

順序構造上の情報幾何的な統計解析

Information Geometric Statistical Analysis with Partial Orders

杉山 磨人^{1,2*} 中原 裕之³ 津田 宏治⁴
Mahito Sugiyama^{1,2} Hiroyuki Nakahara³ Koji Tsuda⁴

¹ 大阪大学 産業科学研究所

¹ ISIR, Osaka University

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 formulate information geometry for a partially ordered set, a hierarchical set of variables, and introduce the orthogonal decomposition of information theoretic quantities. The connection between information geometry and order theory leads to efficient decomposition algorithms. This generalization of Amari's seminal work on hierarchical decomposition of probability distributions on event combinations enables us to analyze high-order statistical interactions.

1 はじめに

事象 e_1, e_2, \dots, e_n があるとき, これらの事象の組合せ全てからなる集合は半順序集合をなし, 階層構造を形成する (図 1a). 甘利 [3] は, 階層構造をもつ事象のべき集合に対して, 確率分布の直交分解を導入した. この研究は, ニューロンの発火パターン [12, 13] や遺伝子間の相互作用 [14], 文章中の単語間の関連 [11] などの様々な応用において, 事象間の高次相関を解析するための理論的基礎を与えている. しかし, 多くの応用分野において, いくつかの事象の組合せは決して起きることはなく, 階層構造は不完全である (図 1b) 例え, e_1 が男性, e_2 が卵巣がんに対応する事象とすると, e_1 と e_2 の組合せは決して起きない. このような不完全な階層構造は, データの不足によっても引き起こされる [9].

本稿では, 半順序集合 (partially ordered set, poset) 上に, 情報幾何で知られている双対座標系を導入する. これによって, 不完全な階層構造上でも, KL ダイバージェンスを分解するための効率的なアルゴリズムが構築でき

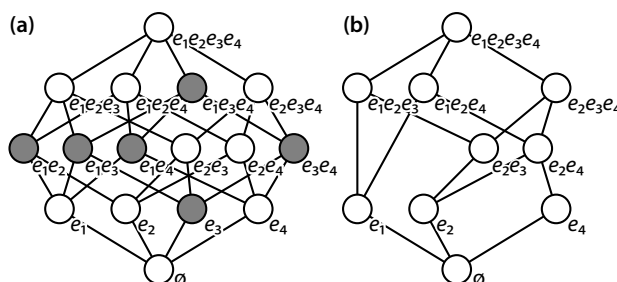


図1 4つの事象 e_1, e_2, e_3, e_4 の組合せからなる階層構造. (a) 完全な階層構造. (b) (a) で灰色で示した組合せを取り除いたときの不完全な階層構造.

る. 提案手法によって, 事象の各組合せを分離して, それらの統計的有意性 [12] を測ることができる. これによって, これまで明らかでなかった, 順序理論と情報幾何のあいだの理論的なつながりを示す.

本稿は, 以下のように構成されている. 2節では, 半順序集合上に双対平坦多様体を導入する. 2.1節において, 本稿で導入する双対平坦構造が, 指数型分布族が持つ双対平坦構造と一致することを示し (数式 (3), (5)), 2.2節では, 半順序集合上での情報量分解アルゴリズムを示す (アルゴリズム 1, 2と定理 1). 3節では, この提案アルゴリズムが, 任意の高次相互作用に関してその情報量を分

* 連絡先: 大阪大学産業科学研究所
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
E-mail: mahito@ar.sanken.osaka-u.ac.jp

離して取り出すことができることを示す。最後に、4節で本稿の内容をまとめて、結論を述べる。

本稿の内容は、文献 [17] に基づいている。

2 半順序集合上の双対平坦多様体

離散確率変数 X と有限集合 S の要素 $s \in S$ に対して、 $p(x) = \Pr(X = x)$ を確率質量関数とする。情報幾何 [3, 4, 5] において、各分布は関数 $p: S \rightarrow \mathbb{R}$ として扱われ、確率分布全体からなる集合は $(|S| - 1)$ 次元の多様体

$$\mathcal{S} = \{P \mid p(x) > 0 \text{ for all } x \in S, \sum_{x \in S} p(x) = 1\}$$

をなす。分布 P の確率 p は \mathcal{S} の座標系をなし、 p 座標系と呼ばれる。この座標系に加えて、情報幾何において、 θ 座標系と η 座標系と呼ばれる \mathcal{S} の座標系が知られており、これらは互いに直交しているという特徴をもつ。特に、 θ と η をともに用いた混合座標系を用いることで、KL ダイバージェンスが分解できる。

