単純同齢林における林木成長パターンのクラスタリング

Clustering Individual Growth Patterns in an Even-Aged Forest Stand

柳原 宏和・吉本　敦
Yanagihara, H. & Yoshimoto, A.

キーワード：　*k*-平均法、情報量規準、成長分析、多変量分散分析モデル、モデル選択

要約：　同一林分内において局所的な地理条件の違いなどにより、個々の林木の成長は異なる。そのような場合、それぞれの林木の成長を同様のパターンにより分類し、分類されるグループ毎に成長を予測できれば、林分全体の成長をより効率的に記述することができる。本研究では、単純同齢林の林木成長に対し、*k*-平均法によりクラスタリングを行い、情報量規準を基に最適なグループ分割の抽出を試みた。ここで用いた方法は、林木の成長曲線のパラメータ推定、推定されるパラメータ値に対するクラスタリング、得られるクラスターを基にした多変量分散分析 (MANOVA) モデルのあてはめ、そして、あてはめ結果から得られる情報量規準による最適グループ分割の決定から構成される方法である。

Abstract:　A growth pattern of an individual tree differs due to such effects as a local geography condition and competition among trees. In such a case, clustering growth patterns of individual trees becomes effective to describe a whole forest stand growth. We proposed a clustering method for tree growth in an even-aged forest stand. Our method consists of 1) estimating parameters of a growth curve, 2) clustering a set of estimated parameters by the *k*-means method, 3) multivariate analysis of variance (MANOVA) on a set of the parameters, and 4) searching for an optimal number of clusters by the smallest information criterion.

Keywords: Growth analysis, Information criterion, k-means method, Multivariate analysis of variance model, Model selection

1. はじめに

同一林分内において局所的な地理条件の違いなどにより、個々の林木の成長は異なる。そのような場合、それぞれの林木の成長を同様のパターンにより分類 (クラスタリング) し、分類されるグループ (クラスター) 毎に成長を予測できれば、林分全体の成長をより効率的に記述することができる。本論文では、個別林木の成長パターンをクラスタリングする手法を提示する。

クラスタリングを行う際、データのクラスターによる分割方法とクラスターの個数の決定が重要な問題である。前者の問題に関しては、数々の研究者により様々な手法の研究が行われている (Gordon, 1999, Everitt, Landau & Leese, 2001 等の総合報告を参照) が、k-平均法 (MacQueen, 1967) が一般的な手法と言える。本論文では、従来よく用いられているトレースを用いた k-平均法ではなく、行列式を用いた k-平均法を用いる。後者のクラスター数の決定に関する問題については、観測値に k-平均法で分割されたクラスターを考慮に入れた異分散性 (Heterogeneity) と正規性を持つ一元配置多変量分散分析 (one-way MANOVA) モデルを仮定し、そのモデルにより情報量規準を計算し、クラスターの個数の最適化する手法を採用する。

本論文の構成は以下の通りである。第2章では、k-平均法を用いたクラスタリング法とそのアルゴリズムに関して述べる。第3章では、使用する正規異分散 MONOVA モデルと情報量規準の関係を述べる。第4章ではこの手法を用いた成長パターンの分類に関する適用例を示す。

2. k-平均法によるクラスタリング

n 個の個体には p 個の観測値があり、それぞれの個体に対して独立に観測値が得られたとし、その観測値のベクトル表示を $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})'$ $(i = 1, \ldots, n)$ とする。ここでの表記 $'$ は、ベクトルや行列の転置を表すものである。観測値には K 個のクラスターがあり、それぞれの個体はいずれかのクラスターに属しているとする。この K 個のクラスターへの各個体の分割

法として最も有名で簡便な手法が、MacQueen (1967) により提案された k-平均法である。

n 個の個体の K 個のクラスターへの分割を $G=\{C_1,\ldots,C_K\}$ とする。ただし、それぞれのクラスターに属する個体の数は $n_k(k=1,\ldots,K;\sum_{k=1}^{K} n_k=n)$ である。このとき、各クラスター $C_k(k=1,\ldots,K)$ の重心と共分散行列は、

[1] $$\overline{\boldsymbol{y}}_k=\frac{1}{n_k}\sum_{i\in C_k}\boldsymbol{y}_i$$

[2] $$\boldsymbol{S}_k=\frac{1}{n_k}\sum_{i\in C_k}(\boldsymbol{y}_i-\overline{\boldsymbol{y}}_k)(\boldsymbol{y}_i-\overline{\boldsymbol{y}}_k)'$$

である。ただし、$\sum_{i\in C_k}$ はクラスター C_k に属する個体だけをすべて足すということを表す記号である。k-平均法では、分散分析における郡内平方和と同様の、クラスター内平方和積和行列 $\boldsymbol{W}(G)$、

[3] $$\boldsymbol{W}(G)=\sum_{k=1}^{K} n_k\boldsymbol{S}_k,$$

の変化量の増減により個体 i が属するクラスターを決定する。

今、第 i 番目の個体がクラスター C_g に属しているような分割 G において、この個体をクラスター $C_h(h\neq g)$ に移動させるという状況を考える。このとき、新しいクラスター分割 $G^*=\{C_1,\ldots,C_h^*,\ldots,C_g^*,\ldots,C_K\}$ におけるクラスター内平方和積和行列 $\boldsymbol{W}(G^*)$ を計算して、個体 i をクラスター G_h に移動させる前の分割 G に基づく $\boldsymbol{W}(G)$ に較べ $\boldsymbol{W}(G^*)$ が小さくなっていれば分割を更新し、小さくなければ更新しないというアルゴリズムを考える。このようなアルゴリズムでクラスター分割の更新を逐次判定し、分割が収束するまで続け、最終的に収束した分割によりクラスターを決定する手法が k-平均法である。

クラスター分割を更新する前のクラスター内平方和積和行列 $\boldsymbol{W}(G)$ と分割を更新した後でのクラスター内平方和積和行列 $\boldsymbol{W}(G^*)$ の差の増減により分割を更新するか否かを決定する手法が k-平均法であるが、実際 $\boldsymbol{W}(G)$ と $\boldsymbol{W}(G^*)$ は行列であるので、大きさの測り方には様々な規準がある。一般的な手法は、行列のトレースにより大きさを測り、

