振り返りと導入

前回までは指数型分布族について調べていた。本稿では次のことを行う:

- 統計モデルを定義する。
- 最尤推定量を定義する。
- Kullback-Leibler ダイバージェンスを定義する。

本稿では指数型分布族は open であるとする。

1 統計モデル

定義 1.1 (統計モデル)**.** $\mathcal{X}$ を可測空間とする。多様体 $\Theta \neq \varnothing$ と写像 $\mathbf{p}\colon \Theta \to \mathcal{P}(\mathcal{X})$ の組 $(\Theta, \mathbf{p})$ が $\mathcal{X}$ 上の**統計モデル (statistical model)** であるとは、ある**実現** (p, μ) が存在して次をみたすことをいう:

(i) $p\colon \Theta \times \mathcal{X} \to \mathbb{R}_{>0},\ (\theta, x) \mapsto p_\theta(x)$
  (i-a) $\forall \theta \in \Theta,\ \ p_\theta$ は可測関数
  (i-b) μ-a.e.$x \in \mathcal{X},\ \ p(\cdot, x)$ は C^∞ 関数

(ii) μ は σ-有限な測度である。

(iii) $\mathbf{p}(\theta) = p_\theta \cdot \mu$

命題 1.2. 指数型分布族 $\mathcal{P}$ とその最小次元実現 (V, T, μ) をひとつ固定したとき、(Θ, P) は $\mathcal{X}$ 上の統計モデルである。ただし Θ は自然パラメータ空間、P は自然パラメータ座標の逆写像である。

証明 $p_\theta(x) = \exp\left[\langle \theta, T(x) \rangle - \psi(\theta)\right]$ とおけばよい。 □

例 1.3 (統計モデルの例)**.**

- Cauchy 分布族

命題-定義 1.4 (スコア関数)**.** 実現 (p, μ) に対し関数 $l\colon \Theta \times \mathcal{X} \to \mathbb{R},\ (\theta, x) \mapsto l_x(\theta) := \log p_\theta(x)$ と定め、実現 (p', μ') に対し同様に l' を定めると、$dl_x = dl'_x$ (a.e.x) が成り立つ。そこで $dl\colon \Theta \times \mathcal{X} \to T^\vee \Theta$ を統計モデル $(\Theta, \mathbf{p})$ の**スコア (score)** という。

証明 $(p, \mu), (p', \mu')$ を考えると、

$$\frac{p'_\theta(x)}{p_\theta(x)} = \frac{d\mu}{d\mu'}(x) \quad \text{(a.e.}x\text{)} \tag{1.1}$$

log をとって θ で微分すれば $dl'_\theta = dl_\theta$ (a.e.x) を得る。 □

命題 1.5. 指数型分布族において

$$E_{\mathbf{p}(\theta)}[(dl_\theta)^2] = g_\theta \tag{1.2}$$

が成り立つ。

証明 $dl_\theta = d(\langle\theta, T\rangle - \psi(\theta)) = T(x) - d\psi_\theta$ だから、$E[T] = d\psi_\theta$ に注意すれば

$$E[(dl_x)_\theta^2] = E[(T - E[T])^2] = \mathrm{Var}[T] = g_\theta \tag{1.3}$$

である。 □

定義 1.6 (Fisher 計量). $E_{\mathbf{p}(\theta)}[(dl_x)_\theta^2]$ を統計モデル $(\Theta, \mathbf{p})$ の **Fisher 計量 (Fisher metric)** という。

2 最尤推定

定義 2.1 (i.i.d. 拡張). $\mathcal{X}$ 上の統計モデル $(\Theta, \mathbf{p})$ に対し、$\mathcal{X}^k$ 上の統計モデル $(\Theta, \mathbf{p}^k)$ を $(\Theta, \mathbf{p})$ の **i.i.d. 拡張 (i.i.d. extension)** という。ただし $\mathbf{p}^k(\theta) := \mathbf{p}(\theta) \times \cdots \times \mathbf{p}(\theta)$ (積測度) である。

命題 2.2. 指数型分布族の i.i.d. 拡張は指数型分布族である。

証明 元の指数型分布族の実現 (V, T, μ) に対し、(V, T', μ^k), $T'(x) := \sum_{i=1}^k T(x_i)$ が i.i.d. 拡張の実現となる。 □

命題-定義 2.3 (最尤推定量). 可測写像 $\hat{\theta}\colon \mathcal{X} \to \Theta$ に関する条件を考える。実現 (p, μ) に対し a.e.$x \in \mathcal{X}$, $\hat{\theta}(x) \in \mathrm{argmax}_{\theta\in\Theta}\, p_\theta(x)$ が成り立つとき、$\hat{\theta}$ を **(p, μ) に関する最尤推定量**と呼ぶことにする。このとき次は同値である:

