ダイバージェンスから導かれる双対構造

いざな (@u50124n4)

2025/12/20

本記事は総研大 統計科学コース Advent Calendar 2025 の 20 日目の記事です.

1. 情報幾何学とは

情報幾何学とは，確率分布の族を多様体とみなし，その上で微分幾何学をするものです. 統計的な推論を幾何学の言葉を使って言い表すことができます（例えばつい最近書いた「最大エントロピー原理の幾何学的解釈、そして情報幾何学へ」という記事（link）をご参照ください).

微分幾何学の数学的な基本構成要素は

- 多様体 $\mathcal{M}$,
- Riemann 計量 g,

です. g から Levi-Civita 接続 $\nabla^{(0)}$ が一意に定まります. Levi-Civita 接続とは捩率 0，計量的な affine 接続のことです. 情報幾何学では「計量的」という条件を外した affine 接続 ∇ を考え，条件を外す代わりにそれと双対的な接続 ∇^* とセットで考えるということをします:

$$Xg(Y,Z) \equiv g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z). \tag{1}$$

すなわち，情報幾何学とは $(\mathcal{M}, g, \nabla, \nabla^*)$ の 4 つ組が織り成す世界を調べる学問です.

$\mathcal{M}, g, \nabla$ の 3 つが決まれば双対接続は一意に決まり，したがって情報幾何学の土台ができます. この 3 つのうち∇ を直接定めるのでなく，対称(0,3)テンソル C を用いて,

$$g\left(\nabla^{(\alpha)}_X Y, Z\right) := g\left(\nabla^{(0)}_X Y, Z\right) - \frac{\alpha}{2} C(X,Y,Z) \quad (\alpha \in \mathbb{R}) \tag{2}$$

のようにして affine 接続 $\nabla^{(\alpha)}$ を作ることもできます. $\nabla^{(\alpha)}$ は捩率 0 で，更にその双対接続は $\nabla^{(-\alpha)}$ なことが簡単な計算からわかります.

このアプローチとして最も有名なのは多様体をパラメータ θ をもつ分布族 $\mathcal{M} = \{p_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$ としたとき，計量 g として Fisher 計量

$$g_\theta(X,Y) = \mathbb{E}_\theta[(Xl_\theta)(Yl_\theta)] = \int (X \log p_\theta(x))(Y \log p_\theta(x)) p_\theta(x)\,\mathrm{d}x, \tag{3}$$

対称テンソル C として Amari-Chentsov テンソル

$$C_\theta(X,Y,Z) := \mathbb{E}_\theta[(Xl_\theta)(Yl_\theta)(Zl_\theta)] \tag{4}$$

を用いるものです. これらから作られる $\nabla^{(\alpha)}$ のうち，特別に $\alpha = 1$ のときを e-接続 $\nabla^{(e)}$，$\alpha = -1$ のときを m-接続 $\nabla^{(m)}$ と呼びます. 計算すると $\nabla^{(\alpha)}$ についての Christoffel 記号は

$$\Gamma^{(\alpha)}{}_{ijk}(\theta) = \frac{1-\alpha}{2}(C_\theta)_{ijk} + \mathbb{E}_\theta\left[(\partial_i \partial_j l_\theta)(\partial_k l_\theta)\right] \tag{5}$$

で，e-接続，m-接続についてはそれぞれ

$$\Gamma^{(e)}{}_{ijk}(\theta) = \mathbb{E}_\theta\left[(\partial_i \partial_j l_\theta)(\partial_k l_\theta)\right], \quad \Gamma^{(m)}{}_{ijk}(\theta) = \mathbb{E}_\theta\left[\frac{1}{p_\theta}(\partial_i \partial_j p_\theta)(\partial_k l_\theta)\right] \tag{6}$$

なことが（更なる計算により）わかります．

$\alpha = \pm 1$ が特別視されるのは，特に $\mathcal{M}$ として指数型分布族

$$p(x;\theta) = \exp\left[\sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)\right] \tag{7}$$

