赤池情報量基準と関数近似

2015SS034 松田 真太朗
指導教員：小藤俊幸

1　はじめに

近年, 赤池弘次博士により発表された赤池情報量基準 (Akaike informaition criterion) は統計的モデルを評価する方法として優れており, 様々な分野で広く活用をされている．その中で,X 線強度に対応するフィルム濃度が情報量基準を用いて曲線近似したものに見あたらないことに気がついた北海道大学の教授らが赤池情報量基準を用いて, 有限フーリエ級数で近似した特製曲線作製を行った．私は, 赤池情報量基準とはどのようなものなのかを追実験を行いながら学んでいき, また私が追実験を行った結果が北海道大学の結果と一致することを目標に取り組んだ。その中で,AIC の値と最良のモデルにはどのような関係があるのかも確認していく．
以後, 赤池情報量基準を AIC と記す.

2　AIC の定義

いま一対の観測データ群を x_1,x_2,$\cdots$,x_n, ならびに y_1,y_2,$\cdots$,y_n, それらを確率変数とみなしたときは x ならびに y と記し、統計的モデルを規定するパラメータを $a = a_1$,a_2,$\cdots$,a_l とする．このとき x と y がそれぞれ独立な場合、統計的モデルの密度関数を $f(x, y|a)$ とするとき
L(a)=$f(x_1, y_1|a)\cdot\cdots\cdot f(x_n, y_n|a)$
を尤度とよび, その自然対数をとった

$$l(a) = \sum_{i=1}^{n} \log f(x_i, y_i|a) \tag{1}$$

を対数尤度とよぶ．ここで n は観測データの数である．l(a) は一連の観測値 x_1,x_2,$\cdots$,x_n,y_1,y_2,$\cdots$,y_n で定まる a の関数である．a に適当な値を与えれば l(a) を最大にすることができる．そのときの $a = \tilde{a}$ を最尤推定量,$l(\tilde{a})$ を最大対数尤度という．このとき最尤推定量によって定められるモデルが良いモデルとされ, 最尤モデルと呼ばれる．[2]
また,
$AIC = -2 \times$ (モデルの最大対数尤度) $+ 2 \times$ (モデルのパラメータ数)
で計算を行う.
いくつかのモデルが存在するとき, 最小の AIC をもつモデルが最適なモデルと考える．[1]

3　AIC と最良モデルの決定法

X 線強度を x, それに対応するフィルム濃度を y とするとき,y を x に依存する正規分布母集団の確率変数とみなすことにする．このときの濃度分布 y の統計的モデルの密度関数 $f(x, y|a)$ を

$$f(x, y|a) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2(x)}} \mathrm{e}^{\frac{-(y-D(x|a))^2}{2\sigma^2(x)}} & (0 \leq \mathrm{x} \leq 1) \\ 0 & (\mathrm{x} < 0, \mathrm{x} > 1) \end{cases} \tag{2}$$

と定めることにする．[2]
上式において,$\sigma^2(x)$ は各 X 線における濃度分布の分散,$D(x|a)$ は濃度分布の平均値で, ともに X 線強度の関数である．a は $D(x|a)$ に含まれるパラメータを示す．以上により特性曲線の近似式は (2) 式中の $D(x|a)$ である．ここでは、$D(x|a)$ を以下に記すような有限フーリエ級数とした.

$$D(x|a) = a_0 + \sum_{m=1}^{M} (a_{2m-1} \sin 2m\pi x + a_{2m} \cos 2m\pi x) \tag{3}$$

この式において,M=1,M=2,· と M の値を変化させていき, 対数尤度と最大対数尤度を求める．最大対数尤度を求めていけば, 最尤モデルと AIC の計算が可能になる．

3.1　数学的計算手順

まず,M に適当な値をあたえて,(1) 式,(2) 式,(3) 式から対数尤度 $l(a)$ をもとめる．

$$\begin{aligned} l(a) &= \sum_{i=1}^{n} \log f(x_i, y_i|a) \\ &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2(x_i)}} \mathrm{e}^{\frac{-(y_i-D(x_i|a))^2}{2\sigma^2(x_i)}} \\ &= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\log\sigma^2(x_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2(x_i)}\{y_i - D(x_i|a)\}^2 \end{aligned}$$

ここで,n は濃度データの総個数で,$l(a)$ は X 線強度と濃度の対のデータ (x_1,y_1),(x_2,y_2),$\cdots$,(x_n,y_n) 各 X 線強度における濃度分布の分散 $\sigma^2(x)$ が決まれば,a=$\{a_0,a_1,\cdots,a_{2M}\}$ の関数になるので, $l(a)$ を最大にする条件 $\tilde{a}$ がもとまる．また,$\tilde{a}$=$\{\tilde{a_0},\tilde{a_1},\cdots,\tilde{a_{2M}}\}$ を求めるには $l(a)$ を a で微分し, 導関数を 3 つもとめる．以下に導関数を記す.

$$\frac{\partial l(a)}{\partial a_0} = \sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)}\{y_i - D(x_i|a)\}$$

$$\frac{\partial l(a)}{\partial a_{2m-1}} = \sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)}\{y_i - D(x_i|a)\}\sin 2m\pi x_i$$

$$\frac{\partial l(a)}{\partial a_{2m}} = \sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)}\{y_i - D(x_i|a)\}\cos 2m\pi x_i$$

これら 3 式についてそれぞれ零にすると,$a_0, a_1, \cdots, a_{2M}$ を未知数とする連立一次方程式がもとまり, それを以下に

記した.

