クラスター化法の統計的評価とその応用

中 村　永 友
Nagatomo Nakamura

博士（学術）

総合研究大学院大学
数物科学研究科
統計科学専攻

平成 6 年度
1995 年 3 月

要旨

本論文の目的は，多変量特性データを始めとする測定データを分類するために用いられてきた階層的分類法と多変量混合分布モデルのそれぞれに見られる，分類法としての主な特性を数理的に考察することと，両者の利点を取り入れた統計的データ診断に有用な新たな“分類方式”を提案することである．まず総論として，本論文で扱う分類法の諸手法についての要約を行い，統計的データ解析の見地から，これらの諸手法について，問題の背景を明らかにする．

クラスター化法（クラスター分析，自動分類法）は，近年急速に進展をみた探索的分類手法の総称である．これらの多くの手法は発見的であり，統計学的な見地からの理論的かつ客観的な補強が不十分なままに，様々な研究分野で広範に利用されてきた．このようなクラスター化法の諸手法は，階層的分類法と非階層的分類法とに大別して考えることが多い．

まず，本論文ではクラスター化法の代表的手法である「組み合わせ的階層分類法」に含まれるいくつかの手法について，クラスターの生成過程にみられる距離空間のひずみの概念を一般化し，これに基づいて手法相互の数理的な関連を明らかにする．また，これは次の多変量正規混合分布モデルに基づく分類方式の改良において，必要な準備である．

次に，別のアプローチとして，統計的分類手法の一つである多変量正規混合分布モデル（以下，正規混合分布モデル）に注目し，これに基づく多変量特性データの分類方式の提案を行う．まず，基礎となる正規混合分布モデルを分類問題に適用するうえでの問題点を整理する．そして，正規混合分布モデルのパラメータ推定およびコンポーネント数推定の計算手続きで問題とされてきた初期値設定において，前述の階層的分類法の諸性質を利用した分類方式を提案し，これによる改良点や有効性を考察する．

ところで階層的分類法にはいくつかの手法があり，分類対象となるデータセットが同一

であっても，選んだ分類手法によって得られる分類結果（クラスター化の過程）が異なるという固有の性質がある．提案する分類方式は，この性質を正規混合分布モデルの初期値設定に積極的に利用することで，分類結果におよぼす影響を客観的に評価する手がかりを与えるところに特徴がある．

さらにこのような議論を通して，正規混合分布モデルの自然な拡張としての多変量 t 混合分布モデル（以下，t 混合分布モデル）の諸性質を考察し，従来とは異なる新たな分類方式を提案する．

最後に，いくつかの事例データ解析を通して，本論文で提案した分類方式の実際データへの適用可能性や実用性を検証する．

以上のように，本論文の主題は，従来個別に議論されてきた複数の分類手法に注目して，それぞれの方法が備える数理的特性を体系的に整理・再構築し，それによって得られる新たな分類方式を提案することにある．

以上に述べたことを考慮し，本論文の構成は次のようになる．

第 I 部　総論

第 I 部は総論として，本論文で扱う階層的分類法および多変量混合分布モデルによる分類法の基本事項とそれらに見られる問題点を要約する．

第 II 部　階層的分類法の性質

第 II 部は，組み合わせ的階層分類法に含まれる諸手法について，距離空間のひずみの概念に基づき，手法相互の数理的な関連を明らかにする．まず，階層的分類法の距離空間のひずみ（保存・拡大・縮小）の一般的な成立条件を与え，それに基づいて階層的分類法の特性を明らかにする．ここで得られた知見から，いくつかの階層的分類法が “可変法” として一般化され，手法相互の相対的な関係が明らかになる．さらにこれらの性質を利用して，新しい手法（一般化可変法）を提案する．

第 III 部　多変量混合分布モデルによる分類法

ここでは，多変量混合分布モデルを分類法として用いる際の主な留意点（EM 法の初期値設定とコンポーネント数の推定）について，実用的な分類方式を提案する．

ここで提案する分類方式は，第 II 部で考察した階層的分類法および k-means 法などの，

従来利用されてきたクラスター化法を，正規混合分布モデルの初期値設定の分類法として利用する．これは，データに内在する構造の特徴に応じて，各々の手法が固有の分類結果を与える（クラスター化の過程が異なる）という性質を利用して，初期値設定のための様々なクラスター化の状況を作り出すことに相当する．そして，これらの分類結果を多変量混合分布モデルの初期値設定に適用する．提案する分類方式は，事後確率（各コンポーネント分布への所属確率），判別率等の指標を用いることにより，得られた分類結果の客観的な比較を可能とし，結果として分類対象のデータ構造のより具体的な診断の手がかりが得られるという利点がある．従来行われてきた多変量混合分布の分類法は，初期値設定が分類結果におよぼす影響の評価は困難であるとされてきたが，ここに提案する方法は，これらの弱点を補うものである．

さらに，情報量規準を用いて正規混合分布モデルのコンポーネント数の推定を行う手続きを提案する．これは提案した分類方式と，ブートストラップ法でバイアス推定を行う手続きを併せて行うところに特徴がある．また，ブートストラップ標本から混合分布モデルのパラメータ推定を行う際の初期値の設定方法について考察し，この方法の有効性を数値実験により検証する．そして，線型近似によるブートストラップ・バイアス推定の変動減少法が，コンポーネント数の推定に際してブートストラップ反復回数の減少を可能とし，あわせてEM法の収束の遅さを補う方法としても有効であることを数値実験から観察する．

さらに多変量正規分布より裾の重いデータへの対処方法として，正規混合分布モデルの自然な拡張としてのt混合分布モデルを提案する．ここで，t混合分布モデルは正規混合分布モデルを包含し，より一般的なモデルとして表現が可能になるという利点がある．

第IV部　データ解析

第IV部は提案した分類方式の有用性を検証するために，いくつかの事例データ解析を行う．

第一の例は，昆虫学の形質に基づく分類法との対比で，興味ある知見が得られた事例である．扱うデータセットはオーストラリアにおける野外調査で計測された「キバハリアリ」の計量データおよび形質データである．これを解析した研究者らは，主に形質データを利用した分岐分類を行い，所与のデータセットが9つの種群からなると結論づけた．

ここでは提案した分類方式を計量データに適用し，種群に相当するコンポーネント分布のその数の推定や，各々の個体（アリ）のコンポーネント分布への所属確率などを求める．この結果を上述の９種群と比較し，興味ある知見が得られた．とくに解析に用いる変数（特性）の選択，データの加工（比率変換など）を含めて，ここで提案した分類方式は，形質に基づく伝統的な分岐分類や系統解析等に先立つ事前処理法として利用できる．この意味で分類結果得られた群（クラスター）の客観的な情報は有用であるとの専門家の意見を得ている．

次に，LANDSAT の画像データへの適用例を取り上げる．ここで提案した分類方式のコンポーネント分布を多変量 t 分布として分類を行う．これを正規混合分布モデルとの比較検討することにより，t 混合分布モデルの有効性を示す．次に，各画素上の多変量特性データの事後確率，確率密度などの情報を用いて，分類結果を効果的に色彩画像化するための配色アルゴリズムを提案する．これら一連の手続きの特徴として，(1) 推定したコンポーネント分布の重なりあう様相が色彩画像として視覚化される，(2) コンポーネント分布やそれらの間の構造が画像上で色彩イメージとして視覚化され，分類結果の画像の解釈が容易になる，(3) 扱う画像データの画素数が比較的大きく（数十万～数百万程度），教師データ (トレーニング・データ) となる地上の詳細な情報が入手困難な場合などに有効である，などが挙げられる．

目次

記号一覧

第II部

$\boldsymbol{E} = \{1, 2, \cdots, i, \cdots, N\}$: 個体の添字の集合

$\boldsymbol{K} = \{1, 2, \cdots, j, \cdots, p\}$: 変量の添字の集合

$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{ip})$: 個体iのp変量観測値ベクトル

$\boldsymbol{X}_N = [x_{ij}]$

$= \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N\}^T$: 多変量特性データ行列$(i \in \boldsymbol{E}, j \in \boldsymbol{K})$

$\boldsymbol{R} = \{1, 2, \cdots, k, \cdots, r\}$: クラスターの添え字の集合

C_k : 第kクラスター$(k \in \boldsymbol{R})$

$P(\boldsymbol{E})$: $\boldsymbol{E}$のべき集合

$\boldsymbol{D} = [d_{ij}]$: 距離行列

$\boldsymbol{S} = [s_{ij}]$: 類似度行列

$\boldsymbol{m}$: 任意の階層的分類手法

$H(\boldsymbol{m})$: $\boldsymbol{m}$により得られる階層構造

$h(\cdot)$: 水準関数

$\boldsymbol{\Theta} = \{\alpha_i, \alpha_j, \beta, \gamma\}$: 組み合わせ的階層分類法のパラメータ集合

$d(C_i, C_j)$: クラスター間距離

$\boldsymbol{\Delta}^{(t)}$: 第tステップにおける更新距離の集合

δ, ε : クラスター間距離の差分

$(\delta = d(C_j, C_k) - d(C_i, C_k), \varepsilon = d(C_i, C_k) - d(C_i, C_j))$

$d(\boldsymbol{m})$: 手法$\boldsymbol{m}$により更新されるクラスター間距離

第III部

N : 標本数

$\boldsymbol{X}_N$: サイズNの標本

$f(\cdot|\cdot)$: 混合分布の確率密度関数

$f_k(\cdot|\cdot)$: 第kコンポーネント分布の確率密度関数

π_k : 第kコンポーネント分布の混合比率

r : $f(\cdot|\cdot)$のコンポーネント数

$\boldsymbol{\Phi}_r = \{\pi_1, \cdots, \pi_{r-1}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_r\}$: $f(\cdot|\cdot)$のパラメータ

$\boldsymbol{\theta}_k$: 第kコンポーネント分布のパラメータ
$\boldsymbol{\mu}_k$: 第kコンポーネント分布(正規分布)の平均ベクトル
$\boldsymbol{\Sigma}_k$: 第kコンポーネント分布(正規分布)の分散共分散行列
$L(\cdot|\cdot)$: $f(\cdot|\cdot)$の対数尤度関数
$Pr(k|\boldsymbol{x}_i) = \tau_{ki}$: $\boldsymbol{x}_i$の第kコンポーネント分布への事後確率
$|\boldsymbol{V}|^{-1/2}h(s_i^2)$: 楕円分布族の確率密度関数
$\boldsymbol{\mu}$: 楕円分布族の位置ベクトル(正規分布のときと同じ)
$\boldsymbol{V}$: 楕円分布族の擬分散共分散行列
ν : 多変量t分布の形状パラメータ
$\mathcal{L}(\cdot)$: 楕円分布族(の混合分布モデル)の対数尤度関数
$s_i^2 = (\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{V}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$: 擬マハラノビス距離
$w(s_i^2|\nu) = w_i$: 楕円分布族のときの各標本へのウェイト
$g_k(\cdot|\cdot)$: 第kコンポーネント分布(楕円分布族)
$\boldsymbol{\mu}_k$: 第kコンポーネント分布の位置ベクトル(楕円分布族)
$\boldsymbol{V}_k$: 第kコンポーネント分布の擬分散共分散行列(楕円分布族)
ν_k : 第kコンポーネント分布の形状パラメータ(多変量t分布)
$s_{ki}^2 = (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\boldsymbol{V}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)$: 第kコンポーネント分布に対する標本$\boldsymbol{x}_i$の擬マハラノビス距離
$w(s_{ki}^2|\nu) = w_{ki}$: 第kコンポーネント分布に対する標本$\boldsymbol{x}_i$へのウェイト
$\Gamma(\cdot)$: ガンマ関数

A_k : 第kコンポーネント分布に対する判別率
N_k : 第kコンポーネント分布から抽出される標本数
A : 全体の正しい判別率
T : Aの推定値

z_{ki} : $\boldsymbol{x}_i$の第kクラスターへの所属を表す変数
$(\boldsymbol{Z}_N = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_i, \cdots, \boldsymbol{z}_N\}),\ \boldsymbol{z}_i = \{z_{1i}, \cdots, z_{ki}, \cdots, z_{ri}\})$
B : ブートストラップ反復回数
$\boldsymbol{X}_N^{*(b)}$: b番目のブートストラップ標本
bias_r : 対数尤度のバイアス(コンポーネント数rのときの)
ICBoot : ブートストラップ法により構成された情報量規準
AIC : 赤池の情報量規準
c : クラスター化法の種類$(\ell = \{1, 2, \cdots, c\})$

第IV部

$\boldsymbol{p}^{DS}$	:	データ空間内の任意の点
$\boldsymbol{g}$	:	混合分布の重心
S_k	:	第kコンポーネント分布に付与される彩度の値
H_{k0}	:	第kコンポーネント分布に付与される色相の値
r_k	:	第kコンポーネント分布に付与される明度の変化幅
I_k	:	$\boldsymbol{p}^{DS}$に付与される明度の値
α_k	:	第kコンポーネント分布に対する$\boldsymbol{p}^{DS}$のパーセント点
$\boldsymbol{p}_k^{HSI} = (H_{k0}, S_k, I_k)$	:	HSI空間内の任意の点

第 I 部

総論

第 1 章

統計的分類法の問題の背景

1.1 はじめに

科学的な調査・研究において, 研究対象となる現象や得られた観測値の分類や類別は重要な操作である. 多くの分野の研究者にとって分類操作は, 得られたデータを要約し, 適切な処理によって有用な情報を与える基本的な手続きとしてきわめて重要な位置を占めている. また, データの客観的記録や理解を助け, 時にはそれ以上の情報を与えてくれる.

古くは図書, 元素, 植物などの分類のように, これらは何らかの対象物があり, それを分類しようとして初めて分類法が工夫・考案されてきた. 分類操作は我々が日常的によく行う行為であり, 現在もさまざまな分野において実際にその操作に直面し, 対象に固有な分類方法が提案されている. しかし, 分類の方法論がデータ解析の道具として研究されるようになったのはここ数十年のことである.

生物分類 (taxonomy) の分野で数値を用いた分類法が積極的に行われるようになったのは, 1950 年代後半とされている. 1963 年には Sokal と Sneath による "*Principles of Numerical Taxonomy*" が出版され, それはコンピュータの普及に伴い, その後の生物分類の分野に大きな影響を与えた. これをきっかけとして, 数値分類, あるいは数量分類 (numerical taxonomy) という用語が一般に広く使われるようになり, 他の分野へ応用されるようになった. 理工学の分野においては, パターン認識, 文字認識, 画像データの分類などにおいて多数の類似の分類手法が提案されている. 一方, 分類に関する方法論に "クラスター分析" と総称される手法がある. これは二十数年ほど前までは他の多変量の統計的手続きとは全く独立に発展して, ここ十数年の間にこれらの間との関係付けが進められてきた方法論である (Gordon

(1981)).

本論文は, クラスター分析の代表的手法である凝集型の階層分類法と, 確率分布モデルを考慮した混合分布モデルによる分類法を扱う. 前者の階層的分類法は, そのアルゴリズムが明解で, コンピュータプログラム化が容易であることから, 現在広く用いられている. 一方, 後者の混合分布モデルによる分類法は, 観測データに複数の確率分布を仮定する方法である. 混合分布モデルは確率分布モデルとしては歴史的には古いが, パラメータ推定等に関して大きな困難があった. しかし, パラメータ推定法としての EM 法 (Expectation Maximization Algorithm; Dempster, Laird and Rubin (1977)) の理論的接近や近年のコンピュータと数値計算技術の発展に伴い, ここ数年様々な分野で用いられつつある分類手法である.

"クラスター分析" という用語は, 本来因子分析や主成分分析という用語に対比されるものとして 1940 年代に登場した. しかし, 近年は一般的に "分類操作" に関する手法の総称として用いられている. 類語としては数値分類, 自動分類 (automatic classification), 教師なしのパターン認識 (unsupervised pattern recognition) などがある. これらの用語は発生の経緯や研究分野が異なるが, 近年は広くクラスタリングあるいはクラスター化法と総称されている.

多変量解析では分類操作に関する代表的な方法として, 判別分析とクラスター化法がある. 前者は, 群がすでに存在し, あるいは設定されており, 各群への所属が既知のデータがあるとき, その所属情報から判別関数が構成される. 判別分析は誤判別率の推定などを中心としてかなり研究されている (Anderson (1984), Seber (1984) などの著書に詳しく説明されている). さらに判別分析はパターン認識の分野で応用・研究され, 統計的パターン認識として位置づけられる (Duda and Hart (1973), Fukunaga (1972) など). これに対してクラスター化法は群の構造に関する事前情報が何もないという設定のもとに, 群 (クラスター) を生成する方法である. この意味でクラスターを生成するための方法であるので, ここでは "クラスター化法" と呼ぶ.

様々な分野でクラスター化法が研究され, ここ 20 ～ 30 年の間にこれに関する文献の量は急速に増加した. しかし異分野間の研究者の交流がほとんどなく, 各分野で独自に研究されてきたという経緯がある. そのためクラスター化法の手法は無数にあり, それらの分類や

総合報告が多くの研究者によりなされている．たとえば, Cormack (1971), Everitt (1980, 1993), Gordon (1981), Hartigan (1975), Jardine and Sibson (1971), Mezzich and Solomon (1980) などがある．Sneath and Sokal (1973) に約 1600 の参考文献が, Duran and Odell (1974) には 409 もの文献が参照されていると Seber (1984) は指摘している．このようにクラスター化法に関連する文献は非常に多い．

クラスター分析を行う場合, 分類対象として個体の分類と変量（変数, 特性）の分類が考えられるが, ここでは前者の個体の分類を考える．そのとき, 解析対象のデータセットはいくつかの個体のある特性の観測値（個体 × 変量の通常の多変量特性値のデータ行列）の場合と, 個体間の親近性を表すデータ（類似度行列など）の場合がある．ここでは主に前者のデータを扱う．

クラスター化法の分類は多くの議論のあるところであるが, ここでは通例にならって階層的分類法と非階層的分類法の二つに大別して考える．階層的分類法はクラスター化の過程のあるステップで併合または分割を行うことにより, 次のステップへ移行する方法である．始めに N 個体を N 個のクラスターと見なして, 最終的に N 個体からなる 1 つのクラスターを作る, またはあらかじめ決めておいたクラスター数まで併合を進める方法が “凝集型の階層的分類法” である．一方, N 個体を 1 つのクラスターと見なして, ここから初めて, 最終的に N 個のクラスターまで細分割を進める, またはある終了条件により分割をやめる方法が “分割型または分枝型の階層的分類法” である．

一方, 非階層的分類法は階層的分類法以外の手法をいうが, ここではとくに分割最適化型手法に議論を限定する．この手法は, あらかじめクラスター数を固定しておいて, 各群の分散共分散行列に関する目的関数を最適化する方法である．Scott and Symons (1971) は, 良く使われているいくつかの分割最適化型の手法が, 多変量正規混合分布モデルの分散共分散行列のパラメータにある制約条件を入れたものと数学的に同値であることを示している．たとえば k-means 法（MacQueen (1967) など）は各クラスターの分散共分散行列を単位行列の定数倍とする条件を入れることと数学的に同値である．実際の応用場面においては, 群内の等質性, 群間の非等質性はとくに重要で, 分割最適化型の手法はこの点において大変優れている．クラスター化法の総合的な報告は Seber (1985) が独自の視点で体系的に整理している．

通常, 複数の分類手法をデータに適用した場合, その解が一意でない（分類結果が同一でない）という性質がある. たとえば凝集型の階層的分類法においては, 個体間の類似度の選択やクラスター間距離の更新アルゴリズムの選択によって, 所与のデータセットのデータ構造の特徴に応じた固有の分類結果が得られる. また, 非階層的分類法の分割最適化型の手法においては, 等質性の基準や最適化する目的関数の設定, 個体の初期配置, 再配置アルゴリズムの選択によって分類結果が異なる. さらに, これらの分類法は記述的かつ探索的な方法であるため, 得られた分類結果の解釈が確率的・統計的に評価できないという点がある. また, 数値分類法などでは類似度や非類似度の選択が重要で, この選択と階層的分類法のアルゴリズムの関係を十分吟味することは重要である.

分布の混合, とくに正規分布の混合は観測された多変量特性データが, 混合比率の異なる正規母集団分布からの実現値と見なすアプローチにおいて, 非常に重要な確率分布モデルである. 近年, 混合分布モデルは外れ値を識別するためのモデル (Aitkin and Tunnicliffe Wilson (1980)) や, Tukey (1960) による混合正規モデル (contaminated normal model) などのようなロバスト推定量の研究にも用いられている (Huber (1964) など).

コンポーネント数が有限個の混合分布を分解すること, つまりモデルのパラメータ推定は非常に困難な問題で, これを手掛けた研究は古くは K. Pearson (1894) が最初といわれている. 彼はモーメント法を使って, 分散が等しくない場合の 2 つの一変量正規分布の 5 個のパラメータ推定の方法を示している. これは 9 次の多項式の根を求めるというもので, コンピュータのなかった当時としては, これを解くことは不可能に近いものであった. その後 Charlier and Wicksell (1924), Cohen(1967) などにより, 代数方程式をより簡単に解く方法が考えられているが, Tan and Chan (1972), Fryer and Robertson (1972) などが示しているように, 混合分布モデルのパラメータ推定において, モーメント推定法は最尤推定法に比べて様々な面で優位でないとされている. 一方, グラフィカルな方法を用いたパラメータ推定は, 1900 年代半ばまで盛んに研究されていて, これは Everitt and Hand (1981) に詳しく記述されている.

この種の問題に最尤法が使われ始めたのは, 1960 年代とされている (Redner and Walker (1984)). コンピュータの処理能力の向上と普及に伴い, さらに数値計算技術の向上により,

近年は最尤法による混合分布モデルのパラメータ推定が主流となっている．混合分布モデルに初めて最尤法を用いたのは Rao (1948) である．彼は2つの分布の分散が等しい一変量のモデルに対して，Fisher のスコア法を用いてパラメータ推定を行った．その後，Rao から Hasselblad (1966, 69) に至る研究をたどると，この種の問題の最尤推定に関する文献はほとんどみあたらず (宮川 (1987))，Hasselblad により初めて g 個の一変量正規分布の混合と，指数分布族の混合分布の問題が扱われた．多変量への拡張が試みられたのは Wolfe (1967, 69, 70) と Day (1969) による．Day は等しい分散共分散行列の仮定の下で，Wolfe は任意の分散共分散行列の仮定の下で研究している．Hasselblad (1967, 69) は EM 法の応用と考えられる反復的（iterative）な解法で最尤推定を行っている．しかし，この方法には EM 法のような定式化された理論はみられない．また広い意味で Wolfe や Day による反復的な方法も EM 法の枠組みでとらえることができる．つまり，彼らは混合分布モデルを EM 法の枠組みである「不完全データの問題」としては認識していなかったようだ．このような認識は Orchard and Woodbury (1972) によって初めてなされた．後に詳しく述べるが，EM 法のアルゴリズムとその理論は Dempster, Laird and Rubin (1977), Redner and Walker (1984) にある．

1980 年代に入り，混合分布モデルに関する興味ある著書が三冊出版されている．Everett and Hand (1981) は混合分布モデルの入門書として，"*Finite Mixture Distributions*" を著した．Titterington, Smith and Makov (1985) の "*Statistical Analysis of Finite Mixture Distributions*" は，より理論的な話題を提供している．McLachlan and Basford (1988) の "*Mixture Models: Inference and Applications to Clustering*" では，よりデータ解析指向の内容に立ち入って議論を展開し，実際のデータ解析における基本的事項と問題点が詳しく述べられている．またこの第一章に混合分布モデルの最近の研究動向がかなり詳しく報告されている．Titterington, Smith and Makov (1985) に示されているように，混合分布モデルは理論的な困難や数多くの解決すべき問題を抱えていて，現実のデータ解析において混合分布モデルを適用する場合，注意すべき多くの問題が存在する．このような現状の研究動向を十分踏まえたうえで，本論文では実際のデータ解析を行う上での実用的な方法論を提案する．

1.2 本論文の扱う問題

本論文が扱う問題とその背景は次のとおりである．まず，凝集型の階層的分類法（以下，単に階層的分類法）はその基本アルゴリズムが個体間の距離または非類似度から出発し，最も近い個体（またはクラスター）を併合して，ボトムアップ式にクラスターを生成する方法である．この際に，最も近い個体を併合することと，併合してできたクラスターとそれ以外の個体あるいはクラスターとの距離の更新にみられる性質に着目する．個々の手法に固有な特徴はこの距離の更新規則で決るが，距離の更新規則と“距離空間のひずみ”の概念は密接な関係がある．これらを関連付けた数理的な考察がこれまでほとんど行われていなかった．更新距離と距離空間のひずみの関係を体系的にまとめることによって各手法の数理的な考察が可能になり，これを利用すると，手法相互の関係を客観的に調べることができる．

次に，通常，推測統計の立場では，得られたデータセットは未知の母集団分布からの実現値であると仮定して様々な推測や予測を進める．これと同様に実際に得られた分類対象となるデータセットが，複数の未知の母集団分布からの実現値であると仮定するとき，混合分布モデルによる分類法は確率分布モデルを考慮した分類法として位置付けられる．しかし，混合分布モデルによる分類法にもいくつかの問題がある．それは，混合分布モデルのパラメータ推定の方法として最尤法を用いることに起因する推定上あるいは実用上の重要な問題である．たとえば，尤度関数には局所的な解が複数存在するという問題がある．収束の遅い EM 法を用いて最尤推定を行うときこの問題がさらに増幅されることがある．そして，コンポーネント数の推定についても多くの問題がある．

実用性を重視して，尤度関数の最適解を得るためには，McLachlan and Basford (1988) が述べるように“数多くの初期値からパラメータ推定を行う方法”があるが，この労力を少しでも減らす効率的かつ合理的な初期値の与え方について考える必要がある．Redner and Walker (1984) の論文の定理 3.1, 定理 3.2 で，かなり自然な条件のもとで，十分大きな標本数のとき，尤度方程式の解は一意に定まり，それは対数尤度の大域的最適解を与えることが理論的に示されている．しかし，実際の問題では，ここでいうところの標本数が十分大きい場合は稀である．さらに，観測データの次元数 (変量，変数の数) とコンポーネント数が多くなると，推定すべき混合分布のパラメータ数は，次元数の 2 乗のオーダーで増加し，それに伴い局

所的な解の数も増大する．しかし，観測データの構造や推定すべきパラメータに関する何らかの事前情報があるとき，それを利用して混合分布を推定すれば，大域的最適解を得る可能性はより高くなると考えられる．これは，EM 法のパラメータの初期値設定は十分考慮しなければならないことを意味している．そこで本論文は，データから混合分布を推定するとき複数のクラスター化法でデータの初期分類を行い，大域的最適解を得る可能性を高める分類方式を提案する．

一方，コンポーネント数の推定は重要な問題であり，従来から様々な方法が提案されている．この問題に対する一つのアプローチは，仮説検定の枠組みの中で尤度比検定統計量を用いることである．このとき，検定統計量の分布は通常の漸近 χ^2 近似が成り立たないことが知られていて (McLachlan (1987), Redner and Walker (1984) など)，この問題に対して様々な工夫がなされている．主な方法としては，尤度比検定統計量の修正 (Wolfe (1971)) や，ベイズによる方法 (混合比率に事前分布を与える方法, Aitkin and Rubin (1985))，ブートストラップ法による尤度比検定統計量の漸近分布の構成 (McLachlan(1987), Thode, Finch and Mendell (1988)) などがある．ここでは情報量規準をブートストラップ法で構成して，モデル選択という立場からコンポーネント数の推定を行う．これはコンポーネント数を変えて情報量規準の値を計算し，これ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方法である．この一連の手続きを，前述の初期値設定法を含めた混合分布モデルのコンポーネント数の推定方法として提案する．

以上のことに基づいて，ここで提案した分類方式の有効性を検証するため，二種類のデータセットの解析を行った．

第一の例は昆虫学（生物学）の形質情報に基づく系統的な分類法（分岐分類法）による結果との対比を行い，混合分布モデルによる分類法の適用可能性を検証した．ここでは，混合分布モデルによる分類結果を分岐分類による分類結果と比較し，興味ある知見が得られた．とくに解析に用いる変数（特性）の選択，データの加工（比率変換など）を含めて，ここで提案した分類方式は，形質に基づく伝統的な分岐分類や系統解析等に先立つ事前処理法として有用であることがわかった．

次は，LANDSAT の画像データへの適用例である．ここではまず，コンポーネント分布が

多変量 t 分布であるモデルを提案し, そのパラメータ推定の方法を示す. そして観測データ (分光反射輝度) に対してこのモデルをあてはめ, 正規混合分布モデルと多変量 t 混合分布モデルの比較を行なった, さらに, 各画素上の多変量特性データの事後確率, 確率密度などの情報を用いて, 分類結果を効果的に色彩画像化するための配色アルゴリズムを提案した. これら一連の手続きの特徴は, (1) 推定したコンポーネント分布の重なりあう様相が色彩画像として視覚化される, (2) コンポーネント分布の構造が画像上で色彩イメージとして視覚化され, 分類結果の画像の解釈が容易になる, (3) 扱う画像データの画素数が比較的大きく (数十万～数百万程度), 教師データとなる地上の詳細な情報が入手困難な場合などに有効である, などが挙げられる.

第 II 部

階層的分類法

第 2 章

階層的分類法の性質

この章は, まず, 分類対象となる多変量特性データや個体の類似や差異を表す測度について説明する. さらに, Lance and Williams (1967) によって議論された "組み合わせ的階層分類法" と呼称される凝集型階層的分類法 (Agglomerative Hierarchical Clustering Algorithm: AHC 手法) に関する基本事項について述べる.

2.1 類似度, 非類似度, メトリック

議論の対象となるデータの構造は, 個々の個体 (objects または individuals) のある特性値からなる, いわゆる多変量特性型のデータである. このとき分類対象としては個体と特性値 (変量, 変数: variables) が考えられるが, ここでは個体の分類を考える. 以下の議論で必要となる記号は次のとおりである.

$$
\begin{array}{rrcl}
\text{個体の集合} & \boldsymbol{E} & = & \{1, 2, \cdots, i, \cdots, N\} \\
\text{変量の添字の集合} & \boldsymbol{K} & = & \{1, 2, \cdots, j, \cdots, p\} \\
\text{個体}i\text{の}p\text{変量観測値ベクトル} & \boldsymbol{x}_i & = & (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{ip}), (i \in \boldsymbol{E}) \\
\text{多変量特性データ行列} & \boldsymbol{X}_N & = & [x_{ij}], (i \in \boldsymbol{E}, j \in \boldsymbol{K}) \\
 & & = & \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N\}^T \\
\text{クラスターの添え字の集合} & \boldsymbol{R} & = & \{1, 2, \cdots, k, \cdots, r\} \\
\text{第}k\text{クラスター} & C_k & & (k \in \boldsymbol{R})
\end{array}
$$

したがって N 個の個体は p 次元ユークリッド空間内の点の集合として表わされる.

ものを分類することとは "類似したものをまとめる" あるいは, "差異のあるものは切り離す" ということである. しかし, このようなあいまいな表現では与えられたデータの客観的な分類は難しい. そこで個体間の類似あるいは差異を表わす測度が必要となる. 通常は二つの個体 i と個体 j との間の親近性の度合を示す測度として, 類似度 (simirality) や非類似度

（差異度: dissimirality）を用いる．類似度は個体間の似ている度合を示し，記号 $s(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $s(C_i, C_j)$, $s(i,j)$, s_{ij} などで表わす．また非類似度は離れ具合（差異）を示し，記号 $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $d(C_i, C_j)$, $d(i,j)$, d_{ij} などで表わす．このとき個体間の類似，非類似の関係は，類似度および非類似度行列

$$\boldsymbol{S} = [s_{ij}], \quad \boldsymbol{D} = [d_{ij}]$$

として与えられる．距離は非類似度の一つであるが，これを $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ として表わす．このとき，$d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ は

$$\boldsymbol{E} \times \boldsymbol{E} \to \boldsymbol{R}$$

という写像である（$\boldsymbol{R}$ は実数の集合）．そして次の条件を満足するとき，とくに“メトリックな距離”という．

定義 **2.1** $\boldsymbol{X}_N$ 上でメトリックとは，実数値関数

$$d : \boldsymbol{E} \times \boldsymbol{E} \to \boldsymbol{R}$$

で表わされる．$\boldsymbol{E}$ 上の個体の観測ベクトル $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ のすべての 2 つの組み合わせは，次の性質を満足する．

(1) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0, {}^{\forall} i, j \in \boldsymbol{E}$,

(2) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$, ただし $\boldsymbol{x}_i = \boldsymbol{x}_j$ のときに限る, ${}^{\forall} i, j \in \boldsymbol{E}$,

(3) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = d(\boldsymbol{x}_j, \boldsymbol{x}_i), {}^{\forall} i, j \in \boldsymbol{E}$,

(4) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_k) + d(\boldsymbol{x}_k, \boldsymbol{x}_j), {}^{\forall} i, j, k \in \boldsymbol{E}$.

ここで，(1) と (2) は非負の実数値関数を，(3) は対称律，(4) は推移律を表す．(4) はいわゆる三角不等式である．(4) の条件をさらに強めたものに次の強メトリック（または超距離，*ultrametric*, *Gordon* (1987) など）不等式がある[1]．

(5) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \max\{d(\boldsymbol{x}_i, \boldsymbol{x}_k), d(\boldsymbol{x}_k, \boldsymbol{x}_j)\}, {}^{\forall} i, j, k \in \boldsymbol{E}$

上の (1) ～ (4) を満たす距離としては，ユークリッド距離などがある．さらに (5) を満たすものとして後で述べる階層構造が与えられる距離がある．

[1] これは後で述べるクラスターの結合距離が単調であるための条件となっている．

2.2 クラスターとクラスター集合

クラスター化法（とくにAHC手法）についての議論を進めるにあったって，“クラスター”を次のように定義する．

定義 **2.2** 記号 C_k はクラスターを表す．ここでクラスターとは，$\boldsymbol{E}$の空でない部分集合であり，$C_k \subseteq \boldsymbol{E}$ である．ある個体 i が $i \in C_k$ であれば，当然 $i \in \boldsymbol{E}$ である．また，$P(\boldsymbol{E})$ を $\boldsymbol{E}$ のべき集合（すべての部分集合の集合）とすると，

$$^{\forall}C_k \in P(\boldsymbol{E})$$

である．また，あるクラスター C_k に含まれる個体の総数を，そのクラスターの“クラスターサイズ”と言い，n_k で表わす．

定義 **2.3** (*DuBien and Warde* (1979)) 次の性質をもつ任意のクラスターの集合 G を“クラスター集合”と名付ける．

$$G = \{C_1, C_2, \cdots, C_k, \cdots, C_r\}$$

このとき，G は次の条件を満足する．

$$\begin{aligned}
&(1)\ ^{\forall}C_k \in G, C_k \neq \phi,\\
&(2)\ C_k, C_{k'} \in G, k \neq k' \text{ならば}, C_k \cap C_{k'} = \phi,\\
&(3)\ \bigcup_{k=1}^{r} C_k = \boldsymbol{E}.
\end{aligned}$$

クラスター集合とは分類過程の各段階で得られるクラスターの集合である．階層分類においてこの定義の(2)は，ある個体が複数のクラスターに同時に所属しないこと，すなわち互いに排反な分割を考えることを示している．

あるクラスター集合により得られる（排反な）組み合わせの数は膨大にある．いま N 個の個体を k 個に分割する組み合わせの数は，次の第2種のスターリング数[2]で与えられることが知られている (Bock (1974) など)．

$$\sum_{i=1}^{k} \frac{1}{k!}(-1)^{k-1} i^N.$$

[2]他にも様々な表記法がある．

たとえば, $k = 2$ とすると $2^{N-1} - 1$ 通りの分割数となる.

クラスター集合 G に含まれるクラスター数のことを "クラスター集合のサイズ" と名づける. ここで G が r 個のクラスターからなるとき, それをとくに G^r と表わすこととする. いま G^N について考えると, 個体 i とクラスター C_i は一致する.

2.3 凝集型階層的分類法の基本アルゴリズム

ここでは AHC 手法の基本アルゴリズムを述べる. この手法の基本アルゴリズムは次のとおりである.

[AHC 手法の基本アルゴリズム]

ステップ **1** N 個の全個体を N 個のクラスターと考え, これらの個体に 1 から N までの番号を付与する.

ステップ **2** すべての個体間の距離を計算して距離行列 $\boldsymbol{D} = [d_{ij}]$ (または類似度行列 $\boldsymbol{S} = [s_{ij}]$) を作る. そして最も近い (似ている) クラスター (個体) を p と q とする. ここで $p < q$ とし, その距離は $d(C_p, C_q)$ である $(p, q \in \boldsymbol{E})$.

ステップ **3** クラスター (個体) C_p と C_q を併合して, 新しいクラスター $C_p \cup C_q$ を作り, クラスター数を一つ減らす. このとき改めてこのクラスターを C_p とする ($p < q$ のもとに, C_q を C_p に吸収する). 併合してできたクラスター C_p とそれ以外のすべてのクラスターとの距離を更新する.

ステップ **4** ステップ 2, 3 を $N - 1$ 回 (またはあらかじめ決めておいた回数だけ) 繰り返す. その都度のステップで併合したときのクラスターの組み合わせと, その結合距離の情報をリンクリスト (結合距離のリスト) として保存する.

ステップ **5** 必要であれば, リンクリストにもとづき併合の過程をデンドロクラム (樹木図) として図示する.

ステップ2の個体間距離の選択とステップ3の距離の更新規則は，用いる手法によって異なる．この選択がAHC手法を特徴付ける．また手法によっては，メトリックでない距離も扱える（たとえばウォード法における平方ユークリッド距離の利用）．

2.4 階層構造と結合の水準関数

ところで，クラスター集合 G は階層構造をもつとすると，その階層構造は次のように定義することができる．

定義 **2.4** (*DuBien and Warde* (1979)) H を集合 $\boldsymbol{E}$ 上で得られた分割 G の階層構造とする．H は入れ子の構造で，分割によって作られる順序系列 (*orderd sequence*) である．これを記号で表わすと次のようになる．

$$H : \{G^N, G^{N-1}, \cdots, G^2, G^1\} \quad \text{ここで} \quad G^N \subseteq G^{N-1} \subseteq \cdots \subseteq G^2 \subseteq G^1.$$

多くのAHC手法は階層構造を作り，デンドログラムとして表現できる．なお，結合距離が単調非減少でないAHC手法は完全な入れ子構造にならない．

定義 **2.5** (*DuBien and Warde* (1979)) AHC手法とは $\boldsymbol{E}$ 上で，ある階層構造を生成するアルゴリズムを言い，これに含まれるいずれかの手法 $\boldsymbol{m}$ である．ここで手法 m は次の条件を満足する．

(1) G^N は初期分割である．

(2) クラスター集合 G^{k-1} $(k \leq N)$ は，G^k の中で二つの“最も近い”クラスターの併合により得られる．もし $C_i, C_j \in G^k$ で，この二つのクラスターが最も近いとすると，$C_i \cup C_j \in G^{k-1}$ である．

以上のことから，データにAHC手法を用いるクラスター化の問題は，分類対象 $\boldsymbol{X}_N$，距離行列 $\boldsymbol{D}$，階層構造 $H(\boldsymbol{m})$ から得られる $(\boldsymbol{X}_N, \boldsymbol{D}, H(\boldsymbol{m}))$ という組により記述される．

次に，クラスター結合時の距離として，つまりデンドログラムにおける結合の高さを表す一つの関数を対応させる．その関数を結合距離の“水準関数”と呼び，次のように定義する．

定義 2.6 水準関数 (*Level function*)

階層構造 H は次の条件を満足する正の実数値関数 h を持ち, これを水準関数という.

(1) $G^i \subseteq G^j$ ならば $h(G^i) \leq h(G^j), {}^\forall G^i, G^j \in H$

(2) $h(G) = 0, {}^\forall G \in \boldsymbol{E}$

この h をクラスターの水準関数とする. また, 水準関数 $h(\cdot)$ のとりうる値を結合の順序に対応させて, h_i $(i = 0, 1, 2, \cdots, N-1)$ と表記する. ここで階層構造 H が完全な入れ子の構造にあるとき h_i は次の関係がある.

$$h_0 = 0 \leq h_1 \leq h_2 \leq \cdots \leq h_i \leq \cdots \leq h_{N-1}, \tag{2.1}$$

このとき h_i は単調増加または単調非減少であるという.

