数理解析研究所講究録
第1916巻 2014年 213-224

アフィン不変ダイバージェンスとその応用

名古屋大学・情報科学研究科　金森 敬文 *
Takafumi KANAMORI
Nagoya University

統計数理研究所　藤澤 洋徳 †
Hironori FUJISAWA
Institute of Statistical Mathematics

概要

統計におけるデータ解析では，損失関数の最小化により推定量や統計量を得ることが多い．損失関数は問題に合わせて適切に定める必要がある．スコアは損失関数の一般的なクラスであり，その性質がよく研究されている．本稿ではスコアを拡張した合成スコアを定義し，その性質について不変性の観点から調べる．さらに，合成スコアから定義される推定量をロバスト推定に応用し，その統計的性質について述べる．

1　はじめに

統計データの解析では，確率分布を基礎に置いてデータの生成機構に関する推論を行う．このとき，データを生成するサンプル分布を近似的に表すために，統計モデルを設定するこが多い．近似の程度を定量的に表すために，確率分布の間の「距離」を適切に定義することが重要である．例えば，データを生成する確率分布を推定するために，統計モデルを用いる状況を考える．代表的な推定法として最尤推定量がある．これは，分布間の「距離」を Kullback-Leibler ダイバージェンスと呼ばれる尺度で測ったとき，サンプル経験分布に最も近い統計モデルの分布を推定量として用いる方法である．多くの統計的手法には，確率分布間の距離 (ダイバージェンス) を対応させることができる．本論文では，統計的推論において有用と考えられる距離尺度を，データ解析において自然に要請されるいくつかの仮定から導出し，その特徴付けを行う．また統計的推論，とくにロバスト推定への応用についても考察する．

* kanamori@is.nagoya-u.ac.jp
† fujisawa@ism.ac.jp

以下で本稿で用いる記号を定義しておく．実数の集合を $\mathbb{R}$, 非負実数の集合を $\mathbb{R}_+$ とする．また関数 $f(x)$ の積分 $\int f(x)dx$ を $\langle f \rangle$ と表す．

2 スコアと合成スコア

統計的推論に用いられるスコアと合成スコアを定義する．本稿では，ユークリッド空間上でルベーグ測度に関して確率密度をもつ分布を扱う．

2.1 定義

確率密度関数 $q(x)$ を用いて予測を行うとする．データ x が観測されたときに被る損失を $\ell(x,q)$ とする．例えば対数損失 $\ell(x,q) = -\log q(x)$ などを考える．

定義 1. 確率密度 p, q に対して，

$$S_0(p,q) = \int \ell(x,q)p(x)dx$$

と表すことができる汎関数 S_0 をスコア *(score; scoring rule)* という．任意の p, q に対して，積分値が存在する場合に不等式 $S_0(p,q) \geq S_0(p,p)$ が成り立つとき，S_0 を適正スコア *(proper scoring rule)* という．さらに等号 $S_0(p,q) = S_0(p,p)$ が成り立つなら $p = q\,(a.s.)$ となるとき，真適正スコア *(strictly proper scoring rule)* という．

真適正スコアを用いて統計的推定を行うことができる．独立に同一の分布にしたがうサンプル $x_1, \ldots, x_n$ が観測されたとき，母集団分布の確率密度 $p(x)$ を統計モデル $p_\theta(x)$ を用いて推定することを考える．統計モデルが分布 $p(x)$ を含み，$p(x) = p_{\theta_0}(x)$ と表せるとする．またサンプルの経験分布を $\widetilde{p}(x) = \frac{1}{n}\sum_{i=1}^n \delta(x - x_i)$ とする．ここで $\delta(x)$ は Dirac のデルタ関数とする．損失関数 $\ell(x,q)$ を用いた真適正スコア S_0 に経験分布を代入すると $S_0(\widetilde{p}, p_\theta) = \frac{1}{n}\sum_{i=1}^n \ell(x_i, p_\theta)$ となる．$S_0(\widetilde{p}, p_\theta)$ を θ に関して最小化したときの最適解を $\widehat{\theta}$ とおくと，これは母集団分布のパラメータ θ_0 をよく近似していることが期待される．なぜなら，データ数が十分多いとき，大数の法則より $S_0(\widetilde{p}, p_\theta)$ は $S_0(p, p_\theta)$ に概収束し，S_0 が真適正スコアであることから $\min_\theta S_0(p, p_\theta)$ の最適解は $\theta = \theta_0$ で与えられるからである．真適正スコアによる推定は M-推定量 [12] の例となっている．このため，真適正スコアから定義される推定量の統計的性質を調べるとき，M-推定量の一般論を適用することができる．

