情報量基準とその応用

M2006MM019 水戸 藍
指導教員　木村 美善

1　はじめに

統計的推測や予測の問題において, 情報量とモデルは大きな役割を果たす. 代表的な情報量基準として, 統計的モデルのよさを評価するために用いられるカルバックライブラー (K-L) 情報量基準があり, この情報量から生まれたのが赤池情報量基準 (AIC) である.AIC は有効なモデル選択の基準として幅広い分野で利用されているが,AIC 以外にも様々な情報量基準がモデル選択に用いられている. 本論文の目的は, それらの情報量基準に関する理論的考察を行うとともに情報量基準を混合モデルに適用した場合の有効性についてシミュレーションにより明らかにする.

2　統計的モデル

データ $\boldsymbol{x}_n = \{x_1, x_2, \cdots, x_n\}$ が分布関数 $G(x)$ から生成されたとする. このとき,$G(x)$ を真の分布という.

$$G(x) = \int_{-\infty}^{x} g(t)dt \tag{1}$$

これに対して真の分布を近似するために想定した分布関数 $F(x)$ をモデルという.

3　統計的モデルの評価

統計モデルのよさ, すなわちモデルの分布 $F(x)$ と真の分布 $G(x)$ の近さを測る尺度として Akaike(1973) はカルバックライブラー (K-L) 情報量

$$I(g;f) = E_G\left[\log\frac{g(X)}{f(X)}\right] \tag{2}$$

$$= E_G\Big[\log g(X)\Big] - E_G\Big[\log f(X)\Big] \tag{3}$$

を用いることを提案した. しかし K-L 情報量は未知の分布 G を含むため, その値を直接計算することができない. ところが K-L 情報量の右辺の第 1 項は真の分布 G だけに依存する定数であるので, 異なるモデルを比較するためには右辺第 2 項だけを考えればよい. これを平均対数尤度という. しかし, 依然として真の分布 G に依存するので未知の確率分布 G をデータ $\boldsymbol{x_n}$ に基づく経験分布関数 $\hat{G}$ で置き換えることで対数尤度

$$E_{\hat{G}}[\log f(X)] = \int \log f(x)d\hat{G}(x) \tag{4}$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n}\log f(x_\alpha) \tag{5}$$

を得る. モデルが, 未知の p 次元パラメータ $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots \theta_p)'$ をもつ確率分布 $f(x|\boldsymbol{\theta})$ で与えられるとき, 対数尤度を $\boldsymbol{\theta} \in \Theta$ の関数とみなして

$$l(\boldsymbol{\theta}) = \sum_{\alpha=1}^{n}\log f(x_\alpha|\boldsymbol{\theta}) \tag{6}$$

と表す. これは対数尤度関数と呼ばれ, この対数尤度関数を最大にする $\hat{\boldsymbol{\theta}}$ を最尤推定量と呼ぶ. しかし対数尤度はバイアスを持ち,K-L 情報量のよい推定量とはならない. そこで, 情報量基準は推定したモデルの最大対数尤度で平均対数尤度を推定したときのバイアス

$$b(G) = E_{\hat{G}}[\log f(X_n|\hat{\boldsymbol{\theta}}(X_n)) - nE_{G(z)}[\log f(Z|\hat{\boldsymbol{\theta}}(X_n))]] \tag{7}$$

を補正した量として導かれる.

3.1　様々な情報量基準

赤池情報量基準 (AIC) AIC は, 真のモデルが想定したモデルの中に含まれ, その推定を最尤法に基づいて行ったときのモデル評価基準である.p は自由パラメータ数である.

$$AIC = -2\sum_{\alpha=1}^{n}\log(X_\alpha|\hat{\boldsymbol{\theta}}) + 2p. \tag{8}$$

修正 AIC Sugiura(1978) は簡単なモデルに対してバイアスを直接評価することでバイアス項の有限補正を行うことで修正 AIC を提案した.

$$CAIC = -2\sum_{\alpha=1}^{n}\log(X_\alpha|\hat{\boldsymbol{\theta}}) + \frac{3p}{2}. \tag{9}$$

ベイズ情報量基準 (BIC) Akaike(1977) および Schwarz(1978) はベイズの事後確率の立場から BIC を提案した.

$$BIC = -2\sum_{\alpha=1}^{n}\log(X_\alpha|\hat{\boldsymbol{\theta}}) + p\log n \tag{10}$$

