I　スコア関数とFisher情報量

abstract　情報幾何学の基礎を支えるスコア関数とFisher情報量を復習します。

1　Introduction

情報幾何学は、確率分布全体の集合を一つの空間（多様体）と見なし、そこに幾何学的な構造を与えて解析する学問分野です。

私たちが馴染んでいるユークリッド空間では、点と点の間に「距離」を定義したり、曲線の「長さ」を測ったりすることができます。情報幾何学は、この考え方を統計モデルの世界に持ち込みます。

- **空間の「点」**：一つ一つの確率分布（例：平均0, 分散1の正規分布）
- **空間の「座標」**：分布を特徴づけるパラメータ（例：正規分布の平均 μ と分散 σ^2）

この「統計多様体」と呼ばれる空間に、距離や角度、曲率といった幾何学的な“ものさし”を導入することで、これまでとは異なる視点から統計モデルの性質を深く理解することを目指します。

実はこのような視点は、すでに数理統計学を学んだことがある方なら、暗に使って来た考え方です。例えば、スコア関数、Fisher情報量そしてKullback-Leibler divergenceを使ったことがあれば、皆さんは暗に情報幾何学的な考え方に触れたことがあるでしょう。そこでこの講義は、これらの概念を情報幾何学を通して理解することを目指します。

2　各概念の基本的な性質

2.1　スコア関数

対数尤度関数の導関数をスコア関数といいます。

定義　スコア関数

パラメータ $\theta = (\theta_1, \cdots, \theta_n)$ を持つ確率分布 $p(x;\theta)$ に対して、そのスコア関数とは以下のように定義される。

$$S(\theta) = \frac{\partial \log p(x;\theta)}{\partial \theta}$$

なぜこのような量が重要なのか、現在は理解が難しいと思いますが、情報幾何学を学んでいくうちにその重要性がよくわかることでしょう。一旦、具体例を与えておきます。

例　正規分布のスコア関数

正規分布 $N(\mu, \sigma^2)$ のスコア関数は $S(\mu), S(\sigma^2)$ はそれぞれ以下のようになる。

$$S(\mu) = \frac{x-\mu}{\sigma^2}, \quad S(\sigma^2) = \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}$$

2.2　スコア関数の性質

スコア関数の重要な性質として、スコア関数の期待値は 0 になるというものがあります。

定理　スコア関数の期待値は0

以下の性質が成り立ちます。

$$\mathbb{E}\left[\frac{\partial \log p(x;\theta)}{\partial \theta}\right] = 0$$

証明 対数微分から、左辺は

$$\mathbb{E}\left[\frac{\partial \log p(x;\theta)}{\partial \theta}\right] = \int p(x;\theta)\frac{\frac{\partial p(x;\theta)}{\partial \theta}}{p(x;\theta)}dx = \int \frac{\partial p(x;\theta)}{\partial \theta}dx$$

と式変形できます。さらに、微分と積分の順序交換が可能だとすると、

$$\int \frac{\partial p(x;\theta)}{\partial \theta}dx = \frac{\partial}{\partial \theta}\int p(x;\theta)dx = \frac{\partial 1}{\partial \theta} = 0$$

です。■

2.3　Fisher情報量

Fisher情報量は各パラメータのスコア関数の内積として定義されます。この内積として定義されているという事実は非常に重要で、ここにFisher情報量と情報幾何学のつながりが垣間見えています。

定義　Fisher情報量

パラメータ $\theta = (\theta_1, \cdots, \theta_n)$ を持つ確率分布 $p(x;\theta)$ に対して、Fisher情報量 $I(\theta)$ の (i,j) 成分を

$$I_{ij}(\theta) = \mathbb{E}\left[\frac{\partial \log p(x;\theta)}{\partial \theta_i}\frac{\partial \log p(x;\theta)}{\partial \theta_j}\right]$$

で定義します。これはスコア関数の分散共分散行列に他なりません。

具体例として、正規分布のFisher情報量を求めておきます。

例　正規分布のFisher情報量

$X \sim N(\mu, \sigma^2)$ とする。X が持つパラメータ (μ, σ^2) のFisher情報量は以下のように与えられる。

$$I(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}$$

2.4　Fisher情報量の性質

Fisher情報量が確率分布にとって本質的かどうかは、以下の定理を知るとひとまずの了解を得ることができるのではないでしょうか。

定理　Fisher情報量は全単射な変数変換に対して不変

確率変数 X がパラメータ θ によって定義される確率分布 $p(x;\theta)$ に従っているとする。変換 $Y = f(X)$ が逆関数 $g = f^{-1}$ を持つとき、X が持つFisher情報量 $I_X(\theta)$ と Y が持つFisher情報量 $I_Y(\theta)$ は等しい。