次に合成スコア (composite score) を定義する．

定義 2. スコア S_0 を用いて $S(p,q) = T(S_0(p,q), q)$ と表せる汎関数 S を合成スコアという．ここで T は $z \in \mathbb{R}$ と確率密度 q に対して実数値を対応させる関数である．スコア

と同様に，適正合成スコア，真適正合成スコアを定義する．======また $S(p,p)$ を p のエントロピーと定義する．======

簡単のため，適正合成スコアや真適正合成スコアの代わりに適正スコアや真適正スコアという．スコアと同様に，合成スコア $S(p,q)$ の p に経験分布 $\tilde{p}$ を代入することができる．したがって，合成スコアを統計的推定に直接用いることができる．真適正スコアの場合と同様に，真適正合成スコアによる推定量は適切な条件の下で統計的一致性をもつことを証明することができる．

スコアや合成スコアを統計的推論に応用することを考える．このとき，合成スコア $S(p,q)$ の p と q にそれぞれ経験分布 $\tilde{p}$ と統計モデル p_θ を代入し，θ に関して最小化するという操作を行う．このとき，最小解は合成スコアの単調変換に対して不変である．この事実から，2 つの合成スコアの等価性を次のように定義する．

定義 3. 合成スコア S_1 と S_2 が等価 *(equivalent)* であるとは，ある単調増加関数 $\xi : \mathbb{R} \to \mathbb{R}$ が存在して $S_1 = \xi(S_2)$ を満たすことである．

真適正合成スコアから，確率分布の間の「距離」を表すダイバージェンスを定義することができる．

定義 4 (ダイバージェンス)**.** $S(p,q)$ を真適正合成スコアとする．このとき $D(p,q) = S(p,q) - S(p,p)$ を S から定義されるダイバージェンスとよぶ．

定義よりダイバージェンスは非負であり，$D(p,q) = 0$ は $p = q$ を意味する．

2.2 スコアと Bregman ダイバージェンス

スコアは Bregman ダイバージェンスと呼ばれるダイバージェンスのクラスと関連が深い．その関連について紹介する．

定義 5 (Bregman スコア)**.** 確率密度関数の集合上で定義された凸関数 $G(p)$ に対して

$$S(p,q) = -G(q) - \int (p(x) - q(x))G^*(x,q)dx$$

を，ポテンシャル G から定義される *Bregman* スコアとよぶ．ここで $G^*(x,q)$ は G の q での劣微分である．

凸関数の性質から Bregman スコアは適正スコアであり，G が強凸関数なら真適正スコアとなる．Bregman スコアから定義されるダイバージェンスを Bregman ダイバージェンスという．Bregman スコアに対して以下の定理が成り立つ．

定理 1 ([4, 7])**.** 正則性の条件のもとで，真適正スコアは *Bregman* スコアとして表せる．

この定理により，正則性の条件のもとで真適正スコアと強凸ポテンシャルから定義される Bregman スコアは同値な概念であることが分かる．Bregman スコアの重要なサブクラスである分離可能 Bregman スコアを定義する．

定義 6 (分離可能 Bregman スコア)．凸関数 $J : \mathbb{R}_+ \Rightarrow \mathbb{R}$ に対して $G(p)$ を $G(p) = \langle J(p) \rangle$ とおく．$G(p)$ をポテンシャルとする *Bregman* スコアを分離可能 *Bregman* スコアという．