を考えたとき，$\nabla^{(e)}, \nabla^{(m)}$ についての曲率が 0 になる[1]からです．曲率が 0 ということは，$\nabla^{(e)}, \nabla^{(m)}$ という視点から見ると指数方分布族は“まっすぐ”な空間であることを意味していて，このとき affine 座標系と呼ばれる，その座標系でみると Christoffel 記号がどこでも 0 になるような座標系が取れることが示せます[2]．特に $\nabla^{(e)}$ についての affine 座標系は θ そのもので，この座標系で見ると例えば測地線は文字通り直線になります[3]:

$$\theta(c(t)) = \theta(p) + tv \quad (c(0) = p,\ v \in \mathbb{R}^n) \tag{8}$$

他にも色々と望ましい性質を持つことがわかりますが，ここでは深入りしません．

さて，$\mathcal{M}, g, \nabla$ の 3 つを決めるもう一つの方法として，ダイバージェンスという距離のような性質を持った関数 $\mathcal{D} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ を用いたものがあります．本記事では，この方法を紹介したいと思います．

2. ダイバージェンスから導かれる双対構造

Definition 1

$\mathcal{M}$ を多様体とする．滑らかな関数 $\mathcal{D} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ が $\mathcal{M}$ 上の**ダイバージェンス**であるとは，

1. （非負性）$\mathcal{D}(p\|q) \geq 0$，等号は $p = q$ のとき，またそのときに限る．
2. 任意の $p \in M$ と，そのまわりのチャート $(U, \{x^i\})$ について，ある正定値対称行列 $\left(g_{ij}(p)\right)_{ij}$ が存在し，p の近傍の任意の点 $q \in U$（$x(q) = x(p) + \Delta x$）について

$$\mathcal{D}(p\|q) = \frac{1}{2} \sum_{i,j} g_{ij}(p)(\Delta x)^i (\Delta x)^j + o(\|\Delta x\|^2). \tag{9}$$

$p \in M$ まわりのチャートを $(U, \{x^i\})$ とし，$\mathcal{M} \times \mathcal{M}$ のチャートを

$$(U \times U, x^1, ..., x^n, x^1, ..., x^n) \tag{10}$$

のようにとります．順番がわかるように，後者の座標系にはダッシュをつけて表すことにします:

$$(U \times U, x^1, ..., x^n, x'^1, ..., x'^n). \tag{11}$$

また，次のような記法を用います:

$$\partial_i \mathcal{D}(p\|p) := \left.\frac{\partial}{\partial x^i} \mathcal{D}\right|_{(p,p)}, \quad \partial'_i \mathcal{D}(p\|p) := \left.\frac{\partial}{\partial x'^i} \mathcal{D}\right|_{(p,p)}. \tag{12}$$

定義の条件 2 から，$\partial_i \mathcal{D}(p\|p) = \partial'_i \mathcal{D}(p\|p) = 0$ がすぐにわかります．更に次が成り立ちます:

[1]一般論として，接続 ∇ について曲率 0 なら，その双対接続 ∇^* についても曲率 0 であることが示せます．

[2]Euclid 空間における直交座標系もそうでした．この意味で，affine 座標系とは，局所的に Euclid 空間的な性質をもったものだと思えます．

[3]測地線方程式が $\frac{\mathrm{d}^2\theta^i}{\mathrm{d}t^2} = 0$ という形になるからです．

Proposition 2

$$g_{ij}(p) = \partial'_i\partial'_j\mathcal{D}(p\|p) = \partial_i\partial_j\mathcal{D}(p\|p) = -\partial_i\partial'_j\mathcal{D}(p\|p) \tag{13}$$

Proof: $\iota : \mathcal{M} \to \mathcal{M} \times \mathcal{M};\ p \mapsto (p,p)$ とする．

$$0 = \partial'_j\mathcal{D}(p\|p) = \left(\partial'_j\mathcal{D} \circ \iota\right)(p) \tag{14}$$