[4] $\mathrm{tr}(\boldsymbol{W}(G)) > \mathrm{tr}(\boldsymbol{W}(G^*))$

であればクラスター分割を更新するというものである。しかしながら、このトレースによる更新判定は、単純に個体 i とクラスター重心とのユーグリット距離の大小で行われ、同一個体内の変数同士が強い相関を持つ場合、つまりクラスターが楕円状であるようなデータでは、この判定規準でうまくクラスターを分割することができない (Everitt, Landau & Leese, 2001, p. 95 の Figure 5.1 を参照)。これは、トレースで大きさを測るとき、単純にクラスター重心からのユーグリット距離が近い方のクラスターに属すると判定されるからである。また、楕円状のクラスターを持つデータでは、個体は必ずしも重心からのユーグリット距離が近い方のクラスターに属しているわけではないため、単純にユーグリット距離が近い方のクラスターに属すると判定されれば、うまくクラスターを分割することができないという状況が起こる。本論文で分析対象とするデータは、同一林分内の個別林木の成長曲線で推定される係数であるため、強い相関が観察されることが予想できる。そこでここでは、行列式により大きさを判断する手法を用いる。つまり、

[5] $|\boldsymbol{W}(G)| > |\boldsymbol{W}(G^*)|$

であればクラスター分割を更新するという手法である。

式 [5] からわかるように、クラスター分割を更新するか否かを判断するためには、新しい分割でのクラスター内平方和積和行列 $\boldsymbol{W}(G^*)$ が必要であり、それを得るためには、毎回個体の入れ替えを行わなければならない。そこで以下のような定理を考える。

定理1. n 個の個体の K 個のクラスターへの分割を $G=\{C_1,\ldots,C_K\}$ とし、第 i 番目の個体が属するクラスターを C_g とする。今、この個体をクラスター C_g から $C_h (h \neq g)$ に移動させたときの新しいクラスター分割 $G^*=\{C_1,\ldots,C_h^*,\ldots,C_g^*,\ldots,C_K\}$ を考える。それぞれの分割におけるクラスター内平方和積和行列 $\boldsymbol{W}(G)$ と $\boldsymbol{W}(G^*)$ の行列式の比、$|\boldsymbol{W}(G^*)|/|\boldsymbol{W}(G)|$ は、

[6] $$\boldsymbol{a}_g = \sqrt{\frac{n_g}{n_g - 1}}\, \boldsymbol{W}(G)^{-1/2}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_g)$$

[7] $$\boldsymbol{a}_h = \sqrt{\frac{n_h}{n_h+1}}\,\boldsymbol{W}(G)^{-1/2}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h)$$

とおくと、

[8] $$\frac{|\boldsymbol{W}(G^*)|}{|\boldsymbol{W}(G)|} = (1+\boldsymbol{a}_h'\boldsymbol{a}_h)(1-\boldsymbol{a}_g'\boldsymbol{a}_g) + (\boldsymbol{a}_g'\boldsymbol{a}_h)^2$$

と表すことができる。ただし $\overline{\boldsymbol{y}}_g$ と $\overline{\boldsymbol{y}}_h$ は式 [1] で与えられる、分割更新前のクラスター C_g と C_h の重心である。

定理 1 の証明については Appendix 1 に記した。この定理を用いると、不等式 $|\boldsymbol{W}(G)|>|\boldsymbol{W}(G^*)|$ と $(1+\boldsymbol{a}_h'\boldsymbol{a}_h)(1-\boldsymbol{a}_g'\boldsymbol{a}_g)+(\boldsymbol{a}_g'\boldsymbol{a}_h)^2<1$ は同値であることがわかる。式 [6] と [7] からわかるように、$\boldsymbol{a}_g$ と $\boldsymbol{a}_h$ には更新後のクラスター分割 G^* は必要ない。つまり式 [8] を用いれば、毎回分割を更新しなくても、分割更新の判定を行うことができる。

以上により、行列式に基づく k-平均法によるクラスタリングのアルゴリズムは、以下のような手順となる。

行列式を用いた k-平均法のアルゴリズム

Step 1. クラスターの個数 K をあらかじめ与え、初期分割を乱数により決定する。

Step 2. 第 i 番目の個体が所属するクラスターを C_g とする。今、個体 i をクラスター C_g から $C_h (h \neq g)$ に移動させたという状況を考える。このとき、

[9] $$(1+\boldsymbol{a}_h'\boldsymbol{a}_h)(1-\boldsymbol{a}_g'\boldsymbol{a}_g) + (\boldsymbol{a}_g'\boldsymbol{a}_h)^2 < 1$$

であれば個体 i が属するクラスターを C_g から C_h に更新する。また上記の関係式を満たさないときは、クラスター分割を更新しない。ただし $\boldsymbol{a}_g$ と $\boldsymbol{a}_h$ は式 [6] と [7] で与えられるベクトルである。

Step 3. Step 2 を $k = 1, \ldots, K$ (ただし $h \neq g$) で繰り返し、さらに $i = 1, \ldots, n$ で繰り返す。

Step 4. Step 3 をクラスター分割が収束するまで繰り返す。

Step 5. 初期配置を換え何度か Step 1 ～ Step 4 を繰り返し、それぞれ得られた分割でのクラスター内平方和積和行列の行列式 $|\boldsymbol{W}(G)|$ が最小となるクラス

ター分割を最適な分割とする。

k-平均法は、クラスター内平方和積和行列の行列式が最小になるようにクラスター分割 G を更新する手法であり、必ずどこかしらの最小値に収束するが、必ずしもそれが真の最小値である保証がない。そこで、ここでは初期分割を変え複数回繰り返す Step 5 が必要となる。

3. 情報量規準を用いたクラスター数の決定法

クラスターの個数は解析者が勝手に決めることができるため、その値により解析結果は変わってくる。そこで、最適な K を決める具体的な規準が必要となる。ここでは、情報量規準を用いたクラスターの個数の決定法を用いる。

今、n 個の個体の K 個のクラスターへの分割 $G=\{C_1,\ldots,C_K\}$ が決まっており、第 i 番目の個体はクラスター $C_k(k=1,\ldots,K)$ に属している $(i\in C_k)$ とする。このとき、$\boldsymbol{y}_i$ に以下のような K 個のグループを持つ正規異分散多変量分散分析 (正規異分散 MANOVA) モデルを仮定する。

[10] $$M_K:\boldsymbol{y}_i \sim i.d.\ \mathrm{N}_p(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\quad (i=1,\ldots,n,\ k=1,\ldots,K)$$

ここで、$\boldsymbol{\mu}_k$ は k 番目のグループの平均を表す $p\times 1$ ベクトルであり、$\boldsymbol{\Sigma}_k$ はそのグループでのばらつきを表す $p\times p$ 分散共分散行列である。このモデル [10] が従う正規分布の確率密度関数は、

