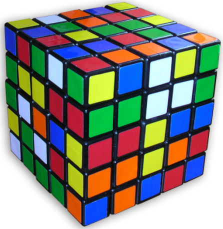
Low Rank Tensor Completion

Ming Yuan

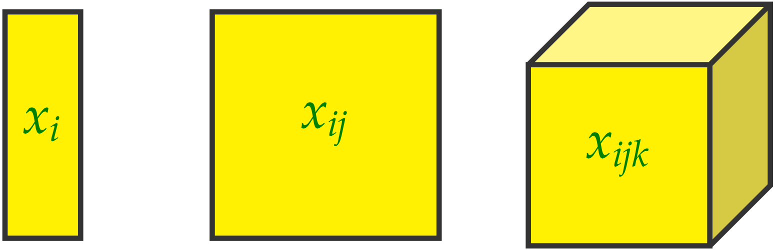
Department of Statistics
Columbia University

ming.yuan@columbia.edu
http://www.columbia.edu/~my2550



(Joint work with Dong Xia and Cun-hui Zhang)

Data in the form of multilinear array



- Spatio-temporal expression data
  [e.g. Kang et al. (2011); Parikshak et al. (2013); Hawrylycz et al. (2015)]
- Imaging (video) data – 3D images, hyper-spectral, and etc.
  [e.g. Liu et al. (2009); Li and Li (2010); Gandy et al. (2011); Semerci et al. (2014)]
- Relational data, recommender system, text mining and etc.
  [e.g. Cohen and Collins (2012); Dunlavy et al. (2013); Barak and Moitra (2016)]
- Latent variable models – topic models, phylogenetic tree, and etc.
  [e.g. Anandkumar et al. (2014)]

$\vdots$

New Challenges

- Algebraic: best low rank approximation may not exist!
  [e.g. de Silva and Lim (2008)]
- Computational: most computations are NP hard!
  [e.g. Hillar and Lim (2013)]
- Probabilistic: different concentration behavior.
  [e.g. Y. and Zhang (2016)]

This talk: implications in *tensor completion*

Overview

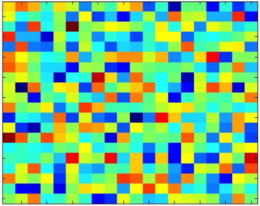
Overview

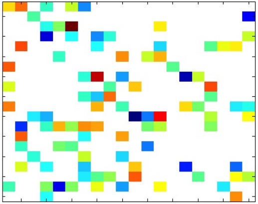
Tensor Completion

- Interested in $\boldsymbol{T} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$
  - $\boldsymbol{T}$ is of high dimension (d_js large)
  - $\boldsymbol{T}$ is (approximately) low rank
- Partial observations:

$$Y_i = T(\omega_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

  To fix ideas:
  - Cubic tensors – $d_1 = \cdots = d_k =: d$
  - ω_is are independently and uniformly sampled
- Our goal:
  - without noise – *exact recovery*
  - with noise – *rates of convergence*





Matrix Completion ($k = 2$)

- Also known as *Netflix problem*
- Incoherence: every observation carries similar amount of information

$$\mu(U) = \frac{d}{r} \max_{1 \le i \le d} \|\boldsymbol{P}_U \boldsymbol{e}_i\|^2$$

- Without measurement error – nuclear norm minimization:

$$\min_M \|M\|_* \qquad \text{subject to } M(\omega_i) = T(\omega_i) \quad \forall i$$

Exact recovery if:

$$n \gg rd \cdot \log(d)$$

- With measurement error – nuclear norm regularization:

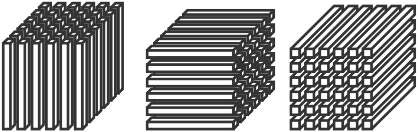
$$\widehat{T} = \arg\min_M \left\{ \frac{1}{n} \sum_{i=1}^{n} [M(\omega_i) - Y_i]^2 + \lambda \|M\|_* \right\}$$

