テンソルのルジャンドル分解

Tensor Decomposition via Legendre Transformation

杉山 麿人[1,2*]　中原 裕之[3]　津田 宏治[4]
Mahito Sugiyama[1,2]　Hiroyuki Nakahara[3]　Koji Tsuda[4]

[1] 国立情報学研究所
[1] National Institute of Informatics
[2] 独立行政法人科学技術振興機構, さきがけ
[2] JST, PRESTO
[3] 理化学研究所　脳科学総合研究センター
[3] RIKEN Brain Science Institute
[4] 東京大学大学院　新領域創成科学研究科
[4] Graduate School of Frontier Sciences, The University of Tokyo

Abstract: We present a novel *nonnegative tensor decomposition* method, called *Legendre decomposition*, which decomposes a given tensor into a multiplicative combination of parameters. Thanks to the well-developed theory of information geometry, the reconstructed tensor is unique and always minimizes the KL divergence from an input tensor. We empirically show that Legendre decomposition can more accurately reconstruct tensors than nonnegative Tucker and CP decompositions.

1 はじめに

行列やテンソルの分解は，多次元配列で表現されたデータを解析する基本的な機械学習手法であり，様々な分野で用いられている．例えば，コンピュータビジョン [29, 30] や推薦システム [26]，信号処理 [9]，脳科学でのデータ処理 [6] などがある．現在の標準的な手法は，行列に対する NMF（非負行列分解）[18, 19]，テンソルに対する CANDECOMP/PARAFAC (CP) 分解 [12] やタッカー分解 [28] である．CP 分解は，入力のテンソルをランク 1 テンソルの和に分解し，タッカー分解は，コアテンソルと行列の積で入力テンソルを近似する．これまで，行列やテンソルの分解の研究が進み，効率的に数少ない要素の組合せで入力を近似するために，多くの関連手法が提案されている [14].

しかし，このような近年の発達にも関わらず，ベクトルや行列を含む任意の次数のテンソルに対して分解を可能とする統計理論は，未だ確立されていない．さらに，CP 分解やタッカー分解は，非凸最適化を含むため，大域最適解は保証されない．これらの分解を凸最適化へ変換するための手法も提案されているが，分散に対する制約などのデータに関する制約が必要となってしまう [20, 27].

本稿では，これらの問題を解決する**情報幾何**（information geometry）[5] を用いた行列やテンソル分解の新しい枠組みとして，**ルジャンドル分解**（Legendre decomposition）を提案する．提案手法では，任意の次数の非負テンソルが，統計多様体上の（離散）確率分布として扱われ，入力テンソルの再構築可能なテンソルからなる部分多様体への**射影**（projection）として分解が実現される．鍵となるのは，テンソルのインデックスによって生成される**半順序構造**（partial order structure）[10, 11] であり，これによってテンソルを多様体上の確率分布として情報幾何的に扱うことが可能となる．

情報幾何の結果を利用することで，提案手法は，以下のような特徴がある．(1) 再構成されるテンソルは，必ず唯一存在し，入力テンソルからの Kullback–Leibler (KL) ダイバージェンスを最小化する．(2) 分解は凸最適化として定式化され，勾配法で計算できる．さらに，**自然勾配法**（natural gradient）を用いることで，より効率的に計算できる．(3) 提案手法は柔軟であり，欠損値や 0 などを取り除いた部分テンソルに対しても適用できる．さらに，提案手法は既存の統計モデルとも密接な関係があり，特

* 連絡先：国立情報学研究所
〒 101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: mahito@nii.ac.jp

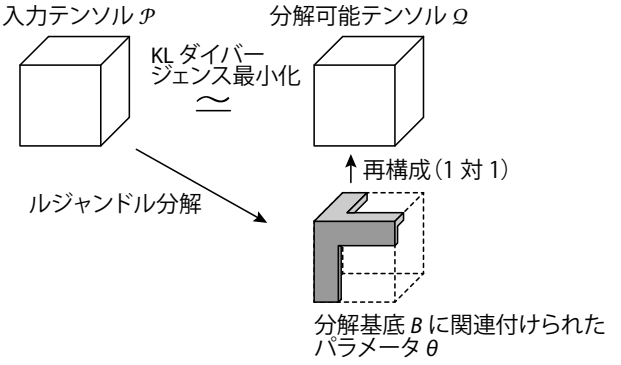

図1　ルジャンドル分解の概要.

に，**ボルツマンマシン**（Boltzmann machine）[1] の拡張として解釈でき，テンソル分解とグラフィカルモデルとの関係 [8, 31, 32] に対して，新しい見方を与える．また，提案手法で用いるモデルは，指数形分布族に含まれており，分解で用いるパラメータ θ は自然パラメータ，制約で用いる η は期待値に，それぞれ対応している．

本稿は，以下のように構成されている．節 2でルジャンドル分解を導入し，節 3ではルジャンドル分解の理論を示す．節 4で実験によって提案手法を検証し，節 5で本稿の貢献をまとめる．

2　ルジャンドル分解

まず，提案手法である**ルジャンドル分解**（Legendre decomposition）を行列に対して導入し，その後一般のテンソルに拡張する．提案手法の概要を図 1に示す．

2.1　定式化

非負行列 $X = (x_{ij}) \in \mathbb{R}_{\geq 0}^{m \times n}$ とする．この行列 X を確率分布として扱うために，以降では常に X を全成分の和で割って正規化した行列 $P = X / \sum_{ij} x_{ij}$ を用いる．自然数 m に対して，$[m] = \{1, 2, \ldots, m\}$ と書く．