である．これが任意の p で成り立つので，$\partial'_j\mathcal{D} \circ \iota = 0$．両辺を x^i で微分すると，連鎖律より

$$0 = \sum_k \left(\partial_k\partial'_j\mathcal{D}(p\|p)\right)\left(\partial_i\iota^k(p)\right) + \sum_k \left(\partial'_k\partial'_j\mathcal{D}(p\|p)\right)\left(\partial_i\iota^k(p)\right) = \partial_i\partial'_j\mathcal{D}(p\|p) + \partial'_i\partial'_j\mathcal{D}(p\|p). \tag{15}$$

したがって

$$\partial'_i\partial'_j\mathcal{D}(p\|p) = -\partial_i\partial'_j\mathcal{D}(p\|p). \tag{16}$$

もう一つの等号も同様にしてわかる． □

Proposition 3

$$g_{ij}(p) := \partial'_i\partial'_j\mathcal{D}(p\|p) = \partial_i\partial_j\mathcal{D}(p\|p) = -\partial_i\partial'_j\mathcal{D}(p\|p) \tag{17}$$

によって $\mathcal{M}$ 上の Riemann 計量 g，

$$\Gamma_{ijk}(p) = -\partial_i\partial_j\partial'_k\mathcal{D}(p\|p), \quad {\Gamma^*}_{ijk}(p) = -\partial'_i\partial'_j\partial_k\mathcal{D}(p\|p) \tag{18}$$

によって affine 接続 ∇，∇^* がそれぞれ定まる．∇, ∇^* は双対接続で，どちらも捩率 0.

Proof: 計量，接続を定めることは，座標に依らないこと（適切な座標変換則に従うこと）を確認すれば良いが，これは単純な計算からわかる．∇, ∇^* の捩率が 0 であることは i, j の対称性から直ちに従う．双対接続になっていることは，前命題の証明と同様に $\iota(p) = (p,p)$ を用いて

$$\begin{aligned}\partial_i g\left(\partial_j, \partial_k\right)|_p &= -\partial_i\left(\partial_j\partial'_k\mathcal{D} \circ \iota\right)|_p = -\partial_i\partial_j\partial'_k\mathcal{D}(p\|p) - \partial'_i\partial_j\partial'_k\mathcal{D}(p\|p) \\ &= \Gamma_{ijk}(p) + {\Gamma^*}_{ikj}(p) = g\left(\nabla_{\partial_i}\partial_j, \partial_k\right)(p) + g\left(\partial_j, \nabla^*_{\partial_i}\partial_k\right)(p)\end{aligned} \tag{19}$$

より． □

これで，多様体 $\mathcal{M}$ とその上のダイバージェンス $\mathcal{D} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$ を決めれば，計量 g，affine 接続とその双対 ∇, ∇^* が自然と定まることがわかりました．計量はダイバージェンスの Taylor 展開の 2 次，接続は 3 次の項と見ることもできて，計量は接ベクトルの長さ/角度などをみるもの，接続はもう一歩踏み込んで接ベクトルの変化率をみるもの，という直感的な解釈とも整合しています．

3. 具体例

パラメトライズされた分布族 $\mathcal{M} = \{p_\theta : \mathcal{X} \to \mathbb{R}_{>0} \mid \theta \in \Theta \subset \mathbb{R}^d\}$ を考えます．Θ は開集合とし，条件として

- 任意の $\theta \in \Theta, x \in \mathcal{X}$ で $p_\theta(x) > 0$,
- 積分と微分は順序交換可能．例えば

$$\int \frac{\partial}{\partial\theta^i} p_\theta(x)\,\mathrm{d}x = \frac{\partial}{\partial\theta^i}\int p_\theta(x)\,\mathrm{d}x. \tag{20}$$

- $x \in \mathcal{X}$ を固定するごとに $\theta \mapsto p_\theta(x)$ は全単射（すなわち，$\mathcal{M}$ は識別可能である）

この条件のもとで，$\mathcal{M}$ を θ を座標系とする多様体とみなします．今後，$p_\theta, p_{\theta'} \in \mathcal{M}$ の間のダイバージェンス $\mathcal{D}(p_\theta \| p_{\theta'})$ を $\mathcal{D}(\theta : \theta')$ などと表すことにします．

