# 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

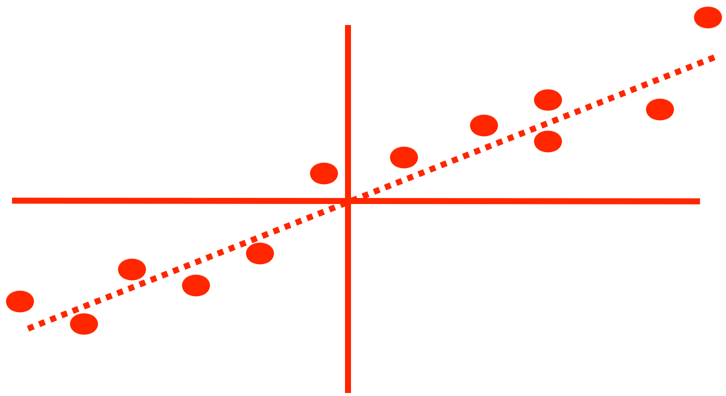
**Our idea + Algorithms**

**Results**

- Full observation
- Missing Data

**Framework** for Robustness in High Dimensions

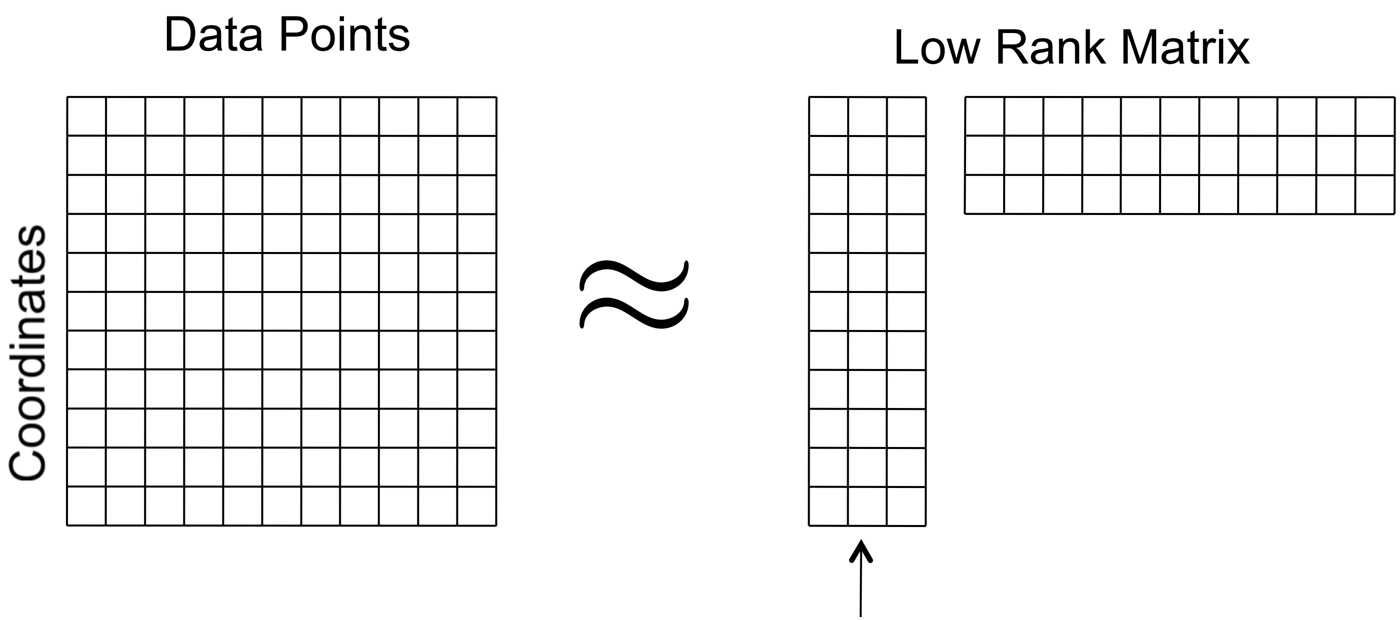