(1) $\hat{\theta}$ はある実現 (p, μ) に関する最尤推定量である。
(2) $\hat{\theta}$ は任意の実現 (p, μ) に関する最尤推定量である。

そこで、上の条件のいずれかが成り立つとき、$\hat{\theta}$ を統計モデル $(\Theta, \mathbf{p})$ の**最尤推定量 (maximum likelihood estimator)** と呼ぶ。

証明 (2) ⇒ (1) は明らか。(1) ⇒ (2) は argmax の定義および $q_\theta = p_\theta \dfrac{d\mu}{d\nu}$ に注意すればわかる。 □

命題 2.4 (尤度方程式). (1) $\hat{\theta}$ が最尤推定量ならば、(2) a.e.x に対し $\hat{\theta}(x)$ は**尤度方程式 (likelihood equation)** $(dl_x)_\theta = 0$ の解である。さらに指数型分布族においては逆も成り立つ。

証明 (1) ⇒ (2) は l_x が θ に関し微分可能であることから従う。指数型分布族の場合、(2) ⇒ (1) は $l_x(\theta) = \langle\theta, T(x)\rangle - \psi(\theta)$ が上に凸であることから従う。 □

例 2.5 (最尤推定量の例).

- 正規分布族の場合

- Cauchy 分布族の場合

3 Kullback-Leibler ダイバージェンス

定義 3.1 (Kullback-Leibler ダイバージェンス). 関数 $D: \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, \infty]$,

$$D(\mu \| \nu) := \begin{cases} E_\mu \left[\log \dfrac{d\mu}{d\nu} \right] & (\mu \sim \nu) \\ \infty & (\mu \nsim \nu) \end{cases} \tag{3.1}$$

を **Kullback-Leibler ダイバージェンス** と呼ぶ。

例 3.2 (KL ダイバージェンスの例).

- 有限集合上の確率分布の族の場合
- 正規分布族の場合

$\mathcal{X} = \{1, \ldots, n\}, n \in \mathbb{N}$ の場合に最尤推定量と KL ダイバージェンスの関係を考える。

定義 3.3 (経験分布). $x = (x_1, \ldots, x_k) \in \mathcal{X}^k$ に対し

$$\hat{p}_x := \frac{1}{k} \sum_{i=1}^{k} \delta^{x_i} \tag{3.2}$$

を x により定まる**経験分布 (empirical distribution)** という。

命題 3.4 (最尤推定量と KL ダイバージェンス). $(\Theta, \mathbf{p})$ を $\mathcal{X}$ 上の統計モデルとし、k 個の i.i.d. 拡張 $(\Theta, \mathbf{p}^k)$ を考える。$x = (x_1, \ldots, x_k) \in \mathcal{X}^k$ とし、$\hat{p}_x$ を x により定まる経験分布とする。このとき、$\mathbf{p}(\Theta)$ が $\hat{p}_x$ と同値な確率測度を含むならば次が成り立つ:

$$\mathrm{argmin}_{\theta \in \Theta} D(\hat{p}_x \| \mathbf{p}^k(\theta)) \subset \mathrm{argmax}_{\theta \in \Theta} p_\theta^k(x) \tag{3.3}$$

証明 argmin を考えるから $\hat{p}_x \sim \mathbf{p}^k(\theta)$ なる θ を考えればよい。

$$\begin{aligned} D(\hat{p}_x \| \mathbf{p}^k(\theta)) &= E_{\hat{p}_x} \left[\log \frac{d\hat{p}_x}{d(\mathbf{p}^k(\theta))} \right] && (3.4) \\ &= E_{\hat{p}_x} \left[\log \frac{d\hat{p}_x}{di} \right] - E_{\hat{p}_x} \left[\log \frac{d(\mathbf{p}^k(\theta))}{di} \right] && (3.5) \\ &= (\theta \text{ によらない項}) - \frac{1}{k} \log p_\theta^k(x) && (3.6) \end{aligned}$$

□

今後の予定

- 射影

参考文献

[AJLS17] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer, **Information Geometry**, Ergebnisse der Mathematik und ihrer Grenzgebiete 34, vol. 64, Springer International Publishing, Cham, 2017.

[Ama16] Shun-ichi Amari, **Information Geometry and Its Applications**, Applied Mathematical Sciences, vol. 194, Springer Japan, Tokyo, 2016 (en).

[AN07] Shun-ichi Amari and Hiroshi Nagaoka, **Methods of Information Geometry**, Translations of Mathematical Monographs, vol. 191, American Mathematical Society, Providence, Rhode Island, April 2007 (en).