

行列およびテンソルデータ に対する機械学習

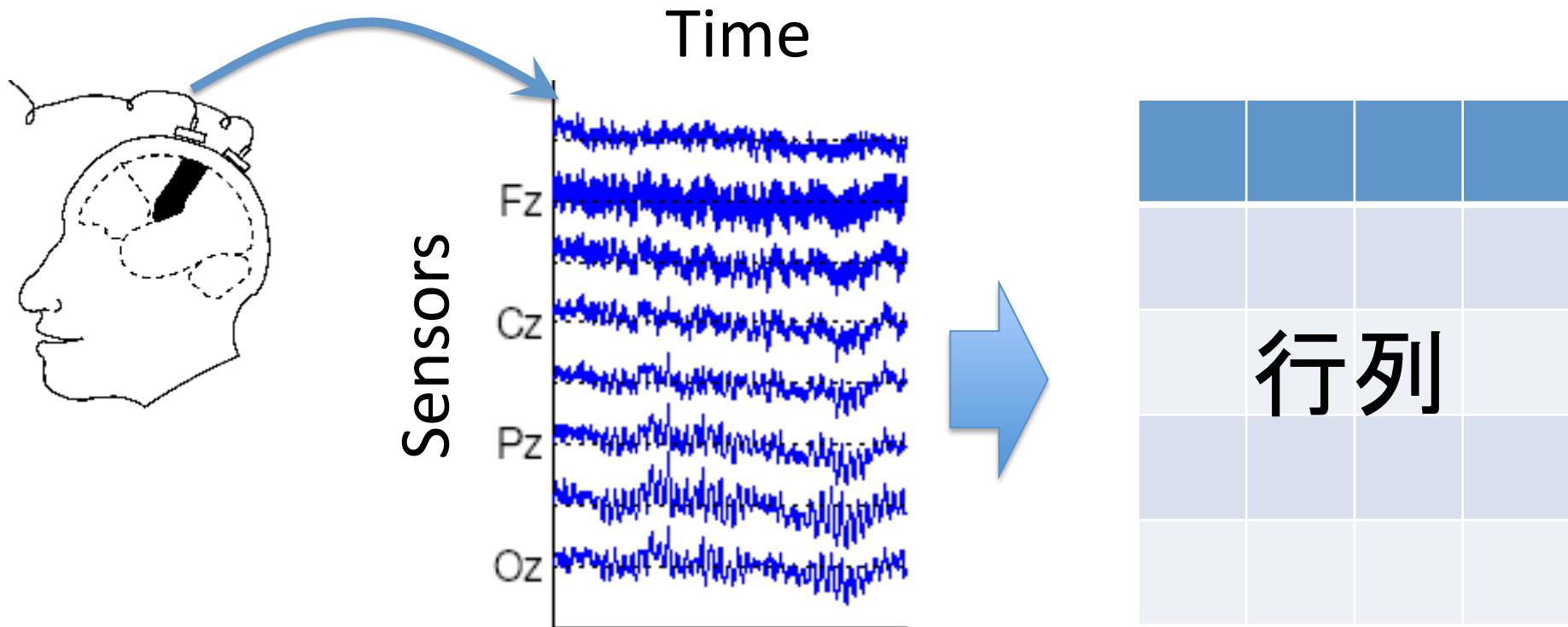
数理6研 富岡亮太

2011/11/28 数理助教の会

自己紹介

- 学部: 計数工学科システムコース
- 卒論 & 修論: 遺伝子ネットワークにおける確率的なゆらぎの研究(木村先生 & 合原先生)
- 博士から機械学習の研究

行列データの例1



- 行と列の構造があるので、ベクトルとして扱うのはもったいない。
- 右と左から行列をかけるのが自然

行列データの例2

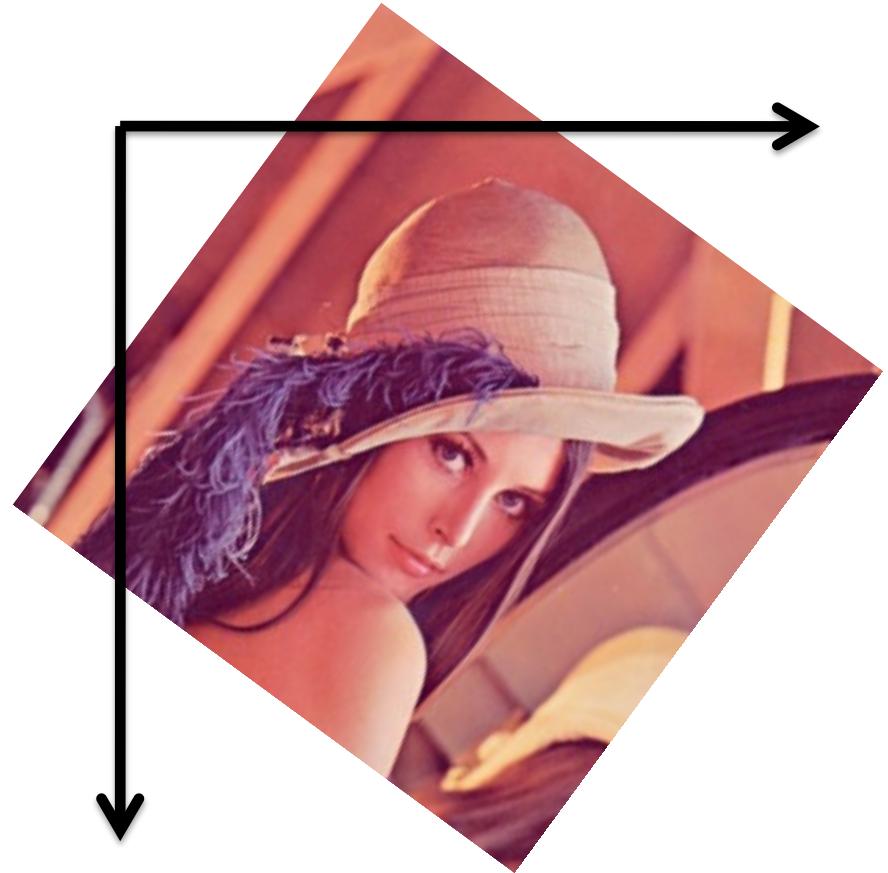
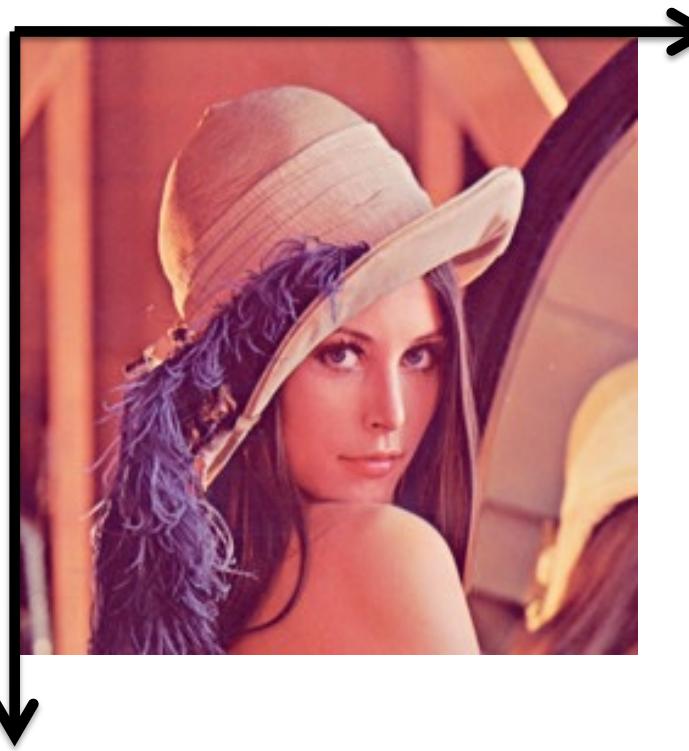
- 商品・ユーザ行列

		映画			
		Star Wars	Titanic	Blade Runner	...
ユーザ	User 1	5	2	4	
	User 2	1	4	2	
	User 3	5	?	?	

(数学的な意味で行列かどうか微妙。)

行列データと考えたくないもの

画像



- 行と列に本質的な差がない

2種類の低ランク分解

- 生成モデル
 - 与えられたひとつの行列を何らかの規準で低ランク近似
 - 応用: 協調フィルタリング、システム同定など

$$X \simeq AB^\top$$

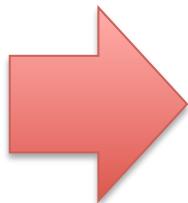
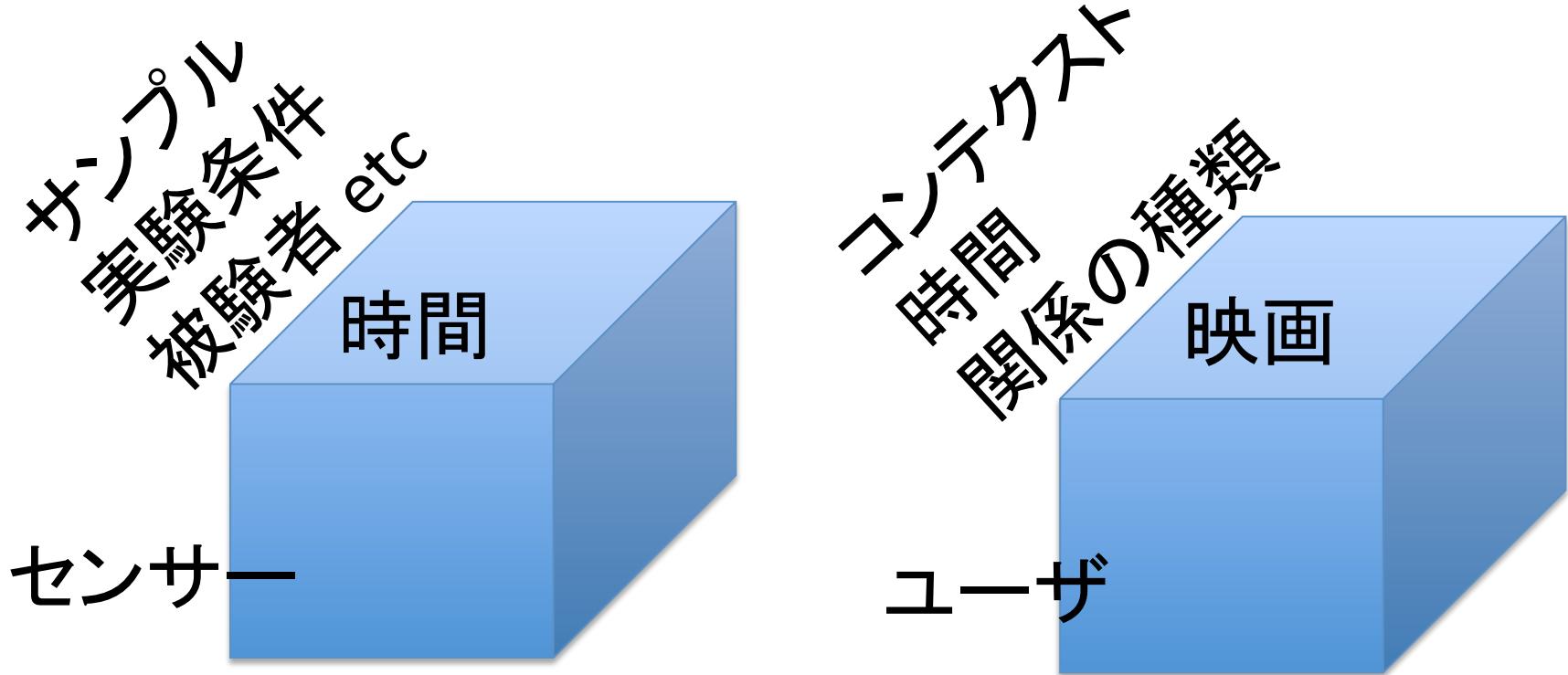
- 判別モデル
 - 行列データを入力として、分類規則を学習
 - 応用: 時空間データの判別(後述)

