Stability of Matrix Factorization for Collaborative Filtering

Yu-Xiang Wang, Huan Xu

National University of Singapore

yuxiangwang@nus.edu.sg, *mpexuh@nus.edu.sg*

29 June 2012

Formulation

What is the problem?

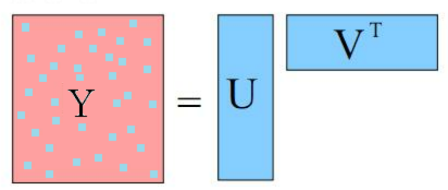
Predict the missing entries of a low-rank matrix, i.e., matrix completion.

$$\begin{pmatrix} 4 & ? & 7 & ? & 2 & ? \\ ? & ? & 3 & 9 & ? & 1 \\ 4 & 4 & ? & ? & ? & 0 \\ ? & 8 & ? & ? & 2 & ? \\ ? & ? & 3 & 0 & ? & 1 \\ 7 & 10 & ? & 6 & ? & ? \end{pmatrix}$$

$$\begin{aligned} \underset{Y}{\text{minimize}} \quad & \frac{1}{2}\left\| P_\Omega(Y - \widehat{Y}) \right\|_F^2 \\ \text{subject to} \quad & \mathrm{rank}(Y) \leq r. \end{aligned} \tag{1}$$

What is the formulation we analyze?

Matrix factorization(MF) that implicitly imposes rank constraint.



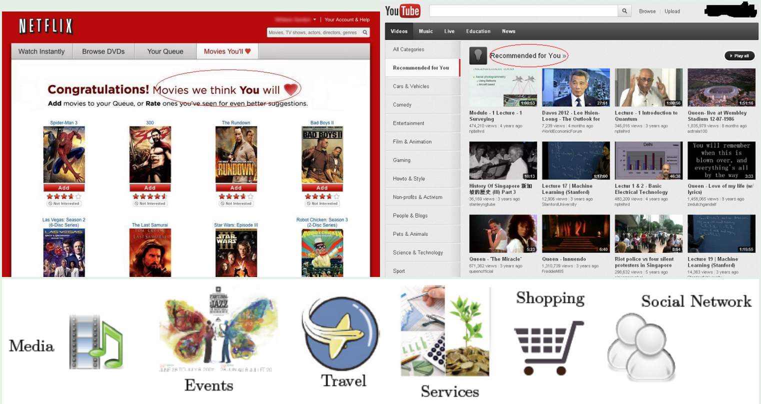
$$\underset{U,V}{\text{minimize}} \quad \frac{1}{2}\left\| P_\Omega(UV^T - \widehat{Y})\right\|_F^2, \tag{2}$$

It is a non-convex formulation that is popular in practice, but theory-free.

Applications

Example (Collaborative filtering/Recommender System)

Predict user preference based on their past ratings.



Emerging CF technology in everyday life.

Example (Collaborative filtering/Recommender System)

Taste of users are influenced only by a small number of latent factors.

$$\begin{array}{cc} & \text{Users} \\ \text{Products} & \begin{pmatrix} 4 & ? & 7 & ? & 2 & ? \\ ? & ? & \boxed{3} & 9 & ? & 1 \\ 4 & 4 & ? & ? & ? & 0 \\ ? & 8 & ? & ? & 2 & ? \\ ? & ? & 3 & 0 & ? & 1 \\ 7 & 10 & ? & 6 & ? & ? \end{pmatrix} \end{array} = \overset{\text{Products Features}}{\begin{pmatrix} 2 & 1 \\ 1 & 3 \\ 0 & 2 \\ 3 & 1 \\ 1 & 0 \\ 3 & 2 \end{pmatrix}} \times \overset{\text{User Features}}{\begin{pmatrix} 1 & 2 & 3 & 0 & 0 & 1 \\ 2 & 2 & 0 & 3 & 2 & 0 \end{pmatrix}}$$

Example (Collaborative filtering/Recommender System)

Taste of users are influenced only by a small number of latent factors.

