

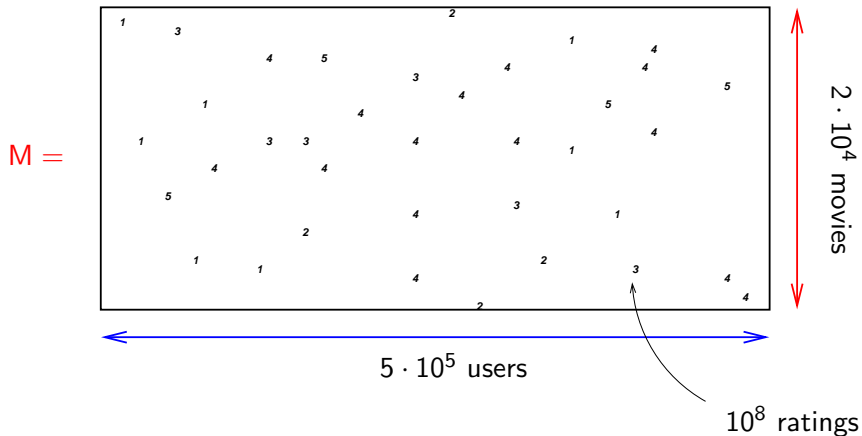
Large Matrices Beyond Singular Value Decomposition

Andrea Montanari

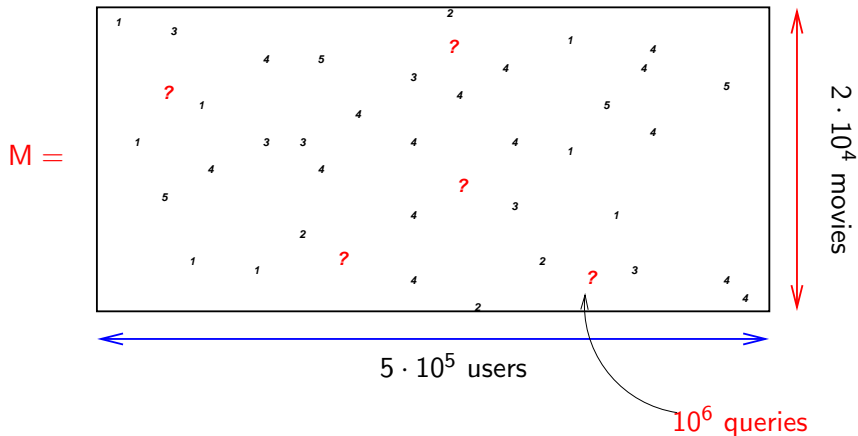
with Raghunandan Keshavan and Sewoong Oh
Stanford University

March 22, 2010

A motivating example: The Netflix challenge



A motivating example: The Netflix challenge



A prize awarded for:

$$\text{RMSE} < 0.8563 \quad ; -)$$

Can we make sense of this?

A prize awarded for:

RMSE < 0.8563 ; -)

Can we make sense of this?

Outline

- 1 The model
- 2 Background
- 3 Algorithm and main theorems
- 4 Numerical simulations
- 5 Further directions

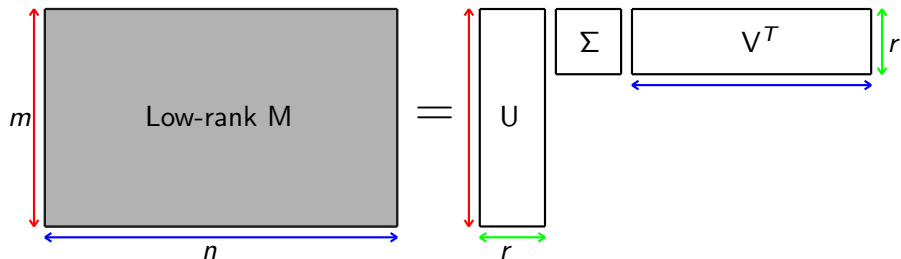
arXiv:0901.3150

arXiv:0906.2027

The model

We need some structure!

