Slides available:

http://www.ibis.t.u-tokyo.ac.jp/ryotat/tensor12kyoto.pdf

Statistical Performance of Convex Tensor Decomposition

Ryota Tomioka

2012/01/26 @ Kyoto University

Perspectives in Informatics 4B

Collaborators: Taiji Suzuki, Kohei Hayashi, Hisashi Kashima

Netflix challenge (2006-2009)

- $1,000,000 prize
- Goal: Improve the performance of a video recommendation system

 (predict who likes which movies)

- Example:

Likes “Star Wars” and “E.T.”,
Doesn’t like “Minority Report”.

Does he like “Blade Runner”?

Matrix completion view

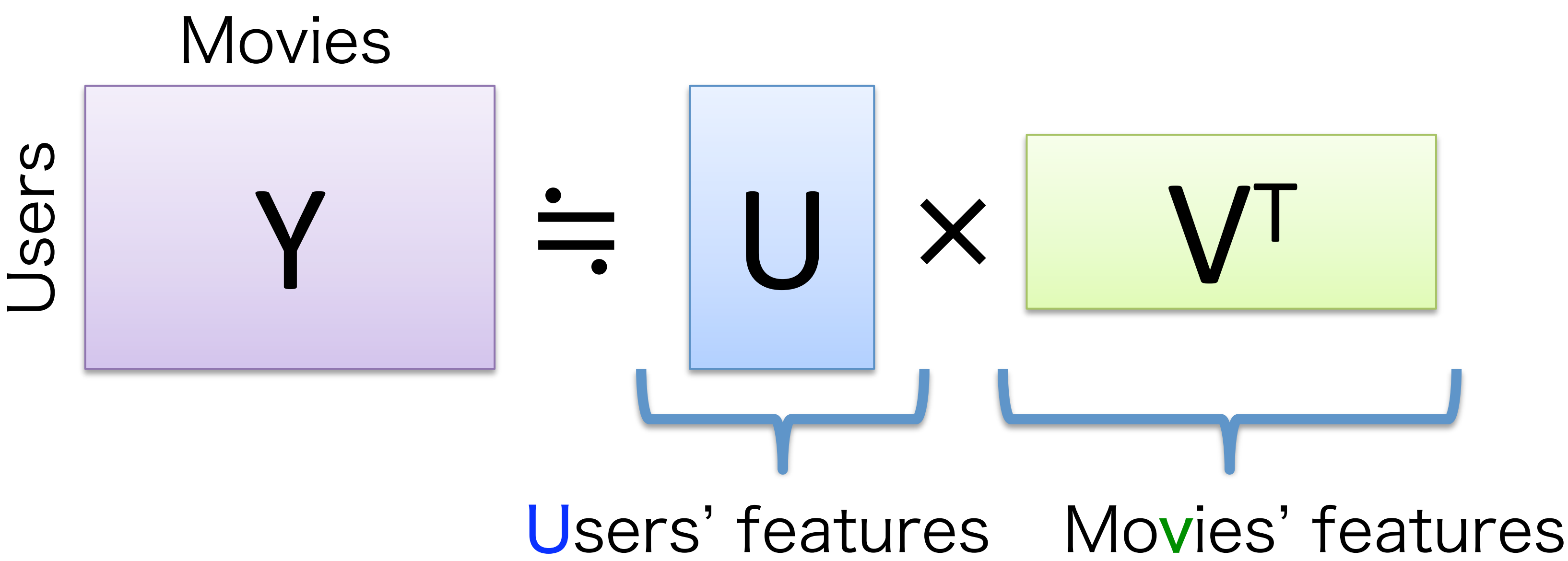
	Star Wars	E.T.	Minority Report	Blade Runner	Monsters Inc.
User A	+1	+1	-1	?	?
User B	+1	?	?	+1	?
User C	?	+1	-1	?	+1
User D	+1	?	?	?	+1
	.	.	.	.	.
	.	.	.	.	.

Goal: fill the missing entries!

Matrix completion

- Impossible without an assumption. (Missing entries can be arbitrary) --- problem is ill-posed
- Most common assumption:

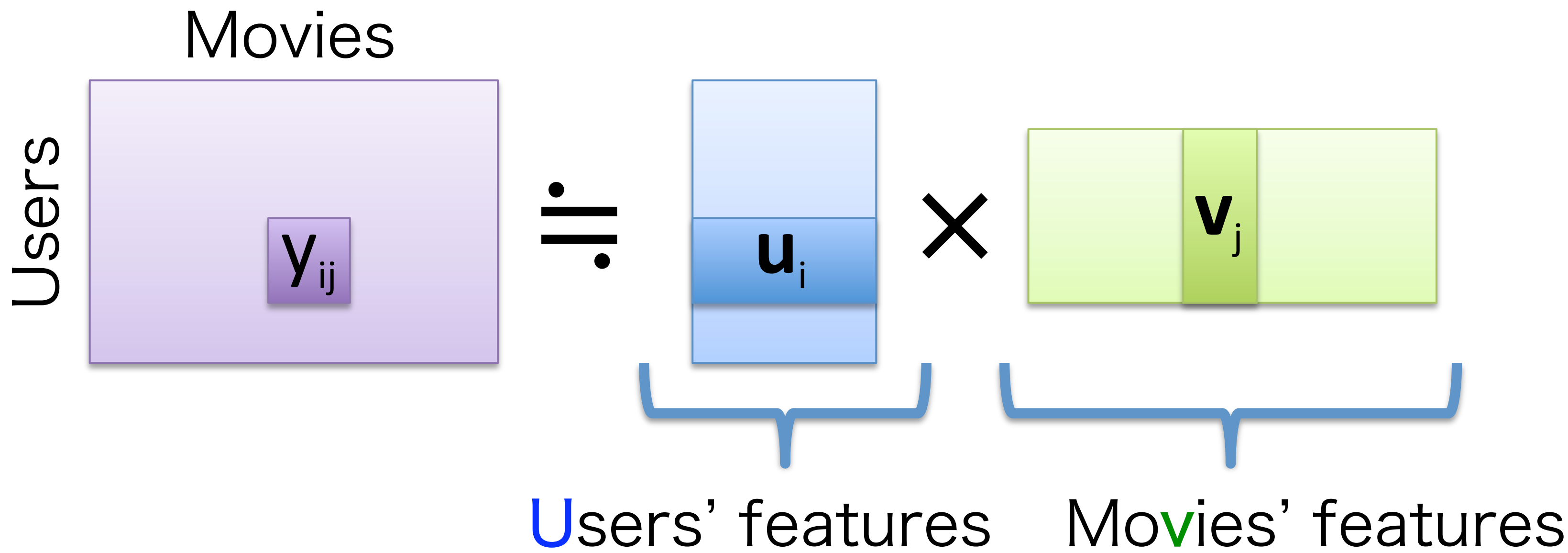
Low-rank decomposition



Matrix completion

- Most common assumption:

Low-rank decomposition (rank r)

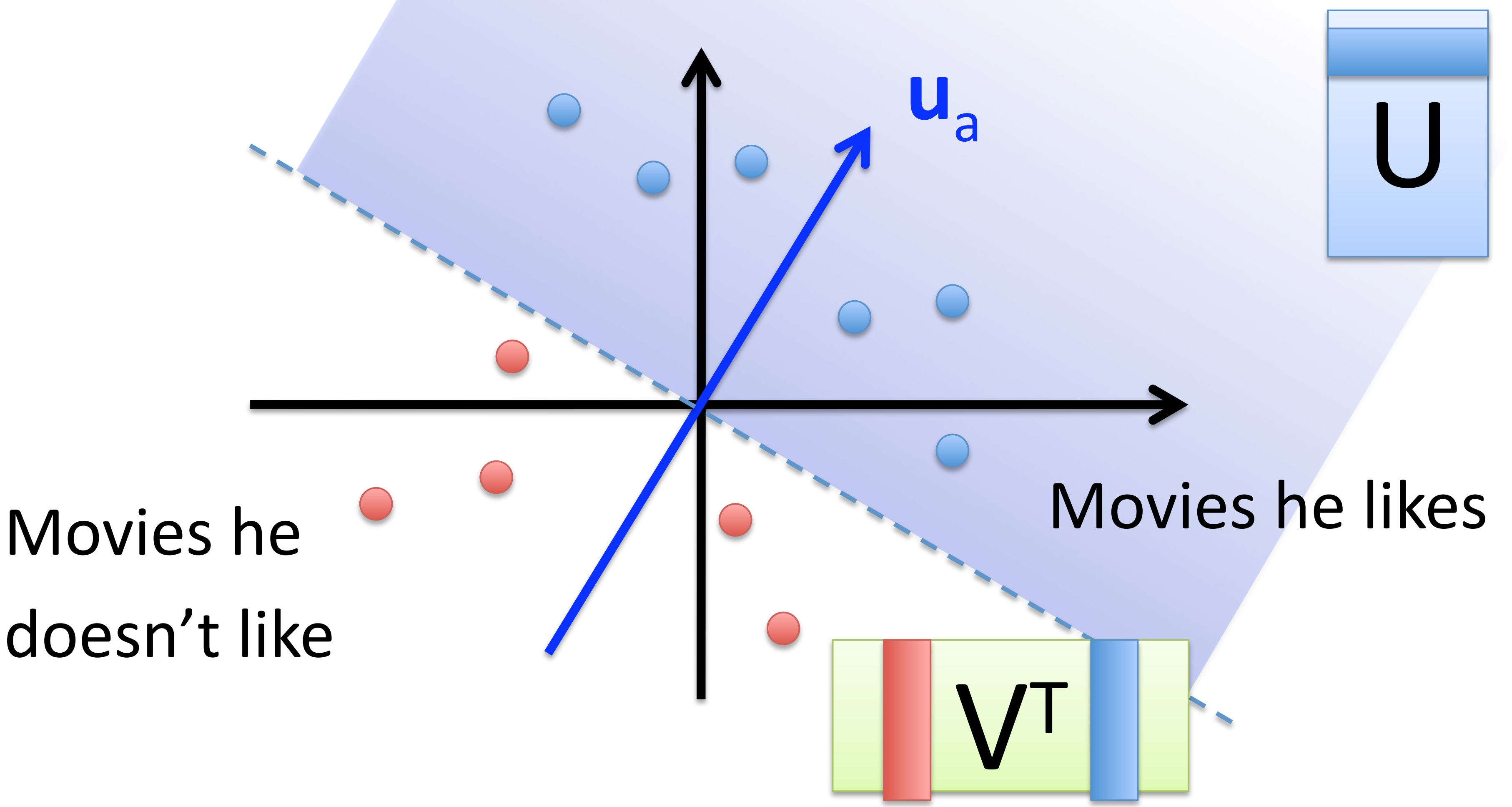


$$y_{ij} = \boldsymbol{u}_i^\top \boldsymbol{v}_j$$ (dot prodcut in r-dim space)