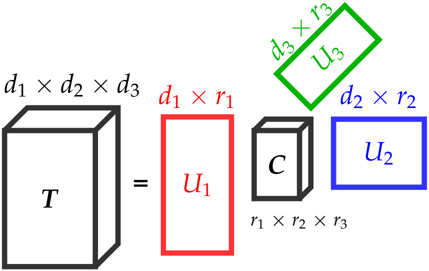
Estimation error:

$$\mathrm{MSE}(\widehat{T}) := \frac{1}{d^2} \|\widehat{T} - T\|_{\mathrm{F}}^2 \lesssim rd \cdot \log(d)/n$$

[e.g., Candes and Recht (2008); Keshavan et al. (2009); Candes and Tao (2010); Gross (2011); Negahban and Wainwright (2011); Recht (2011); Rohde and Tsybakov (2011); Koltchinskii et al. (2012)]

Multilinear Ranks





- Fibers – vectors obtained by fixing two indices
  i.e. mode-1 fibers: $T(:, i_2, \ldots, i_k)$
- Linear space spanned by fibers, i.e. $\mathcal{L}_1(\boldsymbol{T}) = \text{l.s.}\{T(:, i_2, \ldots, i_k) : i_2, \ldots, i_k\}$
- Multilinear ranks

$$r_j = \dim(\mathcal{L}_j(\boldsymbol{T}))$$

- Tucker decomposition

$$\boldsymbol{T} = (U_1, \ldots, U_k) \cdot \boldsymbol{C}$$

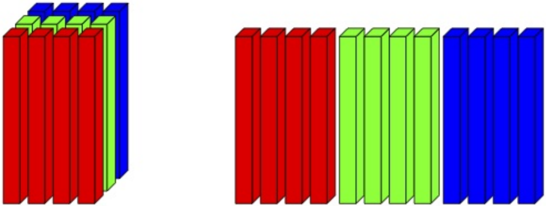
Geometry of Low Rank Tensors

$$\mathcal{A}(r) = \{\boldsymbol{T} \in \mathbb{R}^{d\times\cdots\times d} : r_j(\boldsymbol{T}) \le r\} \cong \underbrace{\mathcal{G}(d,r)\times\cdots\times\mathcal{G}(d,r)}_{k\ \textit{times}} \times \mathbb{R}^{r\times\cdots\times r}$$

- Assume same multilinear ranks ($r_1 = \cdots = r_3 =: r$) for brevity
- Any $\boldsymbol{T} \in \mathcal{A}(r)$ – $\mathrm{rank}(\boldsymbol{T}) \in [r, r^{k-1}]$
- Dimension of $\mathcal{A}(r)$ – $O(r^k + rd)$
- Gold standard:
  - Exact recovery with $\tilde{O}(r^k + rd)$ noiseless entries
  - Estimation error of the order $\tilde{O}_p((r^k + rd)/n)$
- *For matrices*: similar bounds are attainable with nuclear norm minimization/regularization

Overview

Matricization



- Tensor unfolding (matricization)

$$\boldsymbol{T} \in \mathbb{R}^{d\times d\times d} \mapsto \mathcal{M}_1(\boldsymbol{T}) \in \mathbb{R}^{d\times d^2}$$

- Nuclear norm minimization

$$\min_{\boldsymbol{A}\in\mathbb{R}^{d\times d\times d}} \sum_{j=1}^{3} \|\mathcal{M}_j(\boldsymbol{A})\|_* \qquad \text{subject to } A(\omega_i) = T(\omega_i),\ \forall i$$

- Sample size requirement

$$|\Omega| \gg rd^2\text{polylog}(d)$$

[e.g., Signoretto et al. (2010), Tomioka et al. (2010, 2011), Gandy et al. (2011)]

$$\min_{\boldsymbol{A}\in\mathbb{R}^{d\times d\times d}} \|\boldsymbol{A}\|_* \qquad \text{subject to } A(\omega_i) = T(\omega_i),\ \forall i$$

- Tensor nuclear norm
  - Spectral norm

$$\|\boldsymbol{A}\| = \max_{\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=\|\boldsymbol{w}\|=1} \langle \boldsymbol{A}, \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}\rangle$$

  - Nuclear norm

$$\|\boldsymbol{A}\|_* = \max_{\boldsymbol{Y}\in\mathbb{R}^{d\times d\times d}:\|\boldsymbol{Y}\|\le 1} \langle \boldsymbol{Y}, \boldsymbol{A}\rangle.$$