Users

Products Features

User Features

$$
\text{Products}\;\begin{pmatrix} 4 & ? & 7 & ? & 2 & ? \\ ? & ? & \boxed{3} & 9 & ? & 1 \\ 4 & 4 & ? & ? & ? & 0 \\ ? & 8 & ? & ? & 2 & ? \\ ? & ? & 3 & 0 & ? & 1 \\ 7 & 10 & ? & 6 & ? & ? \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ \boxed{1 \quad 3} \\ 0 & 2 \\ 3 & 1 \\ 1 & 0 \\ 3 & 2 \end{pmatrix} \times \begin{pmatrix} 1 & 2 & \boxed{\begin{matrix} 3 \\ 0 \end{matrix}} & 0 & 0 & 1 \\ 2 & 2 & & 3 & 2 & 0 \end{pmatrix}
$$

User 3's overall preference for Product 2

Product 2's feature vector

User 3's feature vector

$$
3 = \begin{pmatrix} 1 & 3 \end{pmatrix} \times \begin{pmatrix} 3 \\ 0 \end{pmatrix}
$$

Performance

Portability

Preference for performance

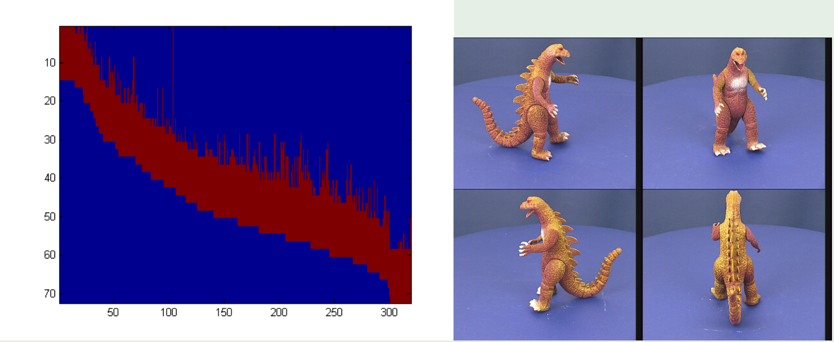
Preference for portability

e.g. Product = Laptop

Applications

Example (3D Structure from Motion in Computer Vision)

Feature track is usually short and incomplete. Full feature matrix can be factorized into the multiplication of camera matrix and structure matrix.



The feature matrix and snapshots of Oxford dinosaur sequence.

Applications

Example (Other applications)

- Localization in wireless sensor network
- System identification in control
- Prediction of missing components in DNA microarray

Example (Other applications)

- Localization in wireless sensor network
- System identification in control
- Prediction of missing components in DNA microarray

Common traits of the applications

- All can be formulated as low-rank matrix completion problem.
- All have researchers who propose to solve by MF (with various algorithms).
- Often with *convincing* empirical results, despite noisy data.

Motivation

Title: Stability of Matrix Factorization for Collaborative Filtering.

Motivation

Title: Stability of Matrix Factorization for Collaborative Filtering.

Why study stability?

Noise and corruptions: Real data are subject to noise and corruptions.

Low-rank as an approximation: How does MF work when data is not exactly low rank?

Manipulator problem in CF: A nasty yet common problem in commercial recommender systems. Also called "Shilling attacks", "Profile-injection attacks".

Motivation

Title: Stability of Matrix Factorization for Collaborative Filtering.

Why study stability?

Noise and corruptions Real data are subject to noise and corruptions.

Low-rank as an approximation How does MF work when data is not exactly low rank?

Manipulator problem in CF A nasty yet common problem in commercial recommender systems. Also called "Shilling attacks", "Profile-injection attacks".

Inherent robustness of MF?

There has been empirical observations that MF is more robust to such attacks compared to kNN. Is there a reason for this?

Model of our analysis

Notations

- $Y \in \mathbb{R}^{m \times n}$: ground truth rank-$r$ matrix
- $\widehat{Y} = Y + E$ is the corrupted observation
- P_Ω: projection to observed matrix entries
- $N^{gnd} \in \mathbb{R}^{m \times r}$ stands for the r dimensional column space of Y.
- $Y^* = U^* {V^*}^T$, N^* represents the optimal solutions.

Assumptions

- Matrix entry bounded by k.
- Sampling is uniformly random.

Main contributions

A comprehensive analysis of MF stability.

Stability metrics

Overall stability $RMSE = \frac{1}{\sqrt{mn}}\|Y^* - Y\|_F$

Subspace stability $\|\sin\Theta\| = \|\sin(\angle(N^*, N^{gnd}))\|$

Individual user stability $RMSE(i) = \frac{1}{\sqrt{m}}\|y_i^* - y_i\|_2$

RMSE bound	Noisy matrix completion; Collaborative filtering (as in Netflix Challenge)
Canonical angle bound	PCA with missing data; dimension reduction; subspace tracking.
Individual RMSE bound	Incremental algorithms; New user without recomputing full factorization; A worst case bound for individual user.