AHC 手法に含まれるいくつかの手法は, "距離の逆転" と呼ばれる現象が生ずる. これはあるクラスター C_i と C_j の包含関係が $C_i \supset C_j$ であるのに, 水準関数の関係が $h(C_i) < h(C_j)$ となる現象をいう. この場合, 階層構造 H は完全な入れ子の構造にならず, 必ずしも (2.1) 式を満足するとは限らない.

2.5 Lance and Williams の組み合わせ的階層分類法の性質

Lance and Williams (1967) は凝集型の階層分類法に含まれる最近隣法, 最遠隣法, 群平均法などが, ある再帰的な更新式で記述できることを提唱した. 彼らはこれを "組み合わせ的な手法"(combinatorial strategies) と表現した. この再帰式はきわめてアルゴリズム的であり計算処理上有効である. この節では, まずこの組み合わせ的手法の考え方について述べ, 階層的分類法にみられる基本的な性質 ("距離空間のひずみ") を調べる.

AHC 手法は 2 つのクラスター C_i と C_j が結合 (併合) されたときに, 他のクラスター C_k $(i \neq j, k \neq i, j, \quad i, j, k \in \boldsymbol{E})$ との距離の更新を, どのように行なうかでその特徴が決まる. Lance and Williams はこの点に着目して, クラスター $C_i \cup C_j$ と, それ以外のクラスター C_k との距離 $d(C_i \cup C_j, C_k)$ は, クラスターサイズ n_i, n_j と, 各クラスター間の距離 $d(C_i, C_j)$, $d(C_i, C_k)$, $d(C_j, C_k)$ という 5 つの既知の値を用いる (2.2) 式の再帰的な距離の更新式を示した.

$$
\begin{aligned}
d(C_i \cup C_j, C_k) \equiv\ & \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) \\
& + \gamma |d(C_i, C_k) - d(C_j, C_k)|, \quad (i \neq j, k \neq i, j)
\end{aligned} \tag{2.2}
$$

なお, 彼らがこの更新式を提案した時点では, クラスター C_k のクラスターサイズ n_k を考慮するような手法, たとえばウォード法などは含まれていなかった. ここで, $\boldsymbol{\Theta} = \{\alpha_i, \alpha_j, \beta, \gamma\}$ は手法によって異なる値をとるパラメータの集合である. この意味で組合せ的手法を $\{\alpha_i, \alpha_j, \beta, \gamma\}$ 族と名付けると, この族に含まれる手法は表2.1のように要約される. パラメータの値は定数である場合や, クラスターサイズの比として表現される変数となる場合がある.

2.6 結合距離の単調性

次に, 結合する距離が単調増加または単調非減少（以下, "単調" と言う）, つまり, 常に

$$d(C_i \cup C_j, C_k) \geq d(C_i, C_j) \tag{2.3}$$

であるための必要十分条件は,

$$\gamma \geq -\min\{\alpha_i, \alpha_j\}, \tag{2.4}$$

$$\alpha_i + \alpha_j + \beta \geq 1, \tag{2.5}$$

$$\alpha_i + \alpha_j \geq 0, \tag{2.6}$$

であることが知られている (Lance and Williams (1966, 1967), Milligan (1979), Batagelj (1981)). ここで一般性を保つためクラスター C_i, C_j, C_k 間の距離の関係は, 次式を保つとする.

$$d(C_i, C_j) < d(C_i, C_k) < d(C_j, C_k) \tag{2.7}$$

さて, (2.2) 式を用いて距離の更新を行なった場合, 重心法やメディアン法などで "距離の逆転" という現象が起こる（結合距離の逆転, 更新距離の逆転などという）. これは次のように説明される. ある2つのクラスター C_i と C_j を結合して, それ以外のあるクラスター C_k

との距離を更新する. このとき, 結合した距離より更新した距離のほうが短くなること, つまり

$$d(C_i \cup C_j, C_k) < d(C_i, C_i) \tag{2.8}$$

という現象を "距離の逆転" という.

2.7 距離空間のひずみの概念

"距離空間のひずみ" の概念は, Lance and Williams(1967) により初めて導入された. 次に示すように, 彼らによるこの概念の記述は多分に直感的なものである.

> *The primary inter-element measures may be regarded as defining a space with know properties. When groups begin to form, it does not follow that the inter group measures define a space with the original model remains unchanged, we describe the strategy as space-conserving. However, with certain strategies the model will behave as though the space in the immediate vicinity of a group has been contracted or dilated; these are the space-distorting strategies. In a space-contracting system a group will appear, on formation, to move nearer to some or all the remaining elements; the chance that an individual element will add to a pre-existing group rather than act as the nucleus of a new group is increased, and the system is said to 'chain.' In a space-dilating system groups appear to recede on formation and growth; individual elements not yet in groups are now more likely to form nuclei of new groups.*

これを要約すると次のようになる：

- "空間の拡大 (space dilating)" とは, 距離の更新時にまだクラスター内に含まれていない個体は, 新しいクラスターの生成核を形成しやすく, クラスター同士をなるべく遠ざけるように働く.
- "空間の縮小 (space contracting)" とは, ある個体がある新しいクラスターの生成核として働くよりも, むしろ既に生成されているクラスターに加わるという傾向を示す.

Lance and Williams の研究は数値実験を通して得られた経験的な知見によるものであったが, この報告に注目した矢島, 王 (1971) は, より具体的にこの考え方を発展させている. それは「群平均法は空間を拡大も縮小もする傾向がなく, 空間を保存する（不変にする）手法と考えよう」という視点から, 群平均法を基準にして空間の拡大・縮小の定義を次のように与えた. 「各々の手法の距離の更新式と, 群平均法の更新式との差をとり, その符号が“正”ならば空間を拡大する手法と定義する. “負”ならば空間を縮小する手法と定義する. 」また, 「ウォード法は空間保存として扱う」, という考え方である. しかし, この考え方には次の点で問題がある.

- “空間の保存” とはなにかの数理的な説明や定義がされておらず, またその概念を曖昧にしたまま扱っている.
- 基準とした群平均法が空間を保存（不変に）することの数理的な説明がされていない.
- ウォード法は空間を保存するとして扱っているが, 彼らの主張に従って, 群平均法とウォード法の更新式との差をとると, 空間を“拡大”するという結果になる.

以上のことから, 次の問題が提起される.

(1) まず, 彼らのいうところの保存（あるいは空間不変）とは, 数値実験における分類過程でみられるクラスターの分類感度（分類の良さ, 分離の程度, デンドログラムの見栄えで判断）などによってなされ, 理論的な裏付けがないこと,

(2) ウォード法との比較の結果から, 群平均法は空間を保存するという基準にはなりえないこと,

などが挙げられる. これらのことより, 彼らの考え方は普遍的とはいえない.

ところで, DuBien and Warde (1979) はこれらの問題をより数学的に, かつ厳密に扱うことを試みた. 彼らは距離空間のひずみを次のように考えた. まず組み合わせ的手法の四つのパラメータ $\alpha_i, \alpha_j, \beta, \gamma$ のうち, α_i と α_j を固定して, β と γ による部分族（sub-family）, つまり β と γ により規定される手法に議論を制限した. このとき, 更新距離 $d(C_i \cup C_j, C_k)$ の

集合を次のように定義する.

$$D(\beta,\gamma) \equiv \{d(C_i \cup C_j, C_k) \text{ at } (\beta,\gamma) | d(C_i,C_j) < d(C_i,C_k) < d(C_j,C_k)\} \tag{2.9}$$

そして更新される距離の単調性と空間の保存・拡大・縮小の条件を, 次のように定義する.

定義 **2.7** 次の条件のときに限り単調増加 (*monotone increasing*) とする.

$$^{\forall} d(C_i \cup C_j, C_k) \in D(\beta,\gamma), d(C_i \cup C_j, C_k) > d(C_i, C_j)$$

定義 **2.8** 次の条件のときに限り空間保存とする.

$$^{\forall} d(C_i \cup C_j, C_k) \in D(\beta,\gamma), d(C_i, C_j) < d(C_i \cup C_j, C_k) < d(C_j, C_k)$$

定義 **2.9** 次の条件のときに限り空間拡大とする.

$$l.u.b.(D(\beta,\gamma)) \geq d(C_j, C_k), (\text{ここで}, l.u.b. : \text{下限})$$

定義 **2.10** 次の条件のときに限り空間縮小とする.

$$g.l.b.(D(\beta,\gamma)) \leq d(C_i, C_k), (\text{ここで}, g.l.b. : \text{上限})$$

しかしここでも, 組み合わせ的手法におけるパラメータを固定しているので, 一般的な定義ではない.

以上でみたように, 階層的分類法には "距離空間のひずみ" と呼ばれる現象が, クラスター間距離を更新するときに生ずる. さらに, AHC 手法には多数の手法があり, 選んだ手法によって得られる分類結果 (クラスター化の過程) が異なるという固有の性質がある. AHC 手法を厳密に評価するためには, 距離空間のひずみの成立条件を数理的な観点から, より一般化する必要がある. 次の第 3章から第 5章において組み合わせ的階層分類法に含まれる諸手法の, 距離空間のひずみの成立条件に基づく手法相互の数理的な関連について詳しく述べる.

表 2.1: 組み合わせ的階層分類法のパラメータ

手法	略記	α_i	α_j	β	γ	$\alpha_i+\alpha_j+\beta$	単調性
Single linkage 最近隣法	SL	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	1	Yes
Complete linkage 最遠隣法	CL	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	1	Yes
Group average 群平均法	GA	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0	1	Yes
Weighted average 加重平均法	WA	$\frac{1}{2}$	$\frac{1}{2}$	0	0	1	Yes
Ward's method ウォード法	WD	$\frac{n_i+n_k}{n_t}$	$\frac{n_j+n_k}{n_t}$	$-\frac{n_k}{n_t}$	0	1	Yes
Centroid method 重心法	CD	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$-\alpha_i\alpha_j$	0	$1+\beta<1$	No
Median method メディアン法	MD	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	$\frac{3}{4}<1$	No
Flexible method 可変法	FX	$\frac{1}{2}(1-\beta)$	$\frac{1}{2}(1-\beta)$	$-\frac{1}{4}$	0	1	Yes

n_i はクラスター C_i のクラスターサイズ, $n_t = n_i + n_j + n_k$

第 3 章

組み合わせ的階層分類法の距離空間のひずみ

“距離空間のひずみ” をより詳しく調べることは, 組み合わせ的階層分類法を特徴付けるパラメータの挙動を調べることに同等である. これらのパラメータの数理的な考察を深めることにより, 組み合わせ的階層分類法における距離空間のひずみの概念を一般化し, 手法の厳密な評価と手法相互の関係が明らかになる. これは, 第 II 部で述べる多変量混合分布モデルに基づく分類方式の提案において必要となる定式化である.

3.1 距離空間のひずみの成立条件

DuBien and Warde (1979) によって与えられた, 距離空間のひずみのより数学的な定義を前章で示した. このとき, 特定のパラメータに制約を与えるなど, 一般的なものではないことはすでに指摘した. これをより厳密にするために, 距離空間のひずみ成立条件を次のように与える.

いま, 第 t ステップの結合によって得られる更新距離 $d(C_i \cup C_j, C_k)$ を要素とする集合を, 次のように表す.

$$\boldsymbol{\Delta}_{\Theta}^{(t)} = \{d(C_i \cup C_j, C_k) \quad at \quad \boldsymbol{\Theta} | d(C_i, C_j) < d(C_i, C_k) < d(C_j, C_k)\} \tag{3.1}$$

ここで $\boldsymbol{\Theta} = \{\alpha_i, \alpha_j, \beta, \gamma\}$ は組み合せ的手法を規定するパラメータ集合である. また,

$$\boldsymbol{\Delta}_{\Theta}^{(0)} = \{d(C_i, C_j)\}$$

とする. ここで次のような条件を設ける.

距離空間のひずみの成立条件

(C1) 距離空間の保存:

$$d(C_i, C_k) < d(C_i \cup C_j, C_k) < d(C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_\Theta^{(t)}$$

(C2) 距離空間の拡大:

$$d(C_j, C_k) \leq d(C_i \cup C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_\Theta^{(t)}$$

(C3) 距離空間の縮小かつ単調:

$$d(C_i, C_j) \leq d(C_i \cup C_j, C_k) \leq d(C_i, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_\Theta^{(t)}$$

(C4) 更新距離の単調性:

$$d(C_i, C_j) \leq d(C_i \cup C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_\Theta^{(t)}$$

(C5) 距離空間の縮小かつ非単調:

$$d(C_i \cup C_j, C_k) < d(C_i, C_j); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_\Theta^{(t)}$$

ここで, $m = 1, 2, \cdots, N-1$, そして N は総個体数である. "距離空間" とは, $\{\alpha_i, \alpha_j, \beta, \gamma\}$ 族の手法によって更新される距離の集合である.

3.2 距離空間のひずみの成立条件の一般化

距離の更新に関係するクラスター間距離に同位 (tie) があるとき, 条件を場合分けして考えなければならず, また議論が複雑になるため, 前節でこのことは考慮しなかった. ここでは, 三つのクラスター C_i, C_j, C_k の間の距離に, 等号の条件が入った場合の距離空間のひずみの成立条件を与える. つまり,

$$d(C_i, C_j) \leq d(C_i, C_k) \leq d(C_j, C_k) \tag{3.2}$$

という関係を仮定すると, (3.1) 式は次のように書き換えられる.

$$\boldsymbol{\Delta}_\Theta'^{(t)} = \{d(C_i \cup C_j, C_k) \quad at \quad \boldsymbol{\Theta} | d(C_i, C_j) \leq d(C_i, C_k) \leq d(C_j, C_k)\} \tag{3.3}$$

等号の入った条件式として次の三つの場合が考えられる.

(i) $d(C_i, C_j) = d(C_i, C_k)$,

(ii) $d(C_i, C_k) = d(C_j, C_k)$,

(iii) $d(C_i, C_j) = d(C_i, C_k) = d(C_j, C_k)$.

これらの場合について距離空間のひずみの成立条件 (C1)–(C3) を, 以下のようにさらに細分して考える. ここで, (iii) 以外は単調性の条件を満足し, また, 条件 (C4), (C5) は共通であるので, ここでは省略する.

(i) $d(C_i, C_j) = d(C_i, C_k)$ のとき

(i-C1) 距離空間の保存:

$$d(C_i, C_j) = d(C_i, C_k) < d(C_i \cup C_j, C_k) < d(C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

(i-C2) 距離空間の拡大:

$$d(C_j, C_k) \leq d(C_i \cup C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

(i-C3) 距離空間の縮小:

$$d(C_i \cup C_j, C_k) = d(C_i, C_j) = d(C_i, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

(ii) $d(C_i, C_k) = d(C_j, C_k)$ のとき

(ii-C1) 距離空間の保存:

$$d(C_i \cup C_j, C_k) = d(C_i, C_k) = d(C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

(ii-C2) 距離空間の拡大:

$$d(C_i \cup C_j, C_k) > d(C_i, C_k) = d(C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

(ii-**C3**) 距離空間の縮小:

$$d(C_i, C_j) \leq d(C_i \cup C_j, C_k) < d(C_i, C_k) = d(C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

(iii) $d(C_i, C_j) = d(C_i, C_k) = d(C_j, C_k)$ のとき

(iii-**C1**) 距離空間の保存:

$$d(C_i \cup C_j, C_k) = d(C_i, C_j) = d(C_i, C_k) = d(C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

(iii-**C2**) 距離空間の拡大:

$$d(C_i \cup C_j, C_k) > d(C_i, C_j) = d(C_i, C_k) = d(C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

(iii-**C3**) 距離空間の縮小:

$$d(C_i \cup C_j, C_k) < d(C_i, C_j) = d(C_i, C_k) = d(C_j, C_k); d(C_i \cup C_j, C_k) \in \boldsymbol{\Delta}_{\Theta}^{\prime(t)}$$

これらの条件を図示すると図 3.1となる. このように等式制約下での距離空間のひずみの成立条件ができたので, AHC 手法のより厳密な評価が可能となる.

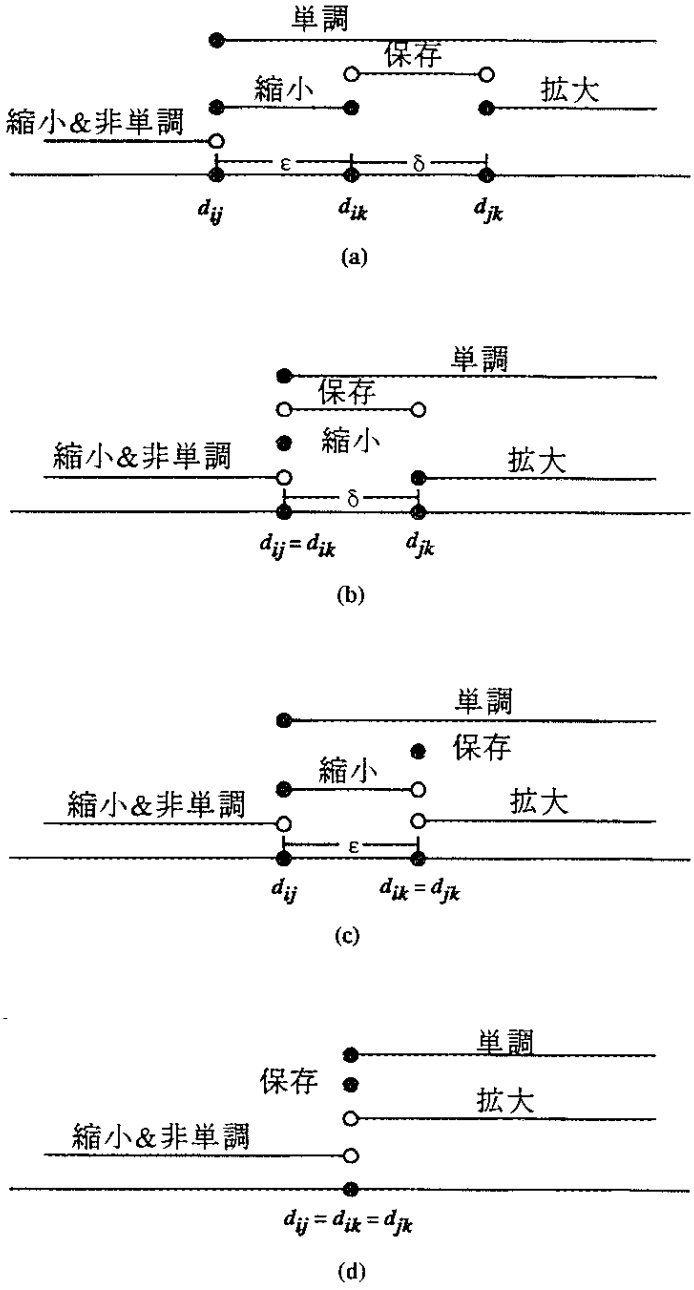

図 3.1: 距離空間のひずみの条件
○はその点を含まない, ●はその点を含む, $d_{ij} = d(C_i, C_j)$. (a) $d(C_i, C_j) < d(C_i, C_k) < d(C_j, C_k)$ の下での成立条件 (C1)-(C5), (b) $d(C_i, C_j) \leq d(C_i, C_k) \leq d(C_j, C_k)$, $d(C_i, C_j) = d(C_i, C_k)$ の下での成立条件 (i-C1)-(i-C3), (C4), (C5), (c) $d(C_i, C_j) \leq d(C_i, C_k) \leq d(C_j, C_k)$, $d(C_i, C_k \leq d(C_j, C_k)$, の下での成立条件 ($ii$-C1)-($ii$-C3), (C4), (C5), (d) $d(C_i, C_j) \leq d(C_i, C_k) \leq d(C_j, C_k)$, $d(C_i, C_j) = d(C_i, C_k) = d(C_j, C_k)$ の下での成立条件 (iii-C1)-(iii-C3), (C4), (C5).

3.3 階層分類法の距離空間のひずみの性質

ここでは, 組み合わせ的手法に含まれる代表的な手法の距離空間のひずみがどうなっているのか, 成立条件に照らし合わせて検討する. なお, 議論の途中で行なう式変形は, クラスター間距離 $d(C_i, C_j)$, $d(C_i, C_k)$, $d(C_j, C_k)$ 等がメトリックでない距離においても成り立つ.

準備としてこれらの 3 つのクラスター間距離の差を次のように定義する.

$$\delta = d(C_j, C_k) - d(C_i, C_k) > 0, \tag{3.4}$$

$$\varepsilon = d(C_i, C_k) - d(C_i, C_j) > 0. \tag{3.5}$$

また, 記号 $D(\boldsymbol{m})$ は更新距離を示し, $\boldsymbol{m}$ は組み合わせ的階層分類法の手法の略記を意味する (表 2.1参照).

最近隣法 (SL: single linkage method or nearest-neighbour method)

この手法は最短距離法, 単連結法 (single linkage) などとも呼称されている. また, 類似あるいは同義の方法として minimum method などがある. グラフ理論で知られている最小張り木問題 (MST:Minimum Spaning Tree) もこの手法と密接な関連がある.

この手法では, 二つのクラスター間距離としてクラスター C_i 内の個体とクラスター C_j 内の個体を結ぶすべての距離のうちの, 最小値 (最短距離) を用いる. (2.2) 式において各パラメータを次のようにおくと最近隣法が得られる.

$$\alpha_i = \alpha_j = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}.$$

この手法はクラスターが併合されると他のクラスター (または個体) が併合したクラスターにより近づくという性質がある. ただし, 結合に関係しないクラスター間の距離は変わらない. こうして, クラスターの併合により距離空間の縮小がおこり, 鎖状現象 (chaining) が現われる. この手法の距離空間のひずみを上で述べた条件にしたがって検証すると次のようになる ((3.2) 式を考慮する).

$$D(\mathrm{SL}) \equiv \min\{d(C_i, C_k), d(C_j, C_k)\} = d(C_i, C_k). \tag{3.6}$$

したがって条件 (C3) から, 最近隣法は空間を縮小させる手法である. $\alpha_i + \alpha_j + \beta = 1$ であるので単調性の条件が保たれる.

最遠隣法 (CL: complete linkage method or farthest-neighbour method)

この手法の別名として complete linkage, maximum method, 最長距離法などがある. また, "maximum complete subgraphs の生成" と関連する手法として知られている.

この手法の 2 つのクラスター間の距離は, クラスター C_i 内の個体とクラスター C_j 内の個体を結ぶすべての距離のうちの, 最大値 (最長距離) を用いる. (2.2) 式においてこの手法のパラメータは, 次のとおりである.

$$\alpha_i = \alpha_j = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}.$$

この手法は, クラスターを併合すると, それ以外のクラスターが併合した部分から離れるという現象がおこる. すなわち距離空間を拡大させる手法である.

最近隣法と同様に, 条件にしたがってこの手法の距離空間のひずみを検証すると次のようになる.

$$D(\mathrm{CL}) \equiv \max\{d(C_i, C_k), d(C_j, C_k)\} = d(C_j, C_k). \tag{3.7}$$

したがって条件 (C2) から, 最遠隣法は空間を拡大させる手法であることがわかる. また, $\alpha_i + \alpha_j + \beta = 1$ であるので単調性の条件が保たれる.

群平均法 (GA: group average method or unweighted group average method)

(2.2) 式において, $\alpha_i = n_i/(n_i + n_j)$, $\alpha_j = n_j/(n_i + n_j)$, $\beta = \gamma = 0$ とおくと群平均法になる. ここで, $d(C_i, C_k) = d(C_j, C_k) - \delta$ とおいて式変形すると次のようになる.

$$\begin{aligned} D(\mathrm{GA}) &\equiv \frac{n_i}{n_i + n_j} d(C_i, C_k) + \frac{n_j}{n_i + n_j} d(C_j, C_k) \\ &= -\frac{n_i}{n_i + n_j}\delta + d(C_j, C_k) < d(C_j, C_k). \end{aligned} \tag{3.8}$$

同様に $d(C_j, C_k) = d(C_i, C_k) + \delta$ とおいて式変形すると次のようになる.

$$D(\mathrm{GA}) = \frac{n_j}{n_i + n_j}\delta + d(C_i, C_k) > d(C_i, C_k). \tag{3.9}$$

これらの式から, $d(C_i, C_k) < D(\mathrm{GA}) < d(C_j, C_k)$ となり, これは条件 (C1) を満足するので, 群平均法は常に距離空間を保存する手法である.

加重平均法 (WA: weighted average method or weighted pair group average)

加重平均法は, パラメータを常に次式のように一定な値として固定する手法である.

$$\alpha_i = \alpha_j = \frac{1}{2}, \beta = \gamma = 0.$$

この手法は群平均法で $n_i = n_j$ と置いた場合, または先に述べる可変法において $\beta = 0$ と置いた場合に相当する. したがって, 条件 (C1) より加重平均法は空間を常に保存する手法である.

重心法 (CD: centroid method)

(2.2) 式でパラメータを $\alpha_i = n_i/(n_i + n_j)$, $\alpha_j = n_j/(n_i + n_j)$, $\beta = -\alpha_i \cdot \alpha_j$, $\gamma = 0$ と与えると重心法になる. この手法は次のように書き直すことができる.

$$\begin{aligned} D(\mathrm{CD}) &\equiv \frac{n_i}{n_i + n_j} d(C_i, C_k) + \frac{n_j}{n_i + n_j} d(C_j, C_k) - \frac{n_i n_j}{(n_i + n_j)^2} d(C_i, C_j) \\ &= d(C_i, C_k) + \frac{n_j}{n_i + n_j} \{ d(C_j, C_k) - d(C_i, C_k) - \frac{n_i}{n_i + n_j} d(C_i, C_j) \} \\ &= d(C_i, C_k) + \tau, \end{aligned} \tag{3.10}$$

ここで, $\tau = \frac{n_j}{n_i+n_j}\{d(C_j, C_k) - d(C_i, C_k) - \frac{n_i}{n_i+n_j} d(C_i, C_j)\}$ である. $\tau \leq 0$ のときこの手法は縮小を示し, $\tau > 0$ のとき保存を示す. つまり, 距離空間の縮小がみられるのは $\delta \leq d(C_i, C_j) \cdot n_i/(n_i + n_j)$ ($\tau \leq 0$ より) を満足するときである. さらに次の式変形をする.

$$\begin{aligned} D(\mathrm{CD}) &= \frac{n_i}{n_i + n_j} (d(C_j, C_k) - \delta) + \frac{n_j}{n_i + n_j} d(C_j, C_k) - \frac{n_i n_j}{(n_i + n_j)^2} d(C_i, C_j) \\ &= d(C_j, C_k) - \frac{n_i}{n_i + n_j} \delta - \frac{n_i n_j}{(n_i + n_j)^2} d(C_i, C_j) \\ &< d(C_j, C_k) \end{aligned}$$

このことから, 重心法は決して拡大しないことがわかる. ここでさらに, 次のような非単調性の条件を導くことができる.

$$\varepsilon + \frac{n_j}{n_i + n_j} \delta < \frac{n_i n_j}{(n_i + n_j)^2} d(C_i, C_j).$$

メディアン法 (MD: median method)

前述の重心法の特徴として次のことがある．もしクラスター C_i と C_j のクラスターサイズ n_i と n_j が $n_i \ll n_j$ という関係があり，これらのクラスターを併合したとき，新たなクラスターの重心はクラスターサイズの大きいクラスター C_j に近ずくことになる．また，クラスターサイズの小さいクラスター C_i の情報は，ほとんど失われてしまうことになる．これらのことを考慮してクラスターサイズに依存しないパラメータを与えるのが，メディアン法である．つまり重心を常にクラスター C_i と C_j の中間にあるように設定する．重心法で $n_i = n_j$ とおくとメディアン法となる．したがって，パラメータは $\alpha_i = \alpha_j = \frac{1}{2}, \beta = -\alpha_i \cdot \alpha_j = -1/4$, $\gamma = 0$ である．このことから，距離空間のひずみの条件は，重心法で得られた条件において $n_i = n_j$ とすればよい．

また，重心法とこのメディアン法は単調性の条件 (2.5) 式を満足しない手法である．つまりこの 2 つの手法は，更新距離の逆転現象がみられる手法であることを示している．

可変法 (FX: flexible method)

Lance and Williams(1967) は単調性の条件 (2.5) 式から，$\alpha_i + \alpha_j + \beta = 1$ と，$\alpha_i = \alpha_j$ の 2 つの条件を考慮して可変法を提案した．通常，可変法のパラメータは次のように与える．

$$\alpha_i = \alpha_j = \frac{(1-\beta)}{2}, \ \beta < 1, \ \text{and} \ \ \gamma = 0.$$

可変法はパラメータ β の値をいろいろと変えることで，距離空間のひずみを制御することができる．彼らはこの手法について数値実験を行い，デンドログラムの観察から次のような考察を行った．

> *The system, in fact, becomes increasingly space-contracting, and, apart only from initial ambiguities, can be made to chain completely by taking β sufficiently close unity. As β falls to zero and then becomes negative, the system ceases to contract and becomes increasingly space-dilating.*

一連の議論の中で，距離空間のひずみとクラスターの分類感度 (デンドログラムで観察される見かけの分類の程度) の概念を取り入れて，クラスター化の過程や鎖状現象の評価を試

みている[1]. しかし, いずれも数値実験による直感的な考察であり理論的な裏付けはない. また, $\beta = 0$ で距離空間を保存すると指摘している (Lance and Williams (1977)). しかしながら, これも数値実験を通して得た結果である.

さて, β の値をいろいろと変えるということは, $(\alpha_i, \alpha_j, \beta)$ で作られるパラメータ空間の中の直線 $\alpha_i + \alpha_j + \beta = 1$ 上を $0 < \alpha_i, \alpha_j < 1$, and $\beta < 1$ の範囲で移動することに他ならない（5章で述べる）.

ここで, これまでに使ってきた δ と ε で, 次のような式変形を行う.

$$\begin{aligned} D(\mathrm{FX}) &\equiv \frac{1}{2}(1-\beta)\{d(C_i, C_k) + d(C_j, C_k)\} + \beta d(C_i, C_j) \\ &= d(C_i, C_k) + \frac{1}{2}\{-(1+\beta)d(C_i, C_k) + (1-\beta)d(C_j, C_k) + 2\beta d(C_i, C_j)\} \\ &= d(C_i, C_k) + \frac{1}{2}\{-\beta(\delta + 2\varepsilon) + \delta\} \\ &= d(C_i, C_k) + \frac{1}{2}\zeta_1 \end{aligned}$$

ここで, $\zeta_1 = -\beta(\delta + 2\varepsilon) + \delta$ である. また同様な式変形をすると,

$$\begin{aligned} D(\mathrm{FX}) &= d(C_j, C_k) + \frac{1}{2}\{(1-\beta)d(C_i, C_k) - (1+\beta)d(C_j, C_k) + 2\beta d(C_i, C_j)\} \\ &= d(C_j, C_k) + \frac{1}{2}\{-\beta(\delta + 2\varepsilon) - \delta\} \\ &= d(C_j, C_k) + \frac{1}{2}\zeta_2 \end{aligned}$$

となる. ここで, $\zeta_2 = -\beta(\delta + 2\varepsilon) - \delta$ である. このように, $D(\mathrm{FX})$ は ζ_1 と ζ_2 を用いて書きかえることができたので, この式を利用して距離空間のひずみの評価を行うと, 次のようになる.

まず, "縮小" となるのは $\zeta_1 \leq 0$ のときで, したがって, β に関する条件

$$0 < \frac{\delta}{\delta + 2\varepsilon} \leq \beta$$

を得る. 次に, "拡大" となるのは $\zeta_2 \geq 0$ のときで, したがって,

$$\beta \leq -\frac{\delta}{\delta + 2\varepsilon} < 0$$

[1]実験の結果 Lance and Willimas (1967) は, $\beta = -1/4$ がよいと提言している

である. これらの条件を共に満足しない場合は "保存" となり, その条件は次のようになる.

$$-\frac{\delta}{\delta+2\varepsilon} < \beta < \frac{\delta}{\delta+2\varepsilon}.$$

ここで示したことは, Lance and Williams の実験に基づく直感的な提言を具体的に証明したことに相当する. さらに, $D(\mathrm{FX}) < d(C_i, C_j)$ により, 次の非単調性の条件を導くことができる.

$$\beta > 1.$$

ウォード法 (**WD: Ward's method**)

Ward (1963) が提案したクラスター間距離の規準が, (2.2) 式により再帰的に表現されることを, Wishart (1969b) が示した. このときウォード法のパラメータは, $n_t = n_i + n_j + n_k$ とおくと,

$$\alpha_i = \frac{n_i+n_k}{n_t}, \alpha_j = \frac{n_j+n_k}{n_t}, \ \beta = -\frac{n_k}{n_t}, \ \gamma = 0$$

である. Lance and Williams (1977) は, この手法は拡大する手法と判断した. 一方, 矢島, 王 (1971) は保存する手法として扱っている.

そこで, これまでと同様に式変形を行うと, 次のようになる.

$$\begin{aligned} D(\mathrm{WD}) &\equiv \frac{n_i+n_k}{n_t}d(C_i,C_k) + \frac{n_j+n_k}{n_t}d(C_j,C_k) - \frac{n_k}{n_t}d(C_i,C_j) \\ &= d(C_i,C_k) + \frac{1}{n_t}[n_j\{d(C_j,C_k)-d(C_i,C_k)\} + n_k\{d(C_j,C_k)-d(C_i,C_j)\}] \\ &= d(C_i,C_k) + \frac{1}{n_t}\{n_j\delta + n_k(\delta+\varepsilon)\} \\ &> d(C_i,C_k). \end{aligned}$$

ここで, $n_j\delta + n_k(\delta+\varepsilon) > 0$ である. この式は, 縮小とはならず, 少なくとも保存となる手法であることを示している. 一方, 次のようにも変形できる.

$$\begin{aligned} D(\mathrm{WD}) &= d(C_j,C_k) \\ &\quad + \frac{1}{n_t}[-n_i\{d(C_j,C_k)-d(C_i,C_k)\} + n_k\{d(C_i,C_k)-d(C_i,C_j)\}] \\ &= d(C_j,C_k) + \frac{1}{n_t}(-n_i\delta + n_k\varepsilon) \\ &= d(C_j,C_k) + \frac{\lambda}{n_t}, \end{aligned}$$

ここで, $\lambda = -n_i\delta + n_k\varepsilon$ である. これは $\lambda \geq 0$ のとき拡大し, $\lambda < 0$ のとき保存することを示している. したがって, 次の結論を得る.

$$d(C_i, C_k) < D(\mathrm{WD}) < d(C_j, C_k) + \frac{\lambda}{n_t}. \tag{3.11}$$

この結果からウォード法は常に距離空間を保存するわけではなく, クラスター間距離 $d(C_i, C_j)$, $d(C_i, C_k)$, $d(C_j, C_k)$ と, クラスターサイズ n_i, n_j, n_k の変動により拡大する場合もあることが示される. したがって, 矢島, 王の主張も, Lance and Williams の主張も共に部分的に正しいことがわかる.

表 3.1は, 各手法を距離空間のひずみの成立条件 (C1)-(C3) に従って検証した結果で, 表 3.2は, 成立条件 (i)-(iii) による結果である. また, 表 3.3は従来の研究で得られた距離空間のひずみの検証結果の一覧である.

表 3.1: 距離空間のひずみの条件

手法	縮小 & 非単調	縮小 ‡	保存	拡大
最近隣法	×	○	×	×
最遠隣法	×	×	×	○
群平均法	×	×	○	×
加重平均法	×	×	○	×
ウォード法	×	×	○ $(\frac{\varepsilon}{\delta} < \frac{n_i}{n_k})$	○ $(\frac{\varepsilon}{\delta} \geq \frac{n_i}{n_k})$
重心法	○ $(\varepsilon + \frac{n_j}{n_i+n_j}\delta < \frac{n_i n_j}{(n_i+n_j)^2} d_{ij})$	○ $(\delta \leq \frac{n_i}{n_i+n_j} d_{ij})$	○ $(\delta > \frac{n_i}{n_i+n_j} d_{ij})$	×
メディアン法	○ $(\varepsilon + \frac{1}{2}\delta < \frac{1}{4} d_{ij})$	○ $(2\delta \leq d_{ij})$	○ $(2\delta > d_{ij})$	×
可変法	○ $(\beta > 1)$	○ $(\frac{\delta}{\delta+2\varepsilon} \leq \beta)$	○ $(\frac{-\delta}{\delta+2\varepsilon} < \beta < \frac{\delta}{\delta+2\varepsilon})$	○ $(\frac{-\delta}{\delta+2\varepsilon} \geq \beta)$
可変法 †	------	○ $(\beta > 0)$	○ $(\beta = 0)$	○ $(\beta < 0)$

$d_{ij} = d(C_i, C_j)$. $d(C_i, C_j) < d(C_i, C_k) < d(C_j, C_k)$, $\delta = d(C_j, C_k) - d(C_i, C_k), \varepsilon = d(C_i, C_k) - d(C_i, C_j)$. † Lance and Williams (1977). ‡ 縮小は “単調 & 縮小” と “非単調 & 縮小” を合わせたものである. つまり非単調 & 縮小は縮小である.

表 3.2: 距離空間のひずみの条件 (2)

	条件		
手法	(i)	(ii)	(iii)
最近隣法	縮小	保存	保存
最遠隣法	拡大	保存	保存
群平均法	保存	保存	保存
加重平均法	保存	保存	保存
ウォード法	保存	拡大	保存
重心法	縮小 $(\delta \leq \frac{n_j}{n_i} d_{ij})$	縮小 $(\varepsilon \leq \frac{n_i n_j}{(n_i+n_j)^2} d_{ij})$	縮小
	保存 $(\delta > \frac{n_j}{n_i} d_{ij})$	非単調 $(\varepsilon < \frac{n_i n_j}{(n_i+n_j)^2} d_{ij})$	
メディアン法	縮小 $(\delta \leq d_{ij})$	縮小 $(\varepsilon \leq \frac{1}{4} d_{ij})$	縮小
	保存 $(\delta > d_{ij})$	非単調 $(\varepsilon < \frac{1}{4} d_{ij})$	
可変法	縮小 $(1 = \beta)$	縮小 $(0 < \beta \leq 1)$	保存
	保存 $(-1 < \beta < 1)$	保存 $(\beta = 0)$	
	拡大 $(\beta \leq -1)$	拡大 $(\beta < 0)$	
	非単調 $(1 < \beta)$	非単調 $(1 < \beta)$	

$d(C_i, C_j) \leq d(C_i, C_k) \leq d(C_j, C_k)$ の条件を満足する．詳細な条件が記載されていない場合は常にその条件を満足する．“非単調” は “空間縮小および非単調” を表す．(i) $d(C_i, C_j) = d(C_i, C_k)$, (ii) $d(C_i, C_k) = d(C_j, C_k)$, (iii) $d(C_i, C_j) = d(C_i, C_k) = d(C_j, C_k)$.

表 3.3: 組合せ的手法の距離空間のひずみの評価

手法	Lance and Williams (1967)	Lance and Williams (1977)	矢島 and 王 (1971)	DuBien and Warde (1979)	Nakamura and Ohsumi (1990)
最近隣法	縮小	縮小	縮小	縮小	縮小
最遠隣法	拡大	拡大	拡大	拡大	拡大
群平均法	保存	保存	保存	—	保存
加重平均法	—	—	—	保存	保存
ウォード法 *1	—	拡大	保存と予想	—	保存と拡大
重心法	保存	保存	縮小 *2	—	保存と縮小
メディアン法	保存	—	縮小 *2	—	保存と縮小
可変法	拡大 ($\beta < 0$) 縮小 ($\beta > 0$)	拡大 ($\beta < 0$) 保存 ($\beta = 0$) 縮小 ($\beta > 0$)	—	拡大 ($\beta = -\frac{1}{4}$)	拡大 *3 保存 *3 縮小 *3

*1 ウォード法が組合せ的手法として表現できることを Wishart(1969) が示した. *2 群平均法を基準にした相対的な比較である. *3 β とクラスター間距離の関係について厳密な条件を求めた (表 3.1参照).