# Principal Components Analysis



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

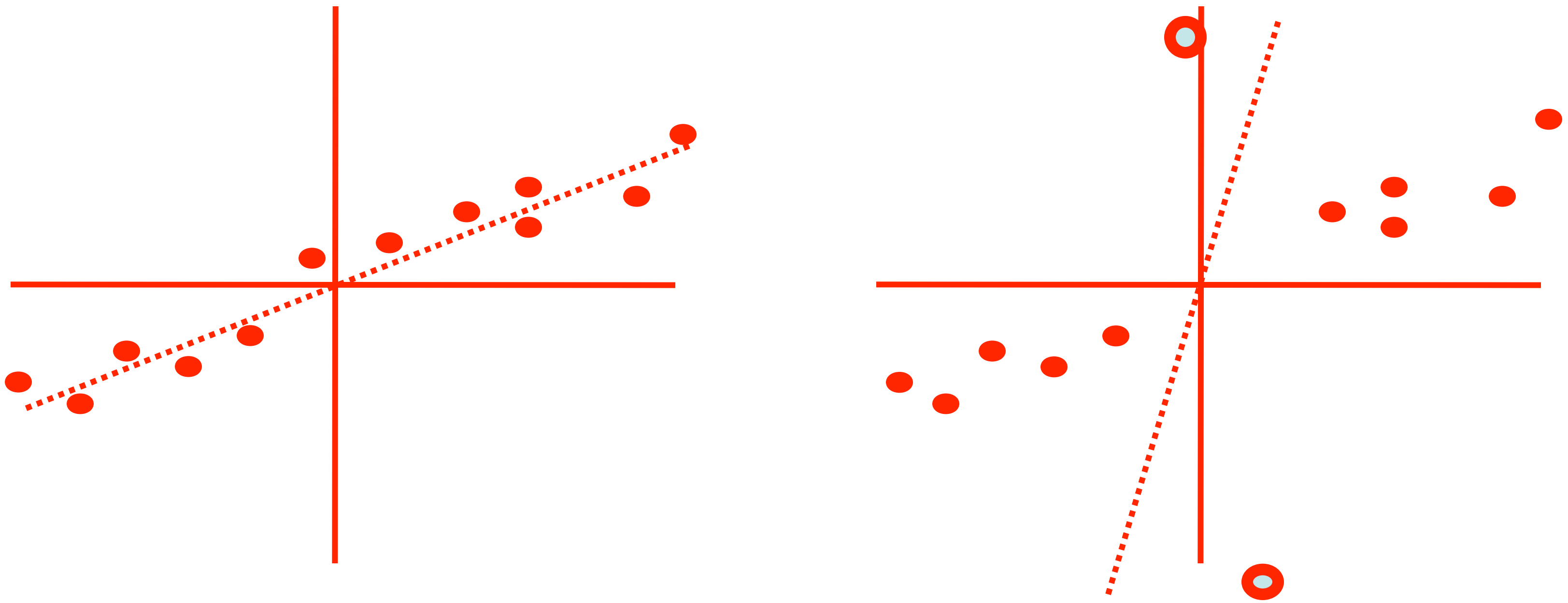
**Classical technique:**

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



# Fragility

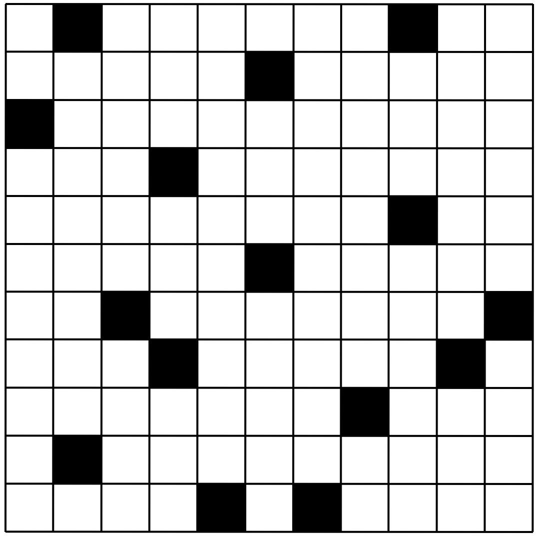
Gross errors of even one/few points can completely throw off PCA