$$\begin{matrix} X_1 & X_2 & X_3 & X_4 & \cdots \end{matrix}$$

$$f(X) = \langle X, \mathbf{W} \rangle + b$$

$$\mathbf{W} = AB^\top \text{ (低ランク)}$$

テンソル（3軸以上のデータ）



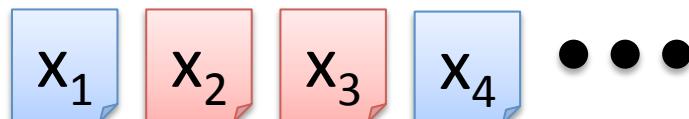
- テンソルの低ランク分解
- テンソルのランクとは？
(応力テンソルとかとは意味が違う)

テンソルの低ランク分解

- 生成モデル
 - 与えられたテンソルを何らかの基準で低ランク近似

$$\mathcal{X} \simeq \mathcal{G} \times_1 U_1 \times_2 U_2 \cdots \times_K U_K$$

- 判別モデル
 - テンソルを入力データとして分類規則を学習



$$f(x) = \underbrace{\langle x, w_1 \rangle}_{\text{1次項}} + \underbrace{\langle xx^\top, W_2 \rangle}_{\text{2次項}} + \cdots$$

行列に対する
生成モデル

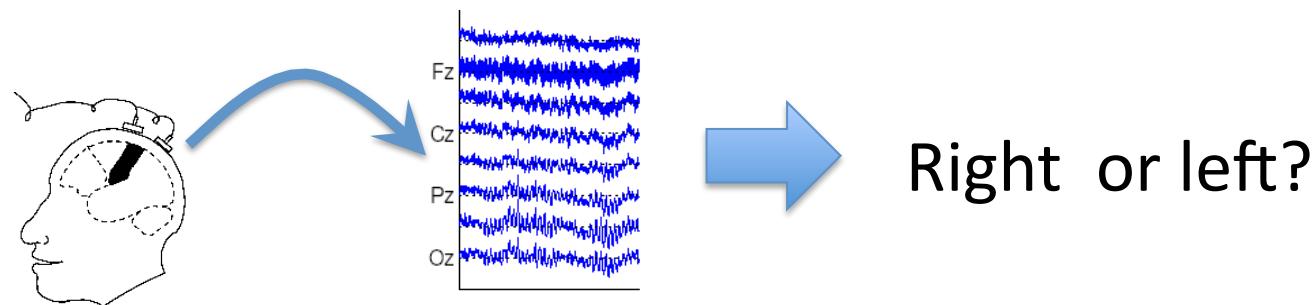
行列に対する
判別モデル
(2006-2009)

テンソルに対する
生成モデル
(2010-)

テンソルに対する
判別モデル
(?)

アジェンダ

- 行列の上の判別モデルに基づくブレイン・コンピュータインターフェースの話(2006-2009)



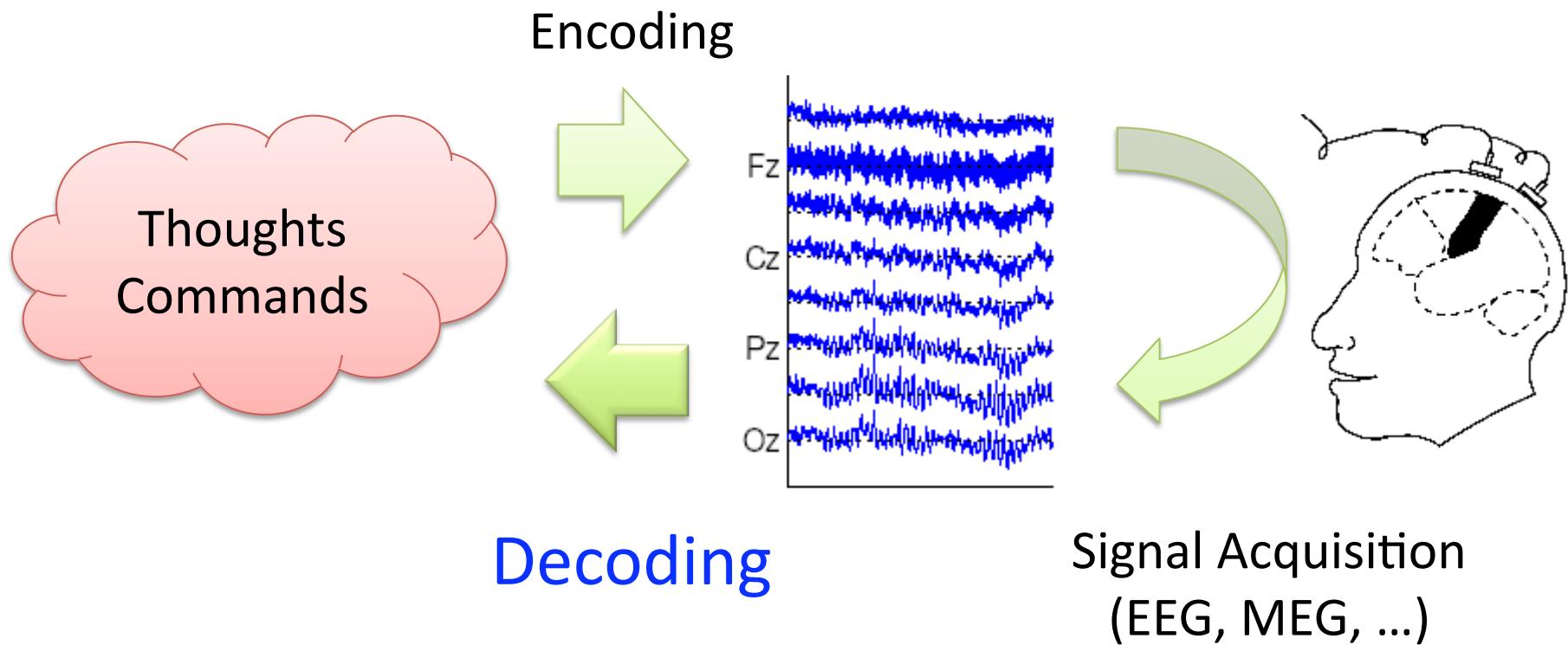
- 高階テンソルのTucker分解を凸化する話(2010-)

$$X = C \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}$$

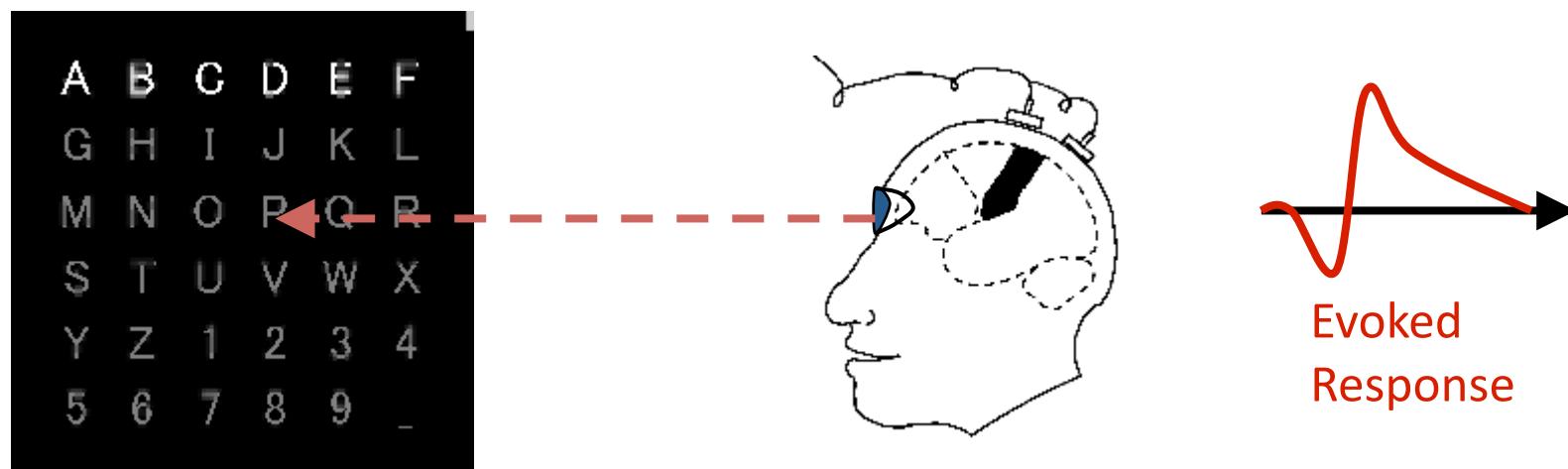
The diagram illustrates the Tucker decomposition of a high-order tensor. On the left, a large blue cube labeled 'X' represents the tensor. Above it, the word 'コア' (core) is written in blue. To the right, the word 'ファクター' (factors) is written in blue. Between the core and the factors, there are three smaller colored cubes: a blue cube labeled 'C' followed by two colored cubes labeled $U^{(1)}$ (green) and $U^{(2)}$ (cyan), which are multiplied together. This is followed by another colored cube labeled $U^{(3)}$ (purple), representing the third factor.

