Nonconvex Optimization for High-Dimensional Estimation (Part 3)

Yuxin Chen

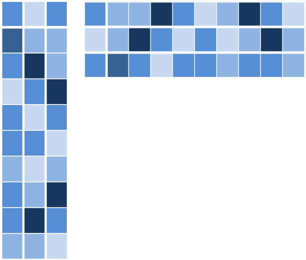
Wharton Statistics & Data Science, Spring 2022

Bridging convex and nonconvex optimization in estimation and inference

Noisy low-rank matrix completion

observations: $\quad M_{i,j} = M^{\star}_{i,j} + \text{noise}, \quad (i,j) \in \Omega$

goal: $\quad$ estimate $\boldsymbol{M}^{\star}$

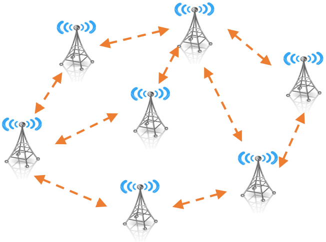


unknown rank-r matrix $\boldsymbol{M}^{\star} \in \mathbb{R}^{n \times n}$

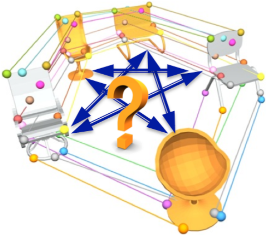
$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

sampling set Ω

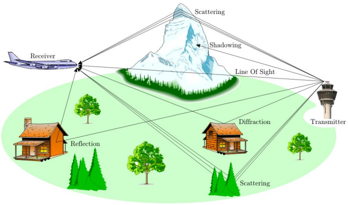
recommendation systems



localization



shape matching



channel estimation

Noisy low-rank matrix completion

observations: $M_{i,j} = M^\star_{i,j} + \text{noise}, \quad (i,j) \in \Omega$

goal: estimate $\boldsymbol{M}^\star$

convex relaxation:

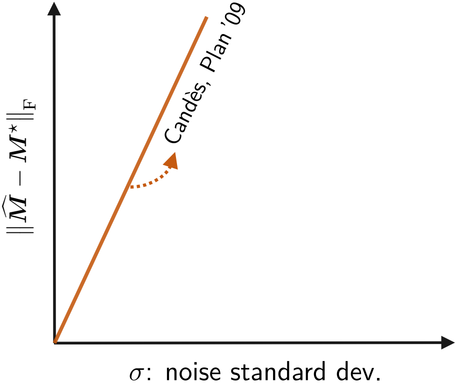
$$\underset{\boldsymbol{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \underbrace{\sum_{(i,j) \in \Omega} \left(Z_{i,j} - M_{i,j}\right)^2}_{\text{squared loss}} + \lambda \|\boldsymbol{Z}\|_*$$

— $\|\boldsymbol{Z}\|_* = \sum_{i=1}^n \sigma_i(\boldsymbol{Z})$

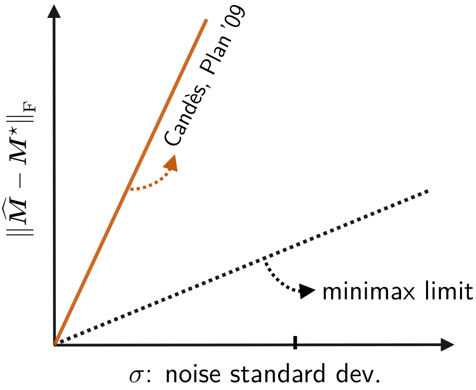
Prior statistical guarantees for convex relaxation

- **random sampling:** each $(i, j) \in \Omega$ indep. with prob. p
- **random noise:** i.i.d. sub-Gaussian noise with variance σ^2
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: rank $r = O(1)$, well-conditioned,...

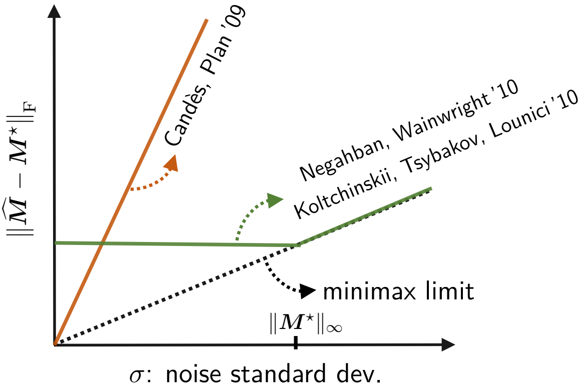
Candès, Plan '09	$\sigma n^{1.5}$

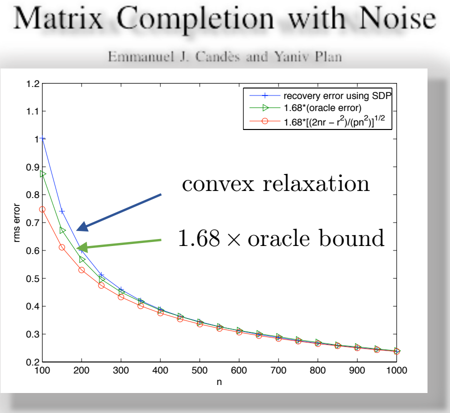


minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$



minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$
Koltchinskii, Tsybakov, Lounici '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$





Existing theory for convex relaxation does not match practice . . .

Matrix Completion with Noise

Emmanuel J. Candès and Yaniv Plan

with adversarial noise. Consequently, our analysis looses a $\sqrt{n}$ factor vis a vis an optimal bound that is achievable via the help of an oracle.

Existing theory for convex relaxation does not match practice . . .

What are the roadblocks?

Strategy: $\boldsymbol{M}^{\mathsf{cvx}}$ is optimizer if $\underbrace{\text{there exists } \boldsymbol{W}}_{\text{dual certificate}}$ s.t.

$(\boldsymbol{M}^{\mathsf{cvx}}, \boldsymbol{W})$ obeys KKT optimality condition

What are the roadblocks?

Strategy: $\boldsymbol{M}^{\mathsf{cvx}}$ is optimizer if $\underbrace{\text{there exists } \boldsymbol{W}}_{\text{dual certificate}}$ s.t.

