

IBIS2016 企画セッション「統計理論」
低ランクテンソルの学習理論と計算理論

^{1,2,3} 鈴木大慈

共同研究者：¹ 金川平志郎，⁴ 清水伸幸，⁴ 小林隼人，⁴ 田頭幸浩

¹ 東京工業大学数理・計算科学系

² JST さきがけ

³ 理研 AIPC

⁴ Yahoo! Japan

2016 年 11 月 16 日
IBIS2016@京都大学吉田キャンパス

Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

高次元パラメータ推定



ベクトル
スパース性

手法

- Lasso
- Sure Screening

応用

- 特徴選択
- 遺伝子データ
解析

高次元パラメータ推定



ベクトル
スパース性

手法

- Lasso
- Sure Screening

応用

- 特徴選択
- 遺伝子データ解析



行列
低ランク性

手法

- PCA
- トレースノルム正則化

応用

- 次元削減
- 推薦システム
- 三層ニューラルネット

高次元パラメータ推定



ベクトル
スパース性

手法

- Lasso
- Sure Screening

応用

- 特徴選択
- 遺伝子データ解析



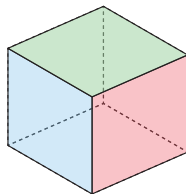
行列
低ランク性

手法

- PCA
- トレースノルム正則化

応用

- 次元削減
- 推薦システム
- 三層ニューラルネット



テンソル
低ランク性

本発表

高次の関係性を記述

概要

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

低ランク行列推定問題

- スパースベクトルの推定 :

スパースな係数の学習 (基底は固定)

- 低ランク行列の推定 :

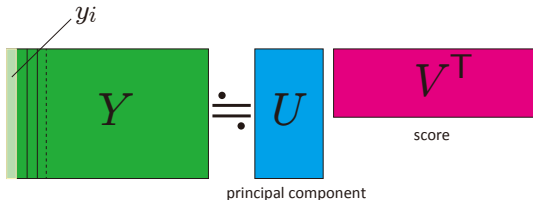
基底の学習 + 係数の学習

係数がスパースになるような基底を学習

例 : 主成分分析 (PCA)

$Y = [\mathbf{y}_1, \dots, \mathbf{y}_n] \in \mathbb{R}^{p \times n}$: 観測値

$$\min_{U \in \mathbb{R}^{p \times k}, V \in \mathbb{R}^{n \times k}} \sum_{i=1}^n \sum_{j=1}^p (Y_{j,i} - U_{i,:} V_{i,:}^{\top})^2 = \min \|Y - UV^{\top}\|_F^2$$



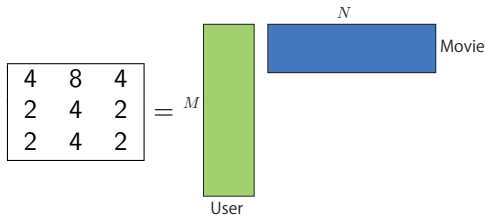
推薦システム

	Movie A	Movie B	Movie C	...	Movie X
User 1	4	8	*	...	2
User 2	2	*	2	...	*
User 3	2	4	*	...	*

推薦システム

ランク 1 (低ランク)

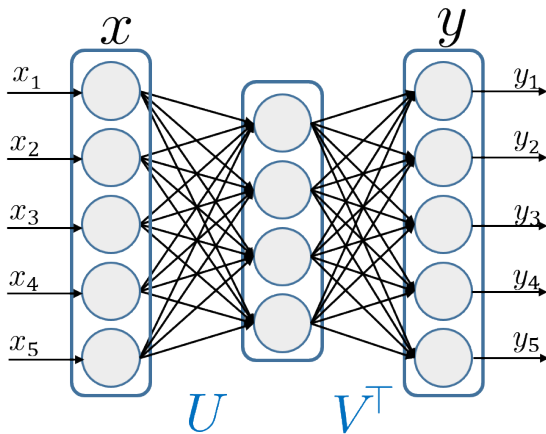
	Movie A	Movie B	Movie C	...	Movie X
User 1	4	8	4	...	2
User 2	2	4	2	...	1
User 3	2	4	2	...	1



$$\min_{U,V} \sum_{(i,j) \in \mathcal{I}} (Y_{i,j} - U_{i,:} V_{j,:}^T)^2$$

(e.g., Srebro et al. (2005); Candès and Tao (2009); Candès and Recht (2009), NetFlix (Bennett & Lanning, 2007))

3層ニューラルネットワーク



$$f(x) = V^\top Ux$$

$$f(x) = V^\top h(Ux)$$

低ランク行列推定の推定精度

真の行列 A^* は $M \times N$ 行列でランク d とする ($d \ll M, N$: 低ランク).

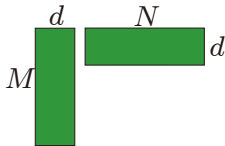
- 最小二乗法:

$$\|\hat{A} - A^*\|_2^2 = O_p\left(\frac{MN}{n}\right)$$

低ランク性を利用できていない.

- トレースノルム正則化:

$$\|\hat{A} - A^*\|_2^2 = O_p\left(\frac{d(M+N)\log(MN)}{n}\right)$$



ただし, $\|A\|_2 := \sqrt{\frac{1}{MN}} \|A\|_F$.

(Negahban et al., 2012; Negahban & Wainwright, 2012; Rohde & Tsybakov, 2011; Koltchinskii, 2012)

$$(\text{最小二乗法}) \quad \frac{MN}{n} \gg \frac{d(M+N)\log(MN)}{n} \quad (\text{正則化学習})$$

概要

① 低ランク行列推定

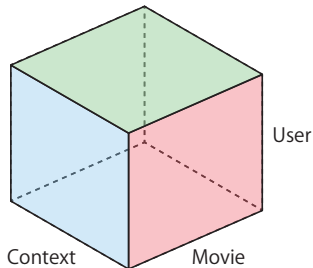
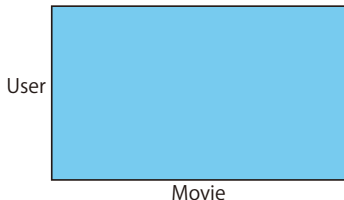
② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

行列からテンソルへ



行列

$$X_{ij} = \sum_{r=1}^d u_r^{(1)} u_r^{(2)}$$

→

テンソル (多次元配列)

$$X_{ijk} = \sum_{r=1}^d u_r^{(1)} u_r^{(2)} u_r^{(3)}$$

- 複数項目間の高次の関係性 (K 次)
- いわゆる座標返還に関して不変な量という意味は薄れている。しかし、多重線形写像という観点は重要。
- さらに高次元 $M_1 \times M_2 \times \cdots \times M_K$
→ 低ランク性 を利用

テンソルのランクとは？

Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

テンソルの分解・ランク

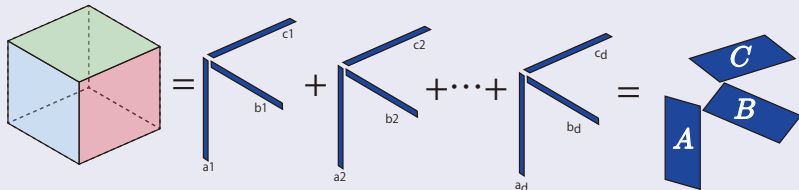
- CP-分解

(Hitchcock, 1927; Hitchcock, 1927; Carroll & Chang, 1970; Harshman, 1970)

- Tucker-分解 (Tucker, 1966)

- テンソルトレイン分解 (Oseledets, 2011)

Tensor rank: CP-rank



CP-分解

Canonical Polyadic 分解 (Hitchcock, 1927; Hitchcock, 1927)
CANDECOMP/PARAFAC (Carroll & Chang, 1970; Harshman, 1970)

$\mathcal{X}_{ijk} = \sum_{r=1}^d a_{ir} b_{jr} c_{kr}$ ($\mathcal{X} = [[A, B, C]]$). $\mathcal{X} = \sum_{r=1}^d a_r \otimes b_r \otimes c_r$ とも書く.