ルジャンドル分解を導入する．まず，インデックスの部分集合 $B \subseteq [m] \times [n]$ を**分解基底**（decomposition basis）として定める．必ず $(1,1) \in B$ としておく．そして，正規化された行列 P を，B に関連付けられたパラメータの乗法的組合せで近似する．例えば，$B = \{(i,j) \mid i = 1 \text{ or } j = 1\}$ とすることで，P を 1 つ目の行と列という 2 つのベクトルで近似できる．数学的には，行列 $Q \in \mathbb{R}_{\geq 0}^{m \times n}$ が B で**分解可能**（decomposable）であることを，$q_{ij}, (i,j) \in \Omega$ が $|B|$ 個のパラメータ $\theta_{ij} \in \mathbb{R}, (i,j) \in B$ を用いて

$$q_{ij} = \prod_{(k,l) \in \downarrow(i,j)} \exp(\theta_{kl}), \tag{1}$$

$$\downarrow(i,j) = \{\, (k,l) \in B \mid k \leq i \text{ and } l \leq j \,\},$$

と表されることと定義する．行列 Q を確率分布として見たとき，インデックスの定義域 Ω は確率質量関数の定義域に対応しており，$\sum_{(i,j) \in \Omega} y_{ij} = 1$ である．ここで，$\Omega = [m] \times [n]$ とするのが最も自然だが，提案手法では $B \subseteq \Omega$ を満たす任意の部分集合 $\Omega \subseteq [m] \times [n]$ を用いることができる．例えば，P 中の欠損値や 0 成分を Ω から取り除いておくことができる．

提案手法は，基底 B を用いて，以下の条件を満たす分解可能な行列 $Q \simeq P$ を見つける．

$$\eta_{ij} = \hat{\eta}_{ij} \quad \forall (i,j) \in B. \tag{2}$$

ここで，$Q \in \mathbb{R}_{\geq 0}^{m \times n}$ に対する制約変数 $\eta_{ij} \in (0,1)$ は

$$\eta_{ij} = \sum_{(k,l) \in \uparrow(i,j)} q_{kl},$$
$$\uparrow(i,j) = \{(k,l) \in \Omega \mid k \geq i \text{ and } l \geq j\}$$

と定義され，$\hat{\eta}_{ij}$ は q_{kl} を p_{kl} と入れ替えることで得られる．節 3の数式(6)で示すように，パラメータ θ と制約 η は**ルジャンドル変換**（Legendre transformation）で接続されており，この関係が鍵となるため，提案手法を**ルジャンドル分解**（Legendre decomposition）と呼ぶ．基底 B に対するただ 1 つの条件は，任意の $(i,j), (k,l) \in B$ に対して $\hat{\eta}_{ij} \neq \hat{\eta}_{kl}$ となることである．これは，P の非零成分のインデックスからなる集合 D に対して，基底 B が $\uparrow(i,j) \cap D \neq \uparrow(k,l) \cap D$ を満たしていれば良い．

節 3で示すように，正規化した行列の集合が持つ双対平坦構造によって，任意の P に対して，数式 (2)の条件を満たす分解可能な行列 Q は必ず以下の性質を持つ．

1. Q は常に存在する．
2. Q は唯一に定まる．
3. Q は Kullback–Leibler (KL) ダイバージェンスの意味で最良近似である．すなわち，

$$Q = \operatorname*{argmin}_{R \in \mathcal{S}_B} D_{\mathrm{KL}}(P, R),$$
$$\mathcal{S}_B = \left\{\, R \in \mathbb{R}_{\geq 0}^{m \times n} \;\middle|\; R \text{ は} B \text{ で分解可能} \,\right\}.$$

ここで，KL ダイバージェンスは，各行列を離散確率分布として捉えて定義される．すなわち，

$$D_{\mathrm{KL}}(P, Q) = \sum_{(i,j) \in \Omega} x_{ij} \log \frac{x_{ij}}{y_{ij}}.$$

例 1. 基底 B として，2 つの極端な例を挙げる．もし B が単集合 $\{(1,1)\}$ であるとき，分解可能な Q は必ず一様であり，任意の $(i,j) \in \Omega$ に対して $q_{ij} = \exp(\theta_{11}) = 1/|\Omega|$ である．対照的に，もし $B = \Omega$ であれば，任意の行列 P それ自身が分解可能となるため，$(i,j) \in \Omega$ に対して $q_{ij} = p_{ij}$ となるため，次元の削減は起きない．

ルジャンドル分解は，入力テンソル P そのものではなく，成分ごとに対数をとった $\log(P)$ の低ランク近似として捉えることができる．分解可能な行列 $Q \simeq P$ に対して，$\log(Q)$ のランクは $(i,j) \in B$ ならば $t_{ij} = \theta_{ij}$，そうでなければ $t_{ij} = 0$ として定まるパラメータ行列 $T = (t_{ij}) \in \mathbb{R}^{m \times n}$ と一致し，

$$\log q_{ij} = \sum_{(k,l) \in \downarrow (i,j)} t_{kl},$$

が成り立つので，$\log(Q)$ のランクは T のランクと一致する．したがって，例えば l 個の行からなる分解基底 B を用いると，$\mathrm{rank}(\log(Q)) \leq l$ が常に成り立つ．

提案手法は，行列から多次元配列，すなわち**テンソル**（tensor）へ，素直に一般化できる．N 階のテンソル $\mathcal{P} \in \mathbb{R}_{\geq 0}^{I_1 \times I_2 \times \cdots \times I_N}$ が与えられたとする．記述を単純化するために，各成分 $p_{i_1 i_2 \ldots i_N}$ を $v = (i_1, i_2, \ldots, i_N) \in \Omega \subseteq [I_1] \times [I_2] \times \cdots \times [I_N]$ を用いて p_v と書く．分解基底 $B \subseteq \Omega$ で $(1,1,\ldots,1) \in B$ を満たすものに対して，分解可能なテンソル $\mathcal{Q} \in \mathbb{R}_{\geq 0}^{I_1 \times I_2 \times \cdots \times I_N}$ は

$$q_v = \prod_{u \in \downarrow v} \exp(\theta_u), \quad \downarrow v = \{\, u \in B \mid u \leq v \,\}, \tag{3}$$

と定まる．ここで，パラメータ $\theta_u \in \mathbb{R}$ に対して，$u \leq v$, $u = (j_1, \ldots, j_N), v = (i_1, \ldots, j_N)$ は $j_1 \leq i_1$, $j_2 \leq i_2$, $\ldots$, $j_N \leq i_N$ と定める．