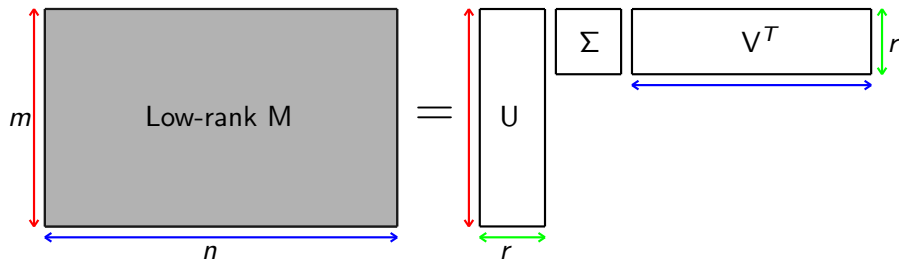
We need some structure!



1. Low-rank matrix M
2. $N = M + Z$
3. Uniformly random sample E

$$N_{ij}^E = \begin{cases} M_{ij} + Z_{ij} & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

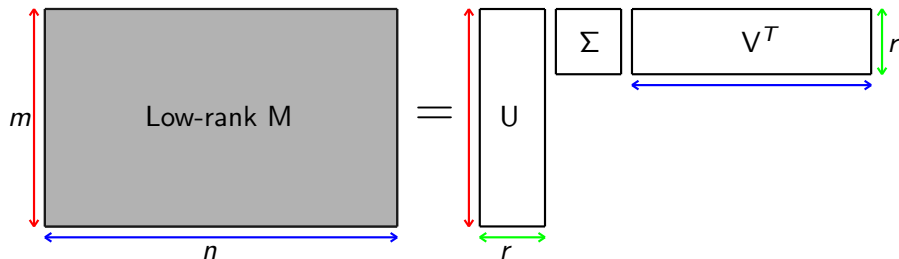
We need some structure!



1. Low-rank matrix M
2. $N = M + Z$
3. Uniformly random sample E

$$N_{ij}^E = \begin{cases} M_{ij} + Z_{ij} & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

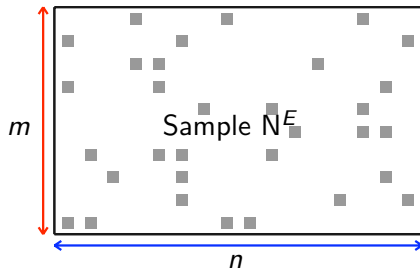
We need some structure!



1. Low-rank matrix M
2. $N = M + Z$
3. Uniformly random sample E

$$N_{ij}^E = \begin{cases} M_{ij} + Z_{ij} & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

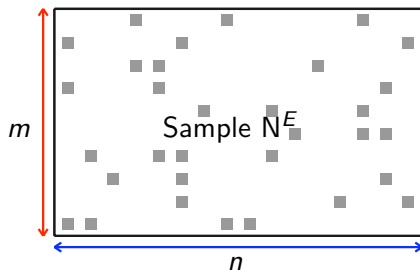
We need some structure!



1. Low-rank matrix M
2. $N = M + Z$
3. Uniformly random sample E

$$N_{ij}^E = \begin{cases} M_{ij} + Z_{ij} & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

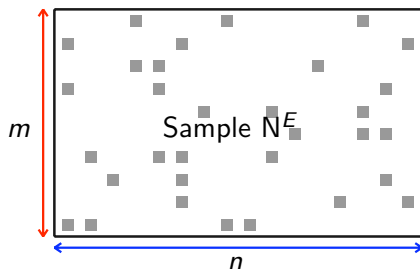
We need some structure!



Goal : Estimation $\hat{M}(E, N^E)$ that minimizes

$$\text{RMSE} \equiv \left(\frac{1}{mn} \sum_{i,j} (M_{ij} - \hat{M}_{ij})^2 \right)^{1/2} .$$

We need some structure!



Goal : Estimation $\hat{M}(E, N^E)$ that minimizes

$$\text{RMSE} \equiv \frac{1}{\sqrt{mn}} \|M - \hat{M}\|_F .$$

Problem Parameters

- Data size $m \times n$, $\alpha \equiv m/n$
- Rank r
- Sample size $|E|$
- Noise Z^E { Running example : $Z_{ij} \sim \text{i.i.d. } N(0, \sigma_z^2)$ }

Background

Unstructured factors

CR1. Incoherence 1

$$\left| \sum_{k=1}^r U_{ik} V_{ak} \right| \leq \mu_0 \sqrt{r}.$$

CR2. Incoherence 2

$$\sum_{k=1}^r U_{ik}^2 \leq \mu_1 r, \quad \sum_{k=1}^r V_{ak}^2 \leq \mu_1 r.$$