第 4 章

一般化可変法とその性質

ここでは, 従来は別の手法として扱わねばならなかった最近隣法と最遠隣法が可変法の変形として扱うことができることを示す. このための準備として, 分類過程におけるクラスター間距離に, ある順序関係を設定し, 前章で示した距離空間のひずみの成立条件を利用する. また, パラメータ γ は, 最近隣法と最遠隣法でしか用いられていないため, パラメータ γ が不必要であることを述べ, 組み合わせ的手法の公式 (2.2) 式が簡略化できることを示す. これらの準備の下に可変法を一般化し, 新しい手法 (一般化可変法) を提案する.

4.1 組み合わせ的手法の簡略化

すでに示したように, 可変法のパラメータを (2.2) 式に代入して, 前章で用いた δ と ε を使って $D(\mathrm{FX})$ 書き直すと次式のようになる.

$$D(\mathrm{FX}) = d(C_i, C_k) + \frac{1}{2}\{-\beta(\delta + 2\varepsilon) + \delta\}. \tag{4.1}$$

いま, 条件 $d(C_i, C_j) < d(C_i, C_k) < d(C_j, C_k)$ があるので, 右辺の第 2 項を常に 0 とおくと (つまり, $\beta = \delta/(\delta+2\varepsilon)$) , 更新距離は常にクラスター $C_i \cup C_j$ とクラスター C_k との距離の最短距離 $d(C_i, C_k)$ を選ぶことになる. これは最近隣法のアルゴリズムそのものである. これは, 可変法のパラメータ β の値を $\delta/(\delta + 2\varepsilon)$ とすれば最近隣法が可変法の変形として記述できることを示している. これと同様にパラメータ β の値を $-\delta/(\delta + 2\varepsilon)$ とすれば最遠隣法に相当する. 以上のことから, β の値を更新に関係のある 3 つのクラスター間距離の差分の調整値として与えると, 最近隣法と最遠隣法は可変法として扱うことができる. 距離の更新式 (2.2) 式の γ の項は, 最近隣法と最遠隣法を記述するためのものであったが, ここで示した

ように γ の項が不要となり, 結果として組み合わせ的手法の簡略化ができることを示している. これは, 4 つのパラメータを考えずに, 3 つのパラメータ $\{\alpha_i, \alpha_j, \beta\}$ で組み合わせ的手法が議論できることに相当する. したがって, Lance and Williams の距離の更新式は, 次の式ですべての組み合わせ的階層分類法が記述できる:

$$d(C_i \cup C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j), \ (i \neq j, k \neq i, j). \tag{4.2}$$

表 4.1に 3 つのパラメータによる組み合わせ的階層分類法の手法を示す. さらに, パラメータ γ を取り除くことで, 次のような利点がある.

- 8 種類の階層的分類法を 3 つのパラメータ $\{\alpha_i, \alpha_j, \beta\}$ で 3 次元空間 (次章で述べるパラメータ空間) に視覚的に表示することができる.
- パラメータ空間の中にそれぞれの手法を位置づけることによって, 距離空間のひずみを的確に表すことができる.
- 結合距離の単調性の条件 ((2.4) 式〜 (2.6) 式) から γ に関する条件を削除することが可能になる.

4.2 可変法の一般化と新しい手法の提案

前節で要約したことを用いてパラメータを次のように与えると, "一般化可変法" (generarized flexible method; GF) が得られる.

$$\alpha_i = \frac{n_i}{n_i + n_j}(1 - \beta), \quad \alpha_j = \frac{n_j}{n_i + n_j}(1 - \beta), \quad \beta < 1. \tag{4.3}$$

ここで可変法の一般化と名づける理由は, このパラメータの値を調整することで, 次の 5 つの手法を表現することができるからである. まず, 距離の更新時に $n_i = n_j$ とおくと従来の可変法となり, $\beta = 0$ とおくと群平均法になる. さらに, $n_i = n_j$, $\beta = 0$ とおくと加重平均法に, $n_i = n_j$, $\beta = \delta/(\delta + 2\varepsilon)$ とおくと最近隣法に, $n_i = n_j$, $\beta = -\delta/(\delta + 2\varepsilon)$ とおくと最遠隣法になる. これをもとにさらにいくつかの変形手法が考えられる.

表 4.1: パラメータ $(\alpha_i, \alpha_j, \beta)$ で再構成された組み合わせ的階層的分類法

手法	α_i	α_j	β
可変法	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	$\lvert\beta\rvert < \infty$
最近隣法	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	$0 < \frac{\delta}{\delta+2\varepsilon} < 1$
最遠隣法	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	$-1 < \frac{-\delta}{\delta+2\varepsilon} < 0$
群平均法	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0
加重平均法	$\frac{1}{2}$	$\frac{1}{2}$	0
ウォード	$\frac{n_i+n_k}{n_t}$	$\frac{n_j+n_k}{n_t}$	$-\frac{n_k}{n_t}$
重心法	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$-\alpha_i\alpha_j$
メディアン法	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$

$\delta = d(C_j, C_k) - d(C_i, C_k) > 0$, $\varepsilon = d(C_i, C_k) - d(C_i, C_j) > 0$.

次に, この一般化可変法の距離空間のひずみの条件を調べると次のようになる. まず, δ と ε を用いて式変形を行うと, 縮小の条件は,

$$\frac{n_j\delta}{n_j\delta + (n_i+n_j)\varepsilon} \leq \beta \tag{4.4}$$

である. 次に, 拡大の条件は,

$$-\frac{n_i\delta}{n_j\delta + (n_i+n_j)\varepsilon} \geq \beta \tag{4.5}$$

である. これらの 2 つの条件を満たさないとき, つまり保存の条件は,

$$-\frac{n_i\delta}{n_j\delta + (n_i+n_j)\varepsilon} < \beta < \frac{n_j\delta}{n_j\delta + (n_i+n_j)\varepsilon} \tag{4.6}$$

となる.

この手法の特性をさらに明らかにするため, 群平均法の距離の更新式

$$D(\mathrm{GA}) = \frac{n_i}{n_i+n_j} d(C_i, C_k) + \frac{n_j}{n_i+n_j} d(C_j, C_k)$$

を利用して式変形を行うと, 次のようになる.

$$D(\mathrm{GF}) = D(\mathrm{GA}) - \frac{\beta}{n_i+n_j}\{n_j\delta + (n_i+n_j)\varepsilon\}. \tag{4.7}$$

表 4.2: 一般化可変法に含まれる手法

パラメータ β の条件	クラスターサイズの条件 $n_i = n_j$	$n_i \neq n_j$
$\beta < 0$	最遠隣法 (CL) 通常の可変法 (FX, $\beta = -\frac{1}{4}$)	修正可変法 (MF, MF*)
$\beta = 0$	加重平均法 (WA)	群平均法 (GA)
$\beta > 0$	最近隣法 (SL)	—

一方, ウォード法に対しても同じような式変形を行うと次のようになる.

$$D(\mathrm{WD}) = D(\mathrm{GA}) + \frac{n_k}{n_t}\frac{1}{n_i + n_j}\{n_i\delta + (n_i + n_j)\varepsilon\}. \tag{4.8}$$

これらの 2 つの式で右辺の δ の係数に注目すると, 一般化可変法は n_j, ウォード法は n_i である. ここで, β の値を $-n_k/n_t$ とすると, この一般化可変法はウォード法とよく似た距離の更新をする手法であることがわかる. また, 次節で述べるパラメータ空間で手法を表示したとき (図 5.1) , ウォード法と同じ領域で定義される. この手法を修正可変法 (modified flexible method) と名付け, MF* と略記する.

また, β の値として $\beta = -(n_i + n_j)/n_k$ と与える方法も考えられ, このとき更新式は

$$D(\mathrm{MF}) = D(\mathrm{GA}) + \frac{1}{n_t}\{n_j\delta + (n_i + n_j)\varepsilon\} \tag{4.9}$$

となる. この手法を MF と略記する.

表 4.2に一般化可変法で示される手法 (generalized flexible methods) をまとめた. これは階層的分類法のうち, 重心法, メディアン法, ウォード法以外の手法が含まれる.

第 5 章

パラメータ空間の特性

組み合わせ的手法のパラメータの性質を調べることは, この手法をより一般的に議論する上で有用である. その主な理由として次のことが挙げられる.

(1) パラメータ空間 $(\alpha_i, \alpha_j, \beta)$ と距離空間のひずみの関係が明らかになる.

(2) パラメータ空間における各手法の位置付けが明らかになる.

(3) パラメータ空間における各手法の領域 (定義域) が明確になり, 手法相互の関係が調べられる.

この章では, まず 4 つのパラメータ $\{\alpha_i, \alpha_j, \beta, \gamma\}$ の定義域を定め, 次にこのパラメータ空間における手法の領域 (定義域) について検討する. また, 先に述べたように, パラメータ γ を除いて議論できるので, $\{\alpha_i, \alpha_j, \beta\}$ の 3 つのパラメータを中心に議論する.

5.1 パラメータ空間における手法間の関係

組み合わせ的手法はパラメータが定数で与えられる手法 (最近隣法, 最遠隣法, 加重平均法, メディアン法など) と, 距離の各更新段階のクラスターサイズの重みとして与えられる可変の手法 (群平均法, 重心法, ウォード法など) がある. 既存のいくつかの組み合わせ的階層分類法のパラメータの定義域を考慮すると, 以下のようなパラメータの定義域が得られる (Ohsumi and Nakamura (1989), Nakamura and Ohsumi (1990)):

$$0 < \alpha_i, \alpha_j < 1; |\beta| < 1; (|\gamma| < 1.)$$

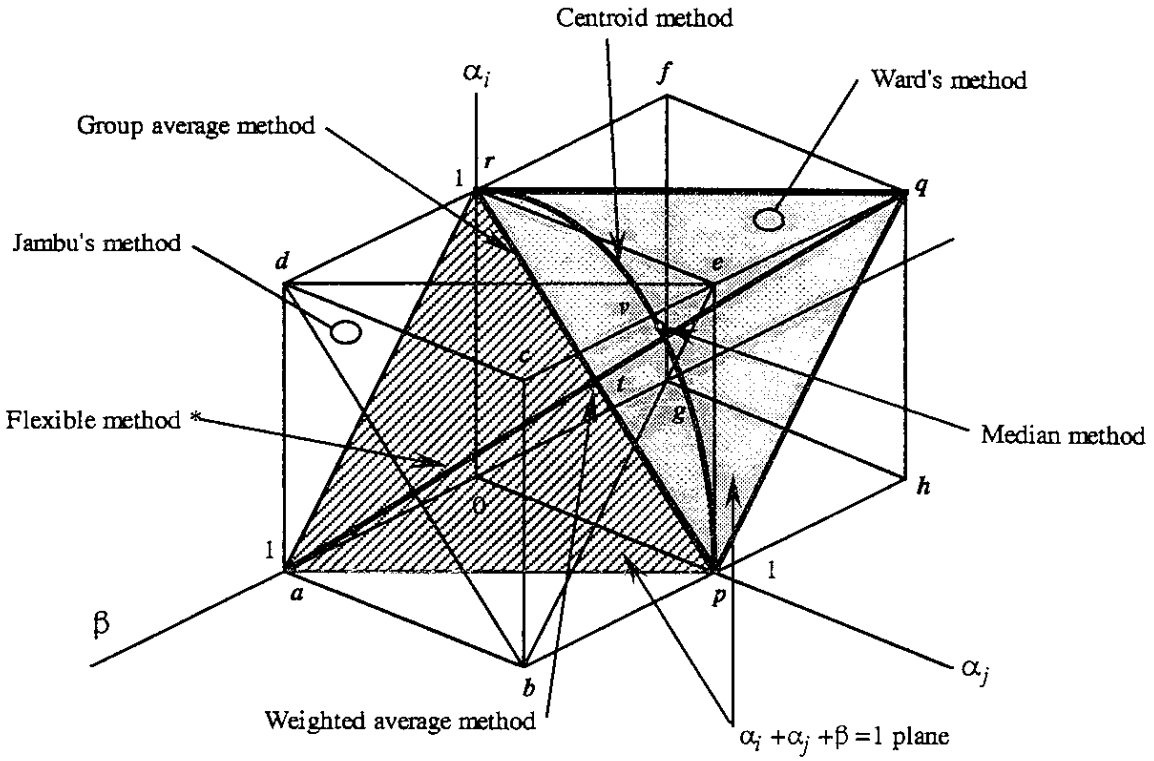

図 5.1: パラメータ空間 $(\alpha_i, \alpha_j, \beta)$
*$\beta > 0$: 最近隣法 (single linkage), $\beta < 0$: (最遠隣法) complete linkage.

図 5.1はパラメータ $(\alpha_i, \alpha_j, \beta)$ で構成される空間において, 組み合わせ的階層分類法に含まれる各手法の対応する (占有する) 領域を示したものである.

このとき各手法が占める領域は次のようになる. まず, 群平均法は, $\alpha_i + \alpha_j = 1$ $(0 < \alpha_i, \alpha_j < 1, \beta = 0)$ の直線上にあり, その領域は図 5.1の線分 $\overline{rp}$ に相当する. したがって, この線分の中点 t は加重平均法に対応する. 重心法は, $\alpha_i + \alpha_j = 1$, $\beta = -\alpha_i \cdot \alpha_j$, $(0 < \alpha_i, \alpha_j < 1)$ の 2 つの曲面の交線の 2 次曲線 $\widehat{rvp}$ がその領域である (2 次曲線の式は $\beta = \alpha_i^2 - \alpha_i = \alpha_j^2 - \alpha_j$ である). この 2 次曲線の頂点 (点 v) がメディアン法に相当する. ウォード法は, 平面 $\alpha_i + \alpha_j + \beta = 1$ $(0 < \alpha_i, \alpha_j < 1; -1 < \beta < 0; \alpha_i, \alpha_j > |\beta|)$ の β が負の領域となる三角形の部分 ($\triangle rpq$) である. 可変法は, 直線 $\alpha_i = \alpha_j = (1 - \beta)/2$ (線分 $\overline{aq}$) である.

最近隣法と最遠隣法が可変法として一般化されているので, この 2 つの方法をパラメータ空間に位置づけると次のようになる. 図 5.1で, 可変法の β の正の領域 (線分 $\overline{at}$) は, 最近

表 5.1: パラメータ空間 $(\alpha_i, \alpha_j, \beta)$(図 5.1) における距離空間のひずみのまとめ

手法	領域	距離空間のひずみ
通常の最近隣法	—	最近隣法を参照 (可変法)
通常の最遠隣法	—	最遠隣法を参照 (可変法)
可変法	線分 aq	縮小, 保存, 拡大
最近隣法 (可変法)	線分 at	縮小
最遠隣法 (可変法)	線分 tq	拡大
群平均法	線分 rp	保存
加重平均法	点 t	保存
ウォード法	三角形 rpq	保存 または 拡大
重心法	弧 rvp	縮小 または 保存 または 縮小 & 非単調
メディアン法	点 v	縮小 または 保存 または 縮小 & 非単調

隣法に相当する. また, β の負の領域 (線分 $\overline{tq}$) は, 最遠隣法である. 以上を要約すると表 5.1となる.

5.2 パラメータ空間における距離空間のひずみ

前節で示したように階層的分類法にみられる距離空間のひずみの特性は, $(\alpha_i, \alpha_j, \beta)$ 空間内に示すことができるが, この空間は 3 つの領域に分割される. ここで, パラメータ γ を考慮しない理由は, 最近隣法と最遠隣法が可変法として一般化される性質によるが, これを視点を変えて "パラメータ空間内における手法と距離空間のひずみの関係" として考えると, 次のようなことが言える.

(1) パラメータ γ を含む空間により各手法の定義域を示すと, $\gamma = 0$ に相当する領域に複数の手法が重複して表示され, その領域の距離空間のひずみを示すことが困難である.

(2) $(\alpha_i, \alpha_j, \beta)$ 空間内に手法を位置づけると, 通常の最近隣法と最遠隣法が加重平均法と同じ領域（点 t）に対応する. したがって, 点 t は距離空間のひずみが保存・拡大・縮小の 3 つの状態を示すことになる. しかし, 可変法に最近隣法と最遠隣法を含めると, 点 t は常に保存の領域となり, パラメータ空間内のひずみの領域を明らかにすることができる.

(3) 多くの手法は距離空間のひずみがパラメータ β に強く依存している.

などが考えられる. 以上の理由により $(\alpha_i, \alpha_j, \beta)$ 空間について, 距離空間のひずみを議論する.

このとき, パラメータ空間 $(\alpha_i, \alpha_j, \beta)$ には次のような特徴がある.

(1) 距離空間のひずみの領域が視覚的に明らかになる.

(2) γ を含むパラメータ空間に比べて手法相互の位置関係の理解が容易である.

手法相互の位置関係と手法の距離空間のひずみの特性から, 次の (R1) ～ (R3) の 3 つの領域にパラメータ空間が分割される. これに加えて, 単調性の条件のみを考えた領域 (R4) がある.

(R1) 常に空間を保存する領域:

$$\begin{aligned} & \{(\alpha_i, \alpha_j, \beta) | \alpha_i + \alpha_j + \beta = 1; 0 < \alpha_i, \alpha_j < 1; \beta = 0\} \\ = \quad & \{(\alpha_i, \alpha_j, \beta) | \alpha_i + \alpha_j = 1; 0 < \alpha_i, \alpha_j < 1\} \end{aligned}$$

これは群平均法, 加重平均法に対応する領域である. また, 可変法で $\beta = 0$ としたときに相当する（線分 rp）.

(R2) 拡大と保存が混在する領域:

$$\begin{aligned} & \{(\alpha_i, \alpha_j, \beta) | \alpha_i + \alpha_j + \beta > 1; 0 < \alpha_i, \alpha_j < 1; -1 < \beta < 1\} \\ \cup \quad & \{(\alpha_i, \alpha_j, \beta) | \alpha_i + \alpha_j + \beta = 1; 0 < \alpha_i, \alpha_j < 1; -1 < \beta < 0\}. \end{aligned}$$

ただし, 距離空間を保存するのは, $\alpha_i + \alpha_j + \beta = 1$ のとき

$$-\frac{\alpha_j \delta}{\varepsilon + \delta} < \beta < \frac{\alpha_j \delta}{\varepsilon}$$

を満足するときである. また, $\alpha_i + \alpha_j + \beta \neq 1$ では,

$$\alpha_j \delta + \beta\varepsilon < (\alpha_i + \alpha_j + \beta - 1)\ d(C_i, C_k) < (1 - \alpha_j)\delta + \beta\varepsilon$$

または

$$(\alpha_i + \beta - 1)\delta + \beta\varepsilon < (\alpha_i + \alpha_j + \beta - 1)\ d(C_j, C_k) < (\alpha_i + \beta)\delta + \beta\varepsilon$$

のときに保存となる (この二つの式は同値で, この条件は (R3) でも同じである).

この領域は, 可変法において β が負の領域とウォード法に対応する領域であり, 立体 ***abp-dcqr*** の $\triangle \boldsymbol{rap}$ を含まない領域である.

(R3) 縮小と保存が混在する領域:

$$\begin{array}{ll} & \{(\alpha_i, \alpha_j, \beta) | \alpha_i + \alpha_j + \beta < 1; 0 < \alpha_i, \alpha_j < 1; -1 < \beta < 1\} \\ \cup & \{(\alpha_i, \alpha_j, \beta) | \alpha_i + \alpha_j + \beta = 1; 0 < \alpha_i, \alpha_j < 1; 0 < \beta < 1\}. \end{array}$$

重心法, メディアン法, および可変法において β が正の領域である. この領域は, ***aphg-frq*** で, 平面 $\alpha_i + \alpha_j + \beta = 1$ の $\beta > 0$ 領域を含み, $\triangle \boldsymbol{rpq}$ を含まない.

(R4) 単調性の条件を満足する領域:

$$\{(\alpha_i, \alpha_j, \beta) | \alpha_i + \alpha_j + \beta \geq 1; 0 < \alpha_i, \alpha_j < 1; -1 < \beta < 1\}.$$

この領域は, (R1), (R2), $\triangle \boldsymbol{rap}$ の和集合である. この領域に入る手法は, 重心法とメディアン法以外の手法である.

ここで, 単調でない領域 ((R4) の補集合) においては更新距離の逆転が起き, 距離空間を縮小することを示している. つまり,

$$d(C_i \cup C_j, C_k) < d(C_i, C_j)$$

となりうることを意味し, 距離空間を縮小する領域 (R3) でも, この可能性がある.

5.3 第 II 部の要約

これまでに議論してきた手法の特徴とパラメータ空間内での手法相互の関係, パラメータ空間内での距離空間のひずみ, 距離空間のひずみと単調性の関係を要約すると次のようになる.

(1) パラメータ空間内で, 平面 $\alpha_i + \alpha_j + \beta = 1$ に近い階層的分類法は, 距離空間を保存する傾向にある.

(2) パラメータ空間を常に保存をする領域に位置する手法は, データ構造によらず常に距離空間を保存する (群平均法と加重平均法).

(3) パラメータ空間内で拡大 (縮小) に領域にある手法は, 与えられたデータ構造を反映して拡大 (縮小) を起こす.

(4) 結合距離の逆転現象は距離空間のひずみの概念と密接な関係にある.

これまで, AHC 手法の距離空間のひずみについていくつか議論されてきたが (表 3.3), いずれも主観的な評価や, 限られた評価でしかなかった. ここでは距離空間のひずみの成立条件を詳しく与えることによって数理的な考察が可能になり, 階層的分類法を客観的に評価することができた. さらに, パラメータ空間内における手法相互の位置関係や, これと距離空間のひずみの関係も明らかになった. また, 距離空間のひずみの成立条件を与えたことにより, AHC 手法に含まれる階層的分類法の客観的な評価が, より一般的に可能となった. さらに, 単調性やクラスター化の過程の距離空間のひずみを考慮することによって, 新しいアルゴリズムの開発の可能性に手がかりを与えることができた.

以上にみたように, 距離空間のひずみの特性に基づいて, 階層的分類法の特徴付けを体系的に考察することができた. これは次の第III部に述べる多変量混合分布モデルにおける検討課題の一つであるパラメータ推定のための初期値設定の方法の予備的考察である.

第 III 部

多変量混合分布モデルによる分類法

二十数年ほど前まではクラスター化法は他の多変量の統計的手続きとはほとんど独立に発展してきたが, ここ十数年の間にこれらの間との関係付けが進められている (Gordon (1981)). とくに初期の研究者は, クラスターのまとまりの度合い (コンパクトさ) の定義として直感的にわかりやすい平方和基準を目的関数として用いた. その後観測値を正規混合分布モデルからの実現値とみて分類することが考えられた (Day (1969), Wolfe (1970)) [1]. 正規混合分布モデルによる分類法は, 確率分布モデル (尤度) を導入した分類手法であり, さらにコンピュータの処理速度と数値計算技術の向上も手伝って, 近年様々な分野で広く使われつつある手法である.

第 III 部では, まず混合分布モデルのパラメータ推定法などの基本的事項についてまとめ, 次にこのモデルによる分類法の特徴や問題点についてまとめる. さらに多変量正規分布より裾の重いデータへの対処方法として, 多変量 t 混合分布モデルを提案する. ここで, 多変量 t 混合分布モデルは正規混合分布モデルの自然な拡張として, 正規混合分布モデルを包含するモデルとして扱えるという利点がある. さらに, 階層的分類法および非階層的分類法 (k-means 法) などの, 従来利用されてきたクラスター化法を, 正規混合モデルの初期値設定の手続きとして利用する分類方式を提案する.

混合分布モデルにおけるパラメータ推定の方法論として, 一般に最尤法が用いられる. このとき, 次の二つの問題が存在する. 第一の問題は尤度関数の局所的な解が複数存在して, 収束の遅い EM 法 (Expectation-Maximization algorithm; Dempster, Laird and Rubin (1977), Redner and Walker (1984)) を用いて推定を行うとき, 最適解を探すための労力が余計にかかるという点である. 第二はコンポーネント数の推定の問題である. データ解析を意識して書かれた McLachlan and Basford (1988) の著書では, この二つの問題に対してとくに前向きの意見や提言があるようには見られない. しかし, 第一の問題に関しては, 彼らは次のように述べている: 「数多くの初期値から出発してパラメータ推定をし, 反復計算を長い時間走らせ, そして, パラメータ空間内で最大尤度に対応するパラメータ値を最尤推定値として選ぶとよい.」

[1] 混合分布モデルの確率分布モデルとしての歴史は大変古い (Peason (1894)) が, 分類法として用いられるようになったのはこのあたりの論文からである.

この主張を受け入れて，ここではデータから混合分布を推定する際に複数のクラスター化法でデータの初期分類を行い，大域的最適解を得る可能性を高めるような手続きを提案する．そのために，まず，通常の EM 法の初期値設定の問題点について第 7章で述べる．第二のコンポーネント数の問題は非常に重要であり，従来様々な方法が提案されている．たとえば，この問題に対する 1 つのアプローチは，この仮説検定の枠組みの中で尤度比検定統計量を用いる方法であるが，検定統計量の分布は，通常の漸近 χ^2 近似が成り立たないことが知られている．そこで，本論文では情報量規準を用いた新たなコンポーネント数の推定手続きを提案する．

第 6 章

多変量 t 分布の混合分布モデル

この章では, まず多変量正規混合分布モデルのパラメータの最尤推定法である, EM 法のアルゴリズムについて述べる. 次に多変量 t 混合分布モデルを提案する. これは正規分布より裾の重いデータへの対処方法として導出したモデルである. そのための準備として楕円分布族のパラメータ推定法である *reweighting* 法について述べる.

6.1 EM 法による正規混合分布モデルのパラメータ推定法

いま, $\boldsymbol{X}_N = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ を p 変量 (次元または変数), 標本の大きさ N の観測値ベクトルとする. この $\boldsymbol{X}_N$ から r 個のコンポーネント分布からなる正規混合分布を推定するとき, その確率分布モデルは次のように書くことができる.

$$f(\boldsymbol{x}|\boldsymbol{\Phi}_r) = \sum_{k=1}^{r} \pi_k f_k(\boldsymbol{x}|\boldsymbol{\theta}_k). \tag{6.1}$$

ここで,

$$\boldsymbol{\Phi}_r = \{\pi_1, \cdots, \pi_{r-1}, \boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_r^T\}^T \subset \boldsymbol{R}^{r-1+\sum_{k=1}^{r} P_k},$$

$$\pi_k \in (0,1), \quad \sum_{k=1}^{r} \pi_k = 1,$$

である. π_k は混合比率, コンポーネント分布 $f_k(\cdot|\cdot)$ は多変量正規分布, $\boldsymbol{\theta}_k$ は $f_k(\cdot|\cdot)$ のパラメータ, P_k は $\boldsymbol{\theta}_k$ の次元, R は実数の集合である.

$f_k(\cdot|\cdot)$ は多変量正規分布を想定しているので, ここで推定すべきパラメータは各コンポーネント分布の平均ベクトル $\boldsymbol{\mu}_k$, 分散共分散行列 $\boldsymbol{\Sigma}_k$, そして混合比率 π_k である. パラメータ

推定は最尤推定法を用いることとすると, このモデルの対数尤度関数は次式で与えられる.

$$\begin{aligned} L(\boldsymbol{\Phi}_r|\boldsymbol{X}_N) &= \log[\prod_{i=1}^{N}\sum_{k=1}^{r}\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)] = \sum_{i=1}^{N}\log\{\sum_{k=1}^{r}\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)\} \\ &= \sum_{i=1}^{N}\log\{\sum_{k=1}^{r}\pi_k(2\pi)^{-p/2}|\boldsymbol{V}|^{1/2}\exp[-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{V}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)]\} \end{aligned}$$

パラメータの最尤推定量はこの関数をパラメータで偏微分してゼロとおけばよいので,

$$\frac{\partial L}{\partial \boldsymbol{\Phi}_r} = \mathbf{0}$$

となる. これを解くと次の推定量が得られる（付録 B）.

$$\hat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i),$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{\pi_k N} Pr(k|\boldsymbol{x}_i)\boldsymbol{x}_i,$$

$$\widehat{\boldsymbol{\Sigma}}_k = \frac{1}{\pi_k N}\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T.$$

ここで,

$$Pr(k|\boldsymbol{x}_i) = \frac{\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{f(\boldsymbol{x}_i|\boldsymbol{\Phi}_r)}$$

は $\boldsymbol{x}_i$ が第 k コンポーネント分布に所属する確率（事後確率）である. これらのパラメータの推定には, 通常 EM 法 (Dempster, Laird and Rubin (1977), Redner and Walker (1984) など) が用いられる. そのアルゴリズムは次のとおりである.

正規混合分布モデルのパラメータ推定法

ステップ **1** (初期値設定) $\boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}, \pi_k^{(0)}$ としてパラメータの初期値を与える. また, 十分小さい正数 $\delta, \varepsilon > 0$ を設定し, 反復計算のカウンタを $t \leftarrow 1$ とする.

ステップ **2** (*E-Step*: 事後確率の推定) 第 t ステップの事後確率を 1 ステップ前のパラメータの推定値を用いて計算する.

$$Pr^{(t)}(k|\boldsymbol{x}_i) = \frac{\pi_k^{(t-1)} f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k^{(t-1)})}{f(\boldsymbol{x}_i|\boldsymbol{\Phi}_r^{(t-1)})}$$

ステップ **3** (*M-Step*: 対数尤度の最大化) 対数尤度の最大化として次のパラメータ値の更新を行う.

$$\pi_k^{(t)} = \frac{1}{N}\sum_{i=1}^{N} Pr^{(t)}(k|\boldsymbol{x}_i),$$

$$\boldsymbol{\mu}_k^{(t)} = \frac{1}{\pi_k^{(t)} N} Pr^{(t)}(k|\boldsymbol{x}_i)\boldsymbol{x}_i,$$

$$\boldsymbol{\Sigma}_k^{(t)} = \frac{1}{\pi_k^{(t)} N}\sum_{i=1}^{N} Pr^{(t)}(k|\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t)})^T.$$

そして, 対数尤度 $L^{(t)}$ を計算する.

ステップ **4** (収束判定) 収束判定として

$$|L^{(t)} - L^{(t-1)}| < \varepsilon \quad \text{または} \quad \|\boldsymbol{\Phi}_r^{(t)} - \boldsymbol{\Phi}_r^{(t-1)}\| < \delta$$

であれば反復計算を終了し, そうでなければステップ 2 へ戻る.

6.2 楕円分布族のパラメータ推定法

標本分散共分散行列を推定する場合, 標本に外れ値が混入するとその推定に大きく影響を与えることは良く知られていることである. これを克服する方法としてロバスト推定法があり (Huber (1981) など), 正規分布より裾の重い分布を表現できる楕円分布族 (狩野 (1992), Kano, Berkane and Bentler (1993) など) を用いたロバストな統計的推定法が考えられてきた. ここではとくに多変量 t 分布を念頭においた, ロバストな推定法について話を進める.

楕円分布の確率密度関数は,

$$|\boldsymbol{V}|^{-1/2} h\{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})|\nu\} \tag{6.2}$$

と表される. ここで $\boldsymbol{V}$ は擬分散共分散行列[1], $\boldsymbol{\mu}$は位置ベクトル[2]である. いま, 互いに独立な標本 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ から, この確率密度関数に基づいて未知のパラメータ $(\boldsymbol{\mu}, \boldsymbol{V})$ の推定を最尤法で行う. ここで ν は $(\boldsymbol{\mu}, \boldsymbol{V})$ 以外のパラメータで, 多変量 t 分布の自由度に相当するパラ

[1]楕円分布族において $\boldsymbol{V}$ は, 正規分布の分散共分散行列のようなものであるので, ここでは擬分散共分散行列と呼ぶことにする (scale matrix: 尺度行列とも言う).

[2]楕円分布族の仮定の下で $\boldsymbol{\mu}$ は, 正規分布の平均ベクトルに相当し, これは位置ベクトルと呼ばれる.

メータである．これは初めに適当な値 $\nu = \nu_0$ が与えられていると仮定して話を進める．ν の推定に関しては改めて後で述べる．

$(\hat{\mu}, \widehat{V})$ は上述の確率密度関数の対数尤度関数の最大化，すなわち各パラメータで偏微分して o とおくことで求められる (付録 B)．尤度関数を次式で定義すると，

$$\mathcal{L}(\mu, \boldsymbol{V}|\nu_0) = \prod_{i=1}^{N} |\boldsymbol{V}|^{-1/2} h\{(\boldsymbol{x}_i - \mu)^T \boldsymbol{V}^{-1} (\boldsymbol{x}_i - \mu)|\nu_0\}, \tag{6.3}$$

$\frac{\partial \log \mathcal{L}}{\partial \mu} = \mathrm{o}, \frac{\partial \log \mathcal{L}}{\partial \boldsymbol{V}} = \mathrm{o}$ より，

$$\hat{\mu} = \frac{1}{\sum_{i=1}^{N} w(\hat{s}_i^2|\nu_0)} \sum_{i=1}^{N} w(\hat{s}_i^2|\nu_0) \boldsymbol{x}_i, \tag{6.4}$$

$$\widehat{V} = \frac{1}{N} \sum_{i=1}^{N} w(\hat{s}_i^2|\nu_0)(\boldsymbol{x}_i - \hat{\mu})(\boldsymbol{x}_i - \hat{\mu})^T. \tag{6.5}$$

ここで，

$$\begin{cases} \hat{s}_i^2 = (\boldsymbol{x}_i - \hat{\mu})^T \boldsymbol{V}^{-1} (\boldsymbol{x}_i - \hat{\mu}), \\ w(u|\nu) = -2 \dfrac{\partial}{\partial u} \log h(u|\nu) \end{cases}$$

である．いま p 次元の多変量 t 分布

$$Mt_p(\mu, \boldsymbol{V}|\nu) = \frac{\Gamma((p+\nu)/2)}{(\pi\nu)^{p/2} \Gamma(\nu/2) |\boldsymbol{V}|^{1/2}} \left\{ 1 + \frac{(\boldsymbol{x} - \mu)^T \boldsymbol{V}_k^{-1} (\boldsymbol{x} - \mu)}{\nu} \right\}^{-(p+\nu)/2} \tag{6.6}$$

を考えると，$w(\cdot|\cdot)$ は

$$w(s^2|\nu) = \frac{\nu + p}{\nu + s^2} \tag{6.7}$$

となる．この $w(\cdot|\cdot)$ は，標本 $\boldsymbol{x}_i$ のウェイトとして解釈できる．各標本の擬マハラノビス距離 $\hat{s}_i^2$ が大きい標本に対してはウェイトが小さくなり，距離が小さければ大きくなる．つまり $Mt_p(\mu, \boldsymbol{V}|\nu)$ に基づくパラメータの推定法は，平均値から遠く離れた (外れ値と思われる) 標本の影響を低減させる効果がある．

とくに多変量 t 分布のウェイト関数を見ると，パラメータ ν の値が正規性からのずれを表し，ロバストネスの程度を示すパラメータであることがわかる．この意味で ν は robustness tuning parameter (Lange, Little and Taylor (1989)) あるいは，extra shape parameter

(Taylor (1992)) などと呼ばれる. ここでは裾の重さの程度を決定するパラメータであるので, "形状パラメータ"(shape parameter) と呼ぶことにする.

(6.4), (6.5) 式のパラメータの推定量の評価方法は, Dempster, Laird and Rubin (1980), Rubin (1983) による *reweighting* 法が有効である. これも EM アルゴリズムの枠組でとらえることができ, 次のように計算される.

Reweighting Algorithm

ステップ **1** 初期値設定

$\boldsymbol{\mu}^{(0)} = \bar{\boldsymbol{x}}$ (標本平均ベクトル), $\boldsymbol{V}^{(0)} = \boldsymbol{S}$ (標本分散共分散行列), $\varepsilon > 0$, $\delta > 0$, $\nu = \nu_0$, $t \leftarrow 1$.

ステップ **2** *E-step*: ウェイトの計算

$$w_i^{(t)} = w((\boldsymbol{x}_i - \boldsymbol{\mu}^{(t-1)})^T (\boldsymbol{V}^{(t-1)})^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}^{(t-1)}) | \nu_0)$$

ステップ **3** *M-step*: パラメータ値の更新

$$\begin{cases} \boldsymbol{\mu}^{(t)} &= \dfrac{1}{\sum_{i=1}^N w_i^{(t)}} \displaystyle\sum_{i=1}^N w_i^{(t)} \boldsymbol{x}_i, \\ \boldsymbol{V}^{(t)} &= \dfrac{1}{N} \displaystyle\sum_{i=1}^N w_i^{(t)} (\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)})(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)})^T \end{cases}$$

を評価し, $\mathcal{L}^{(t)}$ を計算する.

ステップ **4** 収束判定

$|\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}| < \varepsilon$ または, $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(t-1)}\| < \delta$ and $\|\boldsymbol{V}^{(t)} - \boldsymbol{V}^{(t-1)}\| < \delta$ であれば終了, そうでなければ $t \leftarrow t+1$ としてステップ 2 へ戻る.

このアルゴリズムは非常に単純であるが, 収束が遅いという特性がある.