- Exact recovery with high probability if

$$n \gg (r^{1/2}d^{3/2} + r^2 d)\mathrm{polylog}(d)$$

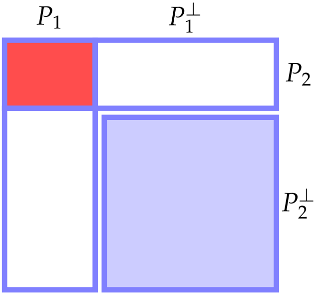
How to get here

Find a dual certificate:

- in the subdifferential of $\partial\|\cdot\|_*(\boldsymbol{T})$
  - characterizing sub-differential of tensor nuclear norm
- supported on $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$
  - concentration inequalities for tensor martingales

If $M = UDV^\top$, then

$$\left.\begin{array}{l} W = \boldsymbol{P}_U^\perp W \boldsymbol{P}_V^\perp \\ \|W\| \leq 1 \end{array}\right\} \Longrightarrow \|X\|_* \geq \|M\|_* + \langle UV^\top + W, X - M\rangle$$



$$\begin{aligned} \mathcal{P}_M^0 &= \boldsymbol{P}_U \otimes \boldsymbol{P}_V \\ \mathcal{P}_M^1 &= \boldsymbol{P}_U \otimes \boldsymbol{P}_V^\perp \\ \mathcal{P}_M^2 &= \boldsymbol{P}_U^\perp \otimes \boldsymbol{P}_V \\ \mathcal{P}_M^\perp &= \boldsymbol{P}_U^\perp \otimes \boldsymbol{P}_V^\perp \end{aligned}$$

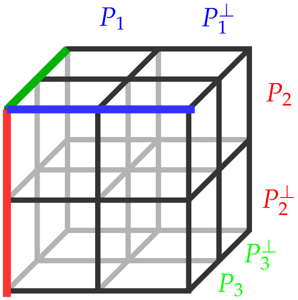
[Watson, 1992]

Projection

- Projection

$$P_1 \otimes P_2 \otimes P_3[A, B, C] = [P_1 A, P_2 B, P_3 C]$$

- Decomposition of space



$$\begin{aligned}
\mathcal{Q}_T^0 &= P_T^1 \otimes P_T^2 \otimes P_T^3 \\
\mathcal{Q}_T^1 &= P_{T^\perp}^1 \otimes P_T^2 \otimes P_T^3 \\
\mathcal{Q}_T^2 &= P_T^1 \otimes P_{T^\perp}^2 \otimes P_T^3 \\
\mathcal{Q}_T^3 &= P_T^1 \otimes P_T^2 \otimes P_{T^\perp}^3
\end{aligned}$$

$$\begin{aligned}
\mathcal{Q}_{T^\perp}^0 &= P_{T^\perp}^1 \otimes P_{T^\perp}^2 \otimes P_{T^\perp}^3 \\
\mathcal{Q}_{T^\perp}^1 &= P_T^1 \otimes P_{T^\perp}^2 \otimes P_{T^\perp}^3 \\
\mathcal{Q}_{T^\perp}^2 &= P_{T^\perp}^1 \otimes P_T^2 \otimes P_{T^\perp}^3 \\
\mathcal{Q}_{T^\perp}^3 &= P_{T^\perp}^1 \otimes P_{T^\perp}^2 \otimes P_T^3
\end{aligned}$$

Characterization of Subdifferential

$$\left.\begin{array}{l} \boldsymbol{W} = \mathcal{Q}_{\boldsymbol{T}}^{0}\boldsymbol{W} \text{ and } \langle \boldsymbol{T}, \boldsymbol{W}\rangle = \|\boldsymbol{T}\|_* \\ \boldsymbol{W}^\perp = \mathcal{Q}_{\boldsymbol{T}}^{\perp}\boldsymbol{W}^\perp \text{ and } \|\boldsymbol{W}^\perp\| \le \frac{1}{2} \end{array}\right\} \Longrightarrow \|\boldsymbol{X}\|_* \ge \|\boldsymbol{T}\|_* + \langle \boldsymbol{W} + \boldsymbol{W}^\perp, \boldsymbol{X} - \boldsymbol{T}\rangle$$

