Convex Tensor Decomposition with Performance Guarantee

Ryota Tomioka

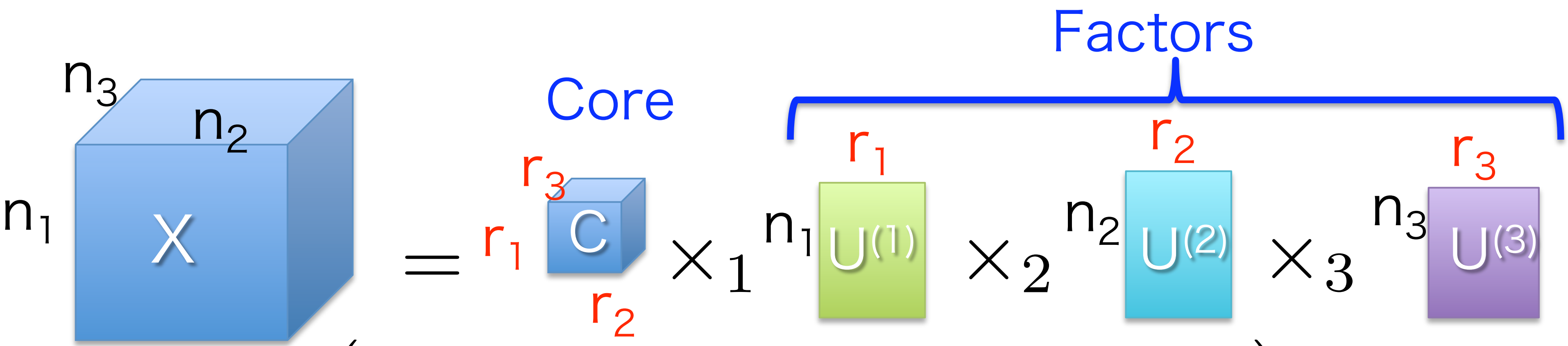
2011/08/16 @ DTU

Collaborators: Taiji Suzuki, Kohei Hayashi, Hisashi Kashima

University of Tokyo & NAIST

Tucker decomposition [Tucker 66]

- Problem: Given a partially observed approximately low-rank tensor *X*, find



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

- Applications: chemo-/psycho-metrics, signal processing, computer vision, neuroscience
- Estimation: alternate minimization (non-convex)

Schatten 1-norm regularization

- Convex optimization problem

$$L(\boldsymbol{W}) + \lambda \|\boldsymbol{W}\|_{S_1}$$

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

(Linear sum of singular-values)

- Applications
  - Collaborative filtering [Srebro et al 05],
  - Multi-task learning [Argyriou et al. 07],
  - Classification over matrices [Tomioka et al. 07]
- Theoretical guarantee
  - Recht et al. 07, Bach 08, Rohde & Tsybakov 11, Negahban & Wainwright 11

Our approach

Matrix

Estimation of
low-rank matrix
(hard)

Trace norm
minimization
(tractable)
[Fazel, Hindi, Boyd 01]

Generalization

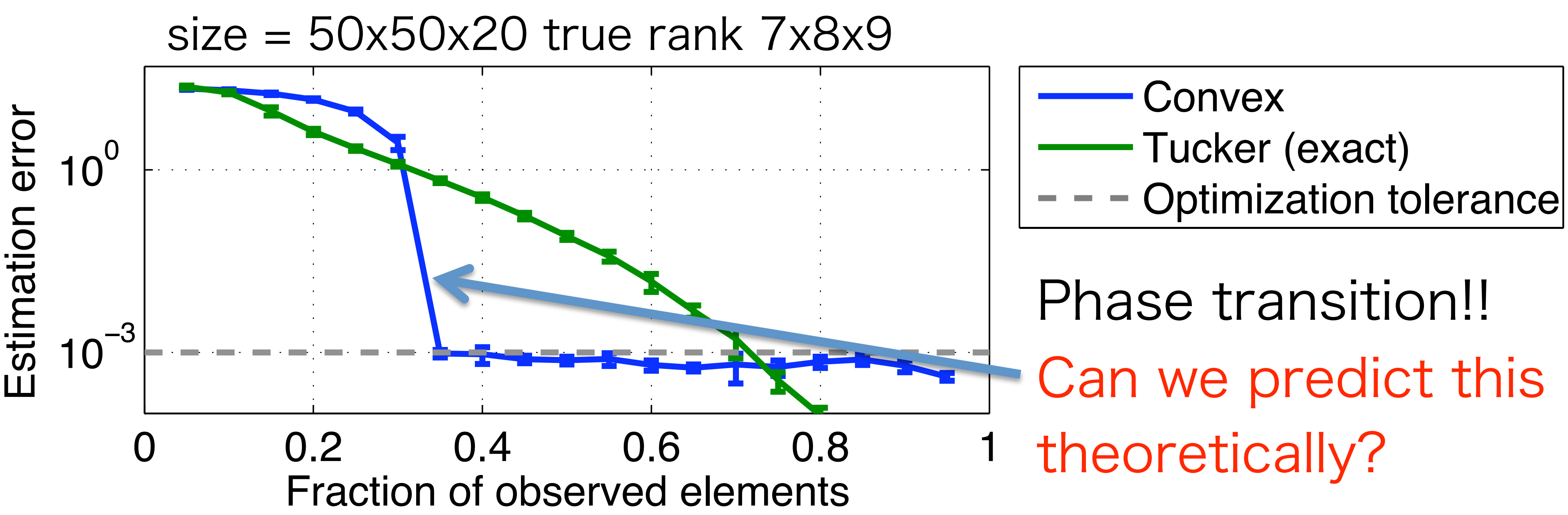
Tensor

Estimation of
low-rank tensor
(hard)
Rank defined in the sense of
Tucker decomposition

Extended
trace norm
minimization
(tractable)

Convex tensor decomposition

- Schatten 1-norm minimization [Liu+09, Signoretto +10, Tomioka+10, Gandy+11]
- Tensor completion result [Tomioka+10]



Phase transition!!

Can we predict this theoretically?

Problem setting

Observation model

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

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

Gaussian noise $N(0, \sigma^2)$

Optimization

Empirical error Regularization

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

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

Observation model

Reg. Const.