分離可能 Bregman スコアの計算には，関数 J と導関数 J' に関連する積分を実行すればよい．計算の容易さから，データ解析に応用されている Bregman スコアのほとんどが分離可能 Bregman スコアである．

これまで，スコアやダイバージェンスを確率密度関数に対して定義したが，自然に非負値関数に対して定義することができる．以下に，非負値関数上で定義される Bregman スコアの例を挙げる．

例 1 (Kullback-Leiber スコア)．非負値関数 f, g に対して

$$S(f, g) = \langle -f \log g + g \rangle$$

を *Kullback-Leiber* スコアという．対応するダイバージェンスを *Kullback-Leiber* ダイバージェンスという．*Kullback-Leiber* スコアは分離可能 *Bregman* スコアであり，ポテンシャルは $G(f) = \langle f \log f - f \rangle$ で与えられる．*Kullback-Leiber* スコアを用いた推定は最尤推定量に一致する．

例 2 (Density-power スコア [1])．*Density-power* スコアは正パラメータ $\gamma > 0$ をもつスコアのクラスであり，非負値関数 f, g に対して

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

と定義される．パラメータ γ を $\gamma \to 0$ とした極限で *Kullback-Leiber* スコアに一致する．*Density-power* スコアは分離可能 *Bregman* スコアであり，ポテンシャルは $G(f) = \langle f^{1+\gamma} \rangle / \gamma$ で与えられる．*Kullback-Leiber* スコアを用いた推定は最尤推定量に一致する．*Density-power* スコアはロバスト推定に用いられる．

例 3 (擬球スコア [5, 3])．擬球スコアは正パラメータ $\gamma > 0$ をもつスコアのクラスであり，恒等的に零でない非負値関数 f, g に対して

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

と定義される．擬球スコア $S(f, g)$ を単調変換した $-\frac{1}{\gamma} \log(-S(f, g))$ をガンマ・スコアという *[3]*．ガンマ・スコアは，パラメータ γ を $\gamma \to 0$ とした極限で *Kullback-*

Leiber スコアに一致する．擬球スコアに対応する *Bregman* スコアのポテンシャルは $G(f) = \langle f^{1+\gamma}\rangle^{\gamma/(1+\gamma)}$ で与えられる．したがって，分離可能 *Bregman* スコアではない．擬球スコアはロバスト推定に用いられる．

3 Hölder スコア

統計的に自然な性質を要請することで，Bregman スコアでは表せない合成スコアを導出することができる．本節では，このような合成スコアである Hölder スコアについて解説する．まず Hölder スコアの定義を以下に示す．

定義 7 (Hölder スコア)**.** 非負実数を γ とする．$\gamma = 0$ のとき非負かつ恒等的に零でない関数 f, g に対して *Hölder* スコアは $S(f,g) = \langle -f \log g + g\rangle$ と定義される．また $\gamma > 0$ に対して

$$S(f,g) = \phi\left(\frac{\langle fg^{\gamma}\rangle}{\langle g^{1+\gamma}\rangle}\right)\langle g^{1+\gamma}\rangle$$

とする．ここで関数 $\phi : \mathbb{R}_+ \to \mathbb{R}$ は $z \geq 0$ に対して $\phi(z) \geq -z^{1+\gamma}$，かつ $\phi(1) = -1$ を満たす．

f が確率密度のとき，Hölder スコア $S(f,g)$ は一般には f に関する期待値として表すことはできない．しかし，$S_0(f,g) = \langle fg^{\gamma}\rangle$, $T(z,g) = \phi(z/\langle g^{1+\gamma}\rangle)\langle g^{1+\gamma}\rangle$ とおくことで合成スコアであることが確認できる．負の γ に対しても，積分値が存在するなら Hölder スコアを定義することが出来る．本稿では簡単のため $\gamma \geq 0$ としておく．まず Hölder スコアが適正スコアであることを証明する．

