AIC[†]

奥田 徹示*

1. K-L 情報量

偶然性にもとづいて発生する現象を、確率分布にしたがう確率変数の実現値が発生することであるとして、ここでは確率分布をモデルともよぶことにする。

さて、いま1つのサイコロがあり、このサイコロの偶数の目および奇数の目の出る確率を推定したいとする。ここで、2つのモデルを考えたとする。すなわち、 $q = \{\text{偶数の目の出る確率, 奇数の目の出る確率}\} = \{q_1, q_2\}$ として、 $q_A = \{0.5, 0.5\}$, $q_B = \{0.4, 0.6\}$ とする。もし、真のモデルを $p = \{p_1, p_2\}$ として、 $p = \{0.45, 0.55\}$ であったとすると、モデル q_A, q_B のどちらがより p に近いかを考える。このような場合、分布間の近さを測る1つの量として、Kullback-Leibler (K-L) 情報量がある。

$$I(p, q) = \sum p_i \log \frac{p_i}{q_i} \quad (1)$$

とすると、 $p_2 = 1 - p_1$, $q_2 = 1 - q_1$ であるから、

$$I(p, q) = p_1 \log p_1 - p_1 \log q_1 + (1 - p_1) \log (1 - p_1) - (1 - p_1) \log (1 - q_1)$$

となる。そこで、 q_1 で微分すると

$$\frac{\partial}{\partial q_1} I(p, q) = -\frac{p_1}{q_1} + \frac{1 - p_1}{1 - q_1}$$

$$\frac{\partial^2}{\partial q_1^2} I(p, q) = \frac{p_1}{q_1^2} + \frac{1 - p_1}{(1 - q_1)^2}$$

となり、 $I(p, q)$ は $q_1 = p_1$ の場合に限り最小値0をとる凸関数($0 < q_1 < 1$)になることが分かる。

したがって、この $I(p, q)$ は分布 p と q との近

さを測る(近くなるほど値が小さくなる)量として考えることができる。これを K-L 情報量という。

ここでの例で考えると

$$I(p, q_A) = 0.45 \log 0.45 - 0.45 \log 0.5 + 0.55 \log 0.55 - 0.55 \log 0.5 = 0.005008$$

$$I(p, q_B) = 0.45 \log 0.45 - 0.45 \log 0.4 + 0.55 \log 0.55 - 0.55 \log 0.6 = 0.005146$$

となり、 q_A の方がより真の分布 p に近いということになる。 $p = (p_1, \dots, p_m)$, $q = (q_1, \dots, q_m)$ の場合も(1)と同様に K-L 情報量を与えることができる。問題は真の分布 p が分からない状況で、なるべく真のモデル p に近いモデル q を求めることにあつた。したがって、与えられたデータから情報量 $I(p, q)$ を推定する必要が生じる。

2. 情報量の推定

いま、事象 $w_1, \dots, w_m$ に対する真の分布は分からないが、その確率分布を p として、

$$I(p, q) = \sum p_i \log p_i - \sum p_i \log q_i$$

であるので、この情報量をなるべく小さくするためには

$$\sum p_i \log q_i$$

をできるだけ大きくするようなモデル q を選ばばよい。ここで実現値として $\log q_i$ をとる確率変数 $\log q$ をとると、各 w_i の観測個数を n_i ($n_1 + \dots + n_m = n$)として、大数の法則から

$$\frac{1}{n} \sum n_i \log q_i \rightarrow E(\log q) \quad (n \rightarrow \infty)$$

となるから、データから計算できる値

$$\sum n_i \log q_i$$

を K-L 情報量の推定に用いることができる。つま

[†] AIC

* Tetsuji OKUDA

大阪工業大学 経営工学科

Osaka Institute of Technology

り $\sum n_i \log q_i$ の値が大きいほどよいモデルとなる。

したがって連続的なモデルでは、未知ではあるが真の密度関数を $g(x)$ 、それに対するモデルを $f(x)$ として、

$$\begin{aligned} I(g, f) &= \int g(x) \log \frac{g(x)}{f(x)} dx \\ &= \int g(x) \log g(x) - \int g(x) \log f(x) dx \end{aligned}$$