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

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

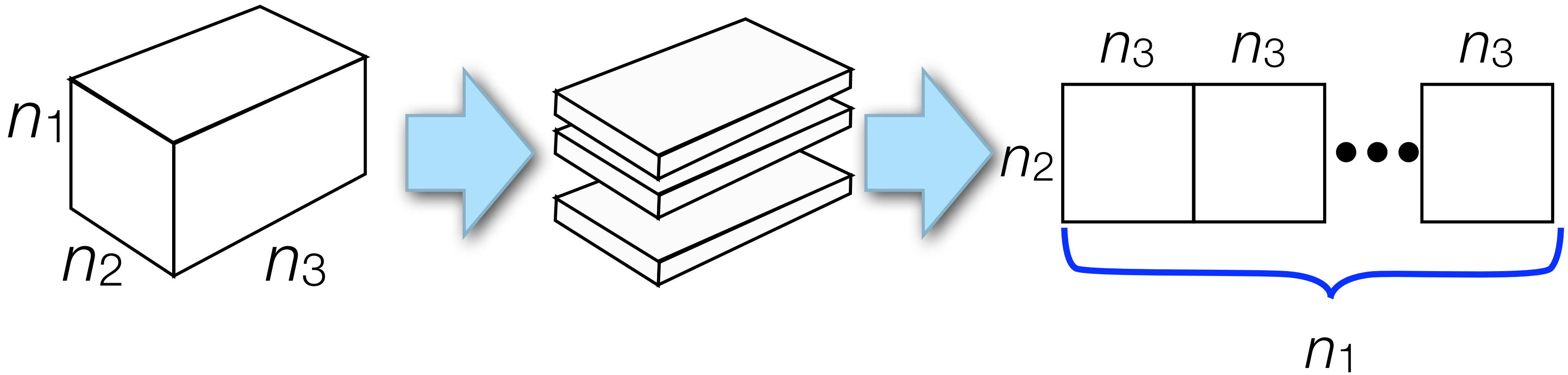
Schatten 1-norm for Tensors

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

Schatten 1-norm for the mode-k unfolding

Example of unfolding (matricization)

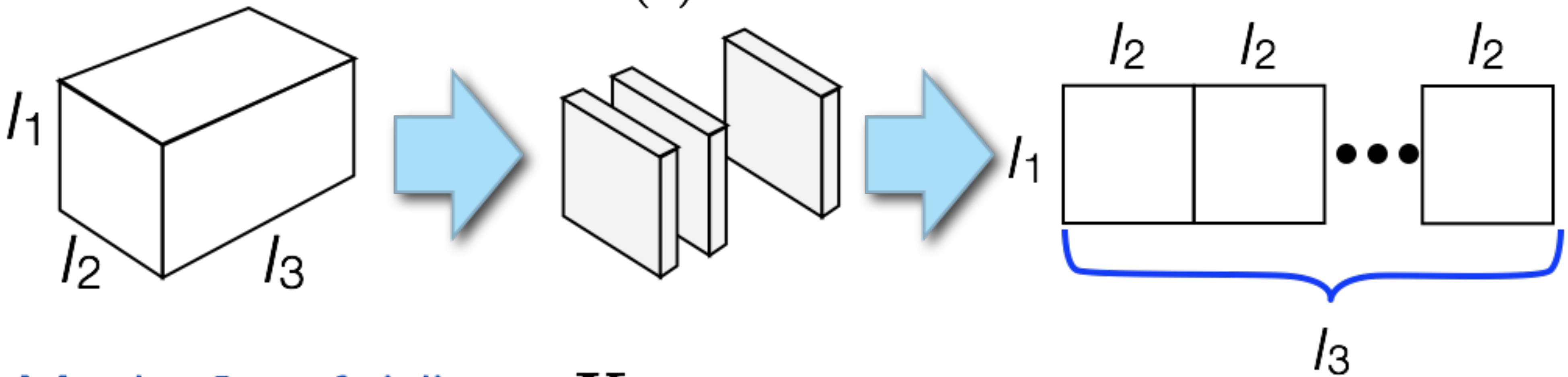
Mode-2 unfolding $\boldsymbol{X}_{(2)}$



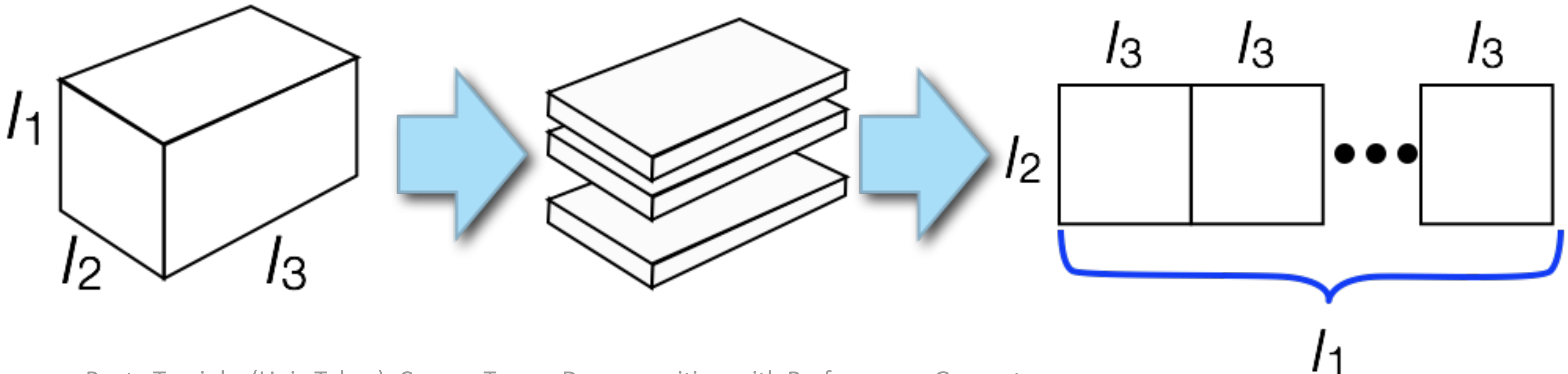
NB: rank of mode-k unfolding = mode-k rank r_k

Mode-k unfolding (matricization)

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



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



Restricted strong convexity (RSC)

(cf. Negahban & Wainwright 11)

- Assume that there is a positive constant κ(X) such that for all tensors $\Delta \in C$

$$\frac{1}{M}\|\mathfrak{X}(\Delta)\|_2^2 \geq \kappa(\mathfrak{X}) |||\Delta|||_F^2$$

(The set C needs to be defined carefully)

Note:

- If $C=R^N$, $\kappa(X)=\min \mathrm{eig}(X^T X)$ $(X \in R^{M \times N})$
- When M<N, restriction is necessary.
- The smaller C, the weaker the assumption.