RMSE Stability

Theorem (RMSE Stability)

There exists an absolute constant C, such that with probability at least $1 - 2\exp(-n)$,

$$\mathrm{RMSE} \leq \frac{1}{\sqrt{|\Omega|}}\|P_\Omega(E)\|_F + \frac{\|E\|_F}{\sqrt{mn}} + Ck\left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}}.$$

RMSE Stability

Theorem (RMSE Stability)

There exists an absolute constant C, such that with probability at least $1 - 2\exp(-n)$,

$$\mathrm{RMSE} \leq \frac{1}{\sqrt{|\Omega|}} \|P_\Omega(E)\|_F + \frac{\|E\|_F}{\sqrt{mn}} + Ck \left(\frac{nr\log(n)}{|\Omega|} \right)^{\frac{1}{4}}.$$

- Sample requirement: $|\Omega| > \Theta(nr\log(n))$ (diminishing sample rate!)
- Only a log factor adding to d.o.f. (Arguably best to hope for.)
- In general, $\mathrm{RMSE} \leq C' \sqrt{1/|\Omega|} \|P_\Omega(E)\|_F$, as long as sample requirement is met.

RMSE Stability

Benchmarking our Stability results

Our result: $\mathrm{RMSE} \leq C\sqrt{1/|\Omega|}\|P_\Omega(E)\|_F$

RMSE Stability

Benchmarking our Stability results

Our result: $\mathrm{RMSE} \leq C\sqrt{1/|\Omega|}\|P_\Omega(E)\|_F$

StableMC[1]: $\mathrm{RMSE} \leq \sqrt{\frac{32\min(m,n)}{|\Omega|}}\|P_\Omega(E)\|_F + \frac{1}{\sqrt{mn}}\|P_\Omega(E)\|_F.$

[1] Candes, E.J. and Plan, Y. Matrix completion with noise.(2010) *Transactions of IEEE*, 98, 925 – 936.

RMSE Stability

Benchmarking our Stability results

Our result: $\mathrm{RMSE} \leq C\sqrt{1/|\Omega|}\|P_\Omega(E)\|_F$

StableMC[1]: $\mathrm{RMSE} \leq \sqrt{\frac{32\min(m,n)}{|\Omega|}}\|P_\Omega(E)\|_F + \frac{1}{\sqrt{mn}}\|P_\Omega(E)\|_F.$

OptSpace[2]: $\mathrm{RMSE} \leq C\kappa^2\frac{n\sqrt{r}}{|\Omega|}\|P_\Omega(E)\|_2.$

[1] Candes, E.J. and Plan, Y. Matrix completion with noise.(2010) *Transactions of IEEE*, 98, 925 – 936.

[2] Keshavan, R.H. and Montanari, A. and Oh, S. Matrix completion with noise.(2010) *IEEE Info. Theory*, 56, 2980 – 2998.

RMSE Stability

Benchmarking our Stability results

Our result: $\mathrm{RMSE} \leq C\sqrt{1/|\Omega|}\|P_\Omega(E)\|_F$

StableMC[1]: $\mathrm{RMSE} \leq \sqrt{\frac{32\min(m,n)}{|\Omega|}}\|P_\Omega(E)\|_F + \frac{1}{\sqrt{mn}}\|P_\Omega(E)\|_F.$

OptSpace[2]: $\mathrm{RMSE} \leq C\kappa^2\frac{n\sqrt{r}}{|\Omega|}\|P_\Omega(E)\|_2.$

Oracle bound: $\mathrm{RMSE} \approx \sqrt{1/|\Omega|}\|P_\Omega(E)\|_F.$

[1] Candes, E.J. and Plan, Y. Matrix completion with noise.(2010) *Transactions of IEEE*, 98, 925 – 936.

[2] Keshavan, R.H. and Montanari, A. and Oh, S. Matrix completion with noise.(2010) *IEEE Info. Theory*, 56, 2980 – 2998.