本稿では、集合 S が半順序集合 (poset) の場合を解析する。ここで、半順序 \leq とは、以下の3つを満たす関係であり、数学や計算機科学における基本的な構造である [8, 10]。任意の $x, y, z \in S$ に対して、(1) $x \leq x$ (反射律)、(2) $(x \leq y \text{ かつ } y \leq x) \Rightarrow x = y$ (反対称律)、(3) $(x \leq y \text{ かつ } y \leq z) \Rightarrow x \leq z$ (推移律)。以下では、 S は常に有限かつ最小元 $\perp \in S$ の存在を仮定し、任意の $x \in S$ に対して、 $\perp \leq x$ である。また、集合 $S \setminus \{\perp\}$ を S^+ と書く。

部分集合 $I \subseteq S$ に対して、下方集合 $\downarrow I = \{x \in S \mid x \leq s, \exists s \in I\}$ 、上方集合 $\uparrow I = \{x \in S \mid x \geq s, \exists s \in I\}$ と表記し、各 $x \in S$ に対して $\downarrow x = \downarrow\{x\}$ 、 $\uparrow x = \uparrow\{x\}$ とする。順序理論では、 $\downarrow x$ を x の主イデアル (principal ideal)、 $\uparrow x$ を x の主フィルター (principal filter) と呼び、これらは半順序集合における基礎的な対象である [8, 10]。

2.1 θ 座標系と η 座標系

まず、多様体 \mathcal{S} の θ 座標系 (θ -coordinate system) を導入する。これは、写像 $\theta: \mathcal{S} \rightarrow \mathbb{R}$ として扱われる。指数型分布においては、 θ は自然パラメータであり、 n 次元のベクトル $\theta = (\theta^1, \theta^2, \dots, \theta^n)$ として扱われ、各分布は、 n 個の関数 F_1, \dots, F_n と正規化定数 $\psi(\theta)$ を用いて

$$p(x; \theta) = \exp\left(\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)\right) \quad (1)$$

という形式で表される [5]。これは、以下のように書き換えることができる。

$$p(x; \theta) = \exp\left(\sum_{s \in S^+} \theta(s) F_s(x) - \psi(\theta)\right). \quad (2)$$

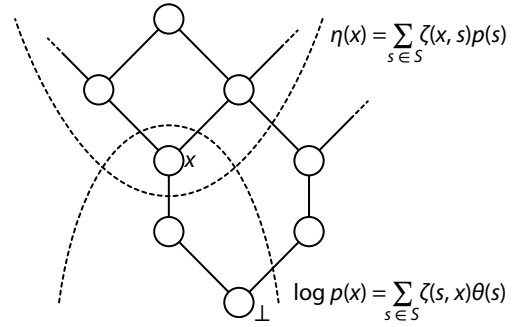


図2 半順序集合における $p(x)$, $\theta(x)$, $\eta(x)$ 。

ここで、 $n = |S^+|$ であり、 $\theta(s)$ と F_s をそれぞれ数式(1)の $\theta^{\omega(s)}$ と $F_{\omega(s)}$ に 1 対 1 に対応づける写像 $\omega: S^+ \rightarrow \{1, 2, \dots, n\}$ が必ず存在する。

与えられた半順序集合 S に対して、 $F_s(x) = \zeta(s, x)$ と定義する。ここで、 $\zeta(s, x)$ はゼータ関数 (zeta function) と呼ばれ、以下のように定義される。

$$\zeta(s, x) = \begin{cases} 1 & \text{if } s \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

また、 $\psi(\theta) = -\log p(\perp)$ とする。すると、数式(2)から、 S 中の x より小さい要素 $s \leq x$ の $\theta(s)$ の和として、 $\log p(x)$ が以下のように展開できる。

$$\log p(x) = \sum_{s \in S} \zeta(s, x) \theta(s) = \sum_{s \leq x} \theta(s). \quad (3)$$

これは、よく知られている n 次元の 2 値ベクトル $\mathbf{x} = (x^1, \dots, x^n) \in \{0, 1\}^n$ に対する対数線形モデル

$$\log p(\mathbf{x}) = \sum_i \theta^i x^i + \sum_{i < j} \theta^{ij} x^i x^j + \sum_{i < j < k} \theta^{ijk} x^i x^j x^k + \dots + \theta^{1 \dots n} x^1 \dots x^n - \psi$$