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

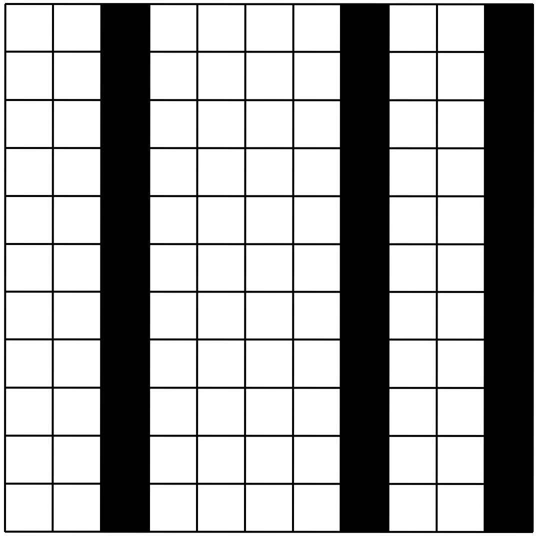
# Two types of gross errors

**Corrupted Features**



**- Individual entries corrupted**

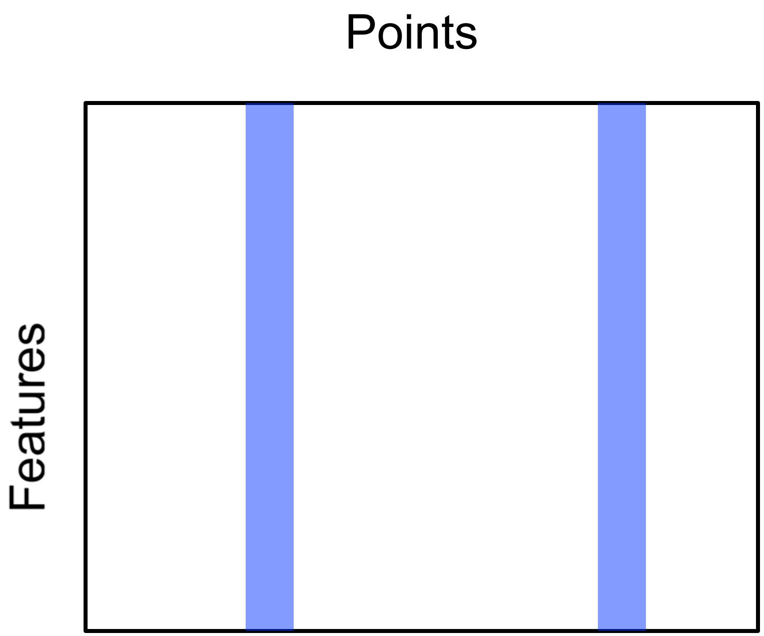
**Outliers**

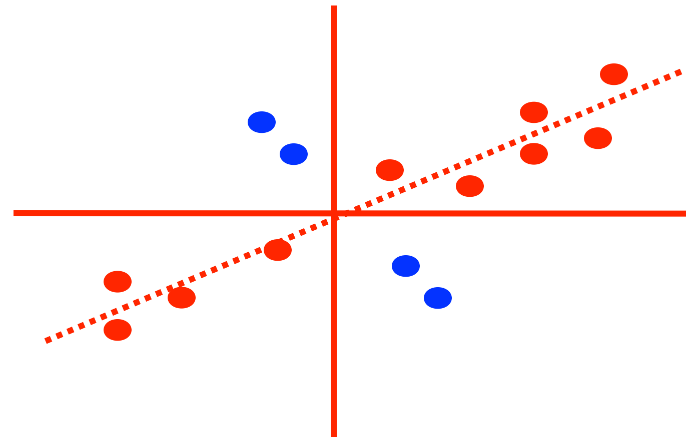


**-Entire columns corrupted**

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

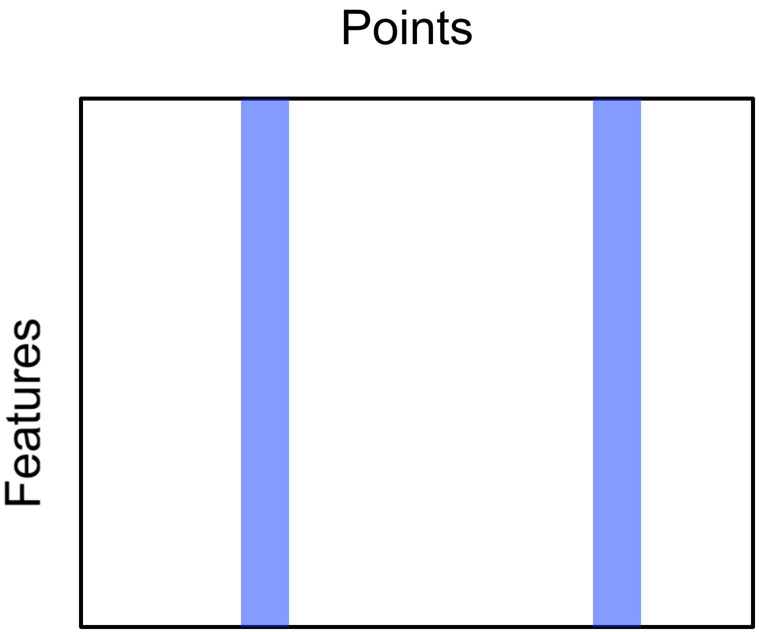
# PCA with Outliers

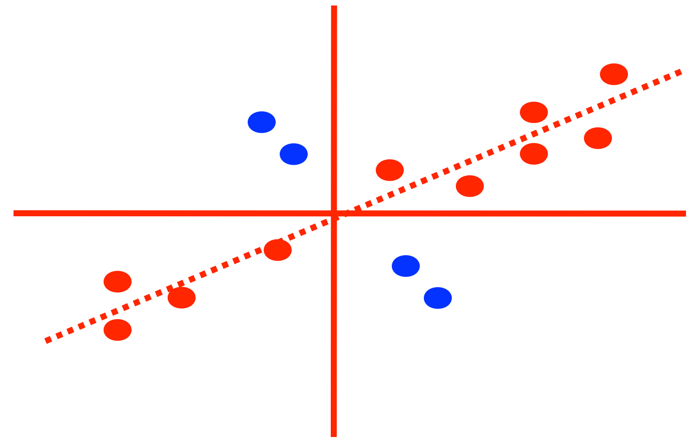