[11] $$f(\boldsymbol{y}_i|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=(2\pi)^{-p/2}|\,\boldsymbol{\Sigma}_k|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_k)\right\}$$

である。このとき、モデル M_K [10] での対数尤度関数は以下のようになる。

[12] $$\begin{aligned}&\ell(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K;\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K)\\&=\sum_{k=1}^{K}\sum_{i\in C_k}\log\{f(\boldsymbol{y}_i|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\}\\&=-\frac{1}{2}\left\{np\log(2\pi)+\sum_{k=1}^{K}n_k\log|\,\boldsymbol{\Sigma}_k|+\sum_{k=1}^{K}\sum_{i\in C_k}(\boldsymbol{y}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_k)\right\}\end{aligned}$$

今、$\mathbf{0}$ と $\boldsymbol{O}$ をすべての成分が 0 であるような零ベクトルと零行列とする。このとき、対数尤度関数 [12] から、それぞれの未知パラメータの最尤推定量

は、偏微分に基づく方程式、

[13] $$\frac{\partial}{\partial \boldsymbol{\mu}_k} \ell(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K) = \mathbf{0}$$

[14] $$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ell(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K) = \boldsymbol{O}$$

を解くと得ることができる。実際に式 [13] と [14] を解くと、k 番目のグループでの平均ベクトルと共分散行列の最尤推定量は、先の章の式[1]のクラスター C_k での重心 $\overline{\boldsymbol{y}}_k$ と式 [2] での共分散行列 $\boldsymbol{S}_k$ に一致することがわかる。

以上のことから、モデル M_K における *AIC* (Akaike Information Criterion, Akaike, 1973) は以下のようになる。

[15]
$$AIC_K = -2\ell(\overline{\boldsymbol{y}}_1, \ldots, \overline{\boldsymbol{y}}_K; \boldsymbol{S}_1, \ldots, \boldsymbol{S}_K) + 2\sum_{k=1}^{K}\left\{p + \frac{1}{2}p(p+1)\right\}$$
$$= \sum_{k=1}^{K} n_k \log|\boldsymbol{S}_k| + np\{\log(2\pi) + 1\} + p(p+3)K$$

モデル M_K でのグループ数はクラスターの個数である。その個数を $K = 1, \ldots, m$ と変化させると、モデルも $M_1, \ldots, M_m$ と m 個存在する。それぞれのモデル $M_1, \ldots, M_m$ の中で最小な *AIC* を持つモデルは、最適なグループ数を持つモデルであり、そのグループ数はそのまま最適なクラスター数となる。つまり、*AIC* が最小になるモデルでのグループ数はそのまま最適なクラスター数となる。しかしながら、*AIC* をクラスター数の最適化に用いるには少し問題がある。式 [15] で定義した *AIC* はカルバック=ライブラーの予測距離によるリスクに関する推定量である。しかしながら、*AIC* は、未知パラメータの個数 (ここではグループ数) が増加すればするほどリスクに対して過小推定をしてしまい、その結果、未知パラメータ数が多いモデルを選び易いと結果に陥る (カルバック=ライブラーの予測距離に基づくリスクと情報量規準の関係は Appendix 2 を参照)。モデル M_K では異分散性を仮定しているので、グループ数が増える毎に、未知パラメータ数も $\{p + p(p+1)/2\}$ 個づつ増える。その結果、比較的グループ数が少ないモデルでも未知パラメータ数は多くなり *AIC* の欠点が露骨に表れてしまうことになり、*AIC* を用いてモデルの最適化を行うことに問題が発生してしまう。

このような *AIC* が持つ欠点を改良したものが、Sugiura (1976)、Hurvich &

Tsai (1989)、Fujikoshi & Satoh (1997) 等で紹介されている修正 *AIC* (Corrected *AIC*、*CAIC* とも呼ばれる) であり、モデル M_K における *CAIC* は以下のように定義される。

[16] $$CAIC_K = \sum_{k=1}^{K} n_k \log |\boldsymbol{S}_k| + np \log (2\pi) + \sum_{k=1}^{K} \frac{n_k (n_k + 1) p}{n_k - p - 2} \quad (n_k > p + 2)$$

本論文ではこちらの *CAIC* を使用し、最適なクラスター数の決定を行う。その方法は以下の通りである。

最適なクラスター数の決定に関するアルゴリズム

Step 1. 最大クラスター数 m を決定する。

Step 2. クラスター数 K を固定して k–平均法を用いて最適なクラスター分割を求め、その分割に基づきデータにモデル M_K [10] をあてはめ、そのモデルでの $CAIC_K$ を計算する。

Step 3. クラスター数 K を 1 から m まで動かし、それぞれのクラスターの個数での *CAIC* を求める。

Step 4. 求められた $CAIC_K (K = 1, \ldots, m)$ を比較し、最も小さくなるときの K を最適なクラスターの個数とする。つまり最適なクラスター数 K_{opt} は、

[17] $$K_{\text{opt}} = \arg \min_{K=1, \ldots, m} CAIC_K$$

である。

上記のアルゴリズムには、$K = 1$ のときが含まれている。$K = 1$ とは、すべての個体が同一のクラスターに属しているということを意味し、モデル M_1 は、グループ層別を行っていないモデルとなる。そのときの $CAIC_1$ は、全てのデータを使った平均と共分散行列の推定量を、

[18] $$\bar{\boldsymbol{y}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i \quad \boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})'$$

とすると、

[19] $$CAIC_1 = n \log |\boldsymbol{S}| + np \log (2\pi) + \frac{n (n + 1) p}{n - p - 2} \quad (n > p + 2)$$

となる。このモデルを比較対照に導入することにより、クラスター分割を行うべきか否かという判断も同時に行っていることになる。

4. 実データへの適用

ここでは、実際の成長データを使った分析結果を用いてクラスタリングを行う。今回解析に使用したデータは、柳原他 (2004) で用いた、福岡県八女群星野村における23年生の無間伐林より抽出した30本から得た成長データである。試験林の形状は図1に示す通りで、プロット内の林木の総数は136本であった。図1の◦は林木立木位置を示し、その大きさは観測時点での胸高直径の相対的な大きさに対応している。ここでは、黒く塗りつぶした円は伐採されたサンプル木、白い円は残存木を表している。また図中の番号は個々の林木に個別に付けたID番号に対応している。