RMSE Stability

Benchmarking our Stability results

Our result: $\mathrm{RMSE} \leq C\sqrt{1/|\Omega|}\|P_\Omega(E)\|_F$

StableMC[1]: $\mathrm{RMSE} \leq \sqrt{\frac{32\min(m,n)}{|\Omega|}}\|P_\Omega(E)\|_F + \frac{1}{\sqrt{mn}}\|P_\Omega(E)\|_F.$

OptSpace[2]: $\mathrm{RMSE} \leq C\kappa^2 \frac{n\sqrt{r}}{|\Omega|}\|P_\Omega(E)\|_2.$

Oracle bound: $\mathrm{RMSE} \approx \sqrt{1/|\Omega|}\|P_\Omega(E)\|_F.$

Our result is optimal up to a constant factor!

[1] Candes, E.J. and Plan, Y. Matrix completion with noise.(2010) *Transactions of IEEE*, 98, 925 – 936.

[2] Keshavan, R.H. and Montanari, A. and Oh, S. Matrix completion with noise.(2010) *IEEE Info. Theory*, 56, 2980 – 2998.

Subspace Stability

Theorem (Subspace stability)

When Y is perturbed by additive error E and observed only on Ω, then there exists a Δ satisfying $\|\Delta\| \leq \sqrt{\frac{mn}{|\Omega|}}\|P_\Omega(E)\|_F + \|E\|_F + \sqrt{mn}|\tau(\Omega)|$, such that:

$$\|\sin\Theta\| \leq \frac{\sqrt{2}}{\delta}\|(\mathbb{P}^{\mathcal{N}^\perp}\Delta)\|; \quad \|\sin\Phi\| \leq \frac{\sqrt{2}}{\delta}\|(\mathbb{P}^{\mathcal{M}^\perp}\Delta^T)\|,$$

where $\|\cdot\|$ is either the Frobenious norm or the spectral norm, and $\delta = \sigma_r^*$, the r^{th} largest singular value of the recovered matrix Y^*, satisfying:

$$\sigma_r - \|\Delta\|_2 \leq \delta \leq \sigma_r + \|\Delta\|_2.$$

Subspace Stability

Explaining Subspace Stability

- Column space and row space are both stable when $\sigma_r \gg \|\Delta\|$
- When matrix is well-conditioned, σ_r is a constant fraction of $\|Y\|_F/\sqrt{r}$. (very large!)
- A better measure for manipulator problem and incremental algorithm.
- The result relies on our RMSE stability and perturbation theory of SVD [3].

[3] Stewart, G.W., Perturbation theory for the singular value decomposition.(1998)

Individual/Prediction error bound

Theorem (Individual/Prediction error)

Let $N_1 \in \mathbb{R}^{|\omega| \times r}$ *denote the restriction of* $N \in \mathbb{R}^{m \times r}$ *on the observed entries of* y. *The least square prediction* $\tilde{y}^*$ *of* $y \in \mathcal{N}^{gnd}$ *via*

$$\tilde{y}^* = N(N_1^T N_1)^{-1} N_1 y_1,$$

has bounded performance:

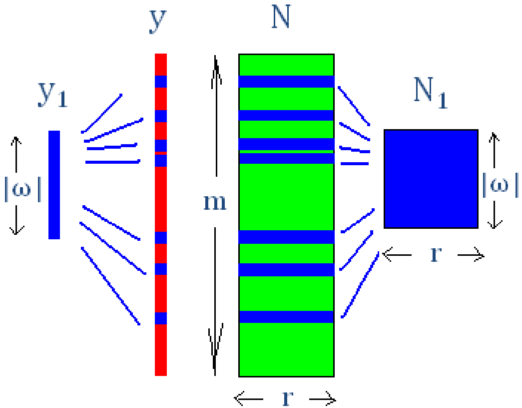
$$\|\tilde{y}^* - y\| \leq \left(1 + \frac{1}{\sigma_{min}}\right) \rho \|y\|,$$

where $\rho = \|\sin \Theta\|$ *(as in Subspace Stability Theorem),* σ_{min} *is the smallest non-zero singular value of* N_1 *(*r^{th} *when* N_1 *is non-degenerate).*

Individual/Prediction error bound

Explaining N_1 and y_1

$$\tilde{y}^* = N(N_1^T N_1)^{-1} N_1 y_1,$$



Shortened vector & shortened subspace

Individual/Prediction error bound

Explaining Individual/Prediction error bound

To understand

$$\|\tilde{y}^* - y\| \leq \left(1 + \frac{1}{\sigma_{min}}\right) \rho \|y\|$$

- $\rho = \|\sin\Theta\|_2$ is bounded by subspace stability.
- σ_{min} is bounded under incoherence assumption.
- For random matrix $\sigma_{min} \approx \sqrt{\frac{|\omega|}{m}} = \sqrt{p}$.