Brain-computer interface

- Aims to “decode” thoughts or commands from human brain signal



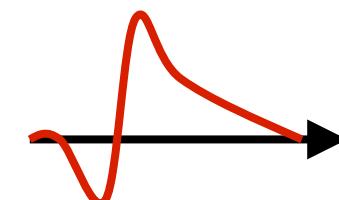
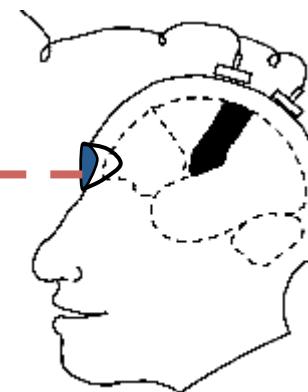
P300 speller system



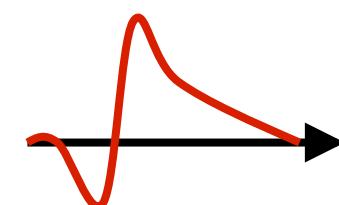
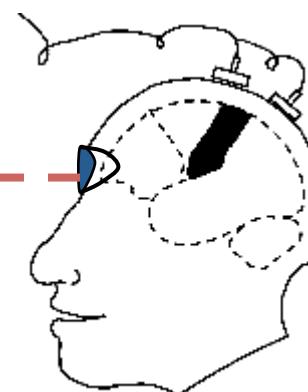
Farwell & Donchin 1988

P300 speller system

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	_



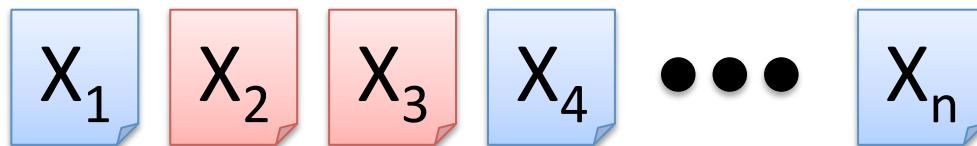
A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	_



The character must be “P”

判別モデル

- 訓練サンプル $(X_1, y_1), \dots, (X_n, y_n)$



- X_i は行列 (センサーの数 × 時間点)
- $y_i = +1$ or -1 (2値分類)

$$\underset{W,b}{\text{minimize}} \quad \sum_{i=1}^n \ell(f(X_i), y_i) + R(W)$$

訓練誤差 正則化

ただし $f(X) = \langle X, W \rangle + b$ (判別器)

低ランク分解を促す正則化

Schatten 1-ノルム (nuclear norm / trace norm)

$$\|\mathbf{W}\|_{S_1} = \sum_{j=1}^r \sigma_j(\mathbf{W}) \quad (\text{特異値の線形和})$$

例えば

$$\begin{aligned} \operatorname{argmin}_{\mathbf{W}} & \left(\frac{1}{2} \|\mathbf{X} - \mathbf{W}\|_F^2 + \lambda \|\mathbf{W}\|_{S_1} \right) \\ & = \mathbf{U} \max(S - \lambda, 0) \mathbf{V}^\top \end{aligned}$$

ただし $\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^\top$ なので、一般の訓練誤差に対しても
低ランク化効果が期待できる（厳密な証明は不明）

低ランク分解する意味

- ・ 時空間フィルタ(特徴抽出器)を学習していることに対応

$$\begin{aligned} W &= U \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} V^\top = \sum_{c=1}^r \sigma_c u_c v_c^\top \\ f(X) &= \left\langle \sum_c \sigma_c u_c v_c^\top, X \right\rangle + b \\ &= \sum_{c=1}^r \color{red}{\sigma_c} \color{blue}{u_c^\top X v_c} + b \end{aligned}$$

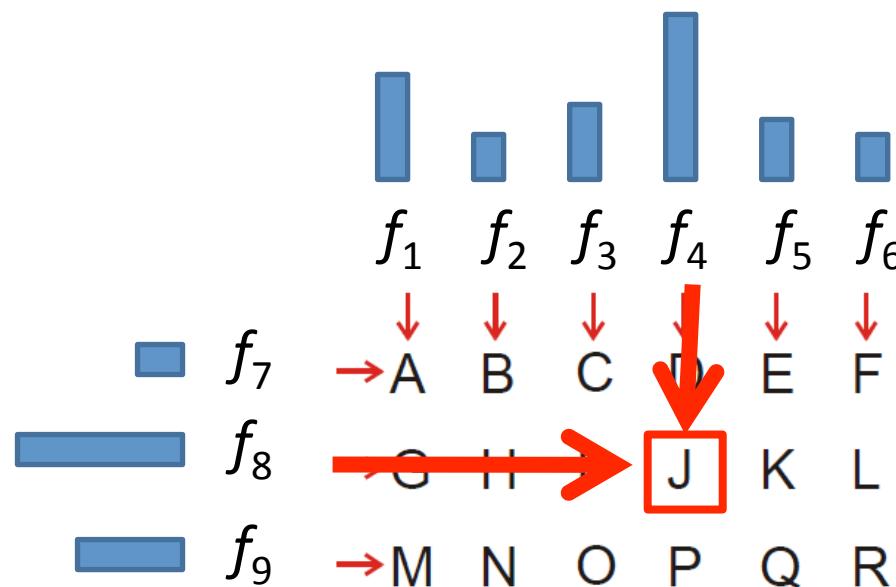
$\uparrow \quad \quad \quad \uparrow$

u_c : 空間フィルタ
 v_c : 時間フィルタ

学習結果がまとまか
「見て」判断できる

Modeling P300 speller (decoding)

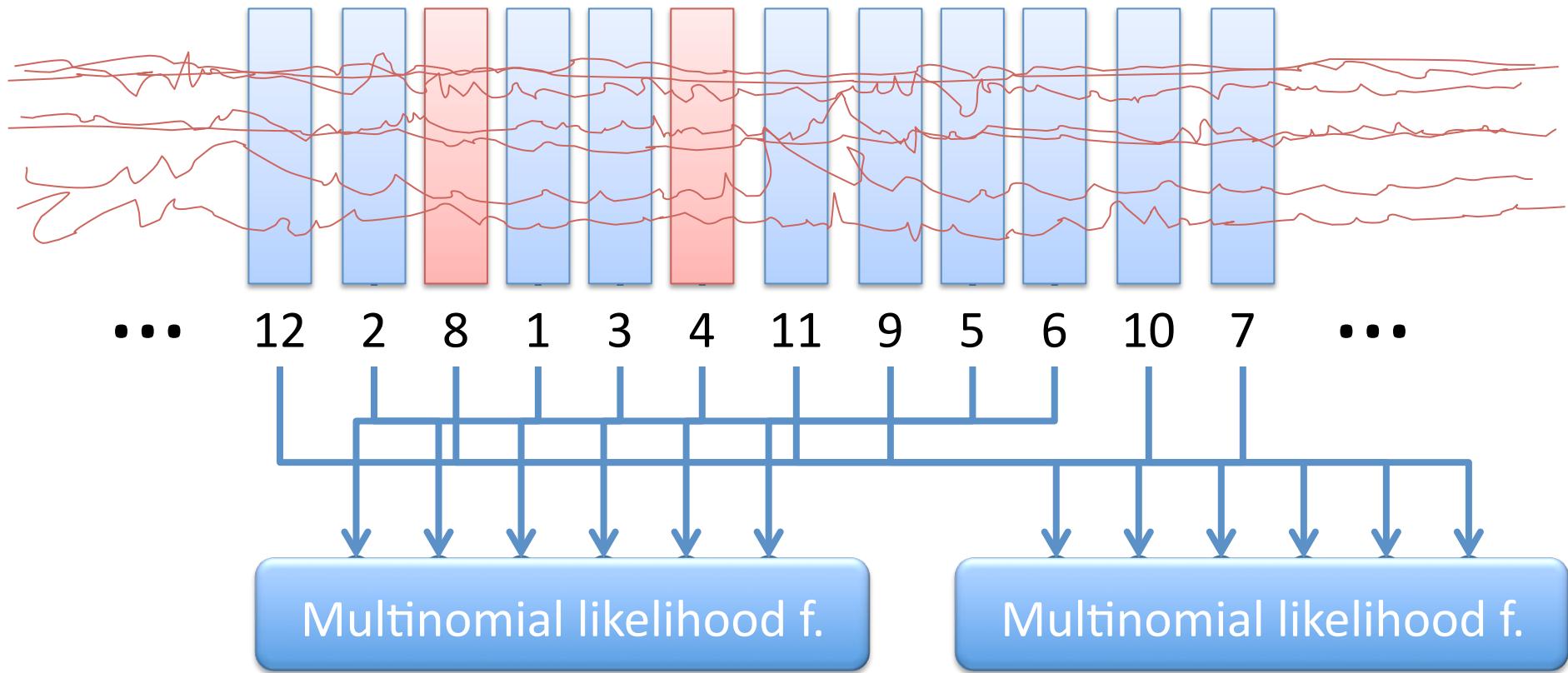
- Suppose that we have a **detector** $f(X)$ that detects the P300 response in signal X .



This is nothing but learning 2×6 -class classifier

f_{11}	\rightarrow	Y	Z	1	2	3	4
f_{12}	\rightarrow	5	6	7	8	9	_

How we do this



$$L(w) = \sum_{i=1}^n (-\log P_w(\text{col} | X_i) + -\log P_w(\text{row} | X_i))$$