ブートストラップ基準量 (EIC) 対数尤度のバイアス式をブートストラップ法を適応して推定すると

$$b(EIC) = E_G\Big[\sum_{\alpha=1}^{n}\log f(X_\alpha|\hat{\boldsymbol{\theta}}) - n\int g(z)\log f(z|\hat{\boldsymbol{\theta}})dz\Big]$$

$$\approx \frac{n}{B}\sum_{i=1}^{B}\{L(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{*}|\boldsymbol{X}_{\boldsymbol{n(i)}}^{*}) - L(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{*}|\boldsymbol{X}_{\boldsymbol{n(i)}})\} \tag{11}$$

となる. したがって,EIC は次式で与えられる.

$$EIC = -2\sum_{\alpha=1}^{n}\log(X_\alpha|\hat{\boldsymbol{\theta}}) + 2\hat{b}(EIC). \tag{12}$$

修正 EIC ブートストラップのバイアス推定の変動減少を目的として提案された方法で, 次式で与えられる.

$$b(CEIC) \approx \frac{n}{B}\sum_{i=1}^{B}\{L(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{*}|\boldsymbol{X}_{\boldsymbol{n(i)}}^{*}) - L(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}|\boldsymbol{X}_{\boldsymbol{n(i)}}^{*}) + L(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}|\boldsymbol{X}_{\boldsymbol{n(i)}}) - L(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{*}|\boldsymbol{X}_{\boldsymbol{n(i)}})\} \quad (13)$$

$$CEIC = -2\sum_{\alpha=1}^{n}\log(X_{\alpha}|\hat{\boldsymbol{\theta}}) + 2\hat{b}(CEIC). \quad (14)$$

4 多変量正規混合分布モデルのコンポーネント数の推定

多変量データに混合分布モデルをあてはめ, その事後確率に基づいてデータを分類する方法は, 分類手法の 1 つとして様々な分野で応用されている. そこで, 情報量基準を混合正規分布モデルのコンポーネント数の推定問題に適用し, シミュレーションを通して, 各情報量基準を比較する.

4.1 混合正規モデル

真の確率密度関数および, その分布関数をそれぞれ $g(x), G(x)$ とし,$g(x)$ に従って観測された大きさ n の標本を $X_n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ とする. 観測されたデータに対する m 個のコンポーネント分布からなる多変量正規分布モデルは

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{m}\epsilon_k N(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}); \ \sum_{k=1}^{m}\epsilon_k = 1 \quad (15)$$

である. ここで ϵ_k は混合比率である.

4.2 推定アルゴリズム

混合分布モデルのパラメータを EM 法で推定するとき, 初期値の設定は重要な問題である. この問題に対して複数のクラスター化法を用いてデータの初期分類を行うことによって,EM 法の初期値を複数設定し, 最大対数尤度を求めることにする (中村 [3]). 情報量基準を用いたコンポーネント数の推定アルゴリズムの基本的な考え方は, コンポーネント数が $m = 1, 2, \cdots, c$ の c 種類のモデルの情報量基準の値を比較して, その値が最小となるモデルのコンポーネント数を推定値とする. その際に,c 種類の各モデルでは複数のクラスター化法による初期値設定を通して, 最大対数尤度を求める.

4.3 データ解析

混合分布のコンポーネント数の推定アルゴリズムを, 真のコンポーネント数が既知のアヤメデータと, 糖尿病データに適用し,AIC,BIC, 修正 AIC,EIC, 修正 EIC のうち, どの情報量基準が真のコンポーネント数をよく推定することができるかについて比較と検証を行った. 糖尿病データ (Reaven and Miller,1979) は, 糖尿病に関する 145 名の対象者が医学的知見から正常群 (76 名), 臨床的症状を伴わない群 (36 名), 臨床的症状を伴う群に分類されている. また, Fisher-Anderson のアヤメデータは 150 個体,4 つの説明変数からなり,3 種のアヤメデータは, Iris setora,Iris versicolor,Iris virginia である. この 2 つのデータセットをコンポーネント数を 1 から 5 として, 多変量正規混合分布モデルにあてはめた.

4.4 分析結果

表 1 各データセットの分析結果

コンポーネント数	AIC	CAIC	BIC	EIC	CEIC
			糖尿病データ		
1	5482.1	5477.6	5508.9	5499.0	5496.7
2	5222.4	5212.9	5278.9	5239.6	5238.7
3	5136.5	5122.0	5222.8*	5159.5*	5161.2*
4	5124.9	5105.4	5241.0	5178.2	5178.4
5	5120.3*	5095.8*	5266.1	5208.4	5209.0
			アヤメデータ		
1	787.84	780.84	830.00	786.27	788.03
2	486.71	472.22	574.02*	494.80	496.93
3	448.37	426.37	580.84	466.96*	470.84*
4	433.53*	404.03*	611.16	491.12	491.06
5	452.83	415.83	675.61	555.65	556.19

糖尿病データでは, 真のコンポーネント数を推定したのは,BIC,EIC, 修正 EIC で, アヤメデータでは, EIC, 修正 EIC であった.EIC と修正 EIC が 2 つのデータとも真のコンポーネント数を推定したが, AIC と修正 AIC はいずれも多めに推定し,BIC は少なめに推定する傾向がある.