3.1. KL ダイバージェンス

Definition 4

KL ダイバージェンスとは

$$\mathcal{D}_{\mathrm{KL}}(\theta : \theta') := \int p_\theta(x) \log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right) \mathrm{d}x. \tag{21}$$

色々なところで見かける量ですが，例えばエントロピーとの関係は前掲記事の中核となるものでした．

Proposition 5

KL ダイバージェンスから導かれる

- 計量は Fisher 計量，
- 接続は e-接続/m-接続．

Proof:

$$\begin{aligned} g_{ij} = \partial_i' \partial_j' \mathcal{D}_{\mathrm{KL}}(\theta : \theta) &= -\int p_\theta (\partial_i \partial_j \log p_\theta) \,\mathrm{d}x = -\int p_\theta \left(-\frac{(\partial_i p_\theta)(\partial_j p_\theta)}{p_\theta^2} + \frac{\partial_i \partial_j p_\theta}{p_\theta} \right) \mathrm{d}x \\ &= \int p_\theta (\partial_i l_\theta)(\partial_j l_\theta) \,\mathrm{d}x - \underbrace{\int \partial_i \partial_j p_\theta \,\mathrm{d}x}_{=0} = \mathbb{E}_\theta \left[(\partial_i l_\theta)(\partial_j l_\theta) \right], \end{aligned} \tag{22}$$

$$\Gamma_{ijk}(\theta) = -\partial_i \partial_j \partial_k' \mathcal{D}_{\mathrm{KL}}(\theta : \theta) = \int (\partial_i \partial_j p_\theta)(\partial_k l_\theta) \,\mathrm{d}x = \mathbb{E}_\theta \left[\frac{1}{p_\theta} (\partial_i \partial_j p_\theta)(\partial_k l_\theta) \right] = \Gamma^{(m)}{}_{ijk}(\theta), \tag{23}$$

$$\Gamma^*{}_{ijk}(\theta) = -\partial_i' \partial_j' \partial_k \mathcal{D}_{\mathrm{KL}}(\theta : \theta) = \int (\partial_k p_\theta)(\partial_i \partial_j l_\theta) \,\mathrm{d}x = \mathbb{E}_\theta \left[(\partial_i \partial_j l_\theta)(\partial_k l_\theta) \right] = \Gamma^{(e)}{}_{ijk}(\theta). \tag{24}$$

□

3.2. α ダイバージェンス

Definition 6

α ダイバージェンスとは，

$$\mathcal{D}_\alpha(\theta : \theta') := \frac{4}{1-\alpha^2}\left(1 - \int p_\theta^{(1-\alpha)/2} p_{\theta'}^{(1+\alpha)/2} \,\mathrm{d}x\right). \tag{25}$$

$\alpha = 0$ で Hellinger 距離

$$\mathcal{H}(\theta : \theta') = 1 - \int \sqrt{p_\theta(x) p_{\theta'}(x)} \,\mathrm{d}x \tag{26}$$

を定数倍したもの，$\alpha \to 1$ で KL ダイバージェンスの引数を逆にしたものに一致します:

$$\begin{aligned}\mathcal{D}_\alpha(\theta:\theta') &= \frac{4}{1-\alpha^2}\int p_{\theta'}\left(1-\left(\frac{p_\theta}{p_{\theta'}}\right)^{\frac{1-\alpha}{2}}\right)\mathrm{d}x = \frac{2}{1+\alpha}\int p_{\theta'}\frac{2}{1-\alpha}\left(1-\left(\frac{p_\theta}{p_{\theta'}}\right)^{\frac{1-\alpha}{2}}\right)\mathrm{d}x \\ &\overset{t=\frac{1-\alpha}{2}}{=} \frac{1}{1-t}\int p_{\theta'}\frac{1}{t}\left(1-\left(\frac{p_\theta}{p_{\theta'}}\right)^{t}\right)\mathrm{d}x \overset{t\to 0}{\to} -\int p_{\theta'}\log\left(\frac{p_\theta}{p_{\theta'}}\right)\mathrm{d}x = \mathcal{D}_{\mathrm{KL}}(\theta':\theta).\end{aligned} \tag{27}$$

同様の計算により $\alpha\to -1$ で KL ダイバージェンス $\mathcal{D}(\theta:\theta')$ に一致することがわかります.

