τ-情報幾何学への誘(いざな)い

田中　勝

ver.1.0

概要

これは，コロナ社から刊行されている「エントロピーの幾何学」で紹介した τ-情報幾何学について，より直感的な導入を与えたものである。また，別に用意した「図解版 τ-情報幾何学」と合わせて読まれることで，τ-情報幾何学がどういうものか理解しやすくなることを期待して作成したものでもある。「図解版 τ-情報幾何学」の方には，敢えて説明をつけることはしなかったので，演習問題のつもりで各図の説明を試みてもらいたい。もちろん解答は，この小論と「エントロピーの幾何学」を読んでいただければわかるものと思う。

τ-情報幾何学では，最初に双対アファイン空間を構成してから幾何学的量を導入していく。したがって，必然的に双対平坦な状況が実現されることになる[*1]。さらに，一般化ピタゴラスの定理を使うだけでよいのなら，ここで与えた「case 0 と case 1 の組合せ」が「甘利流の情報幾何」または「江口流の情報幾何」であると考えても差し支えない。

1　2 つの確率分布の違いを測る

任意の 2 つの確率分布 $p_1(x)$ と $p_2(x)$ の違いを知りたい。どうしよう？

1.1　case 0: $p_2(x) = p_1(x) + u(x)$

差をとってみよう。このとき，$p_2(x) - p_1(x)$ の差が確率変数の実現値すべてについて出てくるので，総和を取ると，

$$\sum_{x\in\mathcal{X}} u(x) = \sum_{x\in\mathcal{X}} (p_2(x) - p_1(x)) = \sum_{x\in\mathcal{X}} p_2(x) - \sum_{x\in\mathcal{X}} p_1(x) = 1 - 1 = 0 \qquad (1)$$

[*1] アファイン空間は平坦である。

となり差が無いことになってしまう。

別の視点からもう一度確認してみよう。そのために，一般の確率分布ではなく確率分布がパラメータの組みと 1 対 1 に対応するようなものを考える。つまり，そのパラメータの組みを確率分布の座標とみなそう。

このとき，$p_2(x)$ を $p_1(x)$ の周りで展開したときの近似誤差として 2 つの確率分布の違いを捉えることにしてみよう。つまり，

$$p_2(x) - p_1(x) = \frac{\partial p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} \frac{\partial^2 p_1}{\partial \theta^i \partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots \tag{2}$$

を考える。ただし，同じ添え字（$i, j, \ldots$）が上下に繰り返し現れたときには和の記号が省略されていることに注意する。

確率変数のすべての実現値についてこの近似誤差の和を取ると，パラメータによる微分と和を取る操作が通常は交換するために，例えば，

$$\sum_{x \in \mathcal{X}} \frac{\partial p_1}{\partial \theta^i} \mathrm{d}\theta^i = \frac{\partial}{\partial \theta^i} \left(\sum_{x \in \mathcal{X}} p_1 \right) \mathrm{d}\theta^i = \frac{\partial 1}{\partial \theta^i} \mathrm{d}\theta^i = 0 \tag{3}$$

となり，やはり当然ながら差が無いことになってしまう。

そこで，2 つの分布の違いを評価するためには，他の方法を考える必要がある。

1.2 case 1: $p_2(x) = e^u p_1(x)$

2 つの確率分布の比で，その違いを評価してみることにしょう。しかし，両辺の対数を取ると

$$\log p_2(x) - \log p_1(x) = u(x) \tag{4}$$

のように対数尤度の差になるので，比を考えるということは，やはり差を考えることになる。

ところが，今度は case 0 とは違い，以下のように確率変数のすべての実現値に対する対数尤度の差の和が 0 になるとは限らない。

$$\sum_{x \in \mathcal{X}} u(x) = \sum_{x \in \mathcal{X}} \left(\log p_2(x) - \log p_1(x) \right) = \sum_{x \in \mathcal{X}} \log \frac{p_2(x)}{p_1(x)} \tag{5}$$