Theorem 1 (deterministic)

- Solution of the opt. problem $\hat{\mathcal{W}}$
- Reg const λ_M satisfies

$$\lambda_M \geq 2 |||\mathfrak{X}^*(\boldsymbol{\epsilon})|||_{\mathrm{mean}} / M$$

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

$$|||\boldsymbol{\mathcal{X}}|||_{\mathrm{mean}} := \frac{1}{K} \sum_{k=1}^{K} \| \boldsymbol{X}_{(k)} \|_{S_\infty}$$

- Under the RSC assumption

$$|||\hat{\mathcal{W}} - \mathcal{W}^*|||_{\mathrm{F}} \leq \frac{32 \lambda_M}{\kappa(\mathfrak{X})} \frac{1}{K} \sum_{k=1}^{K} \sqrt{r_k}$$

A key inequality

$$\mathcal{W}, \mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$$

$$\langle \mathcal{W}, \mathcal{X} \rangle \leq |||\mathcal{W}|||_{S_1} |||\mathcal{X}|||_{\text{mean}}$$

where

$$|||\mathcal{W}|||_{S_1} := \frac{1}{K} \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}\|_{S_1} \qquad |||\mathcal{X}|||_{\text{mean}} := \frac{1}{K} \sum_{k=1}^{K} \|\boldsymbol{X}_{(k)}\|_{S_\infty}$$

K=2: norm duality (tight)

K>2: not tight

$$\|\boldsymbol{X}\|_{S_1} := \sum_{j=1}^{m} \sigma_j(\boldsymbol{X})$$

$$\|\boldsymbol{X}\|_{S_\infty} := \max_{j \in \{1,\ldots,m\}} \sigma_j(\boldsymbol{X})$$

Proof outline

Since $\hat{\mathcal{W}}$ is a minimizer

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

$$\boldsymbol{\Delta}=\hat{\boldsymbol{\mathcal{W}}}-\boldsymbol{\mathcal{W}}^*$$

Error (fixed design)

noise-design correlation

$$\underbrace{\boxed{\frac{1}{2M}\|\mathfrak{X}(\boldsymbol{\Delta})\|_2^2}}_{\text{RIC}\ \geq \frac{\kappa(\mathfrak{X})}{2}\left|\!\left|\!\left|\boldsymbol{\Delta}\right|\!\right|\!\right|_F^2} \leq \underbrace{\boxed{\left|\!\left|\!\left|\mathfrak{X}^*(\boldsymbol{\epsilon})/M\right|\!\right|\!\right|_{\mathrm{mean}}}}_{\leq \frac{\lambda_M}{2}}\left|\!\left|\!\left|\boldsymbol{\Delta}\right|\!\right|\!\right|_{S_1}+\lambda_M\left|\!\left|\!\left|\boldsymbol{\Delta}\right|\!\right|\!\right|_{S_1}$$

Proof outline

Since $\hat{\mathcal{W}}$ is a minimizer

$$\frac{1}{2M}\|\boldsymbol{y}-\mathfrak{X}(\hat{\mathcal{W}})\|_2^2+\lambda_M\left|\!\left|\!\left|\hat{\mathcal{W}}\right|\!\right|\!\right|_{S_1}\le\frac{1}{2M}\|\boldsymbol{y}-\mathfrak{X}(\mathcal{W}^*)\|_2^2+\lambda_M\left|\!\left|\!\left|\mathcal{W}^*\right|\!\right|\!\right|_{S_1}$$

$$\boldsymbol{\Delta}=\hat{\boldsymbol{\mathcal{W}}}-\boldsymbol{\mathcal{W}}^*$$

$$\frac{\kappa(\mathfrak{X})}{2}\left|\!\left|\!\left|\boldsymbol{\Delta}\right|\!\right|\!\right|_F^2\le 2\lambda_M\left|\!\left|\!\left|\boldsymbol{\Delta}\right|\!\right|\!\right|_{S_1}\le 8\lambda_M\left|\!\left|\!\left|\boldsymbol{\Delta}\right|\!\right|\!\right|_F\frac{1}{K}\sum_{k=1}^{K}\sqrt{2r_k}$$

$$\left|\!\left|\!\left|\boldsymbol{\Delta}\right|\!\right|\!\right|_F\le\frac{32\lambda_M}{\kappa(\mathfrak{X})}\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k}$$

Choosing the set C

- We only need the residual Δ to be in C

$$\boldsymbol{\Delta}_{(k)} = \boldsymbol{\Delta}'_k + \boldsymbol{\Delta}''_k$$

mode-k unfolding of the residual

Component spanned by the truth

Orthogonal to the truth

Lemma 2. *Let $\hat{\mathcal{W}}$ be the solution of the minimization problem* (7) *with* $\lambda_M \geq 2\left\|\!\left|\mathfrak{X}^*(\boldsymbol{\epsilon})\right|\!\right\|_{\mathrm{mean}}/M$*, and let* $\Delta := \hat{\mathcal{W}} - \mathcal{W}^*$*, where* $\mathcal{W}^*$ *is the true low-rank tensor. Let* $\boldsymbol{\Delta}_{(k)} = \boldsymbol{\Delta}'_k + \boldsymbol{\Delta}''_k$ *be the decomposition defined in Equation* (4)*. Then for all* $k = 1, \ldots, K$ *we have the following inequalities:*

1. $\mathrm{rank}(\boldsymbol{\Delta}'_k) \leq 2r_k$.

2. $\sum_{k=1}^{K} \|\boldsymbol{\Delta}''_k\|_{S_1} \leq 3\sum_{k=1}^{K} \|\boldsymbol{\Delta}'_k\|_{S_1}$.

Two special cases

- Noisy tensor decomposition (M=N)
  - RSC: trivial.
  - Choose λ depending on the noise-design correlation term $|||\mathfrak{X}^*(\boldsymbol{\epsilon})|||_{\mathrm{mean}}$

- Random Gauss design
  - RSC: more difficult.
  - Choose λ depending on the noise-design correlation term $|||\mathfrak{X}^*(\boldsymbol{\epsilon})|||_{\mathrm{mean}}$

Noisy tensor decomposition

- All the elements are observed once (M=N) with noise.