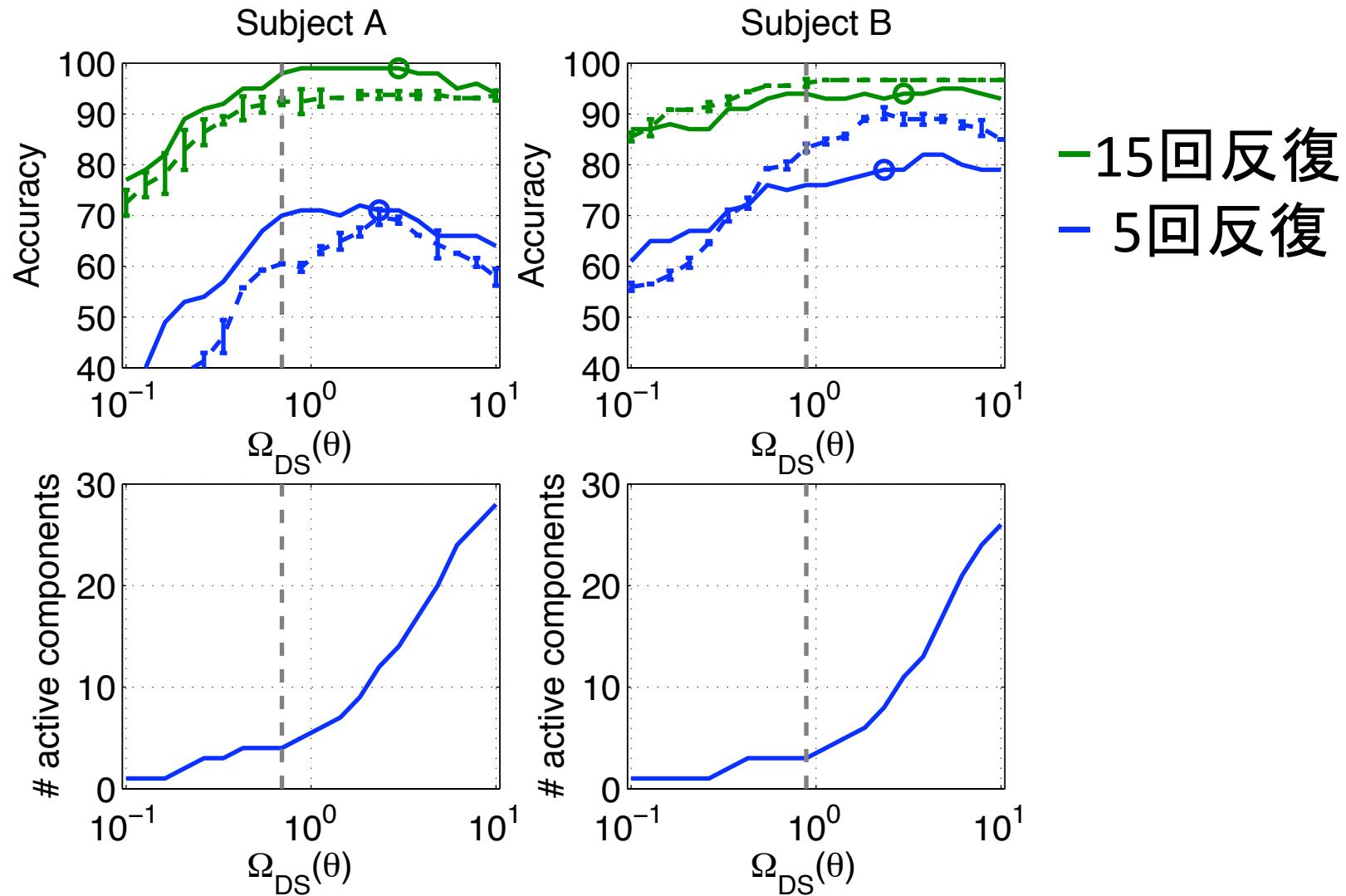
Experiment

- Two subjects (A&B) from BCI competition III
 - 64 channels x 37 time-points (600ms @ 60Hz)
 - 12 epochs x 15 repetitions x 85 letters = 15300 epochs in training set
 - 100 letters for test
- Linear detector function (bias is irrelevant)

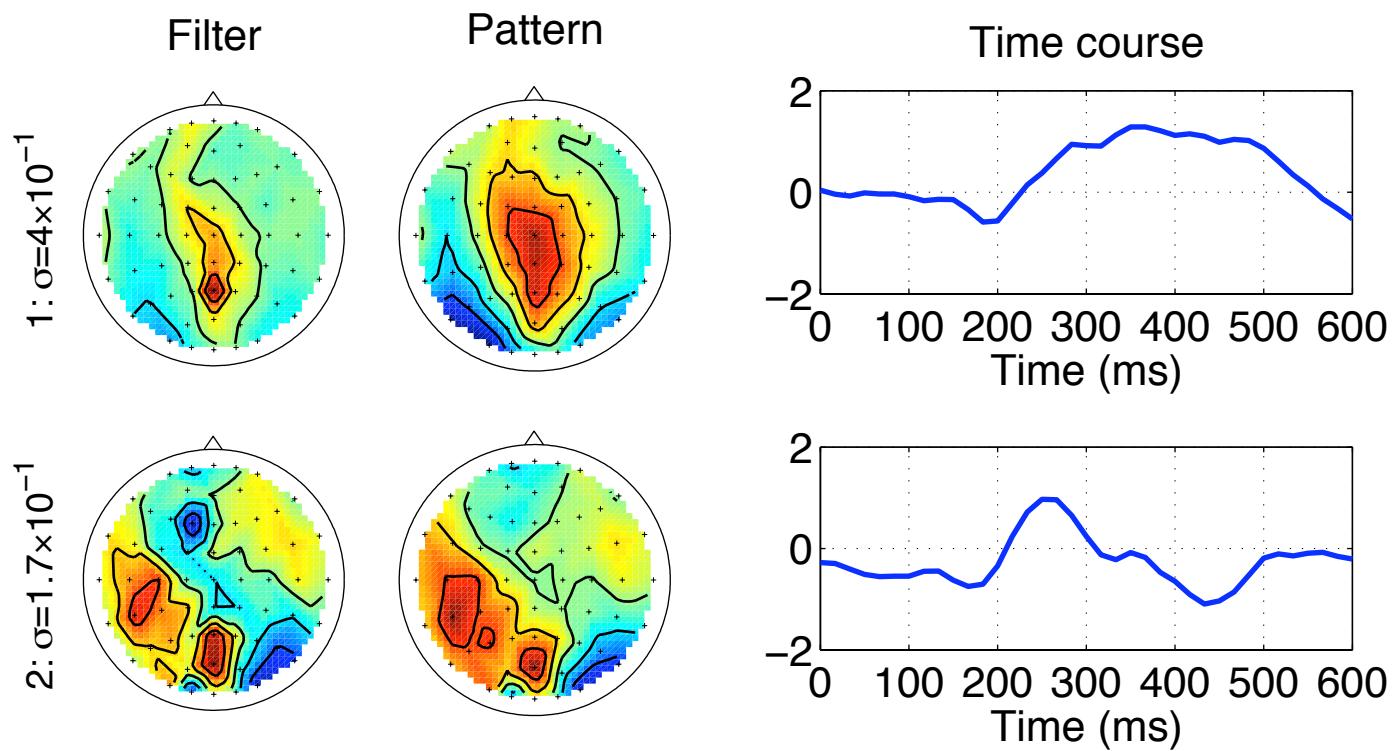
$$f_{\theta}(X) = \langle W, X \rangle$$

$$W \in \mathbb{R}^{64 \times 37}$$

判別性能 (36クラス)

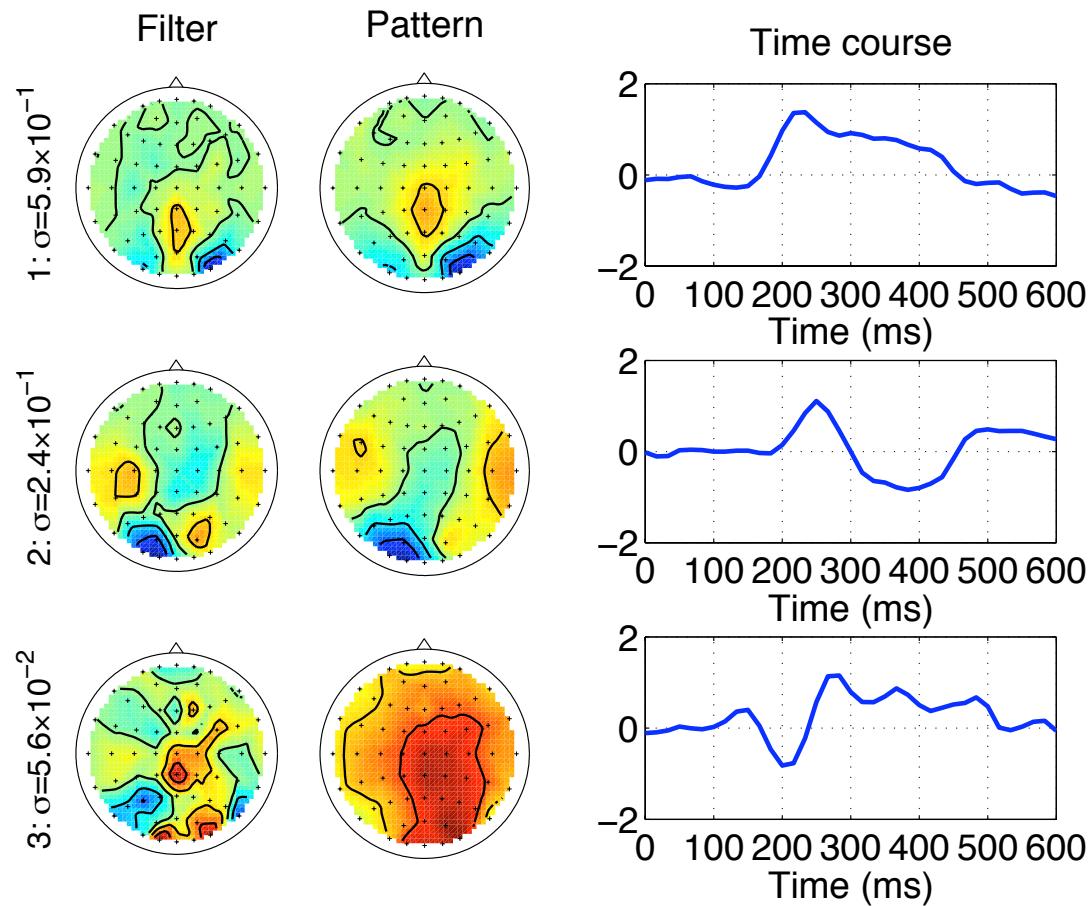


時空間フィルタ (Subject A)



- 300ms以前の早い段階で頭の後ろの方で判別できる
- 300ms以降長く続く頭頂部で判別性がある

時空間フィルタ (Subject B)



- 最初の2つの成分が早い後頭部の成分
- 3つ目の成分が頭頂部の遅い成分に対応

ここまでまとめ

- 損失(訓練誤差)項と正則化項の組み合わせがいろいろ変えられるのが機械学習の(ひとつの)醍醐味
- 今回
 - 損失項: Farwell & Donchin システムの解読方法を素直に表現した形 (2x6クラス分類)
 - 正則化項: 低ランク分解を促すSchatten 1-ノルム
- それなりにもっともらしい時空間フィルタが得られた。

テンソルの低ランク分解

テンソルのランクとは
凸最適化によるテンソル分解
性能の理論解析

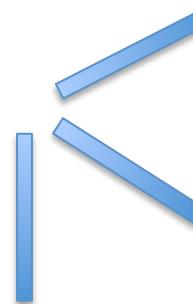
テンソルのランク

定義 $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ (K階テンソル)

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

ただし \mathcal{A}_r はランク1

$$\mathcal{A}_r =$$



(ベクトル
の外積で
書ける)

となる最小の Rを X のランクという。

- ある R で分解が可能かチェックするのはNP完全
- ランクの決定はNP困難
- このような分解を CANDECOMP / PARAFAC 分解 (CP分解)とよぶ

テンソルのランクの不思議な性質1

- 分解の一意性
 - 行列の分解 $X=AB^T$ は一意性がない
 - テンソルのCP分解

$$\mathcal{X} = \sum_{r=1}^R a_r \circ b_r \circ c_r = [[A, B, C]]$$

は一意性を持つ場合がある

(定数倍の自明な自由度をのぞく)

- 一意性の十分条件 (Kruskal 77)

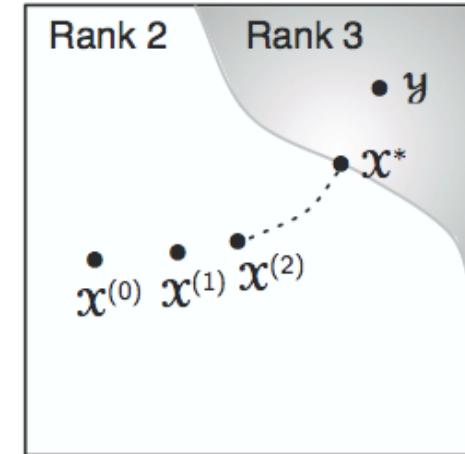
$$k_A + k_B + k_C \geq 2R + 2$$

k_A は行列Aの k-rank (k本の列ベクトルが線形独立となる最大のk)