基底 B を用いて分解可能なテンソル全体からなる集合を $\boldsymbol{\mathcal{S}}_B$ する．ルジャンドル分解は，以下の条件を満たすテンソル $\mathcal{Q} \in \boldsymbol{\mathcal{S}}_B$ を見つける．

$$\eta_v = \hat{\eta}_v, \quad \forall v \in B, \tag{4}$$

ここで，η_v は

$$\eta_v = \sum_{u \in \uparrow v} q_u, \quad \uparrow v = \{\, u \in \Omega \mid u \geq v \,\} \tag{5}$$

として与えられ，$\hat{\eta}_v$ は q_u を p_u で置き換えると得られる．行列のときと同様に，数式 (4)を満たすテンソル $\mathcal{Q}$ は常にただ 1 つ存在し，KL ダイバージェンスを最小かする．すなわち，

$$\mathcal{Q} = \operatorname*{argmin}_{\mathcal{R} \in \boldsymbol{\mathcal{S}}_B} D_{\mathrm{KL}}(\mathcal{P}, \mathcal{R}).$$

テンソル $\log(\mathcal{Q})$ のランクは，$v \in B$ ならば $t_v = \theta_v$，そうでなければ $t_v = 0$ として得られるパラメータテンソル $\mathcal{T}$ のランクと一致する．

2.2 アルゴリズム

ルジャンドル分解における KL ダイバージェンスの最小化を達成するために,2 つの勾配法を提案する.KL ダイバージェンス $D_{\mathrm{KL}}(\mathcal{P}, \mathcal{Q})$ は θ に関して凸なので，アルゴリズム 1に示す標準的な**勾配降下法**（gradient descent）によって大域最適解を求めることができる．ここで，$\varepsilon > 0$ は学習率を表す．数式 (7)で示すように，パラメータ θ_v の勾配は

$$\frac{\partial}{\partial \theta_v} D_{\mathrm{KL}}(\mathcal{P}, \mathcal{Q}) = \hat{\eta}_v - \eta_v,$$

として得られる，各反復において，θ から $\mathcal{Q}$ の計算（アルゴリズム 1の行 5）の計算量が $O(|\Omega||B|)$ で，$\mathcal{Q}$ から η の計算（アルゴリズム 1の行 6）の計算量が $O(|\Omega|)$ なので，一回の反復の計算量は $O(|\Omega||B|)$ となる．しがたって，全体の計算量は，反復回数を h として $O(h|\Omega||B|^2)$ となる．

勾配降下法は，一般的に効率的な学習アルゴリズムであるが，ルジャンドル分解では，各反復において θ からの「デコード」と η への「エンコード」を繰り返す必要があり，効率性が損なわれる．そこで，反復回数を減らすために，2 次収束する最適化手法である**自然勾配法**（natural gradient）[2] を提案する（アルゴリズム 2）．ここでもまた，KL ダイバージェンス $D_{\mathrm{KL}}(\mathcal{P}, \mathcal{Q})$ は θ に関して凸なので，自然勾配法は常に大域最適解を見つけることが保証される．これは，この自然勾配法が凸領域への

Algorithm 1: 勾配降下法

```
GradientDescent(P, B)
    θ を初期化する;          // 例えば θ_v = 0, ∀v ∈ B
    repeat
        foreach v ∈ B do
            現在の θ から Q を計算する;
            Q から η を計算する;
            θ_v ← θ_v + ε(η̂_v − η_v);
    until θ が収束する;
```

Algorithm 2: 自然勾配法

```
NaturalGradient(P, B)
    θ を初期化する;          // 例えば θ_v = 0, ∀v ∈ B
    repeat
        現在の θ から Q を計算する;
        Q から η を計算して，Δη ← η − η̂;
        数式 (8) を用いてフィッシャー情報行列 G と
          その逆行列 G^{-1} を求める;
        θ ← θ − G^{-1}Δη
    until θ が収束する;
```

Bregman アルゴリズムになっているためである [7]．分解基底 $B = \{v_1, v_2, \ldots, v_b\}$ とし，$\theta = (\theta_{v_1}, \ldots, \theta_{v_b})^T$, $\eta = (\eta_{v_1}, \ldots, \eta_{v_b})^T$ とする．現在のパラメータ θ を θ_{next} へ更新する際，自然勾配法は以下を用いる．

$$\Delta\eta = -G\Delta\theta, \quad \begin{cases} \Delta\eta = \eta - \hat{\eta}, \\ \Delta\theta = \theta_{\text{next}} - \theta \end{cases}$$

したがって，更新式は

$$\theta_{\text{next}} = \theta - G^{-1}\Delta\eta$$

となる．ここで，$G = (g_{ij}) \in \mathbb{R}^{b\times b}$ は数式 (8)で与えられるフィッシャー情報行列である．この自然勾配法は，

$$\frac{\partial^2}{\partial\theta_{v_i}\partial\theta_{v_j}} D_{\mathrm{KL}}(\mathcal{P}, \mathcal{Q}) = -\frac{\partial\eta_{v_i}}{\partial\theta_{v_j}} = -g_{ij}$$

であり，フィッシャー情報行列が（負の）ヘシアン行列になっておりニュートン法に一致する．各反復の計算量は $O(|\Omega||B| + |B|^3)$ である．ここで，$\mathcal{Q}$ を θ から求めるための計算量が $O(|\Omega||B|)$ であり，G の逆行列を求めるための計算量が $O(|B|^3)$ である．したがって，全体での計算量は，反復回数を h' とすると $O(h'|\Omega||B| + h'|B|^3)$ となる．

2.3 統計モデルの関係

ルジャンドル分解と他の統計モデルとの興味深い関連を示す．特に，ルジャンドル分解はボルツマンマシンによる学習 [1] の一般化として捉えることができる．これまでにも，テンソル分解とグラフィカルモデルとの関連は議論されていたが [8, 31, 32]，本稿は，グラフィカルモデルそのものではなく，それらのモデルが生成する確率質量関数の定義域（離散確率変数の値域）に着目する．