定理 2. 非零な非負値関数に対して，*Hölder* スコアは適正スコアであり，確率密度に対して，*Hölder* スコアは真適正スコアである．また，$z \neq -1$ に対して $\phi(z) > -z^{1+\gamma}$ が成り立つなら，非負値関数に対して，*Hölder* スコアは真適正スコアである．

Proof. $\gamma = 0$ のとき Kullback-Leibler スコアの性質から言える．以下 $\gamma > 0$ とする．非負値関数 f, g と $\gamma > 0$ に対して，ヘルダー不等式

$$\langle fg^{\gamma}\rangle \leq \langle f^{1+\gamma}\rangle^{1/(1+\gamma)}\langle g^{1+\gamma}\rangle^{\gamma/(1+\gamma)}$$

が成り立つ．ヘルダー不等式と不等式 $\phi(z) \geq -z^{1+\gamma}$ より

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

となるので，Hölder スコアは非負値関数に対して適正スコアである．

関数 f,g が確率密度のとき，$S(f,g)=S(f,f)$ が成り立つと仮定する．このときヘルダー不等式が等式で成り立つので，f,g は線形従属である．f,g が確率密度であることより，線形従属なら $f=g$ となる．したがって，Hölder スコアは確率密度関数に対して真適正スコアである．

関数 $\phi(z)$ が．$\phi(z) > -z^{1+\gamma}, z \neq 1$ を満たすとする．非負値関数 f,g に対して $S(f,g)=S(f,f)$ が成り立つと仮定すると，ヘルダー不等式が等式で成り立つので，f,g は線形従属である．$f=cg$ とおくと，$S(cg,g)=S(cg,cg)$ より $\phi(c)\langle g^{1+\gamma}\rangle = -c^{1+\gamma}\langle g^{1+\gamma}\rangle$ となり，$\langle g^{1+\gamma}\rangle \neq 0$ から $\phi(c)=-c^{1+\gamma}$ となる．仮定より，この等式が成り立つのは $c=1$ のときのみである．したがって $f=g$ となる．以上より，関数 ϕ が仮定を満たすとき，Hölder スコアは非負値関数に対して真適正スコアである． □

Bregman スコアと Hölder スコアの関連を次に示す．

定理 3. $\gamma>0$ の *Hölder* スコアに対して以下が成り立つ

1. *Hölder* スコアと *Bregman* スコアとの共通部分は，ポテンシャル

$$G(f)=\langle f^{1+\gamma}\rangle^{\kappa/(1+\gamma)}, \quad \gamma\geq 0, \quad \kappa\geq 1 \tag{1}$$

をもつ *Bregman* スコアと等価である．

2. *Hölder* スコアと分離可能 *Bregman* スコアとの共通部分は，*density-power* スコアと等価である．

例 4 (Bregman-Hölder スコア)．ポテンシャル (1) をもつ *Bregman* スコアは

$$S(f,g)=\langle g^{1+\gamma}\rangle^{\kappa/(1+\gamma)}\left(1-\frac{1}{\kappa}-\frac{\langle fg^{\gamma}\rangle}{\langle g^{1+\gamma}\rangle}\right), \quad \gamma>0,\ \kappa\geq 1$$

で与えられる．$\kappa=1+\gamma$ とすると *density-power* スコア，$\kappa=1$ とすると擬球スコアが得られる．本稿ではこのスコアを *Bregman-Hölder* スコアとよぶ．

4 Hölder スコアの不変性

Hölder スコアの特徴付けを与える．以下では簡単のため 1 次元確率変数を扱うが，多次元確率変数についても同様の結果が成り立つ．

1 次元確率変数 X の確率密度を $p(x)$ とする．確率変数 X を $Y=(X-\mu)/\sigma$ とアフィン変換すると，確率密度は $p_{\mu,\sigma}(y)=|\sigma|p(\sigma y+\mu)$ と変換される．アフィン変換は，データの測定において単位系を変えることに対応している．単位系を変換すればデータの値は変化するが，データ解析の結果が単位系に依存することは望ましくないと考えら

