数理解析研究所講究録
第2018巻 2017年 70-78

斉次ダイバージェンスとその応用*

Homogeneous Divergences and Applications

金森 敬文†
TAKAFUMI KANAMORI
名古屋大学
NAGOYA UNIVERSITY

竹之内 高志‡
TAKASHI TAKENOUCHI
はこだて未来大学
FUTURE UNIVERSITY HAKODATE

概要

離散集合上で定義される確率分布を推定するための方法を提案する．標本空間上の総和を実際に計算することは困難であり，それを回避するための工夫が必要になる．本稿では，(a) 斉次性を満たすダイバージェンスと (b) 非正規化モデルの局所化を組み合わせることで，計算量が少なく，また統計的有効性を満たす推定量を導出する．

1 問題設定

離散集合 $\mathcal{X}$ 上の確率分布を推定するための枠組を述べる．ここで $\mathcal{X}$ として，例えば $\{+1,-1\}^D$, $D\in\mathbb{N}$ などを想定している．適当な条件のもとで離散無限集合としてもよい．以下，関数 $f:\mathcal{X}\to\mathbb{R}$ に対して $\langle f\rangle$ を $\langle f\rangle=\sum_{x\in\mathcal{X}}f(x)$ と定義する．$\mathcal{X}$ 上での総和 $\langle f\rangle$ を厳密に求めることは，計算量の観点から難しいという状況を考える．関数の集合 $\mathcal{M},\mathcal{P}$ をそれぞれ

$$\mathcal{M}=\{f:\mathcal{X}\to\mathbb{R}_{\geq 0}\,|\,\langle f\rangle<\infty,\ f\text{は恒等的に}0\text{ではない}\},$$
$$\mathcal{P}=\{f\in\mathcal{M}\,|\,\langle f\rangle=1\}$$

と定義する．ここで $\mathbb{R}_{\geq 0}$ は非負実数の集合とする．$\mathcal{P}$ は $\mathcal{X}$ 上の確率関数の集合である．

統計モデル $q_{\boldsymbol{\theta}}$ は $\mathcal{P}$ の部分集合として定義され，パラメータ $\boldsymbol{\theta}$ で指定される関数 $\tilde{q}_{\boldsymbol{\theta}}\in\mathcal{M}$ を用いて

$$q_{\boldsymbol{\theta}}(\boldsymbol{x})=\frac{\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})}{Z_{\boldsymbol{\theta}}},\quad Z_{\boldsymbol{\theta}}=\langle\tilde{q}_{\boldsymbol{\theta}}\rangle$$

と表される．ここで $\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$ を非正規化モデルという．規格化定数 $Z_{\boldsymbol{\theta}}$ を厳密に計算することは実際には不可能であり，さまざまな工夫が必要になる．以下，統計モデルの例を示す．

*本稿について，詳細は [9] を参照のこと．

†kanamori@is.nagoya-u.ac.jp
‡ttakashi@fun.ac.jp

例 1. $D \in \mathbb{N}$ として $\mathcal{X} = \{+1, -1\}^D$ とする．ボルツマンマシンと呼ばれる統計モデル $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ は，以下の非正規化モデルを規格化して定義される．

$$\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left\{ \sum_i x_i \theta_i + \sum_{i<j} x_i x_j \theta_{ij} \right\} \in \mathcal{M},$$
$$\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathcal{X},\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_D, \theta_{12}, \ldots, \theta_{D-1,D}) \in \mathbb{R}^{D(D+1)/2}.$$

□

例 2. 制限ボルツマンマシンと呼ばれる統計モデルは，隠れ変数をもつボルツマンマシンの特別な場合として定義される．変数を $\boldsymbol{v} \in \{+1, -1\}^{D_0}$, $\boldsymbol{h} \in \{+1, -1\}^{D_1}$ とする．非正規化モデル $\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{v})$ はパラメータ $\theta_i, \theta'_j, \theta_{ij}, i = 1, \ldots, D_0, j = 1, \ldots, D_1$ を用いて

$$\begin{aligned}\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{v}) &= \sum_{\boldsymbol{h}} \exp\left\{ \sum_i v_i \theta_i + \sum_j h_j \theta'_i + \sum_{i,j} v_i h_j \theta_{ij} \right\} \\ &= e^{\sum_i v_i \theta_i} \prod_j \left(e^{\theta'_j + \sum_i v_i \theta_{ij}} + e^{-\theta'_j - \sum_i v_i \theta_{ij}} \right)\end{aligned}$$