CP-分解で CP-ランクが決まる.

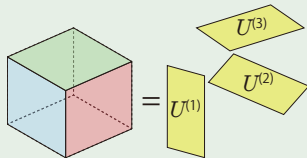
- CP-分解の計算は **NP-困難** (Hillar & Lim, 2013).
- 直交分解できるとは限らない.
- ランク 2 のテンソルでランク 3 のテンソルをいくらでも良く近似できる例が作れてしまう.

CP-分解の一意性

CP-分解が一意であるための十分条件

$$\mathcal{X} = [[U^{(1)}, U^{(2)}, \dots, U^{(K)}]]$$

各 $U^{(k)}$ が列フルランクなら, CP-分解は
スケーリングと列の入れ替えの自由度を除いて一意.
(行列のような回転の自由度はない)

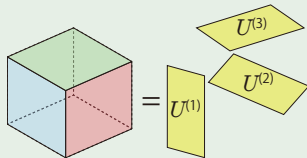


CP-分解の一意性

CP-分解が一意であるための十分条件

$$\mathcal{X} = [[U^{(1)}, U^{(2)}, \dots, U^{(K)}]]$$

各 $U^{(k)}$ が列フルランクなら，CP-分解は
スケーリングと列の入れ替えの自由度を除いて一意。
(行列のような回転の自由度はない)



$\text{k-rank}(A)$: Kruskal ランク，行列 A の劣を任意に r 個とってきたときに線形独立である r の最小値。

より詳細な条件

$\mathcal{X}$ の CP-ランクが d の時，以下が満たされていれば CP-分解は一意：

$$2d + (K - 1) \leq \sum_{k=1}^K \text{k-rank}(U^{(k)}).$$

CP-分解の計算

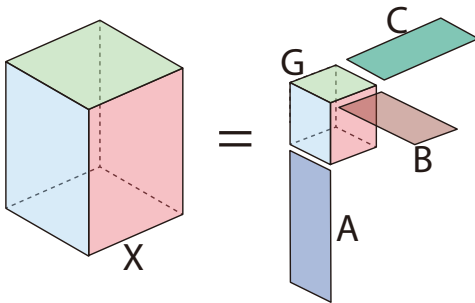
一般的には CP-分解は NP 困難だが..

$$\mathcal{X} = [[U^{(1)}, U^{(2)}, \dots, U^{(K)}]]$$

で, 各 $U^{(k)}$ が列フルランク とする. (このとき, CP-分解は一意)

- すると, 多項式時間で CP-分解は計算できる.
- 有限回の固有値分解や行列の同時対角化などで解ける.
- 確率的アルゴリズム
 - Simultaneous Diagonalization (De Lathauwer, 2006)
 - Simultaneous Generalized Schur Decomposition (De Lathauwer et al., 2004)
 - Generalized Eigenvalue Decomposition (Leurgans et al., 1993)

Tucker-ランク



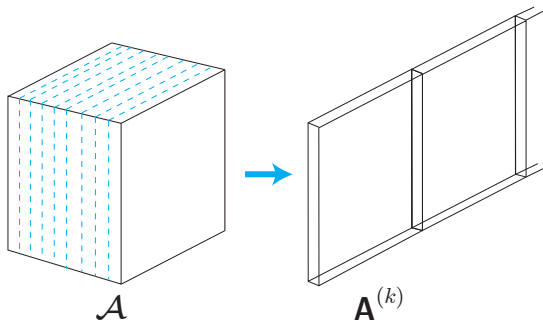
Tucker-分解 (Tucker, 1966)

$$\mathcal{X}_{ijk} = \sum_{l=1}^{r_1} \sum_{m=1}^{r_2} \sum_{n=1}^{r_3} g_{lmn} a_{il} b_{jm} c_{kn}. \quad \mathcal{X} =: [[G; A, B, C]].$$

- G はコアテンソルと呼ばれる.
- G が対角なら CP 分解に帰着.
- Tucker-ランク $= (r_1, r_2, r_3)$. ただし r_k は G の各辺の大きさ. これは, モード k 展開 (次ページ) のランクで与えられる.

行列化

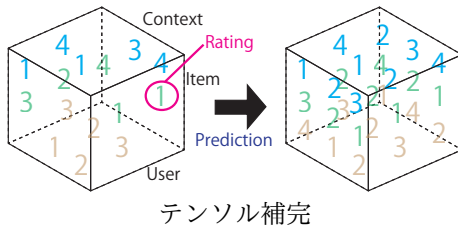
モード k 展開: $\mathbf{A}^{(k)} \in \mathbb{R}^{M_k \times N/M_k}$, $(N = \prod_{k=1}^K M_k)$.



$r_k = \text{rank}(\mathbf{A}^{(k)})$ は Tucker-ランクを与える.

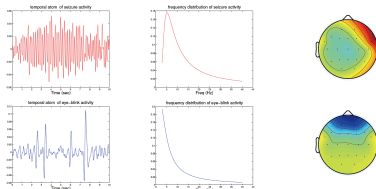
応用

- 推薦システム (ユーザー × アイテム × 店)
- マルチタスク学習
- 信号処理 (空間 (2D) × 時間)
- 自然言語処理 (単語のベクトル表現)



その他の応用

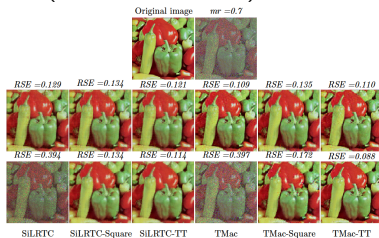
- 脳波解析 (De Vos et al., 2007)



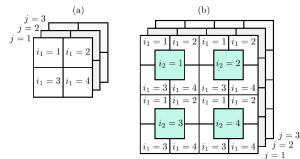
時間 × 周波数 × 空間

EEG monitoring: Epileptic seizure onset localization

- 雑音除去 (Phien et al., 2016)



種々の方法による雑音除去



画像のテンソルトレイン化

Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

テンソル推定の回帰としての定式化

テンソルへの回帰:

$$\underbrace{Y_i}_{\text{応答}} = \langle \underbrace{X_i}_{\substack{\text{入力} \\ \text{(テンソル)}}}, \underbrace{A^*}_{\substack{\text{真} \\ \text{(テンソル)}}} \rangle + W_i$$

$A^*, X_i \in \mathbb{R}^{M_1 \times \dots \times M_K}$: テンソル.

$$\langle X_i, A^* \rangle := \sum_{j_1, \dots, j_K} X_{i, (j_1, \dots, j_K)} A_{j_1, \dots, j_K}^*$$

テンソル間の内積

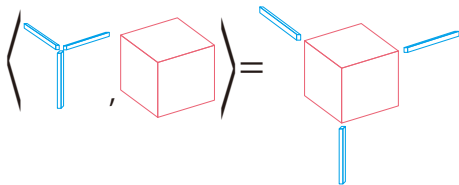
仮定: A^* は「低ランク」.