2.3.1 指数形分布族

正規化されたテンソルからなる集合

$$\boldsymbol{\mathcal{S}} = \left\{ \mathcal{P} \in \mathbb{R}_{>0}^{I_1\times I_2\times\cdots\times I_N} \;\middle|\; \sum\nolimits_{v\in\Omega} p_v = 1 \right\},$$

が，指数形分布族に含まれることを示す．ここで，$B = \Omega$ のとき $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{S}}_B$ であり，全てのテンソルが分解可能である．指数形分布族は，自然パラメータ θ を用いて

$$p(x, \theta) = \exp\left(\sum \theta_i k_i(x) + r(x) - A(\theta)\right)$$

と定義されるため，数式 (3)で示す提案モデル

$$p_v = \prod_{u\in\downarrow v} \exp(\theta_u) = \exp\left(\sum\nolimits_{u\in\Omega} \theta(u)k_u(v) + \psi(\theta)\right),$$

は，明らかに指数形分布族である．ここで，

$$\psi(\theta) = -\theta_{(1,\ldots,1)}, \quad k_u(v) = \begin{cases} 1 & \text{if } (1,\ldots,1) < u \le v, \\ 0 & \text{otherwise,} \end{cases}$$

かつ $r(x) = 0$ である．したがって，提案手法で用いられている θ は指数形分布族の自然パラメータに一致する．さらに，$k_u(v)$ の期待値をとることで，以下のように η が得られる．

$$\mathbb{E}\big[k_u(v)\big] = \sum_{v\in\Omega} k_u(v)p_v = \sum_{v\in\uparrow u} p_v = \eta_u.$$

しがたって，$\mathcal{P}$ のルジャンドル分解は，$\mathcal{P}$ と同じ期待値を持つ分解可能なテンソル $\mathcal{Q}$ を求める操作として捉えることができる．

2.3.2 ボルツマンマシン

ボルツマンマシンは，無向グラフ $G = (V, E)$ によって表される．ここで，V は頂点集合，$E \subseteq V \times V$ 辺集合で，$V = [N] = \{1, 2, \ldots, N\}$ と仮定しておく．ボルツマンマシン G は確率分布 P を定義し，各 N 次元ベクトル $x \in \{0, 1\}^N$ の確率は

$$p(x) = \frac{1}{Z(\theta)} \prod_{i\in V} \exp\left(\theta_i x_i\right) \prod_{\{i,j\}\in E} \exp\left(\theta_{ij} x_i x_j\right)$$

と与えられる．ここで，θ_i はバイアス，θ_{ij} は重み，$Z(\theta)$ は分配関数と呼ばれる．

ボルツマンマシンを提案手法の枠組みで表現するために，$\Omega = \{1, 2\}^N$ とし，

$$B(V) = \left\{ (i_1^a, \ldots, i_N^a) \in \Omega \;\middle|\; a \in V \right\},$$
$$i_l^a = \begin{cases} 2 & \text{if } l = a, \\ 1 & \text{otherwise,} \end{cases}$$
$$B(E) = \left\{ (i_1^{ab}, \ldots, i_N^{ab}) \in \Omega \;\middle|\; \{a, b\} \in E \right\},$$
$$i_l^{ab} = \begin{cases} 2 & \text{if } l \in \{a, b\}, \\ 1 & \text{otherwise.} \end{cases}$$

とする．すると，明らかに，ボルツマンマシン G で表現される確率分布の集合は，$B = B(V) \cup B(E)$ かつ $\psi(\theta) = -\theta_{(1,\ldots,1)} = Z(\theta)$ としたときの $\boldsymbol{\mathcal{S}}_B$，すなわち，数式 (3)で定義される，基底 $B(V) \cup B(E)$ を用いて分解可能な N 階のテンソルの集合と一致する．さらに，N 階のテンソル $\mathcal{P} \in \mathbb{R}_{\ge 0}^{2\times 2\times\cdots\times 2}$ をデータから推定された経験分布とし，各 p_v は 2 値ベクトル $v - (1, \ldots, 1) \in \{0, 1\}^N$ の確率とする．このとき，$B = B(V) \cup B(E)$ としたときにルジャンドル分解で得られるテンソル $\mathcal{Q}$ は，ボルツマンマシン $G = (V, E)$ で学習される分布と一致する．数式 (4)で示した条件は，よく知られているボルツマン

マシンの学習方程式であり，$\hat{\eta}$ と η はそれぞれデータとモデルの分布の期待値に対応する．

したがって，ルジャンドル分解は，以下の 3 点についてボルツマンマシンの一般化となっている．

1. 確率変数の値域は，2 値のみに制限されず，任意の順序集合を用いることができる．すなわち，$\{0,1\}^N$ が任意の $I_1, I_2, \ldots, I_N \in \mathbb{N}$ に対して $[I_1] \times [I_2] \times \cdots \times [I_N]$ へ拡張される．
2. パラメータ θ の基底 B は，$B(V) \cup B(E)$ だけでなく，$[I_1] \times [I_2] \times \cdots \times [I_N]$ の任意の部分集合をとることができ，高次相関 [23] を考慮することができる．
3. 確率変数の値域は，$\{0,1\}^N$ だけでなく，任意の部分集合 $\Omega \subseteq [I_1] \times [I_2] \times \cdots \times [I_N]$ で $B \subseteq \Omega$ を満たすものが可能である．これによって，欠損値などの不必要な成分を取り除き，高速な計算が可能となる．

3 理論解析

ルジャンドル分解を理論的に解析するために，**情報幾何**（information geometry）を用いてルジャンドル分解を再定式化し，その性質を示す．特に，階層的分布の情報幾何 [3, 21] として提案されたモデルの拡張である [25] で提案された対数線形モデルを用いる．この対数線形モデルは，これまでテンソル分解に用いられたことはなく，本研究が初めての適用となる．

