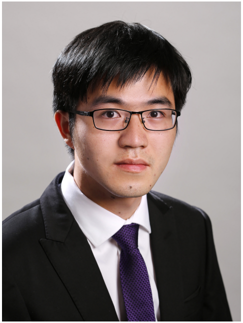
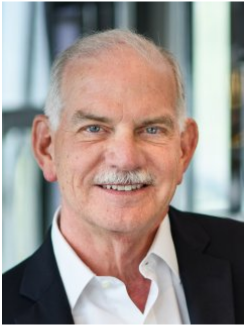
Uncertainty quantification for nonconvex tensor completion

Yuxin Chen

Electrical Engineering, Princeton University

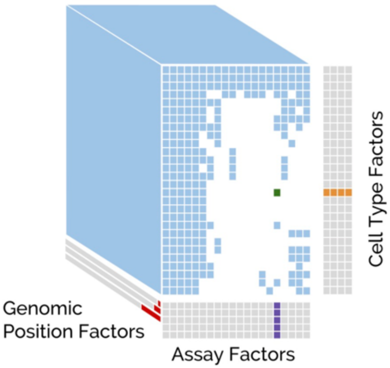


Changxiao Cai
Princeton EE

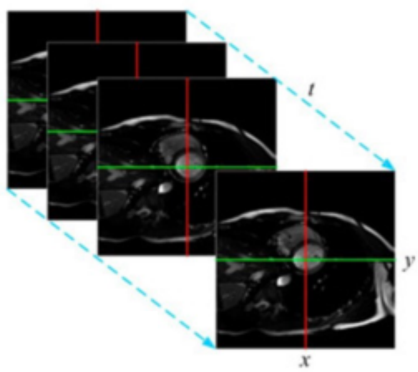


H. Vincent Poor
Princeton EE

Ubiquity of high-dimensional tensor data

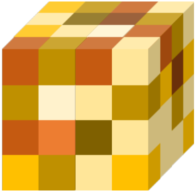


computational genomics
— *fig. credit: Schreiber et al. 19*



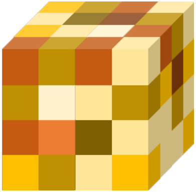
dynamic MRI
— *fig. credit: Liu et al. 17*