(以降、説明を簡単にするために3階のテンソルを考える)

テンソルのランクの不思議な性質2

- 閉じていない



Kolda & Bader 2009

\mathcal{X} はランク3

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

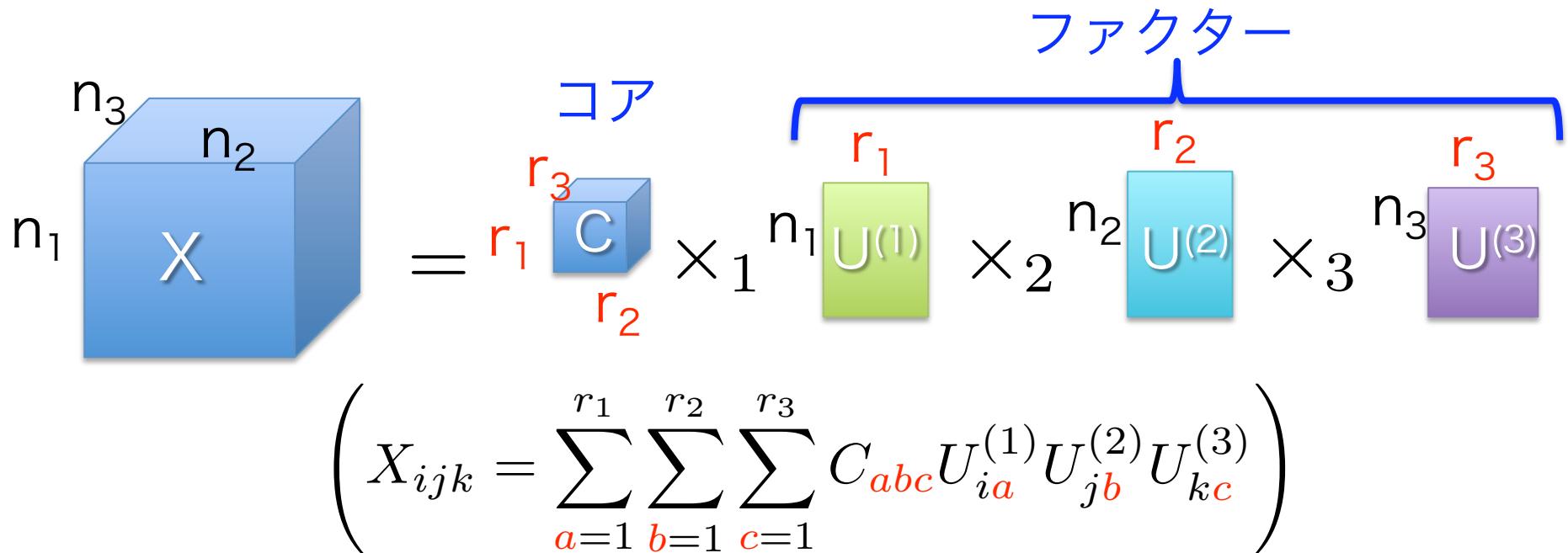
\mathcal{Y} はランク2

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

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

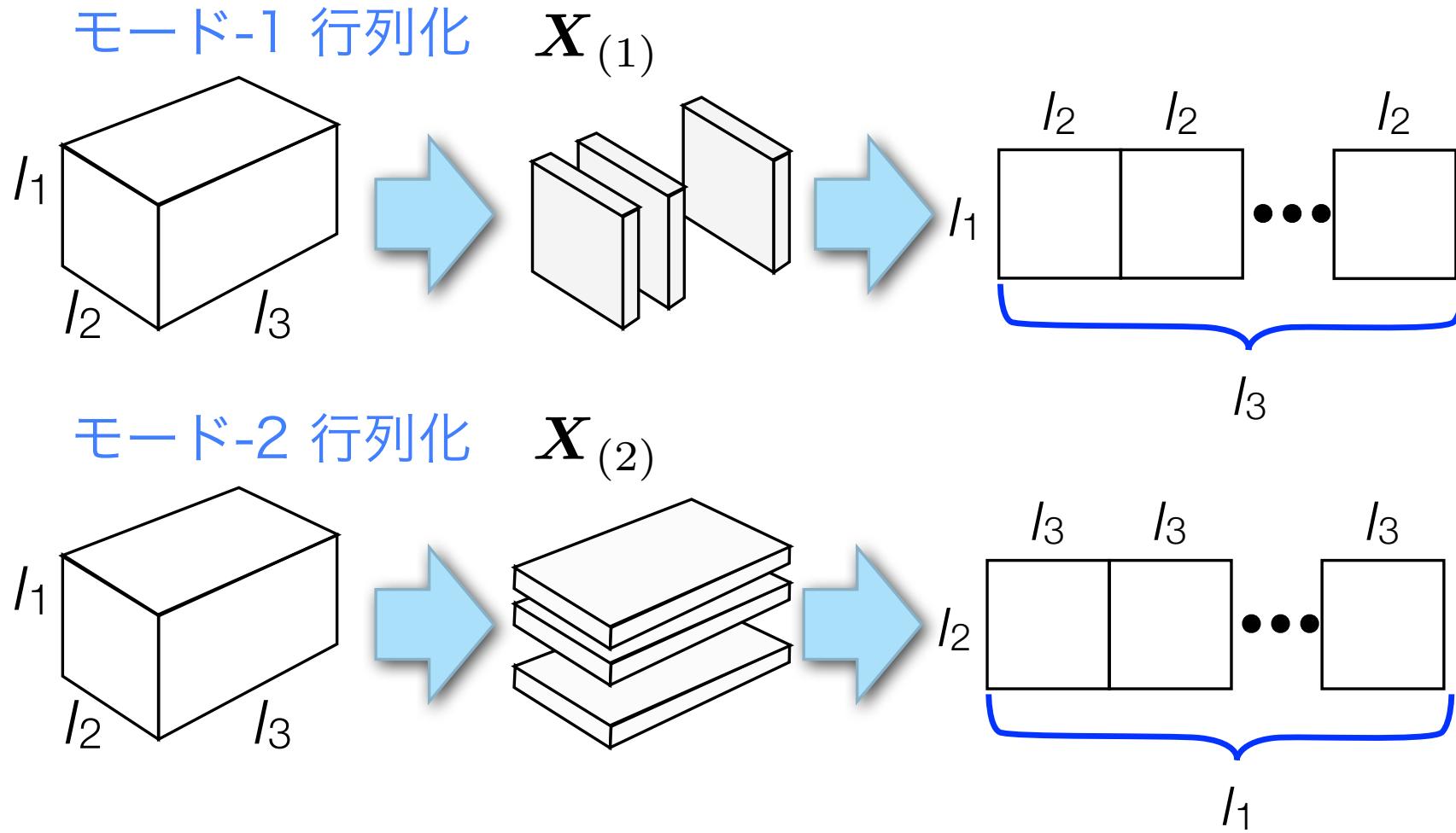
Tucker 分解の意味のランク

- Tucker分解 [Tucker 66]



- CP分解はコアテンソルが対角 ($r_1=r_2=r_3$) の場合
- Tucker 分解のランクは軸(モード)ごとに異なる

テンソルのモード-k 展開 (行列化)



数学的には要素の順番のパーミュテーション

モード-k展開とモード-kランク

$$\mathcal{X} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$$

モード-1 行列化

$$\mathbf{X}_{(1)} = \mathbf{U}_1 \mathbf{C}_{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2)^\top$$

$$\text{rank}(\mathbf{X}_{(1)}) = r_1$$

モード-2 行列化

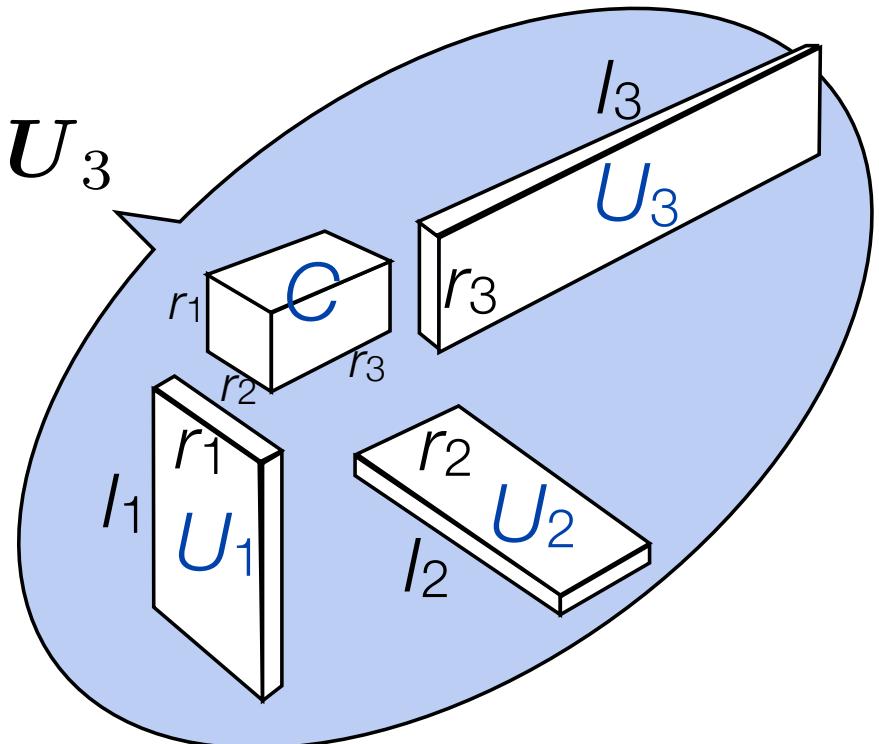
$$\mathbf{X}_{(2)} = \mathbf{U}_2 \mathbf{C}_{(2)} (\mathbf{U}_1 \otimes \mathbf{U}_3)^\top$$

$$\text{rank}(\mathbf{X}_{(2)}) = r_2$$

モード-3 行列化

$$\mathbf{X}_{(3)} = \mathbf{U}_3 \mathbf{C}_{(3)} (\mathbf{U}_2 \otimes \mathbf{U}_1)^\top$$

$$\text{rank}(\mathbf{X}_{(3)}) = r_3$$



Tuckerの意味でのランクは $\mathbf{X}_{(k)}$ の
行列としてのランクに等しい

CP分解 / Tucker分解 ランクの比較

	行列の 特異値分解	CP分解	Tucker分解
ランクの決定	多項式	NP困難	多項式 (各モード展開してSVD)
分解の計算	多項式	NP困難	多項式
分解の一意性	なし	あり	なし
コア	対角	対角	$r_1 \times r_2 \times r_3$ テンソル