Geometric Intuition

r-dimensional space

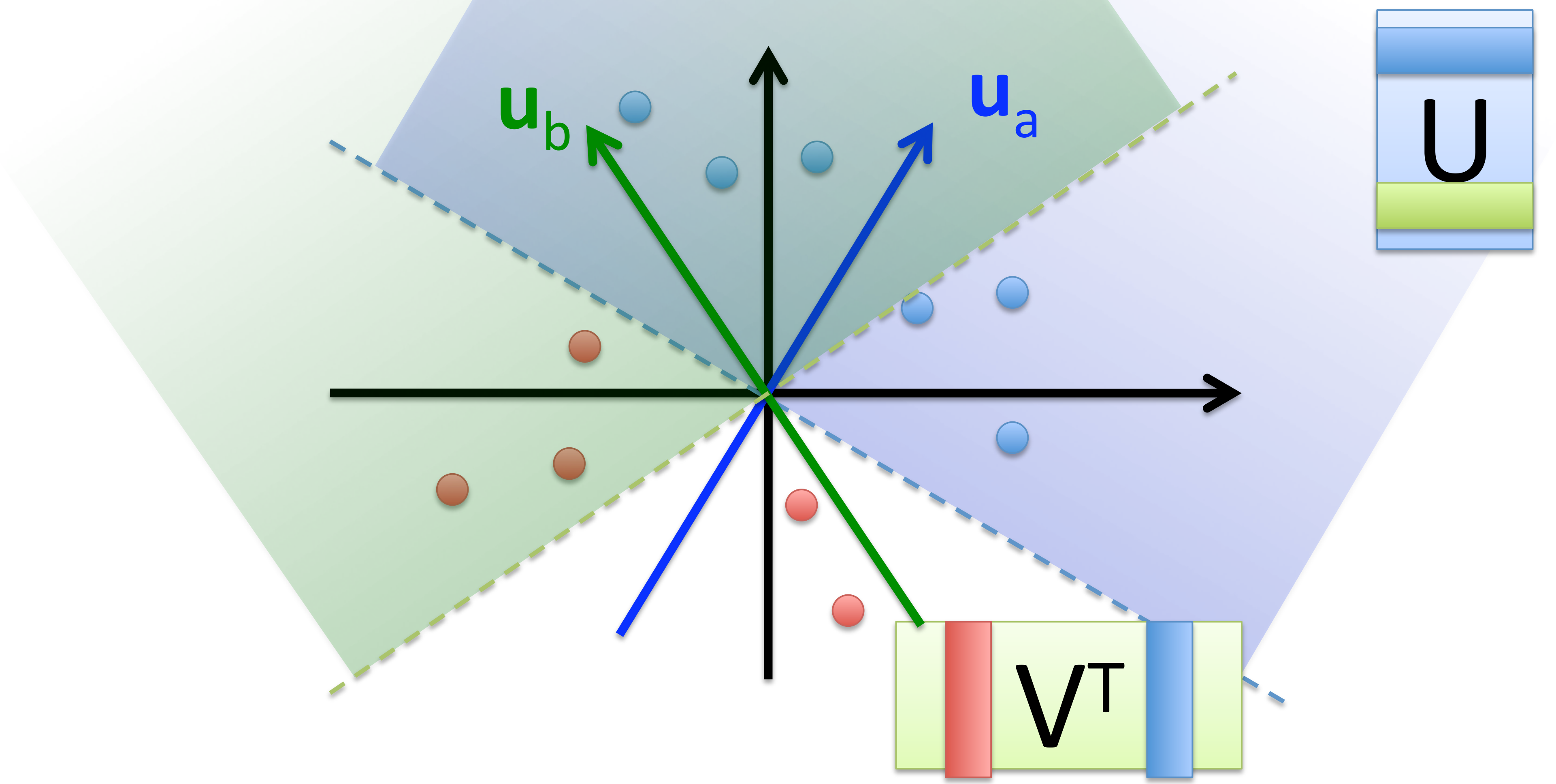
(r: the rank of the decomposition)



Geometric Intuition

r-dimensional space

(r: the rank of the decomposition)