$(\boldsymbol{M}^{\mathsf{cvx}}, \boldsymbol{W})$ obeys KKT optimality condition

David Gross

- **noiseless case**: $\underbrace{\boldsymbol{M}^{\mathsf{cvx}} \leftarrow \boldsymbol{M}^{\star}}_{\text{exact recovery}}$; $\boldsymbol{W} \leftarrow$ golfing scheme

What are the roadblocks?

Strategy: $\boldsymbol{M}^{\mathsf{cvx}}$ is optimizer if $\underbrace{\text{there exists } \boldsymbol{W}}_{\text{dual certificate}}$ s.t.

$(\boldsymbol{M}^{\mathsf{cvx}}, \boldsymbol{W})$ obeys KKT optimality condition

David Gross

- **noiseless case**: $\underbrace{\boldsymbol{M}^{\mathsf{cvx}} \leftarrow \boldsymbol{M}^{\star}}_{\text{exact recovery}}$; $\boldsymbol{W} \leftarrow$ golfing scheme
- **noisy case**: $\boldsymbol{M}^{\mathsf{cvx}}$ is very complicated, hard to construct $\boldsymbol{W}$ . . .

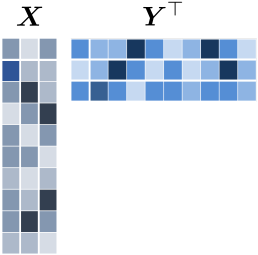
dual certification (golfing scheme)

dual certification (golfing scheme)

nonconvex optimization

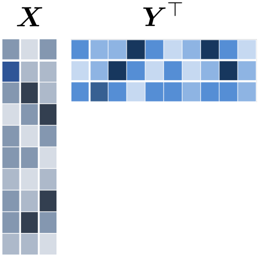
A detour: nonconvex optimization

Burer–Monteiro: represent $\boldsymbol{Z}$ by $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\underbrace{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}}_{\text{low-rank factors}}$



A detour: nonconvex optimization

Burer–Monteiro: represent $\boldsymbol{Z}$ by $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\underbrace{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}}_{\text{low-rank factors}}$



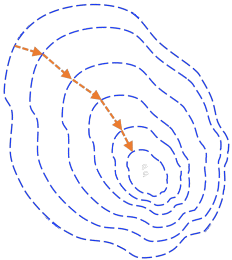
$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \underbrace{\sum_{(i,j)\in\Omega}\Big[\big(\boldsymbol{X}\boldsymbol{Y}^\top\big)_{i,j} - M_{i,j}\Big]^2}_{\text{squared loss}} + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

A detour: nonconvex optimization

- Burer, Monteiro '03
- Rennie, Srebro '05
- Keshavan, Montanari, Oh '09 '10
- Jain, Netrapalli, Sanghavi '12
- Hardt '13
- Sun, Luo '14
- Chen, Wainwright '15
- Tu, Boczar, Simchowitz, Soltanolkotabi, Recht '15
- Zhao, Wang, Liu '15
- Zheng, Lafferty '16
- Yi, Park, Chen, Caramanis '16
- Ge, Lee, Ma '16
- Ge, Jin, Zheng '17
- Ma, Wang, Chi, Chen '17
- Chen, Li '18
- Chen, Liu, Li '19
- ...

A detour: nonconvex optimization

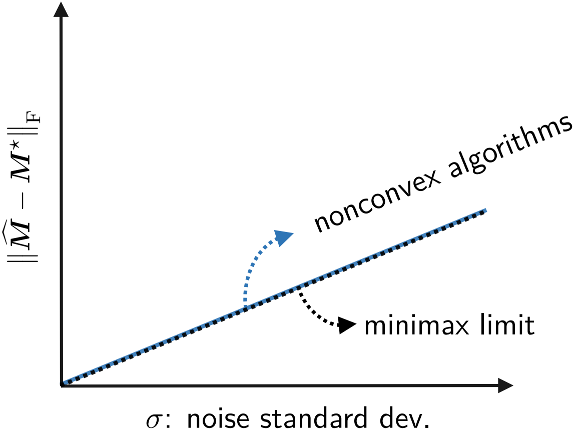
$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \sum_{(i,j)\in\Omega} \Big[\big(\boldsymbol{X}\boldsymbol{Y}^\top\big)_{i,j} - M_{i,j} \Big]^2 + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

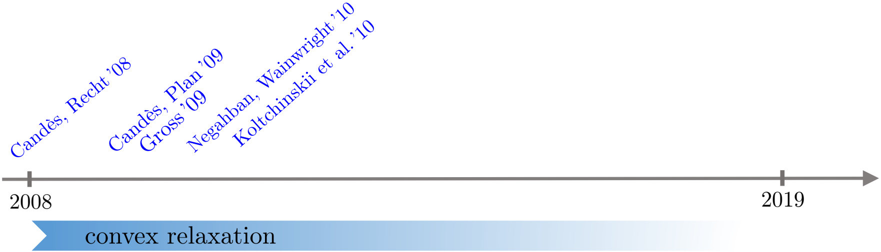


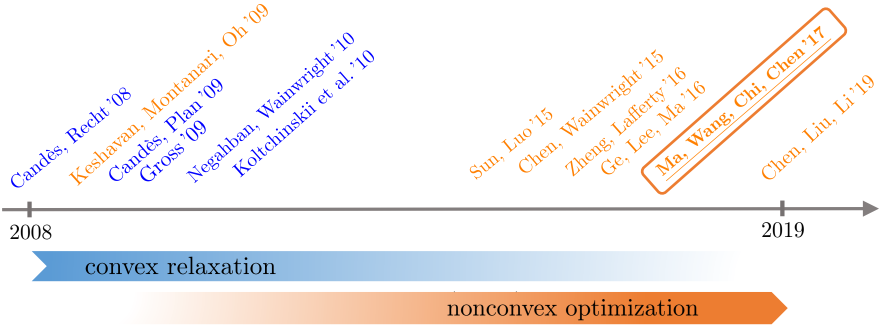
- **suitable initialization:** $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$
- **gradient descent:** for $t = 0, 1, \ldots$