れる．したがって，統計的推論の結果はアフィン変換に対して不変であることが求められる．

上記の条件は，推定量がアフィン変換に対して共変的であると言い換えることもできる．確率密度の推定が共変的であるとは，次の性質が成り立つことである：データ X と統計モデル $q(x)$ から得られる推定量を確率密度 $\widehat{q}(x)$ とし，データ $Y=(X-\mu)/\sigma$ と統計モデル $q_{\mu,\sigma}(y)$ から得られる推定量を確率密度 $\widehat{q_{\mu,\sigma}}(y)$ とするとき，$(\widehat{q})_{\mu,\sigma}=\widehat{q_{\mu,\sigma}}$ が成り立つ．すなわち，元のデータで推定してから推定量を変換しても，変換したデータを用いて推定しても同じ結果を与える．このような推定量は，どのような単位系でデータを測定しても，本質的に同じ推定結果を与えると解釈することができる．

スコアや合成スコアが確率密度の変換 $p\mapsto p_{\mu,\sigma}$ に対して不変であるなら，対応する推定量はデータのアフィン変換に対して共変的になる．以下でこれを示す．合成スコア S が任意の ($\sigma\neq 0$ を満たす) アフィン変換に対して不変であるとき，$S(p,q)=S(p_{\mu,\sigma},q_{\mu,\sigma})$ となる．また合成スコアから導かれる推定量 $\widehat{q}$ について，任意の q に対して $S(p,\widehat{q})\leq S(p,q)$ となる．したがって，任意の q に対して $S(p_{\mu,\sigma},(\widehat{q})_{\mu,\sigma})\leq S(p_{\mu,\sigma},q_{\mu,\sigma})$ となる．したがって，推定量が一意なら $(\widehat{q})_{\mu,\sigma}=\widehat{q_{\mu,\sigma}}$ が成り立つ．

合成スコアに対して，不変性より緩い相対不変性を仮定しても，推定量について同じ結論を得る．合成スコアの相対不変性を，以下のようにダイバージェンスから定義する．

定義 8 (相対不変)．合成スコア $S(p,q)$ を真適正スコアとし，対応するダイバージェンスを $D(p,q)=S(p,q)-S(p,p)$ とする．合成スコアがアフィン変換に対して相対不変であるとは，正値関数 $h(\mu,\sigma)$ が存在して，任意の確率密度 p,q と任意のアフィン変換に対して

$$D(p,q)=h(\mu,\sigma)D(p_{\mu,\sigma},q_{\mu,\sigma})$$

となることである．

Hölder スコアは相対不変である．実際

$$\begin{aligned}D(p_{\sigma,\mu},q_{\sigma,\mu})&=\phi\left(\frac{\langle p_{\sigma,\mu}q_{\sigma,\mu}^{\gamma}\rangle}{\langle q_{\sigma,\mu}^{1+\gamma}\rangle}\right)\langle q_{\sigma,\mu}^{1+\gamma}\rangle+\langle p_{\sigma,\mu}^{1+\gamma}\rangle\\&=\phi\left(\frac{|\sigma|^{\gamma}\langle pq^{\gamma}\rangle}{|\sigma|^{\gamma}\langle q^{1+\gamma}\rangle}\right)\langle q^{1+\gamma}\rangle|\sigma|^{\gamma}+\langle p^{1+\gamma}\rangle|\sigma|^{\gamma}\\&=|\sigma|^{\gamma}D(p,q)\end{aligned}$$

となる．

以下で相対不変な合成スコアについて考察する．統計的推論における計算の簡便さを考慮し，合成スコア S が

$$S(f,g)=\psi(\langle fU(g)\rangle,\langle V(g)\rangle)\qquad(2)$$

と表せると仮定する．ここで U, V は $\mathbb{R}_+$ 上の実数値関数であり，また $\psi : \mathbb{R}^2 \to \mathbb{R}$ は 2 階連続微分可能な関数とする．これは，合成スコア $T(S_0(f,g),g)$ においてスコア S_0 の損失関数を $\ell(x,g) = U(g(x))$ とし，さらに $T(c,g) = \psi(c, \langle V(g) \rangle)$ とした場合に相当する．また (2) の合成スコアのクラスは，分離可能 Bregman スコアのクラスを含む．実際，ポテンシャル $G(p) = \langle J(p) \rangle$ から定義される分離可能 Bregman スコアは，$\psi(a,b) = -a + b$, $U(z) = J'(z)$, $V(z) = J'(z)z - J(z)$ として表すことができる．