証明　変数変換の公式から以下のことがわかる点に注意します。

$$p(y;\theta) = p(x(y);\theta)\frac{dx}{dy}, \quad \log p(y;\theta) = \log p(x(y);\theta) + \log\frac{dx}{dy}$$

特に、dx/dy は θ に依存しないので、

$$\frac{\partial \log p(y;\theta)}{\partial\theta} = \frac{\partial \log p(x(y);\theta)}{\partial\theta}$$

です。X が持つFisher情報量を置換積分で式変形すると、以下のようになります。

$$\begin{aligned} I_X(\theta) &= \int p(x;\theta)\frac{\partial \log p(x;\theta)}{\partial\theta}dx \\ &= \int p(x(y);\theta)\frac{\partial \log p(x(y);\theta)}{\partial\theta}\frac{dx}{dy}dy \\ &= \int p(y;\theta)\frac{\partial \log p(y;\theta)}{\partial\theta}dy \\ &= I_Y(\theta) \end{aligned}$$

以上で定理の主張が確認できました。■

3　Fisher情報量の重要性

Fisher情報量は、統計的推定の理論において中心的な役割を果たします。その代表例がCramer-Raoの下限です。もう一つの重要な内容としてLehmann-Scheffeの定理がありますが、こちらは後ほど詳しく解説します。

3.1　Cramer-Raoの下限

Cramer-Raoの下限は、不偏推定量が持ちうる分散の下限を与えます。

定理　Cramer-Raoの下限

パラメータ θ の任意の不偏推定量 $\hat{\theta}$（すなわち $\mathbb{E}[\hat{\theta}] = \theta$ を満たす推定量）の分散は次を満たします。

$$\mathbb{V}[\hat{\theta}] \geq \frac{1}{I(\theta)}$$

ここで $I(\theta)$ はパラメータ θ に関するFisher情報量です。

この不等式は、Fisher情報量が大きいほど推定量の分散が小さくなる（＝より高精度な推定が可能になる）ことを示します。

3.2　Cramer-Raoの下限の証明

不偏推定量 $\hat{\theta}$ をスコア関数との関係で特徴づけてみましょう。不偏性の定義は

$$\mathbb{E}[\hat{\theta}] = \int \hat{\theta}(x) p(x;\theta)\, dx = \theta$$

でした。ここで両辺を θ で微分してみましょう。微分と積分の順序交換が可能だとすると、

$$\int \hat{\theta}(x) \frac{\partial p(x;\theta)}{\partial \theta}\, dx = 1$$

が得られます。ここで、$\dfrac{\partial p}{\partial \theta} = p \dfrac{\partial \log p}{\partial \theta}$ を用いると

$$\mathbb{E}\left[\hat{\theta} \frac{\partial \log p}{\partial \theta}\right] = 1$$

が得られ、これが不偏推定量とスコア関数との間の関係です。特に、スコア関数の期待値が 0 であることを思い出せば、

$$Cov\left(\hat{\theta}, \frac{\partial \log p}{\partial \theta}\right) = \mathbb{E}\left[\hat{\theta} \frac{\partial \log p}{\partial \theta}\right] - \mathbb{E}[\hat{\theta}]\mathbb{E}\left[\frac{\partial \log p}{\partial \theta}\right] = 1$$

が成り立ちます。

あとは、コーシー・シュワルツの不等式 $Cov(X,Y)^2 \leq \mathbb{V}[X]\mathbb{V}[Y]$ を適用すると

$$1^2 \leq \mathbb{V}[\hat{\theta}]\, \mathbb{V}\left[\frac{\partial \log p}{\partial \theta}\right]$$

が得られ、$\mathbb{V}\left[\dfrac{\partial \log p}{\partial \theta}\right] = \mathbb{E}\left[\left(\dfrac{\partial \log p}{\partial \theta}\right)^2\right] = I(\theta)$ より

$$\mathbb{V}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

が従います。これはCramer-Raoの下限です。