これは一見，良さそうに見えるが，今度は $\log \dfrac{p_2(x)}{p_1(x)}$ の値が正とは限らないため，せっかく得られた差が打ち消されてしまうかもしれない。

そこで，case 0 のときと同様に別の視点からもう一度確認してみよう。つまり座標系を導入して考えてみよう[*2]。このとき，

$$\log p_2(x) - \log p_1(x) = \frac{\partial \log p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} \frac{\partial^2 \log p_1}{\partial \theta^i \partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots \tag{6}$$

ここでも，同じ添え字（$i, j, \ldots$）が上下に繰り返し現れたときには和の記号が省略されていることに注意する。

もちろん，パラメータによる微分と和を取る操作が通常は交換するために，例えば，

$$\sum_{x \in \mathcal{X}} \frac{\partial \log p_1}{\partial \theta^i} \mathrm{d}\theta^i = \frac{\partial}{\partial \theta^i} \left(\sum_{x \in \mathcal{X}} \log p_1 \right) \mathrm{d}\theta^i \tag{7}$$

のようになる。しかし，ここに現れる $\sum_{x \in \mathcal{X}} \log p_1(x)$ が有限の値をとる保証が無いため，対数尤度の差が有限の値[*3]になるとは限らない。つまり，高々 1 次か 2 次で近似しようとしているので，ある程度近いという期待をしているにも関わらず，得られる値が期待通り小さくなる補償がない。

結局，case 0 も case 1 もうまくいかない。そこで case 0 と case 1 を組み合わせることを考えてみよう。

2　case 0 と case 1 の組合せ：双対アファイン空間の導入

まずは，座標系を導入した後の違いを並べて見てみよう。確率分布の差については

$$p_2(x) - p_1(x) = \frac{\partial p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} \frac{\partial^2 p_1}{\partial \theta^i \partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots \tag{8}$$

のようになり，確率分布の比については対数尤度の差として

$$\log p_2(x) - \log p_1(x) = \frac{\partial \log p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} \frac{\partial^2 \log p_1}{\partial \theta^i \partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots \tag{9}$$

のようになる。

さて，それぞれの 1 次の項を組合せてみよう[*4]。

[*2] ここでは，パラメータに関して 1 次か 2 次の近似で十分な場合を考えたい。

[*3] もし，$p_1(x) = 0$ となることがあれば負の無限大に発散してしまう。

[*4] このことを $S1B1$ と表すことにしよう。後でわかるように，S は $Soul$ の S のことであり，B は $Body$ の B のことである。

$$\sum_{x\in\mathcal{X}}\left(\frac{\partial p_1}{\partial\theta^i}\right)\left(\frac{\partial\log p_1}{\partial\theta^j}\right)\mathrm{d}\theta^i\mathrm{d}\theta^j=\sum_{x\in\mathcal{X}}p_1\frac{1}{p_1}\frac{\partial p_1}{\partial\theta^i}\frac{\partial\log p_1}{\partial\theta^j}\mathrm{d}\theta^i\mathrm{d}\theta^j$$
$$=\sum_{x\in\mathcal{X}}p_1\frac{\partial\log p_1}{\partial\theta^i}\frac{\partial\log p_1}{\partial\theta^j}\mathrm{d}\theta^i\mathrm{d}\theta^j=g_{ij}\mathrm{d}\theta^i\mathrm{d}\theta^j \qquad (10)$$

今度は，なんとフィッシャー情報行列が出てきた。つまり，$S1B1$ はフィッシャー情報行列を与えることがわかった。

これまでのことを整理してみよう。まず，case 0 も case 1 もそれぞれアファイン空間になっている。ここで，アファイン空間とは，次のようなものである。

• 空間 A の要素とベクトル空間 U の要素との間に，以下のような “平行移動” $\dashv\!\!+$: $A\times U\to A:(a,u)\mapsto a\dashv\!\!+u$ が定義された 3 つ組 $(A,U,\dashv\!\!+)$ または単に A のことをアファイン空間という：