**Objective: find identities of outliers**
(and hence col. space of true matrix)

# Outlier Pursuit - Idea



Standard PCA

$$\min_{L} \quad \|M - L\|_F$$

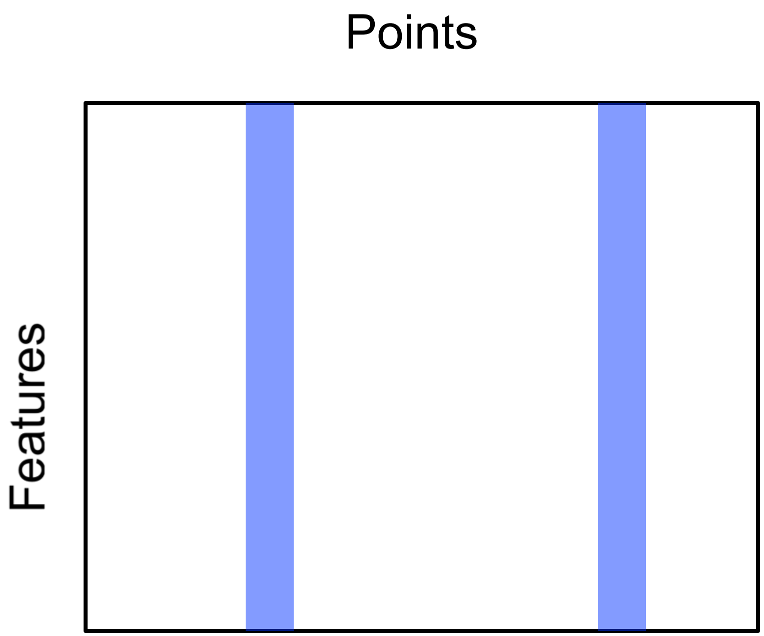
$$s.t. \quad rank(L) = r$$

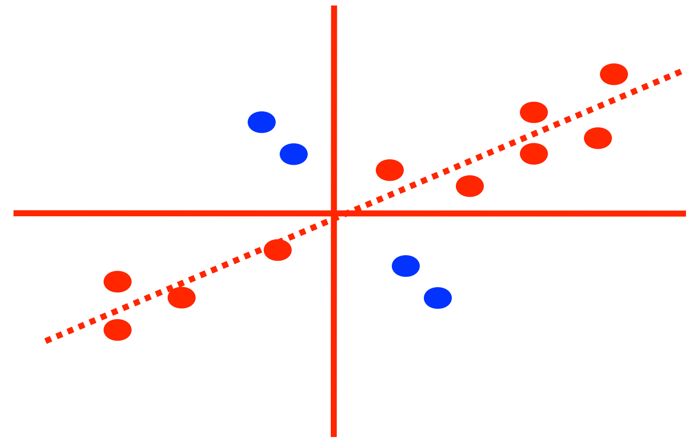
$$\min_{L,C} \quad \|M - L - C\|_F$$

$$s.t. \quad rank(L) = r$$

$$col(C) = c$$

# Outlier Pursuit - Method



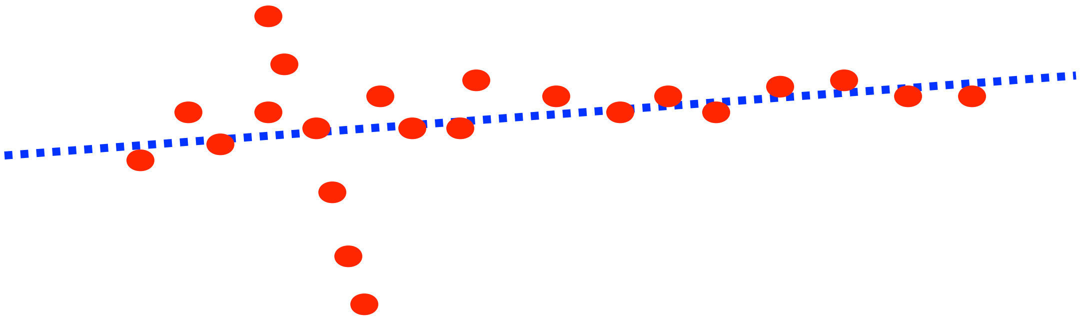



We propose: $$\min_{L,C} \quad \|M - L - C\|_F + \lambda_1 \|L\|_* + \lambda_2 \|C\|_{1,2}$$