の一般化となっている。

ここで、Rota [16] が示したメビウス反転公式 (Möbius inversion formula) を適用することができる。すなわち、

$$\theta(x) = \sum_{s \in S} \mu(s, x) \log p(x) \quad (4)$$

が必ず成り立つ。ここで、 μ は $\mu = \zeta^{-1}$ として定義されるメビウス関数 (Möbius function) で、

$$\mu(x, y) = \begin{cases} 1 & \text{if } x = y, \\ -\sum_{x \leq s < y} \mu(x, s) & \text{if } x < y, \\ 0 & \text{otherwise.} \end{cases}$$

と定義される。これによって、 θ を p から計算できる。

情報幾何 [3] において、指数型分布族の自然パラメータ θ は、 e 平坦な多様体 \mathcal{S} の e アフィン座標として知られている。すなわち、数式(4)で導入した θ の定式化は、 e

アフィン座標であることを意味する。また、 \mathcal{S} における e 座標と双対な座標系として、 m アフィン座標系、すなわち η 座標系 (η -coordinate system) $\eta: \mathcal{S} \rightarrow \mathbb{R}$ が知られており、各 $s \in \mathcal{S}$ に対して $F_s(x) = \zeta(s, x)$ の期待値として定義される。すなわち、 η は以下のように与えられる。

$$\eta(x) = \mathbb{E}[\zeta(x, s)] = \sum_{s \in \mathcal{S}} \zeta(x, s) p(s) = \sum_{s \geq x} p(s). \quad (5)$$

また、メビウス反転公式を用いることで、

$$p(x) = \sum_{s \in \mathcal{S}} \mu(x, s) \eta(s)$$

となる。3つの座標系 p, θ, η の関係を図2に示す。

2つの座標系 θ と η は、ルジャンドル変換によって繋がっており、重要な性質として θ と η の直交性がある。すなわち、任意の $s, s' \in S^+$ に対して

$$\mathbb{E} \left[\frac{\partial}{\partial \theta(s)} \log p(x) \frac{\partial}{\partial \eta(s')} \log p(x) \right] = \delta_{ss'} \quad (6)$$

が成り立つ [3]。ここで、 δ はクロネッカーのデルタで、 $s = s'$ ならば $\delta_{ss'} = 1$ 、そうでなければ $\delta_{ss'} = 0$ である。この関係は、メビウス反転公式から導くことができ、 θ と η の混合座標系を構築する際に欠かせない性質である。

2.2 混合座標系を用いた情報量分解

半順序集合 \mathcal{S} 上の分布を解析する際に鍵となる、 θ と η の混合座標系を導入する。部分集合 $I \subseteq S^+$ に関する混合座標系 $\xi_I: S^+ \rightarrow \mathbb{R}$ とは、 \mathcal{S} の座標系であり、

$$\xi_I(x) := \begin{cases} \eta(x) & \text{if } x \in S^+ \setminus I, \\ \theta(x) & \text{if } x \in I. \end{cases}$$

と定義される。この座標系を用いることで、 P と Q という2つの分布を混ぜることができる。部分集合 $I \subseteq S^+$ に関する分布の組 (P, Q) の**混合分布** (mixed distribution) とは、以下のように定義される分布 $R \in \mathcal{S}$ である。

$$\begin{cases} \eta_R(x) = \eta_P(x) & \text{if } x \in S^+ \setminus I, \\ \theta_R(x) = \theta_Q(x) & \text{if } x \in I, \end{cases} \quad r(\perp) = 1 - \sum_{x \in S^+} r(x).$$

以下では、 $P \in \mathcal{S}$ に対応する θ 座標と η 座標をそれぞれ θ_P, η_P と書く、数式(6)で示した θ と η の直交性により、この混合分布は常に唯一に定まり、well-defined である。

2つの確率分布 P と Q の間の KL ダイバージェンス

$$D_{\text{KL}}[P, Q] = \sum_{x \in \mathcal{S}} p(x) \log \frac{p(x)}{q(x)} \quad (7)$$

を、混合分布 R を使って分解する。

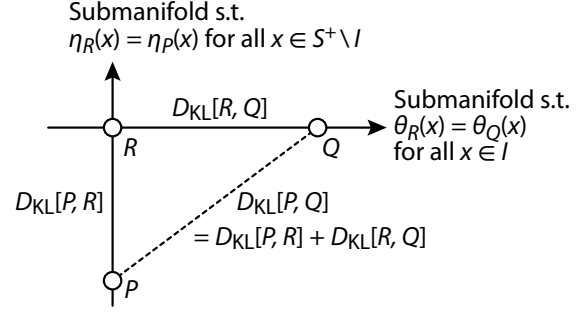


図3 ピタゴラスの定理 (定理1)。

定理1 (ピタゴラスの定理). 2つの確率分布 $P, Q \in \mathcal{S}$ と部分集合 $I \subseteq S^+$ が与えられたとき、 I に関する (P, Q) の混合分布 R と (Q, P) の混合分布 R' に対して、