と書くことができ、 n 個の観測値 $x_1, x_2, \dots, x_n$ を得たとき、

$$\int g(x) \log f(x) dx$$

の評価に

$$\frac{1}{n} \sum \log f(x_i) \quad (2)$$

を用いることができることになる。すなわち、真の分布 p が未知であるので I そのものは推定できないが、(2) によってモデルの相対的な良し悪しは評価できる。

そこで、モデル f を母数 θ に依存させて $f(x, \theta)$ と書くと、結局

$$L(\theta) = \sum \log f(X_i, \theta) \quad (3)$$

をなるべく大きくするように f を選ぶ、すなわち $L(\theta)$ が最大になるように θ を決定することになり、最尤推定量を求めることと同じになる。

すなわち

$$L(\theta_0) = \max \sum \log f(X_i, \theta)$$

なる θ_0 を求めればよい。

以後は連続型分布を対象として議論する。

3. AIC

データ $x = (x_1, \dots, x_n)$ を確率変数 $X = (X_1, \dots, X_n)$ の実現値として、 X_i は互いに独立に同じ真の分布 $g(\cdot) = f(\cdot, \theta^*)$ にしたがうものとする。いま、 Z は X_i と同じ分布にしたがっているとして、前節から、モデルの良し悪しは平均対数尤度

$$E_Z \{ \log f(z, \theta) \} = \int f(z, \theta^*) \log f(z, \theta) dz$$

で評価できたが、(3) との対応で

$$L^*(\theta) = n E_Z \{ \log f(z, \theta) \}$$

を定義する。この $L^*(\theta)$ が大きいほど $f(\cdot, \theta)$ は $f(\cdot, \theta^*)$ に近いということになり、前節の議論からまず第1段階として $L^*(\theta_0)$ で評価することは自然な考えである。しかしながら、現実には θ_0 はデータに依存しているので、 $L^*(\theta)$ の期待値をとって

$$E_X \{ L^*(\theta_0) \} = \int_X L^*(\theta_0) \Pi g(x_i) dx$$

を評価基準とする必要がある。これを期待平均対数尤度とよび、結局この期待平均対数尤度をより大きくするモデルが良いことになる。

ところが、最大対数尤度を期待平均対数尤度の推定量と考えると、それは不偏推定量とならないので、そのバイアスを調整したものが AIC として導かれる。

まず、母数 θ は k 次元ベクトルとして、 $L^*(\theta_0)$ を θ^* のまわりで展開して

$$\begin{aligned} L^*(\theta_0) &\doteq L^*(\theta^*) \\ &+ n(\theta_0 - \theta^*) E_Z \left\{ \frac{\partial}{\partial \theta} \log f(z, \theta) \right\}_{\theta^*} \\ &+ \frac{1}{2} \sqrt{n} (\theta_0 - \theta^*) E_Z \left\{ \frac{\partial^2}{\partial \theta^2} \log f(z, \theta) \right\}_{\theta^*} \sqrt{n} (\theta_0 - \theta^*) \end{aligned}$$

となる。なお、 $\partial/\partial \theta$ の項は k 次元ベクトルを、 $\partial^2/\partial \theta^2$ の項は $\partial^2/(\partial \theta_i \partial \theta_j)$ を要素にもつ $k \times k$ 行列を表す。以後は " $\doteq$ " を " $=$ " と記す。

ここで、 $\theta = \theta^*$ のとき $L^*(\theta)$ は最大となるから、上式の第2項は0となり、

$$\begin{aligned} L^*(\theta_0) &= L^*(\theta^*) \\ &- \frac{1}{2} \sqrt{n} (\theta_0 - \theta^*) J_* \sqrt{n} (\theta_0 - \theta^*) \end{aligned}$$

となる。ここで、

$$J_* = -E_Z \left\{ \frac{\partial^2}{\partial \theta^2} \log f(z, \theta) \right\}_{\theta^*}$$

である。なお、

$$\begin{aligned} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(z, \theta) &= \frac{1}{f(z, \theta)} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(z, \theta) \\ &- \frac{1}{f(z, \theta)^2} \frac{\partial}{\partial \theta_i} f(z, \theta) \frac{\partial}{\partial \theta_j} f(z, \theta) \end{aligned}$$