[Candés, Recht 2008]

Noiseless case

Theorem (Candés, Recht, 2008)

If M is *incoherent*, and

$$|E| \geq C r n^{6/5} \log n$$

then whp

1. M is unique given the observed entries.
2. M is the unique minimum of a SDP.

cf. also [Recht, Fazel, Parrilo 2007]

Noiseless case

Theorem (Candés, Recht, 2008)

If M is *incoherent*, and

$$|E| \geq C r n^{6/5} \log n$$

then whp

1. M is unique given the observed entries.
2. M is the unique minimum of a SDP.

cf. also [Recht, Fazel, Parrilo 2007]

Noiseless case

Theorem (Candés, Recht, 2008)

If M is *incoherent*, and

$$|E| \geq C r n^{6/5} \log n$$

then whp

1. M is unique given the observed entries.
2. M is the unique minimum of a SDP.

cf. also [Recht, Fazel, Parrilo 2007]

Noiseless case

Theorem (Candés, Recht, 2008)

If M is *incoherent*, and

$$|E| \geq C r n^{6/5} \log n$$

then whp

1. M is unique given the observed entries.
2. M is the unique minimum of a SDP.

cf. also [Recht, Fazel, Parrilo 2007]

Noiseless case

Theorem (Candés, Recht, 2008)

If M is *incoherent*, and

$$|E| \geq C r n^{6/5} \log n$$

then whp

1. M is unique given the observed entries.
2. M is the unique minimum of a SDP.

cf. also [Recht, Fazel, Parrilo 2007]

Great, but...

1. $n^{1/5}$ observations for 1 bit of information?
2. RMSE = 0?
3. SDP = $O(n^{4...6})$. Substitute $n = 10^5$...

Great, but...

1. $n^{1/5}$ observations for 1 bit of information?

2. RMSE = 0?

3. SDP = $O(n^{4...6})$. Substitute $n = 10^5$...

Great, but...

1. $n^{1/5}$ observations for 1 bit of information?
2. RMSE = 0?
3. SDP = $O(n^{4...6})$. Substitute $n = 10^5$...

Great, but...

1. $n^{1/5}$ observations for 1 bit of information?
2. RMSE = 0?
3. SDP = $O(n^{4...6})$. Substitute $n = 10^5$...

A movie

Algorithm and main theorems

Naïve Approach

$$N^E = \sum_{k=1}^n x_k \sigma_k y_k^T$$

Rank- r projection :

$$\mathcal{P}_r(N^E) \equiv \frac{mn}{|E|} \sum_{k=1}^r x_k \sigma_k y_k^T$$

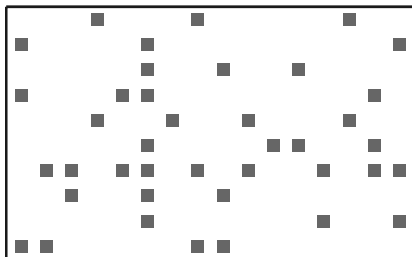
Naïve Approach Fails

- Define : $\text{deg}(\text{row}_i) \equiv \#$ of samples in row i .
- For $|E| = O(n)$, there exists a row with degree $\Omega(\log n / (\log \log n))$.
- *spurious* singular values of $\Omega(\sqrt{\log n / (\log \log n)})$.

Trimming

- Solution : Trimming

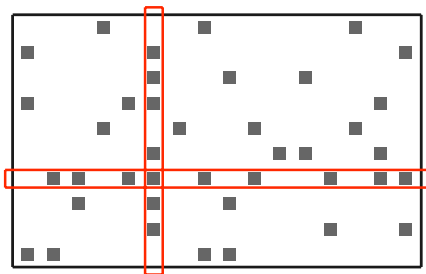
$$\tilde{N}_{ij}^E = \begin{cases} 0 & \text{if } \text{deg}(\text{row}_i) > 2\mathbb{E}[\text{deg}(\text{row}_i)] , \\ 0 & \text{if } \text{deg}(\text{col}_j) > 2\mathbb{E}[\text{deg}(\text{col}_j)] , \\ N_{ij}^E & \text{otherwise.} \end{cases}$$