Challenges in tensor reconstruction

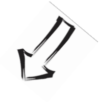


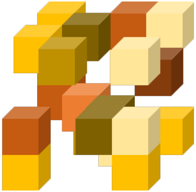
a tensor of interest

Challenges in tensor reconstruction



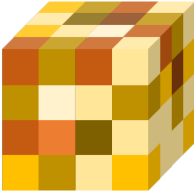
a tensor of interest



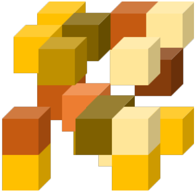


mising data

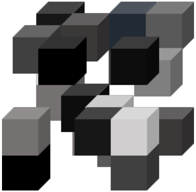
Challenges in tensor reconstruction



a tensor of interest

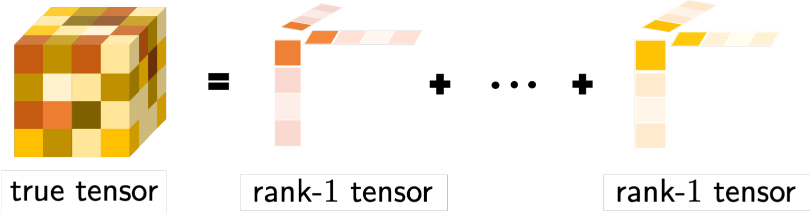


mising data



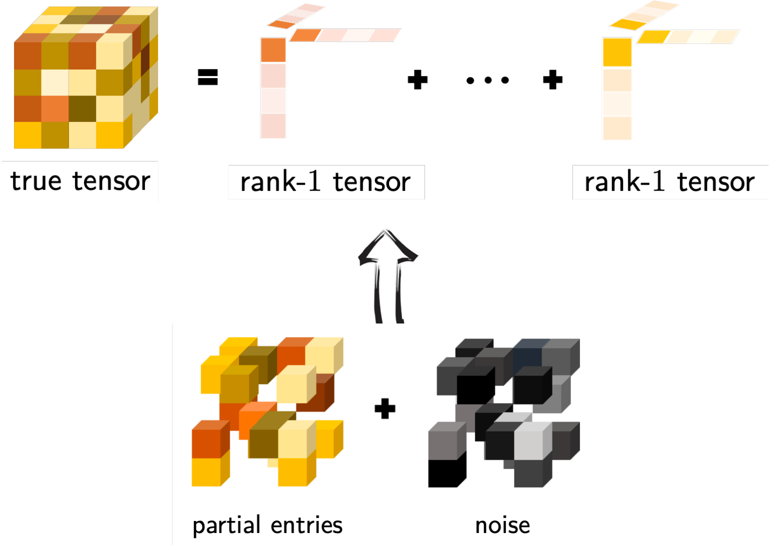
noise

Key to enabling reliable reconstruction from incomplete & noisy data:

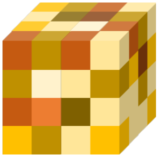


— **exploiting low (CP) rank structure**

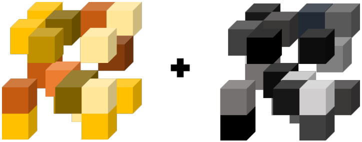
Noisy tensor completion



Mathematical model



$\boldsymbol{T}^{\star}$

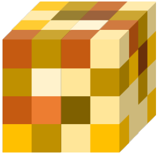


$\boldsymbol{T}^{\mathsf{obs}}$

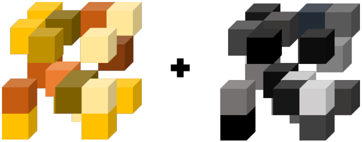
- unknown rank-r tensor $\boldsymbol{T}^{\star} \in \mathbb{R}^{d \times d \times d}$

$$\boldsymbol{T}^{\star} = \sum_{i=1}^{r} \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star}$$

Mathematical model



$\boldsymbol{T}^{\star}$



$\boldsymbol{T}^{\mathsf{obs}}$

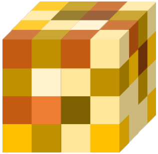
- unknown rank-r tensor $\boldsymbol{T}^{\star} \in \mathbb{R}^{d\times d\times d}$

$$\boldsymbol{T}^{\star} = \sum_{i=1}^{r} \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star}$$

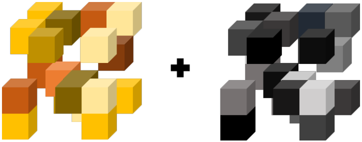
- partial observations over a sampling set Ω

$$T_{i,j,k}^{\mathsf{obs}} = T_{i,j,k}^{\star} + \text{noise}, \qquad (i,j,k) \in \Omega$$

Mathematical model



$\boldsymbol{T}^{\star}$



$\boldsymbol{T}^{\text{obs}}$

- unknown rank-r tensor $\boldsymbol{T}^{\star} \in \mathbb{R}^{d \times d \times d}$

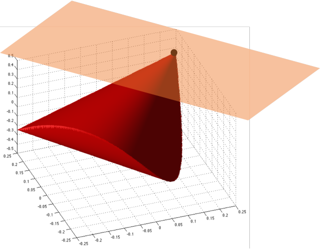
$$\boldsymbol{T}^{\star} = \sum_{i=1}^{r} \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star}$$

- partial observations over a sampling set Ω

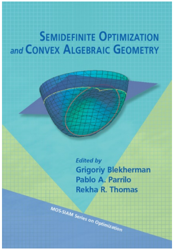
$$T_{i,j,k}^{\text{obs}} = T_{i,j,k}^{\star} + \text{noise}, \qquad (i,j,k) \in \Omega$$

- **goal:** estimate $\{\boldsymbol{u}_i^{\star}\}_{i=1}^{r}$ and $\boldsymbol{T}^{\star}$