と定義される．制限ボルツマンマシンは深層学習において，画像認識のタスクなどを行うための統計モデルとして用いられることがある *[10]*. □

データ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{X}$ が，未知の確率分布 $p \in \mathcal{P}$ から独立に生成されたとする．確率分布 $p(\boldsymbol{x})$ に対して，パラメータ $\boldsymbol{\theta}$ をもつ統計モデル $q_{\boldsymbol{\theta}} \in \mathcal{P}$ を仮定する．統計モデルの中から，$p(\boldsymbol{x})$ をよく近似する確率分布をデータに基づいて推定する問題を考える．よく用いられる方法として最尤推定法がある．これは，負の対数尤度

$$-\frac{1}{N} \sum_{k=1}^{N} \log q_{\boldsymbol{\theta}}(\boldsymbol{x}_k) = \log Z_{\boldsymbol{\theta}} - \frac{1}{N} \sum_{k=1}^{N} \log \tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x}_k)$$

を最小にするパラメータを推定量とする方法である．適当な条件のもとで，漸近一致性や漸近有効性など，理論的に優れた性質をもつことが示されている．しかし規格化定数 $Z_{\boldsymbol{\theta}}$ の計算が困難な場合には，計算上の工夫が必要になる．

最尤推定を求めるための計算アルゴリズムの例を示す．対数尤度を勾配法で最適化することを考える．勾配を計算すると

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left(\log Z_{\boldsymbol{\theta}} - \frac{1}{N} \sum_{k=1}^{N} \log \tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x}_k) \right) = \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial \log \tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}} \right] - \frac{1}{N} \sum_{k=1}^{N} \frac{\partial \log \tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x}_k)}{\partial \boldsymbol{\theta}}$$

となる．右辺第1項の期待値をマルコフ連鎖モンテカルロ法 (MCMC) などを用いて近似し，最適化計算を実行することができる [6]. MCMC を用いる方法は制限ボルツマンマシンでは非常に有効であることが，さまざまな数値例によって示されている．

規格化定数を近似的に計算して最適化する方法の他に，規格化定数の計算を必要としない損失関数を用いて，確率分布を推定する方法も提案されている．本稿では，主にこちらの方法について考察する．

2 スコアとダイバージェンス

最尤推定における対数尤度を他の損失に置き代えることができる．データ $\boldsymbol{x}$ を確率分布 $q(\boldsymbol{x})$ で予測するとき，その損失を $\ell(\boldsymbol{x}, q)$ とする．統計モデル $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ を用いて推定をおこなうとき，経験分布による平均損失

$$\min_{q\in\mathcal{P}} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i, q)$$

を最小にするパラメータを求める．このような方法は，統計的決定理論などで理論的な性質が詳しく調べられている．統計的一致性などの性質を保持しながら，最尤推定とは異なる性質をもつ推定量を構成することができる．

本稿では，スコアを損失の期待値によって定義する．以下に詳細を述べる．

定義 1 (スコア). 次の *1, 2* の性質を満たす $S : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ をスコアという．

1. $f, g \in \mathcal{M}$ に対して

$$S(f, g) \geq S(f, f)$$

が成立する．

2. $p, q \in \mathcal{P}$ に対して次が成り立つ．

(a) ある関数 $\ell(\boldsymbol{x}, q)$ が存在して，

$$S(p, q) = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\ell(\boldsymbol{x}, , q).$$

(b) $S(p, q) = S(p, p)$ なら $p = q$ が成立する．

条件 2(b) を満たすスコアを，厳密には狭義適切スコア (strictly proper scoring rule) という．簡単のため本稿ではスコアとよぶ．定義域 $\mathcal{M} \times \mathcal{M}$ を適当に制限することもある．

スコア $S(p, q)$ の p にデータ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ の経験分布 $\tilde{p}$ を代入すると

$$S(\tilde{p}, q) = \frac{1}{N}\sum_{k=1}^{N} \ell(\boldsymbol{x}_k, q)$$

となる．適当な統計モデル $q_{\boldsymbol{\theta}}$ を q に代入し，パラメータに関して最小化することで，スコアの意味で最適な推定量が得られる．例として最尤推定量 $S(\boldsymbol{x}, q) = -\log q(\boldsymbol{x})$ がある．また，後に示す密度冪スコアや擬球スコアによる推定も，スコアによる推定法の重要な例になっている．

スコアからダイバージェンスを定義する．ダイバージェンスとは距離の 2 乗を一般化した量であり，主に 2 つの確率分布や関数の間の乖離度を測る尺度として，統計学や情報理論，情報幾何学などで重要な役割を果たしている．

定義 2 (ダイバージェンス). $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$ が次の性質を満たすとき，D を $\mathcal{M}$ 上のダイバージェンスという.

