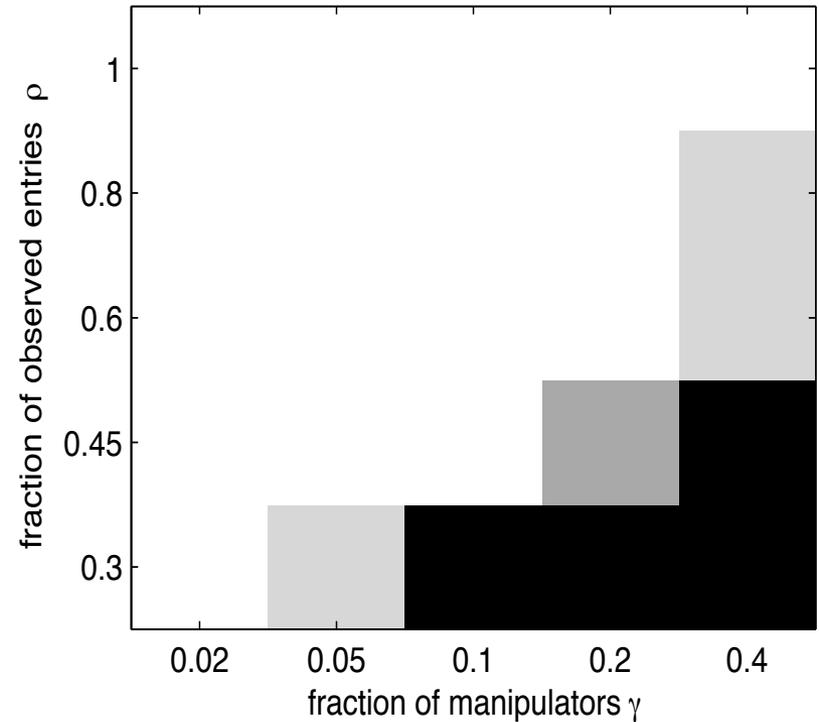
# Robust Collaborative Filtering

---



Algo: Partially observed

Low-rank + Column-sparse



Algo: Partially observed

Sparse + Low-rank

# More generally ...

---

Several methods in  
High-dim. Statistics

$$\min_X \mathcal{L}(y, \mathcal{A}; X) + \lambda r(X)$$

Loss function

regularizer

Our approach:

$$\min_{X_1, X_2} \mathcal{L}(y, \mathcal{A}; X_1 + X_2) + \lambda_1 r_1(X_1) + \lambda_2 r_2(X_2)$$

(same) Loss function

Weighted sum of regularizers

Yields robustness + flexibility in several settings.

**Today: PCA wit Outliers + missing data**

---

Latent factors in time series (**ICML'12**)

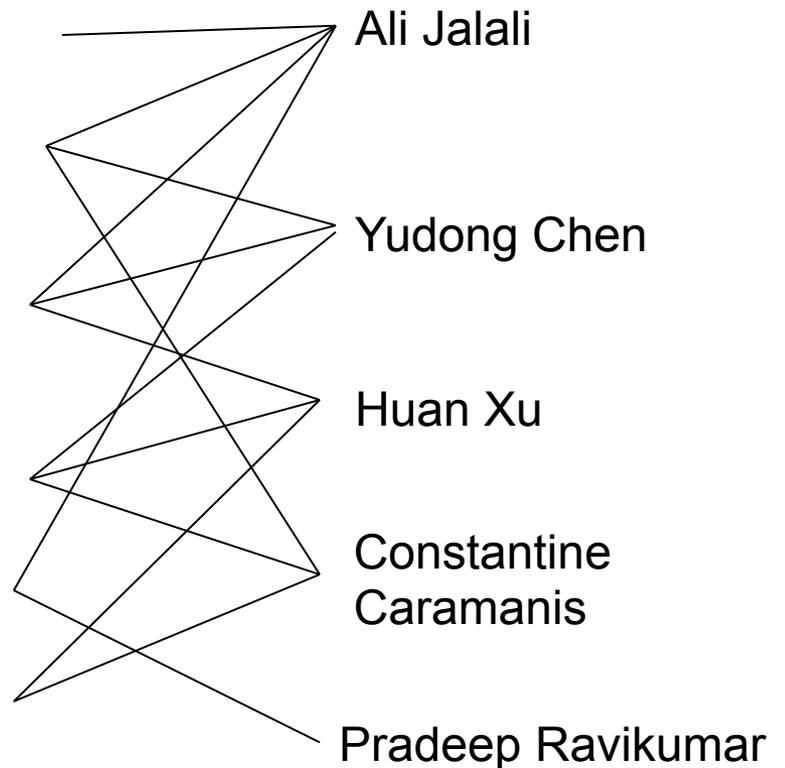
Matrix completion from Errors and Erasures  
(**ISIT 2011, SIAM J. Optim. 2011**)

Graph clustering (**ICML 2011**)

Robust Recommender Systems  
(**ICML 2011**)

Multiple Sparse Regression (**NIPS 2010**)

PCA that is robust to Outliers  
(**NIPS 2010, Trans IT**)



**All papers on my website, Arxiv.**