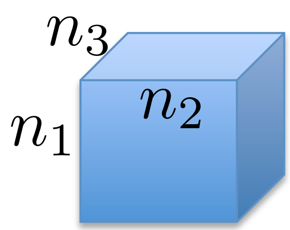
$$\|\mathfrak{X}(\Delta)\|_2^2 = \|\!|\Delta|\!\|_F^2 \quad \Rightarrow \quad \kappa(\mathfrak{X}) = 1/M$$

(RSC)

- Regularization const.

$$\lambda_M \geq 2 |\!|\!| \mathfrak{X}^*(\epsilon) |\!|\!|_{\mathrm{mean}} / M$$

$$\mathbb{E} |\!|\!| \mathfrak{X}^*(\epsilon) |\!|\!|_{\mathrm{mean}} \leq \frac{\sigma}{K} \sum_{k=1}^{K} \left(\sqrt{n_k} + \sqrt{N/n_k} \right)$$



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

(Using random matrix theory)

In addition, $|\!|\!| \mathfrak{X}^*(\epsilon) |\!|\!|_{\mathrm{mean}}$ concentrates around its mean with high probability

Theorem 2

- When all the elements are observed (M=N) and the regularization const. satisfies

$$\lambda_M \geq \frac{2\sigma}{K} \sum_{k=1}^{K} \left(\sqrt{n_k} + \sqrt{N/n_k} \right) / N$$

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

where

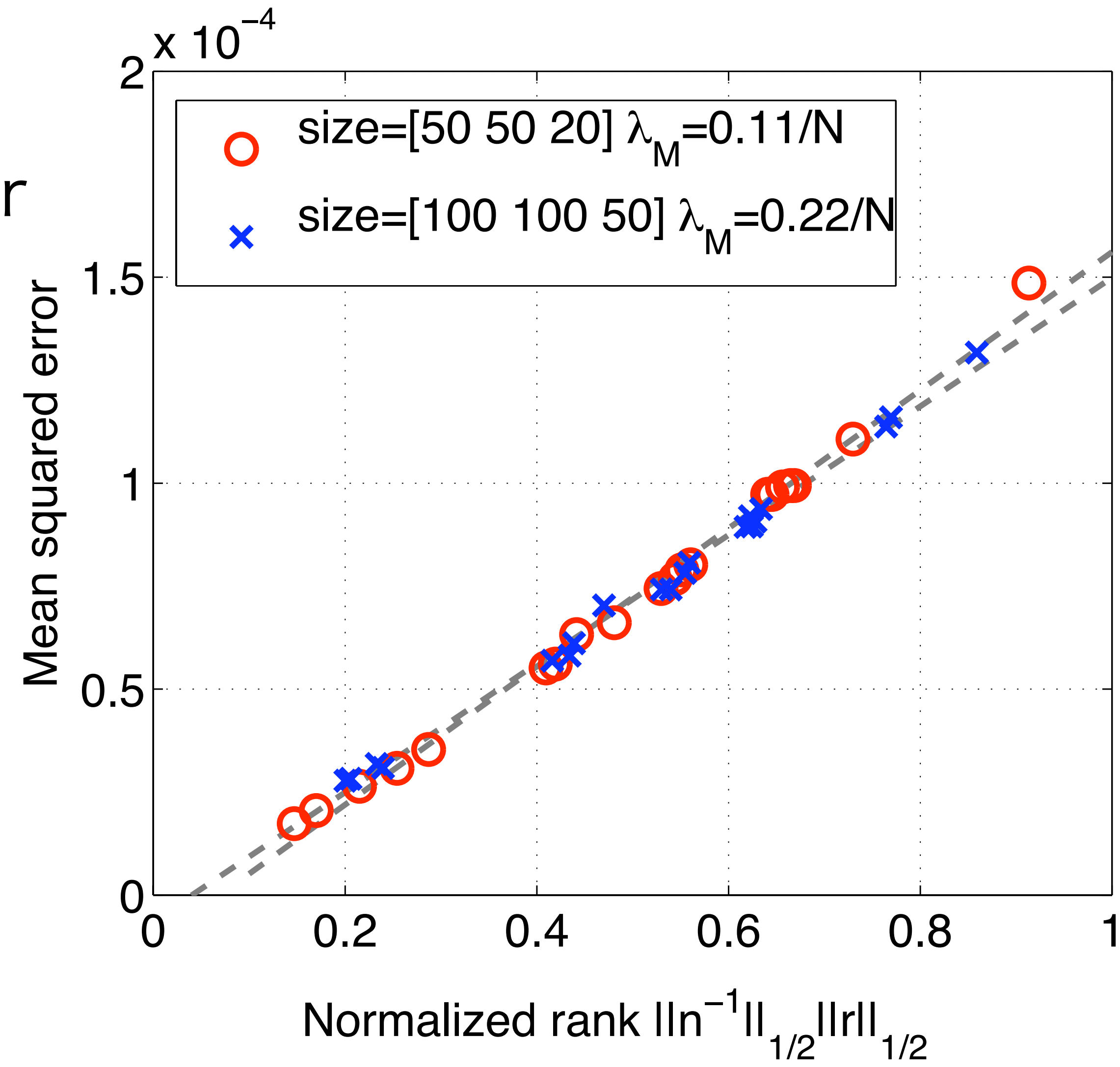
$$\|\boldsymbol{n}^{-1}\|_{1/2} := \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{1/n_k} \right)^2, \quad \|\boldsymbol{r}\|_{1/2} := \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{r_k} \right)^2$$

Noisy tensor decomposition (σ=0.01)