Hölder スコアについて以下が成り立つ．

定理 4 ([8])**.** $S(f,g)$ を式 (2) で表せる合成スコアとし，関数 $V : \mathbb{R}_+ \to \mathbb{R}$ に対して $\lim_{z \searrow 0} V(z) = V(0) = 0$ を仮定する．さらに正則条件を仮定する．$S(f,g)$ がアフィン変換に対して相対不変なら $S(f,g)$ は *Hölder* スコアと等価である．

上の定理における関数 V に対する条件は，$p(x)$ のサポートがコンパクトであっても $\mathbb{R}$ 上の積分 $\langle V(g) \rangle$ が有限値となるために課している．定理 3 と定理 4 より，アフィン変換に対して不変な分離可能 Bregman スコアは density-power スコアに限ることが分かる．

5 さまざまなダイバージェンスにおける不変性

さまざまなクラスのダイバージェンスや分布上の擬距離に対して，不変性などの性質から特徴付けを与える研究が行われている．これに関連する研究をいくつか紹介し，本稿の結果との関連について説明する．

5.1 擬球スコアの特徴付け

例 3 で与えらえる擬球スコアの特徴付けは，ロバスト統計の観点から与えられる [3]. 以下，この結果を紹介する．$S(f,g)$ を式 (2) で表せる合成スコアとし，任意の正数 λ と任意の確率密度 p, q に対して，不等式

$$S(\lambda p, q) \geq S(\lambda p, p) \tag{3}$$

が成り立つことを仮定する．このとき，適当な正則条件の下で $S(p,q)$ は擬球スコアと等価である．

$S(p,q)$ に対する上の条件について補足する．いまデータの分布が $p(x) = (1-\varepsilon)p_0(x) + \varepsilon w(x)$ で与えられるとする．このとき $p_0(x)$ は推定対象の分布であり，$w(x)$ は外れ値の分布である．外れ値の割合 ε が十分小さいとき，外れ値を含むデータから $p_0(x)$ を推定するためのロバスト推定法が多数提案されている．例えば，$p_0(x)$ を正規分布としてデータから期待値を推定することを考える．このとき ε の値が十分小さいなら，たとえ極端に大きな値の外れ値がデータに含まれていても，データの中央値を用いることで期待値を精度よく推定することができる．しかし ε が小さくないとき，無視できない推定バイアスが

生じることがある．これに対して (3) を満たす合成スコア (2) を用いれば，外れ値の割合が (無限小ではなく) 有限の値であっても，推定量のバイアスが非常に小さくなることが保証される．詳細は [3] で解説されている．

一般の Hölder スコアは (3) の条件を満たさない．したがって，Hölder スコアを推定に用いたとき，外れ値の比率が大きいと推定バイアスが生じる可能性がある．一方，6 節で示すロバスト性の基準の下では，擬球スコア以外の Hölder スコアを用いて，ロバストな推定が可能になる場合もある．

5.2 f-ダイバージェンスの特徴付け

f-ダイバージェンス [2] は，確率密度関数 p, q に対して

$$D_f(p,q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx$$

で定義される．ここで f は強凸関数であり，$f(1) = 0$ を満たす．Jensen の不等式から非負性 $D_f(p,q) \geq 0$ が成り立ち，また f の強凸性から，$D_f(p,q) = 0$ なら $p = q$ が成り立つ．

f-ダイバージェンスも，Hölder スコアと同様にデータ変換に対する不変性によって特徴付けることができる．これについて以下で紹介する．詳細については [11] に説明がある．まず，確率密度関数 p, q に対して $D_U(p,q)$ を