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$
$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta_t \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$

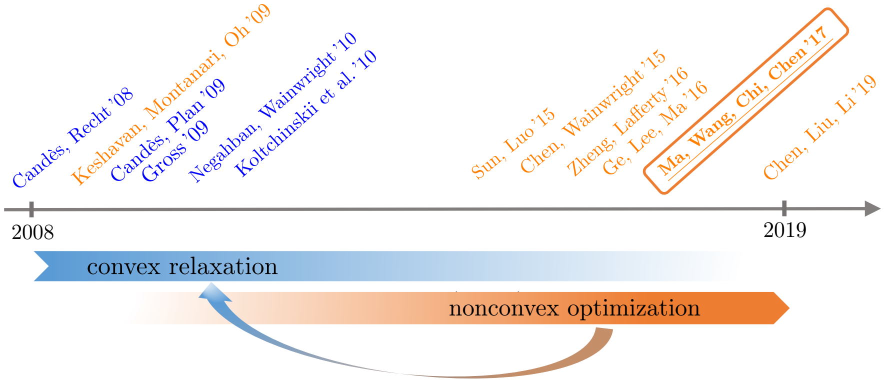
minimax limit	$\sigma\sqrt{n/p}$
nonconvex algorithms	$\sigma\sqrt{n/p}$ (optimal!)




Candès, Recht '08
Candès, Plan '09
Gross '09
Negahban, Wainwright '10
Koltchinskii et al. '10
2008
2019
convex relaxation



Candès, Recht '08
Keshavan, Montanari, Oh '09
Candès, Plan '09
Gross '09
Negahban, Wainwright '10
Koltchinskii et al. '10
Sun, Luo '15
Chen, Wainwright '15
Zheng, Lafferty '16
Ge, Lee, Ma '16
Ma, Wang, Chi, Chen '17
Chen, Liu, Li '19
2008
2019
convex relaxation
nonconvex optimization


Candès, Recht '08
Keshavan, Montanari, Oh '09
Candès, Plan '09
Gross '09
Negahban, Wainwright '10
Koltchinskii et al. '10
Sun, Luo '15
Chen, Wainwright '15
Zheng, Lafferty '16
Ge, Lee, Ma '16
Ma, Wang, Chi, Chen '17
Chen, Liu, Li '19
2008
2019
convex relaxation
nonconvex optimization

An interesting experiment

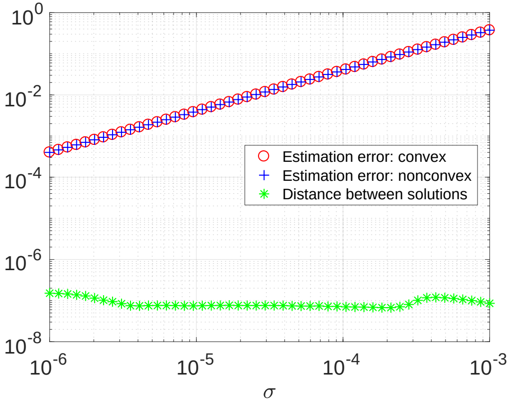
convex: $$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_*$$

nonconvex: $$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{i,j} - M_{i,j}\right]^2 + \underbrace{\frac{\lambda}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2}_{\text{reg}(\boldsymbol{X},\boldsymbol{Y})}$$

— $$\|\boldsymbol{Z}\|_* = \min_{\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{Y}^\top} \tfrac{1}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \tfrac{1}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2$$

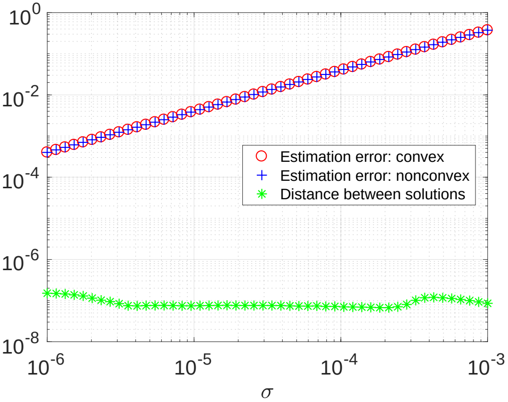
An interesting experiment

$$n = 1000,\ r = 5,\ p = 0.2,\ \lambda = 5\sigma\sqrt{np}$$



An interesting experiment

$$n = 1000, \;\; r = 5, \;\; p = 0.2, \;\; \lambda = 5\sigma\sqrt{np}$$



Convex and nonconvex solutions are exceedingly close!

convex  nonconvex

$$\Downarrow$$

$$\text{stability}\left(\boxed{\text{convex}}\right) \approx \text{stability}\left(\boxed{\text{nonconvex}}\right)$$

Main results: $r = O(1)$

- **random sampling:** each $(i, j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, well-conditioned, incoherent

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, well-conditioned, incoherent

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, well-conditioned, incoherent

$$\underset{\boldsymbol{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Theorem 1 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*

$$\left\|\boldsymbol{M}^{\mathsf{cvx}} - \mathsf{proj}_{\textit{rank-}r}(\boldsymbol{M}^{\mathsf{cvx}})\right\|_{\mathrm{F}} \ll \frac{1}{n^5} \cdot \sigma\sqrt{\frac{n}{p}}$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, well-conditioned, incoherent

$$\underset{\boldsymbol{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \sum_{(i,j) \in \Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma \sqrt{np})$$

Theorem 1 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*
2. $$\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}} \lesssim \sigma \sqrt{\frac{n}{p}}$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, well-conditioned, incoherent

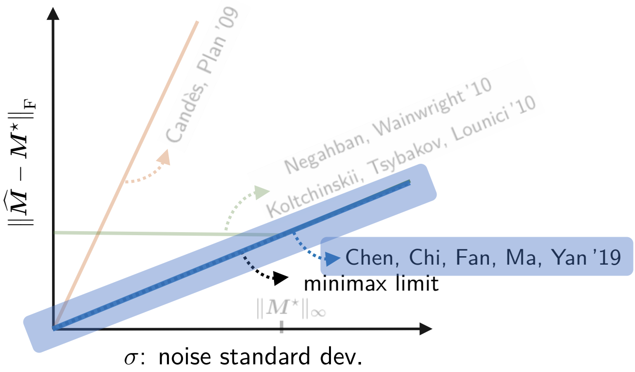
$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda\|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Theorem 1 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