Mean squared error

$$\frac{\left\| \hat{\mathcal{W}} - \mathcal{W}^* \right\|_F^2}{N}$$

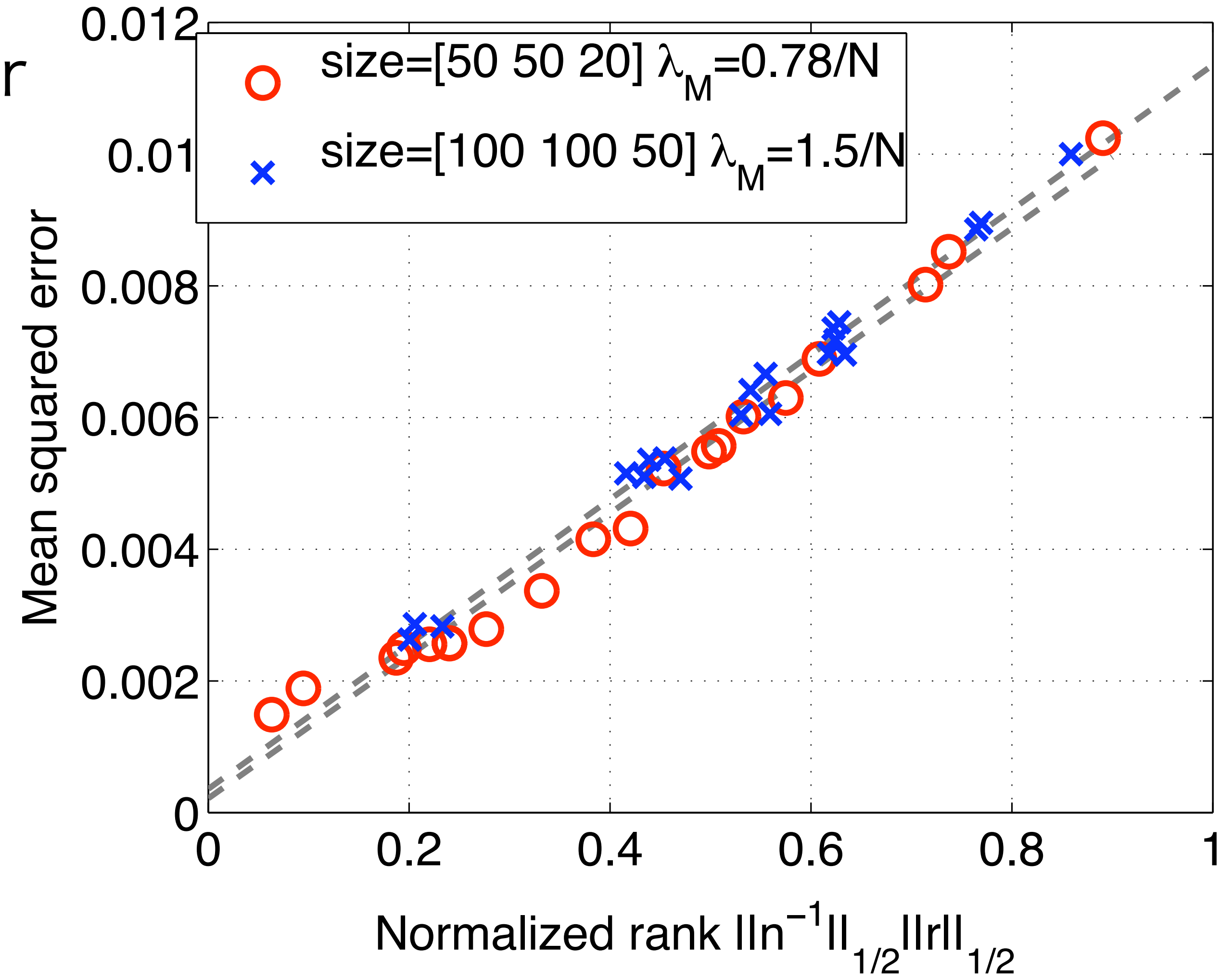
- Theoretical scaling of the reg. const. only depends on the size and not on the rank.
- MSE grows linearly with the normalized rank



Noisy tensor decomposition (σ=0.1)

Mean squared error

$$\frac{\|\hat{\mathcal{W}} - \mathcal{W}^*\|_F^2}{N}$$

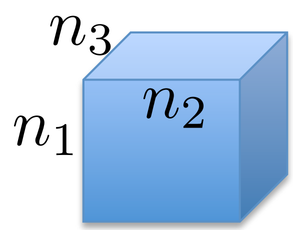


- Theoretical scaling of the reg. const. only depends on the size and not on the rank.
- MSE grows linearly with the normalized rank

Random Gauss design

- Elements of X_i are iid standard Gaussian
- Regularization constant $\lambda_M \geq 2 |||\mathfrak{X}^*(\boldsymbol{\epsilon})|||_{\mathrm{mean}} / M$

$$\mathbb{E}|||\mathfrak{X}^*(\boldsymbol{\epsilon})|||_{\mathrm{mean}} \leq \frac{\sigma\sqrt{M}}{K} \sum_{k=1}^{K} \left(\sqrt{n_k} + \sqrt{N/n_k} \right)$$



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

- RSC (more involved)

Sufficient condition:

$$\frac{M}{N} \geq c \|\boldsymbol{n}^{-1}\|_{1/2} \|\boldsymbol{r}\|_{1/2} \qquad (\kappa(X) = 1/64)$$

constant

Normalized rank

(M=#samples, N=#variables)