- More complex geometry than matrix case
- For matrices $\|\boldsymbol{W}^\perp\| \le 1$ is sufficient and necessary
- For tensor $\|\boldsymbol{W}^\perp\| \le 1/2$ is only sufficient, *not* necessary:

$$\partial\|\cdot\|_*(\boldsymbol{T}) \supseteq \left\{\boldsymbol{W} + \boldsymbol{W}^\perp : \|\boldsymbol{W}^\perp\| \le 1/2\right\}.$$

$*$ For general kth order tensors, upper bound 1, lower bound $2/k(k-1)$.

Concentration of Tensor Martingale

With probability at least $1 - e^{-t}$,

$$\left\| \frac{d^3}{n} \sum_{i=1}^{n} A(\omega_i)\boldsymbol{e}_i - \boldsymbol{A} \right\| \lesssim \|\boldsymbol{A}\|_{\max} \cdot \left(\sqrt{\frac{d^4 t}{n}} + \frac{d^3 t}{n} \right) \cdot \mathrm{polylog}(d).$$

- Contributions from variance and maximum
  - in matrix or vector case – variance dominates
  - for higher order tensors – maximum dominates
- If A is *incoherent*, then $\|A\|_{\max} = O(d^{-3/2})$. Thus

$$n \gg d^{3/2}\mathrm{polylog}(d) \text{ implies } \left\| \frac{d^3}{n} \sum_{i=1}^{n} A(\omega_i)\boldsymbol{e}_i - \boldsymbol{A} \right\| \leq \frac{1}{2}$$

$*$ For general kth order tensors, second term $d^k t/n$.

What happens when $k > 3$?

- Following a similar argument leads to sample size requirement

$$n \gtrsim d^{k/2} \mathrm{polylog}(d).$$

- It can be improved if we incorporate incoherence more explicitly:

$$n \gtrsim (r^{(k-1)/2} d^{3/2} + r^{k-1} d)(\log(d))^2$$

  - Depends on the order k only through the rank r
  - If $r = O(1)$, then the sample size requirement becomes

$$n \gtrsim d^{3/2} (\log(d))^2$$

Incoherent Nuclear Norm Minimization

$$\min_{\boldsymbol{A}\in\mathbb{R}^{d_1\times\cdots\times d_k}} \|\boldsymbol{X}\|_{\star,\boldsymbol{\delta}} \qquad \text{subject to } A(\omega_i) = T(\omega_i) \quad i = 1,\ldots,n$$

- Incoherent rank-one tensors: $\mathscr{U}(\boldsymbol{\delta}) = \cup_{1\le j_1<j_2\le k}\mathscr{U}_{j_1j_2}(\boldsymbol{\delta})$ where

$$\mathscr{U}_{j_1j_2}(\boldsymbol{\delta}) = \left\{\boldsymbol{u}_1\otimes\cdots\otimes\boldsymbol{u}_k : \|\boldsymbol{u}_j\|_{\ell_2}\le 1,\ \forall j; \|\boldsymbol{u}_j\|_{\ell_\infty}\le\delta_j, \forall j\ne j_1,j_2\right\}$$

- Incoherent tensor norms:

$$\|\boldsymbol{X}\|_{\circ,\boldsymbol{\delta}} = \sup_{\boldsymbol{Y}\in\mathscr{U}(\boldsymbol{\delta})}\langle\boldsymbol{Y},\boldsymbol{X}\rangle, \qquad \|\boldsymbol{X}\|_{\star,\boldsymbol{\delta}} = \sup_{\|\boldsymbol{Y}\|_{\circ,\boldsymbol{\delta}}\le 1}\langle\boldsymbol{Y},\boldsymbol{X}\rangle$$

- Encourages solution to be incoherent:
  - In general, $\|\boldsymbol{X}\|_* \le \|\boldsymbol{X}\|_{\star,\boldsymbol{\delta}}$
  - If $\delta_j \ge \mu_j(\boldsymbol{X})$, then $\|\boldsymbol{X}\|_* = \|\boldsymbol{X}\|_{\star,\boldsymbol{\delta}}$

Overview

Polynomial-time Methods?