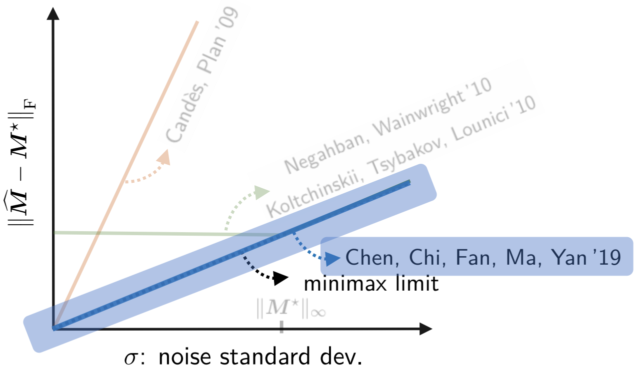
1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*
2. $$\big\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\big\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$
$$\big\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\big\|_{\infty} \lesssim \sigma\sqrt{\frac{n\log n}{p}} \cdot \frac{1}{n}$$

$$\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$



- minimax optimal when $r = O(1)$

$$\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}} \qquad \|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^{\star}\|_{\infty} \lesssim \sigma\sqrt{\frac{n\log n}{p}} \cdot \frac{1}{n}$$



- minimax optimal when $r = O(1)$
- estimation errors are spread out across all entries

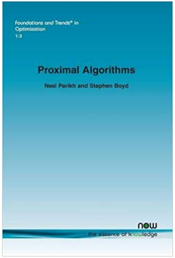
Implicit regularization

No need to enforce spikiness constraint as in Negahban & Wainwright

$$\underset{\color{red}{\|\boldsymbol{Z}\|_\infty \le \alpha}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_*$$

- convex relaxation automatically controls spikiness of solutions

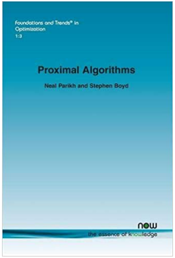
Statistical guarantees for iterative algorithms



$$\underset{\boldsymbol{Z}}{\text{minimize}} \quad g(\boldsymbol{Z}) := \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \quad (1)$$

Many algorithms (e.g. SVT, SOFT-IMPUTE, FPC, FISTA) have been proposed to solve (1), typically without statistical guarantees

Statistical guarantees for iterative algorithms



$$\underset{\boldsymbol{Z}}{\text{minimize}} \quad g(\boldsymbol{Z}) := \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \quad (1)$$

Many algorithms (e.g. SVT, SOFT-IMPUTE, FPC, FISTA) have been proposed to solve (1), typically without statistical guarantees

We provide statistical guarantees for any $\boldsymbol{Z}$ with $g(\boldsymbol{Z}) \leq g(\boldsymbol{Z}_{\mathsf{opt}}) + \varepsilon$ for some sufficiently small $\varepsilon > 0$

Main results: general case

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{r^2 \log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: well-conditioned, incoherent

Main results: general case

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{r^2 \log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: well-conditioned, incoherent

Theorem 2 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*
2. $$\begin{aligned} \left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} &\lesssim \frac{\sigma}{\sigma_{\min}(\boldsymbol{M}^\star)}\sqrt{\frac{n}{p}}\,\|\boldsymbol{M}^\star\|_{\mathrm{F}} \\ \left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\infty} &\lesssim \sqrt{r}\frac{\sigma}{\sigma_{\min}(\boldsymbol{M}^\star)}\sqrt{\frac{n\log n}{p}}\,\|\boldsymbol{M}^\star\|_{\infty} \\ \left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\| &\lesssim \frac{\sigma}{\sigma_{\min}(\boldsymbol{M}^\star)}\sqrt{\frac{n}{p}}\,\|\boldsymbol{M}^\star\| \end{aligned}$$

Main results: general case

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{r^2 \log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: well-conditioned, incoherent

sample complexity bound $O(nr^2 \log^3 n)$ is suboptimal in r

A little analysis:

connection between convex and nonconvex solutions

Link between convex and nonconvex optimizers

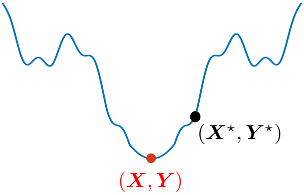
$(\boldsymbol{X}, \boldsymbol{Y})$ is nonconvex optimizer

Link between convex and nonconvex optimizers

$(\boldsymbol{X}, \boldsymbol{Y})$ is nonconvex optimizer $\quad \overset{?}{\Longrightarrow} \quad \boldsymbol{X}\boldsymbol{Y}^\top$ is convex solution

Link between convex and nonconvex optimizers

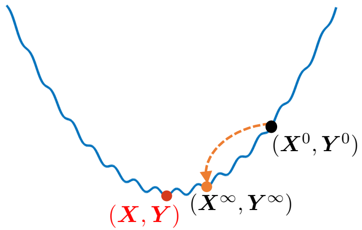
- λ is properly chosen
- $(\boldsymbol{X}, \boldsymbol{Y})$ is close to truth (in $\ell_{2,\infty}$ sense)



$(\boldsymbol{X}, \boldsymbol{Y})$ is nonconvex optimizer $\quad \overset{\checkmark}{\Longrightarrow} \quad \boldsymbol{X}\boldsymbol{Y}^\top$ is convex solution

i.e. dist(convex solution, nonconvex solution) $= 0$

Approximate nonconvex optimizers



Issue: we do NOT know statistical properties of nonconvex optimizers

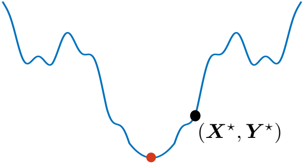
- It is unclear whether nonconvex algorithms converge to optimizers