Trimming

- Solution : Trimming

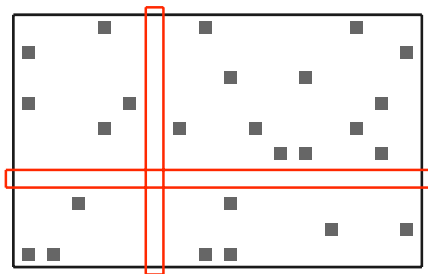
$$\tilde{N}_{ij}^E = \begin{cases} 0 & \text{if } \text{deg}(\text{row}_i) > 2\mathbb{E}[\text{deg}(\text{row}_i)] , \\ 0 & \text{if } \text{deg}(\text{col}_j) > 2\mathbb{E}[\text{deg}(\text{col}_j)] , \\ N_{ij}^E & \text{otherwise.} \end{cases}$$



Trimming

- Solution : Trimming

$$\tilde{N}_{ij}^E = \begin{cases} 0 & \text{if } \text{deg}(\text{row}_i) > 2\mathbb{E}[\text{deg}(\text{row}_i)] , \\ 0 & \text{if } \text{deg}(\text{col}_j) > 2\mathbb{E}[\text{deg}(\text{col}_j)] , \\ N_{ij}^E & \text{otherwise.} \end{cases}$$



The Algorithm

OPTSPACE

Input : sample positions E , sample values N^E , rank r

Output : estimation \hat{M}

- 1: Trim N^E , and let \tilde{N}^E be the output;
 - 2: Compute rank- r projection $\mathcal{P}_r(\tilde{N}^E) = X_0 S_0 Y_0^T$;
 - 3:
-

Main Result

Theorem (Keshavan, Montanari, Oh, 2009)

Assume $|M_{ij}| \leq M_{\max}$. Then, w.h.p., rank- r projection achieves

$$\frac{1}{n} \|M - \mathcal{P}_r(\tilde{N}^E)\|_F = \text{RMSE} \leq CM_{\max} \sqrt{\frac{nr}{|E|}} + C' \frac{n\sqrt{r}}{|E|} \|Z^E\|_2.$$

$$\left(\text{Example: } CM_{\max} \sqrt{\frac{nr}{|E|}} + C' \sigma_z \sqrt{\frac{rn \log n}{|E|}} \right)$$

A comparison

Theorem (Achlioptas, McSherry 2007)

Assume $|E| \geq (8 \log n)^4 n$ and bounded entries. Then

$$\frac{1}{nM_{\max}} \|\mathbf{M} - \mathcal{P}_r(\tilde{\mathbf{M}}^E)\|_F = \text{RMSE} \leq 4\sqrt{r/\epsilon}.$$

with probability larger than $1 - \exp(-19(\log n)^4)$.

(For $n = 10^6$, $(8 \log n)^4 \approx 1.5 \cdot 10^8$)

A comparison

Theorem (Achlioptas, McSherry 2007)

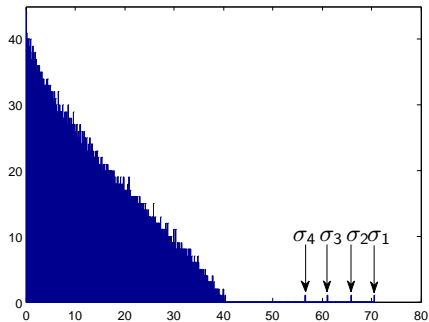
Assume $|E| \geq (8 \log n)^4 n$ and bounded entries. Then

$$\frac{1}{nM_{\max}} \|\mathbf{M} - \mathcal{P}_r(\tilde{\mathbf{M}}^E)\|_F = \text{RMSE} \leq 4\sqrt{r/\epsilon}.$$

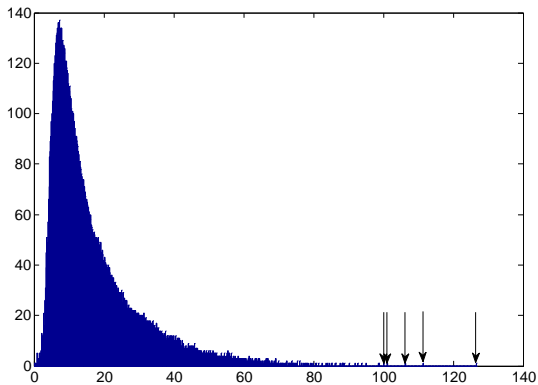
with probability larger than $1 - \exp(-19(\log n)^4)$.