Tucker分解の方が特異値分解の自然な拡張になっているかも

テンソルの低ランク化を促す正則化項

テンソルに対する overlapped Schatten 1-ノルム

$$\|\mathcal{X}\|_{S_1} = \sum_{k=1}^K \gamma_k \|X_{(k)}\|_{S_1}$$



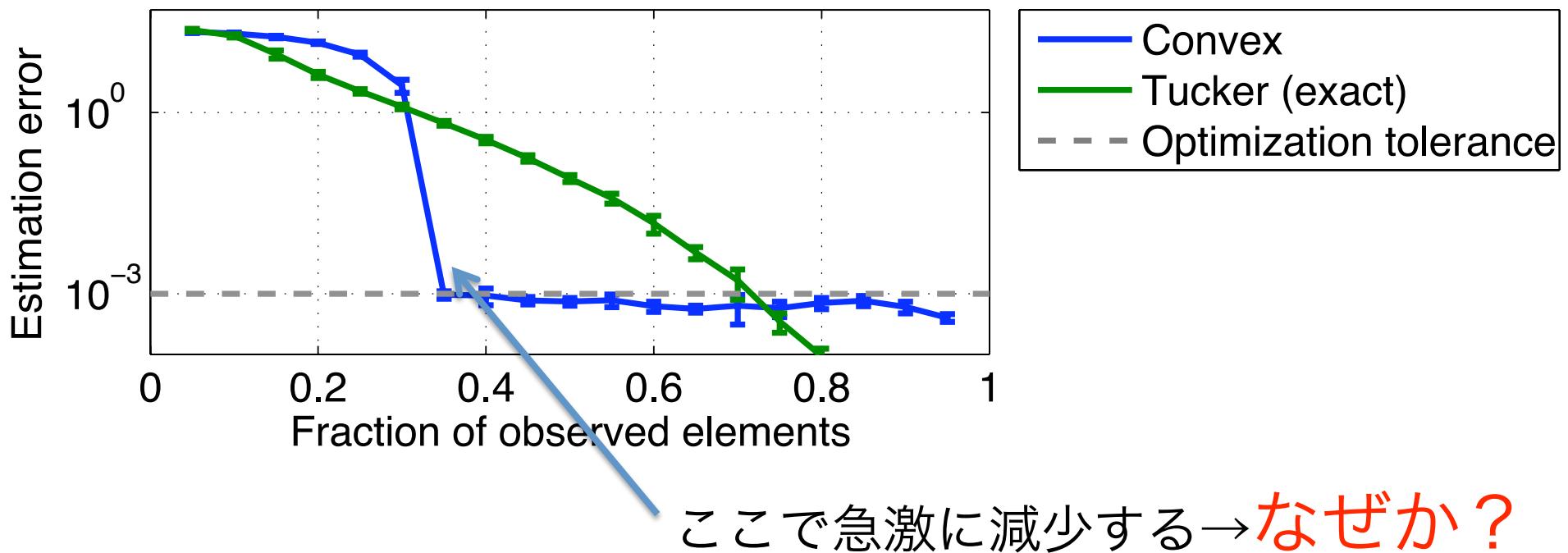
モード-k 行列化の
Schatten 1-ノルム

- ノルムの公理を満たす
- 各モードを γ_k の強さで低ランクになるよう正則化
- これを使えば低ランクテンソルの推定が凸最適化ができる

テンソル補完への応用

- 最適化問題

$$\begin{aligned} & \underset{\mathcal{X}}{\text{minimize}} \quad \| \mathcal{X} \|_{S_1}, \\ & \text{subject to} \quad \mathcal{X}_{ijk} = \mathcal{Y}_{ijk} \quad ((i, j, k) \in \Omega) \end{aligned}$$



圧縮センシング業界における相転移

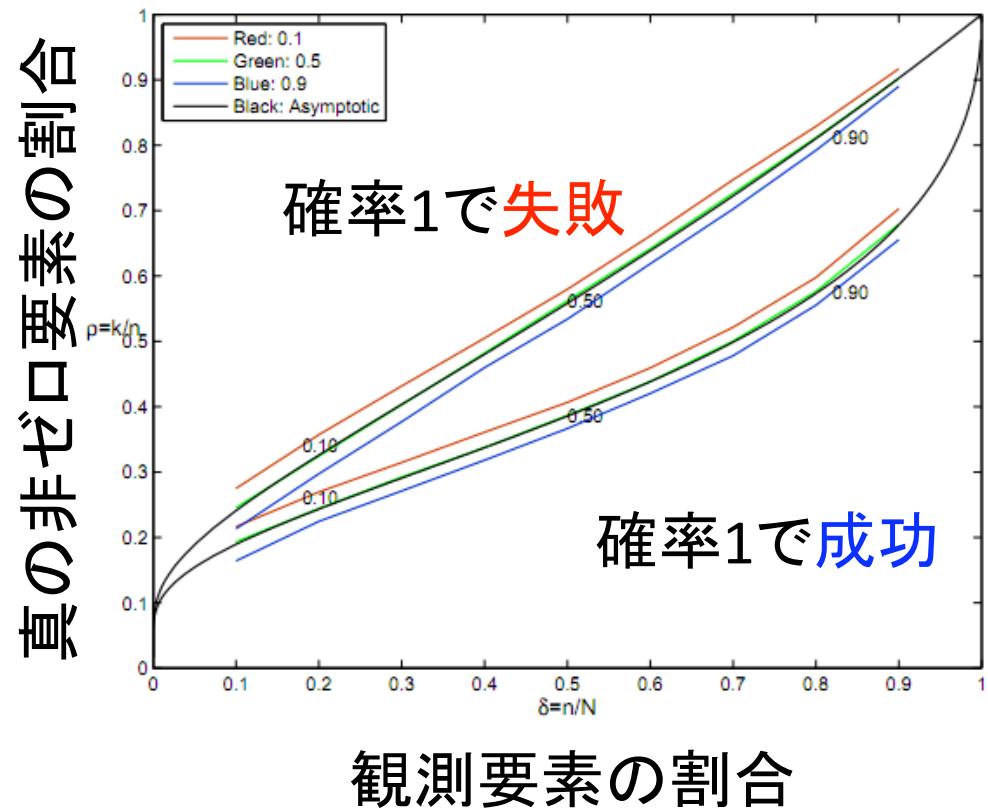
- Donoho-Tanner Phase Transition

$$\underset{x}{\text{minimize}} \quad \|x\|_{\ell_1}$$

$$\text{subject to} \quad Ax = y$$

A: $n \times N$ 行列

$(n \ll N)$



Donoho & Tanner "Precise Undersampling Theorem"

行列補完における相転移

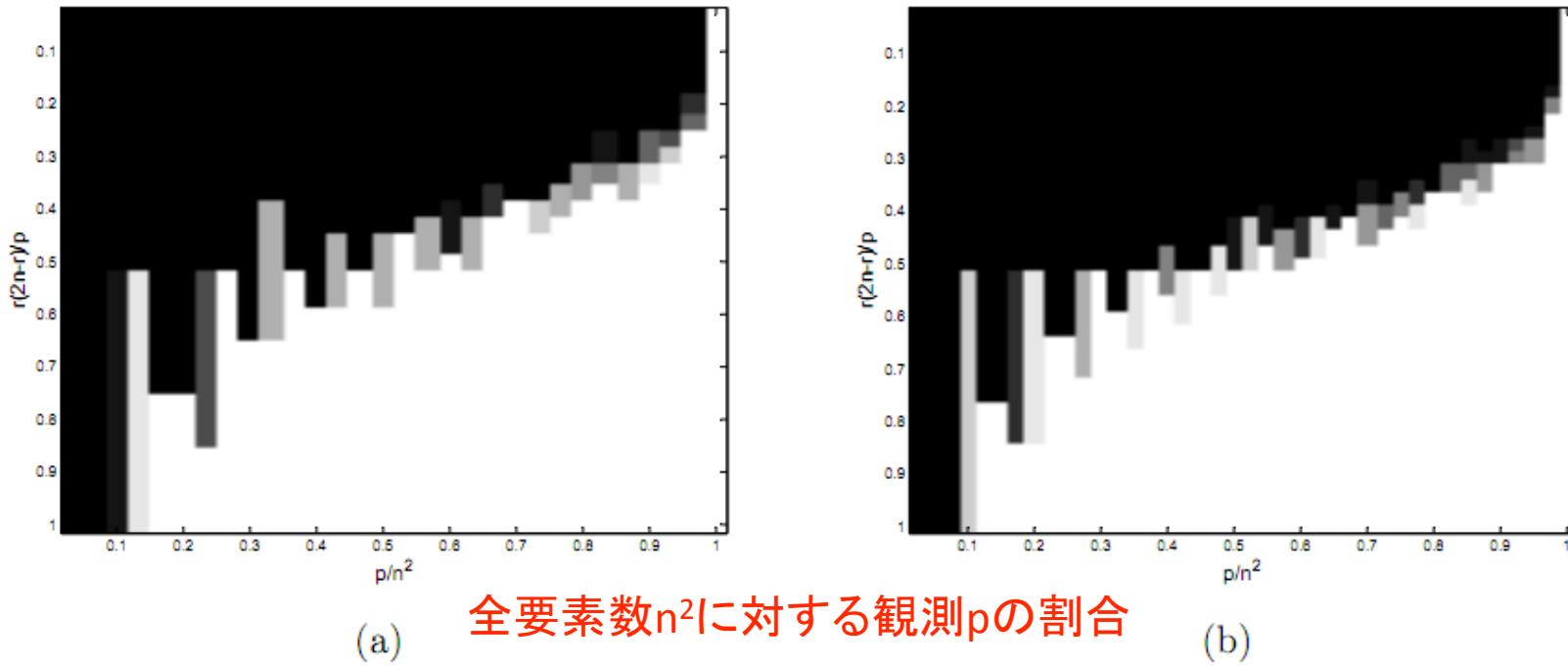


Figure 4: For each (n, p, r) triple, we repeated the following procedure ten times. A matrix of rank r was generated by choosing two random $n \times r$ factors Y_L and Y_R with i.i.d. random entries and set $Y_0 = Y_L Y_R'$. We select a matrix \mathbf{A} from the Gaussian ensemble with p rows and n^2 columns. Then we solve the nuclear norm minimization subject to $\mathbf{A} \operatorname{vec}(X) = \mathbf{A} \operatorname{vec}(Y_0)$. We declare Y_0 to be recovered if $\|X - Y_0\|_F / \|Y_0\|_F < 10^{-3}$. The results are shown for (a) $n = 30$ and (b) $n = 40$. The color of each cell reflects the empirical recovery rate (scaled between 0 and 1). White denotes perfect recovery in all experiments, and black denotes failure for all experiments.