1. $f, g \in \mathcal{M}$ に対して $D(f,g) \geq 0$.
2. $f \in \mathcal{M}$ に対して $D(f,f) = 0$.

同様にして，$\mathcal{P}$ 上のダイバージェンスも定義される. 応用上重要なダイバージェンスでは，上記の性質に加えて $D(f,g) = 0$ なら $f = g$ が成り立つ. スコア S と単調増加関数 $\zeta : \mathbb{R} \to \mathbb{R}$ に対して,

$$D(f,g) = \xi(S(f,g)) - \xi(S(f,f)) \tag{1}$$

はダイバージェンスの定義を満たす. さらにスコアから定義されるダイバージェンス D は，$p, q \in \mathcal{P}$ に対して $D(p,q) = 0$ なら $p = q$ が成り立つ.

以下にダイバージェンスの例をいくつか挙げる.

例 3 (カルバック・ライブラー (KL) ダイバージェンス). *KL* ダイバージェンスは，確率分布間の距離尺度として最もよく用いられるダイバージェンスのひとつである. 統計学における最尤推定量と関連し，また情報理論の分野では冗長度とよばれている. *KL* ダイバージェンスは確率関数 $p, q \in \mathcal{P}$ に対して

$$\mathrm{KL}(p,q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \left\langle p \log \frac{p}{q} \right\rangle$$

と定義される. これは $\mathcal{P}$ 上のダイバージェンスである. $\mathcal{M}$ 上のダイバージェンスに拡張することもできる. スコア

$$S(p,q) = -\langle p \log q \rangle$$

を用いると，$\mathrm{KL}(p,q) = S(p,q) - S(p,p)$ と表せる. データが観測されたとき，その経験分布を p に代入し，統計モデルを q に代入する. *KL* ダイバージェンスの意味で経験分布に最も近い統計モデルを選択する推定法は，最尤推定に一致する. □

例 4 (α ダイバージェンス). 非負値関数 $f, g \in \mathcal{M}$ に対して

$$D_\alpha(f,g) = \frac{1}{\alpha(1-\alpha)} \left\langle \alpha f + (1-\alpha) g - f^\alpha g^{1-\alpha} \right\rangle$$

と定義されるダイバージェンスを α-ダイバージェンスという. ここで $\alpha \in \mathbb{R}$ は定数とする. $\alpha = 0, 1$ のときは，それぞれの極限として

$$\lim_{\alpha \to 1} D_\alpha(f,g) = \mathrm{KL}(f,g), \qquad \lim_{\alpha \to 0} D_\alpha(f,g) = \mathrm{KL}(g,f)$$

と定義される．ダイバージェンスの定義を満たすことは，

$$D_\alpha(f,g)=\left\langle g\left(\frac{1}{1-\alpha}\frac{f}{g}+\frac{1}{\alpha}-\frac{1}{\alpha(1-\alpha)}\left(\frac{f}{g}\right)^\alpha\right)\right\rangle$$

と変形すると，$\langle\cdots\rangle$ の中が非負であることから分かる．一般に α ダイバージェンスは，(1) のようにスコアから導出される形式にはなっていない．α ダイバージェンスは，f ダイバージェンス (もしくは φ ダイバージェンス) とよばれるダイバージェンス・クラスに属している *[1, 4]*. □

例 5 (擬球ダイバージェンス). 次式で定義されるダイバージェンスを擬球ダイバージェンスという：

$$\mathrm{PS}_\gamma(f,g)=\frac{1}{1+\gamma}\log\langle f^{1+\gamma}\rangle+\frac{\gamma}{1+\gamma}\log\langle g^{1+\gamma}\rangle-\log\langle fg^\gamma\rangle,\quad f,g\in\mathcal{M}$$