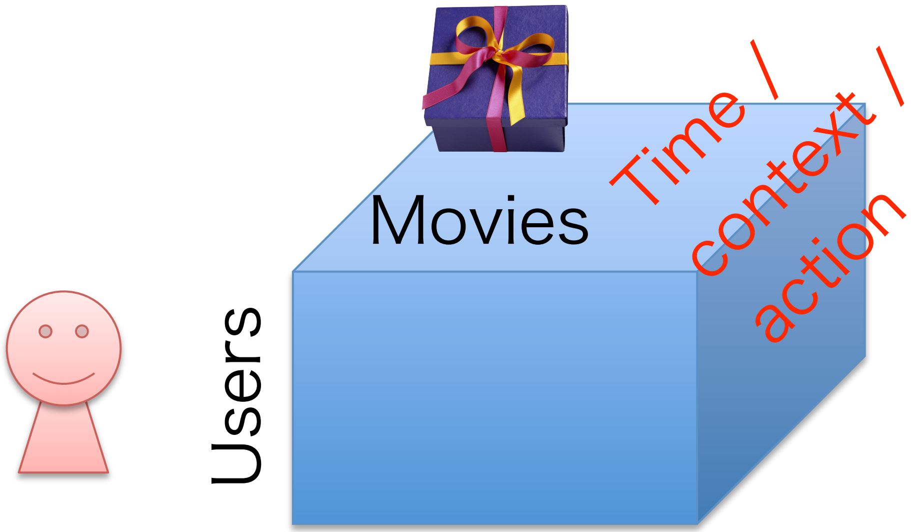


Tensor data completion

- Tensor = Multi-dimensional array
- Beyond 2D

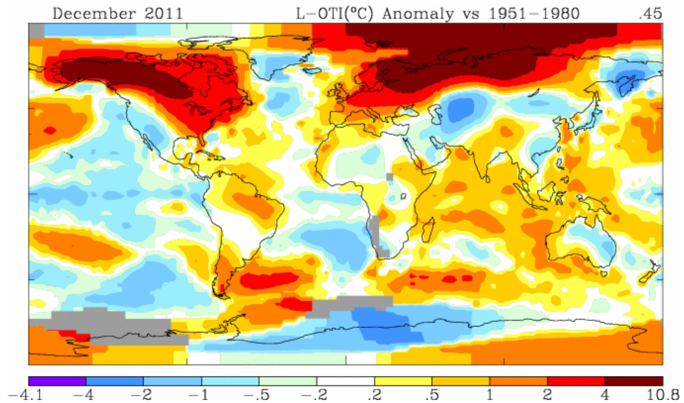
Movie preference

+ time / context / action



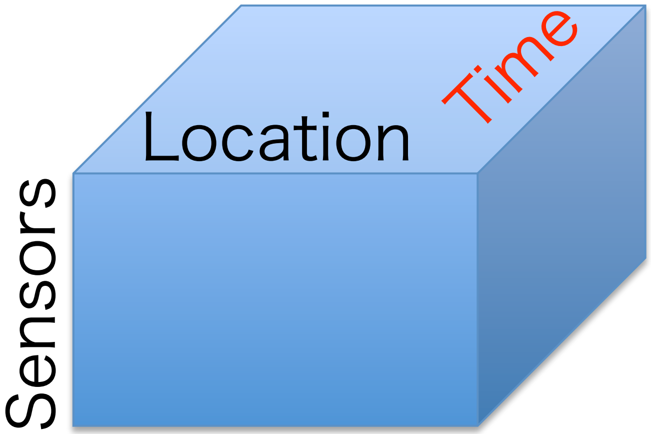
Tensor data completion

- Tensor = Multi-dimensional array
- Beyond 2D



Climate monitoring

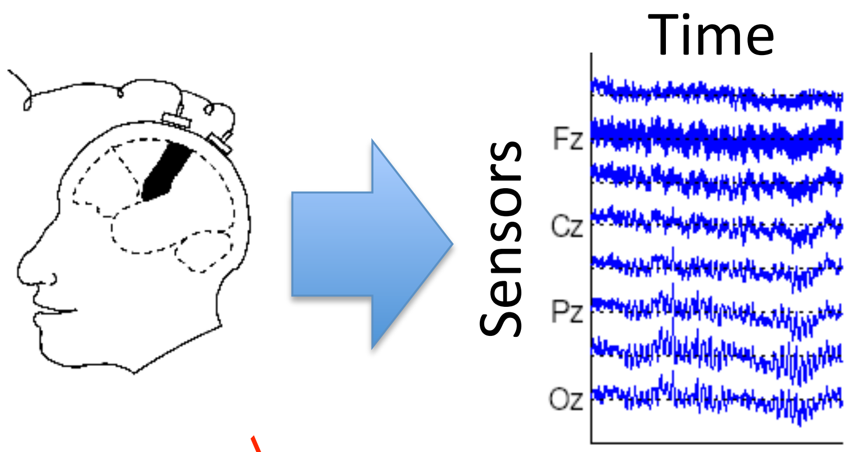
- temperature
- humidity
- rainfall

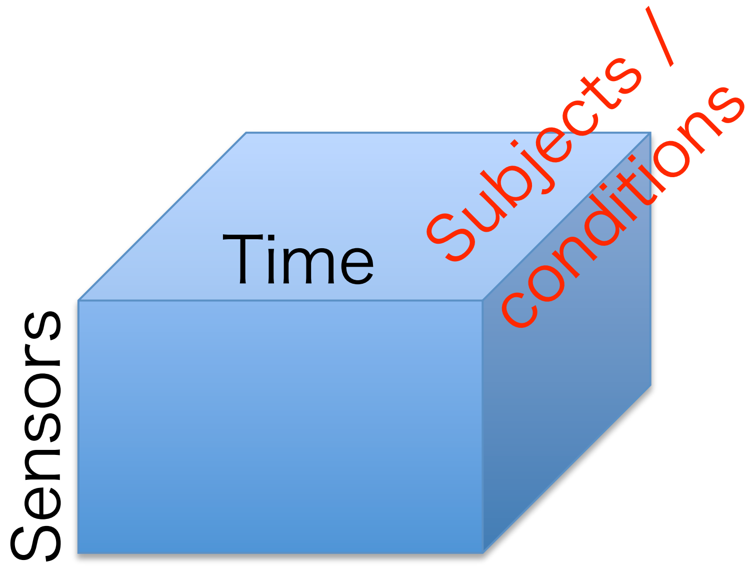


Tensor data completion

- Tensor = Multi-dimensional array
- Beyond 2D

Neuroscience
(brain imaging)





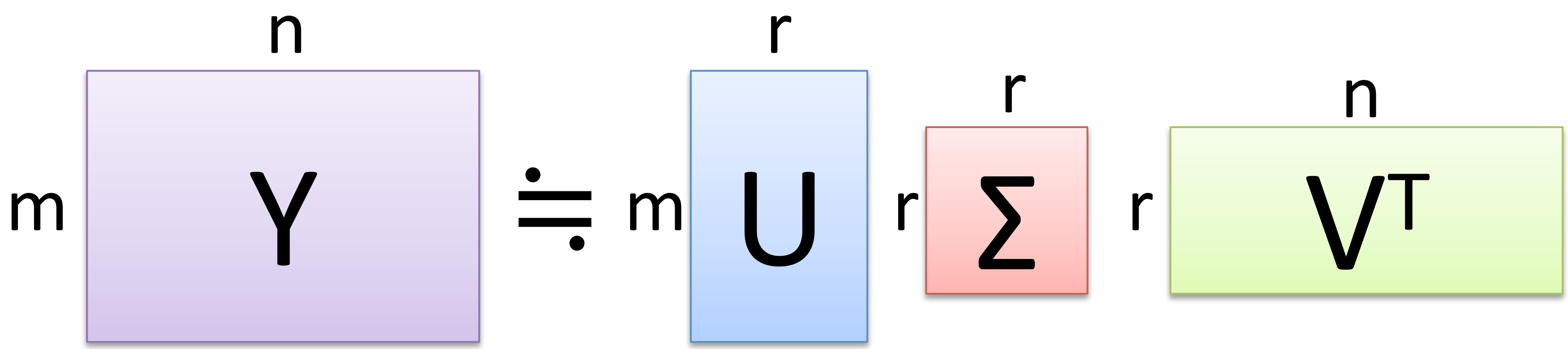
Rest of this talk

- Computing low-rank matrix decomposition
- Generalizing from matrix to tensor
- Analyzing the performance
  - Statistical learning theory

Computing low-rank matrix decomposition

Computing low-rank decomposition

- If all entries are observed (no missing entries)
  - Given Y, compute singular value decomposition (SVD)



where U, V: Orthogonal ($U^TU=I$, $V^TV=I$)

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix}$$

σ_j: jth largest singular value

Tolerating missings

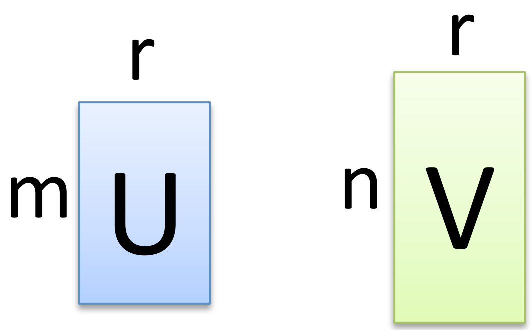
Optimization problem

$$\underset{\boldsymbol{U},\boldsymbol{V}}{\text{minimize}} \sum_{(ij)\in\Omega} (y_{ij} - \boldsymbol{u}_i{}^{\top}\boldsymbol{v}_j)^2$$

Set of observed index pairs

Users' features

Movies' features



Tolerating missings

Optimization problem **Non-convex!**

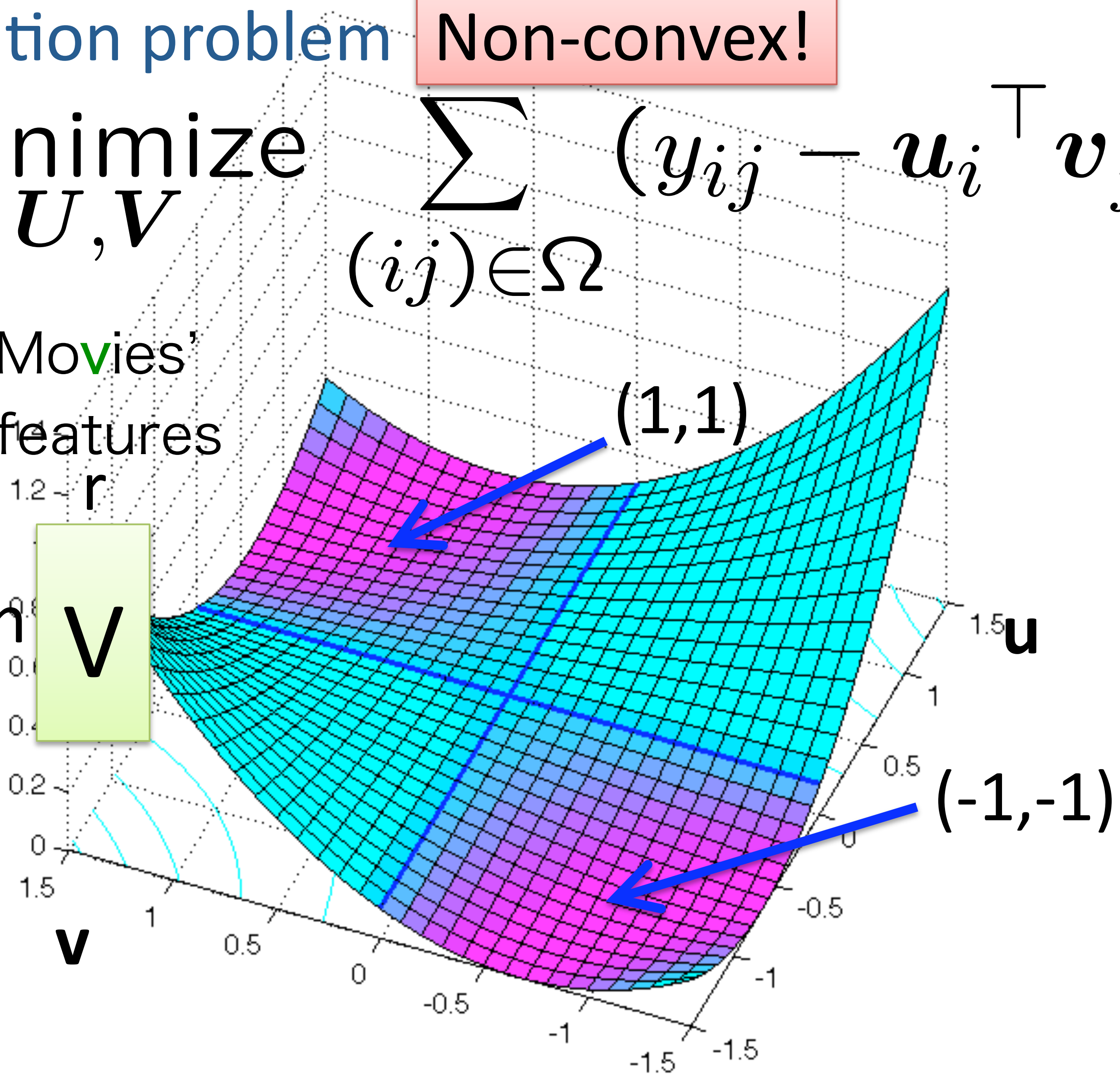
$$\underset{\boldsymbol{U},\boldsymbol{V}}{\text{minimize}} \sum_{(ij)\in\Omega} (y_{ij} - \boldsymbol{u}_i{}^\top \boldsymbol{v}_j)^2$$

Users' features

Movies' features

m U (size $m \times r$)

n V (size $n \times r$)



Tolerating missings

Optimization problem

Still non-convex!