4.5 シミュレーション実験

4.5.1 実験の目的と設定

多変量正規混合分布のコンポーネント数の推定問題において, 各情報量基準を比較することを目的として, シミュレーション実験を行った. 比較する情報量基準は AIC, 修正 AIC,BIC,EIC, 修正 EIC で, コンポーネント数を 1 から 4 までを仮定した. 使用したデータは, あらかじめ構造のわかっているデータを乱数を用いて生成し,2 変数,2 つのコンポーネント分布に制限し, 表 2 の 3 種類のパラメータに従う 2 次元の正規混合分布から乱数を発生し作成した. 標本の大きさは,$n = 50, 200, 500$, シミュレーション回数は 500 回とし, 推定したコンポーネント数をカウントしていく.

表 2 シミュレーションのパラメータ設定

設定	μ_k	Σ_k	π
1	$(0\ \ 0)'$ $(0\ \ 3)'$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$
2	$(0\ \ 0)'$ $(0\ \ 4)'$	$\begin{pmatrix} 16 & 0 \\ 0 & 0.25 \end{pmatrix}$	$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$
3	$(0\ \ 0)'$ $(3\ \ 2)'$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\begin{pmatrix} 16 & 0 \\ 0 & 0.25 \end{pmatrix}$	$\begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix}$

4.6 分析結果と考察

表 3　分析結果

設定	コンポーネント数	AIC	CAIC	BIC	EIC	CEIC
		$n=50$				
1	1	40	7	35	202	425
	2	208	159	205	143	120
	3	167	181	168	105	38
	4	85	153	92	50	7
2	1	0	0	0	0	0
	2	209	85	200	450	460
	3	148	179	145	30	30
	4	143	236	155	20	10
3	1	0	0	0	0	5
	2	220	95	205	440	455
	3	148	159	152	40	30
	4	143	244	143	20	10
		$n=200$				
1	1	4	0	11	15	200
	2	303	221	349	328	195
	3	122	139	94	100	78
	4	71	140	46	57	27
2	1	0	0	0	0	0
	2	309	185	391	368	473
	3	94	133	51	85	25
	4	97	182	58	47	2
3	1	0	0	0	0	0
	2	288	170	396	405	450
	3	136	163	67	70	45
	4	76	167	37	25	5
		$n=500$				
1	1	0	0	0	0	0
	2	335	285	380	373	288
	3	95	110	78	127	125
	4	70	105	42	0	87
2	1	0	0	0	0	0
	2	273	198	318	332	497
	3	170	192	140	152	3
	4	57	110	42	16	0
3	1	0	0	0	0	0
	2	374	278	440	467	498
	3	81	127	38	33	2
	4	45	85	62	0	0

実データの解析と同様に,EIC, 修正 EIC が真のコンポーネント数 2 をよく推定する結果となり, AIC と修正 AIC は多めに推定する傾向が見られる. 設定 1 の共通の球型の分布のとき, シミュレーション回数に関係なくばらつきが見られる. 特に EIC, 修正 EIC の推定が悪い結果となっている. $n=50$ のとき, 全体的に推定にばらつきが見られるが, その中でも EIC と修正 EIC がよい結果となっており, ブートストラップ法の特性が生かされている.$n=200$ のとき,EIC,AIC,BIC に大きな差はない. また,$n=500$ のとき修正 EIC は特別よい結果となった. したがって, この推定問題に対しては修正 EIC が有効といえる.

5 混合回帰モデルのコンポーネント数と説明変数の数の推定問題

前節の多変量正規混合分布モデルのコンポーネント数の推定問題を混合回帰モデルのコンポーネント数と説明変数の数の推定問題へ拡張する. 前節の多変量正規混合分布モデルのコンポーネント数の推定問題において,AIC はコンポーネント数をより多く推定する傾向があることがわかり, この推定問題において, その傾向は正しくない説明変数を選択する結果になりかねない. Naik et al.[4] は, この推定問題に対する新しい情報量基準, 混合回帰基準 (MRC) を提案し, その優位性を主張している. そこで, この MRC について研究し,MRC 値を推定するプログラムを作成することを目的とし, シミュレーション実験を通してその有効性と問題点を示す.