$$D_{\text{KL}}[P, Q] = D_{\text{KL}}[P, R] + D_{\text{KL}}[R, Q], \quad (8)$$

$$D_{\text{KL}}[Q, P] = D_{\text{KL}}[Q, R'] + D_{\text{KL}}[R', P]. \quad (9)$$

が成り立つ。

証明. メビウス反転公式を用いることで、

$$\begin{aligned} & D_{\text{KL}}[P, R] + D_{\text{KL}}[R, Q] - D_{\text{KL}}[P, Q] \\ &= \sum_{s \in \mathcal{S}} (r(s) - p(s)) (\log r(s) - \log q(s)) \\ &= \sum_{s \in \mathcal{S}} (\eta_R(s) - \eta_P(s)) (\theta_R(s) - \theta_Q(s)). \end{aligned}$$

が得られる。混合分布 R は、任意の $x \in \mathcal{S}$ に対して $\theta_R(x) = \theta_Q(x)$ または $\eta_R(x) = \eta_P(x)$ なので、 $D_{\text{KL}}[P, Q] = D_{\text{KL}}[P, R] + D_{\text{KL}}[R, Q]$ が成り立つ。□

さらに、 \mathcal{S} の部分集合の集まり $\{I_0, I_1, \dots, I_k\}$ が $\emptyset = I_0 \subseteq I_1 \subseteq \dots \subseteq I_k = S^+$ を満たすとき、

$$D_{\text{KL}}[P, Q] = \sum_{i=1}^k D_{\text{KL}}[R_{i-1}, R_i] \quad (10)$$

が成り立つ。ここで、各 $i \in \{0, 1, \dots, k\}$ に対して、 R_i は I_i に関する (P, Q) の混合分布であり、 $R_0 = P$ かつ $R_k = Q$ である。

ここで、 P_0 を一様分布とし、 $p_0(x) = 1/|\mathcal{S}|$ とする。これは、任意の $x \in S^+$ に対して $\theta_{P_0}(x) = 0$ となるため、 θ 座標系の原点に対応する。確率分布 P のエントロピー $H(X)$ は

$$H(X) = - \sum_{x \in \mathcal{S}} p(x) \log p(x) = \log |\mathcal{S}| - D_{\text{KL}}[P, P_0]$$

となり、 $\log |\mathcal{S}|$ は定数なので、定理1で示した KL ダイバージェンスの分解によって、エントロピーが

$$H(X) = \log |\mathcal{S}| - (D_{\text{KL}}[P, R] + D_{\text{KL}}[R, P_p]),$$

Algorithm 1: 単集合 $I = \{x^*\}$ に関する (P, Q) の混合分布 R の計算

```

1 COMPUTEMIXEDSINGLE( $x^*$ )
2    $x_1, x_2, \dots, x_m = x^*$  は  $\downarrow x^*$  の要素で, 順序  $\leq$  に従っ
   て小さい順に並んでいるとする
3   2分法などを用いて,
   COMPUTETHETASINGLE( $x^*, r(x^*)$ ) を用いて
    $\theta_R(x^*) = \theta_Q(x^*)$  を満たす  $r(x^*)$  を見つける
4   return  $R$ 
5 COMPUTETHETASINGLE( $x^*, r(x^*)$ )
6   foreach  $s \in \downarrow x^*$  do
7      $f(s) \leftarrow 0$ ; // フラグを初期化
8   for  $i \leftarrow m-1$  to 1 do
9      $r(x_i) \leftarrow \text{COMPUTEPSINGLE}(x_i)$ 
10  for  $i \leftarrow 1$  to  $m$  do
11    Compute  $\theta_R(x_i)$ 
12  return  $\theta_R(x^*)$ 
13 COMPUTEPSINGLE( $x$ )
14   $r(x) \leftarrow p(x)$ 
15  foreach  $s > x$  do //  $x < s$  iff  $x < s, x \leq y < s \Rightarrow x = y$ 
16     $r(x) \leftarrow r(x) + \text{AGGREGATEDIFF}(s, \omega(x))$ 
    //  $\omega(x)$ : index of  $x$ 
17  return  $r(x)$ 
18 AGGREGATEDIFF( $x, i$ )
19  if  $f(x) = i$  or  $x \not\leq x^*$  then
20     $p_{\text{diff}} \leftarrow 0$ 
21  else
22     $f(x) \leftarrow i$ 
23     $p_{\text{diff}} \leftarrow p(x) - r(x)$ 
24    foreach  $s \geq x$  do
25       $p_{\text{diff}} \leftarrow p_{\text{diff}} + \text{AGGREGATEDIFF}(s, i)$ 
26  return  $p_{\text{diff}}$ 