3.1 対数線形モデルの情報幾何

集合 Ω を値域としてもつ離散確率変数 X に対して，その分布を P とする．各 $v \in \Omega$ に対して $p_v = \Pr(X = v) > 0$ かつ $\sum_{v \in \Omega} p_v = 1$ である．ここで，Ω が**半順序集合**（partially ordered set; poset）で，かつ最小の要素 $\bot \in \Omega$ が存在すると仮定する．この $\bot$ を取り除いた集合を $\Omega^+ = \Omega \setminus \{\bot\}$ と書く．

半順序集合 Ω に対して，**ゼータ関数**（zeta function）ζ と**メビウス関数**（Möbius function）μ という 2 つの関数 [22], [16, Chapter 3.1] を用いる．ゼータ関数は $u \leq v$ ならば $\zeta(u,v) = 1$，そうでなければ $\zeta(u,v) = 0$ と定義され，メビウス関数はその畳み込み逆関数として定義される．すなわち，

$$\sum_{v \in \Omega} \zeta(u,v)\mu(v,w) = \sum_{v \in \Omega} \mu(u,v)\zeta(v,w) = \delta(u,w)$$

である．ここで δ はクロネッカーのデルタであり，$u = w$ なら $\delta(u,w) = 1$，そうでなければ 0 をとる．メビウス関数は帰納的に定義することもでき，$u = v$ ならば $\mu(u,v) = 1$，$u < v$ ならば $\mu(u,v) = -\sum_{u \leq w < v} \mu(u,w)$，$u > v$ ならば $\mu(u,v) = 0$ である．

N 階のテンソルに対するルジャンドル分解においては，$\Omega \subseteq [I_1] \times [I_2] \times \cdots \times [I_N]$ はインデックス間の順序 $\leq$ に関して半順序集合であり，$\bot = (1, 1, \ldots, 1)$ である．したがって，各分布 P は，提案手法で扱う（正規化された）テンソル $\mathcal{P}$ に直接対応する．

[25] によって提案された半順序集合 Ω 上の対数線形モデルは，

$$\log p_v = \sum_{u \in \Omega} \zeta(u,v)\theta_u,$$

として与えられ，これは，$B = \Omega$ とすると，$\sum_{u \in \Omega} \zeta(u,v)\theta_u = \sum_{u \in \downarrow v} \theta_u$ となるため，数式 (3)で定義した分解可能なテンソルに対応する．また，θ は

$$\theta_v = \sum_{u \in \Omega} \mu(u,v) \log p_u.$$

と与えられる．さらに，η は

$$\eta_v = \sum_{u \in \Omega} \zeta(v,u) p_u = \sum_{u \in \uparrow v} p_u, \quad p_v = \sum_{u \in \Omega} \mu(v,u)\eta_u,$$

と導入され，これは数式 (5)で導入した制約変数に対応する．

分布全体の集合 $\boldsymbol{\mathcal{S}} = \{P \mid 0 < p_v < 1 \text{ and } \sum p_v = 1\}$ は，**双対平坦**（dually flat）な統計多様体 [25] になっており，これは情報幾何において鍵となる構造である [5, Chapter 6]．双対平坦多様体においては，2 つのパラメータ θ と η が $\boldsymbol{\mathcal{S}}$ の双対な座標系となっており，それらは**ルジャンドル変換**（Legendre transformation）で接続されている．すなわち，2 つの凸関数

$$\psi(\theta) = -\theta_\bot = -\log p_\bot, \quad \varphi(\eta) = \sum\nolimits_{v \in \Omega} p_v \log p_v,$$

があって，

$$\theta = \nabla\varphi(\eta), \quad \eta = \nabla\psi(\theta) \tag{6}$$

という関係が成り立つ．さらに，θ と η は直交しており，

$$\mathbb{E}\left[\frac{\partial \log p_w}{\partial \theta_u} \frac{\partial \log p_w}{\partial \eta_v}\right] = \delta(u,v)$$

となる．したがって，以下が成り立つ．

$$\begin{aligned} \frac{\partial}{\partial \theta_v} D_{\mathrm{KL}}(P,Q) &= \frac{\partial}{\partial \theta_v} \sum_{u \in \Omega} p_u \log q_u \\ &= \frac{\partial}{\partial \theta_v} \sum_{u \in \Omega} p_u \sum_{w \in \downarrow u \setminus \{\bot\}} \theta_w - \frac{\partial}{\partial \theta_v} \sum_{u \in \Omega} p_u \psi(\theta) = \hat{\eta}_v - \eta_v. \end{aligned} \tag{7}$$

さらに，リーマン計量 g は

$$g(\theta) = \nabla\nabla\psi(\theta), \quad g(\eta) = \nabla\nabla\varphi(\eta)$$

となり，これは θ と η の勾配になっており，**フィッシャー情報行列**（Fisher information matrix）と一致する．具体的には，$u, v \in \Omega^+ = \Omega \setminus \{\perp\}$ に対して，

$$g_{uv}(\theta) = \frac{\partial \eta_u}{\partial \theta_v} = \mathbb{E}\left[\frac{\partial \log p_w}{\partial \theta_u}\frac{\partial \log p_w}{\partial \theta_v}\right] = \sum_{w\in\Omega} \zeta(u,w)\zeta(v,w)p_w - \eta_u\eta_v, \tag{8}$$

$$g_{uv}(\eta) = \frac{\partial \theta_u}{\partial \eta_v} = \mathbb{E}\left[\frac{\partial \log p_w}{\partial \eta_u}\frac{\partial \log p_u}{\partial \eta_v}\right] = \sum_{w\in\Omega} \mu(w,u)\mu(w,v)p_w^{-1} \tag{9}$$

となる．この勾配を，自然勾配法（アルゴリズム 2）で用いる．

以下に示すように，$\boldsymbol{\mathcal{S}}$ の部分多様体を 2 つ導入する．

$$\boldsymbol{\mathcal{S}}_T = \{\, P \in \boldsymbol{\mathcal{S}} \mid \theta_v = T_v \text{ for all } v \in \Omega_T \,\},$$
$$\boldsymbol{\mathcal{S}}_H = \{\, P \in \boldsymbol{\mathcal{S}} \mid \eta_v = H_v \text{ for all } v \in \Omega_H \,\}$$