> **Proposition 7**
>
> α ダイバージェンスから導かれる
>
> - 計量は Fisher 計量,
> - 接続は α 接続.

Proof:

$$\begin{aligned} g_{ij} &= \partial'_i\partial'_j\mathcal{D}_\alpha(\theta:\theta) \\ &= \frac{4}{\alpha^2-1}\int p_\theta^{(1-\alpha)/2}\left(\frac{\alpha+1}{2}(\partial_i\partial_j p_\theta)p_\theta^{(\alpha-1)/2} + \frac{\alpha^2-1}{4}(\partial_i p_\theta)(\partial_j p_\theta)p_\theta^{(\alpha-3)/2}\right)\mathrm{d}x \\ &= \int \frac{(\partial_i p_\theta)(\partial_j p_\theta)}{p_\theta}\,\mathrm{d}x = \mathbb{E}_\theta\left[(\partial_i l_\theta)(\partial_j l_\theta)\right],\end{aligned} \tag{28}$$

$$\begin{aligned} {\Gamma^*}_{ijk}(\theta) &= -\partial'_i\partial'_j\partial_k\mathcal{D}_\alpha(\theta:\theta) \\ &= \frac{4}{1-\alpha^2}\int\Bigg[\frac{1-\alpha}{2}(\partial_k p_\theta)p_\theta^{-(1+\alpha)/2}\times \\ &\qquad\qquad \left(\frac{\alpha+1}{2}(\partial_i\partial_j p_\theta)p_\theta^{(\alpha-1)/2} + \frac{\alpha^2-1}{4}(\partial_i p_\theta)(\partial_j p_\theta)p_\theta^{(\alpha-3)/2}\right)\mathrm{d}x\Bigg] \\ &= \int(\partial_i\partial_j p_\theta)\frac{\partial_k p_\theta}{p_\theta}\,\mathrm{d}x + \frac{\alpha-1}{2}\int\frac{(\partial_i p_\theta)(\partial_j p_\theta)(\partial_k p_\theta)}{p_\theta^2}\,\mathrm{d}x \\ &= \int\left(\partial_i\partial_j l_\theta + (\partial_i l_\theta)(\partial_j l_\theta)\right)(\partial_k l_\theta)p_\theta\,\mathrm{d}x + \frac{\alpha-1}{2}(C_\theta)_{ijk} \\ &= \mathbb{E}_\theta\left[(\partial_i\partial_j l_\theta)(\partial_k l_\theta)\right] + \frac{\alpha+1}{2}(C_\theta)_{ijk} = {\Gamma^{(-\alpha)}}_{ijk}(\theta).\end{aligned} \tag{29}$$

$\Gamma_{ijk}(\theta)$ についても同様の計算によって ${\Gamma^{(\alpha)}}_{ijk}(\theta)$ と一致することがわかる. □

4. 最後に

本記事ではダイバージェンスから双対構造が導かれること，その具体例として KL ダイバージェンスと α ダイバージェンスを取り上げました. $\mathcal{M}, \mathcal{D}:\mathcal{M}\times\mathcal{M}\to\mathbb{R}$ という 2 つ（空間である多様体は何はともあれ指定せざるを得ないので，実質 1 つ）さえ指定すれば情報幾何学を始められるというのは非常に嬉しそうです.

一方で，今度はダイバージェンスをどう選ぶか，という問題が生まれます. このアプローチとしてある不変性を満たすようなものを選ぶというものがあり，例えば[1]の第 4 章で紹介されています. 本記事で興味を持った方は，ぜひこちらをお読みください.

5. 参考文献

[1] 甘利俊一, 増補新版 情報幾何学の新展開, サイエンス社, 2025