E.g., $X_i = e_{j_1} \otimes e_{j_2} \otimes \dots \otimes e_{j_K}$ なら, テンソル補完の問題に帰着.

$$\langle X_i, \mathcal{A}^* \rangle := \sum_{j_1, \dots, j_K} X_{i, (j_1, \dots, j_K)} \mathcal{A}_{j_1, \dots, j_K}^*$$

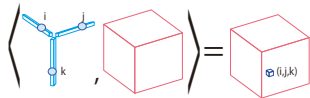
- 特殊ケース: for $u_k \in \mathbb{R}^{M_k}$,

$$\langle u_1 \otimes \dots \otimes u_K, \mathcal{A}^* \rangle = \sum_{j_1=1}^{M_1} \dots \sum_{j_K=1}^{M_K} u_{1,j_1} u_{2,j_2} \dots u_{K,j_K} \mathcal{A}_{j_1, \dots, j_K}^*.$$



- 特に,

$$\langle e_{j_1} \otimes e_{j_2} \otimes \dots \otimes e_{j_K}, \mathcal{A}^* \rangle = \mathcal{A}_{j_1, \dots, j_K}^*.$$



→ テンソル補完

低ランクテンソルの求め方

- 凸正則化法
- 低ランクテンソルの集合上での訓練誤差最小化
 - 交互最適化法
 - Riemannian 最適化
- ベイズ推定法

テンソル推定における統計理論構築の難しさ

- 高次元性
 - 特異性 (パラメータ表現の一意性のなさ)
 - 非凸性
- 単純な統計的漸近理論が使えない.
これらを考慮したバウンドを構築する必要がある.

Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

凸正則化による推定

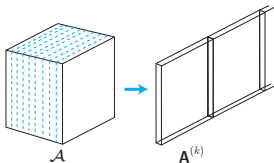
凸正則化アプローチ:

$$\min_{\mathcal{A} \in \mathbb{R}^{M_1 \times M_2 \times \dots \times M_K}} \frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, \mathcal{A} \rangle)^2 + \text{pen}(\mathcal{A}).$$

テンソルを行列に変形してトレースノルム正則化

- Overlapped Schatten 1-norm (Tomioka et al., 2011)
- Latent Schatten 1-norm (Tomioka & Suzuki, 2013; Zheng & Tomioka, 2015)
- Square deal (Mu et al., 2014)
- Scaled latent regularization (Wimalawarne et al., 2014)

$$\frac{1}{N} \|\hat{\mathcal{A}} - \mathcal{A}^*\|_2^2 \leq C \frac{d \sqrt{\prod_k M_k}}{n}$$



Overlapped Schatten 1-norm

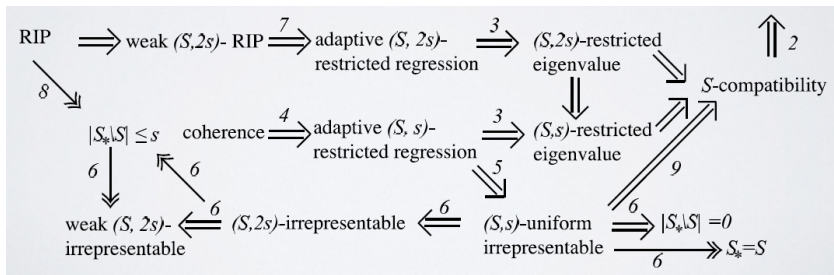
$$\|\mathcal{A}\|_{\overline{S_1/1}} := \sum_{k=1}^K \|\mathbf{A}^{(k)}\|_{\text{Tr}}$$

制限強凸性

制限強凸性（要約）：

$$\frac{\delta}{N} \|\mathcal{A}\|_2^2 \leq \frac{1}{n} \sum_{i=1}^n \langle \mathbf{X}_i, \mathcal{A} \rangle^2$$

がすべてのランク r のテンソル $\mathcal{A}$ に対して成り立つ。



線形回帰における各種制限強凸性に類する条件の関係
(Bühlmann & Van De Geer, 2011)

Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

ベイズ推定

(Suzuki, 2015)

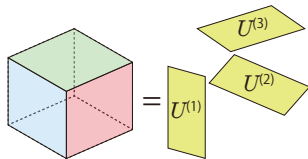
$$\hat{\mathcal{A}} = \int \mathcal{A} d\Pi(\mathcal{A} | X_{1:n}, Y_{1:n}).$$

- **CP-ランク d' のテンソルの事前分布:**

ランク d' テンソル $\mathcal{A}$ の分解

$$\mathcal{A} = [[U^{(1)}, \dots, U^{(K)}]] \quad (U^{(k)} \in \mathbb{R}^{d' \times M_k}).$$

$$\begin{aligned} & \pi(U^{(1)}, \dots, U^{(K)} | d') \\ & \propto \exp \left\{ -\frac{d'}{2\sigma_p^2} \sum_{k=1}^K \text{Tr}[U^{(k)\top} U^{(k)}] \right\}, \end{aligned}$$



- **ランクの事前分布:** ある定数 $1 > \xi > 0$ を用いて

$$\pi(d') \propto \xi^{d'} \quad \left(\sum_{d'=1}^{d^{\max}} \pi(d') = 1 \right).$$

$$\frac{1}{N} \|\hat{\mathcal{A}} - \mathcal{A}^*\|_2^2 \leq C \frac{d(\sum_{k=1}^K M_k)}{n} \log(KnR^K)$$

※制限強凸性条件は不必要.

誤差上界の比較

手法	最小二乗法	行列化による 凸正則化	ベイズ
誤差上界	$\frac{\prod_{k=1}^K M_k}{n}$	$\frac{d\sqrt{\prod_k M_k}}{n}$	$\frac{d(\sum_{k=1}^K M_k) \log(n)}{n}$

大

小

ミニマックス最適
up to $O(\log(n))$
(Suzuki, 2015)

テンソルと行列の間には大きなギャップがある。

数値実験

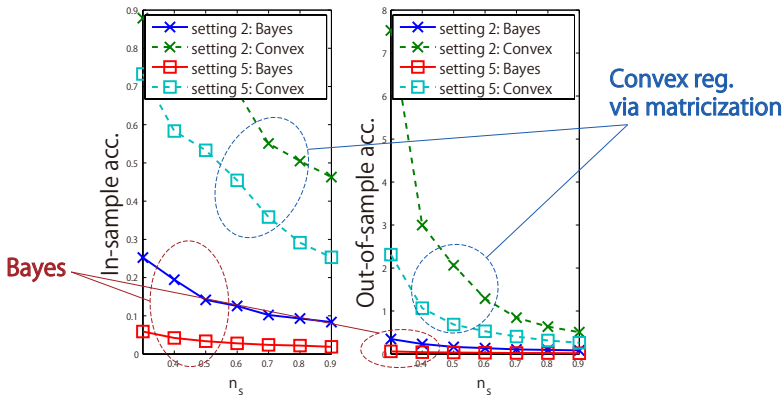


Figure: ベイズ推定法と凸正則化 (overlapped approach) の比較.