これまではパラメータ ν は $\nu = \nu_0$ と固定して議論してきたが, このパラメータの推定方法として以下の方法が考えられる:

1. $\nu = \nu_0$ として固定する,

2. $\hat{\nu}$ をある方法で推定してから $(\boldsymbol{\mu}, \boldsymbol{V})$ の推定を行う（格子点探索 (grid search) を含む），

3. ν と $(\boldsymbol{\mu}, \boldsymbol{V})$ の推定を同時に行う．

とくに多変量 t 分布 $Mt_p(\cdot)$ では，ν は分布の形状を決定するパラメータなので，一般論として，大標本の場合に ν はデータから推定可能であると考えられる．一方，小標本の場合には安定した解が得られにくいと考えられるので，ν に適当な値を与えることが考えられる．また，大変敏感なパラメータなので，その扱いには注意が必要である．Taylor (1992) は ν を推定することの計算量のコストを数値実験により検討している．Lange, Little and Taylor (1989) は **3.** の方法が一番計算量が多く，収束が遅いと報告している．

ここで Lange, Little and Taylor (1989) によるパラメータ ν の推定方法を示しておく．第 t ステップで推定されたパラメータを $\boldsymbol{\mu}^{(t)}, \boldsymbol{V}^{(t)}, \nu^{(t)}$ として与え，*E-step* で $w_i^{(t)}$ を計算する．そして，次の量を計算する：

$$u_i^{(t)} = \psi(\frac{\nu^{(t)}}{2} + \frac{1}{2}) - \log(\nu^{(t)} + s_i^2),$$

ここで，

$$\begin{cases} \psi(x) &= \dfrac{d}{dx}\log\Gamma(x), \quad (digamma\ function), \\ s_i^2 &= (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}). \end{cases}$$

M-step では $\boldsymbol{\mu}^{(t+1)}, \boldsymbol{V}^{(t+1)}$ を計算し，$\nu^{(t+1)}$ として次式を最大にする ν を計算する：

$$\ell(\nu) = \frac{N\nu}{2}\log(\frac{\nu}{2}) - N\log\Gamma(\frac{\nu}{2}) + (\frac{\nu}{2} - 1)\sum_{i=1}^{N} u_i^{(t)} - \frac{\nu}{2}\sum_{i=1}^{N} w_i^{(t)}.$$

これは ν に関する 1 次元探索なので簡単に計算できる (たとえばニュートン法などを用いれば良い)．

6.3 多変量 t 分布の混合分布モデルのパラメータ推定法

6.1節で混合分布モデルの各コンポーネント分布が正規分布であるときの各パラメータの推定方法について述べた．この節では，各コンポーネントが正規分布の拡張である楕円分

布族に従う一般的なモデルを考える．とくに多変量 t 分布を用いて，正規分布より裾が重いデータに対応できる混合分布モデルを提案する．

観測値に対して r 個のコンポーネント分布からなる混合分布モデル

$$f(\boldsymbol{x}|\boldsymbol{\Phi}) = \sum_{k=1}^{r} \pi_k g_k(\boldsymbol{x}|\boldsymbol{\theta}_k) \tag{6.8}$$

のあてはめを考える．ここで，π_k はコンポーネント分布の混合比率パラメータとし ($\sum \pi_k = 1$, $\pi_k \in (0,1)$), $\boldsymbol{\Phi} = \{\pi_1, \cdots, \pi_k, \cdots, \pi_{r-1}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k, \cdots, \boldsymbol{\theta}_r\}$ とおく．

各コンポーネント分布の確率密度関数 $g_k(\cdot)$ を楕円分布族とすると，それは一般的に次のように書くことができる：

$$g_k(\boldsymbol{x}|\boldsymbol{\theta}_k) = |\boldsymbol{V}_k|^{-1/2} h_k\{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{V}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)|\nu_k\}. \tag{6.9}$$

ここで，$h_k(\cdot)$ は $(\boldsymbol{\mu}_k, \boldsymbol{V}_k)$ に依存しない非負の関数，$\boldsymbol{\mu}_k$ は第 k コンポーネント分布の位置ベクトル，$\boldsymbol{V}_k$ は擬分散共分散行列，ν_k は $\boldsymbol{\mu}_k$, $\boldsymbol{V}_k$ とは独立なあるパラメータ，$\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{V}_k, \nu_k\}$ である．

混合分布モデル (6.8) 式のパラメータの推定は，観測値 $\boldsymbol{X}_N = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ を用いて，以下の手順により行う．パラメータ $\nu_k (k = 1, \cdots, r)$ を事前に与えておき ($\nu_k = \nu_k^{(0)}$)，目的関数として擬対数尤度関数 (pseudo log likelihood function)

$$\mathcal{L}(\boldsymbol{\Phi}|\boldsymbol{X}_N) = \sum_{i=1}^{N} \log \sum_{k=1}^{r} \pi_k |\boldsymbol{V}_k|^{-1/2} h_k\{(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T \boldsymbol{V}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)|\nu_k^{(0)}\} \tag{6.10}$$

を構成する．これを $\boldsymbol{\mu}_k$, $\boldsymbol{V}_k$ で偏微分して 0 とおくことにより，各コンポーネント分布の位置ベクトルと擬分散共分散行列の推定量は，次式で与えられる：

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{\sum_{i=1}^{N} \widehat{Pr}(k|\boldsymbol{x}_i) w(\hat{s}_{ki}^2|\nu_k^{(0)})} \sum_{i=1}^{N} \widehat{Pr}(k|\boldsymbol{x}_i) w(\hat{s}_{ki}^2|\nu_k^{(0)}) \boldsymbol{x}_i, \tag{6.11}$$

$$\widehat{\boldsymbol{V}}_k = \frac{1}{N\hat{\pi}_k} \sum_{i=1}^{N} \widehat{Pr}(k|\boldsymbol{x}_i) w(\hat{s}_{ki}^2|\nu_k^{(0)}) (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T. \tag{6.12}$$

ここで，

$$\hat{s}_{ki}^2 = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{V}}_k^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k), \tag{6.13}$$

$$\widehat{Pr}(k|\boldsymbol{x}_i) = \frac{\hat{\pi}_k |\widehat{\boldsymbol{V}}_k|^{-1/2} h_k(\hat{s}_{ki}^2|\nu_k^{(0)})}{f(\boldsymbol{x}_i|\widehat{\boldsymbol{\Phi}})}, \tag{6.14}$$

$$\hat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N}\widehat{Pr}(k|\boldsymbol{x}_i), \tag{6.15}$$

$$w(\hat{s}_{ki}^2|\nu_k^{(0)}) = -2\frac{\partial}{\partial \hat{s}_{ki}^2}\log h_k(\hat{s}_{ki}^2|\nu_k^{(0)}). \tag{6.16}$$

である. $\hat{s}_{ki}^2$ は第 k コンポーネント分布と標本 $\boldsymbol{x}_i$ との擬マハラノビス距離（擬分散共分散行列 $\boldsymbol{V}$ を考慮した距離で, $\boldsymbol{V}$ が分散共分散行列の場合はマハラノビス距離になる）, $\widehat{Pr}(k|\boldsymbol{x}_i)$ は $\boldsymbol{x}_i$ が第 k コンポーネント分布に所属する確率（事後確率）, $w(\hat{s}_{ki}^2|\cdot)$ は第 k コンポーネント分布に対する $\boldsymbol{x}_i$ のウェイトである.

これらの式をもとに, ここではEMアルゴリズム (Dempster, Laird and Rubin (1977), Redner and Walker (1984)) を用いて未知パラメータの推定を行う. すなわち, 事後確率 $Pr(\cdot|\cdot)$ とウェイト $w(\cdot|\cdot)$ を欠測データとして扱い, *E-step*(expectation step) でこれらの期待値推定を行う. 次に, *M-step*(maximization step) として各パラメータの推定を行う. 詳しい手順は後で述べる.

とくに, 各コンポーネント分布が多変量 (p 変量)t 分布の場合, その確率密度関数は,

$$\frac{\Gamma((p+\nu_k)/2)}{(\pi\nu_k)^{p/2}\Gamma(\nu_k/2)|\boldsymbol{V}_k|^{1/2}}\left\{1+\frac{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{V}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}{\nu_k}\right\}^{-(p+\nu_k)/2} \tag{6.17}$$

と書けるので (Lange, Little and Taylor (1989) など), $\boldsymbol{x}_i$ に対する第 k コンポーネント分布のウェイトは次のように計算される：

$$w(s_{ki}^2|\nu_k) = \frac{\nu_k+p}{\nu_k+s_{ki}^2}. \tag{6.18}$$

ν_k は整数値とは限らず正の実数値をとる. ν_k の推定量は陽な形で書くことができないため, π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{V}_k$ と同時には推定せずに, 準ニュートン法などで推定を行う.

次にパラメータ ν_k の推定方法について考える. 基本的な考え方はコンポーネント分布が一つのときと同じであるが, 収束の速さ, パラメータの数などを考慮する必要がある. 標本が十分大きいとき, 前節で述べたパラメータ ν の推定方法 **2.** と **3.** のどちらかを選択すれば良いと思われる. ここでは, $\nu_k(k=1,\cdots,r)$ の推定方法は分布の数が 1 つの場合と同様に, 対数尤度を最大化する方法で行なう. しかし, 6.2節で示した Little, Lange and Taylor(1989) の方法はそのまま混合分布モデルに拡張することは非常に困難である. さらに, ν_k は尤度方

程式を陽な形では解けないので, ν_k に関する尤度関数を最大化する方法で, r 個の ν_k を同時に解く最適化の手法を用いる.

この方法によるパラメータ推定は次のように要約される. 「まず, ν_k を固定し, 他のパラメータ $(\boldsymbol{\mu}_k, \boldsymbol{V}_k)$ を EM 法で推定する. 次に, 最適化手法で $\hat{\nu}_k$ を求め, その値を初期値として再び他のパラメータの推定をする. 収束するまでこれを繰り返す.」

詳しくは次の手順でパラメータ推定を行う.

楕円分布族の混合分布モデルのパラメータ推定アルゴリズム

ステップ 1 [初期値設定]

各コンポーネント分布のパラメータ, 事後確率, ウェイトの初期値を設定する. たとえば事後確率の初期値設定のためには何らかの方法で事前に r 群に分類し,

$$Pr^{(0)}(k|\boldsymbol{x}_i) = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \in C_k \\ 0 & \text{otherwise} \end{cases}$$

$(k = 1, \cdots, r; i = 1, \cdots, N)$ とする. ここで, C_k は第 k クラスターである. 次に, この事後確率の初期値をもとにして, $\pi_k^{(0)}$, $\boldsymbol{\mu}_k^{(0)}$, $\boldsymbol{V}_k^{(0)}$ の計算を行う. ウェイト $w^{(0)}(\cdot|\cdot)$ は (6.16) 式を用いて推定し, 形状パラメータ ν_k の初期値 $\nu_k^{(0)}$ は任意の数値を与える. また, 反復回数のカウンタは $t \leftarrow 1$ とする.

ステップ 2 [E-step]

第 t ステップ $(t \geq 1)$ の *E-step* として, 事後確率 $Pr^{(t)}(\cdot|\cdot)$ とウェイト $w^{(t)}(\cdot|\cdot)$ の推定を行う. すなわち, 事後確率は (6.14) 式, ウェイトは (6.16) 式により推定する. 同時に (6.15) 式により $\pi_k^{(t)}$ の推定を行う.

ステップ 3 [M-step(1)]

第 t ステップの *M-step* は, 尤度の最大化により各コンポーネントのパラメータ推定を行う. すなわち $\boldsymbol{\mu}_k^{(t)}$ と $\boldsymbol{V}_k^{(t)}$ を (6.11), (6.12) 式により推定する.

ステップ 4 [M-step(2)]

t 混合モデルの場合, $\pi_k^{(t)}$, $\boldsymbol{\mu}_k^{(t)}$, $\boldsymbol{V}_k^{(t)}$ を固定して ν_k に関する擬対数尤度関数

の最大化により, $\nu_k^{(t)}$ の推定を行う.

ステップ 5 [収束判定]

次の収束条件を満足すれば反復計算を終了し, そうでなければ $t \leftarrow t+1$ としてステップ 2 へ戻る.

$$|\mathcal{L}^{(t)}(\boldsymbol{\Phi}^{(t)}) - \mathcal{L}^{(t-1)}(\boldsymbol{\Phi}^{(t-1)})| < \varepsilon \quad \text{or} \quad \|\boldsymbol{\Phi}^{(t)} - \boldsymbol{\Phi}^{(t-1)}\| < \delta.$$

ここで $\mathcal{L}^{(t)}(\cdot)$ は第 t ステップで推定された対数尤度の値, $\widehat{\boldsymbol{\Phi}}^{(t)}$ は第 t ステップで推定されたパラメータである. ε と δ は十分に小さい正数である (10^{-5} ～ 10^{-7} 程度).

なお, 正規分布の場合ウェイト $w(\cdot|\cdot)$ は 1 になり, 正規混合分布モデルのパラメータ推定法と一致する (付録 B 参照, その他には Everitt and Hand (1981), McLachlan (1992), McLachlan and Basford (1988) など).

ここで, $w_{ki} = w(s_{ki}^2|\cdot)$ と $Pr(k|\boldsymbol{x}_i) \times w(s_{ki}^2|\cdot)$ の解釈について考える. w_{ki} は推定された第 k 番目のコンポーネント分布から観測されたすべての標本 $\boldsymbol{x}_i (i = 1, \cdots, N)$ への擬マハラノビス距離 s_{ki}^2 により調節される ウェイト である. 第 k コンポーネント分布に対する事後確率 ($Pr(k|\boldsymbol{x}_i)$) が小さい標本や, 外れ値と思われる標本のマハラノビス距離は当然大きくなり, そのウェイトも小さくなる. 事後確率と ウェイト の積も同様な解釈ができる. つまり各コンポーネントが良く分かれている場合, 第 k コンポーネント分布への事後確率が小さい標本は, ウェイトも小さな値になり, これらの積はほとんど 0 に近い値になる.

このように, 楕円分布族の混合分布モデルにおける各パラメータの推定は正規混合分布モデルの場合と比べると, 各標本に ウェイト がかかる以外は, ほとんど同じであることがわかる. 実際 $g_k(\cdot)$ が多変量正規分布の場合は $w_{ki} = 1$ になり, 正規混合分布モデルのパラメータ推定方法と同一になる.

前節でコンポーネント数が 1 つの場合の楕円分布のパラメータ推定が収束が遅いことや, 不安定であることを述べた. これが混合分布モデルでコンポーネント数が複数となる場合, この問題が増幅されるのは容易に想像できる. この問題の解決策の一つとしてまず考えられることは, 十分大きな標本数で, 良いパラメータ (真値に十分近いということ) の初期値を与

えることである. その具体的方法としては,

(1) まず, パラメータの初期値として, 正規混合分布モデルで推定されたパラメータ値を使う.

(2) EM 法の収束特性から, データの次元が高くなると収束が遅くなるので, 次元縮約の手法 (主成分分析や射影追跡法など) で観測データの前処理を行う,

などが考えられる. EM 法の適用に先立って, 十分良い初期値を与えることは, (1) 収束までの反復回数が少なくなる, (2) 最適解に収束する, (3) 尤度が発散するパラメータの境界値に収束しない, などの可能性が高くなると考えられる.

第 7 章

多変量混合分布モデルによる分類法

この章は多変量混合分布モデルを用いる分類方式を提案する (中村 (1994)). それに先立って, 混合分布モデルによる分類法の特徴について述べ, 次に, 多変量データにこの方法を適用する際の問題点である, EM 法の初期値設定について考察する. さらに, 分類結果を評価する尺度として用いる判別率について述べる. データ解析の例として, Iris データ, 糖尿病データを用いて提案分類方式の有効性を検証する. また数値実験を通して, 提案分類方式の性質について詳しく議論する.

7.1 混合分布モデルによる分類法の特徴

標本に混合分布モデルをあてはめ, 推定したパラメータを用いて分類を行うことができる. これは判別分析に類似するが, データに基づき "コンポーネント分布" を推定する点が異なる.

混合分布モデルによる分類は次のように行う. 観測データから EM 法を用いて最尤推定されたパラメータを $\hat{\boldsymbol{\Phi}}_r$ とすると, $\boldsymbol{x}_i$ が第 k コンポーネント分布に所属する確率, つまり事後確率は

$$\hat{\tau}_{ki} = \widehat{Pr}(k|\boldsymbol{x}_i) = \frac{\hat{\pi}_k f_k(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_k)}{f(\boldsymbol{x}_i|\hat{\boldsymbol{\Phi}}_r)} \tag{7.1}$$

で定義される. そこで,

$$\hat{\tau}_{ti} > \hat{\tau}_{ki} \quad (k = 1, \cdots, r; k \neq t) \tag{7.2}$$

のとき $\boldsymbol{x}_i$ を第 t コンポーネント分布に割り当てる. これは最適ルールまたはベイズルール (Anderson (1984), 6 章など) による分類で, 全体の誤判別率を最小にする方法である.

混合分布モデルによる分類法の特徴は, それぞれのクラスターに任意の楕円体構造を仮定することにより, より一般的な分類法として記述できることである. すなわち, 分散共分散行列に制約条件を設けると, $\boldsymbol{W}$ を目的関数とし, これを最適化する大半の分割型分類手法をこの考え方で説明できる (Scott and Symons (1971), Banfield and Raftery (1993)). ここで, $\boldsymbol{W}$ はクラスター内分散共分散行列の総和である. たとえば k-means 法は trace($\boldsymbol{W}$) を最小化する基準の手法 (MacQueen (1967) など) であるが, これは分散共分散行列の構造を球形と仮定 ($\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, $\boldsymbol{I}$ は単位行列) することに同値である (McLachlan and Basford (1988), Everitt (1993, p113) など). また, 別の特徴として, 事後確率を用いることにより, 複数のコンポーネント分布にまたがる個体の所属の度合を確率として評価でき, 分類結果の客観的解釈を容易にするということがある.

従来提案されている数多くの分類手法と, 混合分布モデルによる分類法との大きな違いは次の点にある. 従来の手法は, 個体の所属するグループのラベルを推定して排反な分割を作ることが目的で, それはアルゴリズム (個体の割り当て・入れ換え方法や距離の更新ルール) に依存する部分が大きいことである. 一方, 混合分布モデルによる分類法は, 標本に確率分布をあてはめることで, 個体のコンポーネント分布への所属を確率で表現して, それを推定することである. ここで提案する分類方式は, これら両者の利点を併合して使うことに特徴がある.

7.2 EM 法の初期値設定

EM 法を用いて混合分布モデルのパラメータ推定を行うとき, 良い初期値を与えると, EM 法の収束が早くなる, あるいは大域的最適解が得られる可能性が高められることが知られているので, この問題について考える.

従来行われてきた多変量の混合分布モデルにおける EM 法の初期値設定の方法を要約すると次のようになる. まず, 主観的に初期値を設定する方法として次の二つが考えられる.

(1) 散布図でデータを表示して観察し, 適当なパラメータ値を指定する方法.

(2) 散布図でコンポーネント分布の核となる部分を指定する (McLachlan (1988)) 方法.

これに対して, ある程度客観的に行う方法として次の三つがある.

(i) 何らかの分類法で初期分類を行う. たとえば, Wolfe (1967, 1969, 1970) の NORMIX では, ウォード法や k-means 法を用いて初期分類を行っている.

(ii) ランダムに, またはあるルールで (たとえば系統的に) コンポーネント分布のラベルを個体につける. または, Celeux and Diebolt (1985) の方法などもある.

(iii) ベイズの方法でパラメータの事前分布を与える. これについては Diebolt and Robert (1994) などの論文がある.

また特別な場合としては次の方法がある.

(3) すでにグループのラベルが標本に付与されている場合. たとえば医学データなどで専門的知見に基づく分類がされている場合 (McLachlan (1992), 6.8 節, Ganesalingam and McLachlan (1980) など).

(4) シミュレーションによる数値実験では, サンプルデータを乱数により発生させる際に用いた真のパラメータ値を, 初期値とする方法がある (たとえば Everitt and Hand (1981)). しかしこの方法は実際のデータ解析には使えない.

その他の方法として, 一変量の場合に限られるが, モーメント法による推定値を初期値とする方法が提案されている (Furman and Lindsay (1994a, b)).

ところで実用的な見地からは, 尤度関数の最適解を得るためには McLachlan and Basford (1988) が提案しているような "数多くの初期値からパラメータ推定を行う" という労力を少しでも減らす初期値の与え方について考える必要がある. 混合分布モデルのパラメータ推定に関する数多くの論文では, このことを積極的に議論した研究はほとんどない.

理論的には Redner and Walker (1984) の定理 3.1, 定理 3.2 において, かなり自然な条件のもとで, 十分大きな標本数のとき, 尤度方程式の解は一意に定まり, それは対数尤度の大域的最適解を与えることが示されている. しかし, 実際の問題では, これを満たすほど標本数が十分大きいことは稀である. さらに, 観測値の次元数 (特性値, 変量, 変数の数) とコンポーネント数が多くなると, 推定する混合分布のパラメータ数は, 次元数の 2 乗のオーダーで増加し, 同時に局所的な解の数も増加する. しかしデータ構造やパラメータに関する何らかの

事前情報から初期値を与えることで, 大域的最適解を得る可能性はより高くなると考えられる. これは EM 法におけるパラメータの初期値設定の方式が十分考慮されなければならない要素であることを意味する.

7.3 クラスター化法を用いた EM 法の初期値設定

標本に混合分布モデルをあてはめ, 適当な初期値から出発して EM 法によりパラメータ推定を行った場合, 得られた解が必ずしも大域的な最適解であるという保証はない. この意味で EM 法の初期値設定は大変重要な問題である. この問題に対する一つの方法として, ここでは数理的に性質を考察した複数の階層的分類法を用いる手続きを提案する. これは階層的分類法が所与のデータ構造に応じて固有の分類結果を与えるという特徴を利用して, 様々な初期値設定を行うことである.

ここで提案する分類方式を要約すると次のようになる: 『複数のクラスター化法[1]で混合分布モデルのパラメータの初期値を与える. 次に, それを EM 法の初期値として混合分布のパラメータ推定を行う. こうして得られた複数の解の中から, パラメータ空間内で尤度を最大にするパラメータ値を解として採用する.』 具体的には次の手順で行う.

ステップ **1** 与えられたデータセット $\boldsymbol{X}_N$ に対してクラスター数を r に固定して, クラスター化法で初期分類を行う.

ステップ **2** 初期分類の結果得られた分類結果を EM 法のパラメータの初期値とする.

$$\pi_k^{(0)} = \frac{N_k}{N}$$

$$\boldsymbol{\mu}_k^{(0)} = \frac{1}{N_k} \sum_{i \in C_k} \boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}_k^{(0)} = \frac{1}{N_k - 1} \sum_{i \in C_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(0)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(0)})^T$$

ここで N_k は第 k クラスター C_k のクラスターサイズである $(k = 1, 2, \cdots, r)$.

ステップ **3** ステップ 2 で与えたパラメータを初期値として EM 法を実行する.

[1] ここでクラスター化法とは, 階層的分類法と非階層的分類法のことをさす. とくに第 II 部に述べた AHC 手法と k-means 法を用いる.

ステップ4 用意したc種類のクラスター化法 ($\ell = 1, \cdots, c$) に対して, ステップ1 ～ 3を実行する.

ステップ5 得られたc個の対数尤度の中から, 最大のものを解として採用する.

この分類方式はNORMIX (Wolfe (1967, 1969, 1970)) を拡張したものとして考えることもできる. しかしNORMIXが入力データに対して一意に解が定まるのに対して, ここで提案した分類方式は, "複数の初期値設定を行い, 複数の解の中からできるだけ最適と思われる解を探し, しかも初期値設定が分類結果におよぼす影響を客観的に評価できる" という点で異なる.

ここで, 初期分類にクラスター化法を用いる理由を挙げておく. まず, 混合分布のコンポーネント分布の核となる部分をとらえるため, データの密度の高いところを探す方法が最良と考えられる. その方法として密度推定法 (Silverman (1986)) や, モード法 (Wishart (1969a)), kth nearest neighbour clustering (Wong and Lane (1983)) などがある. しかし, これらの手続きを行うためには様々なパラメータ (密度の閾値等) の設定や, クラスター数の決定問題が関わってくる. これに対してクラスター化法 (階層的分類法および分割型の分類法) は, そのアルゴリズムが明解であることと, クラスター数を指定すると, ある条件下で一意にクラスターができることが最大の利点である. とくに, 階層的分類法はデータが密になっているところからクラスターができる手法であるから, この方法による初期分類は合理的である. さらに, 第II部で考察した階層的分類法の "距離空間のひずみ (保存・拡大・縮小)" の数理に依拠した初期分類を行うことは大きな利点である.

距離空間のひずみの特性を考慮すると, 表7.1にあげた手法の中では拡大する傾向の手法が比較的良い初期分類が与えられるものと考えられる. 非階層的分類法 (分割型最適化手法) としてk-means法を用いるが, すでに7.1節で述べたようにこの手法は混合分布モデルによる分類法の特別な場合に相当するので, これも初期値設定の方法として候補となりうる.

初期分類法としては, 階層的・非階層的分類法の中の手法を用いることができるが, ここでは主に表7.1に挙げた階層的分類法とk-means法を数値実験等で用いることとする.

表 7.1: 階層分類法の距離空間のひずみ

距離空間の ひずみ	手法	略名
常に縮小	最近隣法	SL
(縮小傾向)	メディアン法	MD
縮小と保存	重心法	CD
常に	群平均法	GA
保存	加重平均法	WA
常に拡大	最遠隣法	CL
保存と	可変法	FX
拡大	ウォード法	WD
(拡大傾向)	一般化可変法	MF

この表は組み合わせ的階層的分類法 (Lance and Williams (1967)) で表される階層的分類法を, 距離空間のひずみの観点から分類したものである (Nakamura and Ohsumi (1990), Ohsumi and Nakamura (1994)).

7.4 判別率 — 分類結果を評価する尺度 —

標本から混合分布を推定して分類を行ったとき, その分類結果を評価する統計量として "判別率[2]" (allocation rates) とそのバイアス, および平均自乗誤差 (Basford and McLachlan (1985)) がある. 判別率のバイアスと平均自乗誤差の平方根 (RMSE) が小さいほど良い分類結果と判断できるので, これらの統計量の大きさで初期分類法の比較ができる. 以下これらについて Basford and McLachlan (1985) を引用して簡単な説明をする.

あらかじめコンポーネント分布への所属が分かっている標本があったとする. その所属を表す二値変数を $z_i = (z_{1i}, \cdots, z_{ki}, \cdots, z_{ri})^T$ とし, z_{ki} は $\boldsymbol{x}_i$ がコンポーネント分布 G_k に所属していれば 1, 所属していなければ 0 の値をとる. z_i の推定値 $\hat{z}_i$ $(i = 1, \cdots, N)$ は, 混合分布モデルをあてはめた後, (7.2) 式を満足すれば $\hat{z}_{ki} = 1$ $(k = t)$, 満足しなければ $\hat{z}_{ki} = 0$

[2]この値を 1 から差し引くと, 判別分析における見かけ上の誤判別率に相当するので, "判別率" とした. 以後この用語を使う.

$(k \neq t)$ の値をとる. そこで第 k コンポーネント分布に対する判別率は次式で与えられる.

$$A_k = \frac{1}{N_k} \sum_{i=1}^{N} z_{ki} \delta(z_{ki}, \hat{z}_{ki}).$$

ここで, $\delta(a,b)$ はクロネッカーのデルタで, $a = b$ であれば 1, $a \neq b$ であれば 0 の値をとる. また, N_k は第 k コンポーネント分布から抽出された標本の数である. 混合分布に対する全体の正しい判別率を A とすると,

$$N_k = \sum_{i=1}^{N} z_{ki},$$

$$A = \frac{1}{N} \sum_{k=1}^{r} N_k A_k$$

となる. 標本のコンポーネント分布への所属が事前に既知である場合, つまり判別分析の場合には, $1 - A$ は見かけ上の誤判別率に相当する. これはまた, 標本が本来所属するクラスと分類後の所属クラスのクロス表から得られる単純一致率でもある.

混合分布モデルをあてはめた場合, 判別率 A の推定値は,

$$T = \frac{1}{N} \sum_{i=1}^{N} \max_{k \in \{1,\cdots,r\}} \hat{\tau}_{ki} \tag{7.3}$$

で与えられる. $T - A$ は判別率のバイアスで, これは $N \to \infty$ で 0 に確率収束する. 平均二乗誤差は,

$$MSE(T) = \mathrm{var}(T) + \mathrm{var}(A) + U^2 - 2\mathrm{cov}(T, A) \tag{7.4}$$

である. ここで, $U = \mathrm{E}(T - A)$ である.

パラメータ数に対して標本数が十分大きいと仮定すると, これは分類後のクラスターの“強さ”(個体のまとまりの良さ, クラスター内の個体の類似性の度合い) を示す尺度となる.

7.5 数値例による検証

提案分類方式を二種類のデータセットに適用する. 第一のデータセットは Iris データ, 第二は糖尿病データである. 提案した分類方式で得られた解が大域的最適解であることを確かめるため, “ランダム法” と比較し, 対数尤度の大きい方の解を “最適解” とみなす. それ以外の解は “局所解” と呼ぶことにする. ここでランダム法とは, 『標本に対してランダムに各コンポーネント分布を表すラベルの番号を付与することで初期分類を行い, 混合分布モデルを

表 7.2: Iris データの分類結果

	クラスター化法		正規混合分布モデル				
	対数尤度 *1	A	対数尤度 *1	A	T	Bias	反復回数
OC	−1.2195	1.0000	−1.2012	0.9667	0.9902	0.0235	21
CD	−1.2985	0.9067	−1.2012	0.9667	0.9902	0.0235	32
MD	−1.3424	0.7533	−1.2653	0.7267	0.9994	0.2727	24
CL	−1.3466	0.8400	−1.2012	0.9667	0.9902	0.0235	32
GA	−1.3179	0.7467	−1.2653	0.7267	0.9994	0.2727	10
WA	−1.3934	0.7600	−1.2406	0.8200	0.9795	0.1595	104
WD	−1.3199	0.8933	−1.2012	0.9667	0.9902	0.0235	32
FX	−1.3535	0.8267	−1.2012	0.9667	0.9902	0.0235	34
MF	−1.3207	0.8867	−1.2012	0.9667	0.9902	0.0235	32
KM	−1.3155	0.8993	−1.2012	0.9667	0.9902	0.0235	31
提案分類方式	—	—	−1.2012	0.9667	0.9902	0.0235	32.2(0.98) *2
ランダム法	—	—	−1.2012	0.9667	0.9902	0.0235	48.0(4.24) *2

*1 対数尤度の値は標本数で割ってある. *2 括弧内は標準偏差. ランダム法では 100 回の初期値設定のうち 2 回の最適解を得た. A: 判別率, T: 判別率の推定値, Bias$= T - A$.

あてはめる」という方式をいう. これを 100 回行い, その中の最大対数尤度を解として選ぶことにする.

また, EM 法の収束条件を次のように設定する.

$$\|L^{(t)} - L^{(t-1)}\| < \varepsilon \quad \text{or} \quad \|\boldsymbol{\Phi}_r^{(t)} - \boldsymbol{\Phi}_r^{(t-1)}\| < \delta. \tag{7.5}$$

$L^{(t)}, \boldsymbol{\Phi}_r^{(t)}$ は EM 法の第 t ステップの対数尤度の推定値とパラメータの推定値, ε と δ は十分小さい正数である. ここでは収束条件はいずれも $\varepsilon = \delta = 10^{-9}$ とした.

7.5.1 Iris データ

良く知られている Iris のデータセットは 3 つの種 (Iris setosa, Iris versicolor, Iris virginica) と, 4 つの特性値 (Sepal length, Sepal width, Petal length, Petal width) からなる (原データは Mardia, Kent and Bibby (1979) など). これらすべての個体と特性値を使用し, 任意の分散共分散行列を仮定して, 提案分類方式により分類を行った.

解析結果を表 7.2に示す. 最上段は初期構造 (初期値) として既知の群についての分類情報 (OC: Original Classification) を用いた場合の混合分布モデルによる結果, その次に 9 つのクラスター化法による初期分類の結果, 提案分類方式, そして, 最下段はランダム法による

結果である.

最大対数尤度について, ここでいう最適解が得られた分類手法は以下の 6 手法である: k-means 法 (KM), ウォード法 (WD), 可変法 (FX), 一般化可変法 (MF), 最遠隣法 (CL), 重心法 (CD). 対数尤度の値がランダム法の結果とも一致するのでこれを最適解として採用すると, 提案分類方式は最適解を得ていることがわかる. また, このとき判別率のバイアスの最小値が最適解と一致している. 一方, 以下の 3 手法はここでいう局所解となった: 群平均法 (GA) と加重平均法 (WA), メディアン法 (MD). 最近隣法 (SL) は混合分布モデルのあてはめで, 三つの種の群のうち二つを 1 コンポーネントと判断したため, ここには示さない.

表 7.2において, 通常のクラスター化法と混合分布モデルの各判別率 A (つまり, 単純一致率) の値を比較すると, 前者の場合一番大きい値は重心法 (CD) の 0.9067 に対して, 後者では 9 手法のうち 6 手法が 0.9667 である. また, いずれのクラスター化法を用いても, 混合分布モデルをあてはめると判別率 A は増加している.

次に, 提案分類方式とランダム法のふるまいを EM 法の平均反復回数と判別率をみる. まず平均反復回数は, 提案手続きは 32.2 回, ランダム法では 48.0 回となり, 提案分類方式の方が良い. 最適解として選ばれた回数は, 前者が 9 手法中 6 手法, 後者が 100 回中わずかの 2 回であった.

7.5.2 糖尿病データ

Reaven and Miller (1979) は, 肥満を伴わない 145 名の対象者を糖尿病に関する三種類の検査項目を用いて, 3 群への判別を試みている (原データの出典は Andrews and Herzberg (1985)). 三つの群は一つの正常群 (76 名) と二つの異常群からなり, 異常群の一方は臨床的症状を伴わない群 (36 名) と, 他方は臨床的症状を伴う群 (33 名) である. このデータは五つの検査項目からなり, 原論文ではそのうち 3 変数を使って解析している. さらに医学的知見に基づく分類 (CC: clinical classification) が併記されている. ここでは原論文と同じ 3 つの検査項目 (以降, 変数と呼ぶ) を用いて, 任意の分散共分散行列と仮定して解析を行った. 解析結果は表 7.3に示す.

表 7.3を見ると, 対数尤度が最大になったのは GA, CD, MD の 3 手法である. ここでもランダム法と同じ結果が得られたので, これを最適解とすると, 提案分類方式は有効に働

表 7.3: 糖尿病データの分類結果

	クラスター化法		正規混合分布モデル				
	対数尤度 *1	A	対数尤度 *1	A	T	Bias	反復回数
CC	−17.6607	1.0000	−17.5053	0.8483	0.9528	0.1046	28
CD	−17.7462	0.7310	−17.5053	0.8483	0.9528	0.1046	70
MD	−17.6580	0.8069	−17.5053	0.8483	0.9528	0.1046	45
CL	−17.8998	0.6414	−17.7405	0.6897	0.9481	0.2584	42
GA	−17.7462	0.7310	−17.5053	0.8483	0.9528	0.1046	70
WA	−17.8998	0.6414	−17.7405	0.6897	0.9481	0.2584	42
WD	−17.5437	0.8552	−17.5120	0.8345	0.9545	0.1201	65
FX	−17.5437	0.8552	−17.5120	0.8345	0.9545	0.1201	65
MF	−17.5437	0.8552	−17.5120	0.8345	0.9545	0.1201	65
KM	−17.8659	0.6414	−17.7405	0.6897	0.9481	0.2584	39
提案分類方式	—	—	−17.5053	0.8483	0.9528	0.1046	61.7(14.4)*2
ランダム法	—	—	−17.5053	0.8483	0.9529	0.1046	34.1(20.8)*2

*1 対数尤度の値は標本数で割ってある. *2 括弧内は標準偏差. ランダム法では 100 回の初期値設定のうち 38 回の最適解を得た. A: 判別率, T: 判別率の推定値, Bias= $T - A$.

いていることがわかる. それぞれの試行について判別率のバイアスの値を比較すると, このデータの場合でも最小の手法が提案分類方式の最適解になっている.

初期分類に用いたクラスター化法の分類結果と混合分布モデルによる分類法の結果とを表 7.3の判別率 A で比較する. 提案分類方式で最適解を与えたクラスター化法の GA, CD, MD に比べて, 混合分布モデルによる分類結果の方が A の値が大きくなっている. 一方, WD, FX, MF を初期分類とする混合分布モデルによる分類結果は, 最適解とはならないで, これら三つのクラスター化法の方が A の値が大きくなっている. このデータについてはクラスター化法の方が単純一致率の意味では分類結果は良い.

7.6 数値実験

7.6.1 数値実験の目的

ここでは数値実験を通して以下のことを考察する.

(1) 提案分類方式の有効性の検証

(2) クラスター化法と混合分布モデルの分類結果の比較

(3) 提案分類方式で使用するクラスター化法の比較.

これらを検証するため, あらかじめ構造の分かっている多変量データを乱数を用いて生成して, このデータセットの分類を行った. 前と同様に誤分類率, 平均収束回数, 平均一致率, 対数尤度, 判別率, バイアス, 平均自乗誤差の平方根 (RMSE), 最適解に選ばれた回数を用いて分類結果を評価した. 提案分類方式の初期分類のクラスター化法は, 最近隣法, 最遠隣法, 群平均法, 重心法, ウォード法, 可変法, 修正可変法, k-means 法 (それぞれ略記では SL, CL, GA, CD, WD, FX, MF, KM), それにランダム法を用いた. 階層的分類法の最初の個体間距離は, 平方ユークリッド距離を用いて, 各変数 (変量) の基準化は行なっていない.

7.6.2 サンプリングの方法

数値実験やシミュレーションにおいて, 混合分布のデータセットを作成するとき, 二種類のデータ抽出の方法 (sampling scheme) がある (Ganesalingam and McLachlan (1980) など). 一つは混合サンプリング (mixture sampling) であり, もう一つは分離サンプリング (separate sampling) である. たとえば, 母集団である混合分布のコンポーネント数が 2 つの場合 (Π_1, Π_2), 混合サンプリングとは混合比率 π_1, π_2 に従って, Π_1 と Π_2 からランダムに標本を抽出する方法である. n_k は Π_k の標本数で, これはパラメータを N と π_k とする 2 項分布にしたがう. 分離サンプリングとは標本数 N と混合比率 π_k によって, 各コンポーネント分布の標本数を固定して $(n_k = N \times \pi_k;\ k = 1, 2)$, コンポーネント分布ごとに標本を抽出する方法である.

ここでは前者の混合サンプリングを用いる. 例として, 二つのコンポーネント分布の混合分布の作り方は次のように行う. 標本数を N, 混合比率 $\pi_1 = q$ とすると, まず $[0, 1)$ 上の一様乱数 u を 1 つ生成し, $u \in [0, q)$ であれば第 1 のコンポーネント分布から, $u \in [q, 1)$ であれば第 2 のコンポーネント分布からデータを 1 つ生成する. これを必要とするデータの個数である N 回繰り返す. この方法でデータセットを生成した場合, 第 k 番目の分布のデータの個数は正確に $\pi_k N$ 個生成されない.

一方, 分離サンプリングは次のようにデータセットを作成する. コンポーネント分布を二つとすると, それに対応する混合比率を π_1, π_2, 標本数を N とする. そのとき, 第 1 コンポーネント分布の個体数は $\pi_1 \times N$ となるので, $\pi_1 \times N$ 個の標本をこの分布から抽出する. 同様に第 2 コンポーネント分布から $\pi_2 \times N$ 個の標本を抽出する. これが分離サンプリングであ

表 7.4: シミュレーションのパラメータ設定

設定	μ_k	Σ_k	π	その他
1	$(0,0)^T$ $(0,d)^T$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	$d = 2, 2.5, 3.$ 分散共分散行列は共通
2	$(0,0)^T$ $(d,4)^T$	$\begin{pmatrix} 16 & 0 \\ 0 & 0.25 \end{pmatrix}$	$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	$d = 0, 4, 8.$ 分散共分散行列は共通
3	$(0,0)^T$ $(3,d)^T$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\begin{pmatrix} 16 & 0 \\ 0 & 0.25 \end{pmatrix}$	$\begin{pmatrix} p \\ 1-p \end{pmatrix}$	$d = 2, 3, 4,$ $p = 0.7, 0.3$

る.

ここで, 一様乱数の生成は M 系列乱数 (伏見 (1989) など) を用い, 正規乱数の生成は Box-Müller 法 (伏見 (1989) など) を用いた.

7.6.3 パラメータ設定

ここでは表 7.4に示す三種類のパラメータ設定をした. これらのパラメータにしたがう二つの二次元の正規混合分布から乱数を発生させ, これを分類対象のデータとする. それぞれの標本数は, $N = 30, 50, 100, 200$ ($N = 30$ は実験のパラメータ設定 1, 2 のみ), 実験の回数は $S = 500$ 回とする.

パラメータ設定 1 の特徴は共通の分散共分散行列をもつ球形の分布であり, パラメータ設定 2 は第 1 変量と第 2 変量の分散が大きく異なる細長い分布である. パラメータ設定 3 は設定 1 と 2 を合わせたものである. これはパラメータ推定の制約条件として, パラメータ設定 1, 2 は共通の分散共分散行列を, パラメータ設定 3 では任意の分散共分散行列としてパラメータ推定を行う. EM 法の収束条件は $\varepsilon = \delta = 10^{-6}$ とした.

7.6.4 実験結果と考察

限られたパラメータ設定での数値実験であるが, そこから得られた結果や考察を提案分類方式の有効性, 混合分布モデルによる分類法とクラスター化法の比較, クラスター化法の比較の順に述べる. なお設定 3 では最近隣法 (SL) に対しては鎖状現象[3]が起きたため, 初期

[3]Lance and Williams (1967) を参照.

分類法から除外した.

提案分類方式の有効性

まず, 提案分類方式の有効性について述べる. 図 7.1は設定 2 で $d = 4$, $N = 100$ とした場合に (a) k-means 法, (b) ランダム法, (c) 提案分類方式から得られた対数尤度のヒストグラムである. なお, この設定での真の平均対数尤度に標本数を掛けた値は -422.41 である. したがって, 図の右側のピークは最適解に, 左側は局所解である. 最適解に収束した割合はそれぞれ (a) k-means 法 =82.0%, (b) ランダム =70.0%, (c) 提案分類方式 =97.8% である. 他のクラスター化法では, SL=42.8%, CL=69.2%, GA=49.0%, CD=45.2%, WD=78.0%, FX=68.4%, MF=80.0% であった. 以上のことから提案分類方式は仮定したデータセットの構造をよくとらえると考えられる.

さらに図 7.3(設定 2, $d = 4$) と図 7.4(設定 3, $d = 4$, $p = 0.7$) から提案分類方式における各指標のふるまいを見る. 図 7.3(a) で判別率の推定値 T を見ると, 提案分類方式 (Proposed Proc.) が最も大きいが, 図 7.4(a) では一番小さい. 設定 1 ともあわせて比較すると, 次のような傾向が観察された.