(For $n = 10^6$, $(8 \log n)^4 \approx 1.5 \cdot 10^8$)

Random matrix $r = 4$, $n = 10000$, $\epsilon = 12.5$



Netflix data (trimmed)



The Algorithm

OPTSPACE

Input : sample positions E , sample values N^E , rank r

Output : estimation \hat{M}

- 1: Trim N^E , and let \tilde{N}^E be the output;
 - 2: Compute rank- r projection $\mathcal{P}_r(\tilde{N}^E) = X_0 S_0 Y_0^T$;
 - 3: Minimize RMSE by manifold gradient descent starting at (X_0, S_0, Y_0) .
-

Minimizing RMSE: Naive objective function

$$F_{\text{naive}}(X, Y) \equiv \frac{1}{2} \sum_{(i,j) \in E} \left| M_{ij} - (XY^T)_{ij} \right|^2.$$

$$X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{m \times r}$$

Lots of flat directions!!

Minimizing RMSE: Naive objective function

$$F_{\text{naive}}(X, Y) \equiv \frac{1}{2} \sum_{(i,j) \in E} \left| M_{ij} - (XY^T)_{ij} \right|^2.$$

$$X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{m \times r}$$

Lots of flat directions!!

Minimizing RMSE: Manifold gradient descent (1)

$$\mathcal{F}(X, Y, S) \equiv \frac{1}{2} \sum_{(i,j) \in E} \left| M_{ij} - (XSY^T)_{ij} \right|^2.$$

$$X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{m \times r} \text{ with } X^T X = m\mathbf{1}, Y^T Y = n\mathbf{1}.$$

$$\mathcal{F} : \text{Ortho}(m, r) \times \text{Ortho}(n, r) \times \mathbb{R}^{r \times r} \rightarrow \mathbb{R}$$

$$\text{Ortho}(m, r) \equiv \{X \in \mathbb{R}^{m \times r} : X^T X = m\mathbf{1}\}.$$

Minimizing RMSE: Manifold gradient descent (1)

$$\mathcal{F}(X, Y, S) \equiv \frac{1}{2} \sum_{(i,j) \in E} \left| M_{ij} - (XSY^T)_{ij} \right|^2.$$

$$X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{m \times r} \text{ with } X^T X = m\mathbf{1}, Y^T Y = n\mathbf{1}.$$

$$\mathcal{F} : \text{Ortho}(m, r) \times \text{Ortho}(n, r) \times \mathbb{R}^{r \times r} \rightarrow \mathbb{R}$$

$$\text{Ortho}(m, r) \equiv \{X \in \mathbb{R}^{m \times r} : X^T X = m\mathbf{1}\}.$$

Minimizing RMSE: Manifold gradient descent (2)

$$F(X, Y) \equiv \min_{S \in \mathbb{R}^{r \times r}} \mathcal{F}(X, Y, S),$$
$$\mathcal{F}(X, Y, S) \equiv \frac{1}{2} \sum_{(i,j) \in E} \left| M_{ij} - (XSY^T)_{ij} \right|^2.$$

$$F : \text{Grassmann}(m, r) \times \text{Grassmann}(n, r) \rightarrow \mathbb{R}$$

$\text{Grassmann}(m, r) \equiv \{\text{subspaces of dimension } r \text{ of } \mathbb{R}^m\}.$

Minimizing RMSE: Manifold gradient descent (2)

$$F(X, Y) \equiv \min_{S \in \mathbb{R}^{r \times r}} \mathcal{F}(X, Y, S),$$
$$\mathcal{F}(X, Y, S) \equiv \frac{1}{2} \sum_{(i,j) \in E} \left| M_{ij} - (XSY^T)_{ij} \right|^2.$$

$$F : \text{Grassmann}(m, r) \times \text{Grassmann}(n, r) \rightarrow \mathbb{R}$$

$\text{Grassmann}(m, r) \equiv \{\text{subspaces of dimension } r \text{ of } \mathbb{R}^m\}.$

Main Result

Theorem (Keshavan, Montanari, Oh, 2009)

Let M be an $n \times m$ matrix of rank- r bounded by M_{\max} . Then, w.h.p., rank- r projection achieves

$$\text{RMSE} \leq C M_{\max} \sqrt{nr/|E|} + C' \|Z^E\|_2 n\sqrt{r}/|E|.$$

Theorem (Keshavan, Montanari, Oh, 2009)