Approximate nonconvex optimizers

Strategy: resort to “$\underbrace{\text{approximate stationary points}}_{\nabla f(\boldsymbol{X},\boldsymbol{Y}) \approx \boldsymbol{0}}$” instead

Approximate nonconvex optimizers

Strategy: resort to "$\underbrace{\text{approximate stationary points}}_{\nabla f(\boldsymbol{X},\boldsymbol{Y})\approx\boldsymbol{0}}$" instead



- λ is properly chosen
- $(\boldsymbol{X}, \boldsymbol{Y})$ is close to truth (in $\ell_{2,\infty}$ sense)

$$\nabla f(\boldsymbol{X}, \boldsymbol{Y}) \approx \boldsymbol{0} \quad \overset{\checkmark}{\Longrightarrow} \quad \mathrm{dist}(\boldsymbol{X}\boldsymbol{Y}^\top, \text{convex solutions}) \approx 0$$

Construct approximate nonconvex optimizers via GD

starting from $(\boldsymbol{X}^0, \boldsymbol{Y}^0) =$ truth or spectral initialization:

$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned} \qquad t = 0, 1, \cdots, T$$

Construct approximate nonconvex optimizers via GD

starting from $(\boldsymbol{X}^0, \boldsymbol{Y}^0) =$ truth or spectral initialization:

$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned} \qquad t = 0, 1, \cdots, T$$

- when T is large: there exists point with $\underbrace{\text{very small gradient}}_{\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{\eta T}}}$

Construct approximate nonconvex optimizers via GD

starting from $(\boldsymbol{X}^0, \boldsymbol{Y}^0) =$ truth or spectral initialization:

$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned} \qquad t = 0, 1, \cdots, T$$

- when T is large: there exists point with $\underbrace{\text{very small gradient}}_{\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{\eta T}}}$
- hopefully not far from $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$ (in $\ell_{2,\infty}$ sense in particular)

Analyzing nonconvex GD: leave-one-out analysis

Leave out a small amount of information from data and run GD

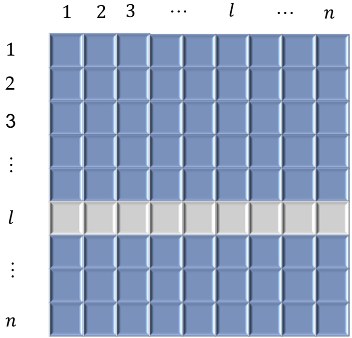
Analyzing nonconvex GD: leave-one-out analysis

Leave out a small amount of information from data and run GD

- Stein '72
- El Karoui, Bean, Bickel, Lim, Yu '13
- El Karoui '15
- Javanmard, Montanari '15
- Zhong, Boumal '17
- Lei, Bickel, El Karoui '17
- Sur, Chen, Candès '17
- Abbe, Fan, Wang, Zhong '17
- Chen, Fan, Ma, Wang '17
- Ma, Wang, Chi, Chen '17
- Chen, Chi, Fan, Ma '18
- Ding, Chen '18
- Dong, Shi '18
- Chen, Liu, Li '19

Analyzing nonconvex GD: leave-one-out analysis

For each $1 \le l \le n$, introduce leave-one-out iterates $\{(\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)})\}$ by replacing l^{th} row (or column) with true values



$\boldsymbol{M}^{(l)}$

$$\Longrightarrow \{(\boldsymbol{X}^{t,(l)}, \boldsymbol{Y}^{t,(l)})\}_{t \ge 0}$$

- exploit partial statistical independence
- exploit leave-one-out stability

Inference and uncertainty quantification

Reasoning about uncertainty

	2		2	
		6		
3	1		4	
		4		1
	0			

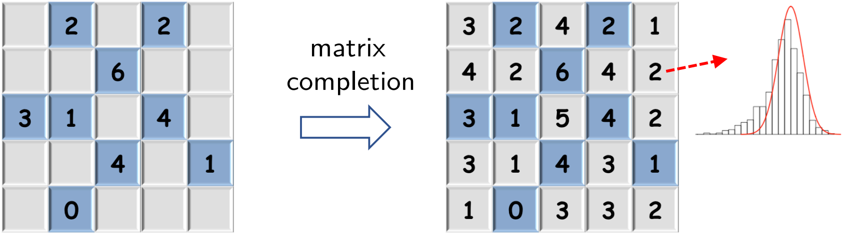
Reasoning about uncertainty

	2		2	
		6		
3	1		4	
		4		1
	0			

matrix completion

3	2	4	2	1
4	2	6	4	2
3	1	5	4	2
3	1	4	3	1
1	0	3	3	2

Reasoning about uncertainty



How to assess uncertainty, or “confidence”, of obtained estimates?

INFERENCE IN HIGH DIMENSIONAL REGRESSION

organized by
Peter Buehlmann, Andrea Montanari, and Jonathan Taylor

American Institute of Mathematics

(3) *Confidence intervals for matrix completion.* In matrix completion, the data analyst is given a large data matrix with a number of missing entries. In many interesting applications (e.g. to collaborative filtering) it is indeed the case that the vast majority of entries is missing. In order to fill the missing entries, the assumption is made that the underlying –unknown– matrix has a low-rank structure.

Substantial work has been devoted to methods for computing point estimates of the missing entries. In applications, it would be very interesting to compute confidence intervals as well. This requires developing distributional characterizations of standard matrix completion methods.

Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\arg\min} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_*$$

- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm

Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda \|\boldsymbol{Z}\|_*$$

- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm
- very challenging to pin down distributions of obtained estimates

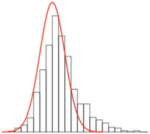
Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_*$$

- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm
- very challenging to pin down distributions of obtained estimates
- existing orderwise bounds come with unspecified (but huge) pre-constants

  $\longrightarrow$ overly wide confidence intervals

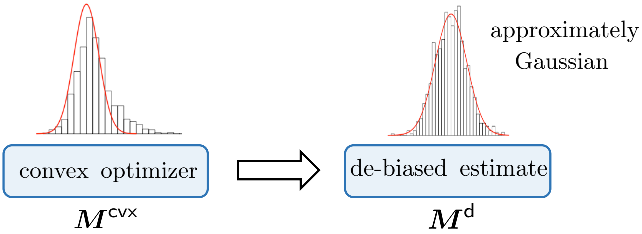
— inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13



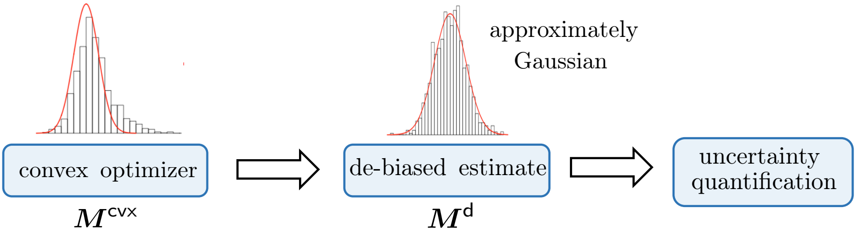
convex optimizer

M^{cvx}

— inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13



— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



De-biasing convex estimate

$$M^{\mathsf{cvx}} \quad \xrightarrow{\text{de-biasing}} \quad \underbrace{M^{\mathsf{cvx}} + \frac{1}{p}\mathcal{P}_{\Omega}(M^{\star} + E - M^{\mathsf{cvx}})}_{\text{(nearly) unbiased estimate of } M^{\star}}$$

De-biasing convex estimate

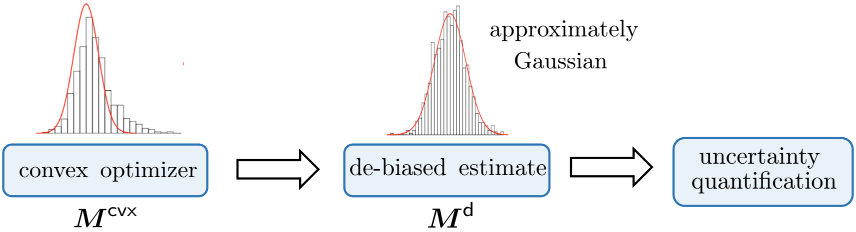
$$\boldsymbol{M}^{\mathsf{cvx}} \quad \xrightarrow{\text{de-biasing}} \quad \underbrace{\boldsymbol{M}^{\mathsf{cvx}} + \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\mathsf{cvx}})}_{\text{(nearly) unbiased estimate of } \boldsymbol{M}^{\star}}$$

- **issue:** high-rank after de-biasing; statistical accuracy suffers

De-biasing convex estimate

$$M^{\mathsf{cvx}} \overset{\text{de-biasing}}{\longrightarrow} \underbrace{\mathsf{proj}_{\mathsf{rank}\text{-}r}\Big(M^{\mathsf{cvx}} + \tfrac{1}{p}\mathcal{P}_\Omega(M^\star + E - M^{\mathsf{cvx}})\Big)}_{\text{1 iteration of singular value projection (Jain, Meka, Dhillon '10)}} =: M^{\mathrm{d}}$$

- **issue:** high-rank after de-biasing; statistical accuracy suffers
- **solution:** low-rank projection


approximately
Gaussian
convex optimizer
M^{cvx}
de-biased estimate
M^{d}
uncertainty
quantification

Distributional guarantees for low-rank factors

$$\boldsymbol{X}^{\mathsf{d}}\boldsymbol{Y}^{\mathsf{d}\top} \leftarrow \underbrace{\text{balanced}}_{\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}=\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}} \text{ rank-}r \text{ decomp. of } \boldsymbol{M}^{\mathsf{d}}$$

$$\boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top} \leftarrow \underbrace{\text{balanced}}_{\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}=\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star}} \text{ rank-}r \text{ decomp. of } \boldsymbol{M}^{\star}$$

Distributional guarantees for low-rank factors

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. $\mathcal{N}(0, \sigma^2)$ (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, well-conditioned, incoherent
- regularization parameter: $\lambda \asymp \sigma\sqrt{np}$

$$\boldsymbol{X}^{\mathsf{d}}\boldsymbol{Y}^{\mathsf{d}\top} \;\leftarrow\; \underbrace{\text{balanced}}_{\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}} = \boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}} \text{ rank-}r \text{ decomp. of } \boldsymbol{M}^{\mathsf{d}}$$

$$\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top} \;\leftarrow\; \underbrace{\text{balanced}}_{\boldsymbol{X}^{\star\top}\boldsymbol{X}^\star = \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star} \text{ rank-}r \text{ decomp. of } \boldsymbol{M}^\star$$

Distributional guarantees for low-rank factors

$$\boldsymbol{X}^{\mathsf{d}}\boldsymbol{Y}^{\mathsf{d}\top} \leftarrow \underbrace{\text{balanced}}_{\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}=\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}} \text{ rank-}r \text{ approx. of } \boldsymbol{M}^{\mathsf{d}}$$

$$\boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top} \leftarrow \underbrace{\text{balanced}}_{\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}=\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star}} \text{ rank-}r \text{ decomp. of } \boldsymbol{M}^{\star}$$

Theorem 3 (Chen, Fan, Ma, Yan '19)

With high prob., there exists global rotation matrix $\boldsymbol{H} \in \mathbb{R}^{r\times r}$ *s.t.*

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- estimation errors for different rows of $\boldsymbol{X}^{\star}$ are nearly independent

$$\boldsymbol{X}^{\mathrm{d}}_{i,\cdot}\boldsymbol{H} - \boldsymbol{X}^{\star}_{i,\cdot} \quad \text{nearly ind. of} \quad \boldsymbol{X}^{\mathrm{d}}_{j,\cdot}\boldsymbol{H} - \boldsymbol{X}^{\star}_{j,\cdot}$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- accurate uncertainty quantification for low-rank factors, e.g.

$$\boldsymbol{X}^{\mathsf{d}}_{i,\cdot}\boldsymbol{H} - \boldsymbol{X}^{\star}_{i,\cdot} \sim \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big) + \text{negligible term}$$

$$\boldsymbol{Y}^{\mathsf{d}}_{i,\cdot}\boldsymbol{H} - \boldsymbol{Y}^{\star}_{i,\cdot} \sim \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big) + \text{negligible term}$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- accurate uncertainty quantification for low-rank factors, e.g.