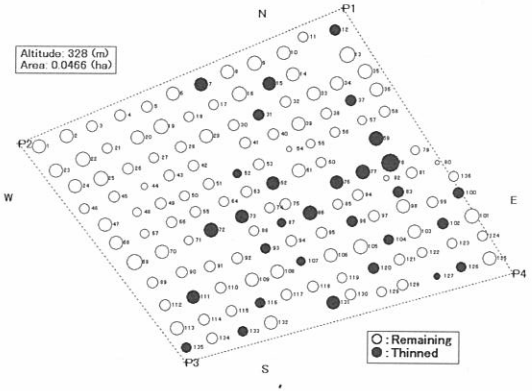

図1. 試験林内の立木位置

30本のサンプルに対しそれぞれ樹幹解析 (Philip, 1994 等参照) を行い、材積の成長データを算出した。そのデータをプロットしたものが図2である。表中のIDは図1のID番号と対応するものである。この図を見るとID No. 78や86の林木のように、12林齢付近で急激に成長する2次関数的な成長パターンと、ほぼ一定に成長する直線的な成長パターンの二種類のパターンが混在しているように見られる。

このような成長データに対し、リチャーズの成長関数 (Richards, 1958) を個々に当てはめ、成長曲線の係数を推定する。なお、このリチャーズの関数をシグモイド型に制約するため、柳原他 (2004) と同様にパラメータに指数変換を行った。つまり、以下のような成長関数を用いることになる。

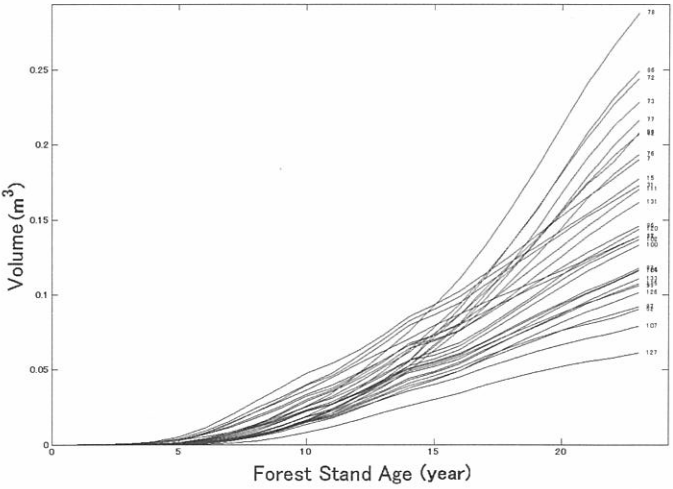

図2. 材積成長データ

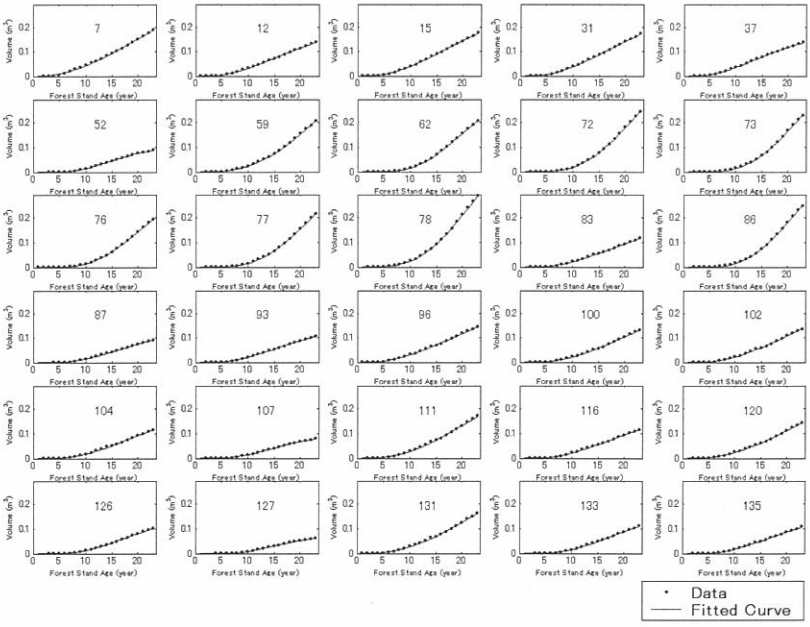

図3. 個々の林木での材積推定成長曲線

表1. 伐採木での材積成長曲線の推定係数

ID	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
7	-0.5056	-3.1368	0.9261
12	-1.4031	-2.5029	1.2579
15	-1.0583	-2.6144	1.2035
31	-1.0562	-2.6178	1.2260
37	-1.6105	-2.2664	1.3671
52	-2.1441	-1.9123	2.0854
59	0.3470	-3.2745	1.2568
62	-0.8119	-2.2821	2.0159
72	-0.2948	-2.5451	1.8091
73	-0.2356	-2.5771	1.8498
76	-0.6876	-2.4175	1.9173
77	-0.2452	-2.6473	1.7606
78	0.1515	-2.7837	1.6081
83	-1.4628	-2.5069	1.4120
86	-0.3324	-2.4380	1.9754
87	-1.9586	-2.1432	1.8145
93	-1.7784	-2.2757	1.5594
96	-1.0334	-2.8046	1.1386
100	-1.0118	-2.7627	1.3205
102	-1.0274	-2.7299	1.3268
104	-1.4438	-2.4863	1.4999
107	-2.2087	-2.0391	1.8715
111	-0.6611	-2.8697	1.2538
116	-1.5193	-2.4582	1.4489
120	-0.7252	-2.9622	1.2041
126	-1.6828	-2.2844	1.7819
127	-2.5377	-1.8313	2.3111
131	-0.5030	-3.0525	1.1614
133	-1.5611	-2.3579	1.6752
135	-1.6306	-2.3656	1.5929
平均	-1.0875	-2.5315	1.5544
標準偏差	0.7164	0.3406	0.3388

[20] $$e^{\theta_1}\left\{1-\exp\left(-e^{\theta_2}t_{i1}\right)\right\}^{\exp(\theta_3)}$$

このような変換を作用させることにより、パラメータ空間は $(0,\infty)$ から $(-\infty,\infty)$ となる。パラメータ推定に関しては、スパイダーアルゴリズム (柳原他, 2004) を用いた。推定された結果を図3に示す。推定された係数の値は表1に示す通りである。また、表中のIDは図1のID番号と対応するものである。