Let M be an $n \times n$ rank- r *incoherent* matrix with $\sigma_1(M)/\sigma_r(M) = O(1)$. If $|E| \geq Cnr \max\{r, \log n\}$, then, w.h.p., OPTSPACE achieves

$$\text{RMSE} \leq C'' \frac{n\sqrt{r}}{|E|} \|Z^E\|_2,$$

provided that the RHS is smaller than $\sigma_r(M)$.

$$\left(\text{Example: } C'' \sigma_z \sqrt{rn \log n / |E|} \right)$$

Two surprises

Can do **much better** than SVD ! ($\mathcal{P}_E(A) \equiv A^E$)

$$\text{minimize } \|\mathcal{P}_E(N - XSY^T)\|_F^2$$

vs

$$\text{minimize } \|\mathcal{P}_E(N) - XSY^T\|_F^2$$

Error = Noise / Sampling factor

Two surprises

Can do **much better** than SVD ! ($\mathcal{P}_E(A) \equiv A^E$)

$$\text{minimize } \|\mathcal{P}_E(N - XSY^T)\|_F^2$$

vs

$$\text{minimize } \|\mathcal{P}_E(N) - XSY^T\|_F^2$$

Error = Noise / Sampling factor

Two surprises

Can do **much better** than SVD ! ($\mathcal{P}_E(A) \equiv A^E$)

$$\text{minimize } \|\mathcal{P}_E(N - XSY^T)\|_F^2$$

vs

$$\text{minimize } \|\mathcal{P}_E(N) - XSY^T\|_F^2$$

Error = Noise / Sampling factor

Comparison: SEMIDEFINITE PROGRAMMING

[Fazel, 2006, Candés, Recht 2008, Candés, Tao, 2009]

Theorem (Candés, Plan, 2009)

Assume *strongly incoherent* matrix M . If $|E| \geq C r n (\log n)^6$ then SEMIDEFINITE PROGRAMMING achieves, w.h.p.,

$$\text{RMSE} \leq C' \sqrt{\frac{n}{|E|}} \|Z^E\|_F + C'' \frac{1}{n} \|Z^E\|_F.$$

$$\left(\text{Example: } C' \sigma_z \sqrt{n} + C'' \sigma_z \frac{\sqrt{|E|}}{n} \right)$$

[Gross, Liu, Flammia, Becker, Eisert; Recht, October 2009:
 $|E| \geq C r n (\log n)^2$]

Comparison: SEMIDEFINITE PROGRAMMING

[Fazel, 2006, Candés, Recht 2008, Candés, Tao, 2009]

Theorem (Candés, Plan, 2009)

Assume *strongly incoherent* matrix M . If $|E| \geq C r n (\log n)^6$ then SEMIDEFINITE PROGRAMMING achieves, w.h.p.,

$$\text{RMSE} \leq C' \sqrt{\frac{n}{|E|}} \|Z^E\|_F + C'' \frac{1}{n} \|Z^E\|_F.$$

$$\left(\text{Example: } C' \sigma_z \sqrt{n} + C'' \sigma_z \frac{\sqrt{|E|}}{n} \right)$$

[Gross, Liu, Flammia, Becker, Eisert; Recht, October 2009:
 $|E| \geq C r n (\log n)^2$]

Comparison: SEMIDEFINITE PROGRAMMING

[Fazel, 2006, Candés, Recht 2008, Candés, Tao, 2009]

Theorem (Candés, Plan, 2009)

Assume *strongly incoherent* matrix M . If $|E| \geq C r n (\log n)^6$ then SEMIDEFINITE PROGRAMMING achieves, w.h.p.,

$$\text{RMSE} \leq C' \sqrt{\frac{n}{|E|}} \|Z^E\|_F + C'' \frac{1}{n} \|Z^E\|_F .$$

$$\left(\text{Example: } C' \sigma_z \sqrt{n} + C'' \sigma_z \frac{\sqrt{|E|}}{n} \right)$$

[Gross, Liu, Flammia, Becker, Eisert; Recht, October 2009:
 $|E| \geq C r n (\log n)^2$]

