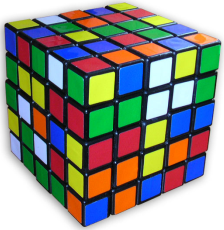
Low Rank Tensor Methods in High Dimensional Data Analysis

Ming Yuan

Department of Statistics
Columbia University

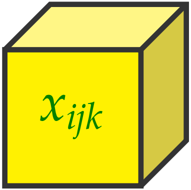
ming.yuan@columbia.edu
http://www.columbia.edu/~my2550



Tensor Everywhere

- Computational biology [Cartwright et al. 2009]
- Computer graphics [Vasilescu and Terzopoulos 2004]
- Computer vision [Shashua and Hazan 2005]
- Neuroimaging [Schultz and Seidel 2008]
- Pattern recognition [Vasilescu 2002]
- Phylogenetics [Allman and Rhodes 2008]
- Quantum computing [Miyake and Wadati2002]
- Scientific computing [Beylkin and Mohlenkamp 1997]
- Signal processing [Comon 1994; 2004]
- Spectroscopy [Smilde et al. 2004]
- Wireless communication [Sidiropoulos et al. 2000]
- ……

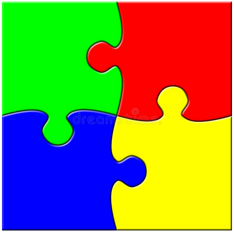
Matrix



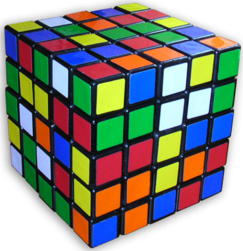
Higher order tensor

High dimensional data analysis

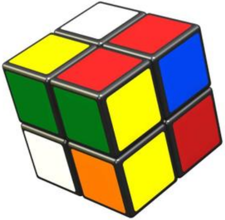
New Challenges



- Conceptual
- Computational
- ……



Ranks, eigenvalues, singular value decomposition



"Most tensor problems are NP hard" [Hillar and Lim, 2013]

Outline

My focus – effect of randomness and high dimensionality

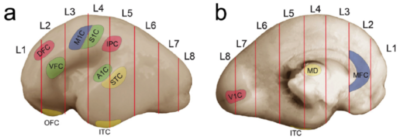
- How they change the nature of the problem
- How they may influence our approach to the problem

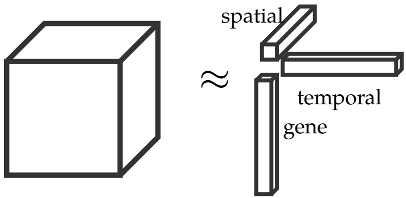
More specifically,

- *Tensor PCA* [Based on joint work with Tianqi Liu and Hongyu Zhao]
- Tensor Completion
- Tensor Sparsification
- Tensor Regression
- ……

Spatio-Temporal Transcriptome of the Brain

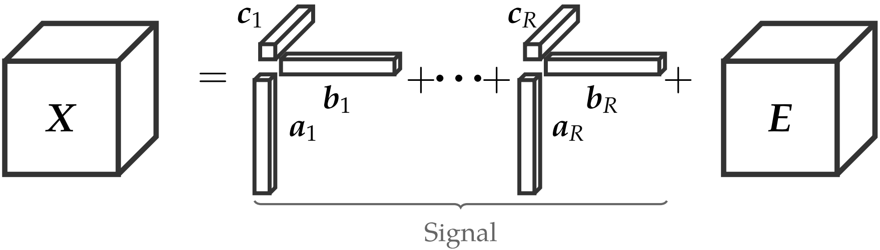
Period	Description	Age
1	Embryonic	4PCW≤Age<8PCW
2	Early fetal	8PCW≤Age<10PCW
3	Early fetal	10PCW≤Age<13PCW
4	Early mid-fetal	13PCW≤Age<16PCW
5	Early mid-fetal	16PCW≤Age<19PCW
6	Late mid-fetal	19PCW≤Age<24PCW
7	Late fetal	24PCW≤Age<38PCW
8	Neonatal and early infancy	0M(birth)≤Age<6M
9	Late infancy	6M≤Age<12M
10	Early childhood	1Y≤Age<6Y
11	Middle and late childhood	6Y≤Age<12Y
12	Adolescence	12Y≤Age<20Y
13	Young adulthood	20Y≤Age<40Y
14	Middle adulthood	40Y≤Age<60Y
15	Late adulthood	60Y≤Age