```

として分解できる。ここで、 R は $I \subseteq S^+$ に関する (P, P_0) の混合分布である。エントロピーは、 I の要素がもつ貢献分 $D_{\text{KL}}(P, R)$ と、それ以外 $D_{\text{KL}}(R, P_0)$ に分解されている。したがって、各 $I \subseteq S$ に対して、 P_I を I に関する (P, P_0) の混合分布とすると、その情報利得を KL ダイバージェンス $D_{\text{KL}}(P, P_I)$ として取り出すことができる。

2.3 混合分布の計算

分布 P と Q の $I \subseteq S^+$ に関する混合分布 R の計算方法について述べる*1。まず、 I が単集合 $I = \{x^*\}$ の場合において R を計算するアルゴリズムを説明する。全ての $x \neq x^*$ に対して $\eta_R(x) = \eta_P(x)$ なので、任意の $x \not\leq x^*$

*1 実装は以下で入手可能：<https://github.com/mahito-sugiyama/information-decomposition>

Algorithm 2: $I \subseteq S^+$ に関する (P, Q) の混合分布 R

```

1 COMPUTEMIXEDMULTI( $I$ )
2   repeat
3     for  $x^* \in I$  do
4       COMPUTEMIXEDSINGLE( $x^*$ ) // Algorithm 1
5   until convergence of  $r$ ;

```

に対して $r(x) = p(x)$ である。したがって、 $x \leq x^*$ を満たす $r(x)$ の計算についてのみ説明する。

まず、 $r(x^*)$ を固定し、 $I_{\geq x} = \{s \in \downarrow x^* \mid s \geq x\}$ かつ $I_{> x} = I_{\geq x} \setminus \{x\}$ とする。各 $x \in \downarrow x^*$ で $x \neq x^*$ を満たすものに対して、条件 $\eta_P(x) = \eta_R(x)$ より $\sum_{s \in I_{\geq x}} p(s) = \sum_{s \in I_{\geq x}} r(s)$ が成り立つ。したがって、 $r(x)$ は $r(x) = p(x) + \sum_{s \in I_{> x}} (p(s) - r(s))$ として得られる。すなわち、もし $\downarrow x^*$ の要素を半順序 \leq に関して小さい順に $\perp = x_0, x_1, \dots, x_m = x^*$ と並べると、順に $r(x_m), r(x_{m-1}), \dots, r(x_0)$ と計算することができる。アルゴリズム 1 の関数 $\text{COMPUTEPSINGLE}(x)$ は、この計算を実行する。さらに、 $r(x^*)$ が固定されているとき、 $r(x)$ を計算した後に $\theta_R(x_0), \theta_R(x_1), \dots, \theta_R(x_m) = \theta_R(x^*)$ が計算できるため、 $r(x^*)$ を入力すると、 $\theta_R(x^*)$ が計算できることを意味する。このプロセスは、アルゴリズム 1 の関数 $\text{COMPUTEETHETASINGLE}(x^*, r(x^*))$ で実行している。この関数は連続なので、2分法などの数値最適化を用いることで、効率的に $\theta_R(x^*) = \theta_Q(x^*)$ を満たす $r(x^*)$ を計算することができる。全体の計算量は、 $h(x^*)$ を反復数とすると、 $O(h(x^*)|\downarrow x^*|^2) \leq O(h(x^*)|S|^2)$ となる。

次に、一般の場合を考える。部分集合 $I = \{x_1^*, x_2^*, \dots, x_l^*\}$ とする。ここでも、 R を解析的に計算するのは困難だが、各 x_i^* に対して $\theta_R(x_j^*)$ ($j \neq i$) を固定して $\theta_R(x_i^*)$ を計算する、というプロセスを繰り返すことで、数値的に求めることができる。全体のアルゴリズムを、アルゴリズム 2 に示す。

補題 1. アルゴリズム 2 は常に (P, Q) の $I \subseteq S^+$ に関する混合分布 $R \in \mathcal{S}$ に収束する。