ミニマックス最適性

状況：テンソル補完

- CP-ランク d かつ max-ノルムが R 以下のテンソルの集合:

$$\mathcal{H}_d(R) := \{ \mathcal{A} \in \mathbb{R}^{M_1 \times \cdots \times M_K} \mid \text{CP-rank } d, \|\mathcal{A}\|_{\max, 2} \leq R \}.$$

- $X_i = e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_K}$ かつ $(i_1, \dots, i_K)$ が一様.

Theorem

$\mathcal{H}_d(R)$ 上のミニマックスリスクは下から次のように抑えられる:

$$\min_{\hat{\mathcal{A}}} \max_{\mathcal{A}^* \in \mathcal{H}_d(R)} \mathbb{E}[\|\hat{\mathcal{A}} - \mathcal{A}^*\|_{L_2}^2] \gtrsim \min \left\{ \sigma^2 \frac{d(M_1 + \cdots + M_K)}{n}, R^{2K} \right\}.$$

概要

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

ベクトルから関数へ

性別：女性
年齢：20代
職業：会社員
住所：都内



監督：アンディ・ウォシャウスキー
ラリー・ウォシャウスキー.
粗筋：ニューヨークの会社でしが
ないコンピュータプログラマーとして
働くトマス・アンダーソンには、裏
世界の凄腕ハッカー“ネオ”...

	映画 A	映画 B	映画 C	...	映画 X
ユーザ 1	4	8	*	...	2
ユーザ 2	2	*	2	...	*
ユーザ 3	2	4	*	...	*

$$f(x^{(1)}, x^{(2)}) = x^{(1)\top} A x^{(2)} = \sum_{r=1}^d \langle x^{(1)}, u_r^{(1)} \rangle \langle u_r^{(2)}, x^{(2)} \rangle$$

$x^{(1)}$: ユーザの特徴, $x^{(2)}$: 映画の特徴.

ベクトルから関数へ

性別：女性
年齢：20代
職業：会社員
住所：都内



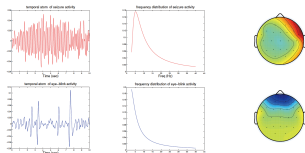
監督：アンディ・ウォシャウスキー
ラリー・ウォシャウスキー.
粗筋：ニューヨークの会社でしが
ないコンピュータプログラマーとして
働くトマス・アンダーソンには、裏
世界の凄腕ハッカー“ネオ”...

	映画 A	映画 B	映画 C	...	映画 X
ユーザ 1	4	8	*	...	2
ユーザ 2	2	*	2	...	*
ユーザ 3	2	4	*	...	*

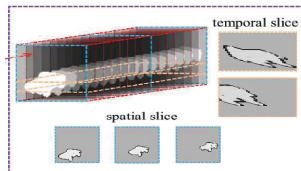
$$f(x^{(1)}, x^{(2)}) = \sum_{r=1}^d f_r^{(1)}(x^{(1)}) f_r^{(2)}(x^{(2)})$$

$x^{(1)}$: ユーザの特徴, $x^{(2)}$: 映画の特徴.

ベクトルから関数へ: 平滑化

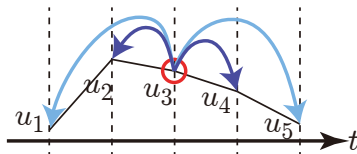


(De Vos et al., 2007)



(Cao et al., 2016)

$$\min_{\{u_r^{(k)}\}_{r,k}} \left\| \mathcal{X} - \sum_{r=1}^d u_r^{(1)} \otimes u_r^{(2)} \otimes u_r^{(3)} \right\|^2 + \underbrace{\sum_{r=1}^d \sum_{k=1}^3 u_r^{(k)\top} G u_r^{(k)}}_{\text{平滑化}}$$



$$u^\top G u = \sum_j (u_j - u_{j+1})^2$$

空間的な平滑化 $\Rightarrow$ カーネル法

(Zdunek, 2012; Yokota et al., 2015a; Yokota et al., 2015b)

非線形テンソルモデル

非線形回帰モデル

$$y_i = f(x_i) + \epsilon_i.$$

Goal: f をデータ $D_n = \{x_i, y_i\}_{i=1}^n$ から推定したい.

非線形テンソルモデル: $f_r^{(k)} \in \mathcal{H}$ (再生核ヒルベルト空間)

$$f(x^{(1)}, \dots, x^{(K)}) = \sum_{r=1}^d f_r^{(1)}(x^{(1)}) \times \dots \times f_r^{(K)}(x^{(K)})$$

線形テンソルモデル:

$$f(x^{(1)}, \dots, x^{(K)}) = \sum_{r=1}^d \langle x^{(1)}, u_r^{(1)} \rangle \times \dots \times \langle x^{(K)}, u_r^{(K)} \rangle$$

$$\text{行列の場合: } \sum_{r=1}^d \langle x^{(1)}, u_r^{(1)} \rangle \langle x^{(2)}, u_r^{(2)} \rangle = (x^{(1)})^\top \underbrace{\left(\sum_{r=1}^d u_r^{(1)} u_r^{(2)\top} \right)}_{\text{matrix}} x^{(2)}.$$

$(x^{(k)})_k$ は必ずしもユークリッド空間の元である必要はない.

推定方法

- 凸正則化法 (Signoretto et al., 2013)
 - 😊 少ない計算量.
 - 😞 (おそらく) ミニマックス最適性を満たない.
- 交互最適化法 (Suzuki et al., 2016) (本研究)
 - 😊 凸最適化の繰り返しで解ける. 少ない計算量.
 - 😞 ミニマックス最適性を満たすには強い条件が必要.
 - 😞 局所最適解の問題.
- ベイズ推定法 (Kanagawa et al., 2016) (本研究)
 - 😊 統計的な性質が良い. ミニマックス最適.
 - 😞 重い計算量.

Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

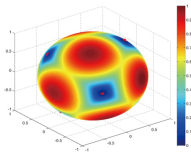
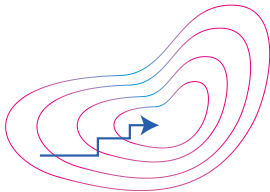
交互最適化法

$$F(\{f_r^{(k)}\}_{r,k}) := \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{r=1}^{d^*} \prod_{k=1}^K f_r^{(k)}(x_i^{(k)}) \right)^2 \quad (\text{経験誤差})$$

ある (r, k) において他の成分を固定したまま $f_r^{(k)}$ のみを更新:

$$\hat{f}_r^{(k)} = \arg \min_{\substack{f_r^{(k)} \in \mathcal{H}}} \left\{ F(f_r^{(k)} | \{\hat{f}_{r'}^{(k')}\}_{(r',k') \neq (r,k)}) + \lambda \|f_r^{(k)}\|_{\mathcal{H}}^2 \right\}.$$

- 目的関数は非凸関数.
- 各成分 $f_r^{(k)}$ に関しては 凸関数.
- 局所最適解へは収束 (e.g. 座標降下).



Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

- 交互最適化法
- **ベイズ推定法**
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

ノンパラメトリックベイズ推定法

非線形テンソルモデル:

$$f(x^{(1)}, \dots, x^{(K)}) = \sum_{r=1}^d f_r^{(1)}(x^{(1)}) \times \dots \times f_r^{(K)}(x^{(K)})$$

$$y_i = f(x_i^{(1)}, \dots, x_i^{(K)}) + \epsilon_i \quad (\epsilon_i \sim N(0, \sigma^2)).$$

再生核ヒルベルト空間: カーネル関数 $k(x, x')$

$\Leftrightarrow$ ガウス過程: 共分散関数 $\text{Cov}(f(x), f(x')) = k(x, x')$

ガウス過程事前分布を各成分 $f_r^{(k)}$ に置く.

事後平均がベイズ推定量:

$$\begin{aligned} \hat{f}(x) &= \int f(x) \, p(df | D_n) \\ \underbrace{p(df | D_n)}_{\text{事後分布}} &\propto \underbrace{p(D_n | f)}_{\text{データ } D_n \text{ の尤度}} \underbrace{p(df)}_{\text{事前分布}} \\ p(df) &= \prod_{r,k} \underbrace{p(df_r^{(k)})}_{\text{ガウス過程事前分布}} \end{aligned}$$

ノンパラメトリックベイズ推定法

非線形テンソルモデル:

$$f(x^{(1)}, \dots, x^{(K)}) = \sum_{r=1}^d f_r^{(1)}(x^{(1)}) \times \dots \times f_r^{(K)}(x^{(K)})$$
$$y_i = f(x_i^{(1)}, \dots, x_i^{(K)}) + \epsilon_i \quad (\epsilon_i \sim N(0, \sigma^2)).$$

非線形テンソルモデルの事前分布:

- 各 $(r, k) \in [d] \times [K]$ において,

$$\lambda_r^{(k)} \sim \text{Exp}(1) \quad (\text{スケーリングパラメータ})$$

$$f_r^{(k)} \sim \text{GP}_{r,k}(\cdot | \lambda_r^{(k)})$$

ただし, $\text{GP}_{r,k}(\cdot | \lambda_r^{(k)})$ はスケーリングされたカーネル関数 $k_{r,k}/\lambda_r^{(k)}$ で定義されるガウス過程.

- ランクの事前分布: $\pi(d) \propto \xi^d$ ($0 < \xi < 1$).

計算: Gibbs サンプルング + Nyström 近似

Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

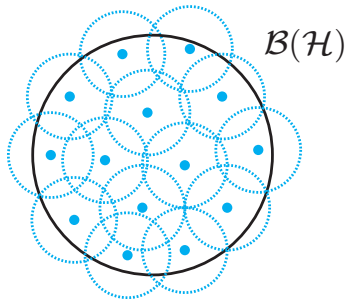
③ 非線形テンソル推定

- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

学習精度（汎化誤差）の評価

RKHS の複雑さ : カバリングナンバー

カバリングナンバー $N(B(\mathcal{H}), \epsilon, L_2(P))$: $L_2(P)$ -ノルムに関する半径 ϵ の球で単位球 $B(\mathcal{H})$ を覆うために必要な最小の個数.



カバリングナンバー条件 (s)

ある $0 < s < 1$ と定数 C が存在して、以下の不等式がすべての $\epsilon > 0$ で成り立っていると仮定する.

$$\log N(B(\mathcal{H}), \epsilon, L_2(P)) \leq C\epsilon^{-2s}.$$

s が小さければ「単純」, 大きければ「複雑」

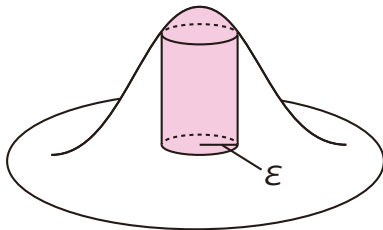
小球確率とカバリングナンバー条件

再生核ヒルベルト空間のカバリングナンバー条件は対応するガウス過程の小球確率の評価を与える.

Proposition (Kuelbs and Li (1993); Li and Shao (2001))

$$-\log(\text{GP}(\{f \mid \|f\|_{L_2(P)} \leq \epsilon\})) \sim \epsilon^{-2s/(1-s)} \quad (\text{小球確率})$$

$$\Leftrightarrow \log N(B(\mathcal{H}), \epsilon, L_2(P)) \sim \epsilon^{-2s}.$$



再生核ヒルベルト空間: カーネル関数 $k(x, x')$

$\Leftrightarrow$ ガウス過程: 共分散関数 $\text{Cov}(f(x), f(x')) = k(x, x')$

交互最適化法の収束

仮定

- f^* は *incoherence* 条件を満たしている (定義は次のスライド).
- $P(X) = P(X_1) \times \cdots \times P(X_K)$.
- $f_r^{*(k)} \in \mathcal{H}$
- その他, 細かい条件.

$\hat{f}^{[t]}$: t -回更新した後の解. s : 再生核ヒルベルト空間 $\mathcal{H}$ の複雑さ.

Theorem

d^*, K に依存した定数 C_1, C_2 が存在して, 初期解が $d(\hat{f}^{[0]}, f^*) \leq C_1$ を満たすならば確率 $1 - \delta$ で次を満たす:

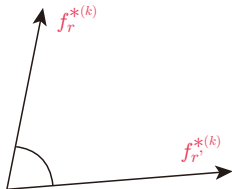
$$\|\hat{f}^{[t]} - f^*\|_{L_2(P)}^2 \leq C_2 \left(\underbrace{d^* K \left(\frac{3}{4}\right)^t}_{\text{最適化誤差}} + \underbrace{d^* K n^{-\frac{1}{1+s}}}_{\text{推定誤差}} \log(1/\delta) \right).$$

- 局所最適解へ線形収束 ($\log(n)$ 回の更新で十分)
- $d(\hat{f}^{[0]}, f^*) = O_p(1) \implies d(\hat{f}^{[\log(n)]}, f^*) = O_p\left(d^* K n^{-\frac{1}{1+s}}\right)$.
「ある程度近いところから始めれば, 最適な推定誤差を達成」

条件の詳細

- Incoherence: $\exists \mu < 1$ s.t.

$$|\langle f_r^{*(k)}, f_{r'}^{*(k)} \rangle| \leq \mu \|f_r^{*(k)}\|_{L_2} \|f_{r'}^{*(k)}\|_{L_2} \quad (r \neq r').$$



- 各成分のノルムに関する上界, 下界:

$$0 < v_{\min} \leq \|f_r^{*(k)}\|_{L_2} \leq v_{\max} \quad (\forall r, k).$$

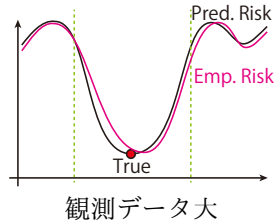
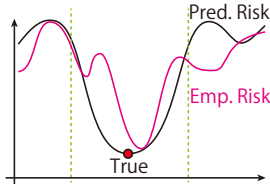
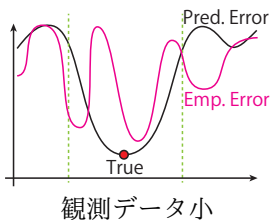
- sup-ノルムに関する条件: $0 < \exists s_2 < 1$ s.t.