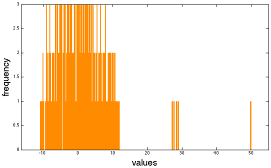
Prior art



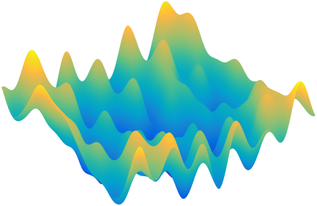
convex relaxation



sum-of-squares hierarchy



spectral methods



nonconvex optimization

Prior art

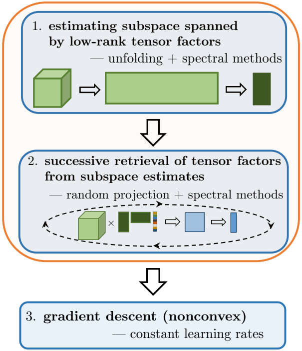
- Gandy, Recht, Yamada '11
- Liu, Musialski, Wonka, Ye '12
- Kressner, Steinlechner, Vandereycken '13
- Xu, Hao, Yin, Su '13
- Romera-Paredes, Pontil '13
- Jain, Oh '14
- Huang, Mu, Goldfarb, Wright '15
- Barak, Moitra '16
- Zhang, Aeron '16
- Yuan, Zhang '16
- Montanari, Sun '16
- Kasai, Mishra '16
- Potechin, Steurer '17
- Dong, Yuan, Zhang '17
- Xia, Yuan '19
- Zhang '19
- **Cai, Li, Poor, Chen '19**
- Cai, Li, Chi, Poor, Chen '19
- Liu, Moitra '20
- ...

A nonconvex approach: Cai et al. (NeurIPS 19)

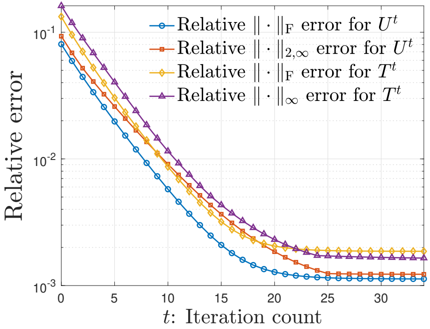
$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega} \Big\{ \Big(\sum\nolimits_{s=1}^{r} \boldsymbol{u}_s^{\otimes 3} \Big)_{i,j,k} - T_{i,j,k}^{\mathsf{obs}} \Big\}^2}_{\text{squared loss}}$$

- proper initializaiton: $\boldsymbol{U}^0$
- gradient descent: for $t = 0, 1, \cdots$

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta_t \nabla f(\boldsymbol{U}^t)$$



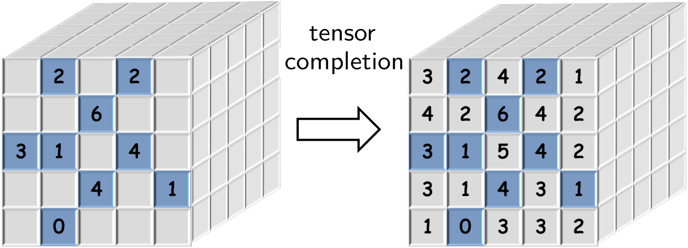
A nonconvex approach: Cai et al. (NeurIPS 19)



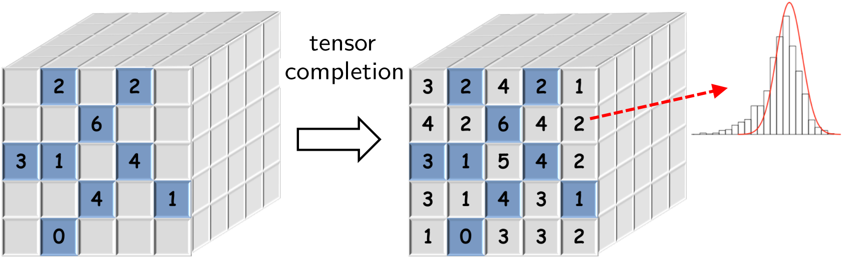
Under mild conditions, this nonconvex algorithm achieves

- linear convergence
- minimax-optimal statistical accuracy (up to log factor)

One step further: reasoning about uncertainty?



One step further: reasoning about uncertainty?



How to assess uncertainty, or "confidence", of obtained estimates due to imperfect data acquisition?

- noise
- incomplete measurements
- $\cdots$

Challenges

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega} \Big\{ \Big(\sum\nolimits_{s=1}^{r} \boldsymbol{u}_s^{\otimes 3} \Big)_{i,j,k} - T_{i,j,k}^{\mathsf{obs}} \Big\}^2}_{\text{squared loss}}$$

- how to pin down distributions of nonconvex solutions?

Challenges

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega}\left\{\left(\sum\nolimits_{s=1}^{r}\boldsymbol{u}_s^{\otimes 3}\right)_{i,j,k} - T_{i,j,k}^{\mathsf{obs}}\right\}^2}_{\text{squared loss}}$$

- how to pin down distributions of nonconvex solutions?
- how to adapt to unknown noise distributions and heteroscedasticity (i.e. location-varying noise variance)?