[Kang et al., 2011]

Low Rank Tensor Approximation



- Alternating least squares [e.g., Kolda and Bader, 2009]
- Smooth optimization methods – Newton methods [e.g., Zhang and Golub, 2011]
- Semidefinite relaxation – sum of squares hierarchy [e.g., Nie and Wang, 2014]
- Spectral – Higher order SVD [De Lathauwer, De Moor and Vandewalle, 2000]
- ……

Tensor PCA

Unexpected difficulties

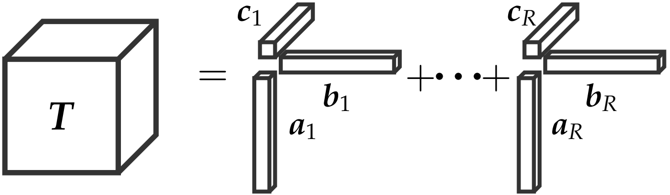
- Different methods yield different results
- It is *ill posed* for a nontrivial set of tensors [de Silva and Lim, 2008]
- It is *NP hard* in general [Hillar and Lim, 2013]

What do we do?

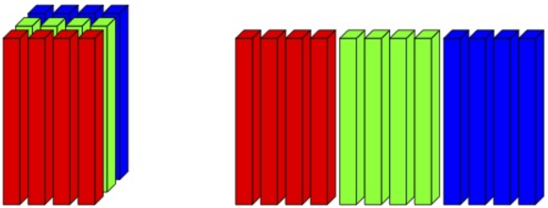
- *Interpretability* – Is it really worth the while?
- *Estimability* – Is it possible to extract the signal?
- *Feasibility* – Is it computable for large-scale data?

Just Another PCA?

Tensor PCA:



Classical PCA:



Interpretability, Interpretability, Interpretability!

- Identifiability – up to permutation and scaling
- Kruskal rank – largest k such that any k column vectors are linear independent

The *PCs* $\boldsymbol{a}_j$s, $\boldsymbol{b}_j$s, and $\boldsymbol{c}_j$s are identifiable if

$$\kappa_{\boldsymbol{a}} + \kappa_{\boldsymbol{b}} + \kappa_{\boldsymbol{c}} \geq 2R + 2$$

[Kruskal, 1977]

Estimability?

- Consider a simple signal+noise models

$$\boldsymbol{X} = \boldsymbol{T} + \boldsymbol{E}, \qquad e_{ijk} \sim_{\text{iid}} N(0,1)$$

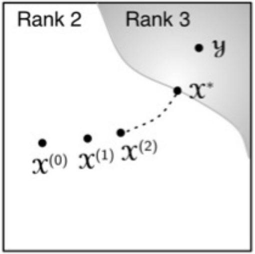
- MLE – best low rank approximation:

$$\widehat{\boldsymbol{T}}^{\text{MLE}} = \underset{\text{rank}(\boldsymbol{A}) \le R}{\arg\min} \|\boldsymbol{X} - \boldsymbol{A}\|_{\text{F}}^2$$

- ~~*How well*~~ does it work – Best low rank approximation may not exist

$$n\left(\boldsymbol{u} + \frac{1}{n}\boldsymbol{v}\right) \otimes \left(\boldsymbol{u} + \frac{1}{n}\boldsymbol{v}\right) \otimes \left(\boldsymbol{u} + \frac{1}{n}\boldsymbol{v}\right) - n\boldsymbol{u} \otimes \boldsymbol{u} \otimes \boldsymbol{u}$$

$$\to \boldsymbol{u} \otimes \boldsymbol{u} \otimes \boldsymbol{v} + \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{u} + \boldsymbol{v} \otimes \boldsymbol{u} \otimes \boldsymbol{u}$$



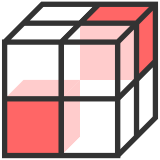
[Kolda and Bader, 2009]

Universal Unestimability

- Signal: $T \in \mathbb{R}^{2\times2\times2}$ such that $\mathrm{rank}(T) = 2$. For *any* such T

$$\pi = \mathbb{P}\left\{\widehat{T}^{\mathrm{MLE}} \text{ is not well defined}\right\} > 0$$

- An example – It is about *signal to noise* ratio, but ......



$$X = \underbrace{\lambda(e_1^{\otimes 3} + e_2^{\otimes 3})}_{T:\ \text{signal}} + E$$

$$\pi \to 0 \qquad \text{if and only if} \qquad \lambda \to \infty$$

- Dead end?