解析 : 問題設定

観測モデル

\mathcal{W}^* : 真のテンソル ランク (r_1, \dots, r_K)

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

ガウス雑音

最適化問題

$$\hat{\mathcal{W}} = \underset{\mathcal{W} \in \mathbb{R}^{n_1 \times \dots \times n_K}}{\operatorname{argmin}}$$

データ尤度

正則化項

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

正則化定数

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

観測作用素 $\mathfrak{X} : \mathbb{R}^N \rightarrow \mathbb{R}^M$

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

仮定：制約強凸性

(cf. Negahban & Wainwright 11)

- 正の定数 $\kappa(X)$ が存在して $\Delta \in C$ なるすべてのテンソル Δ に関して

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

が成り立つ。(集合 C はあとで定義)

注意

- $\kappa(X)$ は X の最小特異値に対応
- $M > N$ (パラメータ数) なら普通の仮定
- 集合 C をいかに狭く取るかがポイント

定理1

- 最適化問題の解 $\hat{\mathcal{W}}$
- ただしノイズ-デザイン相関 $\mathfrak{X}^*(\epsilon) = \sum_{i=1}^M \epsilon_i \mathcal{X}_i$

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

と仮定

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

- 制約強凸性が成り立つと仮定
- この時

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

ランクの2乗根の和が問題の難しさを決めている

補題: Hölderっぽい不等式

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

ただし、 $\|\mathcal{X}\|_{\text{mean}} := \frac{1}{K} \sum_{k=1}^K \|\mathbf{X}_{(k)}\|_{S_\infty}$

モード-k 行列化の
スペクトルノルム

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

注) $\|\cdot\|_{S_1}$ と $\|\cdot\|_{\text{mean}}$ は双対ノルムにはなっていない(もっとタ
イトな不等式が存在する)

2つの特殊な場合

- もうちょっと精密に考える必要
 - 制限強凸性の定数 $\kappa(X)$ は？
 - 正則化定数 λ_M はどう選ぶ？
- ノイズあり行列分解(全部の要素が見えている)
 - 制限強凸性: ほぼ自明
 - 正則化定数 λ_M をノイズ-デザイン相関 $\|\mathcal{X}^*(\epsilon)\|_{\text{mean}}$ の強さに応じて選ぶ
- ランダムガウスデザイン
 - 制限強凸性: あまり自明ではない(ただし Negahban & Wainwright の結果が使える)
 - 正則化定数 λ_M をノイズ-デザイン相関 $\|\mathcal{X}^*(\epsilon)\|_{\text{mean}}$ の強さに応じて選ぶ

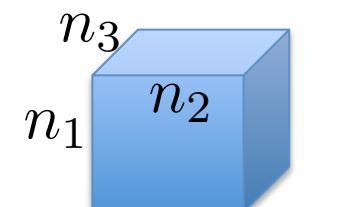
ノイズあり行列分解の場合

- 全部の要素が観測可能 ($M=N$)

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

- 正則化定数 $\lambda_M \geq 2\|\mathfrak{X}^*(\epsilon)\|_{\text{mean}}/M$

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


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

(ランダム行列の理論から)

しかも $\|\mathfrak{X}^*(\epsilon)\|_{\text{mean}}$ は高い確率で平均の周りに集中する

定理2

- 全部の要素が見えている場合 ($M=N$)、

正則化定数 $\lambda_M \geq \frac{2\sigma}{K} \sum_{k=1}^K \left(\sqrt{n_k} + \sqrt{N/n_k} \right) / N$
のように取ると

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

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

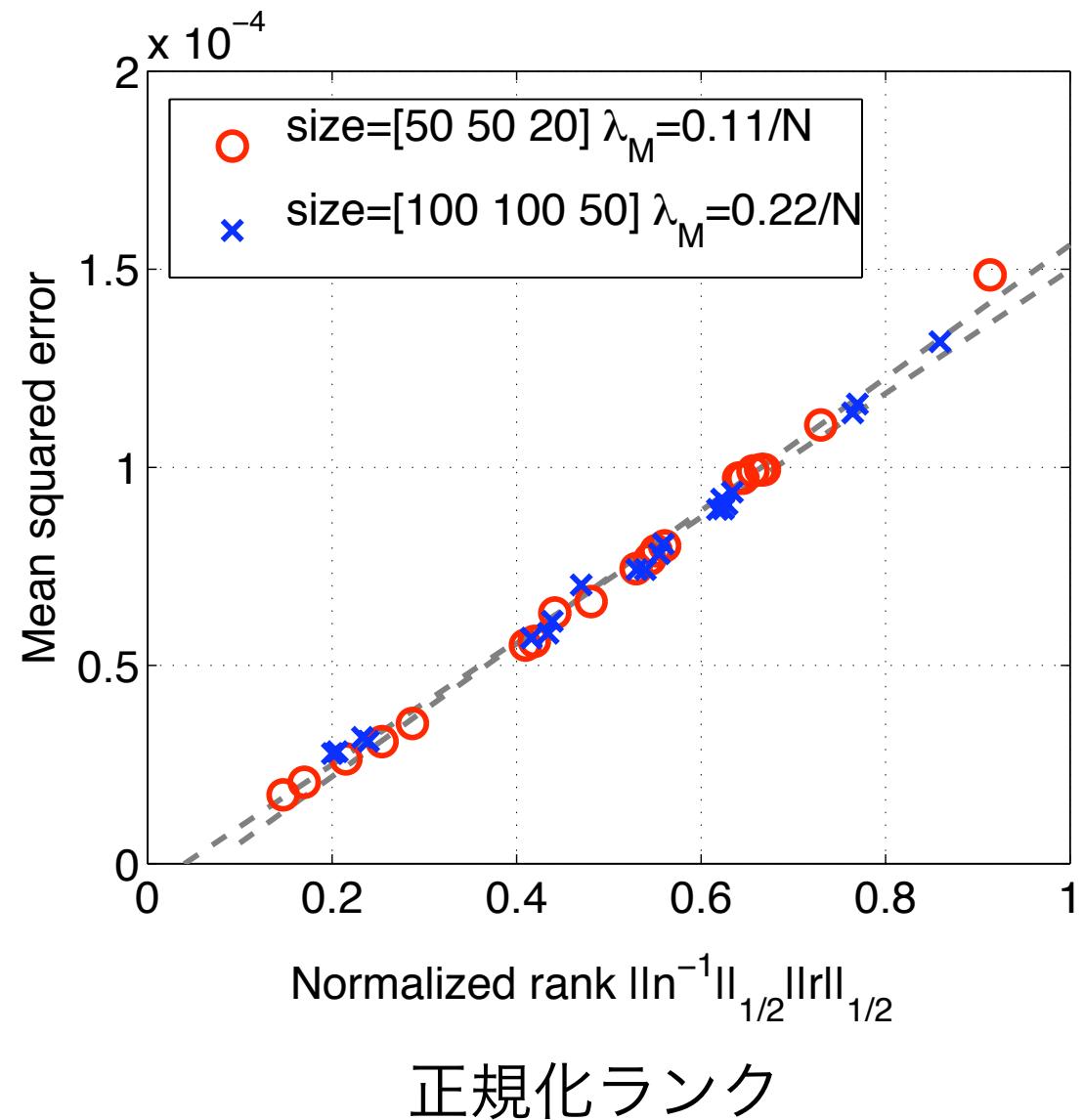
また $\|\mathbf{n}^{-1}\|_{1/2} \|\mathbf{r}\|_{1/2}$ を正規化ランクと呼ぶ

実験結果: ノイズありテンソル分解 (ノイズ小 $\sigma=0.01$)

平均2乗誤差

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

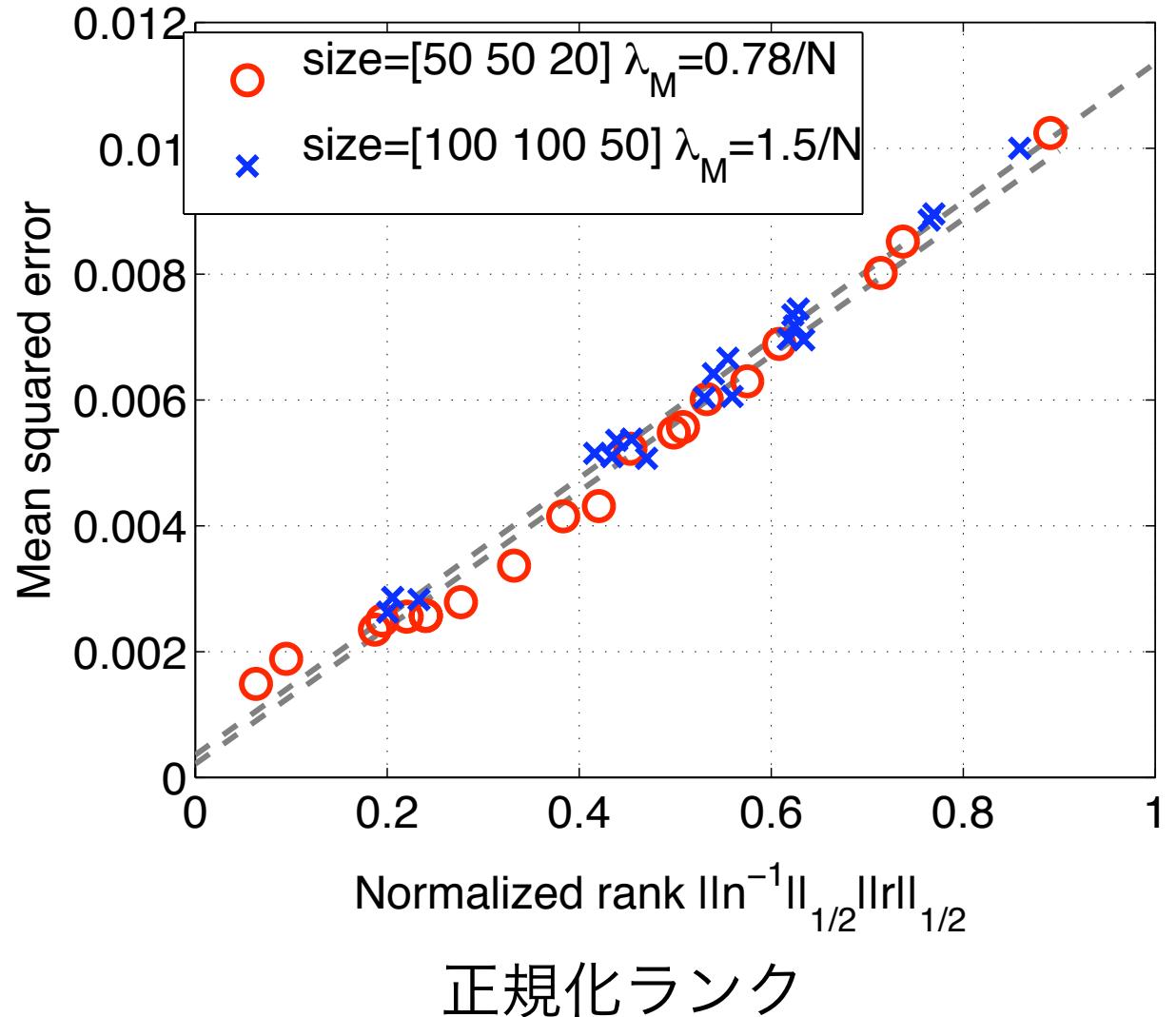
- 正則化定数の取り方は大きさのみに依存し、ランクに依らない
- 平均2乗誤差は正規化ランクに比例