$$\|f\|_{\infty} \leq C \|f\|_{L_2}^{1-s_2} \|f\|_{\mathcal{H}}^{s_2} \quad (\forall f \in \mathcal{H})$$

$$v_r = \|\prod_{k=1}^K f_r^{*(k)}\|_{L_2}, \quad \hat{v}_r = \|\prod_{k=1}^K \hat{f}_r^{(k)}\|_{L_2}, \quad f_r^{**}(k) = \frac{f_r^{*(k)}}{\|f_r^{*(k)}\|_{L_2}}, \quad \hat{f}_r^{(k)} = \frac{\hat{f}_r^{(k)}}{\|\hat{f}_r^{(k)}\|_{L_2}},$$

$$d(\hat{f}, f^*) = \max_{(r,k)} \{|\hat{v}_r - v_r| + v_r \|\hat{f}_r^{(k)} - f_r^{**}(k)\|_{L_2}\}.$$

収束理論の図解



- 予想誤差は真の関数の近傍で局所的に凸関数のような形をしている。
- 経験誤差はサンプルサイズが大きくなるにつれ、予測誤差に近くなってゆく。

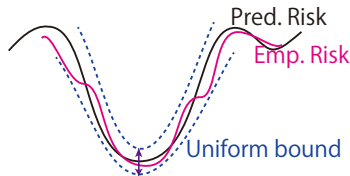
証明に用いた道具

- **Rademacher 複雑度**

$\mathcal{H}$: 関数集合

$$\mathbb{E}_x \left[\sup_{h \in \mathcal{H}} \left| \underbrace{\frac{1}{n} \sum_{i=1}^n h(x_i)}_{\text{経験誤差}} - \underbrace{\mathbb{E}[h]}_{\text{予測誤差}} \right| \right]$$

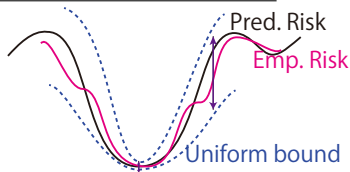
$$\leq 2\mathbb{E}_{x, \sigma} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i) \right| \right] \leq \frac{C}{\sqrt{n}}$$



ただし, $\{\sigma_i\}_{i=1}^n$ は i.i.d. の Rademacher 確率変数
 $(P(\sigma_i = 1) = P(\sigma_i = -1))$.

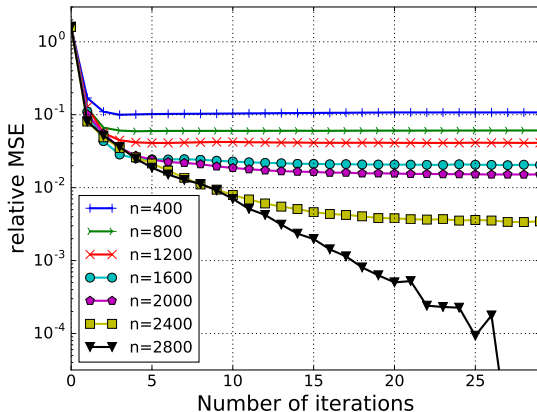
- **局所 Rademacher 複雑度 + peeling device:** 二乗損失の強凸性を利用

$$\mathbb{E}_{x, \sigma} \left[\sup_{h \in \mathcal{H}} \frac{|\frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i)|}{\|h\|_{L_2} + \lambda} \right] \leq C \frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{\lambda^{-\frac{1}{2}}}{n^{-\frac{1}{1+s}}}$$



真の周辺でよりタイトなバウンド

交互最適化の収束



MSE $E[\|\hat{f}^{[t]} - f^*\|^2]$ v.s. 更新回数 t
複数のサンプルサイズ n で比較.

ベイズ推定の理論

事後平均の収束

- $\hat{f}$: 事後平均

$$\hat{f} = \int f \Pi(\mathrm{d}f | (x_1, y_1), \dots, (x_n, y_n)).$$

- f^* : 真の関数

$$f^*(x^{(1)}, \dots, x^{(K)}) = \sum_{r=1}^{d^*} f_r^{*(1)}(x^{(1)}) \times \dots \times f_r^{*(K)}(x^{(K)}).$$

- $\|f\|_n^2 = \sum_{i=1}^n f(x_i)^2$.

Theorem

もし、真の関数が $f_r^{(k)*} \in \mathcal{H}_{r,k}$ ($\forall(r, k)$) かつ $\max_{r,k} \|f_r^{(k)*}\|_\infty \leq R$ を満たすなら、推定誤差は

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^*\|_n^2 \right] \leq C \left\{ (R+1)^{2(K-1)} \sum_{r=1}^{d^*} \sum_{k=1}^K n^{-\frac{1}{1+s_{(r,k)}}} + \frac{d^*}{n} \log \left(\frac{1}{\kappa_\xi} \right) \right\},$$

と抑えられる．ただし、 $0 < s_{(r,k)} < 1$ はカバリングナンバー条件の指数:

$$\log N(\mathcal{B}_{\mathcal{H}_{r,k}}, \epsilon, \|\cdot\|_n) = O(\epsilon^{-2s_{(r,k)}}),$$

ただし、 $\mathcal{B}_{\mathcal{H}_{r,k}}$ は $RKHS\mathcal{H}_{r,k}$ の単位球で $N(\mathcal{B}_{\mathcal{H}_{r,k}}, \epsilon, \|\cdot\|_n)$ はそのカバリングナンバー．

例: Matérn prior

$$\mathcal{X}_k = [0, 1]^p$$

$f_r^{(k)*} \in W^{\alpha,2}[0,1]^p$: ソボレフ空間 (α -回微分可能で, その微分係数が L_2)

- ナイーブな方法: テンソル分解の構造を使わず真の関数 f^* を空間全体 $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$ 上でそのまま推定.

$$n^{-\frac{2\alpha}{2\alpha+Kp}}.$$

☹ 次元の呪い

- 本研究の方法:

$$d^* K n^{-\frac{2\alpha}{2\alpha+p}}.$$

☺ 次元の呪いを回避

ガウス過程回帰のリスクバウンド

観測ノイズに適切な仮定を置くことで, 次の定理を得る.

Theorem (PAC-ベイズバウンド (Dalalyan & Tsybakov, 2008))

事前分布 Π に対して絶対連続なすべての確率測度 ρ に対して, 以下の不等式が成り立つ.

$$\mathbb{E}_{Y_{1:n}|X_{1:n}} \left[\|\hat{f} - f^*\|_n^2 \right] \leq \int \|f - f^*\|_n^2 d\rho(f) + \frac{\beta \mathcal{K}(\rho, \Pi)}{n}.$$

- **平行移動した集合の測度** ガウス測度 μ とすべての原点对称なボレル集合 A について

$$-\log(\mu(A + h)) \leq \frac{1}{2} \|h\|_{\Sigma^{-1}}^2 - \log(\mu(A)).$$

- これらを合わせると,

$$\mathbb{E}[\|\hat{f} - f^*\|_n^2] \lesssim \underbrace{\epsilon_n^2 + \frac{\|f^*\|_{\mathcal{H}}^2}{n} - \frac{1}{n} \log(\text{GP}(\{f \mid \|f\|_n^2 \leq \epsilon_n^2\}))}_{\text{小球確率}}$$

ミニマックス最適性

導かれた上界は ミニマックス最適性 を（定数倍を除いて）満たす.