2 つのコンポーネント分布がよく分かれていて, 群の構造が似ている設定 1, 2 では, 推定された解がデータの構造をとらえたときの判別率は 1 または 1 に近い数値となった. 次に, 提案分類方式は実験の各回ごとに最大の対数尤度を選択しているので, バイアス (図 7.3(b), 図 7.4(b)) と RMSE(図 7.3(c), 図 7.4(c)) のグラフで値が常に小さくなり, 仮定したデータ構造をよくとらえると考えられる.

混合分布モデルによる分類法とクラスター化法の比較

表 7.5 ($N = 200$ のみ表示) の通常のクラスター化法 (CA) と混合分布モデルによる分類法 (MA) での判別率 (A) からわかることについて述べる. パラメータ設定 1 における k-means 法を除いて, 混合分布モデルによる分類での判別率は, クラスター化法の判別率より必ず増加している. これを見かけ上の誤判別率として裏返して見れば, 混合分布モデルをあてはめることで従来のクラスター化法より所与のデータセットの構造をよくとらえていることがわかる. パラメータ設定 1 では k-means 法が有利になり, Ganesalingam (1989) と同じ

表 7.5: クラスター化法と混合分布モデルをあてはめた後の判別率 ($N = 200$ のみ)

設定		SL	CL	GA	CD	WD	FX	MF	KM	Random
設定 1	CA	52.87	75.88	62.23	56.73	79.66	77.37	79.53	83.87	52.91
($d = 2$)	MA	61.17	78.57	66.64	64.36	80.55	79.95	80.56	80.87	68.47
設定 1	CA	52.98	81.84	72.31	65.56	84.89	84.03	85.11	89.23	52.92
($d = 2.5$)	MA	63.67	87.66	77.17	71.48	88.13	88.15	88.16	88.21	73.00
設定 1	CA	53.02	86.60	83.51	77.80	90.00	88.26	90.00	93.36	52.91
($d = 3$)	MA	65.68	92.51	86.46	82.61	92.91	92.91	92.91	92.91	77.36
設定 2	CA	61.06	55.56	54.63	54.15	57.49	56.65	58.15	54.86	52.99
($d = 0$)	MA	75.43	84.40	72.50	71.56	86.72	86.15	86.80	85.88	86.05
設定 2	CA	62.59	66.48	61.93	59.44	71.40	68.36	73.38	73.66	53.00
($d = 4$)	MA	74.92	89.72	78.88	76.05	94.01	92.45	94.62	97.02	85.50
設定 2	CA	67.23	78.47	78.86	75.86	84.93	80.48	86.10	87.58	53.01
($d = 8$)	MA	77.54	95.98	92.31	88.61	99.91	98.17	99.86	100.0	86.12
設定 3	CA	—	63.02	62.28	61.86	66.85	65.10	67.64	65.40	52.83
($d = 2, p = 0.7$)	MA	—	94.72	91.06	90.49	94.90	94.86	94.90	94.90	94.89
設定 3	CA	—	82.12	79.62	79.21	85.27	84.29	86.06	85.66	52.82
($d = 2, p = 0.3$)	MA	—	93.35	91.69	91.43	93.68	93.68	93.71	93.68	93.76
設定 3	CA	—	62.93	63.15	61.76	68.62	66.28	68.15	68.84	52.82
($d = 3, p = 0.7$)	MA	—	99.80	96.78	96.56	99.80	99.80	99.80	99.80	99.80
設定 3	CA	—	83.13	81.08	80.36	87.94	84.96	97.69	89.56	52.74
($d = 3, p = 0.3$)	MA	—	99.52	98.13	97.60	99.72	99.68	99.72	99.72	99.72
設定 3	CA	—	62.99	63.00	62.78	67.62	64.87	68.17	66.91	52.82
($d = 4, p = 0.7$)	MA	—	98.69	95.54	94.11	98.74	98.74	98.74	98.74	98.74
設定 3	CA	—	82.85	79.97	79.74	86.92	85.04	88.93	87.19	52.83
($d = 4, p = 0.3$)	MA	—	98.24	96.60	96.45	98.35	98.35	98.35	98.35	98.35
	CA	58.3	73.5	70.2	67.9	77.6	75.47	79.1	78.8	52.9
全平均		(6.19)	(10.6)	(9.99)	(9.89)	(10.7)	(10.6)	(11.8)	(12.3)	(.0859)
(標準偏差)	MA	69.7	92.7	87.0	85.1	94.0	93.6	94.0	94.2	88.5
		(7.03)	(6.62)	(10.6)	(11.5)	(6.13)	(6.18)	(6.11)	(6.23)	(10.9)

CA: クラスター化法, MA: 混合分布モデルによる分類. SL: 最近隣法, CL: 最遠隣法, GA: 群平均法, MD: メディアン法, WD: ウォード法, FX: 可変法 ($\beta = -0.25$), MF: 一般化可変法, KM:k-means 法, Random: ランダム初期分割.

結果が得られた. しかし, 設定 1, 2 のランダム法の混合分布モデルをあてはめた後の値と他のクラスター化法の結果の値を比べると, その値は多少低い.

多数の文献を調べたが, 混合分布モデルによる分類法の有効性を詳細に言及する論文は見られない. たとえばBayne, Beauchamp, Begovich and Kane (1980) では 14 種類のクラスター化法の比較の中で, NORMIX は余りよい評価が得られていない (これは NORMIX が最適解を推定していないということが考えられる). しかし, ここで示したとおりここで用いたデータセットについては混合分布モデルによる分類法はクラスター化法に比べて, 判別率の意味でよい結果を与えている.

クラスター化法の比較

初期分類法としてのクラスター化法がどのように働いているのか調べる.

図 7.2は今まで述べた数値実験全体 (設定 1 ～ 3) の EM 法の平均反復回数のボックスプロットである. ここで, 反復回数がより少ない手法は, 階層的分類法のうち拡大傾向を示す手法 (CL, WD, FX, MF) と k-means 法である.

一方, 表 7.6 は, 提案分類方式で最適解に選ばれた平均回数を示したものである. この表から初期分類法として好ましい手法は, KM, MF, WD, FX の 4 手法で, これは KM を除くと拡大傾向を示すの階層的分類法である. なお, ここではランダム法は良い結果を与えていない. 全体の傾向として標本数が大きくなると最適解として選ばれる回数は増えるが, 図 7.5 (設定 1, $d = 2.5$) にみるように, 設定 1, 2 のランダム法と縮小傾向を示す階層的分類法は, その逆の傾向がある.

表 7.7は設定 1, 2 において混合分布モデルをあてはめたとき, 2 つのコンポーネント分布であったものを 1 つのコンポーネントのみであると推定した割合の平均を示す. SL はとび抜けて多く, ランダム法は標本数によらず平均的にこの現象が起きている. GA, CD は標本数が多くなるとこの傾向が強くなる. この表から縮小傾向を示す手法は余り良い初期構造を与えていないと判断できる. また, KM と階層的分類法の拡大傾向を示す手法がこの値が極めて小さい.

ここで行なった数値実験全体を通して次の知見が得られた. ここで用いたクラスター化

表 7.6: 提案手続きで最適解に選ばれた回数

設定	条件	SL	CL	GA	CD	WD	FX	MF	KM	Random
設定 1	$d = 2$	190.0	308.3	251.0	228.5	335.5	326.0	335.3	335.3	226.5
	$d = 2.5$	168.0	390.0	325.8	275.5	400.5	400.0	403.5	405.3	248.3
	$d = 3$	158.0	440.0	402.5	367.5	447.3	445.3	447.3	448.8	263.3
設定 2	$d = 0$	235.8	266.3	171.3	171.3	285.8	269.8	289.0	287.3	369.5
	$d = 4$	229.0	323.3	244.0	228.5	370.8	335.0	382.5	390.0	351.8
	$d = 8$	261.0	421.8	388.0	362.5	465.0	434.5	462.3	483.0	344.8
設定 3	$d = 2, p = 0.3$	—	443.3	386.7	371.7	454.7	450.3	457.0	460.3	483.0
	$d = 2, p = 0.7$	—	406.3	369.0	366.3	412.0	415.0	429.3	429.0	474.7
	$d = 3, p = 0.3$	—	480.7	437.0	428.7	488.3	485.3	485.7	489.7	496.3
	$d = 3, p = 0.7$	—	441.0	414.0	406.0	463.0	449.3	485.3	479.0	483.0
	$d = 4, p = 0.3$	—	465.7	415.7	397.3	474.0	470.0	473.7	481.7	494.0
	$d = 4, p = 0.7$	—	426.3	392.3	388.0	438.7	435.0	458.3	458.7	483.0
	全平均	207.0	395.0	342.2	324.0	414.5	403.7	420.2	423.6	380.0
	(標準偏差)	(43.1)	(85.3)	(92.2)	(90.9)	(78.8)	(84.2)	(76.1)	(77.7)	(104.7)

各設定の条件ごとに標本数 $N = 30(50)\sim200$ の平均を示す. SL: 最近隣法, CL: 最遠隣法, GA: 群平均法, MD: メディアン法, WD: ウォード法, FX: 可変法 ($\beta = -0.25$), MF: 一般化可変法, KM:k-means 法, Random: ランダム初期分割.

表 7.7: 2 つのコンポーネントを 1 つのコンポーネントに推定した割合

	SL	CL	GA	CD	WD	FX	MF	KM	Random
設定 1, 2 の平均 (%)	21.81	0.456	1.324	2.054	0.134	0.125	0.111	0.000	6.278
(標準偏差)	(10.85)	(0.443)	(1.660)	(2.314)	(0.168)	(0.166)	(0.202)	(0.000)	(0.845)

法のふるまいを見ると, これらの間にある序列が見られる. たとえば, 表 7.5〜 7.7の平均値を大きさの順に並べたときや, 図 7.3(b), (c), 図 7.4(b), (c) においてクラスター化法の上から下への並びなどに見られる現象である. ここではほぼ SL $\prec$ CD $\prec$ GA $\prec$ CL $\prec$ FX $\prec$ WD $\prec$ MF $\prec$ KM の序列関係がある. 階層的分類法に限れば, この序列関係は表 7.1において縮小から拡大傾向に向かう手法の並びにほぼ対応していると考えられる.

数値実験のまとめ

この節で行なった数値実験をまとめると次のような知見が得られた.

(1) 階層的分類法の拡大傾向を示す手法と k-means 法とが, 良い初期分類を与えるようだ.

(2) 通常のクラスター化法を初期分類とする混合分布モデルによる分類法, とくに提案分類方式は, 通常のクラスター化法のみの場合に比べて判別率の意味で良い分類結果を与える.

(3) ランダム法はコンポーネント分布の構造が類似のときはあまり有効に働かない場合がある (設定 1, 2, 表 7.5から).

(4) 階層的分類法には距離空間のひずみの特性に関連した序列関係がみられる. これは上記 (1) に関係すると考えられる.

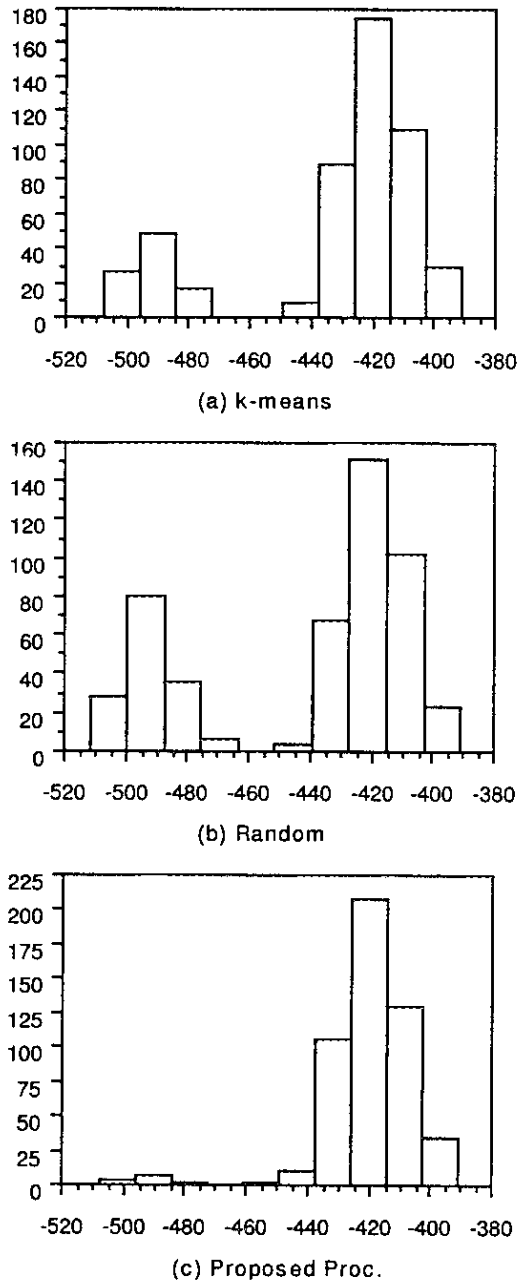

図 7.1: 設定 $2(d = 4, N = 100)$ の対数尤度のヒストグラム
(a) k-means 法, (b) ランダム法, (c) 提案分類方式.

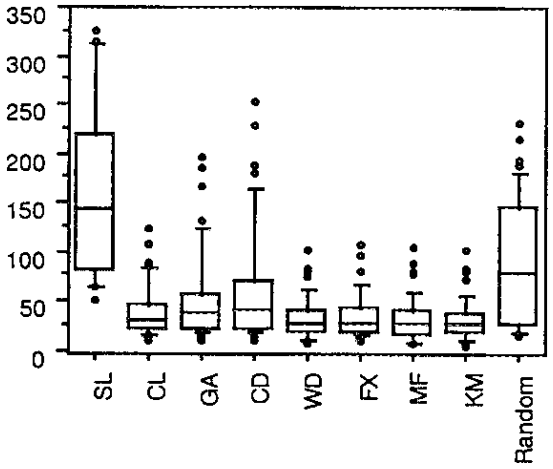

図 7.2: EM 法の平均反復回数のボックスプロット

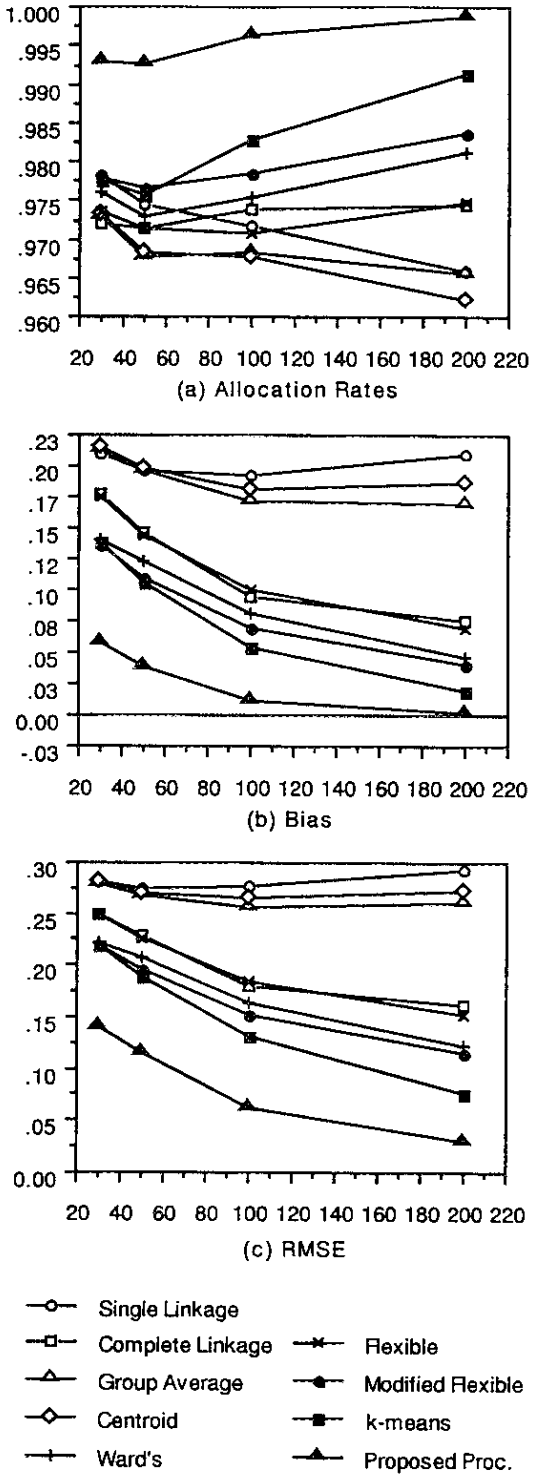

図 7.3: 設定 2 の分類結果
(a) 判別率の推定値, (b) バイアス, (c) 平均自乗誤差の平方根.

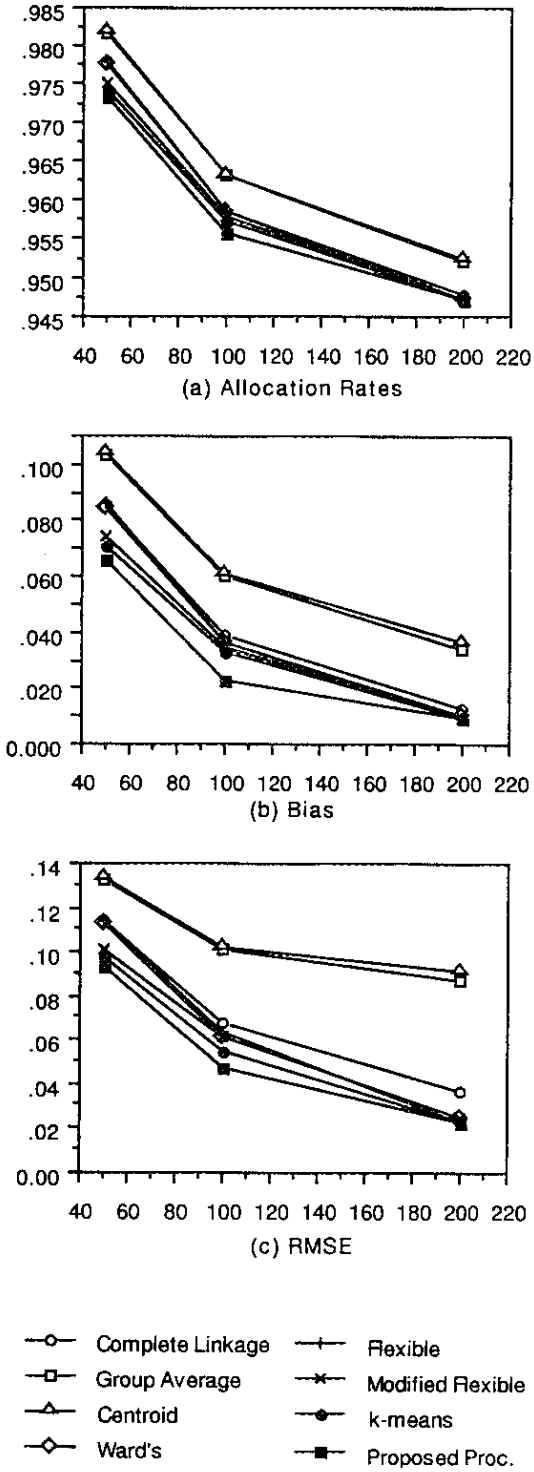

図 7.4: 設定 3 の分類結果
(a) 判別率の推定値, (b) バイアス, (c) 平均自乗誤差の平方根.

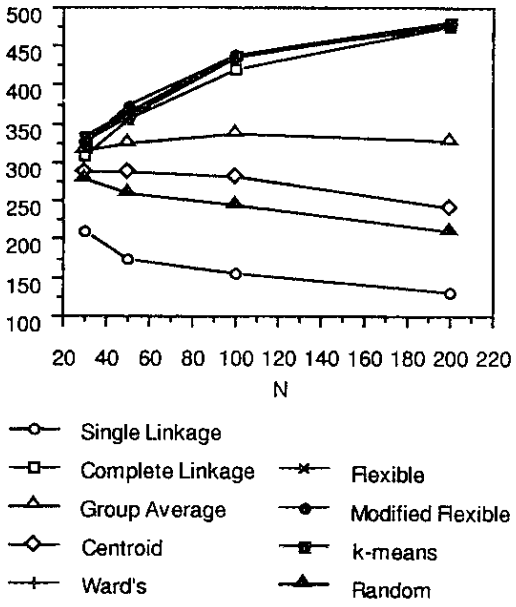

図 7.5: 最適解に選ばれた平均回数

7.7 今後の課題

混合分布モデルによる分類法および提案分類方式の問題点および今後の課題を要約する.

(1) 正規混合モデルに限るが, 推定されたコンポーネント分布の正規性が保たれるかどうかの問題である (McLachlan and Basford (1988), 2.5 節). 実際のデータセットではこの検討が必要と思われる.

(2) 混合分布モデルのコンポーネント数を多くすれば, データに対するあてはまりは良くなるが, 分類 (グループを作る) という観点からは解釈不能になる場合がある.

(3) 推定した複数のコンポーネント分布が一つの分布を取り合う場合がある (推定したパラメータがほぼ同じ値で推定されること).

とくに (2), (3) は, コンポーネント数の推定の問題も同時に考える必要がある.

7.8 まとめ

実際のデータセットの解析と一連の数値実験から, 提案分類方式は所与のデータセットの構造をとらえており, 実用的観点からは有力な方法と考えられる. 一方, ランダム法は数多くの実験が必要であるという点でコストが高い. 少なくとも実験結果からは, データ構造が単純なときやコンポーネント数が多くなったときは, ランダム法に比べて提案分類方式が良い結果を与えた.

一方, 個体間の距離でデータの構造を考えると, ここで使った通常のクラスター化法による分類は平方ユークリッド距離を, 混合分布モデルによる分類法はマハラノビス距離を使っていることになる. とくに数値実験では混合分布からデータを生成しているので, この観点からも混合分布モデルによる分類法が直観的に優位であることが想像できる.

第 8 章

混合分布モデルのコンポーネント数の推定

この章は, 情報量規準を用いた多変量混合分布モデルのコンポーネント数の推定手続きを提案し (小西・中村 (1994a, b)), その有効性を検討することが目的である. 情報量規準の構成に際してはブートストラップ法を適用し, さらに, 線型近似によるブートストラップバイアス推定の変動減少法 (北川・石黒・坂元 (1993), 小西 (1993a)) を導入し, その有用性について検討する. なお, 混合分布モデルのパラメータ推定には, すでに前章で提案した分類方式に基づいて行う. 数値実験と実際データの解析を通して, 提案するコンポーネント数推定の手続きの有効性を検証する.

8.1 混合分布モデルのコンポーネント数の推定

観測データへの混合分布モデルのあてはめにおいて, コンポーネント数をいかに推定するかという問題は重要ではあるが, 実際に適用する上でいくつかの問題点がある. この問題に対する一つのアプローチは, 仮説検定の枠組みの中で尤度比検定統計量を用いる方法であるが, 検定統計量の分布は, 通常の漸近 χ^2 近似が成り立たないことが知られている (McLachlan (1987), Redner and Walker (1984) など). この問題に対する様々な工夫がされていて, 尤度比検定統計量の修正 (Wolfe (1971)), ベイズアプローチによる方法 (Aitkin and Rubin (1985)), ブートストラップ法による尤度比検定統計量の漸近分布の構成 (McLachlan(1987), Thode, Finch and Mendell (1988)) などがある. また, Celeux (1986) の Orlov の方法 (これは nested tests を修正したもので Soromenho (1994) より引用), SEM 法 (EM 法の stochastic 版) による方法 (Celeux and Diebolt (1985)) などがある.

一方, ブートストラップ法 (Efron (1979)) の理論とその有用性が浸透するにしたがって,

コンポーネント数の推定問題に対してもこの手法を用いた論文が見られるようになってきた．たとえばWong(1985)は，k最近隣クラスタリング(kth nearest neighbour clustering, Wong and Lane (1983))を用いた複峰性(multimodality)の検定のために，その有意水準の構成に修正ブートストラップ法を用いている．また，McLachlan(1987), Thode, Finch and Mendell (1988)はブートストラップ法により検定統計量の漸近帰無分布を構成している．Windham and Cutler (1992)はフィッシャー情報行列の比を用いるコンポーネント数の推定方法を示し，ブートストラップ法によりその方法の有効性を示す手続きを提案している．

また，この問題に情報量規準を用いることも考えられる．AIC(Akaike's Information Criterion; Akaike (1973))をこの問題に最初に用いたのはSclove (1983), Bozdgan and Sclove (1984)である．しかし，コンポーネント数を推定する問題に対して，AICは経験的にその数を多めに推定することが知られている．この点に対してBozdgan (1992)は情報量規準ICOMPと修正AICを提案している．この修正AICは，対数尤度のバイアス補正が十分でない量を修正するというものであるが，その理論的根拠は希薄である．このICOMPはvan Emden(1971)のentropic covariance complexity indexの考えを導入したものである．また，Rissanen (1978, 1984)はMDL (Minimum Description Length)という符号理論に基づいた規準を提案し，Bozdgan (1992)は自分の提案した情報量規準とこれとの比較を行っている．

ここではブートストラップ法により構成される情報量規準を用いて，混合分布モデルのコンポーネント数の推定を行う手続きを提案する．この情報量規準はEICと名付けられ，その有効性が検討されつつある（石黒 他(1992), 北川 他(1992), 北川 他(1993), 坂元 他(1992)）．

8.2 コンポーネント数の推定手続き

情報量規準を用いたコンポーネント数の推定手続きを提案する．これはコンポーネント数が$r = 1, 2, \cdots, g$のg種類のモデルを考え，得られる複数の情報量規準の値を比較して最小のモデルを選択する方法である．ここで情報量規準はブートストラップ法で構成するとい

う方式を適用する（付録Cを参照）.

すでに述べたように, 混合分布モデルのパラメータをEM法で推定するとき, 大域的最適解を見つけるという観点からその初期値の設定方法は大変重要である. ここでは, 前章で提案したように, 初期分類のクラスター化法として, 表7.1に示した階層的分類法の各手法と分割型分類法の k-means 法を用いる. このとき, 次の手順によりコンポーネント数の推定を行う.

コンポーネント数の推定手続き

ステップ **1.** 観測値

$$\boldsymbol{X}_N = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\} \overset{iid}{\sim} G(x); g(x)$$

に対して r 個のコンポーネント分布からなる混合分布モデル

$$f(\boldsymbol{x}|\boldsymbol{\Phi}_r) = \sum_{k=1}^{r} \pi_k f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)$$

を想定する. ここで, $G(x)(g(x))$ は未知の混合分布の確率分布関数（密度関数）, π_k は混合比率, $f_k(\cdot|\cdot)$ はコンポーネント分布, $\boldsymbol{\theta}_k$ は $f_k(\cdot|\cdot)$ のパラメータである. また, $\boldsymbol{\Phi}_r = \{\pi_1, \cdots, \pi_{r-1}, \boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_r^T\}^T$ である.

ステップ **2.** クラスター数を $r(r = 1, \cdots, g)$ として $\boldsymbol{X}_N$ に対してクラスター化法で初期分類を行う. その各クラスターへの所属を示す変数 z_{ki} を次のように定義する.

$$z_{ki} = \begin{cases} 1 & x_i \in C_k \\ 0 & \text{others} \end{cases}$$

ここで, C_k は第 k クラスターである. さらに

$$\boldsymbol{Z}_N = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_i, \cdots, \boldsymbol{z}_N\}, \quad \boldsymbol{z}_i = \{z_{1i}, \cdots, z_{ki}, \cdots, z_{ri}\}$$

とすると, $(\boldsymbol{X}_N, \boldsymbol{Z}_N)$ は, 初期分類の状態 (初期構造) を表す "組" である.

ステップ **3.** ステップ2で推定された初期構造から統計量 (各コンポーネント分布の平均ベクトルと分散共分散行列) を計算し, これをEM法の初期値とする. $\boldsymbol{X}_N$ にコンポーネント数を r $(r = 1, \cdots, g)$ として混合分布モデルをあ

てはめる. EM 法により推定されたパラメータを $\hat{\boldsymbol{\Phi}}_r$ とすると, その時の対数尤度関数は

$$(8.1)\quad L(\hat{\boldsymbol{\Phi}}_r|\boldsymbol{X}_N) = \frac{1}{N}\sum_{i=1}^{N}\log f(\boldsymbol{x}_i|\hat{\boldsymbol{\Phi}}_r) = \frac{1}{N}\sum_{i=1}^{N}\log\{\sum_{k=1}^{r}\hat{\pi}_k f_k(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_k)\}$$

である. z_{ki} の推定値, すなわち観測値 $\boldsymbol{x}_i$ が第 k コンポーネント分布に所属する確率 (事後確率) は,

$$(8.2)\quad \hat{\tau}_{ki} = \frac{\hat{\pi}_k f_k(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_k)}{f(\boldsymbol{x}_i|\hat{\boldsymbol{\Phi}}_r)} \quad (= \hat{z}_{ki})$$

である.

ステップ **4.** $\boldsymbol{X}_N$ から標本の大きさ N の (第 b 番目の) ブートストラップ標本 $\boldsymbol{X}_N^{*(b)} = \{\boldsymbol{x}_1^{*(b)}, \cdots, \boldsymbol{x}_N^{*(b)}\}$ を作る $(b = 1, \cdots, B)$ [1].

ステップ **5.** $\boldsymbol{X}_N^{*(b)}$ に対してコンポーネント数を $r(r = 1, \cdots, g)$, EM 法の初期値として $\hat{\boldsymbol{\Phi}}_r$ を用いて混合分布の推定を行う. 推定されたパラメータは $\hat{\boldsymbol{\Phi}}_r^{*(b)}$, その対数尤度関数は $L(\hat{\boldsymbol{\Phi}}_r^{*(b)}|\boldsymbol{X}_N^{*(b)})$ である.

ステップ **6.** このときのバイアスは

$$\mathrm{bias}_r^{(b)} = L(\hat{\boldsymbol{\Phi}}_r^{*(b)}|\boldsymbol{X}_N^{*(b)}) - L(\hat{\boldsymbol{\Phi}}_r^{*(b)}|\boldsymbol{X}_N),$$

バイアスの変動減少法を用いた場合のバイアスは,

$$\mathrm{bias}_{\mathrm{M}r}^{(b)} = L(\hat{\boldsymbol{\Phi}}_r^{*(b)}|\boldsymbol{X}_N^{*(b)}) - L(\hat{\boldsymbol{\Phi}}_r^{*(b)}|\boldsymbol{X}_N) + L(\hat{\boldsymbol{\Phi}}_r|\boldsymbol{X}_N) - L(\hat{\boldsymbol{\Phi}}_r|\boldsymbol{X}_N^{*(b)})$$

である. ここで, たとえば $L(\hat{\boldsymbol{\Phi}}_r^{*(b)}|\boldsymbol{X}_N)$ はブートストラップ標本 $\boldsymbol{X}_N^*$ により推定されたパラメータ $\hat{\boldsymbol{\Phi}}_r^{*(b)}$ を用いて $\boldsymbol{X}_N$ で評価された対数尤度である.

ステップ **7.** ステップ **4.** ~ **6.** をコンポーネント数 $r = 1, 2, \cdots, g$ に対して B 回行う.

---

[1] ブートストラップ標本の作り方　ブートストラップ標本を作成するとき, 有効なシミュレーションを行なうためいくつかの方法が考えられているが (久保川, 江口, 竹村, 小西 (1993, 4.6 節) など), ここでは次の方法で行なう. つまり, 通常の経験分布関数からリサンプリングを行うとき, 各点に $1/N$ の確率を与えて抽出を行う. これはもっとも基本的な方法であり, "一様リサンプリング" と呼ばれる.

ステップ 8. B 回のブートストラップの手続きを行った後の, バイアスの推定値は次のようになる.

$$\widehat{\mathrm{bias}}_r = \frac{1}{B}\sum_{b=1}^{B} \mathrm{bias}_r^{(b)},$$

$$\widehat{\mathrm{bias}}_{\mathrm{M}r} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{bias}_{\mathrm{M}r}^{(b)}.$$

したがってここで用いる情報量規準の値は

$$\mathrm{ICBoot}_r = -2 \times L(\hat{\boldsymbol{\Phi}}_r|\boldsymbol{X}_N) + 2N \times \widehat{\mathrm{bias}}_r,$$

$$\mathrm{ICBoot}_{\mathrm{M}r} = -2 \times L(\hat{\boldsymbol{\Phi}}_r|\boldsymbol{X}_N) + 2N \times \widehat{\mathrm{bias}}_{\mathrm{M}r}$$

となる. ただし, AIC と対応がつくように $-2N$ 倍しておく.

ステップ 9. c 種類のクラスター化法 ($\ell = 1, 2, \cdots, c$) に対してステップ 2. ～ 8. を行う. $g \times c$ 個の ICBoot_r と $\mathrm{ICBoot}_{\mathrm{M}r}$ の中から最小値を選び, 対応する r が推定したコンポーネント数 $\hat{r}$ である.

8.3 ブートストラップ標本の初期値設定方法の考察

前節のコンポーネント数の推定の手続きのステップ 5 では, $\boldsymbol{X}_N^{*(b)}$ に混合分布モデルをあてはめるとき, EM 法の初期値として $\hat{\boldsymbol{\Phi}}_r$ を用いた. いま想定しているデータセットは大標本ではないので, ブートストラップ標本に混合分布モデルをあてはめる際のいくつかの EM 法の初期値設定に方法ついて考察する必要がある[2]. そこで, 次の 3 つの方法を提案する.

方法 1 観測データに混合分布モデルをあてはめる場合と同じようにクラスター化法を用いる.

方法 2 観測データに混合分布モデルをあてはめて推定されたパラメータ $\hat{\boldsymbol{\Phi}}_r$ を用いる.

方法 3 ブートストラップ標本 $\boldsymbol{X}_N^*$ を作るとき, $\boldsymbol{X}_N$ から推定されたコンポーネント分布への所属確率 $\widehat{\boldsymbol{Z}}$ からもブートストラップ標本 $\widehat{\boldsymbol{Z}}^*$ を作る. つまり, ブートストラップ標

[2]すでに述べたように, EM 法でパラメータ推定を行なうとき, 大標本の場合はほとんど初期値に依存しないで最適解に収束するが (Redner and Walker (1984)), 小～中標本の場合は初期値設定が重要である.

本を作るとき, $\boldsymbol{x}_i \to \boldsymbol{x}_j^*$, $z_i \to z_j^*$ のように, 添字 i を対応させる (i は標本の添え字, j はブートストラップ標本の添え字である).

方法 1 は, (1) ブートストラップの反復ごとに $\boldsymbol{X}_N^*$ に初期分類を行うので, そのための時間を要する, (2) $\boldsymbol{X}_N^*$ によっては, その初期分類の結果が $\boldsymbol{X}_N$ の分類結果と大きく異なることがある (クラスター数 (コンポーネント数) を多くするとこの現象は悪化する), (3) ブートストラップ標本 $\boldsymbol{X}_N^*$ は同じ標本が複数回とられる可能性があるため, 階層的分類法は $\boldsymbol{X}_N^*$ の構造を敏感にとらえてしまう, (4) その結果推定される混合分布も局所解に収束している可能性が高いこと, などの欠点がある. 次に, いくつかの論文のシミュレーションにおいて混合分布の推定するとき, その初期値としてデータ生成時の真のパラメータ値をそのまま使うことがあるが, 方法 2 はこれに相当する. つまり, 真の分布と観測データから構成される経験分布関数との関係は, 経験分布関数とブートストラップ標本から構成される経験分布関数との関係に相当し, これに準じた初期値設定を行うことになる. これは合理的な方法と考えられる. 方法 3 は観測データから得られたデータ構造の情報を用いるという点で, 一見方法 2 と変わらないように見える. しかしこれは推定した混合分布のデータ構造 $(\boldsymbol{X}_N, \widehat{\boldsymbol{Z}}_N)$ に対してブートストラップ法を適用していると見ることができる. EM 法により混合分布モデルのパラメータ推定を行うということは, 不完全データの下で最尤推定を行うことに相当するので, この方法は不完全データを推定したデータに対してブートストラップ法を行うことである. 方法 2 が方法 3 と異なるのは $\boldsymbol{X}_N$ からブートストラップ標本 $\boldsymbol{X}_N^*$ を作り, また新たに不完全データの推定からはじめるという点である. さらに, 方法 2 では与えられた $\widehat{\boldsymbol{\Phi}}_r$ から第一回目の反復で事後確率が計算されるが, 方法 3 はすでに事後確率が与えられていて, その上で $\boldsymbol{X}_N^*$ に混合分布モデルをあてはめることになる.

以下の数値例でこれらの三つの方法の比較・検証を行なう.

8.4 数値例による検証

7.5.2節で用いたものと同じ糖尿病データ (Reaven and Miller (1979)) について, コンポーネント数の推定を行う. なお, このデータセットは医学的知見に基づいて事前に分類情報が既知とみられる例である.

8.4.1 コンポーネント数の推定結果

まず, コンポーネント数の推定結果を表 8.1と図 8.1に示す. 表 8.1の数値は次の 5 種類の情報量規準を示し, 太字はそれぞれの中での最小値を表す.

- $\mathrm{ICBoot}_r = -2 \times \hat{\mathcal{L}}(r) + 2N \times \widehat{\mathrm{bias}}_r$,
- $\mathrm{ICBoot}_{\mathrm{M}r} = -2 \times \hat{\mathcal{L}}(r) + 2N \times \widehat{\mathrm{bias}}_{\mathrm{M}r}$,
- $\mathrm{AIC}_r = -2 \times \hat{\mathcal{L}}(r) + 2 \times p(r)$,
- $\mathrm{modAIC}_r = -2 \times \hat{\mathcal{L}}(r) + 3 \times p(r)$; (Bozdgan の修正 AIC, Bozdgan(1992)),
- $\mathrm{MDL}_r = -2 \times \hat{\mathcal{L}}(r) + p(r) \log N$ (MDL 規準, Rissanen(1978, 1984)),

$\hat{\mathcal{L}}(r)$, $p(r)$, N は, それぞれ, コンポーネント数 r のときの最大対数尤度, パラメータ数, 標本数である. 図 8.1は, ICBoot, $\mathrm{ICBoot}_{\mathrm{M}}$, AIC の各値を, 初期分類法に用いたクラスター化法ごとに同時にプロットしたものである.

表 8.1から, 各変数群の ICBoot, $\mathrm{ICBoot}_{\mathrm{M}}$, MDL は, いずれもコンポーネント数 3 を選択している. 一方 AIC と modAIC は 4 を選択している. 図 8.1において, コンポーネント数が 3 の情報量規準の値を見ると (矢印) , 群平均法 (GA) と k-means 法 (KM) が他の規準より値が大きい. つまり, これらの 3 つのクラスター化法から得られた初期値を使って推定されたパラメータが, 他のクラスター化法から出発した収束先とは異なる値 (局所解) に収束したと考えられる.

このデータの解析例は Symons (1981), McLachlan (1992, 6.8 節), Banfield and Raftery (1993) などがあるが, 前者の二つの論文はコンポーネント数を 3 に固定していて, コンポーネント数の推定は行っていない. また, Banfield and Raftery は, クラスター数の推定に尤度比検定統計量の変形版を提案して, このデータに適用しているが, 3 ～ 6 群のうちどれを選んでよいかを決めかねている.

表 8.1: 糖尿病データから推定されたコンポーネント数と対数尤度の値

初期分類法	情報量規準				
	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
MD	**5155.3**[3]	**5156.8**[3]	**5119.2**[4]	**5158.2**[4]	**5220.9**[3]
CD	5158.2[4]	5159.0[4]	**5119.2**[4]	**5158.2**[4]	5235.3[4]
WA	**5155.3**[3]	**5156.8**[3]	**5119.2**[4]	**5158.2**[4]	**5220.9**[3]
GA	5158.9[3]	5159.2[3]	5134.5[3]	5163.5[3]	**5220.9**[3]
CL	5158.2[4]	5159.0[4]	**5119.2**[4]	**5158.2**[4]	5235.3[4]
FX	5157.9[3]	5159.0[3]	**5119.2**[4]	**5158.2**[4]	5222.8[3]
WD	5161.3[3]	5161.7[3]	5136.5[3]	5165.5[3]	5222.8[3]
MF	5162.6[3]	5161.8[3]	5124.9[4]	5163.9[4]	5222.8[3]
KM -	5158.2[4]	5159.0[4]	**5119.2**[4]	**5158.2**[4]	5235.3[4]

表中の太文字は各情報量規準の中で最小値を示す.