実験結果: ノイズありテンソル分解 (ノイズ大 $\sigma=0.1$)

- 平均2乗誤差

$$\frac{\|\hat{\mathcal{W}} - \mathcal{W}^*\|_F^2}{N}$$
- 正則化定数の取り方は大きさのみに依存し, ランクに依らない
- 平均2乗誤差は正規化ランクに比例

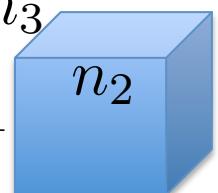


ランダムガウスデザイン (X_i が独立同一なガウス分布から出ている) 場合

- 正則化定数 $\lambda_M \geq 2\|\mathcal{X}^*(\epsilon)\|_{\text{mean}}/M$

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

- 制約強凸性: ちょっと複雑



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

$$\frac{M}{N} \geq c\|\mathbf{n}^{-1}\|_{1/2}\|\mathbf{r}\|_{1/2} \quad \text{ならOK } (\kappa=1/64)$$

↑ ↑

適当な定数 正規化ランク

(M : サンプル数, N : テンソルの要素数)

定理3

ランダムガウスデザインの場合、

1. 観測の要素数に対する割合

$$\frac{M}{N} \geq c \|\mathbf{n}^{-1}\|_{1/2} \|\mathbf{r}\|_{1/2} \quad (\text{制約強凸性の条件})$$

2. 正則化定数

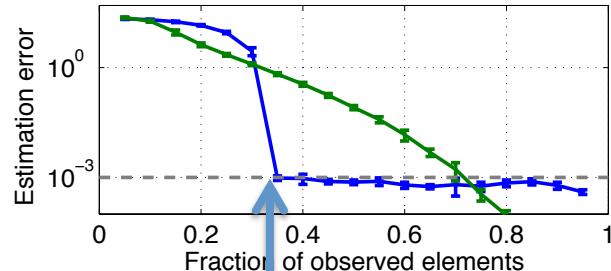
$$\lambda_M \geq \frac{2\sigma}{K} \sum_{k=1}^K \left(\sqrt{n_k} + \sqrt{N/n_k} \right) / \sqrt{M}$$

のもとで

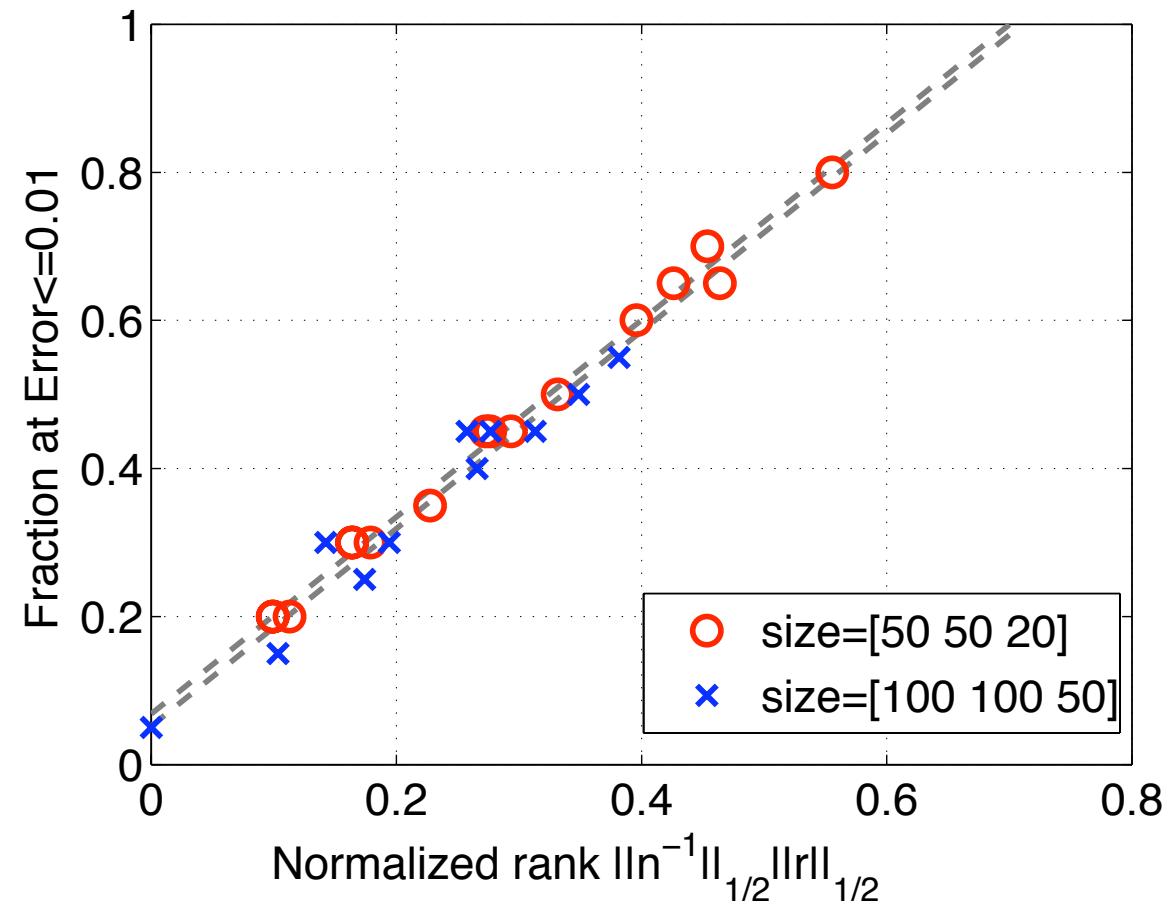
$$\frac{\|\hat{\mathcal{W}} - \mathcal{W}^*\|_F^2}{N} \leq O_p \left(\frac{\sigma^2 \|\mathbf{n}^{-1}\|_{1/2} \|\mathbf{r}\|_{1/2}}{M} \right)$$

実験

テンソル穴埋め (ノイズなし) $\Rightarrow \lambda \rightarrow 0$



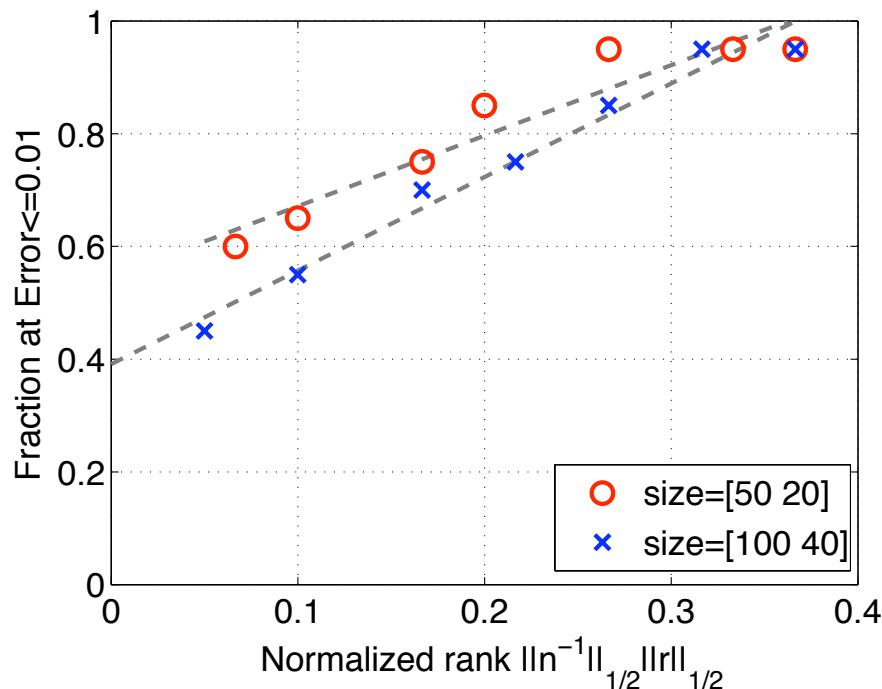
ここでの全
要素に対する
観測要素
の割合



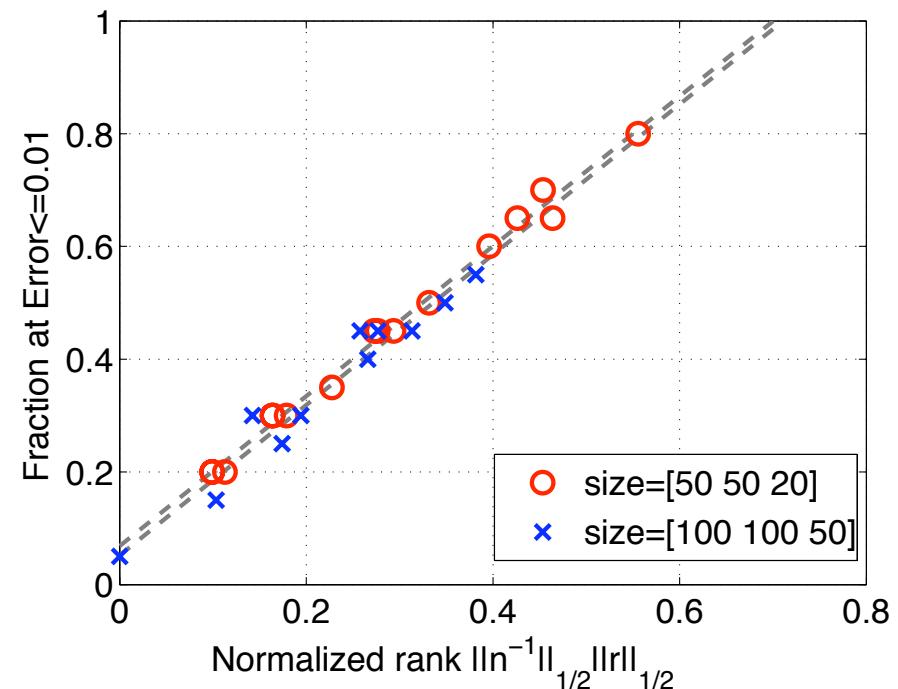
正規化ランク

行列の場合とテンソルの場合

行列穴埋めの場合



テンソル穴埋めの場合



どちらもノイズなし

テンソル分解編のまとめ

- テンソルは数学的対象として非自明な点が多い
- Tucker分解はSVDと関係があるので扱いやすい
- Tucker分解を凸最適化で解く手法を提案した
- 解析結果と実験結果はそこそこ一致している
 - Normalized rank
- 課題：
 - 行列の場合とテンソルの場合の違いはなぜ？
 - 制約強凸性は満たされているのか？
 - サンプル数だけで復元可能性が決まるのか？
 - 一部のモードのみ低ランクな場合は？
 - 応用…