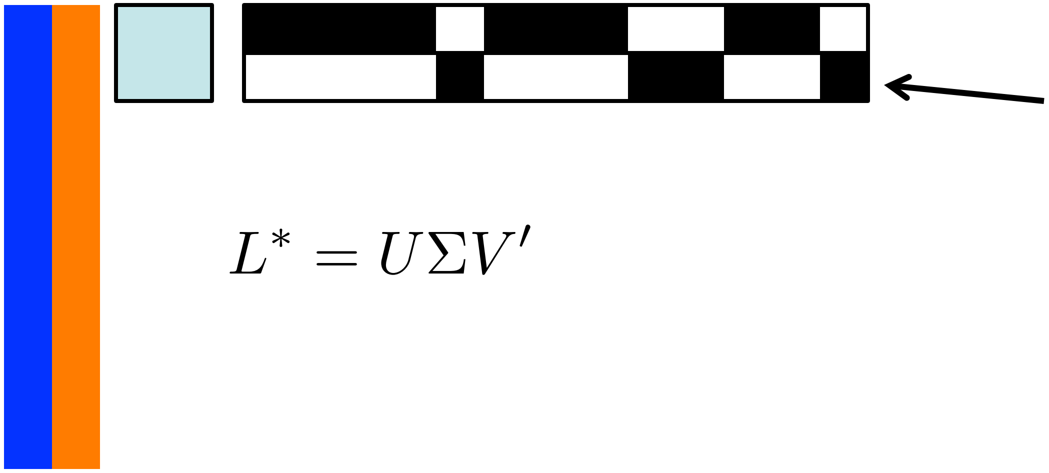
Convex surrogate for
Rank constraint

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

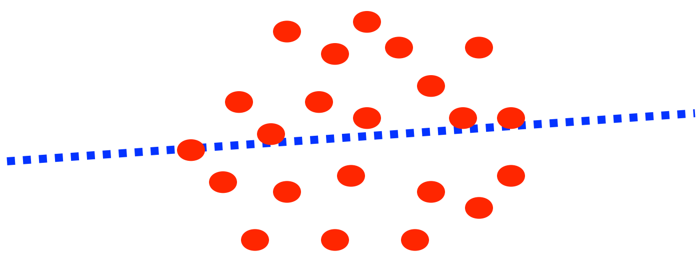
$$L^* = U\Sigma V'$$

$$\max_i \|V' e_i\|$$ 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**

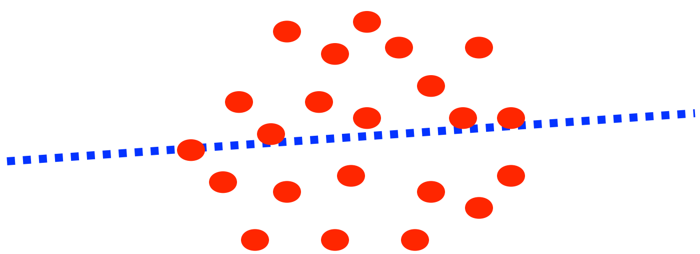
$$\min_{L,C} \quad \|L\|_* + \lambda \|C\|_{1,2}$$