(1) $(a\dashv\!\!+u_1)\dashv\!\!+u_2=a\dashv\!\!+(u_1+u_2)$（繰り返された平行移動は，1 回の平行移動で表すことができる。）

(2) 任意の A の要素 a_1,a_2 に対して，$a_2=a_1\dashv\!\!+u$ となるような U の要素 u が一意に存在する（A のどの要素からも平行移動で好きな A の要素へ移動できる。）

確認してみよう：

• case 0 の場合は，A として確率分布からなる空間 P，U として $\sum_{x\in\mathcal{X}}u=0$ となる要素で構成されるベクトル空間，$\dashv\!\!+$ として通常の足し算 $+$ を取ることで，$(p+u_1)+u_2=p+(u_1+u_2)$ が成り立ち，$p_2=p_1+u$ のとき，$u=p_2-p_1$ と一意に決めることができるので，P はアファイン空間である。そこで，このアファイン空間を P^0 と表すことにする。

• case 1 の場合は，A として確率分布からなる空間 P を拡張して $\sum_{x\in\mathcal{X}}p(x)<\infty$ であるような要素から構成される $\tilde{P}$ を選び，U として適当な関数からなる空間を選び，$\dashv\!\!+$ として $e^u\cdot$ のような演算を選ぶと，$e^{u_2}\cdot(e^{u_1}\cdot p)=e^{u_1+u_2}\cdot p$ が成り立ち，また $p_2=e^u\cdot p_1$ のとき，$u=\log p_2-\log p_1$ となるような要素が一意に存在するので，この拡張された $\tilde{P}$ はアファイン空間である[*5]。そこで，このアファイン空間を P^1 と表すことにする。この P^1 から P を得るためには $\sum_{x\in\mathcal{X}}p(x)<\infty$ なので，その大きさで割るだけでよい。

通常，気安く対数尤度などを考えるが，実はそれがアファイン空間を構成するためのベクトル空間の要素であるという立場からは，確率分布の空間 P を大きさが有限であるよ

[*5] 測度が零になるような箇所での振る舞いを一つ定めているものとする。

うな非負の可測関数からなる空間 $\tilde{P}$ へ拡張すると扱いやすくなる[*6]。また，これは指数型分布族の場合には十分統計量の定義とも関わっている。そもそも，case 0 で任意の 2 つの確率分布の差が 0 になるのは確率分布である（大きさが 1）ということから起こる問題である。

また，フィッシャー情報行列は，対数尤度の 1 次の差として $\tilde{P}$ で得ることができなかった確率分布の差を P に射影（$S1B1$）することで捉えている[*7]。

つまり，これら 2 つのアファイン空間 P^0 と P^1 を組合せることで，フィッシャー情報行列（$S1B1$）が 2 つの確率分布の違いとして現れてきた，ということである。この際にポイントとなるのは，平行移動の定義である。平行移動の定義が変わっても 2 つの確率分布の違い（フィッシャー情報行列（$S1B1$））は変化しないという要請をすれば，どのようなことが可能になるのだろうか。

3 case s（$Body$ と $Soul$）：τ-情報幾何学への入り口

ここでは，確率分布の空間 P を拡張した $\tilde{P}$ で考えていく[*8]。平行移動 $\dashv$ を

$$\left\{1+(1-s)\,u\right\}^{\frac{1}{1-s}} \otimes_s \tag{11}$$

に取る。ここで，$u \in U$ であり[*9]，

$$f \otimes_s g = \left\{f^{1-s}+g^{1-s}-1\right\}^{\frac{1}{1-s}} \tag{12}$$

である。これらは，次のような性質を持っている：

$$\lim_{s\to 0}\left\{1+(1-s)\,u\right\}^{\frac{1}{1-s}} = 1+u \tag{13}$$

$$\lim_{s\to 0} f \otimes_s g = f+g-1 \tag{14}$$

$$\lim_{s\to 1}\left\{1+(1-s)\,u\right\}^{\frac{1}{1-s}} = e^u \tag{15}$$

$$\lim_{s\to 1} f \otimes_s g = fg \tag{16}$$

[*6] もちろん，平行移動の度に規格化を保つように（確率になるように）調整してもよい（P^0）が，平行移動が済んだ後に必要になってから規格化する（確率に戻す）ことにしてもよい（P^1）。