$$\underset{\boldsymbol{W}}{\text{minimize}} \quad \sum_{(ij)\in\Omega} (y_{ij} - w_{ij})^2,$$

$$\text{subject to} \quad \text{rank}(\boldsymbol{W}) \leq r$$

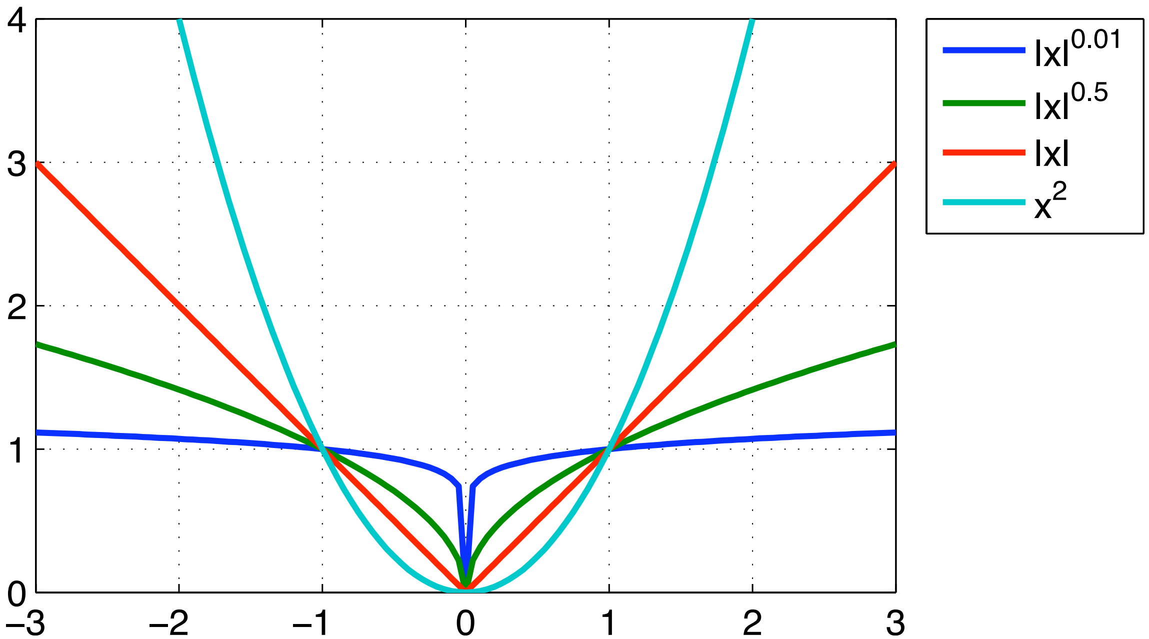
Rank constraint is NP hard

Convex relaxation of rank

Schatten p-norm (to the pth power) $\|\boldsymbol{W}\|_{S_p}^p := \sum_{j=1}^r \sigma_j^p(\boldsymbol{W})$

$\sigma_j(\boldsymbol{W})$: jth largest singular value

$$\|\boldsymbol{W}\|_{S_p}^p \xrightarrow{p \to 0} \mathrm{rank}(\boldsymbol{W})$$



p=1 is the tightest convex relaxation (also known as trace norm / nuclear norm)

Tolerating missings

Optimization problem Convex relaxation

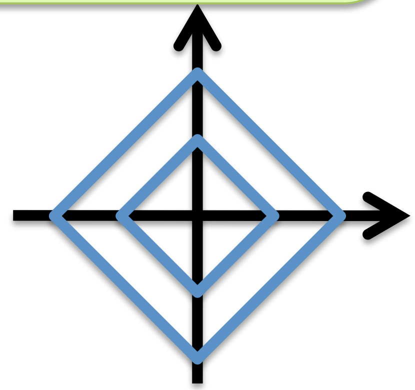
$$\underset{\boldsymbol{W}}{\text{minimize}} \quad \sum_{(ij)\in\Omega} (y_{ij} - w_{ij})^2,$$
$$\text{subject to} \quad \|\boldsymbol{W}\|_{S_1} \leq \tau$$

Schatten 1-norm
(nuclear norm,
trace norm)

$$\|\boldsymbol{W}\|_{S_1} = \sum_{j=1}^{r} \sigma_j(\boldsymbol{W})$$

$\sigma_j(\boldsymbol{W})$: *j*th largest singular value

Cf. Lasso (L_1 norm) for variable selection
= linear sum of abs. coefficients



Take home messages

- Rank constrained minimization is hard to solve (non-convex and NP hard)
- Can be relaxed into a tractable convex problem using Schatten 1-norm.

How about tensors?

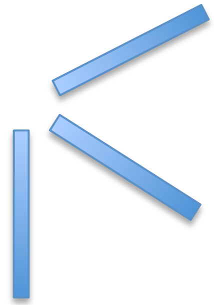
- How to define tensor rank?
- How related to matrix rank?

Ranf of a tensor

Definition. Let $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ (*K*th order tensor)

The smallest number R such that the given tensor X is written as

$$\mathcal{X} = \sum_{r=1}^{R} \mathcal{A}_r \quad \text{where} \quad \mathcal{A}_r =$$



is rank one.

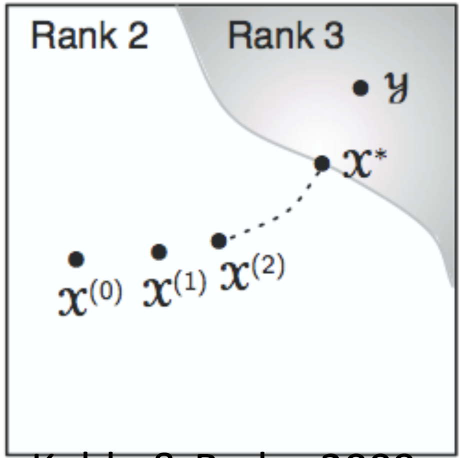
（can be written as an outer product of K vectors）

- Called CP (CANDECOMPO/PARAFAC) decomposition
- Bad news: NP hard to compute the rank R even for a fully observed X.

Bad news 2: Tensor rank is not closed

X is rank 3

$$\mathcal{X} = a_1 \circ b_1 \circ c_2 + a_1 \circ b_2 \circ c_1 + a_2 \circ b_1 \circ c_1$$



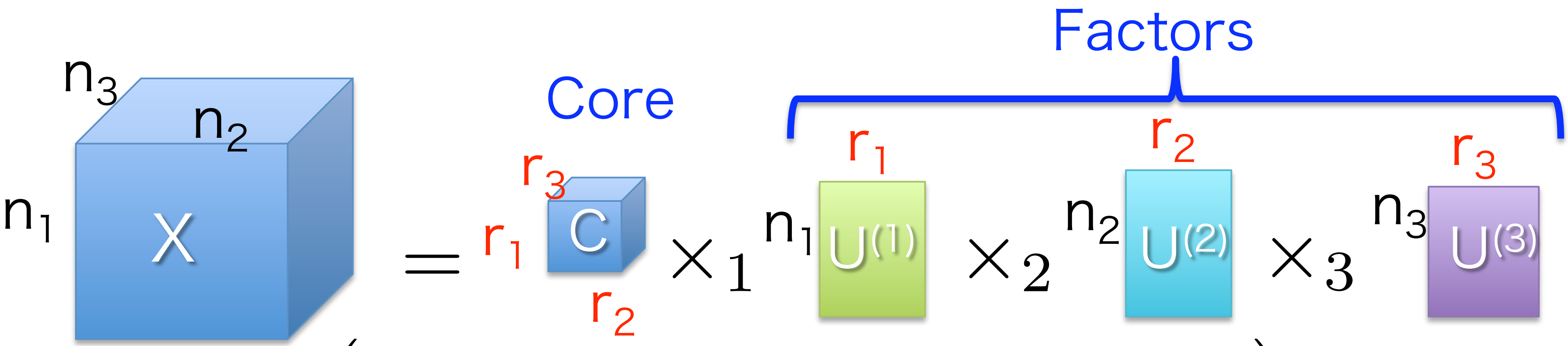
Kolda & Bader 2009

Y is rank 2

$$\mathcal{Y} = \alpha(a_1 + \frac{1}{\alpha}a_2) \circ (b_1 + \frac{1}{\alpha}b_2) \circ (c_1 + \frac{1}{\alpha}c_2) - \alpha a_1 \circ b_1 \circ c_1$$

$$\|\mathcal{X} - \mathcal{Y}\|_F \to 0 \quad (\alpha \to \infty)$$

Tucker decomposition [Tucker 66]

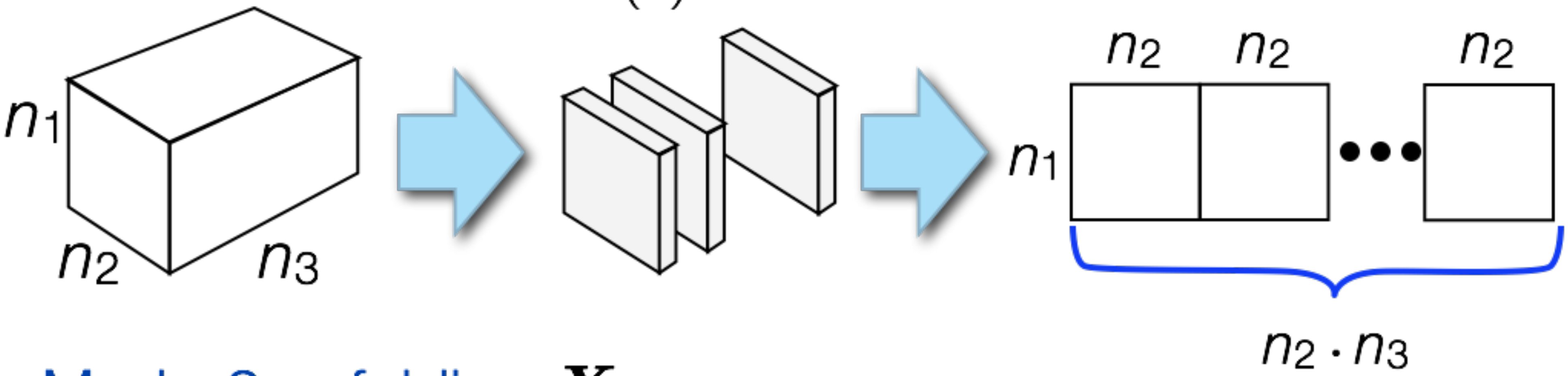


$$\left(X_{ijk} = \sum_{a=1}^{r_1} \sum_{b=1}^{r_2} \sum_{c=1}^{r_3} C_{abc} U^{(1)}_{ia} U^{(2)}_{jb} U^{(3)}_{kc} \right)$$