Individual/Prediction error bound

Explaining Individual/Prediction error bound

To understand

$$\|\tilde{y}^* - y\| \leq \left(1 + \frac{1}{\sigma_{min}}\right) \rho \|y\|$$

- $\rho = \|\sin\Theta\|_2$ is bounded by subspace stability.
- σ_{min} is bounded under incoherence assumption.
- For random matrix $\sigma_{min} \approx \sqrt{\frac{|\omega|}{m}} = \sqrt{p}$.
- When subspace recovery is good and sample rate is significant, prediction for for individual users has guaranteed good results.
- *Exact* when subspace recovery is perfect ($\rho = 0$).

Manipulator problem revisited

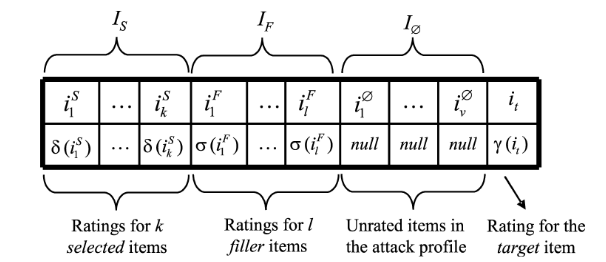
Manipulator injects dummy user profiles to distort the recommendation.

Illustration of manipulator attacks

	Item1	Item2	Item3	Item4	Item5	Item6	Correlation with Alice
Alice	5	2	3	3		?	
User1	2		4		4	1	-1.00
User2	3	1	3		1	2	0.76
User3	4	2	3	1		1	0.72
User4	3	3	2	1	3	1	0.21
User5		3		1	2		-1.00
User6	4	3		3	3	2	0.94
User7		5		1	5	1	-1.00
Attack1	**5**		**3**		**2**	**5**	1.00
Attack2	**5**	**1**	**4**		**2**	**5**	0.89
Attack3	**5**	**2**	**2**	**2**		**5**	0.93
Correlation with Item6	0.85	-0.55	0.00	0.48	-0.59		

Manipulator problem revisited

Attack models[4]



- Push attack/Nuke attack
- Random attack/Average attack
- Bandwagon attack/Segment attack
- Love/Hate attack

[4] Mobasher, B. and Burke, R. and Bhaumik, R. and Williams, C. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness.(2007) *ACM Tran. Info. Tech.* 7, 23.

Manipulator problem revisited

Attack models here

Targeted Attack Push/nuke targeted s items, otherwise pretend to be honest user. $e = e^{gnd} + s$ with sparse s.

Mass Attack General attacks that do not try any form of camouflage. $e = e^{gnd} + e^{gnd^{\perp}}$ where e^{gnd} and $e^{gnd^{\perp}}$ are of similar size.

Manipulator problem revisited

Attack models here

Targeted Attack Push/nuke targeted s items, otherwise pretend to be honest user. $e = e^{gnd} + s$ with sparse s.

Mass Attack General attacks that do not try any form of camouflage. $e = e^{gnd} + e^{gnd^{\perp}}$ where e^{gnd} and $e^{gnd^{\perp}}$ are of similar size.

To apply our theorems:

- E: Dummy user matrix (of width n_e).
- Y: Honest user matrix (of width n).

Manipulator problem revisited

Attack models here

Targeted Attack Push/nuke targeted s items, otherwise pretend to be honest user. $e = e^{gnd} + s$ with sparse s.

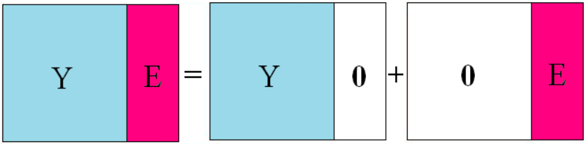
Mass Attack General attacks that do not try any form of camouflage. $e = e^{gnd} + e^{gnd^{\perp}}$ where e^{gnd} and $e^{gnd^{\perp}}$ are of similar size.