Tensor PCA

- Impose orthogonality

$$\mathrm{ODT} = \left\{ \sum_j \lambda_j \boldsymbol{a}_j \otimes \boldsymbol{b}_j \otimes \boldsymbol{c}_j : \boldsymbol{a}_j^\top \boldsymbol{a}_{j'} = \boldsymbol{b}_j^\top \boldsymbol{b}_{j'} = \boldsymbol{c}_j^\top \boldsymbol{c}_{j'} = \delta_{jj'} \right\}$$

- MLE

$$\widehat{\boldsymbol{T}}^{\mathrm{MLE}} = \underset{A \in \mathrm{ODT}, \mathrm{rank}(A) \leq R}{\arg\min} \|\boldsymbol{X} - \boldsymbol{A}\|_{\mathrm{F}}^2$$

Well-defined but *NP hard* to compute!

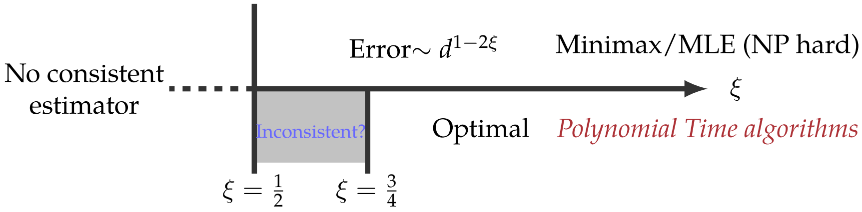
Statistical/Computational Tradeoff

$$X = \lambda u \otimes v \otimes w + E, \qquad e_{ijk} \sim_{\text{iid}} N(0,1)$$

- Dimension – d, signal strength – $\lambda = d^{\xi}$
- Minimax optimal estimator: best rank-one approximation

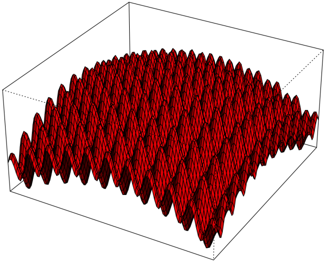
$$\|\widehat{u} \otimes \widehat{v} \otimes \widehat{w} - u \otimes v \otimes w\|_{\mathrm{F}}^2 \sim d^{1-2\xi}$$

- *Feasible* estimator: polynomial time computability



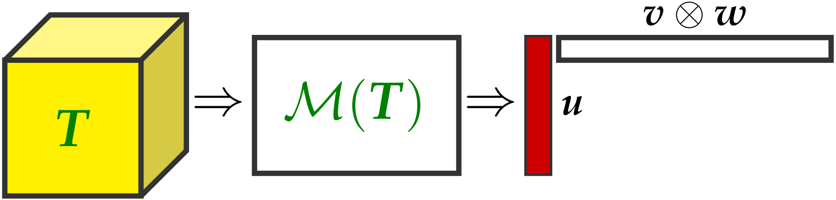
Why is it difficult?

$$\langle \boldsymbol{X}, \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z} \rangle \to \max$$

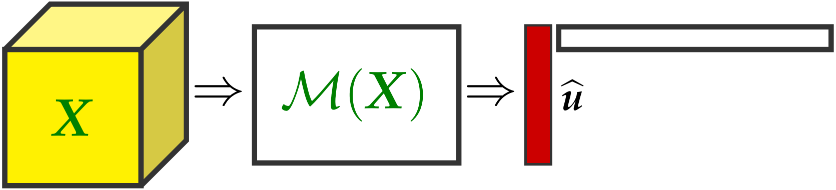


- Polynomial optimization
- Very smooth but highly nonconvex
- *If* we can get close to global optimum, ...