ランク d^* のテンソルの集合:

$$\mathcal{H}_{(d^*, K)}(R) := \left\{ f = \sum_{r=1}^{d^*} \prod_{k=1}^K f_r^{(k)} \mid f_r^{(k)} \in \mathcal{H}_{(r, k)}(R) \right\}.$$

Theorem (ミニマックスリスク)

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{(d^*, K)}(R)} \mathbb{E}[\|f^* - \hat{f}\|_{L_2(P_{\mathcal{X}})}^2] \gtrsim d^* K n^{-\frac{1}{1+s}},$$

ただし, $\mathbb{E}[\cdot]$ は訓練データの出方に関する期待値, $\inf$ はすべての推定量 $\hat{f}$ についてとる.

ベイズ推定量は定数倍を除いてミニマックス最適性を満たす.

Outline

① 低ランク行列推定

② 線形テンソル推定

- テンソルのランク
- テンソル推定の定式化
- 凸正則化による方法
- ベイズ推定法

③ 非線形テンソル推定

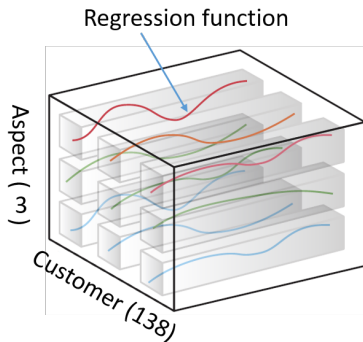
- 交互最適化法
- ベイズ推定法
- ノンパラメトリックテンソル推定の推定理論
- 数値実験: マルチタスク学習

数値実験 (実データ)

マルチタスク学習 [非線形回帰]

- Restaurant data: 138 人の客 × 3 項目 (料理, サービス, 全体)
- レストランの評価 (3 段階) を予測
- レストランは 44 次元の特徴ベクトルで表現 → レストランに対して評価を予測する関数を客と項目ごとに推定.

$$f(\text{客}, \text{項目}, x_{\text{レストラン}}) = \sum_{r=1}^d \alpha_{\text{客}, r} \beta_{\text{項目}, r} f_r(x_{\text{レストラン}})$$

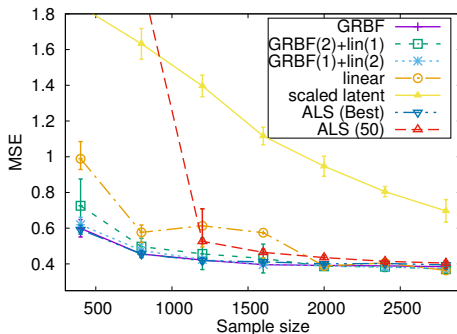


数値実験 (実データ)

マルチタスク学習 [非線形回帰]

- Restaurant data: 138 人の客 × 3 項目 (料理, サービス, 全体)
- レストランの評価 (3 段階) を予測
- レストランは 44 次元の特徴ベクトルで表現 → レストランに対して評価を予測する関数を客と項目ごとに推定.

$$f(\text{客}, \text{項目}, x_{\text{レストラン}}) = \sum_{r=1}^d \alpha_{\text{客}, r} \beta_{\text{項目}, r} f_r(x_{\text{レストラン}})$$

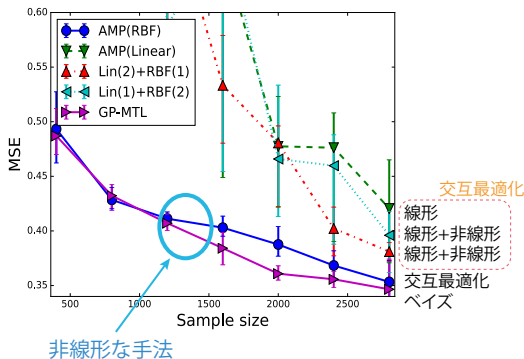


数値実験 (実データ)

マルチタスク学習 [非線形回帰]

- Restaurant data: 138 人の客 × 3 項目 (料理, サービス, 全体)
- レストランの評価 (3 段階) を予測
- レストランは 44 次元の特徴ベクトルで表現 → レストランに対して評価を予測する関数を客と項目ごとに推定.

$$f(\text{客}, \text{項目}, x_{\text{レストラン}}) = \sum_{r=1}^d \alpha_{\text{客}, r} \beta_{\text{項目}, r} f_r(x_{\text{レストラン}})$$

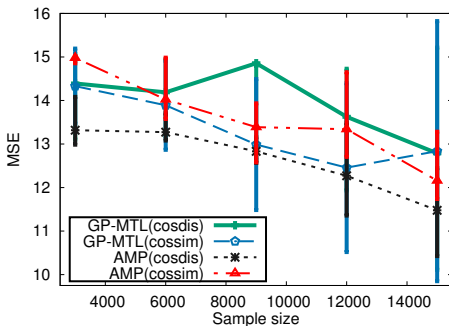


ベイズ法が若干良い. 交互最適化もそれなりに良い精度.

オンラインショッピング購買予測

Yahoo!ショッピングからデータの購買データを取得

- ショップ × 商品 × 客 (508 ショップ, 100 商品)
- その店でその客がその商品をいくつ買うかを予測.
- 客は 65 次元特徴ベクトルで表現. ガウスクERNELを利用.
- ショップも特徴ベクトルで表現し, 類似度を構成.



Figuur: Yahoo! shopping の購買予測. 店間の異なる類似度で比較.

各種手法の比較

手法	行列型正則化法	線形ベイズ	交互最適化	非線形ベイズ
非線形性	線形	線形	非線形	非線形
計算量	◎	×	○	×
統計的性質	×	◎	○	◎

計算量と統計的性質にはトレードオフがある.

まとめ

- 線形テンソルモデルと非線形テンソルモデルで（ほぼ）最適な推定方法を導出.
- ミニマックス最適レート¹を導出.
- 統計的性質と計算量にはトレードオフの関係があった.