5.1 混合回帰モデル

密度関数

$$f(y;\boldsymbol{x},\phi)=\sum_{k=1}^{K}\alpha_k f_k(y;\boldsymbol{x},\boldsymbol{\beta}_k,\sigma_k);\ \sum_{k=1}^{K}\alpha_k=1 \quad (16)$$

をもつ候補モデルを考える. ここで,$f(y;\boldsymbol{x},\phi)$ は平均 $\boldsymbol{x}'\boldsymbol{\beta}$ 分散 σ_k^2 をもつ正規密度で, $\boldsymbol{x}$ は $p\times 1$ ベクトルの説明変数,$\boldsymbol{\beta}_k$ は回帰係数ベクトル, $\phi=\{\alpha_k,\boldsymbol{\beta}_k,\sigma_k;k=1,\cdots,K\}$ である.

5.2 混合回帰基準 (MRC)

混合回帰基準は混合回帰モデルにおいて, 対数尤度のバイアスを評価することで

$$MRC=\sum_{k=1}^{K}\hat{n}_k\log(\hat{\sigma}_k^2)+\sum_{k=1}^{K}\frac{\hat{n}_k(\hat{n}_k+\hat{p}_k)}{\hat{n}_k-\hat{p}_k-2}-2\sum_{k=1}^{K}\hat{n}_k\log(\hat{\alpha}_k) \quad (17)$$

で与えられる. ここで,$\hat{p}_k,\hat{n}_k$ は第 k コンポーネントに対する説明変数の数と所属する要素の数である. MRC の第 1 項目は当てはまりの不足を測り, 候補モデルにおいて多数の変数を含むことで修正することができる. 第 2 項は過度の当てはめに対してペナルティーを加えることでバランスをとる.

5.3 推定アルゴリズム

混合回帰モデルにおいてコンポーネント数 K と説明変数の数 p を同時に決定するために次の手順を使う. はじめに,$\{(K,p):K=1,\cdots,5,p=1,\cdots,7\}$ を推定するために,kmean 法を使用し候補行列 $\boldsymbol{X}$ を分類し, 初期確率を推定する. そして, 混合回帰モデルを推定するために EM 法を応用する. 次にパラメータ推定量によって MRC を計算する.MRC の値が最小になる (K,p) の組み合わせが推定値となる.

5.4 シミュレーション実験

上記のアルゴリズムをもとにシミュレーション実験を行い,MRC の特性を示す. 各設定は以下に示す. シミュ

レーション回数は 1000 回で設定 1, 設定 2 は推定したコンポーネント数と, 説明変数の数の組 (K, p) の数をカウントし, 設定 3 は, 各コンポーネントの説明変数の数の組 (p_1, p_2) をカウントしたものである.

5.5 シミュレーションの設定

設定 1 説明変数の数が同じで, コンポーネント数 $K^0 = 3$ で, 各コンポーネントの標本サイズが $n_k^0 = 100$ の場合を考える. それぞれのコンポーネントは 4 つの説明変数 $(p^0 = 4)$ をもつ回帰モデルである. 真のパラメータは $b_1^0 = (1,1,1,1)', b_2^0 = (1,2,3,4)', b_3^0 = (5,6,7,8)'$. 真の説明変数は $n_k \times 4$ 行列で, $\boldsymbol{X}_1^0 \sim U(0,5), \boldsymbol{X}_2^0 \sim U(5,10), \boldsymbol{X}_3^0 \sim U(10,15)$ から生成した. それぞれのコンポーネントに対する応答変数は $\boldsymbol{Y}_k = \boldsymbol{X}_1^0 \boldsymbol{\beta}_k^0 + \boldsymbol{\epsilon}_k^0$ で, $\boldsymbol{\epsilon}_k^0 \sim N(\boldsymbol{0}, (\sigma_k^0)^2 \boldsymbol{I}_{nk})$ である. $\boldsymbol{X}_k$ は $n_k \times 7$ 行列で,1 列目から 4 列目には $\boldsymbol{X}_k^0$, 5 列目は $U(0,5)$,6 列目は $U(5,10)$,7 列目は $U(10,15)$ が格納されている.

設定 2 設定 1 の標本の大きさを少し変え,$n_1^0 = 50, n_2^0 = 75, n_2^0 = 100$ とし, 説明変数の数が同じで, コンポーネント数 3$(K^0 = 3)$ で, 標本の大きさが各コンポーネントにより異なる場合を考える.