$$\boldsymbol{X}^{\mathsf{d}}_{i,\cdot} - \boldsymbol{X}^{\star}_{i,\cdot}\boldsymbol{H}^{\top} \sim \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}})^{-1}\big) + \text{negligible term}$$

$$\boldsymbol{Y}^{\mathsf{d}}_{i,\cdot} - \boldsymbol{Y}^{\star}_{i,\cdot}\boldsymbol{H}^{\top} \sim \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}})^{-1}\big) + \text{negligible term}$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- accurate uncertainty quantification for low-rank factors, e.g.

$$\boldsymbol{X}^{\mathsf{d}}_{i,\cdot} - \boldsymbol{X}^{\star}_{i,\cdot}\boldsymbol{H}^{\top} \;\sim\; \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}})^{-1}\big) + \text{negligible term}$$

$$\boldsymbol{Y}^{\mathsf{d}}_{i,\cdot} - \boldsymbol{Y}^{\star}_{i,\cdot}\boldsymbol{H}^{\top} \;\sim\; \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}})^{-1}\big) + \text{negligible term}$$

— *asymptotically optimal*

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H}-\boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H}-\boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- accurate uncertainty quantification for matrix entries: if $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ is not too small, then

$$M^{\mathsf{d}}_{i,j} - M^{\star}_{i,j} \;\sim\; \mathcal{N}\big(0, v^{\star}_{i,j}\big) + \text{negligible term}$$

where $v^{\star}_{i,j} \triangleq \frac{\sigma^2}{p}\Big\{\boldsymbol{X}^{\star}_{i,\cdot}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\boldsymbol{X}^{\star\top}_{i,\cdot} + \boldsymbol{Y}^{\star}_{j,\cdot}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\boldsymbol{Y}^{\star\top}_{j,\cdot}\Big\}$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\mathbf{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\mathbf{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- accurate uncertainty quantification for matrix entries: if $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ is not too small, then

$$M^{\mathsf{d}}_{i,j} - M^{\star}_{i,j} \;\sim\; \mathcal{N}\big(0, \widehat{v}_{i,j}\big) + \text{negligible term}$$

where $\widehat{v}_{i,j} \triangleq \frac{\sigma^2}{p}\Big\{\boldsymbol{X}^{\mathrm{d}}_{i,\cdot}(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}})^{-1}\boldsymbol{X}^{\mathrm{d}\top}_{i,\cdot} + \boldsymbol{Y}^{\mathrm{d}}_{j,\cdot}(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}})^{-1}\boldsymbol{Y}^{\mathrm{d}\top}_{j,\cdot}\Big\}$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1})$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1})$$

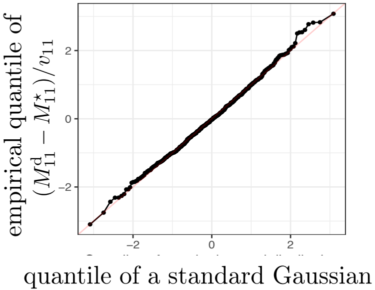
- accurate uncertainty quantification for matrix entries: if $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ is not too small, then

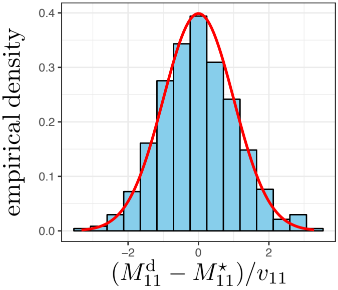
$$M^{\mathsf{d}}_{i,j} - M^{\star}_{i,j} \sim \mathcal{N}\big(0, \widehat{v}_{i,j}\big) + \text{negligible term}$$

where $\widehat{v}_{i,j} \triangleq \frac{\sigma^2}{p}\Big\{\boldsymbol{X}^{\mathrm{d}}_{i,\cdot}(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}})^{-1}\boldsymbol{X}^{\mathrm{d}\top}_{i,\cdot} + \boldsymbol{Y}^{\mathrm{d}}_{j,\cdot}(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}})^{-1}\boldsymbol{Y}^{\mathrm{d}\top}_{j,\cdot}\Big\}$

— *asymptotically optimal*

Numerical experiments





$$n = 1000,\ p = 0.2,\ r = 5,\ \|\boldsymbol{M}^{\star}\| = 1,\ \kappa = 1,\ \sigma = 10^{-3}$$

convex

nonconvex

convex

nonconvex

$$\Downarrow$$

$$\text{inference}\left(\boxed{\text{convex}}\right) \approx \text{inference}\left(\boxed{\text{nonconvex}}\right)$$

Same inference procedures work for both cvx & noncvx estimates!

A bit of intuition

Consider rank-1 PSD case $\boldsymbol{M}^\star = \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$, $p = 1$ (no missing data)

$$\text{minimize}_{\boldsymbol{x}} \qquad \frac{1}{2}\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{x}\|_2^2$$

A bit of intuition

Consider rank-1 PSD case $\boldsymbol{M}^\star = \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$, $p = 1$ (no missing data)

$$\text{minimize}_{\boldsymbol{x}} \qquad \frac{1}{2}\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{x}\|_2^2$$

- first-order optimality condition

$$(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x} + \lambda\boldsymbol{x} = \boldsymbol{0}$$

A bit of intuition

$$(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x} \quad \underbrace{+\lambda\boldsymbol{x}}_{\text{causes bias}} \quad = \boldsymbol{0}$$

A bit of intuition

$$(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x} \quad \underbrace{+\lambda \boldsymbol{x}}_{\text{causes bias}} \quad = \boldsymbol{0}$$

$$\Updownarrow$$

$$(\boldsymbol{x}^{\mathrm{d}}\boldsymbol{x}^{\mathrm{d}\top} - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x}^{\mathrm{d}} = \boldsymbol{0}, \qquad \boldsymbol{x}^{\mathrm{d}} = \sqrt{\frac{\lambda + \|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}}\,\boldsymbol{x}$$