- Gold standard is $\tilde{O}(r^3 + rd)$
- Matricization requires $\tilde{O}(rd^2)$
- Nuclear norm minimization needs $\tilde{O}(r^{1/2}d^{3/2} + r^2 d)$
- *But* tensor nuclear norm is NP hard to compute in the worst case
  - Relaxation – theta norm, sum of squares relaxation
  - Feasible in principle but do not scale well
  - General performance guarantee unclear
- What about nonconvex methods?
  - Success in some practical examples
  - General performance guarantee unclear

Do they work?

$$\min_{A \in \mathcal{A}(r)} \frac{1}{n} \sum_{i=1}^{n} [T(\omega_i) - A(\omega_i)]^2$$

- Recall that we can write

$$A = (W_1, W_2, W_3) \cdot \boldsymbol{G}$$

If $n \gg r^3 \text{polylog}(d)$, the above minimization can be equivalently expressed as

$$\min_{W_1, W_2, W_3 \in \mathcal{G}(d,r)} f(W_1, W_2, W_3)$$

- Smooth optimization techniques to minimize f – practical successes [see, e.g., Vervliet et al., 2014; Kressner et al., 2014]
- But *why*?

If

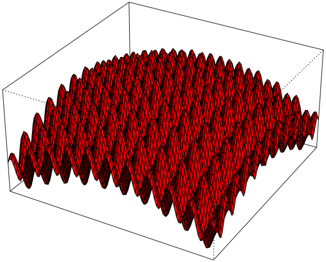
$$n \gg \left(r^{5/2} d^{3/2} + r^4 d\right) \cdot \text{polylog}(d)$$

then f is well-behaved in an *incoherent* neighborhood around truth

What exactly is the problem?

Suppose we want to compute the spectral norm

- of a random matrix – at most d local optima
- of a random tensor – $\exp(\Omega(d))$ local optima [Auffinger and Ben Arous (2013)]



- Polynomial optimization
- Very smooth but highly nonconvex
- *If* we can get close to global optimum, …

Pay a hefty price to get close, pay a little more to get exact!

Initialization

- f is minimized at the linear subspace of $\mathcal{L}_1(\boldsymbol{T})$, $\mathcal{L}_2(\boldsymbol{T})$ and $\mathcal{L}_3(\boldsymbol{T})$
- A first attempt: random initialization – exponentially many tries
- A second attempt: spectral method
  - $\mathcal{L}_1(\boldsymbol{T})$ is the column space of $\mathcal{M}_1(\boldsymbol{T}) \in \mathbb{R}^{d \times d^2}$
  - An unbiased estimate of $\mathcal{M}_1(\boldsymbol{T})$ is

$$\mathcal{M}_1(\widehat{\boldsymbol{T}}) := \mathcal{M}_1\left(\frac{d^3}{n}\sum_{i=1}^{n} T(\omega_i)\boldsymbol{e}_{\omega_i}\right)$$

  - Estimate $\mathcal{L}_1(\boldsymbol{T})$ by applying SVD to the above estimate
  - $n \gg d^2$ to ensure closeness
- A third attempt: "second order" spectral method

Second order spectral method

- $\mathcal{L}_1(\boldsymbol{T})$ is also the eigen-space of $\boldsymbol{S} := \mathcal{M}_1(\boldsymbol{T})\mathcal{M}_1(\boldsymbol{T})^\top \in \mathbb{R}^{d\times d}$
- $\mathcal{M}_1(\widehat{\boldsymbol{T}})\mathcal{M}_1(\widehat{\boldsymbol{T}})^\top$ is a *biased* estimate
- Unbiased estimate – U-statistic

$$\widehat{\boldsymbol{S}} := \frac{d^6}{n(n-1)} \sum_{i\neq j} T(\omega_i)T(\omega_j)\mathcal{M}_1(\boldsymbol{e}_{\omega_i})\mathcal{M}_1(\boldsymbol{e}_{\omega_j})^\top$$