設定 3 説明変数の数が異なる $p_k \neq p$, $K^0 = 2$,$n_k^0 = 100$ の場合を考える. k_1^0 は 2 つの説明変数 $(P_1^0 = 2)$,k_2^0 は 4 つの説明変数 $(P_1^0 = 4)$ をもつ回帰モデルである. 真のパラメータは $b_1^0 = (1,2)', b_2^0 = (5,6,7,8)'$. 真の説明変数はコンポーネント 1 は $n_k \times 2$ 行列で, コンポーネント 2 は $n_k \times 4$ 行列で $\boldsymbol{X}_1^0 \sim U(5,10), \boldsymbol{X}_2^0 \sim U(10,15)$ から生成した. それぞれのコンポーネントに対する応答変数は $\boldsymbol{Y}_k = \boldsymbol{X}_1^0 \boldsymbol{\beta}_k^0 + \boldsymbol{\epsilon}_k^0$ である. $\boldsymbol{X}_k$ は $n_k \times 5$ 行列で,1 列目から 4 列目には $\boldsymbol{X}_k^0$ が格納されており, 5 列目は $U(0,5)$, から生成された. 真のモデルを選択するために候補コンポーネント $1 \leq K \leq 3$, 候補説明変数の数 $1 \leq p \leq 5$ をを考える. この問題はすべての k に対して 155 の場合を考えなければならない. これを縮小するために, まず, 全ての変数を含むことによって最適なコンポーネント数を決定し, 選択されたコンポーネント数に対して最適な係数の組を決定する.

5.6 実行結果

設定 1 と設定 2 の実行結果より, 説明変数の数が同じ場合, 各コンポーネントの標本数に関係なくよい推定結果となった. したがって, 説明変数の数が同じと思われる観測データに対しては MRC は有効に働くことが言える. 設定 3 では第 1 ステップですべてがコンポーネント数を 2 と推定したが, ばらつきがあることがわかる. 真の説明変数の数の組は $(p_1, p_2) = (2,4)$ だが, 不思議とその逆である $(p_1, p_2) = (4,2)$ も多く推定されている. 実際の問題では設定 3 のように, 説明変数の数が異なる場合の方が多く, この MRC では最適なモデルを選択することができないかもしれない. また, コンポーネント数の推定に加え, データの次元が大きくなる場合, 説明変数の組ををすべて考えることはより複雑で大変な作業になる.

表 4　実行結果

	(K,p)	1	2	3	4	5	6	7
	1	0	0	0	0	0	0	0
設定 1	2	0	0	0	0	0	0	0
	3	0	0	0	896	80	20	4
	4	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0
設定 2	2	0	0	0	0	0	0	0
	3	0	0	0	813	184	2	0
	4	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	1
	(p_1, p_2)	1	2	3	4	5		
	1	0	0	0	0	0		
設定 3	2	0	0	0	319	65		
	3	0	0	0	56	2		
	4	0	275	56	27	30		
	5	0	51	4	18	5		

6 おわりに

本論文で取り扱っている情報量基準は, 従来から利用されているモデル選択の際の有効な手法であり, 次々と新しい情報量基準が提案されている. 本研究では, 情報量基準を用いて様々なモデルを評価することで各情報量基準の特徴や有効性を示すことができた. また, 有限混合回帰のコンポーネント数と説明変数の数の推定問題においては, パラメータを多く推定しなければならないので, 推定の精度はあまりよくならないと考えられるが, 新しい情報量基準 MRC は, 説明変数の数が同じとき有効な手段であるといえる. しかし説明変数の数が異なるとき効率があまりよくなかったので, 今後更なる研究が必要である.

参考文献

[1] 小西貞則・北川源四郎 (2004). シリーズ予測と発見の科学 2 情報量基準, 朝倉書店.

[2] 中村永友 (1995). 多変量正規混合分布に基づく分類法, 計算機統計学, 第 8 巻第 2 号.

[3] 中村永友・小西貞則 (1998). 情報量基準に基づく多変量正規混合分布モデルのコンポーネント数の推定, 応用統計学,Vol.27,No.3.

[4] Naik,P.A.,Shi,p.and Tsai,C.L(2007).Extending the Akaike Information Criterion to Mixture Regression Models,*Journal of the American Statistical Association,*Vol.102.

[5] Sugiura,N.(1978).Further analysis of the data by Akaike's information criterion and the finite corrections. *Communication in Statistics Series A*,Vol.7.