Theorem: random Gauss design

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

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

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

Theorem: random Gauss design

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

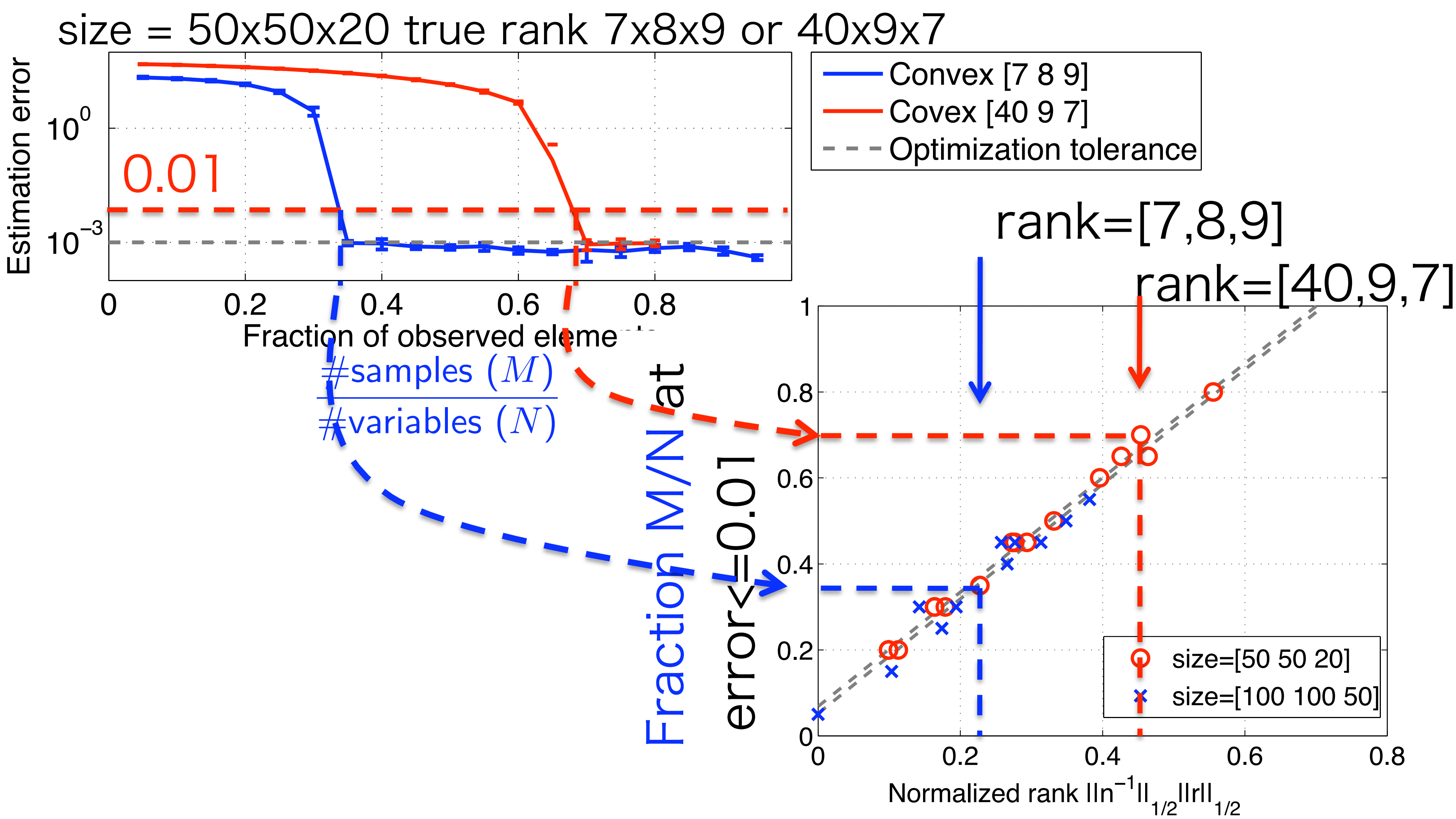
$$\frac{\#\text{samples } (M)}{\#\text{variables } (N)} \geq c_1 \underbrace{\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}_{\text{Normalized rank}}$$

Convergence!