これらは，それぞれ $\Omega_T, \Omega_H \subseteq \Omega^+$ を満たす制約 T, H で定まる．1 つめの部分多様体 $\boldsymbol{\mathcal{S}}_T$ は θ 座標に制約があり，2 つめの部分多様体 $\boldsymbol{\mathcal{S}}_H$ は η 座標に制約がある．情報幾何では，$\boldsymbol{\mathcal{S}}_H$ は e 平坦，$\boldsymbol{\mathcal{S}}_H$ は m 平坦となる [5, Chapter 2.4]．もし $\Omega_T \cup \Omega_H = \Omega^+$ かつ $\Omega_T \cap \Omega_H = \emptyset$ が満たされれば，それらの共通部分 $\boldsymbol{\mathcal{S}}_T \cap \boldsymbol{\mathcal{S}}_H$ は必ず単集合 $\{Q\}$ となる．これは，$Q \in \boldsymbol{\mathcal{S}}_T$ 及び $Q \in \boldsymbol{\mathcal{S}}_H$ を満たす分布 Q は必ず唯一存在することを意味する．さらに，KL ダイバージェンスに関する**ピタゴラスの定理**（Pythagorean theorem）が成り立つ．任意の $P \in \boldsymbol{\mathcal{S}}_T$ と $R \in \boldsymbol{\mathcal{S}}_H$ に対して，

$$D_{\mathrm{KL}}(P,R) = D_{\mathrm{KL}}(P,Q) + D_{\mathrm{KL}}(Q,R). \tag{10}$$

となる．

3.2 射影としてのルジャンドル分解

ルジャンドル分解をテンソルの部分多様体への**射影**（projection）として定式化する．部分多様体 $\boldsymbol{\mathcal{S}}_B$ を，

$$\boldsymbol{\mathcal{S}}_B = \{\, \mathcal{Q} \in \boldsymbol{\mathcal{S}} \mid \theta_v = 0 \text{ for all } v \in \Omega \setminus B \,\},$$

と定義する．これは，分解可能なテンソルからなる集合に対応しており，θ 座標に関する制約を持つので e 平坦である．（正規化された）テンソル $\mathcal{P} \in \boldsymbol{\mathcal{S}}$ に対して，分解可能なテンソル $\mathcal{Q} \in \boldsymbol{\mathcal{S}}_B$ は，以下の条件を満たすとき，かつそのときに限り，$\mathcal{P}$ のルジャンドル分解である．

$$\begin{cases} \theta_v = 0 & \text{if } v \in \Omega \setminus B, \\ \eta_v = \hat{\eta}_v & \text{if } v \in B, \end{cases} \tag{11}$$

ここで，$\hat{\eta}_v$ は $\mathcal{P}$ から定まる．1 つ目の条件は $\mathcal{Q}$ が分解可能であることを保証しており，2 つ目の条件は数式 (4)に対応する．

ここで，2 つ目の条件は m 平坦な部分多様体に対応する．

$$\boldsymbol{\mathcal{S}}'_B = \{\, \mathcal{Q} \in \boldsymbol{\mathcal{S}} \mid \eta_v = \hat{\eta}_v \text{ for all } v \in B \,\}.$$

したがって，ルジャンドル分解は，$\mathcal{P}$ が与えられたとき，共通部分 $\boldsymbol{\mathcal{S}}_B \cap \boldsymbol{\mathcal{S}}'_B$ を見つける操作として解釈できる．これは，情報幾何において，$\mathcal{P}$ の部分多様体 $\boldsymbol{\mathcal{S}}_B$ への m **射影**（m-projection）として知られている．多様体 $\boldsymbol{\mathcal{S}}$ の θ と η に関する双対平坦な構造によって，$\mathcal{Q} \in \boldsymbol{\mathcal{S}}_B \cap \boldsymbol{\mathcal{S}}'_B$ が必ず唯一存在し，かつ KL ダイバージェンスを最小化することが保証される [4, Theorem 3]．すなわち，

$$\mathcal{Q} = \operatorname*{argmin}_{\mathcal{R}\in\boldsymbol{\mathcal{S}}_B} D_{\mathrm{KL}}(\mathcal{P},\mathcal{R})$$

が成り立つ．これに対して，分解可能なテンソル $\mathcal{R} \in \boldsymbol{\mathcal{S}}_B$ が与えられたとき，数式 (11)を満たすテンソル $\mathcal{Q}$ を見つける操作は,$\mathcal{R}$ の部分多様体 $\boldsymbol{\mathcal{S}}'_B$ への e **射影**（e-projection）と呼ばれる．実際には，アルゴリズム 1 とアルゴリズム 2 は，この e 射影を実行している．これは，最適化すべきパラメータの数が，e 射影では $|B|$ に対して，m 射影では $|\Omega \setminus B|$ となり，通常は $|B| \le |\Omega \setminus B|$ が成り立つためである．部分多様体 $\boldsymbol{\mathcal{S}}_B$ 及び $\boldsymbol{\mathcal{S}}'_B$ が凸集合なので，提案アルゴリズムの大域的収束性が保証される．

この射影において，$\mathcal{P} \in \boldsymbol{\mathcal{S}}$ であることを仮定していたが，もし基底 B が非零のインデックス集合に含まれていなかった場合は，$\mathcal{P} \notin \boldsymbol{\mathcal{S}}$ となる可能性がある．この場合においても，任意の $u, v \in B$ に対して $\hat{\eta}_u \neq \hat{\eta}_v$ が満たされていれば，ルジャンドル分解は有効であり，$\mathcal{P}$ からの KL ダイバージェンスを最小化する唯一の $\mathcal{Q}$ を求めることができる．これは，任意の $\varepsilon > 0$ に対して，$\mathcal{P}' \in \boldsymbol{\mathcal{S}}$ が存在し，$D_{\mathrm{KL}}(\mathcal{P},\mathcal{P}') < \varepsilon$ かつ各 $v \in B$ に対して $\eta_v = \eta'_v$ が成り立つ．したがって，$\mathcal{P}'$ を $\boldsymbol{\mathcal{S}}_B$ に e 射影して得られるテンソル $\mathcal{Q}$ に対しても，$\mathcal{Q} = \operatorname{argmin}_{\mathcal{R}\in\boldsymbol{\mathcal{S}}_B} D_{\mathrm{KL}}(\mathcal{P},\mathcal{R})$ が成り立つ．