$$D_U(p,q) = \int U(p(x), q(x)) dx$$

で定義される非負値関数とする．さらに $D_U(p,q) = 0$ なら $p = q$ が成り立つとする．データ x が，関数 τ により $y = \tau(x)$ に 1 対 1 変換されたとき，確率密度 $p(x)$ が $p_\tau(y)$ に変換されるとする．このとき D_U に対して不変性 $D_U(p,q) = D_U(p_\tau, q_\tau)$ を要請する．関数 τ の可微分性など適当な正則条件の下で，D_U はある f-ダイバージェンス D_f と等価，すなわち単調増加関数 ξ が存在して $\xi(D_U(p,q)) = D_f(p,q)$ となることを示すことができる．

f-ダイバージェンス D_f は任意の可微分な 1 対 1 変換に対して不変である．このため f-ダイバージェンスに基づく推定は，どのようなデータ変換に対しても共変的であると期待される．しかし一般には，Dirac のデルタ関数を用いて表現される経験分布 $\widetilde{p}$ を f-ダイバージェンスにを直接代入することはできないため，f-ダイバージェンスを推定に用いるためには工夫が必要となる．このため，本来 f-ダイバージェンスが持っている不変性は，推定量では失われてしまうと考えられる．一方，Hölder スコアは経験分布を直接代入できる形式であるため，合成スコアが持っている不変性を推定量が直接引き継ぐことになる．しかし Hölder スコアは一般の 1 対 1 変換に対して不変ではないため，アフィン

変換以外の変換を考える必要があるときには，推定量の共変性が成立せず，注意が必要である．

6 ロバスト推定への応用

Hölder スコアを統計的推定に応用する．データ $x_1, \ldots, x_n$ は，理想的な状況では分布 $p_0(x)$ から独立に得られるとする．しかし，データを観測する過程で外れ値などが混入し，実際の観測値は $p_{\varepsilon,z}(x) = (1-\varepsilon)p_0(x) + \varepsilon\delta(x-z)$ から得られたとする．ここで $\delta(x)$ は Dirac のデルタ関数であり，z が外れ値である．このように，外れ値などが混入したデータから目標となる $p_0(x)$ を推定するために，ロバスト推定量が用いられる．外れ値の比率 ε が非常に小さいとき，推定量のバイアスを評価することで，外れ値に対する推定量の頑健さを定量化することができる．

推定量のバイアスを測るために影響関数を定義する．統計モデル $p_\theta(x), \theta \in \Theta \subset \mathbb{R}^d$ を用いて，ターゲットの分布である $p_0(x)$ を推定する．ここで $p_0(x)$ は統計モデルに含まれ，$p_0(x) = p_{\theta_0}(x)$ が成り立つと仮定する．推定量 $\widehat{\theta}$ を統計的汎関数とみなして，分布 p からパラメータ $\theta \in \Theta$ への対応関係を $p \mapsto \widehat{\theta}(p) \in \Theta$ と記述する．実際の推定ではデータの経験分布 $\widetilde{p}(x)$ が得られるため，推定パラメータは $\widehat{\theta}(\widetilde{p}) \in \Theta$ と表せる．統計モデル p_θ の下での推定量 $\widehat{\theta}$ の一致性，すなわち $\widehat{\theta}(p_\theta) = \theta$ が任意の $\theta \in \Theta$ に対して成り立つことを仮定する．データの分布が $p_{\varepsilon,z}(x)$ のとき，推定量は $\widehat{\theta}(p_{\varepsilon,z})$ となる．これは，目標である θ_0 とは一般に一致しない．その差 $\widehat{\theta}(p_{\varepsilon,z}) - \theta_0 = \widehat{\theta}(p_{\varepsilon,z}) - \widehat{\theta}(p_{\theta_0})$ を推定量 $\theta(p)$ の分布 $p_{\varepsilon,z}$ の下での (パラメータ θ_0 における) バイアスとよぶ．推定量 p の影響関数 $\mathrm{IF}(z;\theta,S)$ を，バイアスの極限

$$\mathrm{IF}(z,\theta_0;\widehat{\theta}) = \lim_{\varepsilon \searrow 0} \frac{\widehat{\theta}(p_{\varepsilon,z}) - \widehat{\theta}(p_0)}{\varepsilon}$$