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

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

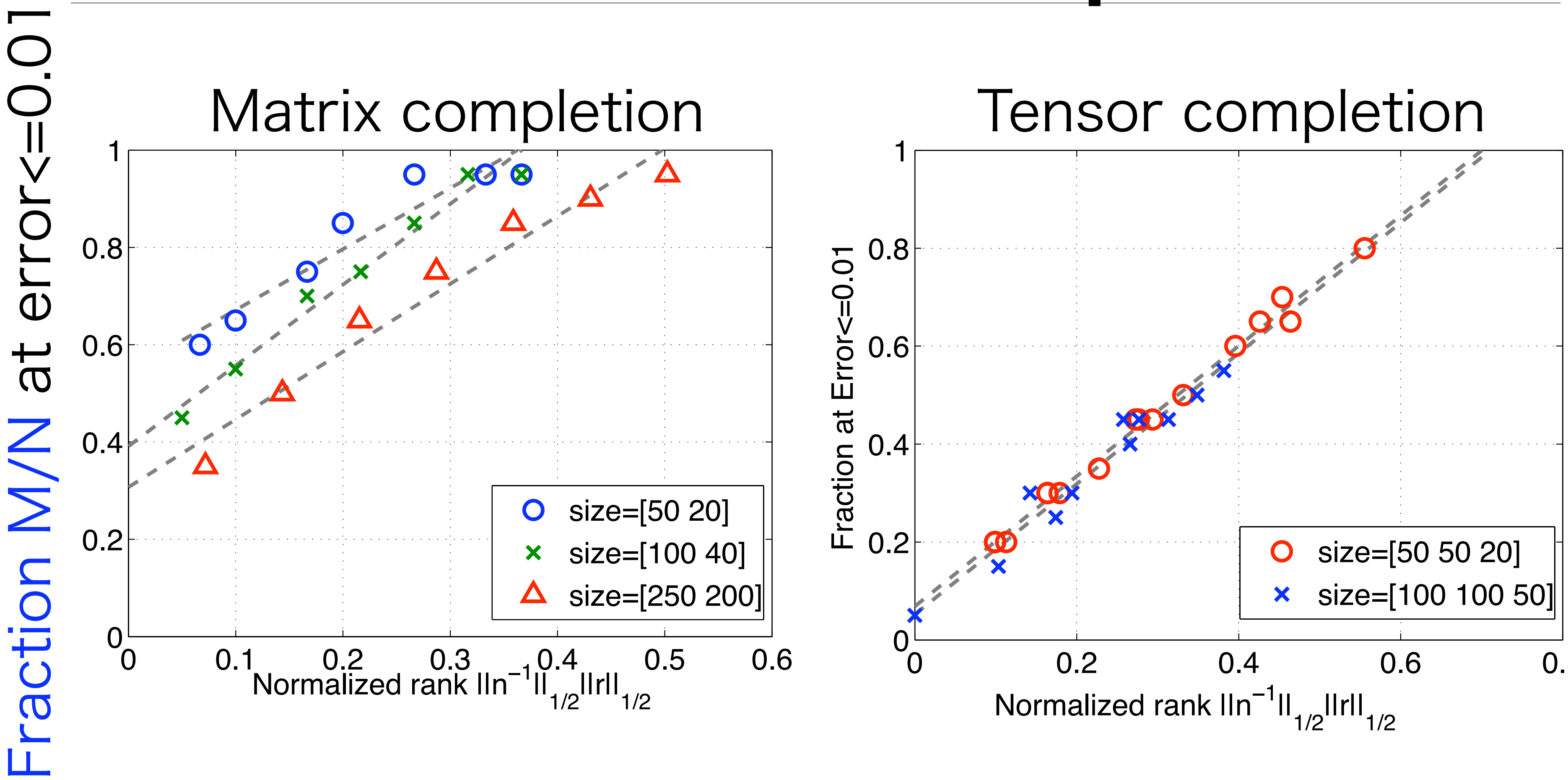
Tensor completion results



No observation noise

Normalized rank

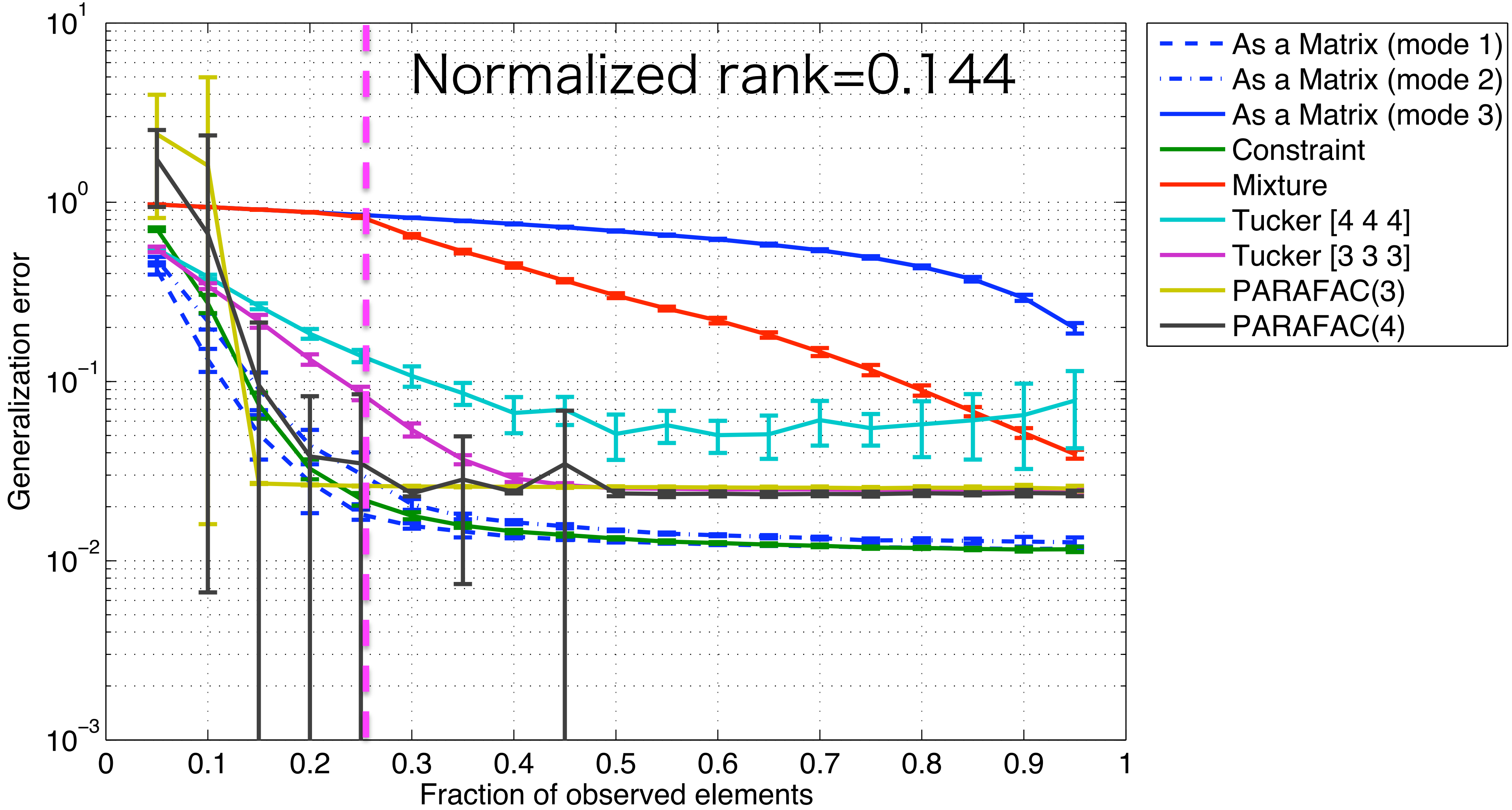
Matrix / tensor completion



Tensor completion *easier* than matrix completion!?

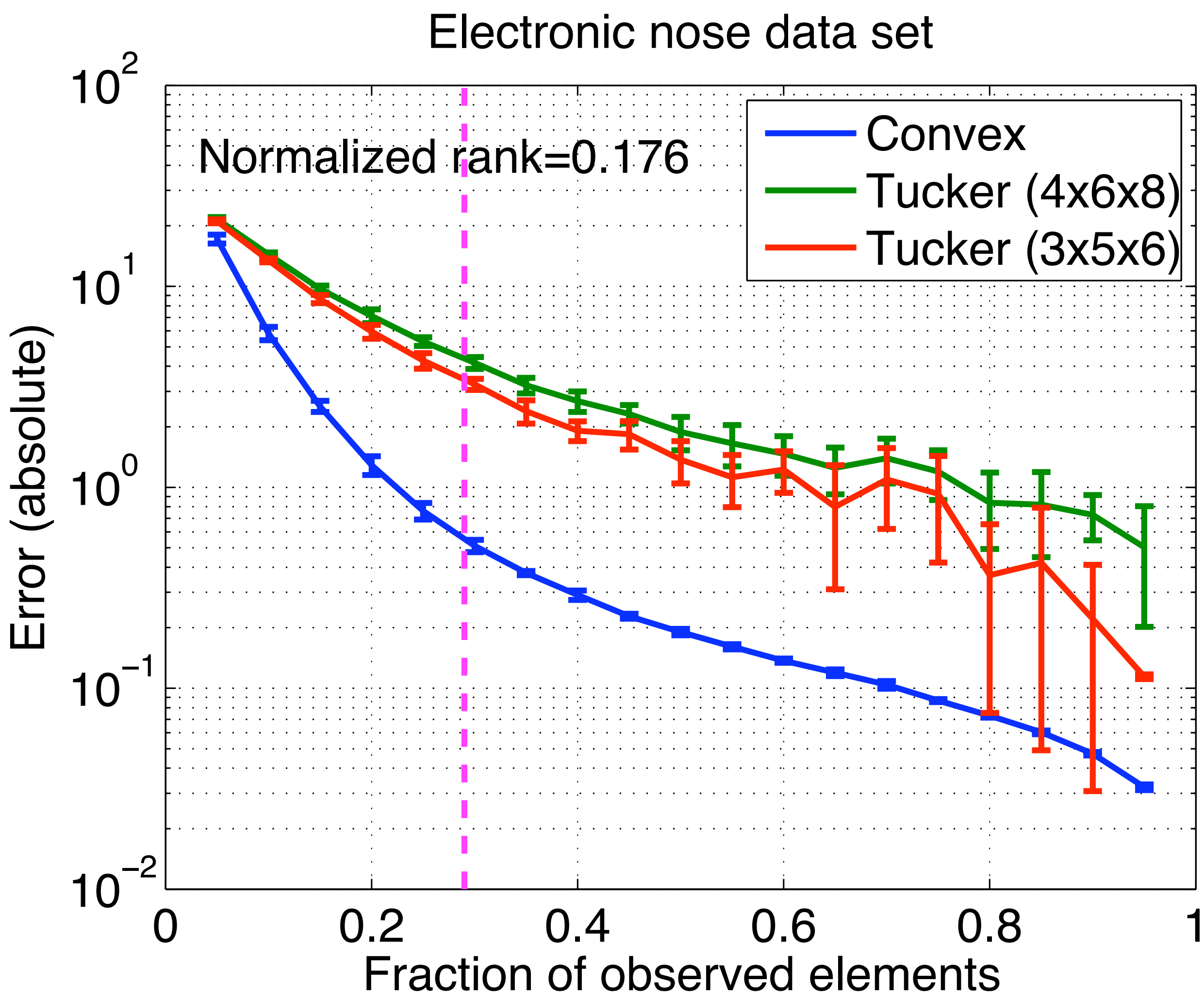
Amino acid fluorescence dataset [Bro 97]