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

# 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)

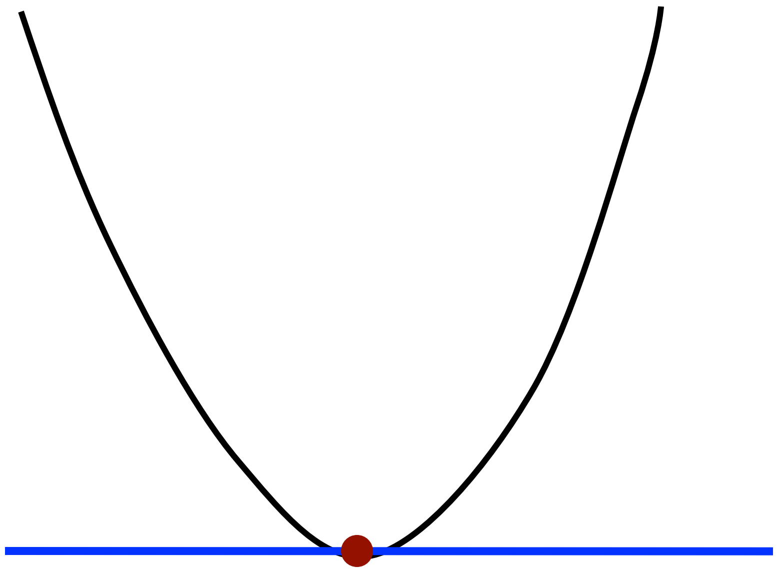
Our convex program can identify upto a fraction $\gamma$ of outliers as long as

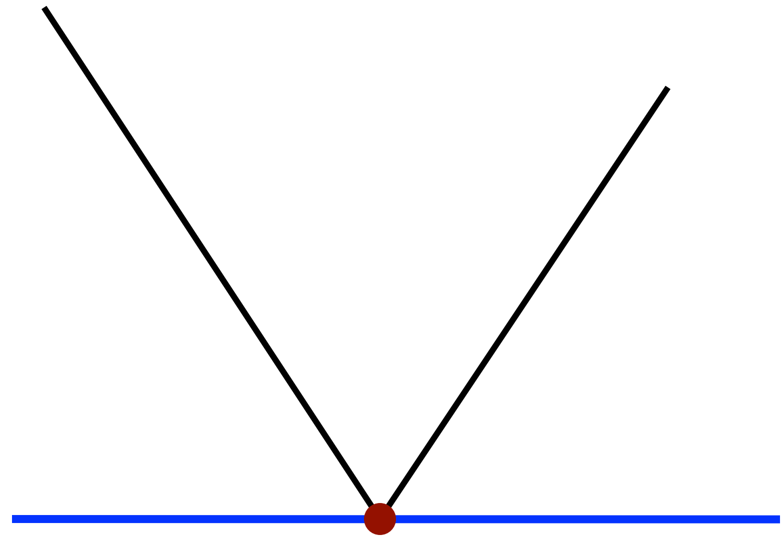
$$\frac{\gamma}{1-\gamma} \leq \frac{c}{\mu r}$$

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

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

# Proof Technique





A point $x$ is the optimum of a convex function $f$ $\Leftrightarrow$ 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:**

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

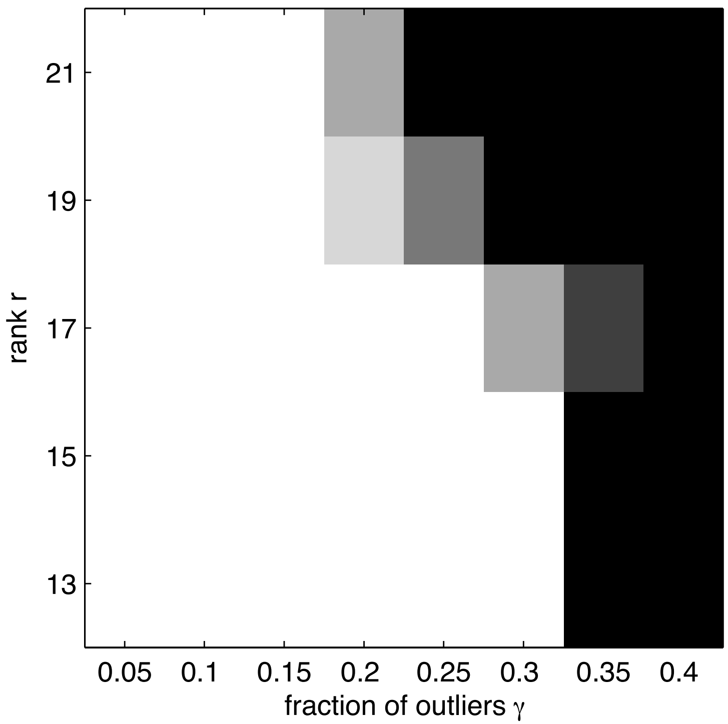
$$s.t. \;\; ColSupp(C) \subset ColSupp(C^*)$$