となる。したがってフィッシャー情報行列を

$$I_* = E_z \left\{ \frac{\partial}{\partial \theta_i} \log f(z, \theta) \frac{\partial}{\partial \theta_j} \log f(z, \theta) \right\}_{\theta_*}$$

と表すと、 $J_* = I_*$ となる。一方、最尤推定量のよく知られた性質から、 $\sqrt{n}(\theta_0 - \theta_*)$ は n が大きくなると $N(0, I_*^{-1})$ に近づくので、結局 $\sqrt{n}(\theta_0 - \theta_*)$ は $N(0, J_*^{-1})$ に近づく。

また、一般に X が $N(\mu, \Sigma)$ の k 次元正規分布にしたがうとき、 $(X - \mu)^T \Sigma^{-1} (X - \mu)$ は自由度 k の χ^2 分布にしたがう。すなわち、 n が大きくなれば、近似的に

$$E_x \{L^*(\theta_0)\} = L^*(\theta^*) - \frac{k}{2}$$

が成立する。

また、同様の計算をすることにより、

$$E_x \{L(\theta^*)\} = E_x \{L(\theta_0)\} - \frac{k}{2}$$

が導かれる。さらに、

$$\begin{aligned} L^*(\theta^*) &= n \int f(z, \theta^*) \log f(z, \theta^*) dz \\ &= \sum \int f(x_i, \theta^*) \log f(x_i, \theta^*) dx_i \\ &= \int_x \Pi g(x_i) (\sum \log f(x_i, \theta^*)) dx \\ &= E_x \{L(\theta^*)\} \end{aligned}$$

となる。

したがって、

$$E_x \{L^*(\theta_0)\} = E_x \{L(\theta_0)\} - k$$

が成立するから、期待平均対数尤度の不偏推定量として $L(\theta_0) - k$ が考えられる。

そこで、

$$\text{AIC} = -2 L(\theta_0) + 2k$$

として、このAICを最小にするモデルを選択することにすれば、期待平均対数尤度を最大にするという目的のために、バイアスのかからない推定量としてAICを用いることができる。ただし、歴史的な事情で推定量は -2 倍されるようである。

以上は[1]にもとづいて説明したが、AICは期待平均対数尤度の不偏推定量の1つとして導かれており、K-L情報量以外にも基準として考えられ

る可能性があり、AICはモデル選択のために利用可能な1つの基準であると理解できる。

ただ本節では、 $g(\cdot) = f(\cdot, \theta^*)$ として議論したが、実際問題として真の分布 g は未知であるので、想定するモデルのクラス $f(\cdot, \theta)$ で $g(\cdot)$ が表現できない場合が考えられる。

この場合は、

$$\max \int g(z) \log f(z, \theta) dz$$

となる θ_* を求めても、 $f(z, \theta_*)$ は必ずしも $g(z)$ ではない。したがって、 $J_* = I_*$ が一般的に成立しない。このような場合には、期待平均対数尤度の補正項としてパラメータ数以外に補正項が生じることが[2]で述べられている。

ただし、モデル $f(\cdot, \theta)$ に多くの母数を含むのであれば、真の分布に近づけるためにかなりの調整が可能であるから、わざわざ別のモデルを選択する必要はそれほどないので、この場合にはあまり問題にならないであろう。

また、AICは n を大きくしたときに近似的に導かれているが、現実にはデータ数は有限であるから、データ数 n に関するもう少し詳しい議論も望まれる。これらいくつかの問題点も考えられるが、これらについては[2]を参照されたい。

いずれにしても、AICは基本的に最尤推定量を用いており、またモデル選択に対する統計理論があまりない現状では、バイアスの補正を施した段階においてこのAICはかなり有力な手法であると思われる。

参 考 文 献

- [1] 坂元, 石黒, 北川: 情報量統計学, 共立出版, 1983
- [2] 竹内 啓: 現象と行動のなかの統計数理(13章), 新曜社, 1980

[問い合わせ先]

〒535 大阪市旭区大宮5-16-1
大阪工業大学 経営工学科
奥田 徹示
☎: 06-952-3131(内)3714
FAX: 06-952-6197