- Size=5x51x201 rank=3x3x3



Eletronic nose data [Skob & Bro 04]

- Size=18x241x12 rank=3x5x6 (guessed)



Conclusion

- Convex tensor decomposition --- now with performance guarantee
- Normalized rank predicts empirical scaling behavior well

Issues

- Why matrix completion more difficult than tensor completion?
- Worst case analysis-> average case analysis (stat physics method?)

More issues

- 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)
- Other loss functions
  - Sparse noise （anomaly detection from video）
  - Low-rank classifier over tensors

Approach 3: Mixture of low-rank tensors

- Each mixture component Z_k is regularized to be low-rank only in mode-k.

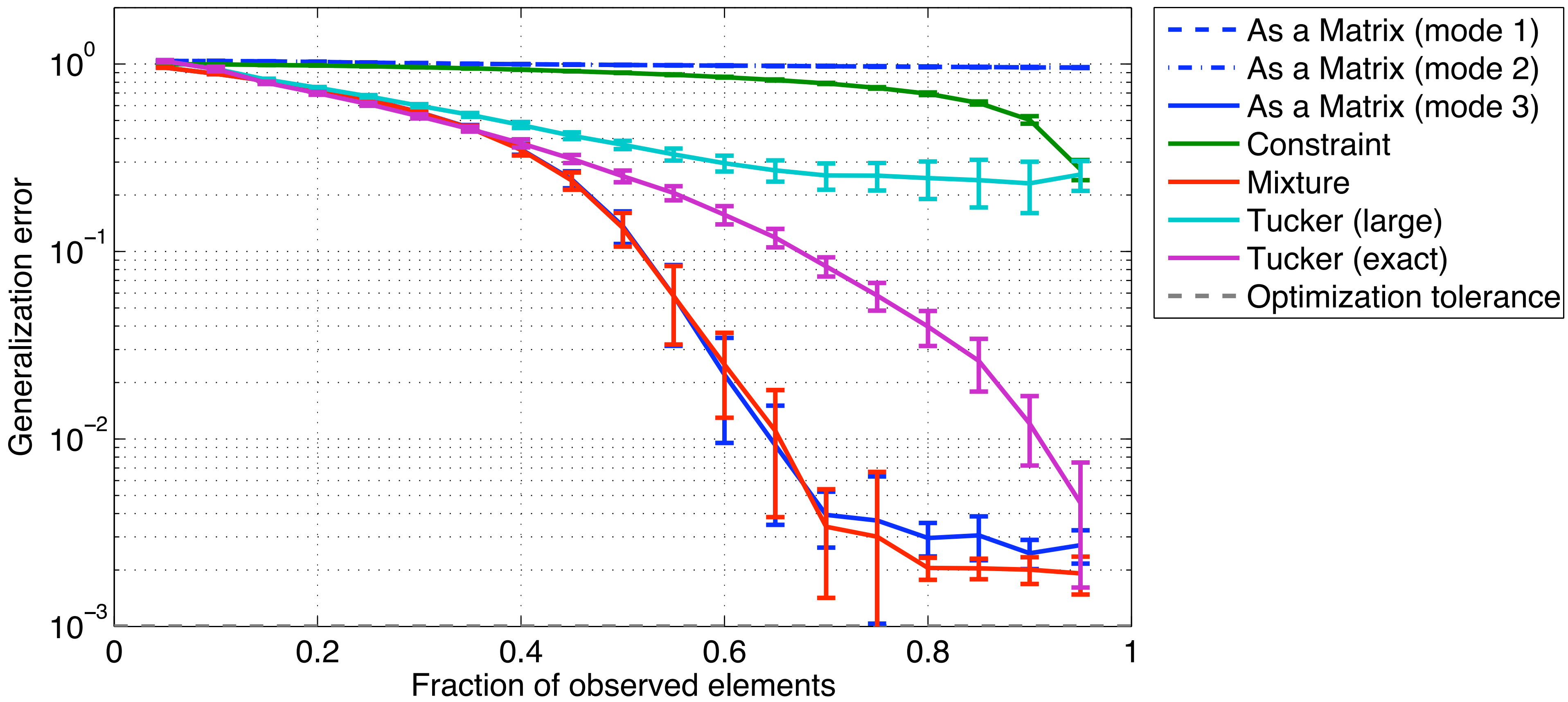
$$\underset{\mathcal{Z}_1,\ldots,\mathcal{Z}_K}{\text{minimize}} \quad \frac{1}{2\lambda}\left\|\Omega \circ \left(\mathcal{Y} - \sum\nolimits_{k=1}^{K} \mathcal{Z}_k\right)\right\|_F^2 + \sum_{k=1}^{K} \gamma_k \|\boldsymbol{Z}_{k(k)}\|_*,$$

Pro: Each Z_k takes care of each mode

Con: Sum is not low-rank

Mixture is sometimes better

True tensor: Size 50x50x20, rank 50x50x5. No noise (λ=0).



Singular value shrinkage

$$\mathrm{softth}(\boldsymbol{X}) = \underset{\boldsymbol{Z}\in\mathbb{R}^{R\times C}}{\mathrm{argmin}} \left(\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{X}\|_F^2 + \lambda\|\boldsymbol{Z}\|_*\right)$$

$$= \boldsymbol{U}\max(\boldsymbol{S}-\lambda, 0)\boldsymbol{V}^\top$$

where $X=USV^T$