ここで γ は正の定数とする．ダイバージェンスの定義を満たすことは，ヘルダーの不等式から分かる．さらに $f,g\in\mathcal{M}$ に対して $\mathrm{PS}_\gamma(f,g)=0$ なら，f と g は *1* 次従属である．擬球ダイバージェンスは，擬球スコア

$$S(f,g)=-\frac{\langle fg^\gamma\rangle}{\langle g^{1+\gamma}\rangle^{\gamma/(1+\gamma)}}$$

から $-\log(-S(f,g))+\log(-S(f,f))$ によって導出される．データの経験分布を $\tilde{p}$，統計モデル q_θ として，擬球スコア $S(\tilde{p},q_\theta)$ を最小する推定量は，外れ値に対して非常にロバストであることが示されている *[5, 7]* □

例 6 (密度冪ダイバージェンス). 次式で定義されるダイバージェンスを密度冪ダイバージェンスという：

$$D(f,g)=\langle\gamma g^{1+\gamma}+f^{1+\gamma}-(1+\gamma)fg^\gamma\rangle,\quad f,g\in\mathcal{M}.$$

ここで γ は正の定数とする．ダイバージェンスの定義を満たすことは，関数 $z\mapsto z^{1+\gamma}$ の凸性から分かる．さらに $f,g\in\mathcal{M}$ に対して $D(f,g)=0$ なら $f=g$ が成り立つ．密度冪ダイバージェンスは，密度冪スコア

$$S(f,g)=\langle\gamma g^{1+\gamma}-(1+\gamma)fg^\gamma\rangle,\quad f,g\in\mathcal{M}$$

から (1) によって導出される．データの経験分布を $\tilde{p}$，統計モデルを q_θ として，スコア $S(\tilde{p},q_\theta)$ を最小する推定量は外れ値に対してロバストであることが示されている *[3]*. ロバスト推定において，擬球スコアとの関連が調べられている *[8]*. □

上記のスコアやダイバージェンスでは規格化定数の計算が必要となる．このため，これらの推定法を本稿の問題設定においてそのまま用いることはできない．

3 斉次ダイバージェンスとその局所化

統計的性質だけでなく計算量も考慮して，規格化定数を計算する必要がないスコアやダイバージェンスを導出する．まずダイバージェンスの斉次性を定義する．

定義 3 (ダイバージェンスの斉次性). $\mathcal{M}$ 上のダイバージェンス $D(f,g)$ が次の条件を満たすとき，斉次性をもつという．

1. $f,g\in\mathcal{M}$ と任意の $c>0$ に対して $D(f,g)=D(f,c\cdot g)$.
2. $D(f,g)=0$ なら f と g は *1* 次従属.

例 7. 擬球ダイバージェンスは斉次性をもつ．一方，*KL* ダイバージェンス，α ダイバージェンス，密度冪ダイバージェンスは斉次性を満たさない． □

確率分布 p と非正規化モデル $\tilde{q}$ に対して，擬球ダイバージェンスが $\mathrm{PS}_\gamma(p,\tilde{q})=0$ となるとき，$\tilde{q}$ は p に比例する関数になる．したがって，データの経験分布を $\tilde{p}$ として，$\mathrm{PS}_\gamma(\tilde{p},\tilde{q})$ を非正規化モデル $\tilde{q}$ に関して最小化することで，定数倍を除いてデータの分布を推定することができる．しかし擬球ダイバージェンスでは，$\langle\tilde{q}^{1+\gamma}\rangle$ の計算で $\mathcal{X}$ 上の総和が必要になる．これは規格化定数を求めるのと同程度の計算量が必要であり，実用的ではない．

総和の計算を避けるために，経験分布 $\tilde{p}$ を用いて非正規化モデルの定義域を局所化することを考える．具体的には，非正規化モデル $\tilde{q}(\boldsymbol{x})\in\mathcal{M}$ の代わりに，適当な実数 α に対して $\tilde{p}(\boldsymbol{x})^\alpha\tilde{q}(\boldsymbol{x})^{1-\alpha}$ を擬球ダイバージェンスに代入する．ここで $\tilde{p}(\boldsymbol{x})^\alpha\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}$ は，$\tilde{p}(\boldsymbol{x})$ が経験分布のとき観測データ上でのみ非ゼロの値を取る．したがって，$\langle\tilde{p}(\boldsymbol{x})^\alpha\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}\rangle$ は

$$\langle\tilde{p}(\boldsymbol{x})^\alpha\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}\rangle=\sum_{\boldsymbol{x}\in\mathcal{X}}\left(\frac{N_{\boldsymbol{x}}}{N}\right)^\alpha\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}=\sum_{\substack{\boldsymbol{x}:\text{データに}\\ \text{現れるパターン}}}\left(\frac{N_{\boldsymbol{x}}}{N}\right)^\alpha\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}$$