Comparison

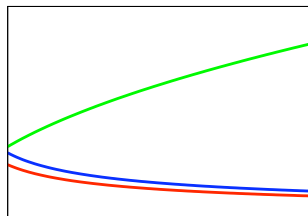
When Z is i.i.d $N(0, \sigma_z^2)$,

$$\text{ORACLE: RMSE} \simeq C \sigma_z \sqrt{\frac{r n}{|E|}}$$

$$\text{OPTSPACE: RMSE} \leq C' \sigma_z \sqrt{\frac{r n \log n}{|E|}}$$

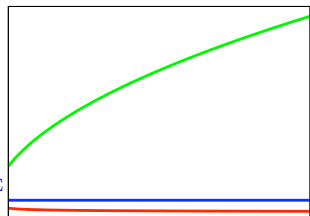
$$\text{SDP: RMSE} \leq C'' \sigma_z \left\{ \sqrt{n} + \frac{\sqrt{|E|}}{n} \right\}$$

RMSE



$|E|$

RMSE

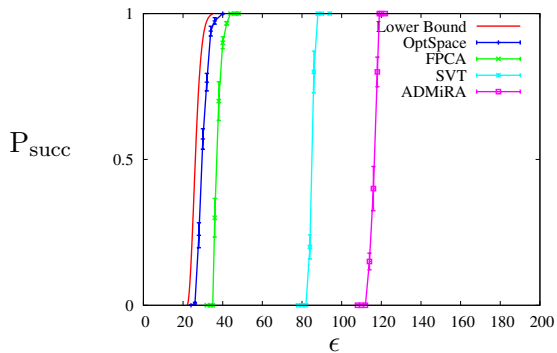


n

Numerical simulations

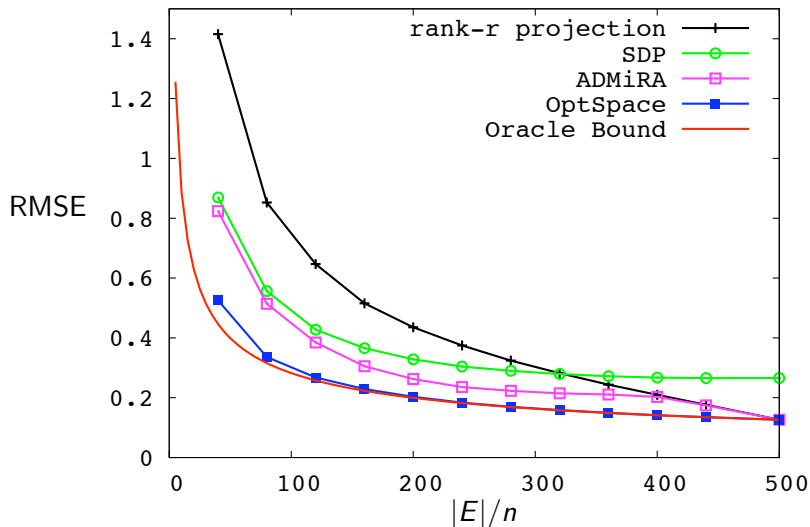
A noiseless example

- $m = n = 1000, r = 10$



A noisy example

- $n = 500, r = 4, \sigma_z = 1$, example from [Candés, Plan, 2009]



Further directions

Direction 1: Regularization

1. $N = M + Z$
2. Uniformly random sample E

$$N_{ij}^E = \begin{cases} M_{ij} + Z_{ij} & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Large Z \rightarrow Overfitting

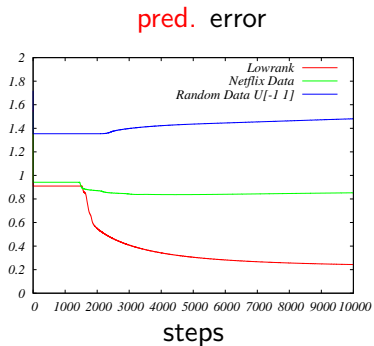
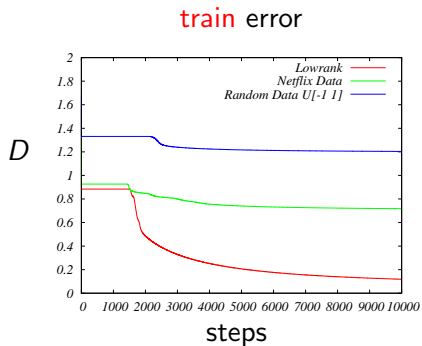
Direction 1: Regularization

1. $N = M + Z$
2. Uniformly random sample E

$$N_{ij}^E = \begin{cases} M_{ij} + Z_{ij} & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Large Z \rightarrow Overfitting