Challenges

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega} \Big\{ \Big(\sum\nolimits_{s=1}^{r} \boldsymbol{u}_s^{\otimes 3} \Big)_{i,j,k} - T_{i,j,k}^{\mathsf{obs}} \Big\}^2}_{\text{squared loss}}$$

- how to pin down distributions of nonconvex solutions?
- how to adapt to unknown noise distributions and heteroscedasticity (i.e. location-varying noise variance)?
- existing estimation guarantees are highly insufficient
  $\longrightarrow$ overly wide confidence intervals

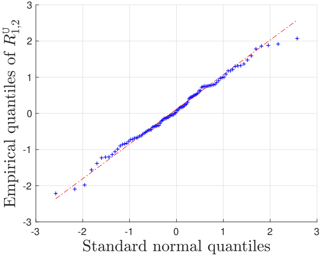
Assumptions

$$\boldsymbol{T}^\star = \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \in \mathbb{R}^{d\times d\times d}$$

- **random sampling**: each entry is observed independently with prob. $p \gtrsim \frac{\mathrm{polylog}(d)}{d^{3/2}}$
- **random noise**: independent zero-mean sub-Gaussian with variance of roughly the same order (but not identical)
- **ground truth**: low-rank ($r = O(1)$), incoherent (tensor factors are de-localized and nearly orthogonal to each other), and well-conditioned

Main results: distributional theory

- random sampling
- independent sub-Gaussian noise
- ground truth: low-rank, incoherent, well-conditioned



Theorem 1

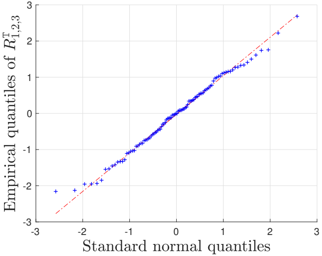
With high prob., there exists permutation matrix $\mathbf{\Pi} \in \mathbb{R}^{r \times r}$ s.t.

$$\boldsymbol{U}\mathbf{\Pi} - \boldsymbol{U}^{\star} \sim \mathcal{N}(\mathbf{0}, \textit{Cramér-Rao}) + \textit{negligible term}$$

— *asymptotically optimal*

Main results: distributional theory

- random sampling
- independent sub-Gaussian noise
- ground truth: low-rank, incoherent, well-conditioned



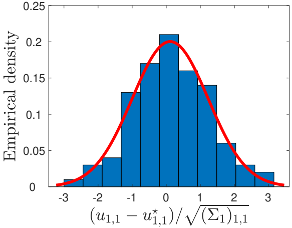
Theorem 2

Consider any (i, j, k) s.t. the corresponding "SNR" is not exceedingly small. Then with high prob.,

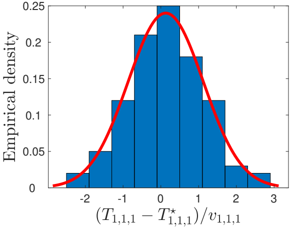
$$T_{i,j,k} - T^{\star}_{i,j,k} \sim \mathcal{N}(0, \textit{Cramér-Rao}) + \textit{negligible term}$$

— *asymptotically optimal*

- **Gaussianality and optimality:** estimation error of nonconvex approach is zero-mean Gaussian, who (co)-variance is “minimal”

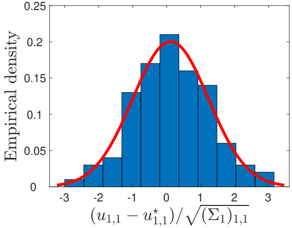


tensor factor entry

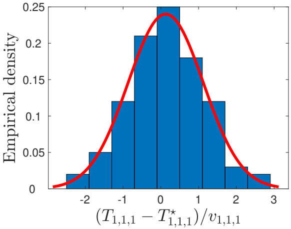


tensor entry

- **Gaussianality and optimality:** estimation error of nonconvex approach is zero-mean Gaussian, who (co)-variance is "minimal"
- **Confidence intervals:** error (co)-variance can be accurately estimated, leading to valid CI construction



tensor factor entry



tensor entry

- **Gaussianality and optimality:** estimation error of nonconvex approach is zero-mean Gaussian, who (co)-variance is "minimal"
- **Confidence intervals:** error (co)-variance can be accurately estimated, leading to valid CI construction
- **Adaptivity:** our procedure is data-driven, fully adaptive to unknown noise levels and heteroscedasticity