A bit of intuition

$$\left(\boldsymbol{x}\boldsymbol{x}^{\top} - \boldsymbol{x}^{\star}\boldsymbol{x}^{\star\top} - \boldsymbol{E}\right)\boldsymbol{x} \quad \underbrace{+\lambda\boldsymbol{x}}_{\text{causes bias}} \quad = \boldsymbol{0}$$

$$\Updownarrow$$

$$\left(\boldsymbol{x}^{\mathrm{d}}\boldsymbol{x}^{\mathrm{d}\top} - \boldsymbol{x}^{\star}\boldsymbol{x}^{\star\top} - \boldsymbol{E}\right)\boldsymbol{x}^{\mathrm{d}} = \boldsymbol{0}, \qquad \boldsymbol{x}^{\mathrm{d}} = \sqrt{\tfrac{\lambda + \|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}}\,\boldsymbol{x}$$

$$\Updownarrow$$

$$\boldsymbol{x}^{\mathrm{d}} - \boldsymbol{x}^{\star} = \underbrace{\frac{1}{\|\boldsymbol{x}^{\mathrm{d}}\|_2^2}\boldsymbol{E}\boldsymbol{x}^{\mathrm{d}}}_{\text{nearly Gaussian}} + \underbrace{\frac{\left(\boldsymbol{x}^{\star} - \boldsymbol{x}^{\mathrm{d}}\right)^{\top}\boldsymbol{x}^{\mathrm{d}}}{\|\boldsymbol{x}^{\mathrm{d}}\|_2^2}\boldsymbol{x}^{\star}}_{\text{hopefully small}}$$

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

Theorem 4 (Chen, Fan, Ma, Yan '19)

$$\frac{\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}}^2}{n^2} = \underbrace{\frac{(2 + o(1)) nr\sigma^2}{n^2 p}}_{\textit{Oracle lower bound}} \quad \textit{with high prob.}$$

Back to estimation: de-biased estimator is optimal

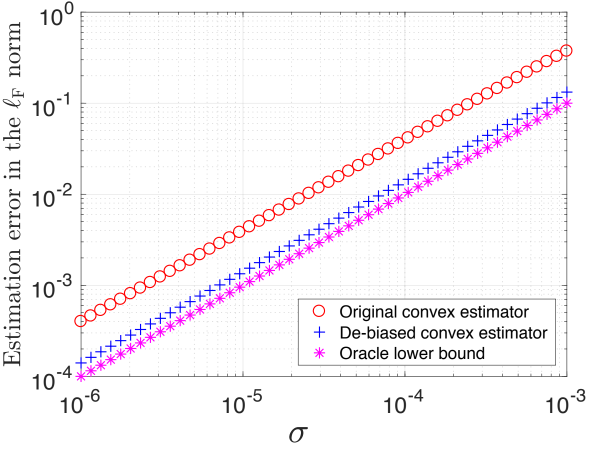
Distributional theory in turn allows us to track estimation accuracy

Theorem 4 (Chen, Fan, Ma, Yan '19)

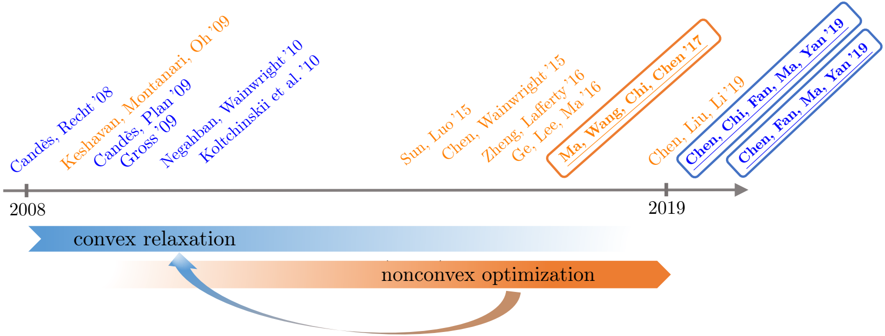
$$\frac{\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}}^2}{n^2} = \underbrace{\frac{(2+o(1))nr\sigma^2}{n^2 p}}_{\textit{Oracle lower bound}} \quad \textit{with high prob.}$$

- precise characterization of estimation accuracy
- achieves full statistical efficiency (including pre-constant)

Numerical evidence ($r = 5$, $p = 0.2$, $n = 1000$)



Euclidean estimation error vs. noise standard deviation σ


Candès, Recht '08
Keshavan, Montanari, Oh '09
Candès, Plan '09
Gross '09
Negahban, Wainwright '10
Koltchinskii et al. '10
Sun, Luo '15
Chen, Wainwright '15
Zheng, Lafferty '16
Ge, Lee, Ma '16
Ma, Wang, Chi, Chen '17
Chen, Liu, Li '19
Chen, Chi, Fan, Ma, Yan '19
Chen, Fan, Ma, Yan '19
2008
2019
convex relaxation
nonconvex optimization

Reference

- "*Matrix completion with noise*," E. J. Candès, Y. Plan, *Proceedings of the IEEE*, vol. 98, no. 6, 2010.
- "*Restricted strong convexity and weighted matrix completion: Optimal bounds with noise*," S. Negahban, M. Wainwright, *Journal of Machine Learning Research*, vol. 13, no. 1, 2012.
- "*Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion*," V. Koltchinskii,K. Lounici, A. B. Tsybakov, *Annals of Statistics*, vol. 39, no. 5, 2011.
- "*Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization*," Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, *SIAM Journal on Optimization*, vol. 30, no. 4, 2020.

Reference

- "*Inference and uncertainty quantification for noisy matrix completion*," Y. Chen, J. Fan, C. Ma, Y. Yan, *Proceedings of the National Academy of Sciences*, vol. 116, no. 46, 2019.
- "*Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution*," C. Ma, K. Wang, Y. Chi, Y. Chen, *Foundations of Computational Mathematics*, vol. 20, no. 3, 2020.
- "*Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data*," Y. Chen, J. Fan, C. Ma, Y. Yan, *Annals of Statistics*, vol. 49, no. 5, 2021.