となる．ここで $N_{\boldsymbol{x}}$ はパターン $\boldsymbol{x}$ に一致するデータ数とする．よって観測データ数のオーダーで計算できる．上記のような，経験分布との積型混合 $\tilde{p}(\boldsymbol{x})^\alpha\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}$ を経験分布による局所化とよぶ．

確率分布 $p(\boldsymbol{x})$ と非正規化モデル $\tilde{q}(\boldsymbol{x})$ に対して，2つの積型混合を

$$p(\boldsymbol{x})^\alpha\tilde{q}(\boldsymbol{x})^{1-\alpha},\quad p(\boldsymbol{x})^{\alpha'}\tilde{q}(\boldsymbol{x})^{1-\alpha'}$$

とし，これらを擬球-ダイバージェンスに代入する．ここで α と α' は異なる実数とする．このとき，もし

$$\mathrm{PS}_\gamma(p(\boldsymbol{x})^\alpha\tilde{q}(\boldsymbol{x})^{1-\alpha},p(\boldsymbol{x})^{\alpha'}\tilde{q}(\boldsymbol{x})^{1-\alpha'})=0$$

なら $p^\alpha\tilde{q}^{1-\alpha}\propto p^{\alpha'}\tilde{q}^{1-\alpha'}$，すなわち $p\propto\tilde{q}$ が得られる．さらに，p に経験分布 $\tilde{p}$ を代入すると，ダイバージェンスを計算するための計算量はデータ数のオーダー程度になる．ここで $p\in\mathcal{P},q\in\mathcal{M}$ に対して

$$\mathrm{LPS}_{\alpha,\alpha',\gamma}(p,q)=\mathrm{PS}_\gamma(p^\alpha q^{1-\alpha},p^{\alpha'}q^{1-\alpha'})$$

とおき，局所擬球ダイバージェンスとよぶ．以上の考察から，経験分布 $\tilde{p}$ と非正規化モデル $\tilde{q}_{\boldsymbol{\theta}}$ の間の局所擬球ダイバージェンスを最小にする推定法

$$\min_{\boldsymbol{\theta}} \mathrm{LPS}_{\alpha,\alpha',\gamma}(\tilde{p}, \tilde{q}_{\boldsymbol{\theta}})$$

は，$\mathcal{X}$ 上での総和を必要とせず，統計的にはフィッシャー一致性をもつことが分かる．

最適化の観点から局所擬球ダイバージェンスの性質を調べると，次の定理が得られる．

定理 1 ([9]). 非正規化モデル $\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$ は，規格化されていない指数型分布族として

$$\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\{\boldsymbol{\theta}^T \boldsymbol{h}(\boldsymbol{x})\}, \quad \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$$

と定義されるとする．また $p \in \mathcal{P}$ とする．このとき $\mathrm{LPS}_{\alpha,\alpha',\gamma}(p, \tilde{q}_{\boldsymbol{\theta}})$ がパラメータ $\boldsymbol{\theta}$ に関して凸関数になることと，$(\alpha + \gamma\alpha')/(1+\gamma) = 1$ が成り立つことは同値である．

局所擬球ダイバージェンスのパラメータが，定理1の $(\alpha + \gamma\alpha')/(1+\gamma) = 1$ を満たすとき，γ を α, α' で表して，

$$\mathrm{LPS}_{\alpha,\alpha'}(p, q) = \frac{1-\alpha'}{\alpha-\alpha'} \log\langle p^{\alpha} q^{1-\alpha} \rangle + \frac{\alpha-1}{\alpha-\alpha'} \log\langle p^{\alpha'} q^{1-\alpha'} \rangle$$