Empirical coverage rates (CR)

tensor factor

(r,σ)	Mean(CR)	Std(CR)
$(2,10^{-2})$	0.9481	0.0201
$(2,10^{-1})$	0.9477	0.0228
$(2,1)$	0.9478	0.0215
$(4,10^{-2})$	0.9450	0.0218
$(4,10^{-1})$	0.9472	0.0231
$(4,1)$	0.9462	0.0234

tensor entries

(r,σ)	Mean(CR)	Std(CR)
$(2,10^{-2})$	0.9494	0.0218
$(2,10^{-1})$	0.9513	0.0218
$(2,1)$	0.9475	0.0222
$(4,10^{-2})$	0.9434	0.0225
$(4,10^{-1})$	0.9494	0.0220
$(4,1)$	0.9494	0.0219

$d=100,\ p=0.2,$ heteroscedastic

Back to estimation: ℓ_2 optimality

Distributional theory in turn allows us to track estimation accuracy

Back to estimation: ℓ_2 optimality

Distributional theory in turn allows us to track estimation accuracy

Theorem 3

Suppose noise is i.i.d. Gaussian. $\exists$ *some permutation* $\pi(\cdot)$ *s.t.*

$$\|\boldsymbol{u}_{\pi(l)} - \boldsymbol{u}_l^\star\|_2^2 = \underbrace{\frac{(2+o(1))\sigma^2 d}{p\|\boldsymbol{u}_l^\star\|_2^4}}_{\text{Cram er-Rao lower bound}}, \quad 1 \le l \le r$$

$$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_{\mathrm{F}}^2 = \underbrace{\frac{(6+o(1))\sigma^2 rd}{p}}_{\text{Cramér-Rao lower bound}}$$

Back to estimation: ℓ_2 optimality

Distributional theory in turn allows us to track estimation accuracy

Theorem 3

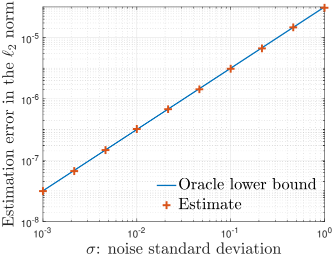
Suppose noise is i.i.d. Gaussian. $\exists$ some permutation $\pi(\cdot)$ s.t.

$$\|\boldsymbol{u}_{\pi(l)} - \boldsymbol{u}_l^\star\|_2^2 = \underbrace{\frac{(2+o(1))\sigma^2 d}{p\|\boldsymbol{u}_l^\star\|_2^4}}_{\text{Cram er-Rao lower bound}}, \quad 1 \le l \le r$$

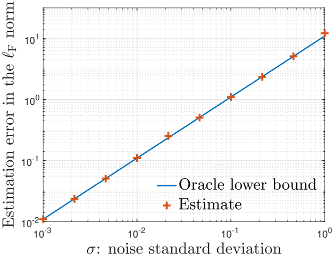
$$\|\boldsymbol{T} - \boldsymbol{T}^\star\|_{\mathrm{F}}^2 = \underbrace{\frac{(6+o(1))\sigma^2 r d}{p}}_{\text{Cramér-Rao lower bound}}$$

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

Numerical ℓ_2 errors vs. Cramér–Rao bounds



tensor factor estimation



tensor estimation

$$r = 4,\ p = 0.2,\ d = 100$$

Concluding remarks

nonconvex optimization

optimal estimation guarantees

fast, adaptive to unknown noise levels

optimal uncertainty quantification

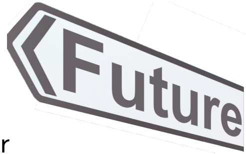
Concluding remarks

nonconvex optimization

optimal estimation guarantees

fast, adaptive to unknown noise levels

optimal uncertainty quantification



future directions

- improve dependency on rank & cond. number
- more general sampling patterns
- other tensor-type problems

Paper:

"Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality," C. Cai, H. V. Poor, Y. Chen, ICML 2020

Other related papers:

"Nonconvex low-rank symmetric tensor completion from noisy data," C. Cai, G. Li, H. V. Poor, Y. Chen, NeurIPS 2019

"Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees," C. Cai, G. Li, Y. Chi, H. V. Poor, Y. Chen, accepted to Annals of Statistics, 2019