- with probability at least $1-e^{-t}$,

$$\|\widehat{\boldsymbol{S}}-\boldsymbol{S}\| \lesssim \|\boldsymbol{T}\|_{\max}^2 \cdot \left(\frac{d^6t^2}{n^2} + \frac{d^{9/2}t}{n}\right) \cdot \mathrm{polylog}(d)$$

  - For incoherent $\boldsymbol{T}$, $\|\widehat{\boldsymbol{S}}-\boldsymbol{S}\| = o_p(1)$ if $n \gg d^{3/2}\mathrm{polylog(d)}$
  - Sharper concentration around $\boldsymbol{S}$ than $\mathcal{M}_1(\widehat{\boldsymbol{T}})$ around $\mathcal{M}_1(\boldsymbol{T})$
- Consistent estimates iff

$$n \gg \left(rd^{3/2} + r^2 d\right)\log d.$$

Effect of Noise

$$Y_i = T(\omega_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

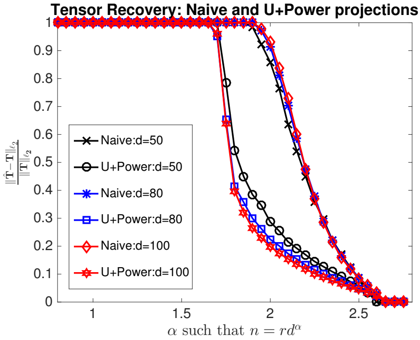
- More difficult – control the effect of noise
- But easier – suffices to get close to the target
- And subtlety – scaling

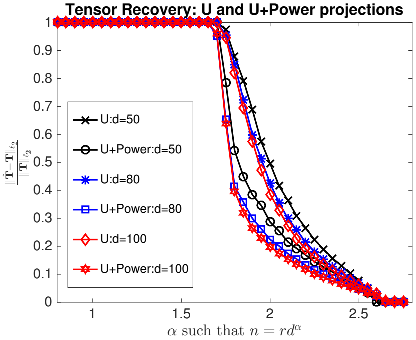
$$\inf_{\widehat{\boldsymbol{T}}} \sup_{\boldsymbol{T} \in \Theta(r_0, \beta_0)} \left(\frac{1}{d^3} \|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_{\ell_p}^p \right)^{1/p} \asymp (\|\boldsymbol{T}\|_{\ell_\infty} + \sigma_\varepsilon) \sqrt{\frac{rd\log(d)}{n}},$$

provided that

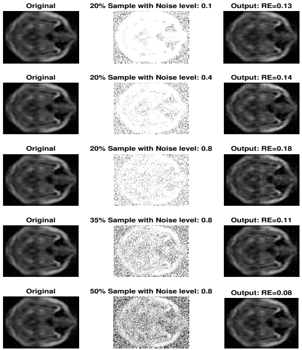
$$n \gg \left(rd^{3/2} + r^2 d \right) \log d.$$

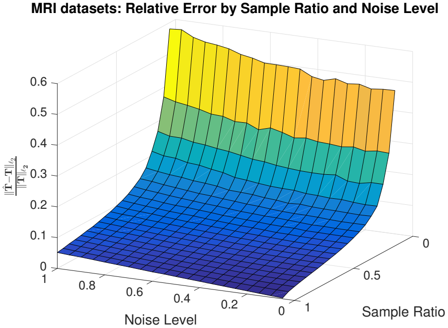
Effect of Initialization and Power Iteration





A Data* Example





∗ Taken from BrainWeb [Cocosco et al., 1997; 217 × 181 × 181].

Concluding Remarks

Methods	Tractable?	Sample size requirement
Matricization	Yes	$\tilde{O}(rd^2)$
Nuclear Norm minimization	No	$\tilde{O}(r^{1/2}d^{3/2} + r^2 d)$
Nonconvex	Yes	$\tilde{O}(rd^{3/2} + r^2 d)$

- Polynomial-time methods with better dependence on d? *Possibly no.*
  - "Equivalence" to random k-SAT problem [Barak and Moitra (2016)]
- Polynomial-time methods with better dependence on r? *I don't know.*
- Is it worth the while? *Definitely yes!*
  - Regression [Chen, Raskutti and Y. (2015, 2016)]
  - PCA [Liu, Y. and Zhao (2017)]

$\vdots$