$$ColSpace(L) \subset ColSpace(L^*)$$

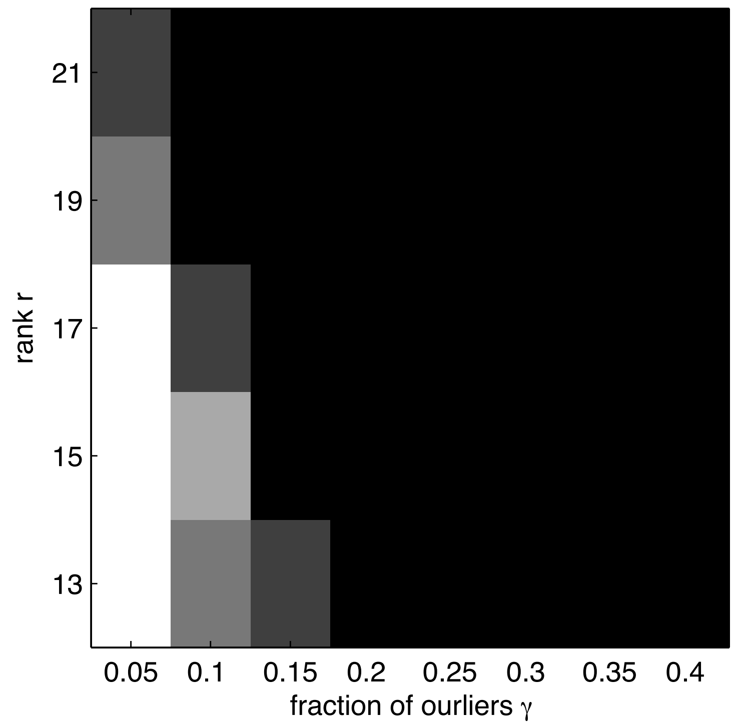
$(\widehat{L}, \widehat{C})$ is, by definition, a nice point.

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

**Standard PCA** of M is solution of

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

$$rank(L) \leq r$$

**Median** is solution of

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

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

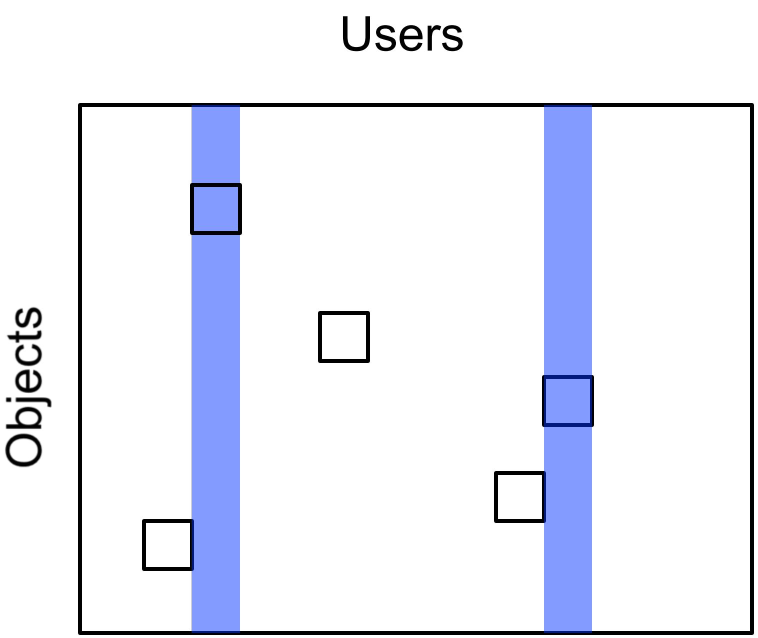
**Our method** is (convex rel. of)