手法	線形モデル		
	最小二乗法	行列化による凸正則化	ベイズ
誤差上界	$\frac{\prod_{k=1}^K M_k}{n}$	$\frac{d\sqrt{\prod_{k=1}^K M_k}}{n}$	$\frac{d(\sum_{k=1}^K M_k) \log(n)}{n}$

ノンパラメトリックベイズ推定量の予測誤差バウンド

$$\mathbb{E}[\|\hat{f} - f^*\|_n^2] \leq Cd^*Kn^{-\frac{1}{1+s}}.$$

ミニマックス最適レート

$$\min_{\hat{f}} \max_{f^* \in \mathcal{H}_{(d^*, \kappa)}(R)} \mathbb{E}[\|\hat{f} - f^*\|_{L_2}^2] \gtrsim d^*Kn^{-\frac{1}{1+s}}.$$

- Barak, B., & Moitra, A. (2015). Tensor prediction, rademacher complexity and random 3-xor. *arXiv preprint arXiv:1501.06521*.
- Bennett, J., & Lanning, S. (2007). The netflix prize. *Proceedings of KDD Cup and Workshop 2007*.
- Bhojanapalli, S., & Sanghavi, S. (2015). A new sampling technique for tensors. *arXiv preprint arXiv:1502.05023*.
- Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
- Candès, E., & Tao, T. (2009). The power of convex relaxations: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56, 2053–2080.
- Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9, 717–772.
- Cao, W., Wang, Y., Sun, J., Meng, D., Yang, C., Cichocki, A., & Xu, Z. (2016). Total variation regularized tensor RPCA for background subtraction from compressive measurements. *IEEE Transactions on Image Processing*, 25, 4075–4090.
- Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition. *Psychometrika*, 35, 283–319.
- Dalalyan, A., & Tsybakov, A. B. (2008). Aggregation by exponential weighting sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72, 39–61.

- De Lathauwer, L. (2006). A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM journal on Matrix Analysis and Applications*, 28, 642–666.
- De Lathauwer, L., De Moor, B., & Vandewalle, J. (2004). Computation of the canonical decomposition by means of a simultaneous generalized schur decomposition. *SIAM journal on Matrix Analysis and Applications*, 26, 295–327.
- De Vos, M., Vergult, A., De Lathauwer, L., De Clercq, W., Van Huffel, S., Dupont, P., Palmini, A., & Van Paesschen, W. (2007). Canonical decomposition of ictal scalp eeg reliably detects the seizure onset zone. *NeuroImage*, 37, 844–854.
- Hardt, M., & Wootters, M. (2014). Fast matrix completion without the condition number. *Proceedings of The 27th Conference on Learning Theory* (pp. 638–678).
- Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16, 1–84.
- Hillar, C. J., & Lim, L.-H. (2013). Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60, 45.
- Hitchcock, F. L. (1927). Multiple invariants and generalized rank of a p-way matrix or tensor. *Journal of Mathematics and Physics*, 7, 39–79.

- Jain, P., Meka, R., & Dhillon, I. S. (2010). Guaranteed rank minimization via singular value projection. *Advances in Neural Information Processing Systems* (pp. 937–945).
- Jain, P., & Oh, S. (2014). Provable tensor factorization with missing data. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Weinberger (Eds.), *Advances in neural information processing systems 27*, 1431–1439. Curran Associates, Inc.
- Kanagawa, H., Suzuki, T., Kobayashi, H., Shimizu, N., & Tagami, Y. (2016). Gaussian process nonparametric tensor estimator and its minimax optimality. *Proceedings of the 33rd International Conference on Machine Learning (ICML2016)* (pp. 1632–1641).
- Koltchinskii, V. (2012). *Sharp oracle inequalities in low rank estimation* (Technical Report). arXiv:1210.1144.
- Kuelbs, J., & Li, W. V. (1993). Metric entropy and the small ball problem for gaussian measures. *Journal of Functional Analysis*, 116, 133–157.
- Leurgans, S., Ross, R., & Abel, R. (1993). A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14, 1064–1083.
- Li, W. V., & Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic Processes: Theory and Methods*, 19, 533–597.

- Mu, C., Huang, B., Wright, J., & Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. *Proceedings of the 31th International Conference on Machine Learning* (pp. 73–81).
- Negahban, S., Ravikumar, P., Wainwright, M. J., & Yu, B. (2012). A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Statistical Science*, 27, 538–557.
- Negahban, S., & Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13, 1665–1697.
- Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33, 2295–2317.
- Phien, H. N., Tuan, H. D., Bengua, J. A., & Do, M. N. (2016). Efficient tensor completion: Low-rank tensor train. *arXiv preprint arXiv:1601.01083*.
- Rohde, A., & Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39, 887–930.
- Shah, P., Rao, N., & Tang, G. (2015). Optimal low-rank tensor recovery from separable measurements: Four contractions suffice. *arXiv preprint arXiv:1505.04085*.
- Signoretto, M., Lathauwer, L. D., & Suykens, J. A. K. (2013). Learning tensors in reproducing kernel Hilbert spaces with multilinear spectral penalties. *CoRR*, abs/1310.4977.

- Srebro, N., Alon, N., & Jaakkola, T. (2005). Generalization error bounds for collaborative prediction with low-rank matrices. *Advances in Neural Information Processing Systems (NIPS) 17*.
- Sun, R., & Luo, Z.-Q. (2014). Guaranteed matrix completion via non-convex factorization. *arXiv preprint arXiv:1411.8003*.
- Sun, W., Wang, Z., Liu, H., & Cheng, G. (2015). Non-convex statistical optimization for sparse tensor graphical model. *Advances in Neural Information Processing Systems* (pp. 1081–1089).
- Suzuki, T. (2015). Convergence rate of Bayesian tensor estimator and its minimax optimality. *Proceedings of the 32nd International Conference on Machine Learning (ICML2015)* (pp. 1273–1282).
- Suzuki, T., Kanagawa, H., Kobayashi, H., Shimizu, N., & Tagami, Y. (2016). Minimax optimal alternating minimization for kernel nonparametric tensor learning. *The 30th Annual Conference on Neural Information Processing Systems (NIPS2016)* (p. accepted).
- Tomioka, R., & Suzuki, T. (2013). Convex tensor decomposition via structured Schatten norm regularization. *Advances in Neural Information Processing Systems 26* (pp. 1331–1339). NIPS2013.
- Tomioka, R., Suzuki, T., Hayashi, K., & Kashima, H. (2011). Statistical performance of convex tensor decomposition. *Advances in Neural Information Processing Systems 24* (pp. 972–980). NIPS2011.

- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279–311.
- Wimalawarne, K., Sugiyama, M., & Tomioka, R. (2014). Multitask learning meets tensor factorization: task imputation via convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Weinberger (Eds.), *Advances in neural information processing systems 27*, 2825–2833. Curran Associates, Inc.
- Yokota, T., Zdunek, R., Cichocki, A., & Yamashita, Y. (2015a). Smooth nonnegative matrix and tensor factorizations for robust multi-way data analysis. *Signal Processing*, 113, 234–249.
- Yokota, T., Zhao, Q., & Cichocki, A. (2015b). Smooth parafac decomposition for tensor completion. *arXiv preprint arXiv:1505.06611*.
- Zdunek, R. (2012). Approximation of feature vectors in nonnegative matrix factorization with gaussian radial basis functions. *International Conference on Neural Information Processing* (pp. 616–623).
- Zhang, Z., & Aeron, S. (2015). Exact tensor completion using t-svd. *arXiv preprint arXiv:1502.04689*.
- Zhao, T., Wang, Z., & Liu, H. (2015). A nonconvex optimization framework for low rank matrix estimation. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett and R. Garnett (Eds.), *Advances in neural information processing systems 28*, 559–567. Curran Associates, Inc.

Zheng, Q., & Tomioka, R. (2015). Interpolating convex and non-convex tensor decompositions via the subspace norm. *Advances in Neural Information Processing Systems* (pp. 3088–3095).