この得られた係数の推定量を観測値 $\boldsymbol{y}$ とし、成長パターンのクラスタリングを行う。実際には、第 i 番目の林木での推定された成長曲線の係数をそれぞれ $\hat{\theta}_{i1}$、$\hat{\theta}_{i2}$、$\hat{\theta}_{i3}$ とすると、$\boldsymbol{y}_i=(y_{i1},y_{i2},y_{i3})'=(\hat{\theta}_{i1},\hat{\theta}_{i2},\hat{\theta}_{i3})'$ となる。これらの散布図を図4に示す。図中の番号は図1のID番号と対応するものである。この散布図を見れば、データは二本の直線上に乗っているように見える。とくに、y_1 と y_2、y_1 と y_3 に関する散布図においてその傾向は強い。これによりデータには二つのグループがあることがはっきり予想される。

この $\boldsymbol{y}_i$ に対して、前述したクラスタリング手法を適用する。最大クラスター数 m を4として、それぞれのクラスターの個数において行列式に基づく k-平均法によりクラスター分割を行い、その分割を基に $\boldsymbol{y}_i$ に正規異分散MANOVAモデルをあてはめ、そのモデルでの情報量規準 *CAIC* を計算した。それぞれのクラスターの個数における情報量規準 *CAIC* を図5に示す。この

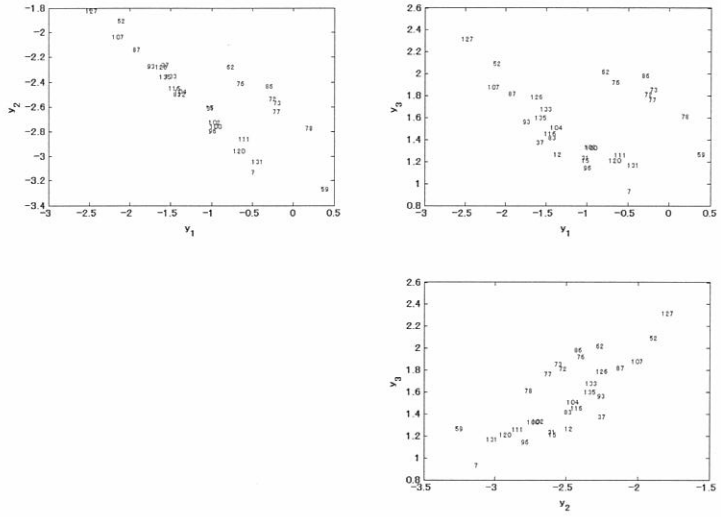

図4. y_1、y_2、y_3の散布図

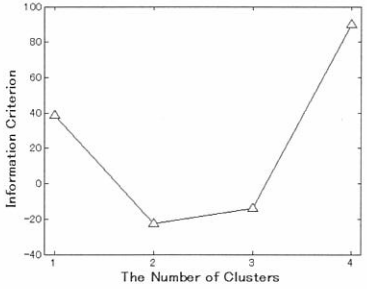

図5. クラスターの個数と*CAIC*

図より最適なクラスター数は2であることがわかる。図6に最終的に得られた最適なクラスター分割を示す。この図は、図4と同じ$\boldsymbol{y}_i$に関する散布図であり、図中において、•は第1クラスターに属する個体、×は第2クラスターに属する個体を表している。また、この最適なクラスター分割により、図2の成長データをグループに別けて表示すると図7のようになる。ここで、点線は第1クラスターに属する林木の成長データ、実線は第2クラスターに属する林木の成長データである。この図から、12年目付近で急激に成長す