証明. 分布の列 R_1, R_2, \dots において、各 R_i はアルゴリズム 2 の関数 $\text{COMPUTEETHETASINGLE}$ の i 回目の反復で得られた分布とする。定理 1 から、任意の i に対して $D_{\text{KL}}[R_i, R] = D_{\text{KL}}[R_i, R_{i+1}] + D_{\text{KL}}[R_{i+1}, R]$ が成り立つので、 $D_{\text{KL}}[R_i, R] \geq D_{\text{KL}}[R_{i+1}, R]$ が成り立ち、等号は $D_{\text{KL}}[R_i, R_{i+1}] = 0$ のときのみ成り立つ。ここで、任意の $\varepsilon > 0$ に対して、もし $D_{\text{KL}}[R_i, R] > \varepsilon$ ならば $D_{\text{KL}}[R_i, R_j] > 0$ を満たす R_j ($j > i$) が必ず存在するので、アルゴリズム 2 は混合分布 R に収束する。□

各 $x^* \in I$ について, $r(x^*)$ の計算量が $O(h(x^*)\lfloor x^* \rfloor^2)$ なので, 全体の計算量は $O(h \sum_{x^* \in I} h(x^*)\lfloor x^* \rfloor^2) \leq O(h|S|^3 \sum_{x^* \in I} h(x^*))$ である. ここで, h は r が取束するまでに要する反復数である.

2.4 θ の統計的有意性

半順序集合 S 上の分布 P に対して KL ダイバージェンスの分解を用いることで, 尤度比検定, より具体的には G 検定によって, パラメータ θ の統計的有意性 (statistical significance) を測ることができる. 各 $\theta(x)$ は, P の対数線形モデルにおける係数であり, 頻度 η に直交しているので, 分布 P における要素 x の貢献度を表している.

部分集合 $I \subseteq S$ の検定では, 帰無仮説と対立仮説は

$$H_0: \theta(x) = 0, \forall x \in I, \quad H_1: \theta(x) \neq 0, \forall x \in I$$

となる [12, 14]. これは, 数式(3)で導入した対数線形モデル $\log p(x) = \sum_{s \in S} \zeta(s, x)\theta(s)$ において, I 中の全ての要素 $x \in I$ の貢献を $\theta(x) = 0$ とすることでノックダウンしていることを意味する. このとき統計量 λ は,

$$\lambda = 2N \sum_{s \in S} \left(p(s) \log \left(\frac{p(s)}{r(s)} \right) \right) = 2ND_{\text{KL}}[P, R]$$

となることが知られている. ここで, N はサンプルサイズ, R は I に関する (P, P_0) の混合分布であり, 帰無分布に対応する. したがって, λ はアルゴリズム 1 と 2 によって計算できる. この統計量 λ は自由度 $|S| - 1$ の χ^2 分布にしたがうので, p 値が計算できる.

3 データからの学習

3.1 高次相互作用の直交分解

代表的な応用として, 事象の組合せの直交分解について紹介する. はじめに述べたように, n 個の事象 e_1, \dots, e_n を考える. 各部分集合 $x \subseteq [n] = \{1, \dots, n\}$ に対して, $p(x)$ を組合せ $\bigcap_{i \in x} e_i$ の確率とする. ここでの目的は, $\log p(x)$ を部分集合 $s \subseteq x$ の係数の和へと分解することであり, これは $\log p(x) = \sum_{s \in S} \zeta(s, x)\theta(s)$ で与えられる θ 座標系である. 順序 \leq は, $x \subseteq s$ ならば $x \leq s$ として, 包含関係によって定まる. 係数 $\theta(x)$ は, 頻度, すなわち η 座標系 $\eta(x) = \sum_{s \in S} \zeta(x, s)p(s)$ と直交しているため, それぞれの相互作用 $\bigcap_{j \in x} e_j$ の貢献度を表している.

サンプルサイズ N のデータ t_1, t_2, \dots, t_N を仮定し, 各データ点 t_i を同時に起こった事象の集合とする. 各確率 $p(x)$ は, $\hat{p}(x) = |\{i \in [N] \mid t_i = x\}| / N$ として推定でき

表1 例1のデータ.

t_1	e_2
t_2	e_2
t_3	e_4, e_5
t_4	e_1, e_2, e_4, e_5
t_5	e_1, e_2, e_4, e_5
t_6	e_3
t_7	e_1, e_2, e_4, e_5
t_8	e_4, e_5
t_9	e_1, e_2, e_4, e_5
t_{10}	e_2

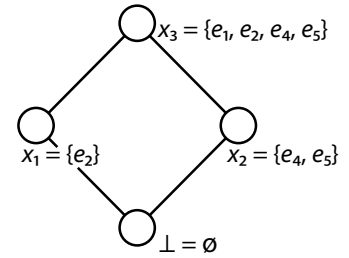


図4 表1のデータから生成される半順序集合.