$$\sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)} D(x_i|a) = \sum_{i=1}^{n} \frac{y_i}{\sigma^2(x_i)} \quad (4)$$

$$\sum_{i=1}^{n} \frac{D(x_i|a)\sin 2m\pi x_i}{\sigma^2(x_i)} = \sum_{i=1}^{n} \frac{y_i \sin 2m\pi x_i}{\sigma^2(x_i)} \quad (5)$$

$$\sum_{i=1}^{n} \frac{D(x_i|a)\cos 2m\pi x_i}{\sigma^2(x_i)} = \sum_{i=1}^{n} \frac{y_i \cos 2m\pi x_i}{\sigma^2(x_i)} \quad (6)$$

3.2 解法

具体的にどのように解いていくのかをしめす。
(4)-(6) 式を $Ax = b$ の行列式に直して考え、M=1 から順次変換して解いていく. M=1 のときを以下に記す. ここからは表 1 のデータ 12 個を用いて計算を行うので、n=12 とする. その他、x_i は x 線強度の強さ、y_i は平均値、$\sigma^2(x_i)$ は分散の値である.

$$A = \sum_{i=1}^{12} \frac{1}{\sigma^2(x_i)} \begin{bmatrix} 1 & \sin 2\pi x_i & \cos 2\pi x_i \\ \sin 2\pi x_i & \sin^2 2\pi x_i & \sin 2\pi x_i \cos 2\pi x_i \\ \cos 2\pi x_i & \sin 2\pi x_i \cos 2\pi x_i & \cos^2 2\pi x_i \end{bmatrix}$$

$$x = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} \quad b = \sum_{i=1}^{12} \frac{1}{\sigma^2(x_i)} \begin{bmatrix} y_i \\ y_i \sin 2\pi x_i \\ y_i \cos 2\pi x_i \end{bmatrix}$$

とそれぞれ表し, を行い表 1 のデータを代入すると

$$\begin{bmatrix} 2613.6527 & 292.7214 & 543.5363 \\ 292.7214 & 1218.7881 & -49.37913 \\ 543.5363 & -49.3791 & 1394.7398 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 2686.6952 \\ -59.4094 \\ 71.0064 \end{bmatrix}$$

これにより、a_0, a_1, a_2 の解を求めることができた.
$a_0 = 1.151, a_1 = -0.342, a_2 = -0.410$
最尤推定量が求めることができたので、(3) に代入することで特製曲線近似式の $D(x|a)$ が決定される. これをを以下の式に代入し最大対数尤度を求める.

$$l(a) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\log\sigma^2(x_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2(x_i)}\{y_i - D(x_i|a)\}^2$$

計算を行った結果、$l(a) = -1631.1153$ という値が得られた. またパラメータ数とは連立一次方程式の本数であり、AIC の定義式に従い AIC を求めると
AIC=-2 × (-1631.115)+2 × 3=3268.23 となった.
また、M=5 のとき
$a_0 = 1.283, a_1 = -0.381, a_2 = -0.463, a_3 = -0.262, a_4 = 0.073, a_5 = 0.239, a_6 = -0.025, a_7 = 0.871, a_8 = 0.277, a_9 = -2.916, a_{10} = 1.252$ となる.
l(a)=-463.28,AIC=-2 × (463.28)+2 × 11=-904.64
これを $M = 8$ まで行い、各 AIC を求める. 結果を表 2 に記した. ここで、モデルのパラメータ数を A とおく.

4 結果と比較

追実験の結果は M=2 の値から一致しなくなった. 算出方法は何度も確認を行ったので間違っていない. 北海道大学は X 線フィルム濃度のグラフと一致するときは、はじめにあらわれる極小値（A=11）と停留点（A=13）を見つけることが最良のモデルと考察しており、値は異なるが極小値と停留点の A は一致しているので A=11,A=13 が最良のモデルと考察できる.

表 1 濃度データの統計的性質 [2]

X 線強度	平均値	分散
0.0	0.225	0.00267
0.2	0.236	0.00211
0.4	0.279	0.00250
0.6	0.388	0.00325
0.8	0.549	0.00369
1.0	1.043	0.00602
1.2	1.772	0.00991
1.4	2.520	0.01299
1.6	3.180	0.01321
1.8	3.483	0.00784
2.0	3.641	0.00833
2.4	3.714	0.00839

表 2 A の個数と AIC の値

A の個数	AIC
3	3268.23
5	1564.74
7	-702.56
9	-832.77
11	-904.64
13	-901.64
15	-1043.44
17	-1045.07

表 3 北海道大学の結果 [2]

A の個数	AIC
3	3268.23
5	1222.86
7	-952.93
9	-1058.84
11	-1129.22
13	-1127.39
15	-1135.79
17	-1135.07

5 おわりに

今回、このような実験を私が自分で手順をおい、実際に数値をもとめていくことで情報量基準の 1 つである AIC の導き方を学べた.、最良のモデルを選択するとき、一般的に AIC が最小なものが最良のモデルとされているが、極小値や停留点にも注目することでより正確に良いモデルを選択できる. 今後も AIC がどこに利用されているのかを調べ、時間があるときにはまた追実験を行いたい.

参考文献

[1] 赤池弘次・甘利俊一・北川源四郎・樺島祥介・下平英寿:「赤池情報量基準 AIC -モデリング・予測・知識発見-」. 共立出版, 東京,2007

[2] 花田博之・関之山勝博・加藤浩:「赤池情報量基準による X 線フィルム特性曲線の決定法」北海道大学医療技術短期大学部紀要,4:33-40,1991
URL:http://hdl.handle.net/2115/37532