[*7] この射影は逆向きでもよい。つまり，P で得ることができなかった確率分布の違いを $\tilde{P}$ に射影することで捉えているとみてもよい。

[*8] これを正錐（positive cone）という。

[*9] ベクトル空間 U は平行移動のパラメータ s に応じて適切なものが選択される必要がある。

平行移動の仕方を決定しているパラメータ s の値で区別されるようなアファイン空間 P^s を一つ選ぶと，選ばれたアファイン空間 P^s の要素は互いに

$$p_2 = \{1 + (1-s)\,u\}^{\frac{1}{1-s}} \otimes_s p_1 \tag{17}$$

のような関係で繋がっている。

このように定義しておくと，例えば $s \to 0$ のとき，式 (13) と (14) を用いると

$$p_2 = (1+u) + p_1 - 1 = p_1 + u \tag{18}$$

のようになり，P^0 での状況を再現する。

また，$s \to 1$ のとき，式 (15) と (16) を用いると

$$p_2 = e^u p_1 \tag{19}$$

のようになり，P^1 での状況を再現する。

したがって，先ほどの case 0 と case 1 はここで定義されたアファイン空間の特殊な場合ということになる。

さて，p_2 を p_1 で近似したときの状況が知りたい。そのために τ-対数尤度を考えると

$$\ln_s p_2 = \ln_s p_1 + u \tag{20}$$

のようになる。ただし，

$$\ln_s f = \frac{1}{1-s}\left(f^{1-s} - 1\right) \tag{21}$$

である。また，

$$\lim_{s\to 1} \ln_s f = \log f \tag{22}$$

である。

このとき，

$$\begin{aligned}
&\ln_s p_2 - \ln_s p_1 \\
&= \frac{\partial \ln_s p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2}\frac{\partial^2 \ln_s p_1}{\partial \theta^i \partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots \\
&= p_1^{1-s} \frac{\partial \log p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} p_1^{1-s} \left(\frac{\partial^2 \log p_1}{\partial \theta^i \partial \theta^j} + (1-s)\, \frac{\partial \log p_1}{\partial \theta^i} \frac{\partial \log p_1}{\partial \theta^j} \right) \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots
\end{aligned} \tag{23}$$

ここで，

$$\frac{1}{p_1}\frac{\partial^2 p_1}{\partial \theta^i \partial \theta^j} = \frac{\partial^2 \log p_1}{\partial \theta^i \partial \theta^j} + \frac{\partial \log p_1}{\partial \theta^i}\frac{\partial \log p_1}{\partial \theta^j} \tag{24}$$

であることに注意すると，$s \to 0$ のとき，

$$(p_2 - 1) - (p_1 - 1) = p_2 - p_1 = \frac{\partial p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} \frac{\partial^2 p_1}{\partial \theta^i \partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots \tag{25}$$

を再現する。

また，$s \to 1$ のとき，

$$\log p_2 - \log p_1 = \frac{\partial \log p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} \frac{\partial^2 \log p_1}{\partial \theta^i \partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots \tag{26}$$

が再現される。

P^1 の相棒が P^0 だったので，P^s の相棒を P^{1-s} に選んでみよう[*10]。このとき，

$$\begin{aligned}
&\ln_{1-s} p_2 - \ln_{1-s} p_1 \\
&= \frac{\partial \ln_{1-s} p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} \frac{\partial^2 \ln_{1-s} p_1}{\partial \theta^i \partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots \\
&= p_1^s \frac{\partial \log p_1}{\partial \theta^i} \mathrm{d}\theta^i + \frac{1}{2} p_1^s \left(\frac{\partial^2 \log p_1}{\partial \theta^i \partial \theta^j} + s \frac{\partial \log p_1}{\partial \theta^i} \frac{\partial \log p_1}{\partial \theta^j} \right) \mathrm{d}\theta^i \mathrm{d}\theta^j + \cdots
\end{aligned} \tag{27}$$