により定義する [6]．影響関数は，数学的には汎関数 $\widehat{\theta}$ のガトー微分である．定義より，

$$\widehat{\theta}(p_{\varepsilon,z}) = \theta_0 + \varepsilon \cdot \mathrm{IF}(z,\theta_0;\widehat{\theta}) + o(\varepsilon)$$

となるので，影響関数は外れ値 z に対する推定量 $\widehat{\theta}$ の感度を表している．

影響関数から，推定量のロバスト性を測るための規準がいくつか提案されている．例えば影響関数のノルムを外れ値に関して最悪評価した gross error sensitivity $\sup_z \|\mathrm{IF}(z,\theta_0;\widehat{\theta})\|$ などがある [6]．ここでは再下降性 (redescending property)[6, 9]

$$\forall \theta_0 \in \Theta, \quad \lim_{\|z\| \to \infty} \|\mathrm{IF}(z,\theta_0;\widehat{\theta})\| = 0$$

について考える．再下降性は，あまりにも大きな外れ値は自動的に無視される，という性質であり，これは実データ解析において有用と考えられる．正規分布モデルで期待値を推

定するとき，擬球スコアは再下降性をもつが，density-power スコアはもたないことが分かっている [1, 3].

Hölder スコアから定義される推定量のロバスト性について考える.

定理 5. 正則条件の下で，以下は同値である.

1. 関数 ϕ とパラメータ $\gamma > 0$ をもつ *Hölder* スコアから定義される推定量が正則条件を満たす任意の統計モデルに対して再下降性をもつ.
2. 等式 $\phi''(1) = -\gamma(1+\gamma)$ が成り立つ.

詳細な証明は [8] にある．擬球スコアは $\phi''(1) = -\gamma(1+\gamma)$ を満たすが，density-power スコアは $\phi''(1) = 0$ となり条件を満たさない．また Bregman-Hölder スコアを等価な Hölder スコアとして表現したとき，$\phi''(1) = -\gamma(1+\gamma) + (\kappa - 1)(1+\gamma)$ となるので，$\kappa = 1$，すなわち擬球スコアのときのみ，再下降性をもつことが分かる.

7 考察

本稿では，統計的推論のための損失として Hölder スコアを導入した．Hölder スコアは，Bregman スコアや局所スコア [10] などとは異なるクラスの合成スコアである．また Hölder スコアは，アフィン変換に対する不変性という性質によって特徴付けられることを示した．さらに，Hölder スコアをロバスト推定に用いたとき，再下降性をもつ推定量のクラスをについて考察した.

本稿では，特に (2) で表せる合成スコアに対して不変性を仮定し，Hölder スコアを導出した．より一般の合成スコアに対して本稿の結果を拡張することは重要な課題である．また確率密度関数や非負値関数だけでなく，行列や作用素に対するダイバージェンスへと拡張することも，応用上重要な課題となっている.

参考文献

[1] 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.

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

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

[4] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and

estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[5] I. J. Good. Comment on "measuring information and uncertainty," by R. J. Buehler. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, page 337339, Toronto: Holt, Rinehart and Winston, 1971.

[6] F. R. Hampel, P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel. *Robust Statistics. The Approach based on Influence Functions.* John Wiley and Sons, Inc., 1986.

[7] A. D. Hendrickson and R. J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42:19161921, 1971.

[8] T. Kanamori and H. Fujisawa. Affine invariant divergences associated with composite scores and its applications. *Bernoulli*, to appear.

[9] R. Maronna, R.D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods.* Wiley, 2006.

[10] M. Parry, A. P. Dawid, and S. Lauritzen. Proper local scoring rules. *Annals of Statistics*, 40:561–592, 2012.

[11] Yu Qiao and N. Minematsu. A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, pages 3884–3890, 2010.

[12] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.