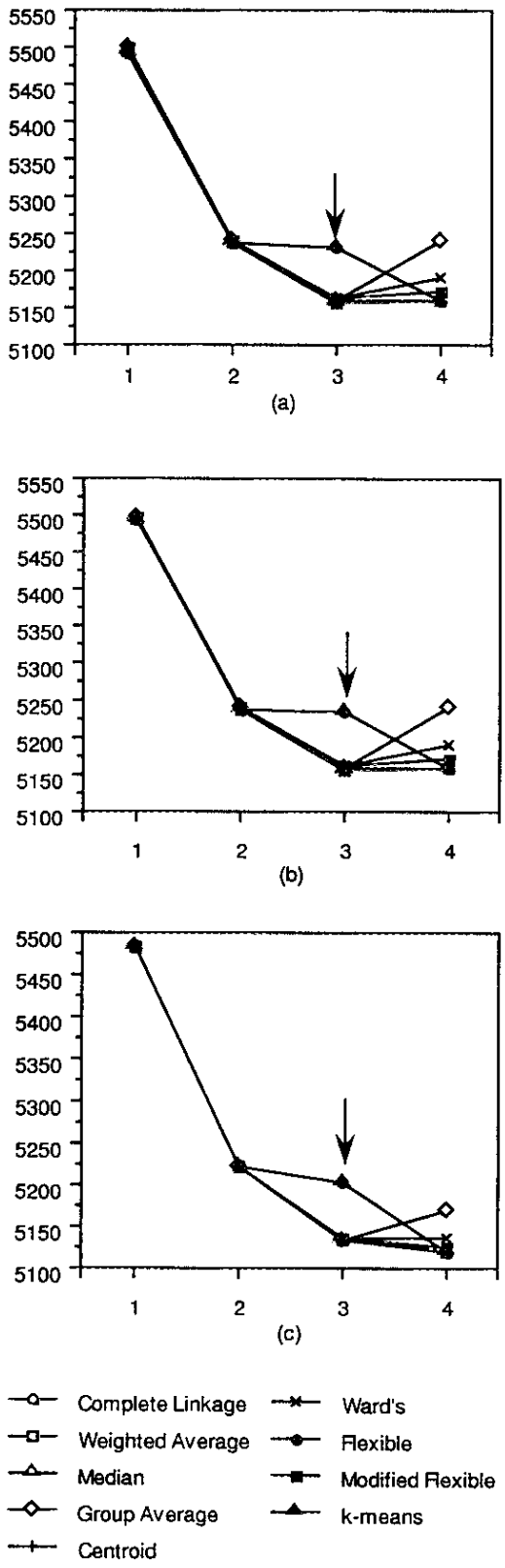

図 8.1: 糖尿病データのコンポーネント数の推定結果
(a) ICBoot, (b) $\mathrm{ICBoot}_{\mathrm{M}}$, (c) AIC.

表 8.2: 糖尿病データのブートストラップ反復の所要時間・回数

	実行時間 (秒)	回数 *1
方法 1	1221.49	500+646
方法 2	601.24	500+105
方法 3	590.18	500+105

*1 500 回のブートストラップ反復を行うのに発生させたブートストラップ標本の数. 使用機種は HITACHI ワークステーション 3050RX.

8.4.2 ブートストラップ標本の初期値設定

次に, 同じデータセットを使って 8.3節で考察したブートストラップ標本に混合分布モデルをあてはめるときの初期値設定の方法の検証を行う. 観測データの初期分類法としてメディアン法を用い[3], ブートストラップ反復回数は $B = 500$ とした.

表 8.2に実行時間と 500 回のブートストラップ反復を行うのに発生させたブートストラップ標本の数を示す. この表から, 方法 1 は実行時間, ブートストラップ標本の抽出回数のいずれについても, 他の方法に比べてロスが多いことがわかる. 次に, 各方法で観測データに 1 ～ 4 群をあてはめたときの対数尤度の推定値を表 8.3に示す. また, 表 8.4はブートストラップ標本における対数尤度の推定値 (平均値) とその標準偏差, 対数尤度の最大値, 平均収束回数とその標準偏差である. 図 8.2はコンポーネント数が 1 ～ 4 に対して, 方法 1 で行ったときのブートストラップ標本から推定された対数尤度のヒストグラムである. コンポーネント数が 2 の (b) のときに 2 つの山ができている. 小さい方の山は局所解である. ここで, 8.3節で考察した局所解に収束する可能性が高いという現象が起きている. 他の方法ではコンポーネント数 1 ～ 4 に対していずれも 1 つの山であった. 以上の数値実験から次の知見が得られた.

(1) 方法 2 と 3 はほとんど同じ結果を示し, 方法 3 が丁度 1 回収束が早い.

(2) 方法 1 は局所解に収束する可能性が高い. また平均的に対数尤度の値も方法 1, 2 に比べて小さく, その標準偏差も大きく, EM 法の反復回数も多い.

(3) 観測データで推定されたコンポーネント分布 f_k と, ブートストラップ標本から推定さ

[3] メディアン法は表 8.1で情報量規準が最小の初期分類法であるので, この分類法でテストを行なってみた.

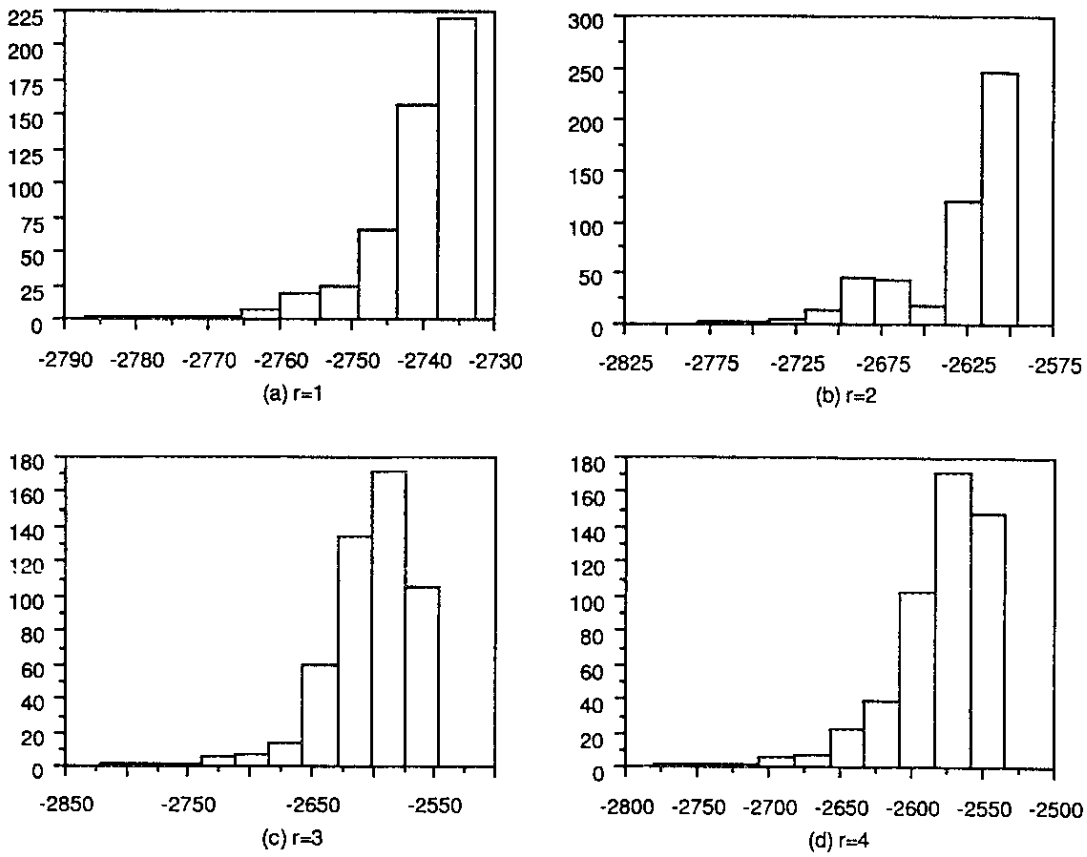

図 8.2: 糖尿病データの対数尤度のヒストグラム
(a) 1 群, (b) 2 群, (c) 3 群, (d) 4 群, 初期分類法はメディアン法. $B = 500$.

れたコンポーネント分布 $f^*_{k'}$ との対応関係[4] が取れた回数は, 方法 2, 3 では 100% であるが, 方法 1 は平均して 80% 程度である.

これらのことから, 次の数値実験においてブートストラップ標本に混合分布モデルをあてはめるときは, 観測データから推定されたパラメータを初期値とする方法を用いる.

8.5 シミュレーション実験

混合分布のデータセットを乱数により作成し, これに対して混合分布モデルをあてはめ, いくつの分布が混合しているかを推定する. ここでは 7.6節の数値実験の結果を参考に, 初期

[4] ここで対応関係とは $\sum_{k,k'} \|\theta_k - \hat{\theta}_{k'}\| < \varepsilon$ (k'は$k = \{1, \cdots, r\}$の順列) となったときをさす. ここで $\theta_k = \{\mu_k, \Sigma_k\}$, ε は任意の正数である ($10^0 \sim 10^{-3}$ 程度).

表 8.3: 1 ～ 4 群での対数尤度の推定値

コンポーネント数	対数尤度の推定値
1	−2732.0
2	−2592.2
3	−2538.3
4	−2520.6

表 8.4: 1 ～ 4 群での対数尤度の推定値

	コンポーネント数	対数尤度の推定値の平均と標準偏差	対数尤度の最大値	EM 法の反復回数の平均と標準偏差	対応回数 †
方法 1	1	−2741.1(7.59)	−2732.7	2.0(0)	500
	2	−2629.7(35.68)	−2596.0	34.1(25.1)	419
	3	−2602.6(34.6)	−2547.3	51.2(67.7)	230
	4	−2580.0(34.6)	−2534.2	56.5(42.7)	423
方法 2	1	−2740.0(6.90)	−2732.2	2.0(0)	500
	2	−2606.7(9.07)	−2594.4	30.8(18.2)	500
	3	−2561.0(11.6)	−2544.9	46.2(31.7)	500
	4	−2553.7(18.8)	−2527.8	54.9(69.1)	500
方法 3	1	−2740.0(6.90)	−2732.2	2.0(0)	500
	2	−2606.7(9.07)	−2594.4	29.8(18.2)	500
	3	−2561.0(11.6)	−2544.9	45.2(31.7)	500
	4	−2553.7(18.8)	−2527.8	53.9(69.1)	500

† 観測データで推定されたコンポーネント分布 f_k とブートストラップ標本から推定されたコンポーネント分布 $f_{k'}^*$ との対応がとれた回数 (103ページの脚注を参照).

分類法として k-means 法のみを用いた. そして, 線型近似によるブートストラップバイアス推定の変動を減少させる方法 (北川・石黒・坂元 (1993), 小西 (1993a)) を用いて, 通常のバイアス推定との比較を行う.

シミュレーションの目的は次の 4 点にある：

(1) ブートストラップ法を用いた情報量規準により混合分布モデルのコンポーネント数を推定する. また, 他の情報量規準との比較を行う.

(2) 対数尤度のバイアス ($\widehat{\text{bias}}$) および変動減少法を用いた場合のバイアス ($\widehat{\text{bias}}_{\text{M}}$) の比較を行う. とくに, $\widehat{\text{bias}} \fallingdotseq \widehat{\text{bias}}_{\text{M}}$ であるが, $\text{Var}(\widehat{\text{bias}}) > \text{Var}(\widehat{\text{bias}}_{\text{M}})$ となることを検証する.

(3) ブートストラップバイアス推定の変動減少法により, ブートストラップの回数を減らすことができることを検証する.

(4) 混合分布モデルのパラメータの計算には膨大な時間を要するため, 有効なブートストラップ・シミュレーション[5]を行ない, そのシミュレーション回数の検討を行なう.

シミュレーションを行う場合のコンポーネント数の推定手続きにおいて, 8.2節のステップ1. に, 混合分布 $G(\cdot)$ にしたがうデータを発生させるステップを加え, この推定手続きを S 回繰り返す. そしてシミュレーション全体の情報量規準の平均値（期待値）とバイアスの平均値（期待値）とその分散を計算する. なお, 混合分布のデータセットの抽出方法は, 7.6.2節と同様に混合サンプリングを用いる.

8.5.1 数値実験の準備

2変量の混合分布モデルを考える. コンポーネント数として $r = 3$, 各コンポーネント分布の平均ベクトルの位置関係は正三角形とし, 1辺の長さ（分布間の距離）を d とする. 次に示す2種類のパラメータ設定のデータ生成に共通の条件は, 標本数は $N = 200$, 各コンポーネント分布の分散共分散行列は単位行列（$\boldsymbol{\Sigma}_k = \boldsymbol{I}$, k=1,2,3）である. したがって, パラメータ推定に関する条件として, 共通の分散共分散行列を仮定する. その他の条件は表8.5のように設定する. なお, 混合比率は整数比で表してある.

ここで設定1は, コンポーネント数の推定と, ブートストラップ反復の有効な回数を調べるための設定である. 設定2は有効なシミュレーションの回数を調べることを目的とする. また, 設定3は分布間の距離と混合比率を変えたときの情報量規準のふるまいを調べるためのものである.

[5] "有効なブートストラップシミュレーション" とは, 標本が与えられたもとで対数尤度のバイアスの分散を可能な限り小さくする手法のことである. ここでは "線形近似法" と呼ばれる方法を用いる (久保川, 江口, 竹村, 小西 (1993) の第4章などを参照).

表 8.5: シミュレーションのデータセットの設定

条件	設定 1	設定 2	設定 3
データ生成			
分布間の距離 (d)	3.0	3.0	3.0, 2.5, 2.0
混合比率 (π)	1 : 1 : 1	1 : 1 : 1	1 : 1 : 1, 2 : 1 : 1, 2 : 2 : 1, 3 : 1 : 1, 3 : 2 : 1, 4 : 1 : 1
標本数 (N)	200	200	50, 100, 200[*2]
推定手続			
あてはめる分布の数 (r)	1 ～ 6	1 ～ 6	1 ～ 6
シミュレーションの回数 (S)	200	100, 150, 200, 500, 1000[*1]	100, 75, 50[*2]
ブートストラップの回数 (B)	5, 10, 20, 30, 40, 50, 100	5, 20	20

[*1] $S = 1000$は$B = 5$ の場合のみ, [*2] NとSは対応する.

表 8.6: AIC と ICBoot が選んだコンポーネント数のクロス表 (設定 1 の場合)

		上段:ICBoot 下段:ICBoot$_M$						
		1	2	3	4	5	6	Sum
AIC	1	0	0	0	0	0	0	0
		0	0	0	0	0	0	
	2	0	0	0	0	0	0	0
		0	0	0	0	0	0	
	3	0	0	128	1	0	0	128
		0	0	128	1	0	0	
	4	0	0	18	20	0	0	38
		0	0	17	21	0	0	
	5	0	0	9	0	5	0	14
		0	0	10	0	4	0	
	6	0	0	11	3	0	5	19
		0	0	11	2	1	5	
	Sum	0	0	166	24	5	5	200
		0	0	166	24	5	5	

$N = 200, S = 200, B = 50, d = 3.0$.

表 8.7: 各種の情報量規準が選んだコンポーネント数の回数 (設定 1)

コンポーネント数	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
1	0	0	0	0	9
2	0	0	0	0	3
3	166	166	129	184	187
4	24	24	38	15	1
5	5	5	14	0	0
6	5	5	19	1	0

表 8.8: 各種の情報量規準の値

コンポーネント数	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
1	1506.0343	1506.2642	1507.5723	1512.5723	1524.0639
2	1493.8840	1494.1968	1494.0606	1502.0606	1520.4471
3	1469.1783	1469.6610	1467.7527	1478.7527	1504.0342
4	1473.0023	1473.4548	1469.7487	1483.7487	1515.9251
5	1477.2882	1477.7112	1471.4415	1488.4415	1527.5128
6	1481.3352	1481.7025	1473.1398	1493.1398	1539.1061

8.5.2 実験結果

設定 1 による実験

図 8.3(a) は $S = 200, B = 50, d = 3.0$ の設定に対するシミュレーションの結果を示す. この図は 5 種類の情報量規準 ICBoot, ICBoot_{M}, AIC, modAIC, MDL と, $-2N \times \hat{\mathcal{L}}(r)$(図では LL) の値をプロットしたものである. $\hat{\mathcal{L}}(r)$, N は, それぞれコンポーネント数 r のときの最大対数尤度と標本数である. また, MDL は AIC と対応がつくように -2 倍してある. 横軸はコンポーネント分布の数, 縦軸はこれらの値である. 5 種類の情報量規準はいずれも 3 群で最小となり, もとのコンポーネント数を正しく推定している. 図 8.3(b) は同じ設定 1 における第 1 回目のシミュレーションでの 5 種類の情報量規準の値をプロットしたものである. この場合, いづれもコンポーネント数 3 を選んでいるが, AIC はコンポーネント数 3 と 6 の値がほぼ同じである.

図 8.3(a) はシミュレーション全体を通しての平均であるため, 推定した対数尤度の変動は少なくなり, コンポーネント数が 3 以上ではいずれも滑らかな変化である. これに比べ図 8.3(b) は変動が大きく, 1 回のみの試行では対数尤度の推定にかなりの変動があることがわかる.

シミュレーション全体に対する結果が表 8.6である ($N = 200$, $S = 200$, $B = 50$, $d = 3.0$). これは各シミュレーション毎に情報量規準が選んだ (推定した) コンポーネント数について, AIC に対する ICBoot, ICBoot_{M} のクロス集計である. 表から分かるように, この場合いずれの情報量規準もコンポーネント数 3 の頻度が一番多いが, シミュレーションの各回で常にコンポーネント数を 3 に選んでいるわけではない. また, 表 8.7ともあわせて見ると, ICBoot や ICBoot_{M} に比べ, 経験的に言われているように AIC がコンポーネント数を多めに推定している様子がわかる. また, そのときのシミュレーション全体の情報量規準の値を表 8.8に示す. いずれも全体としてはコンポーネント数 3 で最小値となり, 正しく推定している.

図 8.4は $r = 1, \cdots, 5$ に対してブートストラップの回数を $B = 5$〜100 回として得られたバイアスの値, バイアス $\pm 1 \times$ S.D.(標準偏差) の値 (左側), 標準偏差の値 (右側) をそ

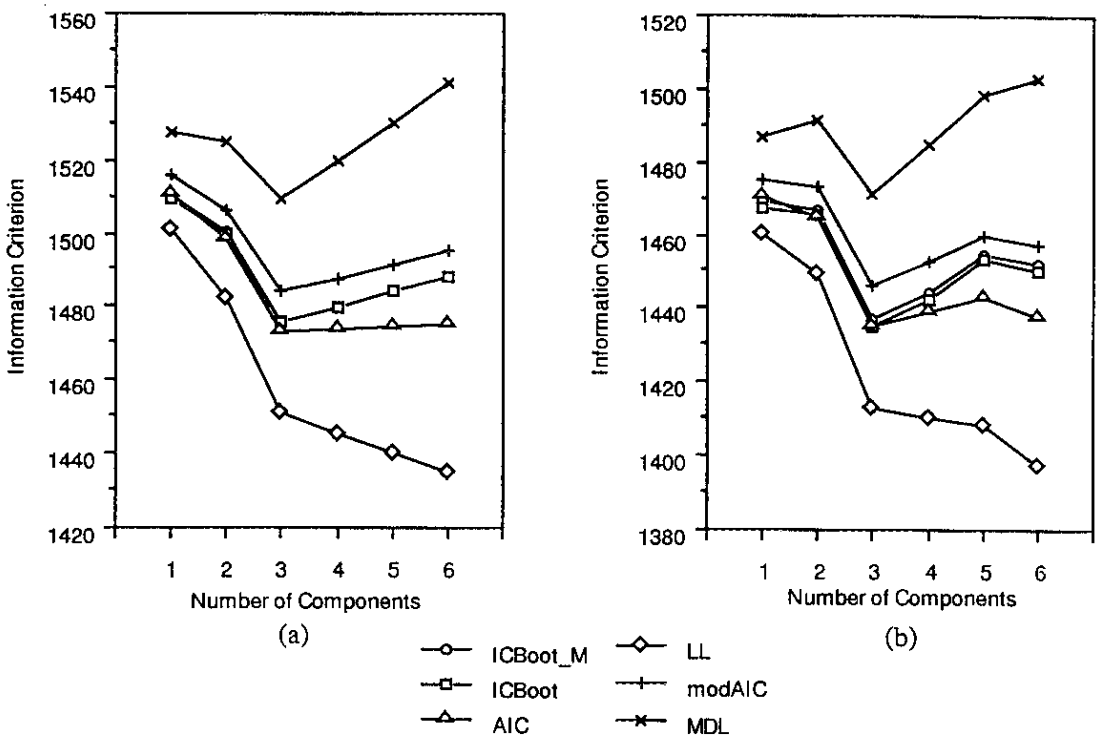

図 8.3: コンポーネント数推定のシミュレーション結果
(a) シミュレーション全体, (b) シミュレーションの第 1 回目, 設定は
$S = 200, B = 50, d = 3.0$.

れぞれプロットしたものである. 横軸はブートストラップの回数で, 点線は ICBoot, 実線は ICBoot_{M} に対応する. まず, これらの図から $\widehat{\text{bias}} \fallingdotseq \widehat{\text{bias}}_{\text{M}}$ であることがわかるが, ICBoot_{M} が多少大きくなる傾向がある. そして, 左側のいずれの図も ICBoot_{M} が ICBoot の内側にある. 右側の図でも ICBoot_{M} が常に下方にある. これは, すべてのブートストラップの回数設定に対してバイアス推定の変動が減少していることを示している. 図の右側の点線はほぼ単調に減少し, $B = 100$ で最小になっている. $B = 100$ での ICBoot の標準偏差の最小値とほぼ同じ ICBoot_{M} の実線の回数を見ると, コンポーネント数が 1 のときを除いてほぼ 30 ～ 50 回の間である. つまり, バイアス推定の変動減少法を用いることで, 通常ブートストラップの回数を 100 回行うところを 30 ～ 50 回程度に減らしても同じ効果が得られることがこの実験から読みとれる.

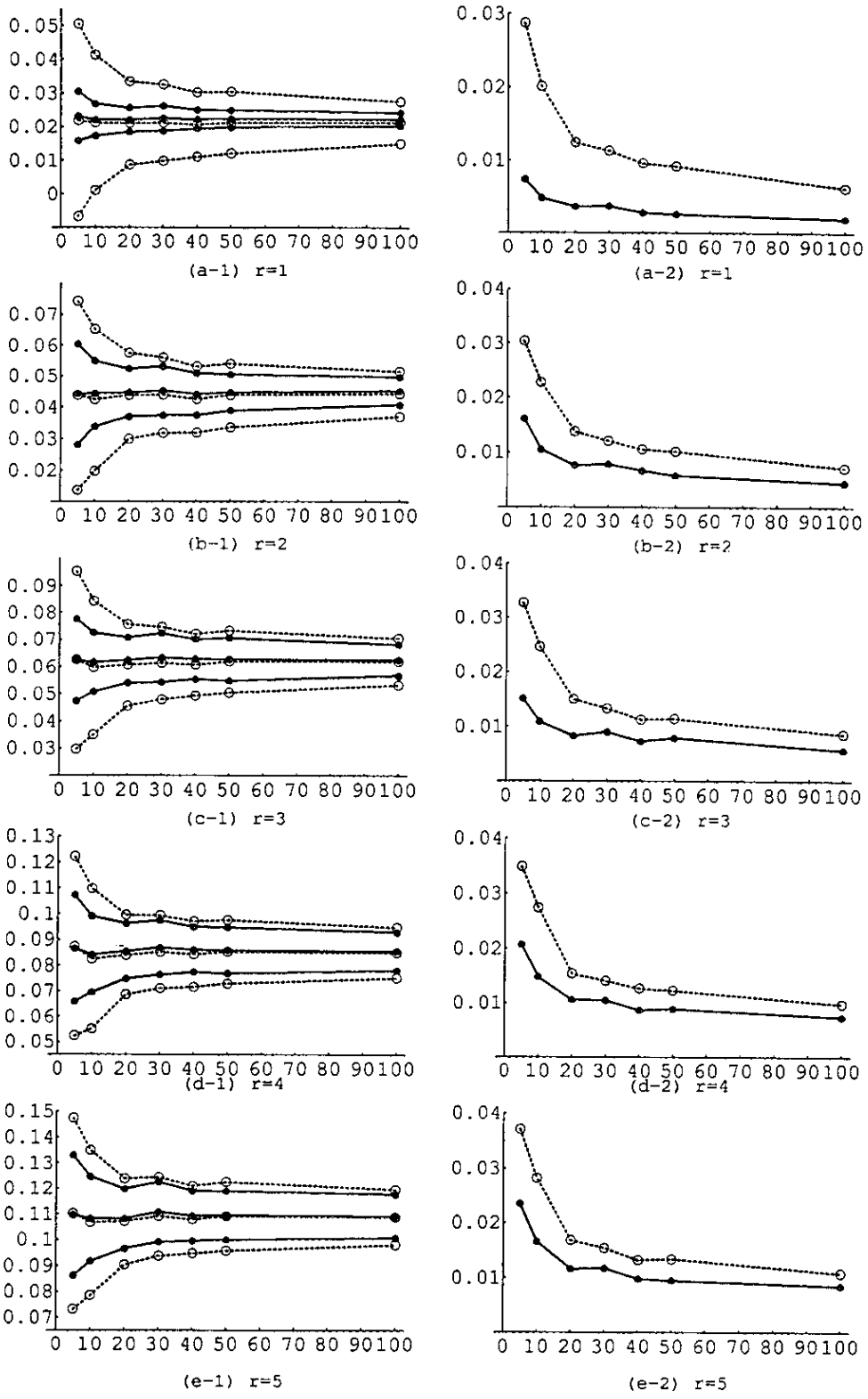

図 8.4: 推定されたバイアスとその標準偏差
(a) コンポーネント数 $r = 1$, (b) $r = 2$, (c) $r = 3$, (d) $r = 4$, (e) $r = 5$, 左側の (?-1) はバイアス $\pm 1 \times$ S.D.(標準偏差), 右側の (?-2) はバイアス, 点線は ICBoot, 実線は ICBoot_{M}.

設定 2 による実験

図 8.5 (a) は $B = 5, S = 100, 150, 200, 500, 1000$ 回の場合のバイアスの標準偏差, 図 8.5 (b) は $B = 20, S = 100, 150, 200, 500$ のバイアスの標準偏差のプロットで, 横軸はコンポーネント数である. 二つの図の上方の折れ線の束は, 通常のブートストラップ法によるバイアスの標準偏差で, 下方の折れ線の束はバイアスの変動減少法による標準偏差である. 図 8.5(a) では, 変動減少法を用いたバイアスの変動は (図の下), $S = 150$ 以上でほぼ同じ傾向を示しているが, $S = 100$ はコンポーネント数が 3 と 6 において他のものと傾向が異なる. 同様に, 通常の方法によるバイアスの変動 (図の上) も $S = 100$ 以外はほぼ同じ変化をしていて, シミュレーションの回数が多くなるほど, バイアスの標準偏差の値も大きくなる. しかし, $S = 1000$ では変化は滑らかで, $S = 150$〜500 よりバイアスの標準偏差の値は小さい. 一方, 図 8.5(b) では, シミュレーションの回数が多くなるほどバイアスの標準偏差の値も大きくなる傾向が見られる. いずれもほぼ同じ傾向で変化しているが, $S = 500$ の場合はその変化は滑らかである. 以上のことから, シミュレーションの回数は, 少なくとも $S = 500$ 程度以上は必要であると思われる.

設定 3 による実験

設定 3 はコンポーネント分布間の距離, 混合比率, 標本数をそれぞれ変化させたとき, 情報量規準がどのようなふるまいを示すかを調べた実験である. 各シミュレーションの設定条件を表 8.5に示す. 各種情報量規準の正しいコンポーネント数を推定した割合は表 8.9に, さらに詳しい結果を表 8.10〜表 8.12に示す. これらの実験結果から次の知見が得られた.

- ICBoot と $\mathrm{ICBoot_M}$ はほぼ同じふるまいをする.
- コンポーネント分布間の距離を小さくすると, 他の情報量規準に比べて AIC は正解率が良いが, これはコンポーネント数を多めに推定していることに起因していると考られる.
- MDL は d が大きいとき推定は非常に良いが, 小さくなると急速に悪くなる.

表 8.9: 正しいコンポーネント数を推定した割合 (%, 設定 3)

π	Information criterions	$d=3$			$d=2.5$			$d=2$		
		200	100	50	200	100	50	200	100	50
1:1:1	ICBoot	83	79	38	78	32	20	30	9	3
	ICBoot_{M}	89	84	35	77	37	17	32	7	4
	AIC	67	56	32	64	48	29	40	22	19
	modAIC	92	81	33	77	32	16	23	11	5
	MDL	95	47	13	30	4	7	0	0	1
2:1:1	ICBoot	86	74	29	69	41	25	20	16	0
	ICBoot_{M}	88	75	29	69	39	25	21	15	0
	AIC	71	62	34	63	35	27	31	27	9
	modAIC	91	70	38	69	32	22	23	12	3
	MDL	87	46	23	33	8	7	0	1	0
2:2:1	ICBoot	83	67	36	70	48	13	25	8	5
	ICBoot_{M}	83	67	36	68	51	13	24	9	6
	AIC	61	62	40	59	45	26	30	25	12
	modAIC	88	70	38	70	44	10	24	7	4
	MDL	93	49	20	33	10	4	3	1	1
3:1:1	ICBoot	87	73	30	63	41	12	24	12	1
	ICBoot_{M}	86	71	33	68	42	12	24	10	2
	AIC	67	53	41	51	41	25	35	19	18
	modAIC	91	64	28	62	32	11	17	10	2
	MDL	95	47	15	26	5	5	0	1	0
3:2:1	ICBoot	86	73	28	65	27	11	22	14	4
	ICBoot_{M}	83	72	31	67	30	9	22	15	4
	AIC	33	57	62	65	27	27	31	28	16
	modAIC	83	70	32	63	25	17	18	10	3
	MDL	83	40	19	24	8	5	3	0	0
4:1:1	ICBoot	78	72	26	65	37	7	23	11	5
	ICBoot_{M}	76	70	26	62	39	8	23	8	5
	AIC	66	48	38	53	30	22	29	20	18
	modAIC	83	59	27	60	28	9	19	9	6
	MDL	81	38	12	21	7	1	2	1	1

表 8.10: シミュレーション設定 3 の結果 ($d = 3$)

$d = 3, N = 50, \pi = 1:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	34	39	6	31	68
2	22	19	7	21	17
3	38	35	32	33	13
4	5	5	15	5	2
5	1	2	26	9	0
6	0	0	14	1	0

$d = 3, N = 100, \pi = 1:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	6	5	0	5	43
2	4	2	1	4	10
3	79	84	56	81	47
4	9	6	19	7	0
5	2	3	12	2	0
6	0	0	12	1	0

$d = 3, N = 200, \pi = 1:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	0	0	0	0	2
2	1	1	0	1	1
3	83	89	67	92	95
4	9	5	12	5	2
5	6	5	17	2	0
6	1	0	4	0	0

$d = 3, N = 50, \pi = 2:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	48	45	15	41	63
2	18	17	11	15	12
3	29	29	34	38	23
4	3	8	13	5	2
5	2	1	15	1	0
6	0	0	12	0	0

$d = 3, N = 100, \pi = 2:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	12	12	1	13	44
2	4	6	0	5	9
3	74	75	62	70	46
4	6	3	17	8	1
5	2	2	10	2	0
6	2	2	10	2	0

$d = 3, N = 200, \pi = 2:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	0	0	0	0	5
2	1	1	0	1	4
3	86	88	71	91	87
4	8	8	13	7	4
5	5	3	15	1	0
6	0	0	1	0	0

$d = 3, N = 50, \pi = 2:2:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	35	36	6	32	62
2	23	23	12	22	18
3	36	36	40	38	20
4	5	3	19	7	0
5	1	1	14	0	0
6	0	1	9	1	0

$d = 3, N = 100, \pi = 2:2:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	6	5	0	6	37
2	12	13	5	12	13
3	67	67	62	70	46
4	10	9	15	7	3
5	5	5	11	4	1
6	0	1	7	1	0

$d = 3, N = 200, \pi = 2:2:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	0	0	0	0	3
2	0	0	0	0	4
3	83	83	61	88	93
4	11	10	19	9	0
5	4	6	13	2	0
6	2	1	7	1	0

$d = 3, N = 50, \pi = 3:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	49	44	12	42	68
2	14	16	13	21	15
3	30	33	41	28	15
4	4	4	19	7	1
5	2	2	10	1	0
6	1	1	5	1	1

$d = 3, N = 100, \pi = 3:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	5	6	0	8	31
2	9	9	6	13	19
3	73	71	53	64	47
4	12	12	26	12	3
5	1	2	7	2	0
6	0	0	8	1	0

$d = 3, N = 200, \pi = 3:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	0	0	0	0	3
2	0	0	0	0	2
3	87	86	67	91	95
4	8	9	12	7	0
5	3	3	13	2	0
6	2	2	8	0	0

$d = 3, N = 50, \pi = 3:2:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	46	44	8	35	66
2	23	22	16	24	15
3	28	31	33	32	19
4	1	1	15	5	0
5	2	2	19	4	0
6	0	0	9	0	0

$d = 3, N = 100, \pi = 3:2:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	11	9	0	7	40
2	9	8	6	14	19
3	73	72	57	70	40
4	6	6	17	7	1
5	1	5	10	2	0
6	0	0	10	0	0

$d = 3, N = 200, \pi = 3:2:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	0	0	0	0	5
2	2	1	0	2	10
3	86	83	62	83	83
4	10	12	19	12	2
5	1	4	9	3	0
6	1	0	10	0	0

$d = 3, N = 50, \pi = 4:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	46	42	12	44	70
2	23	25	11	26	17
3	26	26	38	27	12
4	5	7	18	3	1
5	0	0	9	0	0
6	0	0	12	0	0

$d = 3, N = 100, \pi = 4:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	8	6	0	7	36
2	12	13	6	20	24
3	72	70	48	59	38
4	4	6	19	6	2
5	1	2	14	5	0
6	3	3	13	3	0

$d = 3, N = 200, \pi = 4:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	0	0	0	0	5
2	2	3	0	4	13
3	78	76	66	83	81
4	14	15	20	12	1
5	5	5	8	1	0
6	1	1	6	0	0

表 8.11: シミュレーション設定 3 の結果 ($d = 2.5$)

$d = 2.5, N = 50, \pi = 1:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	69	66	28	68	85
2	10	16	10	12	8
3	20	17	29	16	7
4	0	0	15	2	0
5	1	1	12	2	0
6	0	0	6	0	0

$d = 2.5, N = 100, \pi = 1:1:1$

1	45	39	13	47	92
2	19	19	13	15	4
3	32	37	48	32	4
4	3	4	11	4	0
5	1	1	10	2	0
6	0	0	5	0	0

$d = 2.5, N = 200, \pi = 1:1:1$

1	8	9	0	11	58
2	6	6	4	7	12
3	78	77	64	77	30
4	5	6	11	4	0
5	3	2	13	1	0
6	0	0	8	0	0

$d = 2.5, N = 50, \pi = 2:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	57	58	19	53	80
2	17	15	13	18	11
3	25	25	27	22	7
4	1	1	20	6	2
5	0	0	10	1	0
6	0	1	11	0	0

$d = 2.5, N = 100, \pi = 2:1:1$

1	27	28	7	39	80
2	24	25	17	19	12
3	41	39	35	32	8
4	4	4	14	7	0
5	4	3	16	3	0
6	0	1	11	0	0

$d = 2.5, N = 200, \pi = 2:1:1$

1	12	13	4	14	55
2	5	4	3	7	11
3	69	69	63	69	33
4	11	10	15	8	1
5	1	2	5	1	0
6	2	2	10	1	0

$d = 2.5, N = 50, \pi = 2:2:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	68	68	28	74	89
2	15	14	10	9	5
3	13	13	26	10	4
4	1	2	13	2	1
5	3	3	13	4	1
6	0	0	10	1	0

$d = 2.5, N = 100, \pi = 2:2:1$

1	31	30	9	36	78
2	15	15	21	16	12
3	48	51	45	44	10
4	4	3	14	3	0
5	2	1	3	1	0
6	0	0	8	0	0

$d = 2.5, N = 200, \pi = 2:2:1$

1	9	9	2	11	56
2	12	11	1	10	11
3	70	68	59	70	33
4	8	7	19	6	0
5	1	5	10	3	0
6	0	0	9	0	0

$d = 2.5, N = 50, \pi = 3:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	62	65	19	63	86
2	25	22	24	21	9
3	12	12	25	11	5
4	0	0	12	4	0
5	1	1	13	1	0
6	0	0	7	0	0

$d = 2.5, N = 100, \pi = 3:1:1$

1	40	38	11	45	83
2	10	12	13	15	12
3	41	42	41	32	5
4	5	4	12	6	0
5	2	2	10	1	0
6	2	2	13	1	0

$d = 2.5, N = 200, \pi = 3:1:1$

1	14	11	3	14	60
2	7	9	5	12	13
3	63	68	51	62	26
4	12	9	21	11	1
5	3	3	16	1	0
6	1	0	4	0	0

$d = 2.5, N = 50, \pi = 3:2:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	66	68	24	65	86
2	21	21	13	14	9
3	11	9	27	17	5
4	1	2	10	3	0
5	1	0	15	1	0
6	0	0	11	0	0

$d = 2.5, N = 100, \pi = 3:2:1$

1	45	40	13	42	80
2	17	22	21	25	12
3	28	30	27	25	8
4	8	6	20	6	0
5	1	2	8	0	0
6	1	0	11	2	0

$d = 2.5, N = 200, \pi = 3:2:1$

1	9	8	2	10	60
2	11	10	7	14	16
3	65	67	56	63	24
4	12	12	18	9	0
5	2	1	11	3	0
6	1	2	6	1	0

$d = 2.5, N = 50, \pi = 4:1:1$

g	ICBoot	ICBoot_M	AIC	modAIC	MDL
1	73	72	28	69	89
2	18	18	18	16	10
3	7	8	22	9	1
4	2	2	17	3	0
5	0	0	7	2	0
6	0	0	8	1	0

$d = 2.5, N = 100, \pi = 4:1:1$

1	28	26	7	33	74
2	28	28	22	34	19
3	37	39	30	28	7
4	6	6	21	4	0
5	1	1	13	1	0
6	0	0	7	0	0

$d = 2.5, N = 200, \pi = 4:1:1$

1	8	7	1	10	55
2	15	18	11	23	24
3	65	62	53	60	21
4	10	12	17	7	0
5	0	0	10	0	0
6	2	1	8	0	0

表 8.12: シミュレーション設定 3 の結果 ($d = 2$)

$d = 2, N = 50, \pi = 1:1:1$

g	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
1	87	87	40	89	96
2	8	8	15	5	3
3	3	4	19	5	1
4	2	1	8	1	0
5	0	0	10	0	0
6	0	0	8	0	0

$d = 2, N = 100, \pi = 1:1:1$

1	69	72	32	74	94
2	18	17	22	17	6
3	10	9	22	9	0
4	3	2	12	0	0
5	0	0	8	0	0
6	0	0	4	0	0

$d = 2, N = 200, \pi = 1:1:1$

1	51	48	21	58	98
2	14	15	12	17	2
3	30	32	40	23	0
4	4	3	12	2	0
5	0	0	8	0	0
6	1	2	7	0	0

$d = 2, N = 50, \pi = 2:1:1$

g	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
1	84	84	38	82	91
2	16	16	23	15	9
3	0	0	9	3	0
4	0	0	13	0	0
5	0	0	10	0	0
6	0	0	7	0	0