Higher order SVD

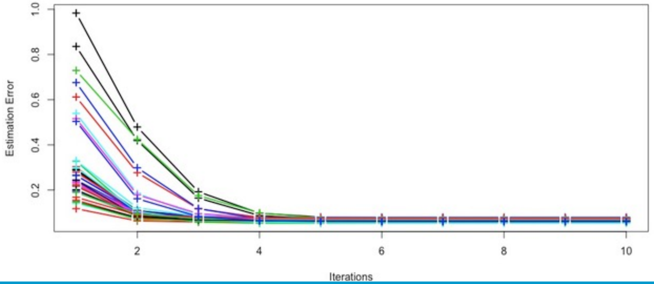


With noise

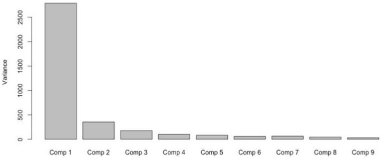


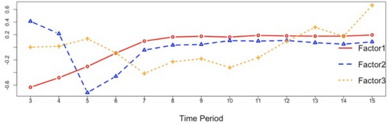
Consistent but *suboptimal* if $\lambda \gg d^{3/4}$

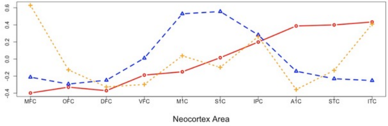
- Let $\boldsymbol{a}^{[m]} = \boldsymbol{a}/\|\boldsymbol{a}\|$ where
$$\boldsymbol{a} = \boldsymbol{X} \times_2 \boldsymbol{b}^{[m-1]} \times_3 \boldsymbol{c}^{[m-1]} - \boldsymbol{a}^{[m-1]};$$
- Let $\boldsymbol{b}^{[m]} = \boldsymbol{b}/\|\boldsymbol{b}\|$ where
$$\boldsymbol{b} = \boldsymbol{X} \times_1 \boldsymbol{a}^{[m]} \times_3 \boldsymbol{c}^{[m-1]} - \boldsymbol{b}^{[m-1]};$$
- Let $\boldsymbol{c}^{[m]} = \boldsymbol{c}/\|\boldsymbol{c}\|$ where
$$\boldsymbol{c} = \boldsymbol{X} \times_1 \boldsymbol{a}^{[m]} \times_2 \boldsymbol{b}^{[m-1]} - \boldsymbol{c}^{[m-1]}.$$

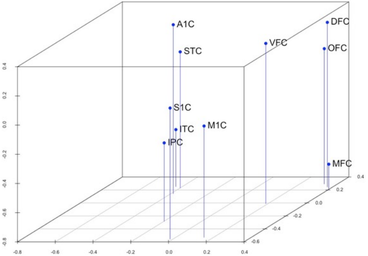


Spatio-Temporal Transcriptome of the Brain

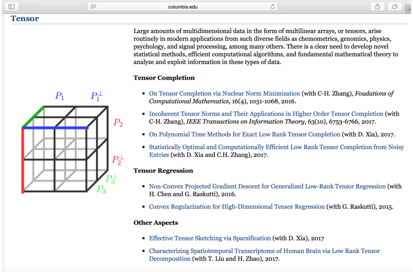








More General Treatment


columbia.edu

Tensor

Large amounts of multidimensional data in the form of multilinear arrays, or tensors, arise routinely in modern applications from such diverse fields as chemometrics, genomics, physics, psychology, and signal processing, among many others. There is a clear need to develop novel statistical methods, efficient computational algorithms, and fundamental mathematical theory to analyze and exploit information in these types of data.

P_1 $P_1^\perp$ P_2 $P_2^\perp$ $P_3^\perp$ P_3

Tensor Completion

- On Tensor Completion via Nuclear Norm Minimization (with C-H. Zhang), *Foudations of Computational Mathematics*, 16(4), 1031-1068, 2016.
- Incoherent Tensor Norms and Their Applications in Higher Order Tensor Completion (with C-H. Zhang), *IEEE Transactions on Information Theory*, 63(10), 6753-6766, 2017.
- On Polynomial Time Methods for Exact Low Rank Tensor Completion (with D. Xia), 2017.
- Statistically Optimal and Computationally Efficient Low Rank Tensor Completion from Noisy Entries (with D. Xia and C.H. Zhang), 2017.

Tensor Regression

- Non-Convex Projected Gradient Descent for Generalized Low-Rank Tensor Regression (with H. Chen and G. Raskutti), 2016.
- Convex Regularization for High-Dimensional Tensor Regression (with G. Raskutti), 2015.

Other Aspects

- Effective Tensor Sketching via Sparsification (with D. Xia), 2017
- Characterizing Spatiotemporal Transcriptome of Human Brain via Low Rank Tensor Decomposition (with T. Liu and H. Zhao), 2017.

`http://www.columbia.edu/~my2550/project.html#tensor`

Real applications are more complex but the *message* remains:

- Do not always rely on intuition from matrices
- Identify "easier" cases so that we can develop more efficient methods