のようになる。

すぐにわかるように，P^s 側と P^{1-s} 側の $\mathrm{d}\theta^i$ の 1 次の項を組合せる（$S1B1$）と

$$\begin{aligned}
&\sum_{x \in \mathcal{X}} \left(p_1^s \frac{\partial \log p_1}{\partial \theta^i} \right) \left(p_1^{1-s} \frac{\partial \log p_1}{\partial \theta^j} \right) \mathrm{d}\theta^i \mathrm{d}\theta^j \\
&= \sum_{x \in \mathcal{X}} p_1 \frac{\partial \log p_1}{\partial \theta^i} \frac{\partial \log p_1}{\partial \theta^j} \mathrm{d}\theta^i \mathrm{d}\theta^j \\
&= g_{ij} \mathrm{d}\theta^i \mathrm{d}\theta^j
\end{aligned} \tag{28}$$

が得られ，フィッシャー計量が平行移動のパラメータ s に依らずに不変な量になっていることが確認できる。

そこで，P^s と P^{1-s} の組みを τ-アファイン構造（双対アファイン構造）と呼ぶことにする。実は，このような双対性の下で情報幾何学を展開するのが τ-情報幾何学である。

ちなみに，τ-情報幾何学では，エントロピーは $Body$ と $Soul$ の対応する原点（始点）の組合せ（$S0B0$）で求めることができる。しかし，発散する項が現れるのでちょっとした工夫（くり込み）が必要になる。また，ダイバージェンスは，$Body$（P^s）側で p_2 を u で 1 次近似したときの誤差，つまり 2 次以上の誤差の期待値を取ることで得られる。

[*10] P^s を $Body$ と呼び，P^{1-s} を $Soul$ と呼ぶ。

実際，p_2 を u で直接展開することで，

$$p_2 = \left\{p_1^{1-s} + (1-s)\,u\right\}^{\frac{1}{1-s}} = p_1 + p_1^s u + F^{(2)}(x) \tag{29}$$

となるので，$u = \ln_s p_2 - \ln_s p_1$ を利用して $F^{(2)}(x)$ を評価すると，

$$F^{(2)}(x) = (p_2 - p_1) - p_1^s \left(\ln_s p_2 - \ln_s p_1\right) \tag{30}$$

が得られる。これに $\dfrac{1}{s}$ を掛けて確率変数の取り得るすべての値について和を取ると

$$\begin{aligned}\sum_{x\in\mathcal{X}} \frac{1}{s} F^{(2)}(x) &= \frac{1}{s}\sum_{x\in\mathcal{X}} \left\{p_2 - p_1 - p_1^s \left(\ln_s p_2 - \ln_s p_1\right)\right\} \\ &= \frac{1}{s\,(1-s)} \sum_{x\in\mathcal{X}} \left\{(1-s)\,p_2 + s p_1 - p_1^s p_2^{1-s}\right\}\end{aligned} \tag{31}$$

のようになる。これがダイバージェンスである。

甘利または江口流の情報幾何学で有名な双対接続は，実は，$Body$（P^s）側での $\mathrm{d}\theta^i$ の 2 次の項と $Soul$（P^{1-s}）側での $\mathrm{d}\theta^i$ の 1 次の項の組合せ（$S1B2$），または $Body$ 側での $\mathrm{d}\theta^i$ の 1 次の項と $Soul$ 側での $\mathrm{d}\theta^i$ の 2 次の項の組合せ（$S2B1$）により簡単に得ることができる。このとき，$S1B2$ の組合せは甘利または江口流での $\alpha = 1$ の場合の接続に対応しており，$S2B1$ の組合せは $\alpha = -1$ の場合の接続に対応している。

以上で，この「τ-情報幾何学への誘い」を終えることにする。この続きはコロナ社より刊行された「エントロピーの幾何学」で是非お楽しみ頂きたい。