$d = 2, N = 100, \pi = 2:1:1$

1	65	64	24	62	96
2	16	19	22	23	3
3	16	15	27	12	1
4	3	2	13	3	0
5	0	0	8	0	0
6	0	0	6	0	0

$d = 2, N = 200, \pi = 2:1:1$

1	53	50	26	54	95
2	19	22	16	18	5
3	20	21	31	23	0
4	4	4	14	4	0
5	2	1	5	1	0
6	2	2	8	0	0

$d = 2, N = 50, \pi = 2:2:1$

g	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
1	77	78	37	80	94
2	17	16	26	15	5
3	5	6	12	4	1
4	1	0	13	1	0
5	0	0	7	0	0
6	0	0	5	0	0

$d = 2, N = 100, \pi = 2:2:1$

1	72	65	27	69	96
2	18	23	28	20	3
3	8	9	25	7	1
4	1	2	5	2	0
5	1	1	8	2	0
6	0	0	9	0	0

$d = 2, N = 200, \pi = 2:2:1$

1	55	54	26	60	93
2	15	16	16	12	4
3	25	24	30	24	3
4	4	4	11	3	0
5	0	2	11	0	0
6	1	0	6	1	0

$d = 2, N = 50, \pi = 3:1:1$

g	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
1	76	74	37	75	94
2	23	24	22	22	6
3	1	2	18	2	0
4	0	0	13	0	0
5	0	0	5	1	0
6	0	0	5	0	0

$d = 2, N = 100, \pi = 3:1:1$

1	63	68	36	70	95
2	24	20	16	18	4
3	12	10	19	10	1
4	1	2	8	1	0
5	0	0	8	1	0
6	0	0	13	0	0

$d = 2, N = 200, \pi = 3:1:1$

1	45	45	13	54	95
2	21	20	18	23	5
3	24	24	35	17	0
4	7	8	17	5	0
5	3	3	11	1	0
6	0	0	6	0	0

$d = 2, N = 50, \pi = 3:2:1$

g	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
1	78	80	40	84	96
2	17	15	19	10	4
3	4	4	16	3	0
4	1	1	8	3	0
5	0	0	7	0	0
6	0	0	10	0	0

$d = 2, N = 100, \pi = 3:2:1$

1	69	67	36	78	98
2	15	14	15	11	2
3	14	15	28	10	0
4	0	2	9	0	0
5	1	1	6	1	0
6	1	1	6	0	0

$d = 2, N = 200, \pi = 3:2:1$

1	48	48	22	56	95
2	18	19	20	18	2
3	22	22	31	18	3
4	9	9	11	7	0
5	2	1	5	1	0
6	1	1	11	0	0

$d = 2, N = 50, \pi = 4:1:1$

g	ICBoot	ICBoot_{M}	AIC	modAIC	MDL
1	75	73	32	72	89
2	17	17	21	18	10
3	5	5	18	6	1
4	1	3	10	2	0
5	1	1	11	1	0
6	1	1	8	1	0

$d = 2, N = 100, \pi = 4:1:1$

1	73	72	36	74	96
2	16	19	20	17	3
3	11	8	20	9	1
4	0	0	6	0	0
5	0	0	9	0	0
6	0	1	9	0	0

$d = 2, N = 200, \pi = 4:1:1$

1	47	47	16	54	92
2	23	22	21	22	6
3	23	23	29	19	2
4	7	6	17	5	0
5	0	2	9	0	0
6	0	0	8	0	0

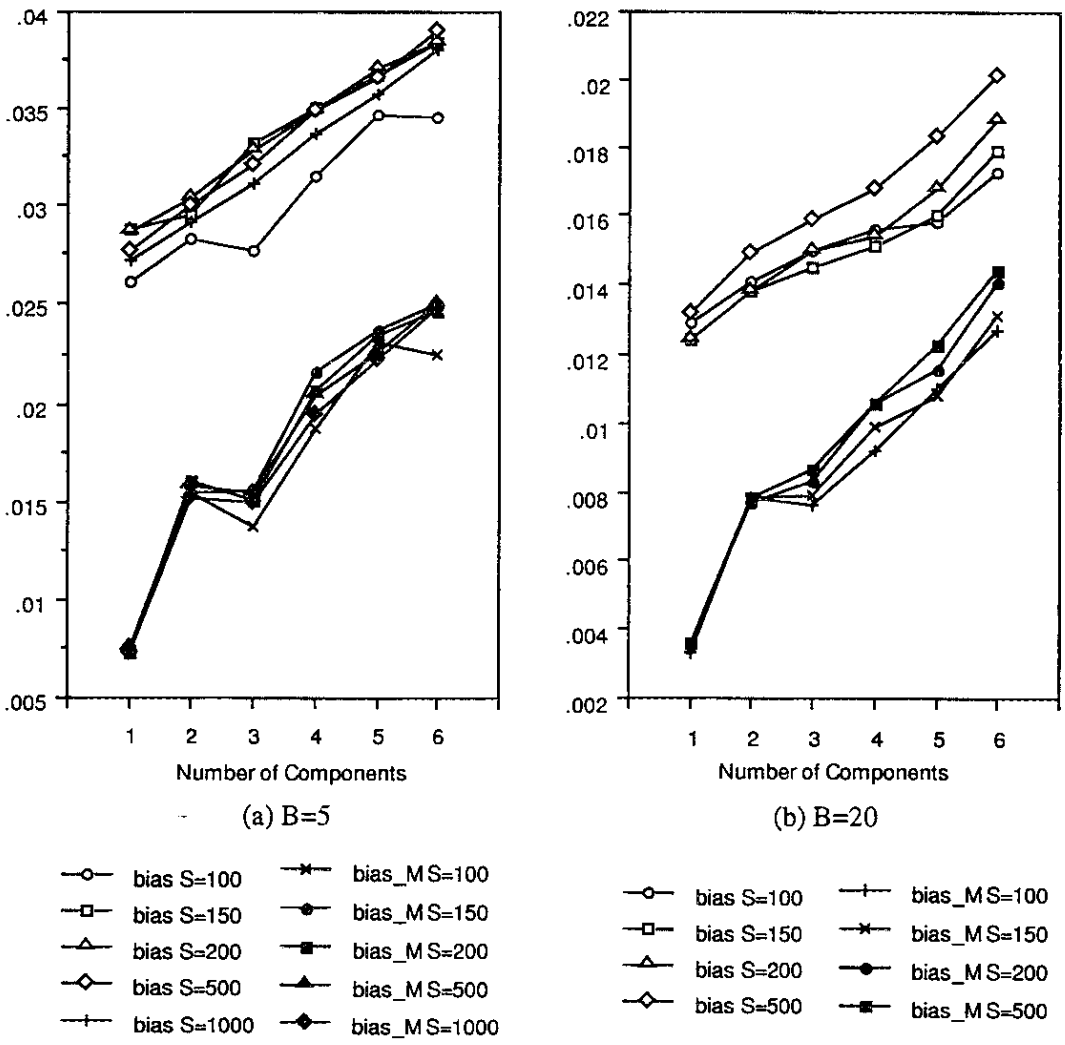

図 8.5: シミュレーション回数を変えたときのバイアスの標準偏差の変動
(a) ブートストラップ反復回数 $B = 5$, シミュレーションの回数
$S = 100, 150, 200, 500, 1000$, (b) $B = 20$,$S = 100, 150, 200, 500$.
上方は ICBoot, 下方は ICBoot_{M} のバイアスである.

8.6 仮説検定手法との比較

Soromenho (1994) は, コンポーネント数の決定について以下に挙げる 5 種類の仮説検定の手法（以下単に検定手法）の比較をシミュレーションにより行っている. いま, ここで提案の方法とこれらの検定手法との比較のため, Soromenho と同一のパラメータ設定でシミュレーションを行った.

(1) Orlov の nested test の修正した方法 (Celeux (1987)),

(2) 尤度比検定統計量を修正した方法 (Wolfe (1970)),

(3) パラメータ空間の境界上での検定を避けるため, 混合比率で事前に積分する方法 (4 種類の事前分布を与えている, Aitkin and Rubin (1985)),

(4) SEM 法 (EM 法の Stochastic 版) による, F 分布に従う統計量を用いる方法 (Celeux (1986), Celeux and Diebolt (1985)),

(5) 尤度比検定統計量の漸近帰無分布をブートストラップ法により構成し, それを用いた検定法 (McLachlan (1987)).

なお, 原論文の Aitkin and Rubin の方法ではさらに 4 つの事前分布の設定があるため, 合計 8 種類についての比較を行っている. 検定の仮説は, 帰無仮説 $\mathrm{H}_0 : g = 1$ に対して 対立仮説 $\mathrm{H}_1 : g = 2$ とし, SEM 法を除いて有意水準を 5% としている. 実験は正しいコンポーネント数を推定した割合（正解率 (%)）で示している.

シミュレーションのパラメータ設定は, 標本数 $N = 50, 100, 200$ に対応してシミュレーション回数は $S = 100, 75, 50$ である. また, 2 つのコンポーネント分布の場合の混合比率は $\pi = (p, 1-p)^T$ で, p =0.5, 0.4, 0.3, 0.2, 0.1 の 5 種類である. さらに細かい設定は以下のとおりである.

設定 1: 1 変量, 1 つのコンポーネント分布

分散 $\sigma^2 = 1$.

設定 2: 1 変量, 2 つのコンポーネント分布 (等分散のとき)

分布間の距離 $d = 1, 2, 3$, 分散 $\sigma^2 = 1$.

設定 3: 1 変量, 2 つのコンポーネント分布 (等分散でないとき)

平均値 $\mu_1 = \mu_2$, 分散 $\sigma_1^2 = 1$, $\sigma_2^2 = 2, 3$.

設定 4: 2 変量, 1 つのコンポーネント分布

分散共分散行列 $\boldsymbol{\Sigma} = \boldsymbol{I}$(単位行列のとき).

設定 5: 2 変量, 2 つのコンポーネント分布 (等分散)

分布間の距離 $d = 2, 3$, 分散共分散行列 $\boldsymbol{\Sigma} = \boldsymbol{I}$.

設定 6: 2 変量, 2 つのコンポーネント分布 (等分散でないとき) [6]

平均値 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, 分散共分散行列 $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$.

これらの設定で, 通常のブートストラップ法を用いて, コンポーネント 数を$g = 1, 2$ としてコンポーネント数を推定し, 他の手法と比較した結果が表 8.13～ 8.17である (5 種類の検定手法の結果の数値は Soromenho (1994) から引用) . ブートストラップの回数は $B = 200$, シミュレーションの回数は Soromenho の設定と同じである. また, 等分散の設定に対しては, 等分散性とそうでないときの仮定をおいてパラメータ推定を行った. また, ICBoot_M は ICBoot とほぼ同じふるまいをするので, ここでは ICBoot_M は割愛する.

シミュレーションの結果は次のとおりである. 設定 1 と設定 2 の 1 変量の場合に共通して, ICBoot は他の検定手法の中の正解率の一番高い手法 (以下, 最適な検定手法) に対して, 多少正解率は低いようである. 設定 3 の 1 変量で分散が異なる場合, 標本数が $N = 200$ で ICBoot と最適な検定手法はほとんど同じ結果を与える. しかし, 標本数が小さくなるにつれて ICBoot も, 検定手法のいづれも正解率は低くなるが, 検定手法の急激な下がり方に対して, ICBoot はの変化は緩やかである. また, 設定 1 ～ 3 の AIC については, ICBoot とほぼ同じふるまいである.

[6] Soromenho の論文では $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ と記述しているが, これは $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$ の間違いと考えられる. 原著のとおりに数値実験を行なうと全く違う結果が得られる.

設定 4 の 2 変量で 1 コンポーネントの場合 (図 8.16), ICBoot と AIC は標本数に関係なくほぼ同じ割合の正解率であるが (ICBoot は 90% 前後で, AIC は 77% 前後), 他の検定手法は標本数が小さくなると同時に正解率も低下している. 設定 5 の 2 変量で 2 コンポーネントの場合は (図 8.17), 標本数が小さくなるにつれて ICBoot も検定手法も正解率は低くなる. さらに混合比率 p が小さくなるにつれて検定手法の正解率は低くなるが, ICBoot はほぼ同じ正解率を保っている. 設定 6 の 2 変量で 2 コンポーネント, 分散共分散行列が異なる場合では (図 8.17), 標本数が $N = 200$ のときに ICBoot と最適な検定手法とはほぼ同じである. 標本数と混合比率 p が小さくなるにつれて ICBoot も, 5 種類の検定手法も正解率は低くなるが, 検定手法は 1 桁から 0% までの割合まで低減するが, ICBoot は最低でも 20% 前後までしか下がらない. 設定 5, 6 に関しては, AIC は ICBoot よりその正解率の割合は常に大きい. これと前節のシミュレーションの結果を考え合わせると, やはりコンポーネント数を多めに推定していると考えられる.

全体を通して, 標本数や混合比率の変化に対する ICBoot のふるまいをみると, 従来の検定手法とほぼ同様であるが, 多少 ICBoot の方が正解率は高い. 1 変量で等分散とした設定では, ICBoot は最適な検定手法に比べて正解率は低いが, 分散が等しくないときは, ICBoot は正解率は高い. 一方, 多変量では相対的に ICBoot の正解率は高い. したがって, 実用的側面からは ICBoot のほうが適用範囲が広いようである.

表 8.13: 設定 1: モデルが一変量標準正規分布 ($r = 1$) のとき正しいコンポーネント数を当てた割合 (%)

	N		
検定手法	200^{*1}	100^{*2}	50^{*3}
ORLOV	72	64	50
WOLFE	86	82	70
A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	70	68	72
A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	60	62	62
A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	60	58	58
A-R empirical prior	76	72	74
SEM	96	94	90
MCLACHLAN	94	92	90
ICBoot$^{\dagger 1}$	82	91	90
AIC$^{\dagger 1}$	78	89	84
MDL$^{\dagger 1}$	100	100	98
ICBoot$^{\dagger 2}$	82	87	83
AIC$^{\dagger 2}$	74	81	80
MDL$^{\dagger 2}$	98	100	99

^{*1}: $B = 200, S = 50$, ^{*2}: $B = 200, S = 75$, ^{*3}: $B = 200, S = 100$, $^{\dagger 1}$: 等分散性の条件下でパラメータ推定を行った, $^{\dagger 2}$: 分散が等しくないという条件下でパラメータ推定を行った.

表 8.14: 設定 2: モデルが一変量標準正規分布 ($r = 2$) のとき正しいコンポーネント数を当てた割合 (%)

p	検定手法	$d=3$			$d=2$			$d=1$		
		N			N			N		
		200	100	50	200	100	50	200	100	50
0.5	ORLOV	96	92	83	64	48	45	12	10	8
	WOLFE	100	94	81	44	33	26	6	6	4
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	54	44	39	22	20	17	8	6	3
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	46	40	36	26	20	16	8	6	3
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	40	32	22	18	22	16	2	0	3
	A-R empirical prior	58	46	40	30	22	18	10	8	5
	SEM	96	89	59	58	41	20	48	21	12
	MCLACHLAN	98	90	59	38	30	12	6	4	6
	ICBoot	100	96	71	56	37	23	14	12	10
	AIC	100	99	82	56	47	32	10	16	14
	MDL	98	73	54	6	11	8	0	1	2
0.4	ORLOV	98	88	79	64	46	40	20	12	6
	WOLFE	100	94	77	52	30	17	8	6	4
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	50	40	38	20	18	19	8	4	3
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	46	38	37	30	16	18	6	4	2
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	58	30	21	20	20	16	4	2	0
	A-R empirical prior	60	44	38	32	22	19	12	8	7
	SEM	96	89	59	62	34	25	28	9	5
	MCLACHLAN	98	92	58	44	28	9	0	4	0
	ICBoot	100	96	66	50	40	21	16	9	8
	AIC	100	100	75	62	47	37	16	8	9
	MDL	96	81	50	8	11	9	0	1	3
0.3	ORLOV	100	93	75	74	48	36	18	16	6
	WOLFE	100	94	75	62	28	29	6	4	2
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	48	38	33	18	16	17	6	4	3
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	40	35	32	26	14	16	4	4	2
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	62	30	22	28	20	17	6	6	0
	A-R empirical prior	64	41	34	40	24	20	12	10	6
	SEM	98	86	49	72	33	22	34	10	9
	MCLACHLAN	98	93	53	56	22	14	4	6	4
	ICBoot	100	99	79	64	36	22	14	11	10
	AIC	100	100	86	70	55	29	12	11	16
	MDL	100	84	57	16	12	10	2	0	2
0.2	ORLOV	98	90	83	76	49	43	20	18	4
	WOLFE	100	94	82	64	38	31	4	3	2
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	44	32	26	12	17	13	2	2	2
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	36	30	26	20	16	13	2	2	1
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	68	28	16	30	22	14	8	5	0
	A-R empirical prior	68	34	29	42	22	18	10	6	4
	SEM	90	73	41	50	18	12	18	10	6
	MCLACHLAN	98	89	56	54	29	11	4	8	4
	ICBoot	100	96	78	76	59	32	8	11	9
	AIC	100	97	81	72	57	37	7	12	14
	MDL	98	81	60	16	21	11	1	3	2
0.1	ORLOV	94	84	67	74	46	34	16	14	3
	WOLFE	98	80	63	60	21	20	4	6	1
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	36	25	17	8	14	11	0	0	0
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	32	22	11	18	10	10	0	0	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	60	20	10	18	18	14	4	6	0
	A-R empirical prior	62	28	20	38	19	18	6	8	1
	SEM	76	42	14	42	6	7	4	5	2
	MCLACHLAN	96	69	37	52	12	10	4	5	1
	ICBoot	98	87	58	60	39	26	16	7	9
	AIC	100	89	65	58	39	23	18	7	16
	MDL	90	65	36	14	4	5	0	0	7

表 8.15: 設定 3: 正しいコンポーネント数を当てた割合 (%, 一変量, 等分散でなく, 等平均の二つの正規分布のとき)

p	検定手法	$\sigma_1 = 1; \sigma_2 = 3$			$\sigma_1 = 1; \sigma_2 = 2$		
		N			N		
		200	100	50	200	100	50
0.5	ORLOV	96	50	10	72	32	14
	WOLFE	96	48	6	72	26	8
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	50	18	6	34	12	2
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	44	14	4	30	10	2
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	38	10	2	29	6	0
	A-R empirical prior	54	18	8	36	14	4
	SEM	84	52	12	74	46	14
	MCLACHLAN	92	46	8	76	46	8
	ICBoot	100	93	70	86	53	46
	AIC	100	91	68	88	53	37
	MDL	92	61	35	14	8	9
0.4	ORLOV	88	48	8	70	30	10
	WOLFE	96	42	6	72	22	7
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	45	12	6	28	11	2
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	40	12	2	24	8	1
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	35	9	4	26	6	2
	A-R empirical prior	52	16	6	34	12	3
	SEM	88	50	10	72	42	8
	MCLACHLAN	92	44	6	73	44	8
	ICBoot	98	81	52	70	32	33
	AIC	98	80	56	72	33	27
	MDL	64	27	19	2	0	6
0.3	ORLOV	76	45	8	54	28	8
	WOLFE	88	38	4	56	20	4
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	42	10	4	24	5	0
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	35	8	1	20	2	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	34	8	2	22	4	1
	A-R empirical prior	44	12	5	30	8	1
	SEM	60	46	10	56	35	5
	MCLACHLAN	80	42	4	59	34	3
	ICBoot	86	52	54	34	27	30
	AIC	92	63	54	36	29	23
	MDL	28	13	18	2	3	5
0.2	ORLOV	40	18	4	32	10	4
	WOLFE	52	16	2	38	8	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	31	6	2	15	3	0
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	26	4	0	14	0	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	28	8	0	16	0	0
	A-R empirical prior	36	8	1	18	4	0
	SEM	32	26	5	39	14	4
	MCLACHLAN	52	22	0	15	2	2
	ICBoot	42	41	34	16	13	19
	AIC	56	48	37	24	24	27
	MDL	4	9	8	0	1	3
0.1	ORLOV	12	10	2	8	5	0
	WOLFE	24	10	1	12	4	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	10	2	1	4	1	0
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	8	2	0	0	0	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	4	2	0	0	0	0
	A-R empirical prior	10	4	1	6	2	0
	SEM	14	15	4	12	10	2
	MCLACHLAN	16	12	1	14	8	1
	ICBoot	28	12	26	10	13	18
	AIC	32	25	37	20	20	29
	MDL	0	3	6	0	0	3

表 8.16: 設定 4: 1 つの 2 次元正規分布のとき ($r = 1$) 正しいコンポーネント数を当てた割合 (%)

検定手法	N		
	200[*1]	100[*2]	50[*3]
ORLOV	52	44	20
WOLFE	54	48	17
A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	40	28	12
A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	38	22	10
A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	30	18	8
A-R empirical prior	46	32	14
SEM	96	84	52
MCLACHLAN	94	82	50
ICBoot[†1]	88	92	92
AIC[†1]	78	75	77
MDL[†1]	100	99	98
ICBoot[†2]	96	97	95
AIC[†2]	80	75	73
MDL[†1]	100	100	96

[*1]: $B = 200, S = 50$, [*2]: $B = 200, S = 75$, [*3]: $B = 200, S = 100$, [†1]: 等分散性の条件下でパラメータ推定を行った, [†2]: 分散が等しくないという条件下でパラメータ推定を行った.

表 8.17: 設定 5, 6: 2 次元正規分布のとき正しいコンポーネント数を当てた割合 (%)

		Homocodastic case $d=3$			Homocodastic case $d=\sqrt{8}$			Heterocodastic case		
		N			N			N		
p	検定手法	200	100	50	200	100	50	200	100	50
0.5	ORLOV	82	62	42	48	28	20	80	48	10
	WOLFE	82	68	49	78	38	20	68	32	4
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	52	32	16	32	21	8	38	16	7
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	42	27	12	28	18	6	35	12	4
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	38	2S	13	26	17	6	28	12	4
	A-R empirical prior	56	38	18	36	26	12	42	21	10
	SEM	96	82	51	92	70	21	81	52	12
	MCLACHLAN	94	84	50	90	68	20	82	48	8
	ICBoot	100	87	42	100	72	34	86	41	20
	AIC	100	100	75	100	88	57	98	65	64
	MDL	88	59	27	78	33	20	26	8	9
0.4	ORLOV	80	58	39	44	23	18	78	43	6
	WOLFE	78	62	44	72	34	14	64	24	2
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	50	28	12	30	17	6	36	10	5
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	36	26	10	22	15	4	31	8	2
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	36	24	12	24	14	4	24	14	3
	A-R empirical prior	54	34	16	33	24	8	38	18	7
	SEM	94	80	47	88	66	16	78	47	9
	MCLACHLAN	94	80	50	88	62	12	76	41	6
	ICBoot	100	85	53	100	72	33	80	32	17
	AIC	100	97	70	100	89	59	98	73	59
	MDL	94	63	31	84	40	17	30	8	8
0.3	ORLOV	72	52	32	40	20	13	54	28	2
	WOLFE	68	58	36	60	27	13	33	18	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	48	24	8	24	14	2	20	6	2
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	32	20	6	16	10	0	18	4	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	34	24	10	22	11	4	20	10	2
	A-R empirical prior	50	27	13	30	19	6	36	10	4
	SEM	90	75	45	80	60	12	62	32	6
	MCLACHLAN	88	78	48	82	59	8	68	28	2
	ICBoot	100	85	62	100	81	45	68	39	19
	AIC	100	96	21	100	92	67	94	68	62
	MDL	98	65	61	92	43	22	28	9	6
0.2	ORLOV	38	26	17	22	16	7	32	12	0
	WOLFE	42	34	22	38	18	10	20	10	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	30	16	6	20	8	0	18	2	1
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	24	14	4	12	6	0	10	1	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	22	18	8	12	8	0	15	4	0
	A-R empirical prior	35	20	10	25	15	4	24	6	2
	SEM	82	72	42	75	54	8	48	24	S
	MCLACHLAN	82	70	38	78	53	8	45	17	0
	ICBoot	100	91	59	96	88	44	56	25	22
	AIC	100	95	72	100	93	62	82	61	54
	MDL	100	67	35	90	55	26	4	7	9
0.1	ORLOV	33	12	8	10	8	4	28	8	0
	WOLFE	38	28	16	22	10	6	16	4	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.0$	22	10	2	12	4	0	10	0	0
	A-R $\alpha_1 = 0.5; \alpha_2 = 0.5$	18	8	0	8	3	0	6	0	0
	A-R $\alpha_1 = 1.0; \alpha_2 = 1.5$	10	6	0	6	2	0	4	0	0
	A-R empirical prior	26	16	6	16	6	1	19	1	1
	SEM	78	68	36	71	50	6	36	16	2
	MCLACHLAN	77	64	34	72	48	4	34	12	1
	ICBoot	98	87	60	94	80	55	40	21	21
	AIC	98	92	68	96	85	62	70	53	49
	MDL	78	52	32	70	37	23	4	3	6

8.7 まとめと考察

ブートストラップ法に基づいて対数尤度のバイアス補正を行うことによって得られる情報量規準が, 混合分布モデルのコンポーネント数の推定に適用可能かを調べることが本章の主な目的であった. 限られた設定での数値実験と 2 種類の実際のデータセットにより検証した結果から, 次の知見が得られた.

(1) 提案した混合分布モデルのコンポーネント数の推定手続きは有効に働くと思われる.

(2) 多変量の標本のとき, ブートストラップ法により構成された情報量規準は, 各種の検定手法に比べて良い結果が得られた.

(3) ブートストラップバイアス推定の変動減少法は, とくに混合分布モデルのコンポーネント数の推定において有効に働いている. これは EM 法の収束の遅さを補うという実用上の利点である.

今後の研究課題は, データの特性値が多数の多変量 (高次元) の場合等について検討を進める必要がある.

第 IV 部

データ解析

第 9 章

キバハリアリデータの混合分布モデルによる分類

本章は, キバハリアリというアリの部位計測データ[1]に第 III 部で提案した分類方式を適用した解析例を示す.

Ogata (1991) (緒方一夫氏, 九州大学熱帯農学センター) はキバハリアリ属 (*Myrmecia*) の種群レベルでの分類 (taxonomy) を行っている. オーストラリアにおいて野外採集を行ない, 主に Austrarian National Insect Collection, CSIRO での標本調査をした. キバハリアリ属はオーストラリア大陸とその周辺に生息するアリで, 現存する種として約 100 種以上が知られており, この属だけで一つのキバハリアリ亜科を構成する. 彼は 82 種以上, 252 個体以上のアリの形質特性データと計測データを採集して, 従来行われてきた形質特性に関する研究を再検討している. 種群の系統関係の検証は, 子孫形質の共有による分岐的関係を推定する方法論 (cladistic methodology) で行われ, その結果所与のデータセットを 9 つの種群に分類している. 専門的見地からの見解・意見やキバハリアリ属の特性については Ogata(1991), 緒方 (1995), Ogata and Taylor (1991) に詳しく記載されている.

ここではキバハリアリの計測データの分類を混合分布モデルにより行い, 分岐分類で行われた分類結果との対比を行うことが大きな目的である.

ここで行うデータ解析の目的は, 次のとおりである.

(1) キバハリアリの部位計測データを混合分布モデルにより分類すること.

(2) 主成分分析で次元縮約を行ない, キバハリアリデータの特徴抽出を行い, これを利用し

[1]ここで用いたキバハリアリのデータセットは統計数理研究所 共同研究 (6- 共研 A-57) でまとめられたものを用いた.

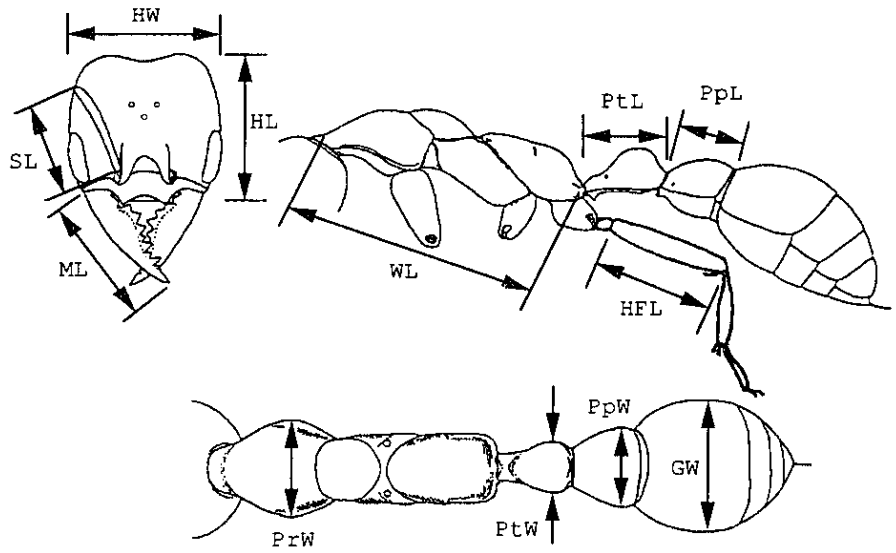

図 9.1: キバハリアリの計測部位

て分類に必要な変数の選択を行うこと.

(3) 混合分布モデルで分類を行なうとき, いくつのコンポーネント分布のモデルのあてはめが適当であるか検討を行なうこと.

9.1 キバハリアリの特徴

キバハリアリはオーストラリア大陸およびその周辺に分布（生息）するアリで, キバハリアリ属として知られている. このアリは, "bulldog ants" と呼ばれ, 多くのアリ研究者や生物学者の注目を集めている. その理由はキバハリアリがアリ科 (family Formicidae) の中でも祖先的な形質を多く持っていて, "原始的" なアリとされているためである. 現生種は約 100 種が知られていて, この属だけで 1 つの亜科（キバハリアリ亜科）を構成する. このアリの体長は 30mm を越える大変大きな種もあるが, それに比べて 8mm より小さい種もある. ほとんどの種は盛り上がった巣をつくり, 大きなものはその直径が 1m を越すものもある. いくつかの種は "jack jumper" と言われるように, 飛び跳ねることができる. ほとんどの種はかなり攻撃的で, 人が刺されるとアレルギーになったり, 死ぬことさえある. このアリの外見の特徴を緒方 (1995) から引用すると, (1) 非常に大きなあご, (2) 発達した複眼と単眼, (3) 明瞭に区別される胸部の亜区分, (4) 2 つの節からなる腹柄, (5) 強力で機能的な刺針,

表 9.1: 計測値の基本統計量

	HW	HL	SL	ML	WL	PrW	HFL	PtW	PtL	PpW	PpL	GW
個体数	247	242	233	233	114	238	125	240	235	245	234	245
欠測	5	10	19	19	138	17	127	12	17	7	18	7
平均	59.0	58.2	57.9	59.4	78.0	39.4	62.1	24.6	33.9	33.4	23.2	57.8
標準偏差	14.7	15.3	22.2	17.7	12.7	8.54	14.2	4.38	12.0	6.45	5.84	15.3
変動係数	24.9	26.2	38.3	29.8	32.2	21.7	22.9	17.8	35.4	19.3	25.2	26.5
最小値	20	23	19	21	41	15	29	9	14	15	10	22
最大値	95	90	103	100	97	60	95	33	37	48	42	93

HW: Head width (頭幅), HL: Head length (頭長), SL: Scape length (柄節長), ML: Mandible length (大あご長), WL: Weber's length of mesosoma (ウェーバーの胸長), PrW: Pronotal width (前胸幅), HFL: Hind femoral length (後脚腿節長), PtW: Petiole width (腹柄節幅), PtL: Petiole length (腹柄節長), PpW: Postpetiole width (後柄節幅), PpL: Postpetiole length (後柄節長), GW: Gastral width (腹部幅). 計測値は 20 で割ると mm の単位となる.

など特徴的な形態を有していて, 翅のないスズメバチといった印象を受ける.

9.2 データセットの特徴

キバハリアリのデータセットは 1 人の研究者 (緒方氏) がすべての個体を計測したデータであることから, 計測による誤差は小さいと考えられる. このデータセットは全部で 12 の特性 (変数) からなる. 付録 Dに原データの一覧を, 図 9.1に計測部位を, 表 9.1にこれらの特性の基本統計量を示す. また, 表 9.2に緒方氏が分岐分類で 9 つの種群に分類した際の名称と対応する種群の番号, および各種群の個体数を示す (Ogata (1991)). 以後この数字で "種群 {1,2}" のように用いる.

計測データの統計量は表 9.1に示す. 全個体数 $N = 252$ に対して, 部位 WL の欠測の個体数は 欠測数 $= 252 - 114 = 138$, 部位 HFL は 欠測数 $= 252 - 125 = 127$ と, 約半数の個体しか使えないので, 10 変数を対象データとする. その結果, 欠測値を考慮すると 200 ~ 230 程度の個体が利用可能である. 以降, "特性" のことを変数と呼ぶことにする.

次に, データセット全体の特徴を把握するため, 10 変数を用いて相関係数行列から主成分分析を行った. 第 1 主成分の寄与率は 85.3%, 第 2 主成分は 10.6%, 第 3 主成分は 1.3% である. 第 1 ~ 3 主成分スコアの散布図を図 9.2に示す. データセットの基本的な特徴は図 9.2(a) のように, A の細長い楕円体の部分, C の平たい楕円体の部分, 図 9.2(b) の B の飛び出た部

表 9.2: 各種群の計測した種数と個体数

番号	種群の名称	種数	個体数	原変数の選択 HL, SL, PpW	比率変数の選択 MI, PpI, PtW/PtL
1	*aberrans* group	6	14	14	13
2	*cephalotes* group	3	12	8	8
3	*gulosa* group	37	123	106	107
4	*mandibularis* group	7	17	17	17
5	*nigrocincta* group	3	9	8	8
6	*picta* group	2	6	6	5
7	*pilosula* group	15	42	38	37
8	*tepperi* group	5	18	17	17
9	*urens* group	4	11	11	10
合計		82	252	225	222

分から成る．A を構成する主な種群は {2, 4, 6, 7, 8, 9} である．それの主要部分である {2, 4, 7, 8} の 4 種群は混在して，1 つの楕円体を構成している（図 9.3(a)）．また，これらの各種群ごとに散布図で表示すると，ほぼ一直線上に並ぶことが確認できる．以上のことから，第 1 主成分はほぼ “サイズに関する因子”(個体の大きさを表す指標) であることがわかる．

9.3 計測値 (原変数) を用いた分類

9.3.1 変数の選択

提案分類方式をこのデータセットに適用するにあたって，10 変数の中から変数選択を行った．その主な理由としては次のことが挙げられる．

(1) 仮にすべての種群を識別するとした場合，1 種群あたりの最小個体数が 6 であるので，2 ～ 4 程度の変数で行わなければならない（全変数を使うと 1 群あたりの情報が足りない）．

(2) 2 ないし 3 個の変数を用いると散布図の観察が容易になる．

(3) 変数の数とグループ数が多くなると，推定すべきパラメータ数が増大し，混合分布モデルの局所解が増し，大域的最適解の推定が困難になる．

これらの理由から，3 個程度の変数を選択するのが適当と判断した．

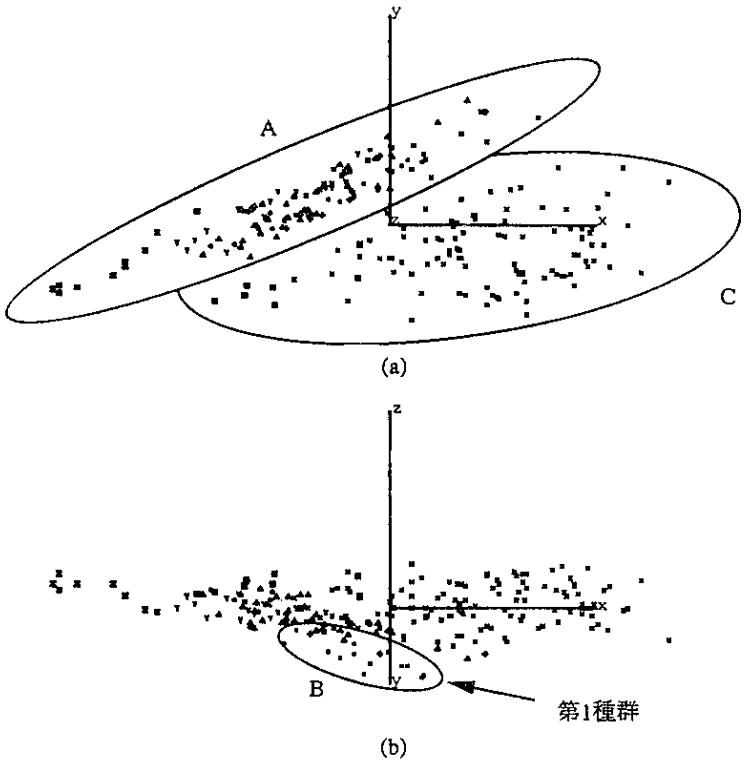

図 9.2: 原変数からの第 1 ～ 3 主成分スコアの散布図

回帰分析や判別分析などにおける変数選択の問題は様々な方法が提案されている．とくに主成分分析に基づいた方法では Jolliffe (1972, 73) や, GCD(Generalized coefficient of determination) による前向き選択の手続き (Yanai (1980)) 等がある．Jolliffe(1972, 73) に従うと変数選択の方法は次のような種類が考えられる (竹内, 竹村 (1986)).

(1) 回帰分析に基づく方法 (他の変数への回帰によって説明できる変数は取り除く),

(2) 主成分分析による方法 (小さな固有値の主成分と高い相関を持つ変数は取り除く),

(3) 変数のクラスタリングによる方法 (主に相関係数に基づいて類似変数を分類し, 各グループから一つづつの変数を残す)

すでに種々の実験結果から見たように, 混合分布モデルに基づく分類方式では数個の変数 (2, 3 個) を用いることが, もっとも効果的である．このことから, ここで用いる変数選択

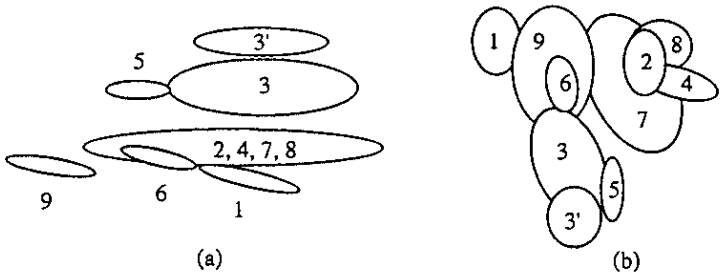

図 9.3: 原変数と比率変数のデータ構造の簡略図

(a) 原変数, (b) 比率変数. 番号は主群を示す.

の方針として, 上述の 10 変数を用いた主成分スコアの散布図の観察で得られたデータ構造の特徴をよく代表する変数を選ぶ, という考え方を採用する. 具体的には次の手順で行う.

ステップ 1 相関係数行列に基づく主成分分析を行い, 因子負荷量を計算する (表 9.3).

ステップ 2 相関係数行列を観察して, 相関が低い変数を選ぶ (サイズの因子の影響が見られることから, 変数相互の相関はいづれも大きいので, 相対的に小さいものを目安として選別する).

ステップ 3 第 1 ～ 3 主成分の因子負荷量の散布図を描き, 変数の関係を観察する.

⟶ 図 9.4の第 2 と第 3 主成分の組み合わせの散布図に注目すると 2 つのグループとそれ以外の散らばっている様子が観察できるので, 次のようにグループ化する. {SL, PtL, ML}, {HL, GW, HW, PrW}, そして {PpL, PtW, PpW}.

ステップ 4 サイズの因子の影響がなるべく小さい変数を選ぶ. すなわち第 1 主成分の因子負荷量の絶対値が相対的に小さいものを選ぶ. そして, サイズの因子以外の因子と相関が高い変数を選ぶ. つまり, 第 2,3 主成分に対する因子負荷量が大きい変数を選ぶ.