とおく．このとき簡単な考察から，$\alpha > 1 > \alpha' \neq 0$ として一般性を失わない．

4 局所擬球ダイバージェンスによる推定

局所擬球ダイバージェンスから得られる推定量の統計的性質について述べる．

定理 2 ([9]). 非正規化モデルを $\tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$ とする．データがしたがう確率分布は $p(\boldsymbol{x}) = q_{\boldsymbol{\theta}_0}(\boldsymbol{x}) = \tilde{q}_{\boldsymbol{\theta}}(\boldsymbol{x})/Z_{\boldsymbol{\theta}}$ と表せるとする．局所擬球ダイバージェンス $\mathrm{LPS}_{\alpha,\alpha',\gamma}$ の最小化によって得られる推定量 $\widehat{\boldsymbol{\theta}}$ が漸近的に正規分布にしたがうと仮定する．また正規化された統計モデル $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ のフィッシャー情報量を $I(\boldsymbol{\theta})$ とする．このとき

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1})$$

が成り立つ．

標準的な漸近展開によって，上記の結果を確認することができる．この定理から，局所擬球ダイバージェンスが有効推定量を与えることが分かる．

有効推定量が得らえる理由を直感的に説明する．まず，局所擬球ダイバージェンスと例 4 で示した α ダイバージェンスとの関連を述べる．α ダイバージェンス D_{α} を用いて，次に示すように規格化定数を計算しない推定を行うことができる．確率分布 $p \in \mathcal{P}$ と非

正規化モデル $q \in \mathcal{M}$ の間の α ダイバージェンス $D_\alpha(p,q)$ には $\langle q \rangle$ が現れるため，直接計算することは困難である．パラメータが異なる2つの α ダイバージェンスの差を考え，

$$\begin{aligned} D_{\alpha,\alpha'}(p,q) &= D_\alpha(p,q) - \frac{\alpha'}{\alpha} D_{\alpha'}(p,q) \\ &= \left\langle \left(\frac{1}{1-\alpha} - \frac{\alpha'}{\alpha(1-\alpha')}\right)p - \frac{1}{\alpha(1-\alpha)} p^\alpha q^{1-\alpha} + \frac{1}{\alpha(1-\alpha')} p^{\alpha'} q^{1-\alpha'} \right\rangle \end{aligned}$$

のように定義すると，$\langle q \rangle$ の項がキャンセルする．さら p と q の積型混合が現れる．$D_{\alpha,\alpha'}(p,q)$ を混合 (α,α') ダイバージェンスとよぶ．このとき

$$\min_{z>0} D_{\alpha,\alpha'}(p, q/z) = c_{\alpha,\alpha'} \left(\exp(\mathrm{LPS}_{\alpha,\alpha'}(p,q)) - 1\right)$$

が成り立つ．ここで $c_{\alpha,\alpha'} = (\alpha - \alpha')/(\alpha(1-\alpha')(\alpha-1)) > 0$ で与えられる．したがって，混合 (α,α') ダイバージェンスを定数倍 z について最小化した乖離度は，局所擬球ダイバージェンスと単調変換で関連付けられている．すなわち，混合 (α,α') ダイバージェンスの最小化と局所擬球ダイバージェンスの最小化は等価である．

α ダイバージェンスは情報幾何において重要な役割を果たしている [2]．実際，指数型分布族上で α ダイバージェンスの (自然パラメータに関する) ヘッセ行列は，フィッシャー情報量に一致する．したがって，α ダイバージェンスに基づく推定法はフィッシャー有効である．混合 (α,α') ダイバージェンスでも同様であり，よってそれと等価な局所擬球ダイバージェンスがフィッシャー有効な推定量を導出することが結論される．定理2の結果は，この対応関係を3つのパラメータをもつ局所擬球ダイバージェンス $\mathrm{LPS}_{\alpha,\alpha',\gamma}$ の場合に拡張したとみなせる．

本稿で提案された統計的手法について，数値実験の結果などは [9] に詳しい．

参考文献

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28(1):131–142, 1966.

[2] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, 2000.

[3] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[4] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[5] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.*, 99(9):2053–2081, 2008.

[6] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, August 2002.

[7] T. Kanamori and H. Fujisawa. Affine invariant divergences associated with composite scores and its applications. *Bernoulli*, 20(4):2278–2304, 2014.

[8] T. Kanamori and H. Fujisawa. Robust estimation under heavy contamination using unnormalized models. *Biometrika*, 102(3):559–572, 2015.

[9] T. Takenouchi and T. Kanamori. Empirical localization of homogeneous divergences on discrete sample spaces. In *The Neural Information Processing Systems (NIPS 2015)*, 2015.

[10] 岡谷貴之. 深層学習. 講談社, 2015.