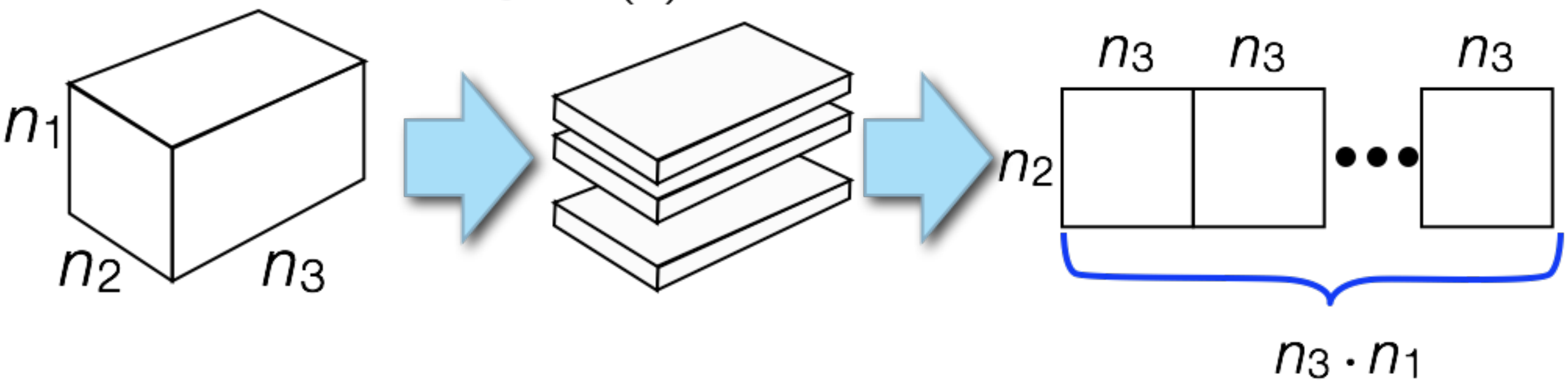
- Also known as higher-order SVD [De Lathauwer+00]
- Rank (r_1, r_2, r_3) can be computed in polynomial time using unfolding operations.

Mode-k unfoldings (matricization)

Mode-1 unfolding $\boldsymbol{X}_{(1)}$



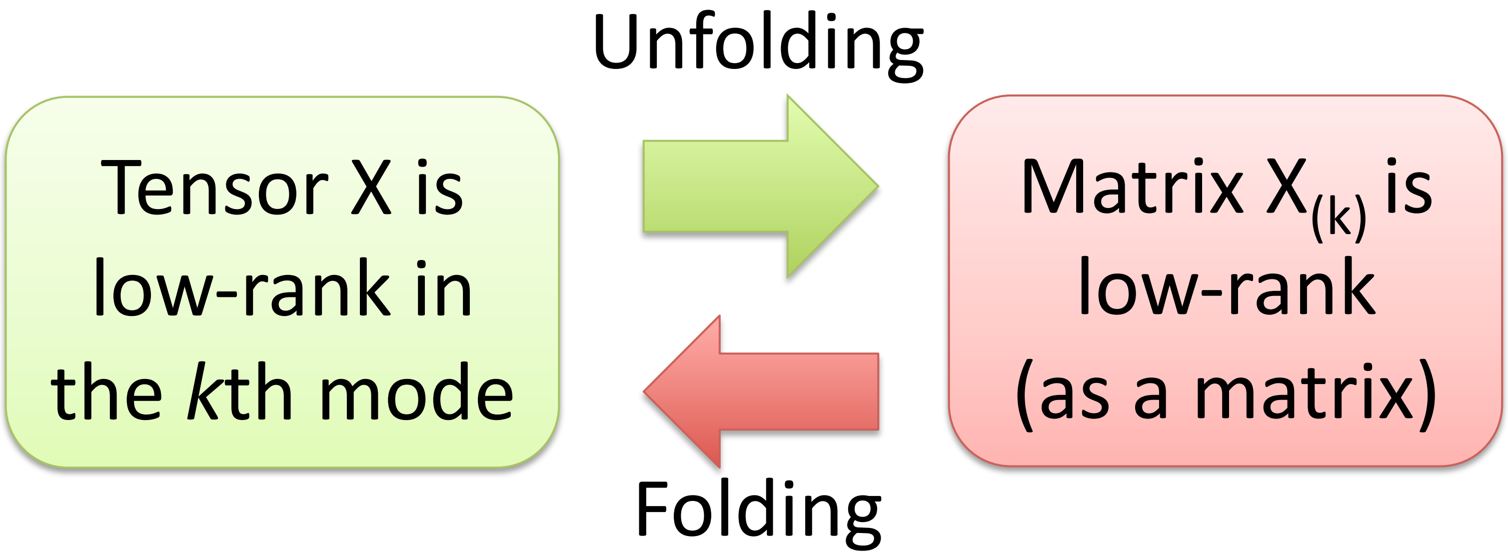
Mode-2 unfolding $\boldsymbol{X}_{(2)}$



Computing Tucker rank

- For each k=1,…,K
  - Compute the mode-k unfolding $X_{(k)}$
  - Compute the (matrix) rank of $X_{(k)}$

$$r_k = \mathrm{rank}(\boldsymbol{X}_{(k)})$$



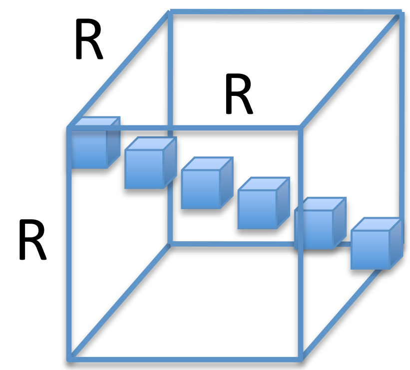
Computing Tucker rank

- For each k=1,…,K
  - Compute the mode-k unfolding $X_{(k)}$
  - Compute the (matrix) rank of $X_{(k)}$

$$r_k = \text{rank}(\boldsymbol{X}_{(k)})$$

- Difference between Tensor rank and Tucker rank
  - Tensor rank is a single number R (may not be easy to compute)
  - Tucker rank is defined for each mode (easy to compute)
- CP decomp is a special case of Tucker decomp with $R=r_1=r_2=\ldots=r_K$ and diagonal core

C =



Basic idea

- We know how to do matrix completion with Schatten 1-norm (tractable convex optimization)
- We know how to compute Tucker rank (=the rank of the mode-k unfolding)

Schatten 1-norm + Unfolding = Convex, tractable tensor completion

Overlapped Schatten 1-norm for Tensors

$$|||\mathcal{W}|||_{S_1} := \frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{W}_{(k)}\|_{S_1}$$

Schatten 1-norm of the mode-k unfolding

Measures the overall low-rank-ness

(not just a single mode)

Convex Tensor Estimation

Matrix

Estimation of *low-rank* matrix (hard)

Convex relaxation

Schatten 1-norm minimization (tractable) [Fazel, Hindi, Boyd 01]

Generalize

Tensor

Estimation of *low-rank* tensor (hard)