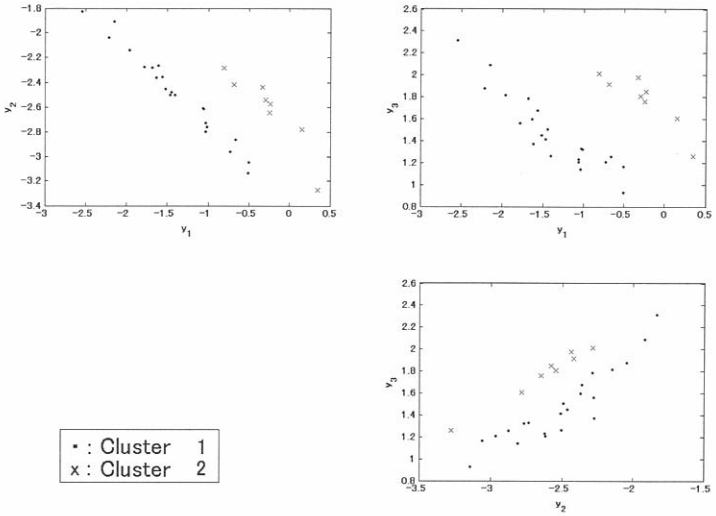

図6. 最適なクラスター

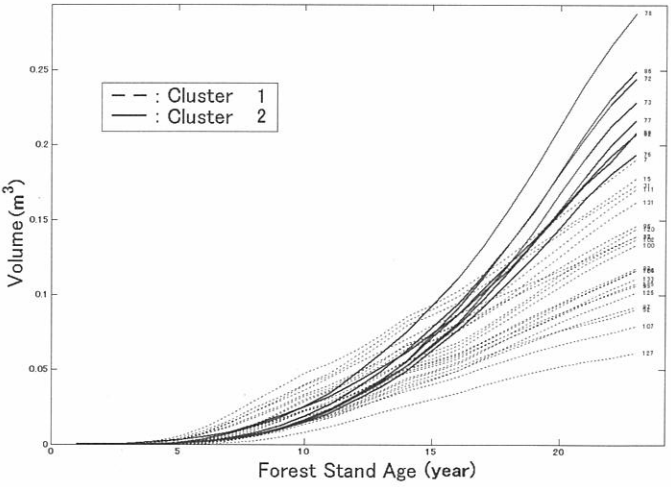

図7. 最適なクラスターによる成長データの分割表示

る2次関数の様な成長パターンと、ほぼ一定に成長する直線の様な成長パターンの2種類の成長が混在していることが分かる。

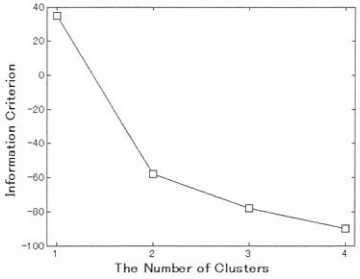

図8．クラスターの個数と*AIC*

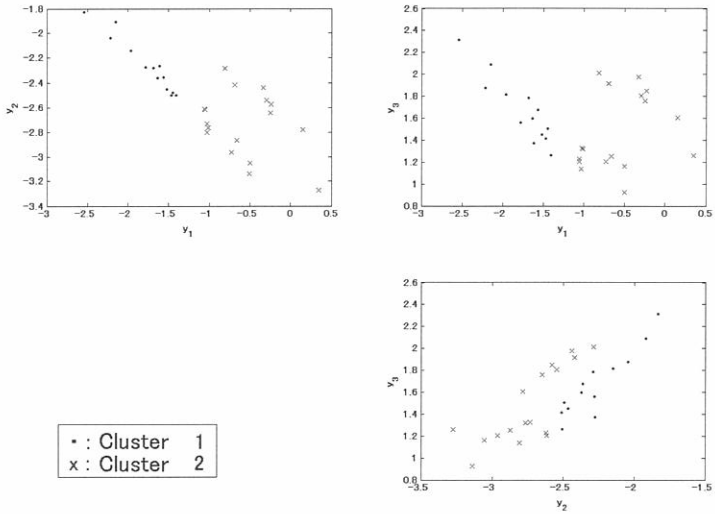

図9．tr $(\boldsymbol{W}(G))$を用いたクラスタリング結果 $(K=2)$

最後に他の手法でクラスター分析を行った結果を示す。まず最適なクラスターの個数の決定に*CAIC*ではなく*AIC*を使った場合の結果を図8に示す。図5と同じように最大クラスター数を4として解析を行った。この場合、最適なクラスター数は4となっている。これは第3章で述べたように、*AIC*はパラメータ数が大きいモデルを選び易い傾向があるため、最適なクラスターを4と選んでしまったという典型的な例である。

次にクラスター内平方和積和行列 $\boldsymbol{W}(G)$ と $\boldsymbol{W}(G^{*})$ の差を測る規準に、行

列式ではなくトレースを用いて k-平均法を行った結果を図9に示す。クラスター数は $K=2$ である。トレースで最適化を行った場合、単純にデータ間のユーグリット距離が近いもの同士を同一クラスターと決定してしまう。林分成長データの様に、強い相関を持つデータでは、まったくうまく機能していないことが観察できる。

5. おわりに

本論文では、同一林分内の個々の林木を複数の成長パターンによりクラスタリングし、林分成長を把握する方法を提示した。クラスター分割の最適化には、行列式に基づいた k-平均法を用い、クラスターの個数の最適化には、正規異分散 MANOVA モデルとそのモデルでの情報量規準 *CAIC* を用いた。近年、林分成長データに対する多変量モデルの開発が行われている (柳原・吉本, 2003, Yanagihara & Yoshimto, 2004, 柳原他, 2004等)。それらすべてのモデルでは、同一林分内に複数の成長パターンがある場合、そのグループ別けは既知でなければデータに適用することができない。そのため、これらのモデルを用いる際には、成長パターンによるグループを把握することが必要不可欠となる。本論文で提示した手法はそのような場合に有効になるものと考えられる。

引用文献

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *2nd. International Symposium on Information Theory* (B. N. Petrov & F. Csáki Eds.), 267-281, Akadémia Kiado, Budapest.

Everitt, B. S., Landau, S. & Leese, M. 2001. *Cluster Analysis* (4th. ed.). Edward Arnold, London.

Fujikoshi, Y. & Satoh, K. 1997. Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707-716.

Gordon, A. D. 1999. *Classification* (2nd. ed.). Chapman & Hall/CRC, New York.

Hurvich, C. M. & Tsai, C. L. 1989. Regression and times series model selection in small samples. *Biometrika*, **50**, 226-231.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman Ed.), **1**, 281-298, Berkeley.

Richards, F. J. 1958. A flexible growth function to empirical use. *J. Exp. Bot.*, **10**, 290-300.

Siotani, M., Hayakawa, T. & Fujikoshi, Y. 1985. *Modern multivariate statistical analysis: a graduate course and handbook*. American Sciences Press, Columbus, Ohio.

Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. Theory Methods*, **7**, 13-26.

柳原宏和, 吉本 敦 2003. 一般化多変量分散分析モデルの林木直径成長分析への適用可能性. 統計数理, **51**, 19-35.

Yanagihara, H. & Yoshimoto, A. 2004. Statistical procedure for assessing the amount of carbon sequestrated by sugi (*Cryptomeria japonica*) plantation. *Discussion Paper Series* No. **1076**, Institute of Policy and Planning Sciences, University of Tsukuba, Tsukuba.

柳原宏和・吉本 敦・能本美穂 2004. 林分成長分析のための一般化非線形混合効果モデル. 森林資源管理と数理モデル Vol.3 (鹿又秀聡・吉本 敦 編集), 14-46, 森林計画学会出版局, 東京.

Appendices

A. 1. 定理1の証明

ここでは、定理1の式 [8] の導出法を述べる。まず個体 i をクラスター C_g から C_h に移動させると、新しいクラスター C_g^* と C_h^* での重心はそれぞれ

$$[21] \qquad \overline{\boldsymbol{y}}_g^* = \frac{1}{n_g - 1}(n_g \overline{\boldsymbol{y}}_g - \boldsymbol{y}_i) \quad \overline{\boldsymbol{y}}_h^* = \frac{1}{n_h + 1}(n_h \overline{\boldsymbol{y}}_h + \boldsymbol{y}_i)$$

となり、新しい重心と観測値 $\boldsymbol{y}_j (j = 1, \ldots, n)$ との差はそれぞれ

$$[22] \qquad \begin{aligned} \boldsymbol{y}_j - \overline{\boldsymbol{y}}_g^* &= \boldsymbol{y}_j - \overline{\boldsymbol{y}}_g + \frac{1}{n_g - 1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_g) \\ \boldsymbol{y}_j - \overline{\boldsymbol{y}}_h^* &= \boldsymbol{y}_j - \overline{\boldsymbol{y}}_h - \frac{1}{n_h + 1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h) \end{aligned}$$

となる。特に、$j = i$ の時、

$$[23] \qquad \boldsymbol{y}_i - \overline{\boldsymbol{y}}_g^* = \frac{n_g}{n_g - 1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_g) \quad \boldsymbol{y}_i - \overline{\boldsymbol{y}}_h^* = \frac{n_h}{n_h + 1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h)$$