4 実験

ルジャンドル分解の性能を人工データと実データを用いた実験で検証する．すべての実験は，Amazon Linux AMI 2017.09 を用いて，2.3 GHz Intel Xeon CPU E7-8880 v3 と 256 GB のメモリで実行した．ルジャンドル分解は `C++` で実装し，`icpc` 18.0.0 でコンパイルした．

実験を通して，分解基底 B を

$$B = B_1 \cup B_2, \quad \begin{cases} B_1 = \{(i,j,1) \mid i \le l \text{ or } j \le l\}, \\ B_2 = \{(i,1,k) \mid i \le l \text{ or } k \le l\}, \end{cases}$$

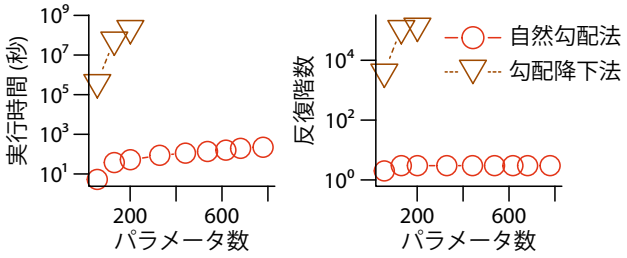

図2　人工データにおける実験結果.

とした．この基底の形を図 1に示す．基底 B の要素数 $|B|$ は，分解に用いるパラメータ数に対応する．実験では，l を用いることでパラメータ数を変化させ，分解の性能を検証した．

4.1　人工データでの実験結果

2 つの提案アルゴリズムである，勾配降下法（アルゴリズム 1）と自然勾配法（アルゴリズム 2）について，それらの効率性を人工データで検証した．サイズが $20\times20\times20$ の 3 階のテンソルを一様分布からランダムに生成し，パラメータ数 $|B|$ を変化させながら実行時間と反復回数を計測した．比較のために，アルゴリズム 2においては，外側のループ（行 3から 8）を 1 回の反復として計測し，学習率は $\varepsilon = 0.1$ とした．

結果を図 2に示す．この結果から，自然勾配法が勾配降下法よりも劇的に高速であることがわかる．パラメータ数が 400 あたりでは，自然勾配法は勾配降下法よりも 6 桁以上高速である．この違いは，反復回数の減少に起因する．自然勾配法では，すべての場合で，収束までの反復回数は 2, 3 回であったが，勾配降下法では同じ結果を得るために 10 万回以上の反復を必要とした．以下では，常に自然勾配法を用いる．

4.2　実データでの実験結果

次に，3 階テンソルの実データを用いて，ルジャンドル分解の性能を検証した．標準的なテンソル分解手法である，非負タッカー分解 [13] と非負 CANDECOMP/PARAFAC (CP) 分解 [24] と比較した．これらの手法は，`TensorLy` [15] の実装を用いて実行した．非負タッカー分解では，常にランク (l,l,l) のタッカー分解を用いて，非負 CP 分解ではランク l の分解を用いた．したがって，ランク (l,l,l) タッカー分解では $(I_1+I_2+I_3)l+l^3$ 個のパラメータがあり，ランク l CP 分解では $(I_1+I_2+I_3)l$ 個のパラメータがある．分解の性能を測るために，入力データと再構成したテンソル間の RMSE (Root Mean Squared Error) を用いた．

手書き数字の画像データである MNIST [17] を用いて実験をおこなった．各数字について，最初 500 個の画像を取り出し，サイズ $28\times28\times500$ の 3 階テンソルを 10 個作成した．結果を図 3に示す．“1” を除き，すべての数字でルジャンドル分解が最良の性能を示し，“1” においてもパラメータ数が 4,000 以下のときは最も性能が高かった．これは，ルジャンドル分解が行列のインデックス順序を有効に利用できているためと考えられる．

5　おわりに

本稿では，情報幾何を利用した非負テンソル分解の手法である**ルジャンドル分解**（Legendre decomposition）を提案した．ルジャンドル分解では，テンソルをルジャンドル変換を介して 2 つのパラメータ θ と η に変換し，そのパラメータ空間で最適化を実行する．また，ルジャンドル分解とボルツマンマシンの学習との関係を示した．実験によって，ルジャンドル分解は標準的なテンソル分解手法である非負タッカー分解と非負 CP 分解と比べて，より高精度でテンソルを分解できることを示した．本研究は，テンソル分解に対する情報幾何的アルゴリズムのより深い理論的解析につながる．

謝辞．本研究は，JSPS 科研費 26880013 (MS)，26120732 (HN)，15H05711 (KT)，及び JST CREST, JST ERATO, RIKEN PostK, NIMS MI2I, KAKENHI Nanostructure (KT) の助成を受けたものです．

参考文献

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.
[2] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
[3] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.
[4] S. Amari. *Information Geometry and Its Applications: Convex Function and Dually Flat Manifold*, pages 75–102. Springer, 2009.
[5] S. Amari. *Information Geometry and Its Applications*. Springer, 2016.
[6] C. F. Beckmann and S. M. Smith. Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, 25(1):294–311, 2005.
[7] Y. Censor and A. Lent. An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications*, 34(3):321–353, 1981.
[8] J. Chen, S. Cheng, H. Xie, L. Wang, and T. Xiang. Equivalence of restricted Boltzmann machines and tensor network states. *Physical Review B*, 97:085104, Feb 2018.
[9] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163, 2015.