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

$$rank(L) \leq r$$

# Collaborative Filtering w/ Adversaries

# Collaborative Filtering w/ Adversaries




Low-rank matrix that

- Is partiallly 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

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

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

- since we are doing matrix completion and manipulator identification

# Our Result

**Theorem:**

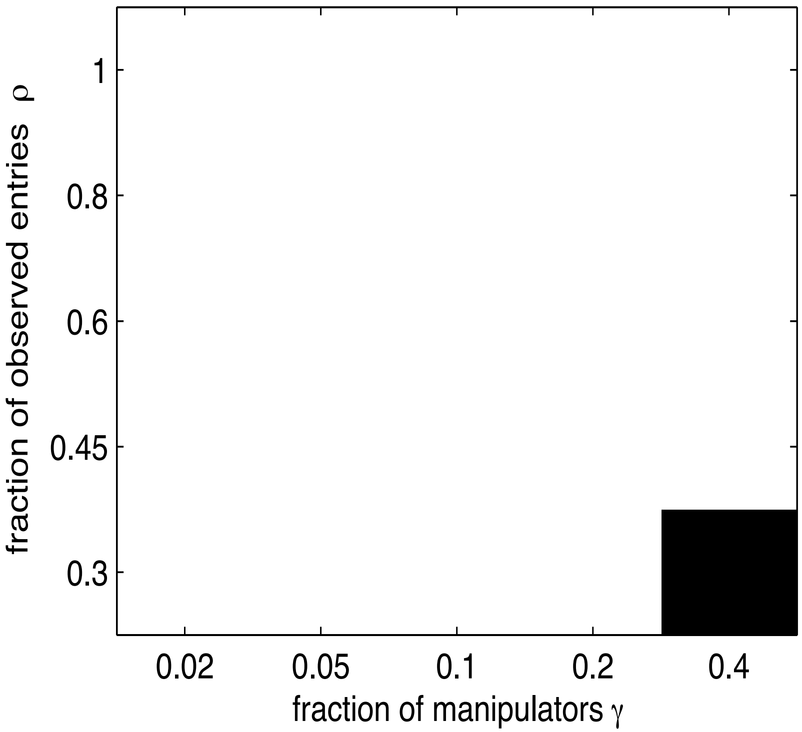
Convex program optimum $(\widehat{L}, \widehat{C})$ is such that $\widehat{L}$ has the correct column space and the support of $\widehat{C}$ is exactly the set of manipulators, whp, provided $n \geq p$

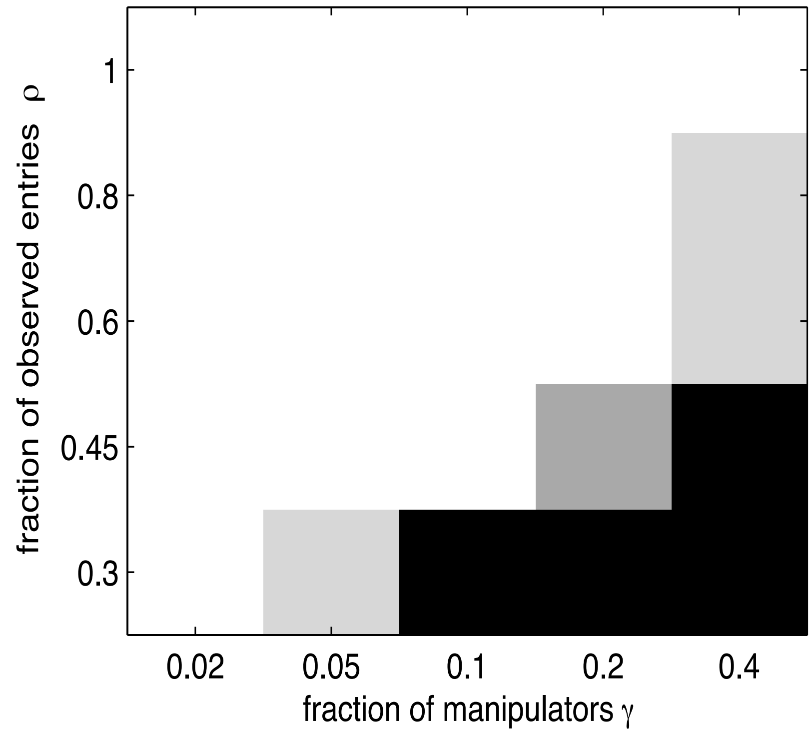
Sampling density $$\rho \geq c_1 \frac{\mu^2 r^2 \log^3(4n)}{p}$$

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

*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

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

Loss function regularizer

Our approach:

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

(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)**

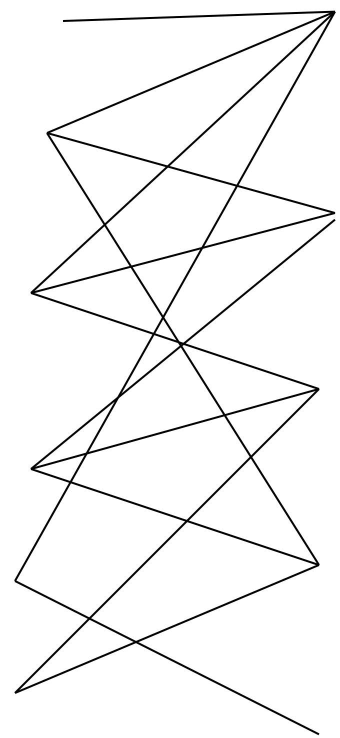
Matrix completion from Errors and Erasures
**(ISIT 2011, SIAM J. Optim. 2011)**

Graph clustering **(ICML 2011)**

Robust Recommender Systems
**(ICML 2011)**

Multiple Sparse Regression **(NIPS 2010)**

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



Ali Jalali

Yudong Chen

Huan Xu

Constantine Caramanis

Pradeep Ravikumar

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