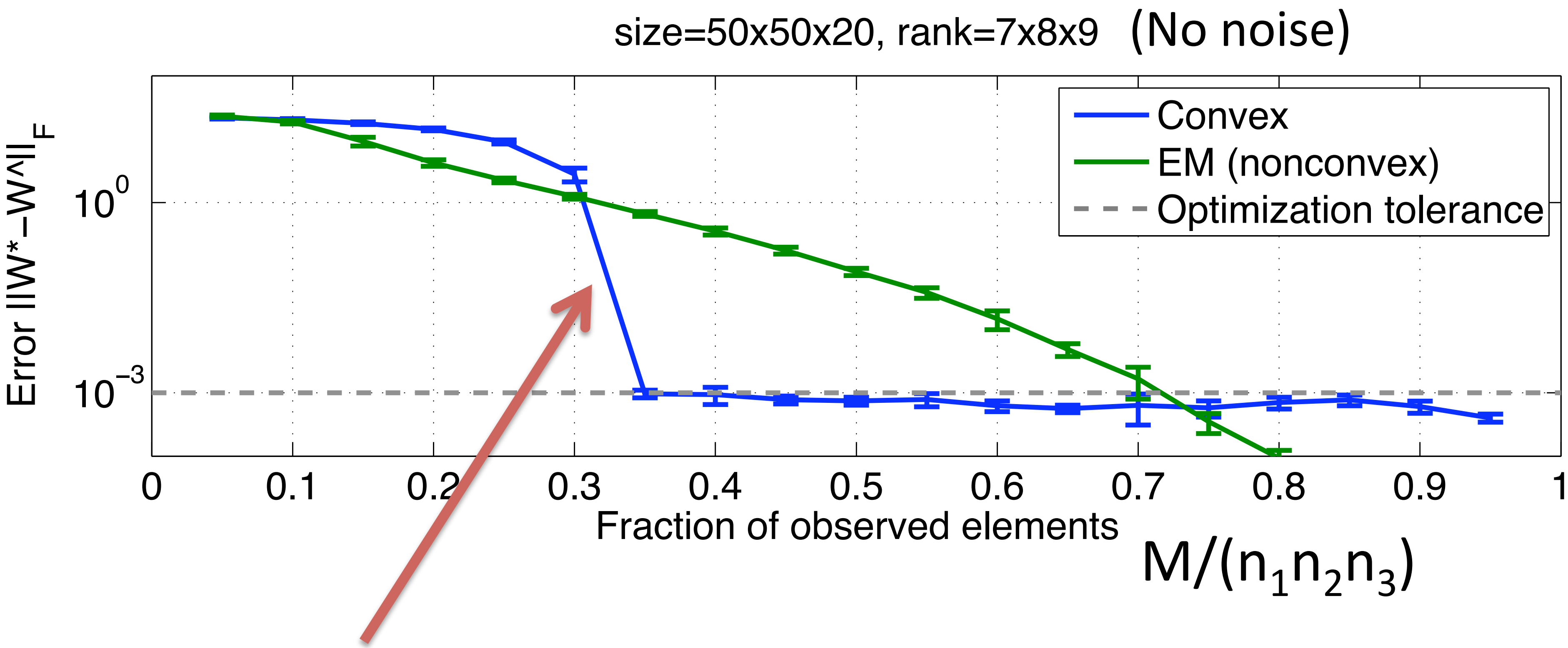
Convex relaxation

Overlapped Schatten 1-norm minimization

[Liu+09, Signoretto+10, Tomioka+10, Gandy+11]

Empirical performance

Tensor completion result [Tomioka et al. 2010]



Phase transition!!

Can we predict this theoretically?

Analyzing the performance of convex tensor decomposition

Problem setting

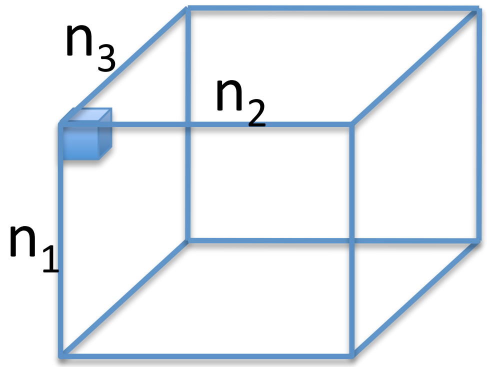
Observation model

$$y_i = \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{W}}^* \rangle + \epsilon_i \quad (i = 1, \ldots, M)$$

$\boldsymbol{\mathcal{W}}^*$ true tensor rank-(r_1,...,r_K)

ϵ_i Gaussian noise

Example (tensor completion)

X_1 = 

Problem setting

Observation model

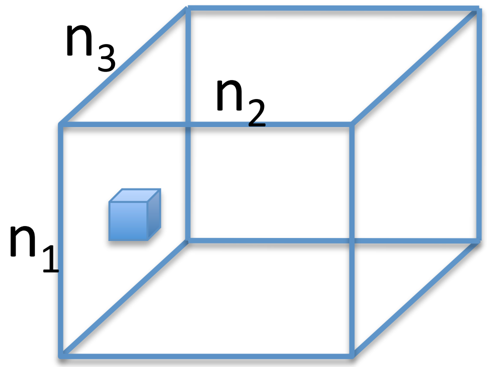
$$y_i = \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{W}}^* \rangle + \epsilon_i \quad (i = 1, \ldots, M)$$

$\boldsymbol{\mathcal{W}}^*$ true tensor rank-(r_1,...,r_K)

ϵ_i Gaussian noise

Example (tensor completion)

X_2 =



Problem setting

Observation model

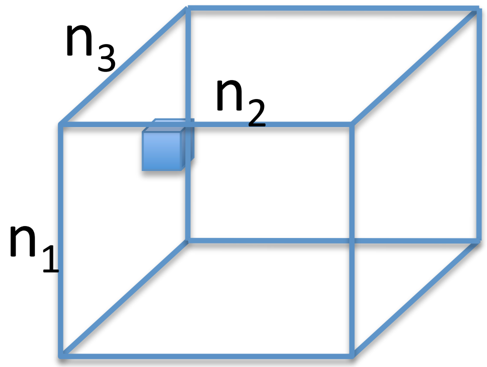
$$y_i = \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{W}}^* \rangle + \epsilon_i \quad (i = 1, \ldots, M)$$

$\mathcal{W}^*$ true tensor rank-(r_1,...,r_K)

ϵ_i Gaussian noise

Example (tensor completion)

X_3 =



Problem setting

Observation model

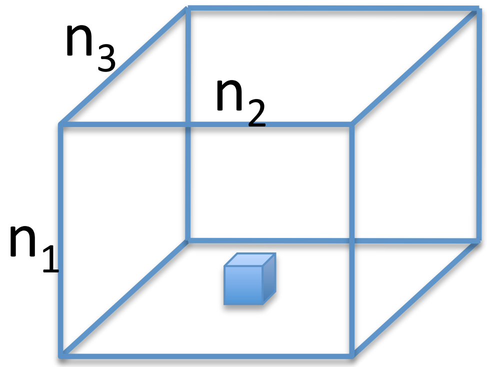
$$y_i = \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{W}}^* \rangle + \epsilon_i \quad (i = 1, \ldots, M)$$

$\mathcal{W}^*$ true tensor rank-(r_1,...,r_K)

ϵ_i Gaussian noise

Example (tensor completion)

X_4 =



and so on...

Problem setting

Observation model

$$y_i = \langle \mathcal{X}_i, \mathcal{W}^* \rangle + \epsilon_i \quad (i = 1, \ldots, M)$$

$\mathcal{W}^*$ true tensor rank-(r_1,…,r_K)

ϵ_i Gaussian noise N(0,σ^2)

Optimization

Empirical error Regularization

$$\hat{\mathcal{W}} = \underset{\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}}{\operatorname{argmin}} \left(\frac{1}{2M} \| \boldsymbol{y} - \mathfrak{X}(\mathcal{W}) \|_2^2 + \lambda_M \left|\left|\left| \mathcal{W} \right|\right|\right|_{S_1} \right)$$

Observation model Reg. Const.

$$(N = \textstyle\prod_{k=1}^{K} n_k)$$

$$\mathfrak{X} : \mathbb{R}^N \to \mathbb{R}^M$$

$$\mathfrak{X}(\mathcal{W}) = (\langle \mathcal{X}_1, \mathcal{W} \rangle, \ldots, \langle \mathcal{X}_M, \mathcal{W} \rangle)^\top$$

Analysis objective

- We would like to show something like

Estimated tensor

True low-rank tensor

Mean squared error

$$\frac{\left\|\widehat{\mathcal{W}} - \mathcal{W}^*\right\|_F^2}{N} \leq O_p\left(\frac{c(\boldsymbol{n}, \boldsymbol{r})}{M}\right)$$

The size $\boldsymbol{n} = (n_1, \ldots, n_K)$

The rank $\boldsymbol{r} = (r_1, \ldots, r_K)$

Number of samples M

Theorem: random Gauss design

Assume elements of X_i are drown iid from standard normal distribution. Moreover

$$\frac{\#\text{samples } (M)}{\#\text{variables } (N)} \geq c_1 \underbrace{\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}_{\text{Normalized rank}} \approx \frac{r}{n}$$

$$\|\boldsymbol{n}^{-1}\|_{1/2} := \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{1/n_k}\right)^2, \quad \|\boldsymbol{r}\|_{1/2} := \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k}\right)^2$$

Theorem: random Gauss design

Assume elements of X_i are drown iid from standard normal distribution. Moreover

$$\frac{\#\text{samples } (M)}{\#\text{variables } (N)} \geq c_1 \underbrace{\|\boldsymbol{n}^{-1}\|_{1/2} \|\boldsymbol{r}\|_{1/2}}_{\text{Normalized rank}} \approx \frac{r}{n}$$

Convergence!

$$\frac{\left\| \hat{\mathcal{W}} - \mathcal{W}^* \right\|_F^2}{N} \leq O_p\left(\frac{\sigma^2 \|\boldsymbol{n}^{-1}\|_{1/2} \|\boldsymbol{r}\|_{1/2}}{M} \right)$$

$$\|\boldsymbol{n}^{-1}\|_{1/2} := \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{1/n_k} \right)^2, \quad \|\boldsymbol{r}\|_{1/2} := \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{r_k} \right)^2$$

Proof idea

Since $\hat{\mathcal{W}}$ minimizes the objective,

Estimated tensor

True low-rank tensor

$$\mathrm{Obj}(\hat{\mathcal{W}}) \leq \mathrm{Obj}(\mathcal{W}^*)$$

It is not so hard to see:

$$\mathfrak{X}^*(\boldsymbol{\epsilon}) = \sum_{i=1}^{M} \epsilon_i \mathcal{X}_i$$

$$\frac{1}{2M}\|\mathfrak{X}(\hat{\mathcal{W}} - \mathcal{W}^*)\|_2^2 \leq \left\langle \mathfrak{X}^*(\boldsymbol{\epsilon})/M, \hat{\mathcal{W}} - \mathcal{W}^* \right\rangle + \lambda_M \left\| \hat{\mathcal{W}} - \mathcal{W}^* \right\|_{S_1}$$

What we want to derive:

$$\frac{\left\| \hat{\mathcal{W}} - \mathcal{W}^* \right\|_F^2}{N} \leq O_p\left(\frac{c(\boldsymbol{n}, \boldsymbol{r})}{M}\right)$$