となることに注意して欲しい。式 [22] を用いると、個体 i をクラスター C_g から C_h に移動させたときのクラスター内共分散行列は、

$$[24] \qquad \begin{aligned} \boldsymbol{S}_g^* &= \frac{1}{n_g - 1}\left\{ \sum_{j \in C_g}^{n} (\boldsymbol{y}_j - \overline{\boldsymbol{y}}_g^*)(\boldsymbol{y}_j - \overline{\boldsymbol{y}}_g^*)' - (\boldsymbol{y}_i - \overline{\boldsymbol{y}}_g^*)(\boldsymbol{y}_i - \boldsymbol{y}_g^*)' \right\} \\ &= \frac{1}{n_g - 1}\left\{ n_g \boldsymbol{S}_g - \frac{n_g}{n_g - 1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_g)(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_g)' \right\} \end{aligned}$$

$$[25] \qquad \begin{aligned} \boldsymbol{S}_h^* &= \frac{1}{n_h + 1}\left\{ \sum_{j \in C_h}^{n} (\boldsymbol{y}_j - \overline{\boldsymbol{y}}_h^*)(\boldsymbol{y}_j - \overline{\boldsymbol{y}}_h^*)' + (\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h^*)(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h^*)' \right\} \\ &= \frac{1}{n_h + 1}\left\{ n_h \boldsymbol{S}_h + \frac{n_h}{n_h + 1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h)(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h)' \right\} \end{aligned}$$

となる。今、クラスター内平方和積和行列は $\boldsymbol{W}(G) = \sum_{k=1}^{K} n_k \boldsymbol{S}_k$ であり、g と h 以外のクラスターの重心は変化しないことに注意すれば、個体 i を移動させた後のクラスター内平方和積和行列は、

[26] $$\boldsymbol{W}(G^*) = \sum_{k=1}^{K} n_k^* \boldsymbol{S}_k^* = \sum_{k \neq g,h}^{K} n_k \boldsymbol{S}_k + (n_g - 1)\boldsymbol{S}_g^* + (n_h + 1)\boldsymbol{S}_h^*$$

となる。この式 [26] に式 [24] と [25] を代入すれば、

[27] $$\boldsymbol{W}(G^*) = \boldsymbol{W}(G) + \frac{n_h}{n_h + 1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h)(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_h)' - \frac{n_g}{n_g - 1}(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_g)(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_g)'$$

となることがわかる。ここで、第2章の式 [6] と [7] でのベクトル $\boldsymbol{a}_g$ と $\boldsymbol{a}_h$ を用いると、式 [27] の両辺の行列式は、

[28] $$|\boldsymbol{W}(G^*)| = |\boldsymbol{W}(G)|\,|\,\boldsymbol{I}_p + \boldsymbol{a}_h\boldsymbol{a}_h' - \boldsymbol{a}_g\boldsymbol{a}_g'\,|$$

となる。今、$\boldsymbol{D} = \boldsymbol{I}_p + \boldsymbol{a}_h\boldsymbol{a}_h'$ とおくと、

[29] $$|\boldsymbol{I}_p + \boldsymbol{a}_h\boldsymbol{a}_h' - \boldsymbol{a}_g\boldsymbol{a}_g'| = |\boldsymbol{D}|\,|\,\boldsymbol{I}_p - \boldsymbol{D}^{-1/2}\boldsymbol{a}_g\boldsymbol{a}_g'\boldsymbol{D}^{-1/2}| = |\boldsymbol{D}|\left(1 - \boldsymbol{a}_g'\boldsymbol{D}^{-1}\boldsymbol{a}_g\right)$$

であり、

[30] $$|\boldsymbol{D}| = 1 + \boldsymbol{a}_h'\boldsymbol{a}_h \quad \boldsymbol{D}^{-1} = \boldsymbol{I}_p - \frac{\boldsymbol{a}_h\boldsymbol{a}_h'}{1 + \boldsymbol{a}_h'\boldsymbol{a}_h}$$

であることに注意すれば、

[31] $$|\boldsymbol{I}_p + \boldsymbol{a}_h\boldsymbol{a}_h' - \boldsymbol{a}_g\boldsymbol{a}_g'| = (1 + \boldsymbol{a}_h'\boldsymbol{a}_h)(1 - \boldsymbol{a}_g'\boldsymbol{a}_g) + (\boldsymbol{a}_g'\boldsymbol{a}_h)^2$$

となる (行列式、逆行列の公式は、それぞれ、Siotani, Hayakawa & Fujikoshi, 1985, p. 591 の A. 1. 3, p. 592 の A. 2. 2 等参照)。よって式 [31] を式 [28] に代入すると、

[32] $$|\boldsymbol{W}(G^*)| = |\boldsymbol{W}(G)|\left\{(1 + \boldsymbol{a}_h'\boldsymbol{a}_h)(1 - \boldsymbol{a}_g'\boldsymbol{a}_g) + (\boldsymbol{a}_g'\boldsymbol{a}_h)^2\right\}$$

となり、式 [32] の両辺を $|\boldsymbol{W}(G)|$ で割ることにより式 [8] を得ることができる。

A. 2. カルバック=ライブラーの予測距離に基づくリスクと情報量規準

ここでは、モデル M_K [10] でのカルバック=ライブラーの予測距離に基づくリスクと情報量規準の関係を述べる。今、観測値 $\boldsymbol{y}_i$ は以下のような真のモデルに従っていると仮定する。

[33] $$M^*: \boldsymbol{y}_i \sim i.d.\,\mathrm{N}_p(\boldsymbol{\eta}_i^*, \boldsymbol{\Psi}_i^*) \quad (i = 1, \ldots, n)$$

もちろん、この真のモデルは未知である。今、$\boldsymbol{u}_i$ を $\boldsymbol{y}_i$ と独立に同一な分布に従う確率変数 (未来の観測値) とし $\boldsymbol{Y}=(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)'$、$\boldsymbol{U}=(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)'$ とする。このとき、モデル M_K でのカクバック=ライブラーの予測距離に基づくリスクは、

$$R_K=-2\sum_{k=1}^{K}\sum_{i\in C_k}\mathrm{E}_U^*\mathrm{E}_Y^*\left[\log\{f(\boldsymbol{u}_i|\,\overline{\boldsymbol{y}}_k,\boldsymbol{S}_k)\}\right]$$

[34]
$$=\sum_{k=1}^{K}n_k\mathrm{E}_Y^*\left[\log|\,\boldsymbol{S}_k|\right]+np\log(2\pi)$$
$$+\sum_{k=1}^{K}\sum_{i\in C_k}\mathrm{E}_Y^*\left[\mathrm{tr}\,(\boldsymbol{S}_k^{-1}\boldsymbol{\Psi}_i^*)+(\overline{\boldsymbol{y}}_k-\boldsymbol{\eta}_i^*)'\boldsymbol{S}_k^{-1}(\overline{\boldsymbol{y}}_k-\boldsymbol{\eta}_i^*)\right]$$