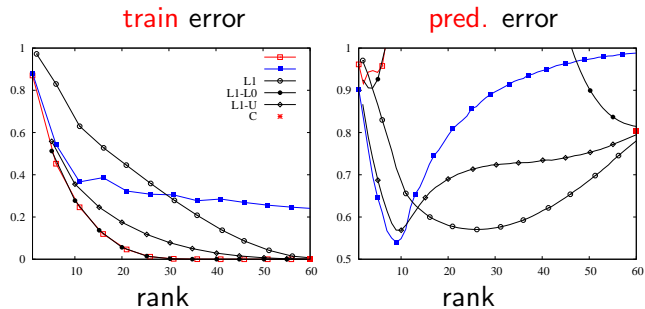
Example: Rank = 5



Regularization

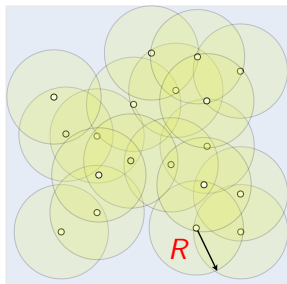
$$\mathcal{F}(X, Y, S) \equiv \frac{1}{2} \sum_{(i,j) \in E} \left| M_{ij} - (XSY^T)_{ij} \right|^2 + \frac{\lambda}{2} \sum_{a,b=1}^r S_{ab}^2.$$

$m = n = 100, r = 10, |E| = 0.5 \cdot mn, \text{SNR} = 1$



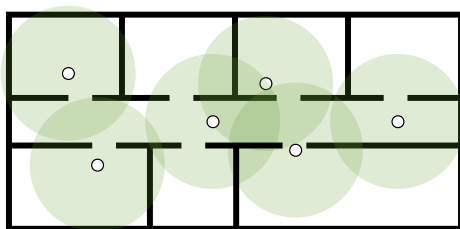
(cf. Mazumder, Hastie, Tibshirani, 2009)

Direction 2: Distributed positioning



Determine positions from pairwise distance measurements.

Direction 2: Distributed positioning



Determine positions from pairwise distance measurements.

What is the connection?

$$M_{ij} \equiv \|x_i - x_j\|^2$$

$$\text{rank}(M) = 4$$

What is the connection?

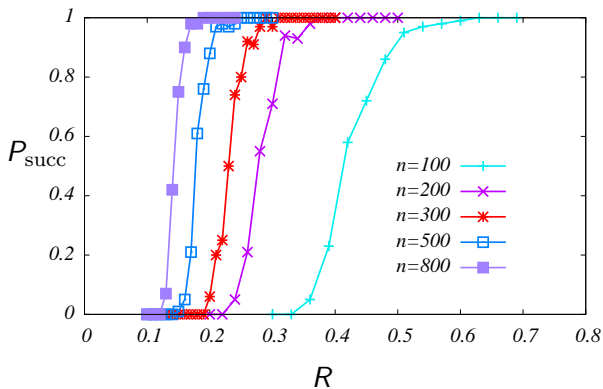
$$M_{ij} \equiv \|x_i - x_j\|^2$$

$$\text{rank}(M) = 4$$

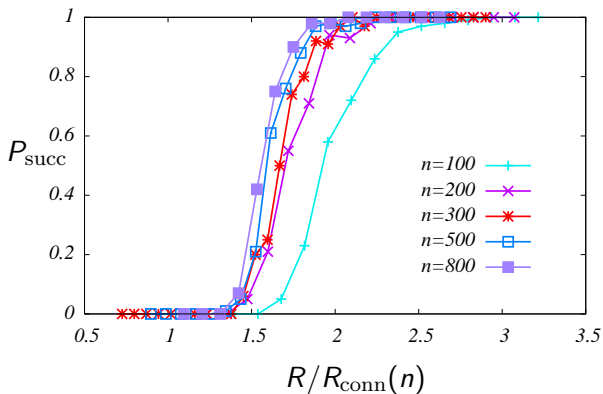
Challenges

- Entries are not sampled uniformly.
- Noise model.
- Distributed algorithm.

A small simulation



A phase transition



Conclusion

Conclusion and future directions

1. Low-rank approximation: there is better than SVD
2. Precise characterization, still open
3. Distributed positioning, recommendation systems, etc.

Thanks

Conclusion and future directions

1. Low-rank approximation: there is better than SVD
2. Precise characterization, still open
3. Distributed positioning, recommendation systems, etc.

Thanks