る. ここで, 効率的に \hat{p} を推定して $\theta_{\hat{p}}$ や $\eta_{\hat{p}}$ を計算するため, 組合せ全体からなる集合 $2^{[n]}$ を枝刈りして, データに頻出していない組合せを取り除く. 頻度のしきい値 $\sigma \in \mathbb{R} (0 \leq \sigma \leq 1)$ を用いて, $S^+ = \{x \subseteq [n] \mid \hat{p}(x) \geq \sigma\}$ と定め, $\hat{p}(\perp) = 1 - \sum_{x \in S^+} \hat{p}(x)$ とする. したがって, 多様体 \mathcal{S} の次元は 2^n から N へ落ちる. べき集合 $2^{[n]}$ の任意の部分集合は半順序集合なので, 2節で提案した分解アルゴリズムを用いて $\theta_{\hat{p}}$ や $\eta_{\hat{p}}$, それらの混合分布が計算できる. ここで, 各データ点 t_i はデータベースのトランザクションに対応しており, $\eta_{\hat{p}}(I)$ はパターンマイニング [1] で用いられている I のサポート (support) そのものである.

例1. 表1に示したデータが与えられたと仮定する. しきい値 $\sigma = 0.2$ とする. このとき, 図4に示すように半順序集合 $S = \{\perp, x_1, x_2, x_3\}$ となり, $\perp = \emptyset$, $x_1 = \{e_2\}$, $x_2 = \{e_4, e_5\}$, $x_3 = \{e_1, e_2, e_4, e_5\}$ である. 推定される確率は, $\hat{p}(\perp) = 0.1$, $\hat{p}(x_1) = 0.3$, $\hat{p}(x_2) = 0.2$, $\hat{p}(x_3) = 0.4$ となる. したがって, $\theta_{\hat{p}}$ は $\theta_{\hat{p}}(\perp) = -2.303$, $\theta_{\hat{p}}(x_1) = 1.099$, $\theta_{\hat{p}}(x_2) = 0.693$, $\theta_{\hat{p}}(x_3) = -0.405$ と得られる. ここで, \hat{R}_x を (\hat{P}, P_0) の $\{x\} \in S$ に関する混合分布とする. 高次相互作用に対する KL ダイバージェンスは $D_{\text{KL}}[\hat{P}, \hat{R}_{x_1}] = 0.0523$, $D_{\text{KL}}[\hat{P}, \hat{R}_{x_2}] = 0.0170$, $D_{\text{KL}}[\hat{P}, \hat{R}_{x_3}] = 0.0040$ と求まる. これらの p 値を実際に計算すると, どれも 0.99 より大きくなってしまいが, これはサンプルサイズが $N = 10$ と小さいためであり, 例えば $N = 300$ だと x_1 の p 値は 0.001 となり, 有意水準 $\alpha = 0.05$ のもとで有意である.

3.2 非負整数ベクトル

同様の戦略は, n 次元の非負整数のベクトル $\mathbb{Z}_{\geq 0}^n$ からなる半順序集合 S にも適用できる. 集合 S を $\mathbb{Z}_{\geq 0}^n$

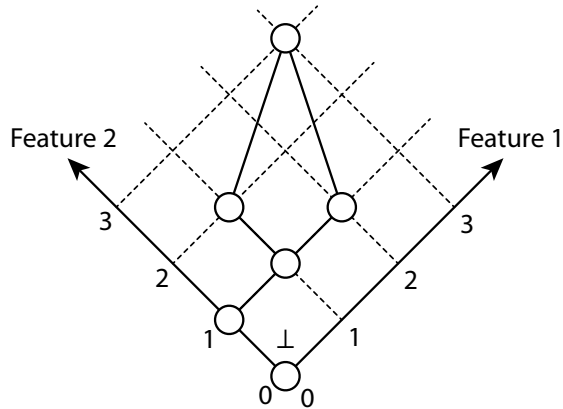


図5 例 2 の非負整数ベクトルから形成される半順序集合.

の部分集合とし, 各ベクトルの組 $x, y \in \mathbb{Z}_{\geq 0}^n$ において $x = (x^1, x^2, \dots, x^n)$ かつ $y = (y^1, y^2, \dots, y^n)$ とする. このとき, 半順序 \leq を任意の $i \in [n]$ に対して $x^i \leq y^i$ なら $x \leq y$ と定める. すると, $0 = (0, 0, \dots, 0) \in \mathbb{Z}_{\geq 0}^n$ は最小元 \perp となり, 任意の部分集合 $S \subset \mathbb{Z}_{\geq 0}^n$ は半順序集合となる.