Proof outline (1/3)

Estimated tensor

True low-rank tensor

$$\frac{1}{2M}\|\mathfrak{X}(\widehat{\mathcal{W}}-\mathcal{W}^*)\|_2^2 \le \left\langle \mathfrak{X}^*(\boldsymbol{\epsilon})/M, \widehat{\mathcal{W}}-\mathcal{W}^*\right\rangle + \lambda_M \left\|\widehat{\mathcal{W}}-\mathcal{W}^*\right\|_{S_1}$$

Inequality 1: upper-bound the dot product

$$\left\langle \mathfrak{X}^*(\boldsymbol{\epsilon})/M, \widehat{\mathcal{W}}-\mathcal{W}^*\right\rangle \le O_p\left(\sqrt{\frac{\sigma^2 N\|\boldsymbol{n}^{-1}\|_{1/2}}{M}}\left\|\widehat{\mathcal{W}}-\mathcal{W}^*\right\|_{S_1}\right)$$

(optimization duality / random matrix theory)

Proof outline (1/3)

Estimated tensor

True low-rank tensor

$$\frac{1}{2M}\|\mathfrak{X}(\hat{\mathcal{W}} - \mathcal{W}^*)\|_2^2 \leq \left(\sqrt{\frac{\sigma^2 N \|\boldsymbol{n}^{-1}\|_{1/2}}{M}} + \lambda_M\right) \left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_{S_1}$$

Inequality 1: upper-bound the dot product

$$\left\langle \mathfrak{X}^*(\boldsymbol{\epsilon})/M, \hat{\mathcal{W}} - \mathcal{W}^* \right\rangle \leq O_p\left(\sqrt{\frac{\sigma^2 N \|\boldsymbol{n}^{-1}\|_{1/2}}{M}} \left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_{S_1}\right)$$

Trade-off between $\sqrt{\frac{\sigma^2 N \|\boldsymbol{n}^{-1}\|_{1/2}}{M}}$ and λ_M

⇨ Optimal reg. const $\lambda_M \simeq O_p\left(\sqrt{\frac{\sigma^2 N \|\boldsymbol{n}^{-1}\|_{1/2}}{M}}\right)$

Proof outline (2/3)

Estimated tensor

True low-rank tensor

$$\frac{1}{2M}\|\mathfrak{X}(\widehat{\mathcal{W}}-\mathcal{W}^*)\|_2^2 \leq \sqrt{\frac{\sigma^2 N\|\boldsymbol{n}^{-1}\|_{1/2}}{M}}\left\|\widehat{\mathcal{W}}-\mathcal{W}^*\right\|_{S_1}$$

Inequality 2: relate the schatten 1-norm with the Frobenius norm

$$\left\|\widehat{\mathcal{W}}-\mathcal{W}^*\right\|_{S_1} \leq \sqrt{\|\boldsymbol{r}\|_{1/2}}\left\|\widehat{\mathcal{W}}-\mathcal{W}^*\right\|_F$$

(relation between L1- and L2-norm)

Proof outline (2/3)

Estimated tensor

True low-rank tensor

$$\frac{1}{2M}\|\mathfrak{X}(\hat{\mathcal{W}} - \mathcal{W}^*)\|_2^2 \leq \sqrt{\frac{\sigma^2 N \|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}{M}} \left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_F$$

Inequality 2: relate the schatten 1-norm with the Frobenius norm

$$\left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_{S_1} \leq \sqrt{\|\boldsymbol{r}\|_{1/2}} \left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_F$$

(relation between L1- and L2-norm)

Proof outline (3/3)

Estimated tensor

True low-rank tensor

$$\frac{1}{2M}\|\mathfrak{X}(\hat{\mathcal{W}} - \mathcal{W}^*)\|_2^2 \le \sqrt{\frac{\sigma^2 N \|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}{M}}\left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_F$$

Inequality 3: lower-bound the left hand-side

$$\kappa\left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_F^2 \le \frac{1}{M}\|\mathfrak{X}(\hat{\mathcal{W}} - \mathcal{W}^*)\|_2^2$$

If $\dfrac{\#\text{samples } (M)}{\#\text{variables } (N)} \ge c_1 \|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}$

(Gordon-Slepian Theorem in Gaussian process theory)

Proof outline (3/3)

Estimated tensor

True low-rank tensor

$$\kappa\left\|\widehat{\mathcal{W}}-\mathcal{W}^*\right\|_F^2 \le \sqrt{\frac{\sigma^2 N\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}{M}}\left\|\widehat{\mathcal{W}}-\mathcal{W}^*\right\|_F$$

Inequality 3: lower-bound the left hand-side

$$\kappa\left\|\widehat{\mathcal{W}}-\mathcal{W}^*\right\|_F^2 \le \frac{1}{M}\|\mathfrak{X}(\widehat{\mathcal{W}}-\mathcal{W}^*)\|_2^2$$

If $\dfrac{\#\text{samples } (M)}{\#\text{variables } (N)} \ge c_1\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}$

(Gordon-Slepian Theorem in Gaussian process theory)

Back to the theorem statement

Assume elements of X_i are drown iid from standard normal distribution. Moreover

$$\frac{\#\text{samples } (M)}{\#\text{variables } (N)} \geq c_1 \underbrace{\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}} \approx \frac{r}{n}$$

Normalized rank

Convergence!

$$\frac{|||\hat{\mathcal{W}} - \mathcal{W}^*|||_F^2}{N} \leq O_p\left(\frac{\sigma^2\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}{M}\right)$$

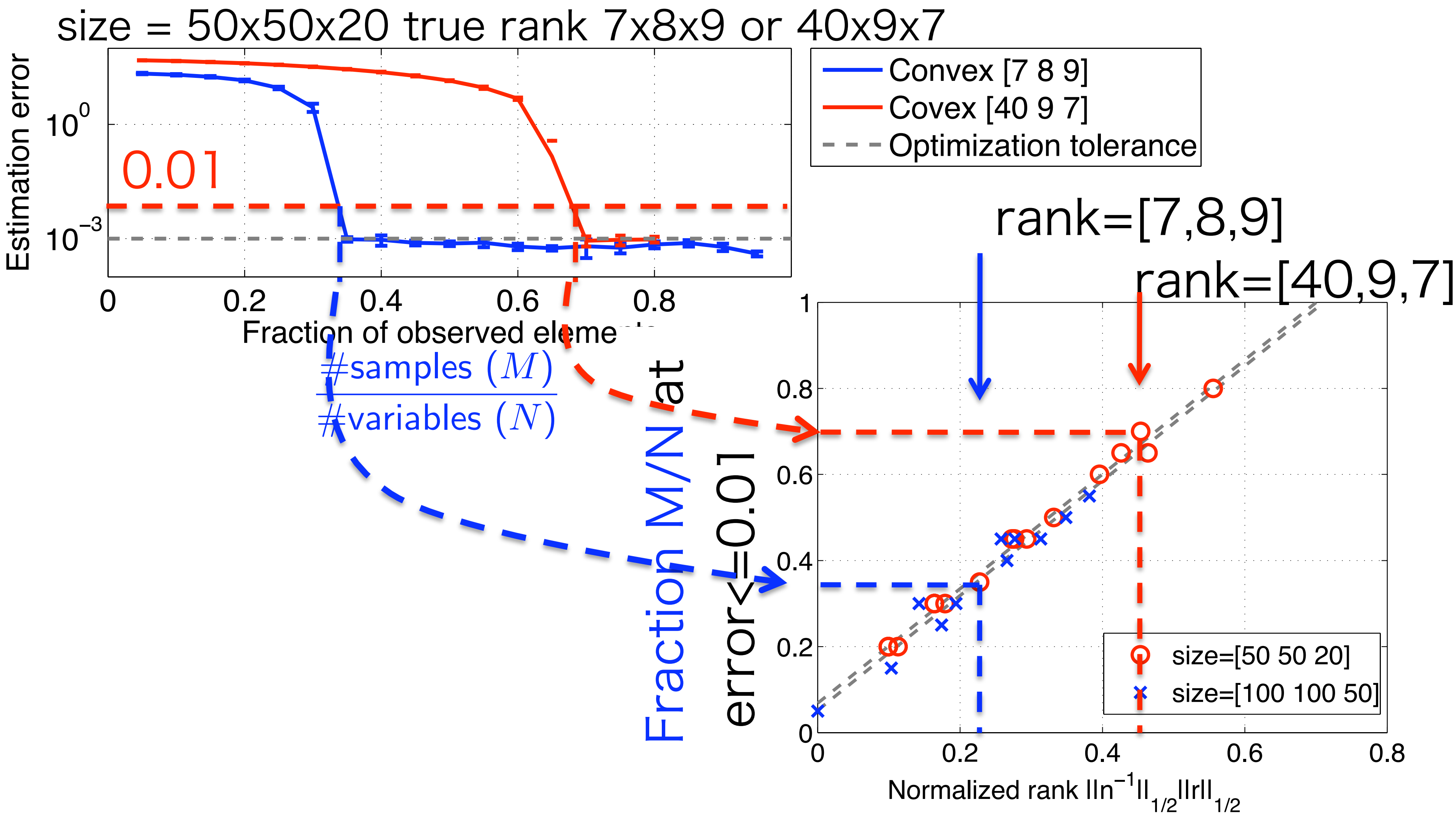
$$\|\boldsymbol{n}^{-1}\|_{1/2} := \left(\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\sqrt{1/n_k}\right)^2, \quad \|\boldsymbol{r}\|_{1/2} := \left(\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\sqrt{r_k}\right)^2$$

Notice:

- Sample-size condition independent of noise σ^2.
- Bound RHS proportional to σ^2.