To apply our theorems:

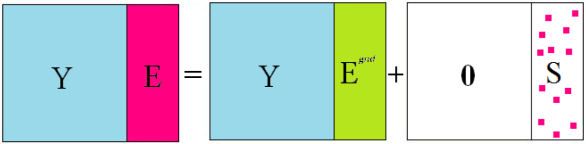
- E: Dummy user matrix (of width n_e).
- Y: Honest user matrix (of width n).

Key concept: Only prediction for Y block matters! We may assign arbitrary ground truth to E block.

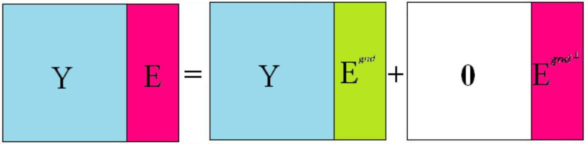
Instead of:



TargetAttack:



MassAttack:



Robustness of MF to manipulator attacks

Robustness to Targeted Attacks

Proposition 3: MF is *strongly robust* to Targeted Attacks.

$$\mathrm{RMSE} \leq 4k\sqrt{\frac{s_{max} n_e}{|\Omega|}} + Ck\left(\frac{(n+n_e)r\log(n+n_e)}{|\Omega|}\right)^{\frac{1}{4}}.$$

Ideas and implications:

- The bulk of Targeted Attacks is still inside the true subspace.
- When s is small, its impact $E^{gnd^{\perp}}$ to the recovered subspace is small.
- Essentially, RMSE converge to 0 when dimension m increase.
- n_e can be as large as the number of honest users n.

Robustness of MF to manipulator attacks

Robustness to Mass Attacks

Proposition 4: MF is only *weakly robust* to Mass Attacks.

If $n_e < \frac{\sqrt{n}}{\kappa^2 r}(\frac{\mathbf{E}|Y_{i,j}|^2}{k^2})$ and $|\Omega| = pm(n+n_e)$ satisfying $p > 1/m^{1/4}$

$$\mathrm{RMSE}_Y \leq C_1 \kappa k \left(\frac{r^3 \log(n)}{p^3 n}\right)^{1/4}, \qquad \mathrm{RMSE}_E \leq \frac{C_2 k}{\sqrt{p}},$$

Ideas and implications:

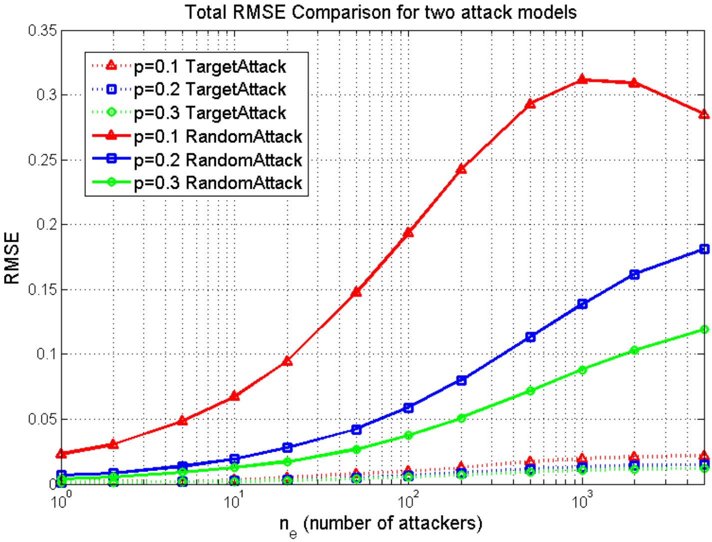
- Idea is that when number of attacks are small, recovered subspace error $\|\sin(\Theta)\|$ is small.
- Error impact concentrates on E(dummy user) block.
- n_e can only be a small fraction of $\sqrt{n}$ for $RMSE_Y$ to go to 0 asymptotically.