データとして, n 次元ベクトル $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ が N 個与えられたとする. 事象の組合せの直交分解の場合と同様に, 半順序集合 S^+ はデータからしきい値 $\sigma \in \mathbb{R}$ を用いて $S^+ = \{\mathbf{x} \in \mathbb{Z}_{\geq 0}^n \mid \hat{p}(\mathbf{x}) \geq \sigma\}$ として求まる. したがって, S と経験分布 \hat{P} を用いて, 情報量分解が適用できる.

例 2. データ $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{25} \in \mathbb{Z}_{\geq 0}^2$ を, $\mathbf{x}_1, \dots, \mathbf{x}_3 = (0, 1)$, $\mathbf{x}_4 = (1, 0)$, $\mathbf{x}_5, \dots, \mathbf{x}_8 = (1, 1)$, $\mathbf{x}_9, \dots, \mathbf{x}_{11} = (1, 2)$, $\mathbf{x}_{12}, \dots, \mathbf{x}_{21} = (2, 1)$, $\mathbf{x}_{22}, \dots, \mathbf{x}_{25} = (3, 3)$ とする. このとき, もし $\sigma = 2/25$ ならば $S = \{(0, 0), (0, 1), (1, 1), (1, 2), (2, 1), (3, 3)\}$ となる (図 5).

4 おわりに

本稿では, 情報幾何と順序理論の間の理論的つながりを示した. 具体的には, 事象間に半順序関係をもつ確率分布が, 必ず指数型分布族の多様体が持つ双対平坦構造と同一の構造を持つことを示した. さらに, 既存研究 [6, 7, 15, 19] とは異なり, 任意の半順序集合に対して適用可能な, 効率的な情報量分解アルゴリズムを示した.

代表的な応用例として, 事象の高次相互作用の直交分解を示し, 半順序集合が直接データから効率的に構築できることを示した. これによって, n 変数の場合の多様体の次元は, これまでのアプローチでは必ず 2^n であったが, 必ずサンプルサイズ N 以下となる. したがって,

ゲノムワイド関連解析における SNP データ [18] や, 神経科学におけるニューロン発火パターンのデータ [2] など, 変数の数が数千から数百万にもなりうる高次元データに対して, 直交分解を適用することができる.

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] C. C. Aggarwal and J. Han, editors. *Frequent Pattern Mining*. Springer, 2014.
- [2] A. Alivisatos, M. Chun, G. Church, R. Greenspan, M. Roukes, and R. Yuste. The brain activity map project and the challenge of functional connectomics. *Neuron*, 74:970–974, 2012.
- [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*. Springer, 2016.
- [5] S. Amari and H. Nagaoka. *Methods of information geometry*. American Mathematical Society, 2007.
- [6] N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost. Shared information—new insights and problems in decomposing information in complex systems. In *Proceedings of the European Conference on Complex Systems 2012*, pages 251–269. Springer, 2013.
- [7] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.
- [8] B. A. Davey and H. A. Priestley. *Introduction to lattices and order*. Cambridge University Press, 2 edition, 2002.
- [9] E. Ganmor, R. Segev, and E. Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences*, 108(23):9679–9684, 2011.
- [10] G. Gierz, K. H. Hofmann, K. Keimel, J. D. Lawson, M. Mislove, and D. S. Scott. *Continuous Lattices and Domains*. Cambridge University Press, 2003.
- [11] Y. Hou, X. Zhao, D. Song, and W. Li. Mining pure high-order word associations via information geometry for information retrieval. *ACM Transactions on Information Systems*, 31(3):12:1–12:32, 2013.
- [12] H. Nakahara and S. Amari. Information-geometric measure for neural spikes. *Neural Computation*, 14(10):2269–2316, 2002.
- [13] H. Nakahara, S. Amari, and B. J. Richmond. A comparison of descriptive models of a single spike train by information-geometric measure. *Neural computation*, 18(3):545–568, 2006.
- [14] H. Nakahara, S. Nishimura, M. Inoue, G. Hori, and S. Amari. Gene interaction in DNA microarray data is decomposed by information geometric measure. *Bioinformatics*, 19(9):1124–1131, 2003.
- [15] E. Olbrich, N. Bertschinger, and J. Rauh. Information decomposition and synergy. *Entropy*, 17(5):3501–3517, 2015.
- [16] G.-C. Rota. On the foundations of combinatorial theory I: Theory of Möbius functions. *Z. Wahrscheinlichkeitstheorie*, 2:340–368, 1964.
- [17] M. Sugiyama, H. Nakahara, and K. Tsuda. Information decomposition on structured space. In *2016 IEEE International Symposium on Information Theory*, pages 575–579, July 2016.
- [18] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.
- [19] P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *arXiv:1004.2515*, 2010.