Threshold behavior in the limit $\sigma^2 \rightarrow 0$

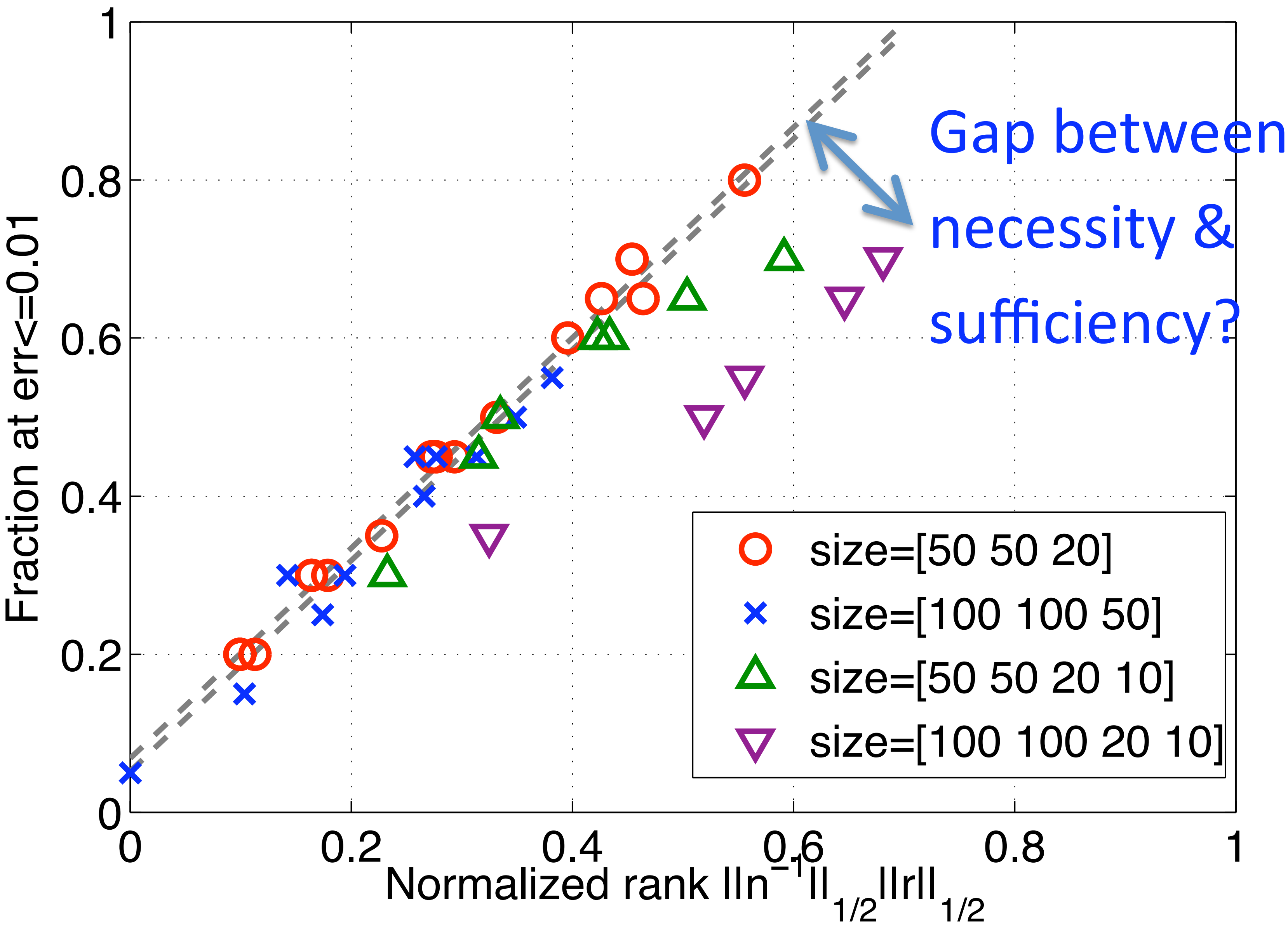
Tensor completion results



No observation noise

Normalized rank

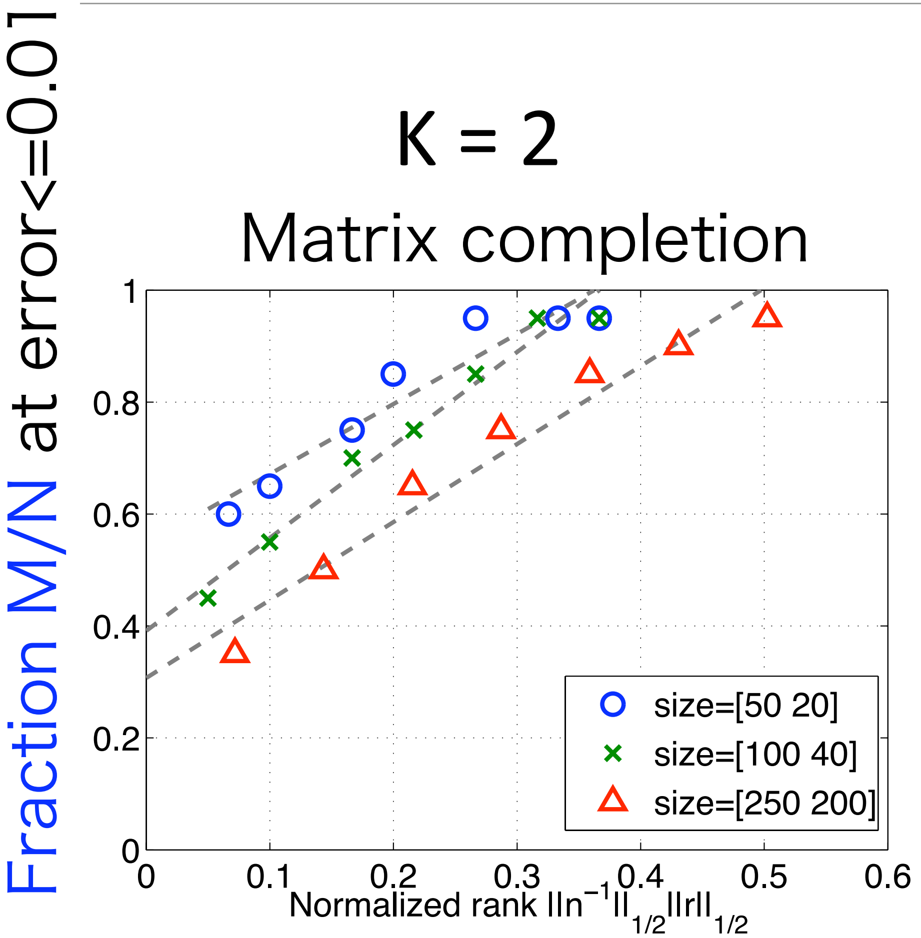
Including 4th order tensors



Matrix / tensor completion

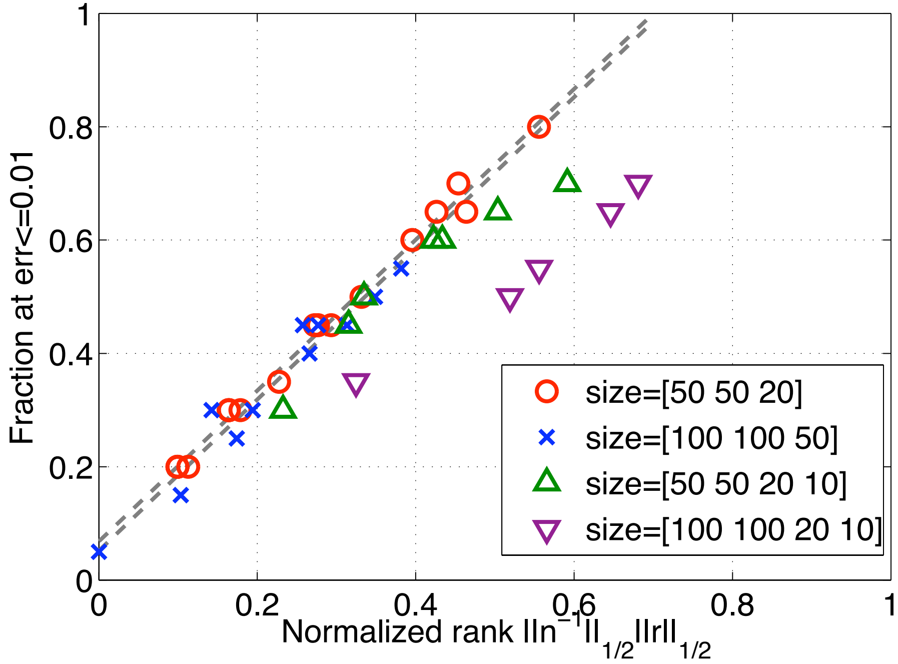
K = 2

Matrix completion



K ≧ 3

Tensor completion



Tensor completion *easier* than matrix completion!?

Conclusion

- Many real world problems can be cast into the form of tensor data analysis.
- Convex optimization is a useful tool also for the analysis of higher order tensors.
- Proposed a convex tensor decomposition algorithm with performance guarantee
- Normalized rank predicts empirical scaling behavior well

Issues

- Why matrix completion more difficult than tensor completion?
- How big the gap between necessity and sufficiency?
- Random Gaussian design ≠ tensor completion

  ⇒ Incoherence (Candes & Recht 09)

  ⇒ Spikiness (Negahban et al 10)
- When only some modes are low-rank
  - Schatten 1-norm is too strong ⇒ Mixture approach
  - E.g. Mode 1, 4 is low rank but the rest is not (combinatorial problem)

Thank you!