となる。ただし、$\overline{\boldsymbol{y}}_k$ と $\boldsymbol{S}_k$ は式 [1] と [2] により定義されるクラスター C_k における重心と共分散行列、また、f は式 [11] で定義される密度関数、E_Y^* と E_U^* は、真のモデル M^* [33] の下での $\boldsymbol{Y}$ と $\boldsymbol{U}$ に関する期待値を表す。モデル選択問題における最適なモデルとは、このリスク R_K [34] が最小になるモデルである。しかしながら、このリスク R_K には未知パラメータが含まれるため、直接評価する事ができない。そのため、R_K の推定量を考え、その推定量により R_K を評価することになる。この R_K の推定量が情報量規準と呼ばれるものである。リスク R_K の最も簡単な推定量は、対数尤度関数 ℓ [12] を用いて、

[35]
$$-2\ell(\overline{\boldsymbol{y}}_1,\ldots,\overline{\boldsymbol{y}}_K;\boldsymbol{S}_1,\ldots,\boldsymbol{S}_K)=\sum_{k=1}^{K}n_k\log|\,\boldsymbol{S}_k|+np\left\{\log(2\pi)+1\right\}$$

と表される。さらに、この推定量には定数バイアスが存在し、そのバイアスは、

$$B_K=R_K-\left\{-2\mathrm{E}_Y^*[\ell(\overline{\boldsymbol{y}}_1,\ldots,\overline{\boldsymbol{y}}_K;\boldsymbol{S}_1,\ldots,\boldsymbol{S}_K)]\right\}$$

[36]
$$=\sum_{k=1}^{K}\sum_{i\in C_k}\mathrm{E}_Y^*\left[\mathrm{tr}\,(\boldsymbol{S}_k^{-1}\boldsymbol{\Psi}_i^*)+(\overline{\boldsymbol{y}}_K-\boldsymbol{\eta}_i^*)'\boldsymbol{S}_k^{-1}(\overline{\boldsymbol{y}}_K-\boldsymbol{\eta}_i^*)\right]-np$$

である。このバイアス B_K [36] をその推定量 $\hat{B}_K$ で補正した

[37] $-2\ell(\overline{\boldsymbol{y}}_1,\ldots,\overline{\boldsymbol{y}}_K;\boldsymbol{S}_1,\ldots,\boldsymbol{S}_K)+\hat{B}_K=\sum_{k=1}^{K}n_k\log|\boldsymbol{S}_k|+np\{\log(2\pi)+1\}+\hat{B}_K$

で情報量規準は定義される。

真のモデル M^* と候補のモデル M_K が一致する、すなわち $\boldsymbol{\eta}_i^*=\boldsymbol{\mu}_k$、$\boldsymbol{\Psi}_i^*=\boldsymbol{\Sigma}_k$ のとき、B_K の n に関するテーラー展開は、

[38] $$B_K=p(p+3)K+\mathrm{O}(n^{-1})\quad(n\to\infty)$$

となる。ただし $\mathrm{O}(n^{-1})$ は n が大きくなれば0に近づく定数項であり、$n\mathrm{O}(n^{-1})$ は $n\to\infty$ のとき定数に収束する。Akaike (1973) により提案された *AIC* は、[38] のテーラー展開式で $\mathrm{O}(n^{-1})$ の項を無視して $B_K\approx p(p+3)K$ と近似し、$\hat{B}_K=p(p+3)K$ と定義している。そのため、モデル M_K での AIC_K は式 [15] のようになる。ところが、$B_K\approx p(p+3)K$ では、標本数 n が小さいとき無視した $\mathrm{O}(n^{-1})$ の項が大きくなり、その近似が悪くなり、結果、*AIC* のリスクに対するバイアスも大きくなる。簡単な例として、真のモデル M^* と候補のモデル M_K が一致しているときを考えてみよう。このとき、$\overline{\boldsymbol{y}}_k$ と $\boldsymbol{S}_k(k=1,\ldots,K)$ は独立であり、それぞれ、

[39] $$\sqrt{n_k}(\overline{\boldsymbol{y}}_k-\boldsymbol{\mu}_k)\sim i.d.\,\mathrm{N}_p(\boldsymbol{0},\boldsymbol{\Sigma}_k)\quad n_k\boldsymbol{S}_k\sim i.d.\,\mathrm{W}_p(n_k-1,\boldsymbol{\Sigma}_k)$$

である。ただし $\mathrm{W}_p(m,\boldsymbol{A})$ は、平均行列 $\boldsymbol{A}$ を持つ自由度 m の p 次元ウィッシャート分布を示す (ウィッシャート分布については、Siotani, Hayakawa & Fujikoshi, 1985, p. 59-109 等参照)。[39] 式とウィッシャート分布に従う確率変数行列の逆行列の期待値の公式 (Siotani, Hayakawa & Fujikoshi, 1985, p. 74 の Theorem 2. 4. 6 等参照) により、B_K は、

[40] $$B_K=\sum_{k=1}^{K}\frac{n_k(n_k+1)p}{n_k-p-2}-np\quad(n_k>p+2)$$

と計算できる。今、不等式、

[41] $$\frac{n_k}{n_k-p-2}>1+\frac{1}{n_k}(p+2)\quad\sum_{k=1}^{K}\frac{1}{n_k}>\frac{1}{n}K^2$$

を用いると、式 [40] は、

[42] $$B_K > p(p+3)K + \frac{1}{n}p(p+2)K^2$$

となる。[42] 式はバイアス B_K の下限値である。この式を用いると、*AIC* のリスク R_K に対するバイアスは、

[43] $$B_{AIC_K} = R_K - \mathrm{E}_Y^*[AIC_K] = B_K - p(p+3)K > \frac{1}{n}p(p+2)K^2$$

と評価できる。これにより、*AIC* の期待値はリスクに対して小さくなること、その相違はクラスター数 K が大きくなるほど増大することがわかる。その結果、AIC_K は K が大きくなるほどリスク R_K に対して過小推定をしてしまい、K が大きいモデルを選び易くなる。

B_K の推定量として近似値 $p(p+3)K$ を用いているため、*AIC* はリスクに対して過小推定してしまう。そこで、$\hat{B}_K$ として式 [40] での正確な値を用い、バイアスを補正した情報量規準が *CAIC* である。そのため、モデル M_K での $CAIC_K$ は式 (14) のようになる。この *CAIC* はモデル M_K が真のモデルを含む (overspecified model: 詳しい定義は、Fujikoshi & Satoh, 1997 参照) 場合、リスク R_K に対する不偏推定量となり、*AIC* を用いたときに起こる過度な過小推定を回避することができる。