Numerical verification

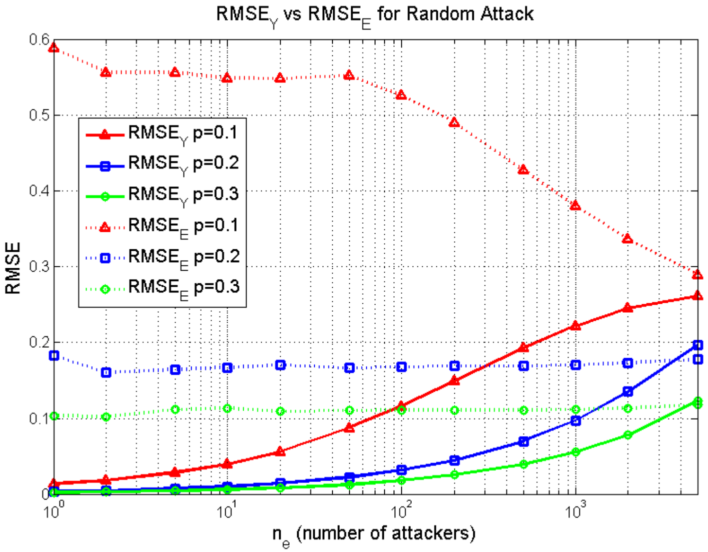
Setting of the simulation

1. $Y \in \mathbb{R}^{1000\times 1000}$, $rank(Y) = 10$. $E \in \mathbb{R}^{1000\times n_e}$.
2. Targeted Attack: randomly copy a column of Y, assign 2 "push" and 2 "nuke" targets.
3. Mass Attack: random column, assign 2 "push" and 2 "nuke" targets.
4. Algorithm: Alternating Least Square (ALS).

Numerical verification

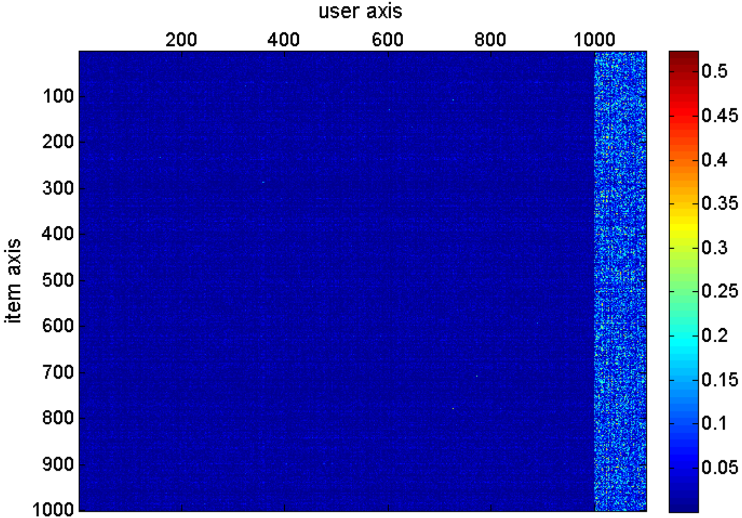


Numerical verification



Numerical verification

Illustration of error distribution of random attacks at $n_e = 100$.



Conclusions and future works

Conclusions

- A comprehensive study of the stability of MF (first of its kind).
- A near-optimal stability bound, a subspace stability bound and a worst-case bound for individual columns.
- A insightful illustration of MF's inherent robustness to manipulators.

Future directions

- Theoretical front: Under what conditions can MF reach global optimal? With which algorithm?
- Algorithmic front: Develop robust variation of MF that *provably* handles *arbitrary* attacks.

Questions and answers

Ask me more at Poster 83 outside the LT.

Additional slide: Algorithms for MF

Incomplete list of algorithms for MF

- Alternating: ALS, PowerFactorization, IRLS;
- Second order: Wiberg, Damped Newton, LM_X;
- Incremental/stochastic: GROUSE/GRASTA;
- Convex relaxation: SVT, APG, FPCA, ALM (Not necessarily rank-r)
- Other methods: MF-LRSDP(low-rank SDP), LMaFit(Alternating & SOR-like), OptSpace(SVD-based with theoretical guarantee)

Additional slide: Algorithms for MF

Performance evaluation MF algorithms

- Solution depends on algorithm and initialization.
- Factorization/Grassmannian methods empirically performs better than convex relaxation at the trade-off of losing the global optimality.
- Convincing empirical results are demonstrated (LM_X, MF-LRSDP and LMaFit).
- Some algorithms have larger basin of convergence (LM_X).

Additional slide: Global optimal assumption

Analysis independent to algorithms

- We analyze the *global solution* of MF-formulation (not specific algorithm).
- We assume that under certain conditions global optimal can be reached with high probability (at least for some algorithm).
  - Random low-rank matrices are *always* exactly completed (in MF-LRSDP paper).
  - LM_X: 90% of random initializations end up at global minimum on real noisy SfM data.
- Our results might be over-optimistic, but not completely unrealistic.