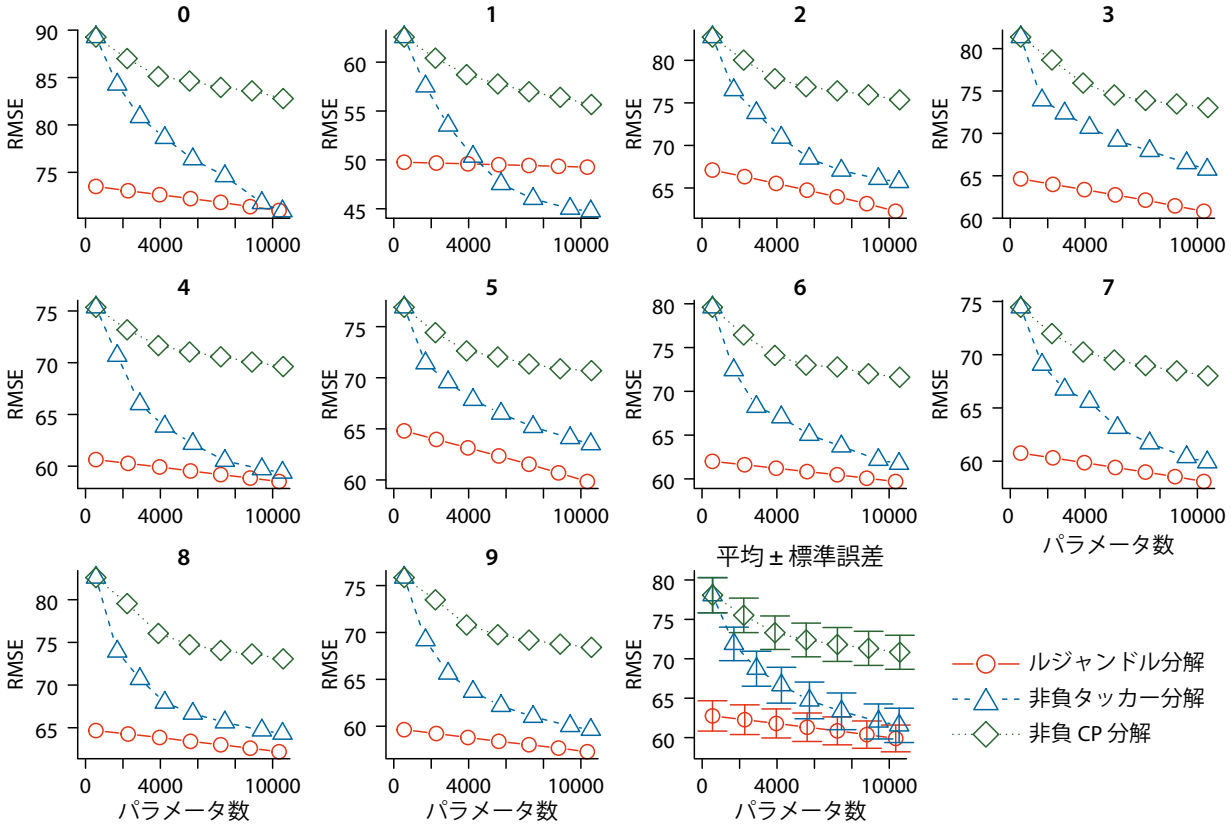

図3　MNIST データにおける，0 から 9 までの各数字についての結果と，全体にわたっての平均 ± 標準誤差．

[10] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2 edition, 2002.

[11] G. Gierz, K. H. Hofmann, K. Keimel, J. D. Lawson, M. Mislove, and D. S. Scott. *Continuous Lattices and Domains*. Cambridge University Press, 2003.

[12] R. A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. Technical report, UCLA Working Papers in Phonetics, 1970.

[13] Y. D. Kim and S. Choi. Nonnegative Tucker decomposition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[14] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[15] J. Kossaifi, Y. Panagakis, and M. Pantic. TensorLy: Tensor learning in Python. *arXiv:1610.09555*, 2016.

[16] J. P. S. Kung, G.-C. Rota, and C. H. Yan. *Combinatorics: The Rota Way*. Cambridge University Press, 2009.

[17] Y. LeCun, C. Cortes, and C. J. C. Burges. The MNIST database of handwritten digits, 1998.

[18] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[19] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information processing Systems 13*, pages 556–562, 2001.

[20] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

[21] H. Nakahara and S. Amari. Information-geometric measure for neural spikes. *Neural Computation*, 14(10):2269–2316, 2002.

[22] G.-C. Rota. On the foundations of combinatorial theory I: Theory of Möbius functions. *Z. Wahrseheinlichkeitstheorie*, 2:340–368, 1964.

[23] T. J. Sejnowski. Higher-order Boltzmann machines. In *AIP Conference Proceedings*, volume 151, pages 398–403, 1986.

[24] A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 792–799, 2005.

[25] M. Sugiyama, H. Nakahara, and K. Tsuda. Tensor balancing on statistical manifold. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3270–3279, 2017.

[26] P. Symeonidis. Matrix and tensor decomposition in recommender systems. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 429–430, 2016.

[27] R. Tomioka and T. Suzuki. Convex tensor decomposition via structured schatten norm regularization. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1331–1339. 2013.

[28] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[29] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: TensorFaces. In *Proceedings of The 7th European Conference on Computer Vision (ECCV)*, volume 2350 of *LNCS*, pages 447–460, 2002.

[30] M. A. O. Vasilescu and D. Terzopoulos. Multilinear (tensor) image synthesis, analysis, and recognition [exploratory dsp]. *IEEE Signal Processing Magazine*, 24(6):118–123, 2007.

[31] K. Y. Yılmaz, A. T. Cemgil, and U. Simsekli. Generalised coupled tensor factorisation. In *Advances in Neural Information Processing Systems 24*, pages 2151–2159. 2011.

[32] Y. K. Yılmaz and A. T. Cemgil. Algorithms for probabilistic latent tensor factorization. *Signal Processing*, 92(8):1853–1863, 2012.