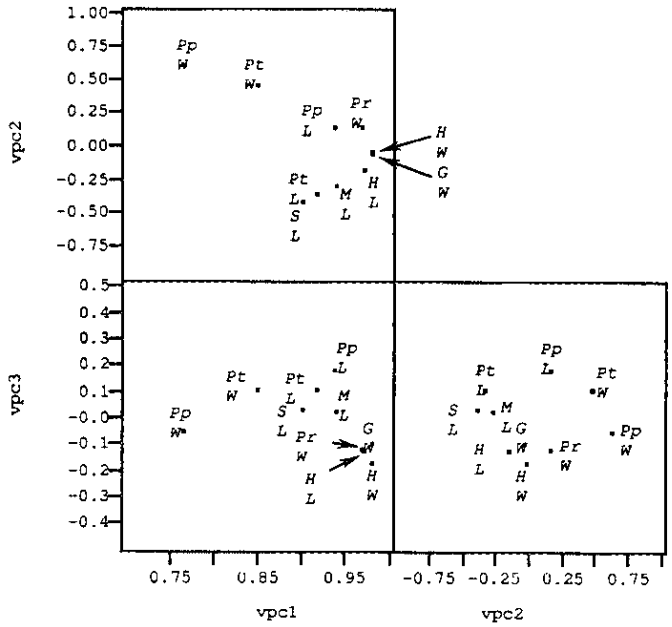

図 9.4: 原変数での因子負荷量の散布図

⟶ PpW が第 1 主成分に対しては値が一番小さく, 第 2 主成分に対しては一番大きい.

ステップ **5** ステップ 2, 4 で注目した変数の中から候補を選ぶ.

⟶ {SL, PtL, ML} から SL と ML を, {HL, GW, HW, PrW} から HL を, そして {PpL, PtW, PpW} から PtL と PpW を選んだ.

ステップ **6** 第 1 ～ 3 主成分スコアの散布図の布置と, 表 9.1の変動係数の値 (なるべく大きい値) を参考に, 3 つの変数を選ぶ.

⟶ HL, SL, PpW を選択.

この結果, 各種群の有効個体数は表 9.2(原変数の選択) のようになった.

選択した変数について散布図を描き, データ構造の観察を行うと, 次のような特徴が見られる (図 9.5参照).

- 細長い楕円体 (A) とその脇にある部分 (B), そして円盤が 1/4 ほど欠けた領域 (C) がある (図 9.5(a),(b)).

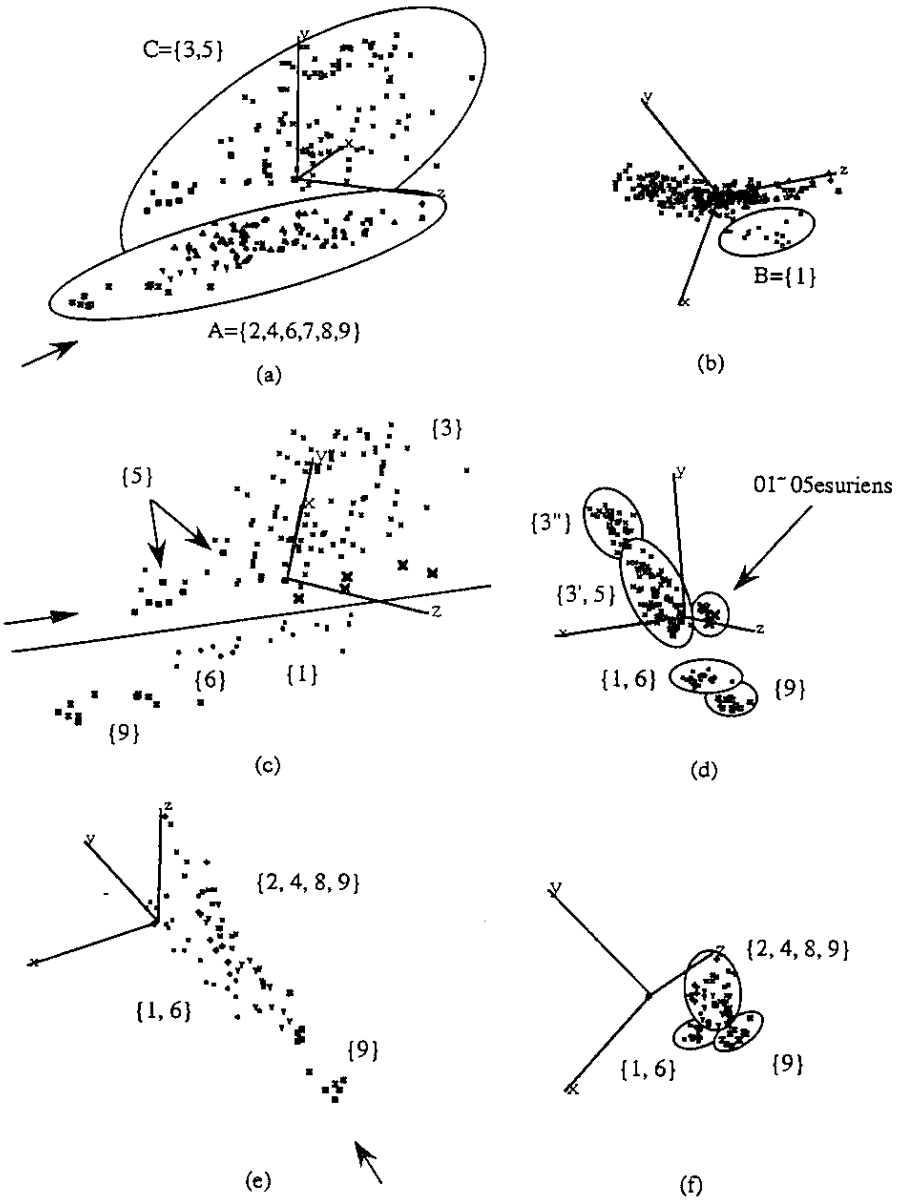

図 9.5: 原変数の散布図
(a) 2 群に見える方向から, (b) (a) の矢印の方向から見たもの (第 1 種群が飛び出ている), (c) {1, 3, 5, 6}+{9} 種群の散布図, (d) (c) の矢印の方向から見たもの, (e) {1, 2, 4, 6, 8, 9} 種群の散布図, (f) (e) の矢印の方向から見たもの.

表 9.3: 原変数の相関係数行列と因子負荷量行列

Variable	HW	HL	SL	ML	PrW	PtW	PtL	PpW	PpL	GW
HW	1.0000									
HL	0.9776	1.0000								
SL	0.8791	0.9293	1.0000							
ML	0.9226	0.9457	0.9582	1.0000						
PrW	0.9631	0.9238	0.8024	0.8644	1.0000					
PtW	0.8020	0.7344	0.5882	0.6751	0.8779	1.0000				
PtL	0.8812	0.9305	0.9710	0.9588	0.8224	0.6374	1.0000			
PpW	0.7397	0.6432	0.4400	0.5375	0.8383	0.9295	0.4837	1.0000		
PpL	0.8965	0.8770	0.7837	0.8347	0.9169	0.8498	0.8109	0.7970	1.0000	
GW	0.9648	0.9630	0.9049	0.9188	0.9436	0.8047	0.9030	0.7264	0.8910	1.0000
vpc1	0.9808	0.9717	0.9020	0.9394	0.9700	0.8496	0.9164	0.7647	0.9376	0.9800
vpc2	-0.0184	-0.1590	-0.4048	-0.2871	0.1563	0.4687	-0.3490	0.6290	0.1481	-0.0435
vpc3	-0.1583	-0.1134	0.0468	0.0384	-0.1077	0.1257	0.1194	-0.0391	0.1999	-0.0838
vpc4	-0.0458	-0.0489	0.0499	0.0402	-0.0297	0.1793	0.0758	0.0117	-0.2409	0.0343
vpc5	-0.0514	0.0028	0.0728	-0.1412	-0.0570	-0.0353	0.0268	0.0436	0.0112	0.1342
vpc6	-0.0424	-0.0850	0.0242	0.0588	0.0458	-0.0934	0.0353	0.1017	-0.0203	-0.0092
vpc7	0.0016	0.0424	-0.0431	-0.0794	0.0469	-0.0111	0.1208	0.0095	-0.0165	-0.0692
vpc8	0.0344	0.0535	-0.0048	0.0272	-0.1190	-0.0168	0.0132	0.0728	-0.0093	-0.0372
vpc9	0.0192	-0.0008	0.1000	-0.0367	0.0172	0.0102	-0.0367	0.0118	-0.0046	-0.0716
vpc10	-0.0676	0.0582	0.0037	0.0147	0.0153	-0.0008	-0.0193	0.0086	-0.0048	-0.0060

- A の領域は約 6 種群が含まれている. → 種群 {2, 4, (6), 7, 8, 9}
- B は約 2 種群が含まれている. → 種群 {1, (6)}
- C は 2 種群が含まれている. 種群 {3, 5}　(種群 {6} は種群 {1} の延長上にあると見ることができる. あるいは A の一部とも見ることもできる)
- 種群 {3} はさらに 3 つに分かれているようである. 図 9.5(c),(d) にその状況を示す. 図中の太い×印は 3 から多少離れている 5 つの個体である (01 ~ 05 esuriens 種).
- 種群 {5} は種群 {3} の一部分の延長上にある (図 9.5(b)).
- B の種群 {1, 6} はほぼ一直線上に布置しているが, これらの種群間は重なっていない (図 9.5(c)).
- A の領域は各種群が入り組んでいる. とくに種群 {7} は A の大半を占め, これを取り除いて表示したのが図 9.5(c) ~ (f) である.
- 各種群はほぼ一直線上に並んでいる (図 9.5(e), (f)).

9.3.2 提案分類方式による分類

コンポーネント数を 2 ～ 7 としたときの分類結果を, 表 9.4に示す. これは, Ogata (1991) による分岐分類により得られた 9 種群の結果と, 混合分布モデルによる分類結果とをクロス集計したものである. なお, 表中の "AR" は判別率の推定値である.

分類結果 ($r = 2$ 群のとき)

コンポーネント数を 2 として分類を行ったときの最適解は,

$$\{\{1,3,5,6\},\{2,4,7,8,9\}\} \Longrightarrow \{\{B,C\},A\}$$

のようになった. この分類結果は次のような特徴がある.

(1) 種群 {1, 3, 5, 6} が 1 つのコンポーネント分布を構成しているが, 種群 {3, 5} と種群 {1, 6} の種群間は離れている (図 9.5(c)) . ここで, 種群 {3, 5} の C の領域のかけた部分を補う形で種群 {1, 6} があるため (図 9.5(d)) , これらを 1 コンポーネントとしてとらえた. 種群 {1, 3, 5, 6} の間には, 一平面に乗るような何らかの関係があると考えられる.

(2) 種群 {2, 4, 7, 8, 9} が 1 つの細長い楕円体のコンポーネント分布となっている.

また, 局所解の一つとして

$$\{\{3,5\},\{1,2,4,6,7,8,9\}\} \Longrightarrow \{C,\{A,B\}\}$$

のように分類された. 局所解は散布図で観察したとおりの分類のされ方である (図 9.5(a)). 種群 {3, 5} は分岐分類で最も近いとされている二つの種群で, これが 1 コンポーネントとして推定されている.

分類結果 ($r = 3$ 群のとき)

コンポーネント数が 3 のときの最適解は,

$$\{\{1,6\},\{3,5\},\{2,4,7,8,9\}\} \Longrightarrow \{B,C,A\}$$

のようになった. 最適解は2群のときの円盤状のコンポーネント分布を (種群 {1, 3, 5, 6}), 2つに分ける形で分類された. これは, 原変数の主成分スコアで観察された, {A, B, C} の部分をとらえている.

分類結果 ($r = 4$〜6 群のとき)

4 〜 6 群の分類結果は次のようになった. $r = 4$ では $r = 3$ の分類の C 部分ががさらに2つに分かれた. $r = 5$ とすると A がさらに2つに分かれ, $r = 6$ では C がさらに2つに分かれた.

コンポーネント数が2のときの最適解は, いくつかの種群が同一平面に分布する構造の特徴があると推測される. また, 局所解はおおむねデータの目視と一致しているのが特徴である.

表 9.4のクロス表にみるように, コンポーネント数が増えてくると, とくに種群 {2, 4, 6, 7, 8} の所属が複数のクラスターにまたがるようになり, 曖昧になってくる. それはこれらの種群が互いに入り組んでいるためである.

9.4 比率変数を用いた分類

前節の考察から, いくつかの種群が入り組んでいることや, 変数間にサイズの因子が含まれていることがわかった. そこでサイズの因子を取り除くために原変数を比率変数として加工し, 新たな変数で分類を試みる.

これについてはすでに5種類の加工変数が Ogata により作成されていて, 各々に指標としての名称が与えられている（表 9.5上段）. ここではさらに5種類の変数を作成した（表 9.5下段）. 表 9.6に比率変数の統計量を示す. このうち, 指標 "LI" には欠測値が約半数あるので, これを除き, 9 変数を分析対象とする.

これらの 9 変数から相関係数行列を計算し, そこから主成分分析を行い, 原変数と同様な手続きで変数選択を行った.

相関係数行列と因子負荷量行列を表 9.7に, 因子負荷量の散布図を図 9.6に示す. これらの図表から, MI と PpI の2つの変数を中心に, 第3の変数を選ぶことにした. その理由とし

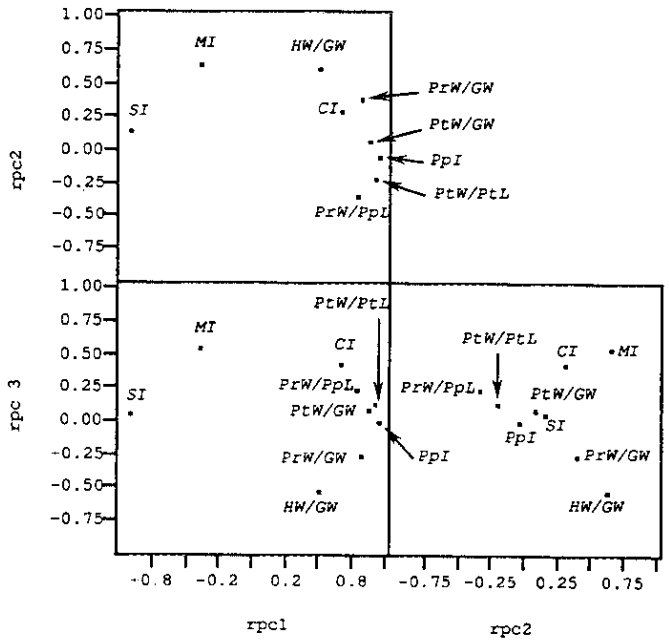

図 9.6: 比率変数での因子負荷量の散布図

て, MI は第 1 主成分に対する相関が一番低く, 第 2, 3 主成分に対しては一番高いこと, PpI は第 1 主成分に対して一番相関が高いことが挙げられる. 選択の方法は MI, PpI との組み合わせで, 変動係数が大きい変数を選んだ. その結果, MI, PpI, PtW/PtL を選択した.

選択した変数に基づく散布図のデータ構造には次のような特徴がある.

- 種群 {1} は他の種群からは離れている.
- 種群 {3, 5} はほぼ 1 かたまりである. 種群 {5} の脇に種群 {3} が位置する.
- 種群 {2, 4, 7, 8} がほぼ 1 かたまりであるが, 原変数の場合と比べると各種群の領域が良く分離している (図 9.3(b)) .
- 種群 {6, 9} は種群 {1}, {3, 5}, {2, 4, 7, 9} の間に位置する. とくに種群 {9} は分散が大きく, これらの 3 つとオーバーラップしている (図 9.3(b), 図 9.7, z 印は {9} 種群, ◆印は {6} 種群) .

次に, 分類結果を表 9.8に示す. これは表 9.4と同様な方法によるクロス表である. コンポーネント数を 2 ～ 7 として分類すると次のような結果が得られた.

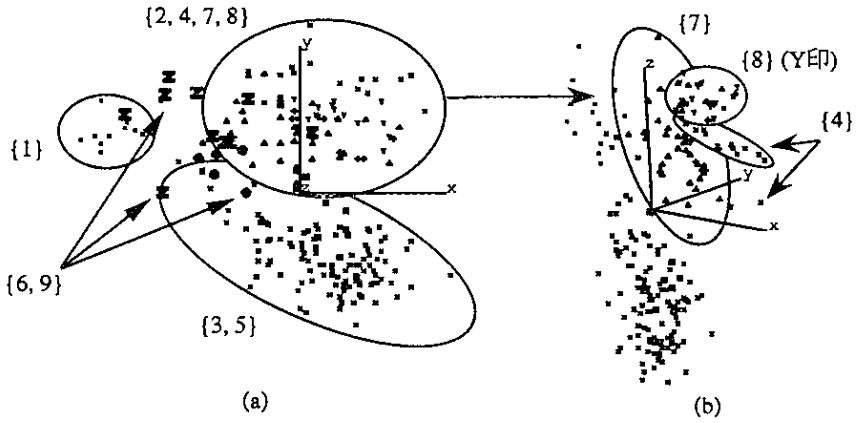

図 9.7: 比率変数の散布図

- 2 群: {{1, 2, 4, 6, 7, 8, 9}, {3, 5}}
- 3 群: {{1, 9'},{2, 4, 6, 7, 8, 9"}, {3, 5}}
- 4 群: {{1, 9'}, {2, 4, 6, 7, 8, 9"}, {3, 5'}, {5"}}
- 5 群: {{1, 9'}, {7', 8}, {2, 4, 6, 7", 9"}, {5"}, {3, 5'}}
- 6, 7 群: 各種群が複雑に入り組んでいる.

ここで, クォーテションマーク (') とダブルクォーテションマーク (") はある種群が 2 つ以上に分類されたことを示す.

これらの分類結果と, 散布図の観察から得られた種群 {6, 9} が 3 つの種群 {1}, {3, 5}, {2, 4, 7, 8} の分布の延長上の共通部分であるという知見を参考に, 種群 {6, 9} を取り除いて分類を行った. このときコンポーネント数を 2 ～ 6 として分類した結果は次のようになった (表 9.9) .

- 2 群: {{1, 2, 4, 7, 8}, {3, 5}}
- 3 群: {{1},{2, 4, 7, 8}, {3, 5}}
- 4 群: {{1}, {2, 4, 7', 8'},{7", 8"}, {3, 5}}

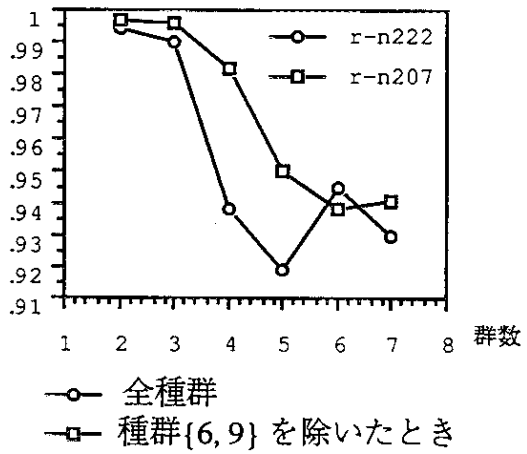

図 9.8: 比率変数での各コンポーネント数の判別率の値

- 5 群: {{1}, {2, 4, 7', 8'},{7", 8"}, {3'}, {3", 5}}
- 6 群: {{3", 5}, {1}, {2', 4, 7', 8'}, {3'}, {2", 7", 8"}, {2"', 7"'}}

この分類結果は 2 ～ 5 群までは, ほぼ入れ子の形に分類されていることが特徴である.

比率変数による分類結果から, 次のような知見が得られた.

- 種群 {6, 9} を取り除いても, 種群 {2, 4, 7, 8} はきれいに分類されず, これらの種はほぼ 1 かたまりと認識された.
- 原変数から得られた主成分スコアのデータ構造の観察で, 種群 {3} の 5 個体 (01 ～ 05 esuriens 種) がこの種群から離れていたが, 比率変数によるとさらに離れて, 種群 {6, 9} とほぼ同じ位置にある (図 9.7(b)). 結果として, これらの 5 個体は {3, 5} 種群と同じ群には分類されていない.

次に, 判別率の値で分類結果を比較してみると, 次のようになる.

- $r = 2$～5, 7 群では種群 {6, 9} を取り除くと判別率が高くなる (図 9.8) .
- 種群 {6, 9} を取り除いた分類の場合, $r = 7$ で判別率がわずかに増加している. これは所属個体数の小さいコンポーネント分布が生じるためである.

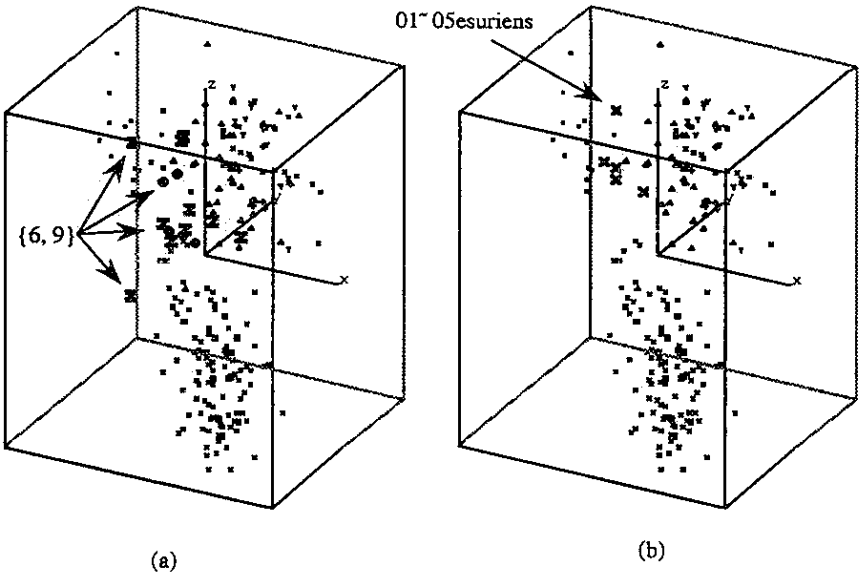

図 9.9: 比率変数の散布図

(a) 全種群のとき, (b) {6, 9} 種群を取り除いたとき.

9.5 コンポーネント数の推定

第 III 部で提案した混合分布モデルのコンポーネント数の推定手続きを, 比率変数に対して適用してみた.

比率変数を用いたときのコンポーネント数の推定結果を表 9.10と図 9.10(a) に示す. ここに見るように, 推定されたコンポーネント数は 7 であった (* 印). 一方, 種群 {6, 9} を除去したときのコンポーネント数の推定結果を表 9.11, 図 9.10(b) に示す. これらの二種群をとり除いて推定されたコンポーネント数は 5 であった. とり除く前の推定値は 7 であるから, 二種群を取り除いた結果, 数字の上では $7-2=5$ となるが, これらの分類のされ方は全く異なることがわかる (図 9.10を参照). これは種群 {6, 9} の影響がなくなることによって, ある程度まとまっている種群の特徴が出てきたためである.

9.6 考察

まず, 解析結果全体を通して得られた総合的知見は次の二つである.

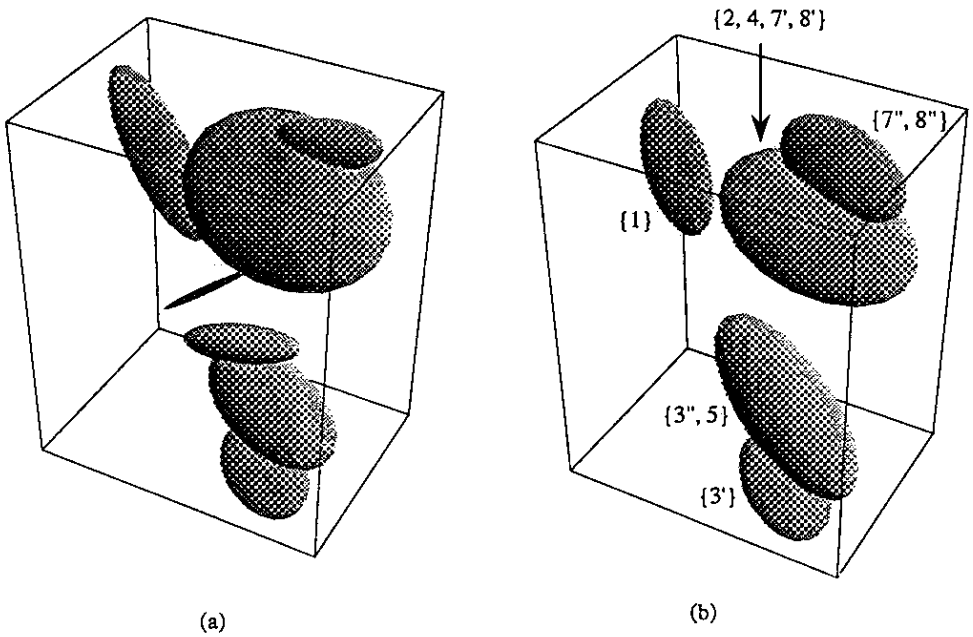

図 9.10: 比率変数に混合分布モデルをあてはめた結果

(a) 全種群 (7 群), (b) {6, 9} 種群を取り除いた (5 群).

- 種群 {1} は他の種群からは離れている,
- 種群 {3} は 3 つに分かれると考えるのが妥当と思われる.

さらに, 原変数による分類結果からの知見は,

- 2 群のとき種群 {3, 5} と種群 {1, 6} は離れているが, 種群 {3, 5} の円盤 (平面) の延長上に種群 {1, 6} が位置する.
- 種群 {1, 3, 5, 6} は一平面上に乗るような何らかの関係があるのか.
- 種群 {2, 4, 7, 8} は互いに入り組んでいて分離することは困難である.
- 種群 {1, 6} がほぼ 1 直線上に分布し, これは 3 群で分類したとき, 1 つのコンポーネント分布に相当する.

また, 比率変数による分類結果からの知見はとして次のことがある.

- 比率変数を用いると, サイズの因子を取り除くという意味で効果があった.

- 種群 {6, 9} を取り除いて分類すると, 3 つの種群 {1}, {3, 5}, {2, 4, 7, 8} はよく分かれる. その結果, コンポーネント数ごとの判別率の値も相対的に大きくなっているので, できたクラスターはよりまとまりが良いと考えられる.

おわりに, 提案した分類方式による分類結果とコンポーネント数の推定結果 (以下, あわせて提案手続きと言う) を緒方氏の分岐分類の結果 (Ogata(1991)) と比較すると, 次のように要約される.

まず, 分岐分類で推定された系統関係は次のとおりである.

$$\{\{\{\{\{3, 5\}, 6\}, 9\}, 2, \{\{4, 8\}, 7\}\}, 1\}$$

分岐分類では {3, 5} は近い種群とされたが, 提案手続きによると {{3', 5}, {3"}} として分類された. とくに, 最初の下線部について比較すると, 比率データの場合には散布図上で種群 {3, 5} の延長上に種群 {6, 9} があり, さらに種群 {6, 9} とほぼ同じ位置に esuriens 種の 5 個体がある. 一方, 次の下線部は, 原変数, 比率変数で共にほぼ 1 かたまりになっている. 比率変数では, この部分は {2, 4, 7', 8'} と {7", 8"} の, 2 つの群に分類された ({6, 9} を除いたとき).

9.7 おわりに

今後の研究課題としては, 次の 3 点が挙げられる.

(1) 比率変数で主成分分析を行い, そのスコアデータを使った分類と, ここで行った比率変数による分類の比較・検討.

(2) 種群 {2, 4, 7, 8} の分離は可能であるか, あるいは変数の加工, 次元縮約の別な方法を検討し, さらに解析を進めること.

(3) 分岐分類に用いられた形質に基づく分類法と, ここで提案した分類方式との関連性や, 二値変数と連続変数をあわせた混合分布モデルには興味がある.

最後に, 解析結果の考察について 3 人の生物学者[2]から次のような意見が出された. この分析結果から得られた群（クラスター）の情報は, データの予備解析として有用であるとのコメントを得ている. とくに解析に用いる変数（特性）の選択, データの加工（比率変換など）を含めて, ここで提示した分類方式は, 形質に基づく伝統的な分岐分類, 系統解析等に先立つ事前処理法として利用できる. 種群 {1} が他の種群から離れているという知見に対しては, Emery (1911) [3]がキバハリアリ属に 3 亜種を認め, その一つを Promyrmecia 亜属としたことと一致する. また, 種群 {6, 9} を取り除いたことによって種群 {3, 5} が他と分かれているのは, Clark (1951)[3] がキバハリアリ類を 2 属に分けて, 一つを *Myrmecia* としたこととほぼ一致する. 種群 {6, 9} については次のような意見が出された. 「一つの仮説として種群 {6, 9} は新しいニッチ (生態的地位) に進出した後, 形態を様々に広げたために他の種群とまたがるような計測値の分布をしたということが考えられる. しかし, Ogata (1991) の分岐分類によれば, それらよりも後に種群 {3, 5} が分岐したと推定されている (Ogata (1991), Fig.51). この食い違いは表型分類と分岐分類の接点となるところであり, より詳細な分析を必要とする今後の研究課題である.」また, 分岐の系統関係は分子 (遺伝子) レベルでの変化が作用しているので, この方法に基づく比較・検討は必要とされる (三中 (1995)).

[2]統計数理研究所 共同研究 (6- 共研 A-57) のメンバーの小野山 敬一氏 (帯広畜産大学), 三中 伸宏氏 (農業環境技術研究所), 寺山 守氏 (東京大学教養学部).

[3]この論文は Ogata (1991) を参照.

表 9.4: 原変数での分類結果 ($r = 2 \sim 7$)

$r = 2$, 最適解

	1	2	Sum
1	0	14	14
2	8	0	8
3	5	101	106
4	17	0	17
5	0	8	8
6	2	4	6
7	36	2	38
8	17	0	17
9	11	0	11
Sum	96	129	225
AR	0.983	0.991	0.988

$r = 2$, 準最適解

	1	2	Sum
1	14	0	14
2	8	0	8
3	5	101	106
4	17	0	17
5	0	8	8
6	6	0	6
7	38	0	38
8	17	0	17
9	11	0	11
Sum	116	109	225
AR	0.999	0.984	0.992

$r = 3$

	1	2	3	Sum
1	13	0	1	14
2	0	0	8	8
3	0	101	5	106
4	0	0	17	17
5	0	8	0	8
6	0	0	2	6
7	0	1	37	38
8	0	0	17	17
9	3	0	11	11
Sum	17	110	98	225
AR	0.961	0.973	0.963	0.968

$r = 4$

	1	2	3	4	Sum
1	13	1	0	0	14
2	0	8	0	0	8
3	0	5	38	63	106
4	0	17	0	0	17
5	0	0	0	8	8
6	4	2	0	0	6
7	0	38	0	0	38
8	0	17	0	0	17
9	0	11	0	0	11
Sum	17	99	38	71	225
AR	0.976	0.993	0.996	0.991	0.992

$r = 5$

	1	2	3	4	5	Sum
1	13	1	0	0	0	14
2	0	8	0	0	0	8
3	0	0	75	31	0	106
4	0	10	0	0	7	17
5	0	0	1	7	0	8
6	2	4	0	0	0	6
7	0	28	0	0	10	38
8	0	5	0	0	12	17
9	0	0	0	0	11	11
Sum	15	56	76	38	40	225
AR	0.977	0.912	0.948	0.952	0.931	0.938

$r = 6$

	1	2	3	4	5	6	Sum
1	13	1	0	0	0	0	14
2	0	8	0	0	0	0	8
3	0	0	38	57	11	0	106
4	0	10	0	0	0	7	17
5	0	0	0	1	7	0	8
6	2	4	0	0	0	0	6
7	0	27	0	0	0	11	38
8	0	5	0	0	0	12	17
9	0	0	0	0	0	11	11
Sum	15	55	38	58	18	41	225
AR	0.979	0.904	0.996	0.951	0.957	0.944	0.947

$r = 7$

	1	2	3	4	5	6	7	Sum
1	13	1	0	0	0	0	0	14
2	0	8	0	0	0	0	0	8
3	0	0	38	33	34	0	1	106
4	0	10	0	0	0	7	0	17
5	0	0	0	1	1	0	6	8
6	2	4	0	0	0	0	0	6
7	0	27	0	0	0	11	0	38
8	0	5	0	0	0	12	0	17
9	0	0	0	0	0	11	0	11
Sum	15	55	38	34	35	41	7	225
AR	0.979	0.904	0.998	0.933	0.972	0.944	1.000	0.949

表 9.5: 比率変数

指標名	加工処理	有効個体数 N
Cephalic index (CI)	HW/HL*100	242
Scape index (SI)	SL/HW*100	233
Mandibulo-cephalic index (MI)	ML/HL*100	233
Hind femoral index (LI)	PpW/GW*100	124
Postpetiolar index (PpI)	HFL/PrW*100	244
	PtW/PtL*100	235
	PpW/PpL*100	233
	PtW/GW*100	237
	PrW/GW*100	235
	HW/GW*100	240

表 9.6: 比率変数の基本統計量

	CI	SI	MI	LI	PpI	PtW/PtL	PpW/PpL	PtW/GW	PrW/GW	HW/GW
個体数	242	233	233	124	244	235	233	237	235	240
欠測	10	19	19	128	8	17	19	15	17	12
平均	101.8	97.8	102.6	179.7	59.8	68.4	129.9	38.4	69.5	103.1
標準偏差	5.9	20.1	9.9	33.2	10.1	16.6	17.0	5.8	6.2	7.1
変動係数	5.8	20.6	9.6	18.5	16.9	24.3	13.1	15.1	8.9	6.9
最小値	80.0	61.5	72.6	130.3	40.0	38.6	94.1	26.0	54.4	80.0
最大値	117.8	141.4	123.5	296.3	82.9	103.6	182.6	53.7	87.8	129.2

表 9.7: 比率変数の相関係数行列と因子負荷量行列

	Index	CI	SI	MI	PpI					
	Variable	HW /HL	SL /HW	ML /HL	PpW /GW	PtW /PtL	PpW /PpL	PtW /GW	PrW /GW	HW /GW
比率変数	CI=HW/HL	1.0000								
	SI=SL/HW	−0.5812	1.0000							
	MI=ML/HL	0.0606	0.5254	1.0000						
	PpI=PpW/GW	0.5380	−0.8624	−0.3697	1.0000					
	PtW/PtL	0.5968	−0.8622	−0.4179	0.8815	1.0000				
	PpW/PpL	0.5470	−0.7238	−0.3894	0.7184	0.7615	1.0000			
	PtW/GW	0.4926	−0.7326	−0.1900	0.8971	0.8464	0.5872	1.0000		
	PrW/GW	0.4879	−0.6797	−0.1845	0.7825	0.6283	0.4227	0.7357	1.0000	
	HW/GW	0.3622	−0.4176	−0.1058	0.3899	0.2302	0.1304	0.3585	0.6804	1.0000
比率変数の主成分	rpc1	0.6672	−0.9276	−0.4115	0.9527	0.9237	0.7773	0.8764	0.8127	0.4932
	rpc2	0.3049	0.1492	0.6596	−0.0399	−0.2085	−0.3383	0.0783	0.3954	0.6284
	rpc3	0.4551	0.0833	0.5704	0.0112	0.1577	0.2627	0.1043	−0.2371	−0.5087
	rpc4	−0.4455	0.1130	0.1922	0.2037	0.0908	−0.2098	0.3987	0.1630	−0.2345
	rpc5	−0.2049	0.1077	0.1164	0.0273	−0.1234	0.4082	−0.0367	0.0393	0.1122
	rpc6	−0.0435	−0.0138	0.0277	−0.0276	0.0749	−0.0080	0.1679	−0.3044	0.1847
	rpc7	0.0943	0.2858	−0.1307	−0.0323	0.0387	0.0251	0.1007	0.0365	0.0124
	rpc8	−0.0584	−0.0000	0.0395	−0.1631	0.2007	0.0151	−0.0513	0.0596	0.0201
	rpc9	−0.0130	0.0646	0.0201	0.1415	0.0872	−0.0278	−0.1169	−0.0361	0.0301

表 9.8: 比率変数での分類結果 (g=2 ～ 7)

$r = 2$, 最適解

	1	2	Sum
1	13	0	13
2	8	0	8
3	5	102	107
4	17	0	17
5	0	8	8
6	4	1	5
7	36	1	37
8	17	0	17
9	10	0	10
Sum	110	112	222
AR	0.992	0.996	0.994

$r = 3$

	1	2	3	Sum
1	13	0	0	13
2	0	8	0	8
3	0	5	102	107
4	0	17	0	17
5	0	1	7	8
6	0	5	0	5
7	0	36	1	37
8	0	17	0	17
9	3	7	0	10
Sum	16	96	110	222
AR	1.000	0.988	0.990	0.990

$r = 4$

	1	2	3	4	Sum
1	13	0	0	0	13
2	0	8	0	0	8
3	0	5	31	71	107
4	0	17	0	0	17
5	0	0	6	2	8
6	0	5	0	0	5
7	0	36	1	0	37
8	0	17	0	0	17
9	3	6	1	0	10
Sum	16	94	39	73	222
AR	1.000	0.990	0.803	0.936	0.938

$r = 5$

	1	2	3	4	5	Sum
1	12	0	1	0	0	13
2	0	2	6	0	0	8
3	0	0	5	70	32	107
4	0	2	15	0	0	17
5	0	0	0	2	6	8
6	0	0	5	0	0	5
7	0	7	29	0	1	37
8	0	13	4	0	0	17
9	3	0	6	0	1	10
Sum	15	24	71	72	40	222
AR	0.968	0.955	0.953	0.929	0.808	0.919

$r = 6$

	1	2	3	4	5	6	Sum
1	13	0	0	0	0	0	13
2	0	5	3	0	0	0	8
3	0	5	0	4	37	61	107
4	0	1	14	2	0	0	17
5	0	0	0	0	1	7	8
6	0	3	0	2	0	0	5
7	0	19	17	0	0	1	37
8	0	2	15	0	0	0	17
9	3	4	1	2	0	0	10
Sum	16	39	50	10	38	69	222
AR	1.000	0.996	0.970	0.884	0.938	0.933	0.945

$r = 7$

	1	2	3	4	5	6	7	Sum
1	2	0	11	0	0	0	0	13
2	0	5	0	0	0	0	3	8
3	0	5	0	71	31	0	0	107
4	10	1	0	0	0	1	5	17
5	0	0	0	2	6	0	0	8
6	0	3	0	0	1	1	0	5
7	2	20	0	0	1	1	13	37
8	0	3	0	0	0	0	14	17
9	1	3	3	0	0	3	0	10
Sum	15	40	14	73	39	6	35	222
AR	0.965	0.943	0.996	0.939	0.854	0.999	0.951	0.930

表 9.9: 比率変数での分類結果 (6,9 種群除いて, g=2 ～ 7)

$r = 2$

	1	2	Sum
1	13	0	13
2	8	0	8
3	5	102	107
4	17	0	17
5	0	8	8
7	36	1	37
8	17	0	17
Sum	96	111	207
AR	0.999	0.995	0.997

$r = 3$

	1	2	3	Sum
1	13	0	0	13
2	0	8	0	8
3	0	5	102	107
4	0	17	0	17
5	0	0	8	8
7	0	36	1	37
8	0	17	0	17
Sum	13	83	111	207
AR	1.000	0.996	0.996	0.996

$r = 4$

	1	2	3	4	Sum
1	13	0	0	0	13
2	0	6	2	0	8
3	0	5	0	102	107
4	0	15	2	0	17
5	0	0	0	8	8
7	0	29	7	1	37
8	0	4	13	0	17
Sum	13	59	24	111	207
AR	1.000	0.966	0.944	0.997	0.982

$r = 5$

	1	2	3	4	5	Sum
1	13	0	0	0	0	13
2	0	6	2	0	0	8
3	0	5	0	39	63	107
4	0	15	2	0	0	17
5	0	0	0	1	7	8
7	0	29	7	0	1	37
8	0	4	13	0	0	17
Sum	13	59	24	40	71	207
AR	1.000	0.965	0.944	0.927	0.945	0.950

$r = 6$

	1	2	3	4	5	6	Sum
1	0	13	0	0	0	0	13
2	0	0	3	0	2	3	8
3	31	0	5	71	0	0	107
4	0	0	15	0	2	0	17
5	6	0	0	2	0	0	8
7	1	0	28	0	5	3	37
8	0	0	5	0	11	1	17
Sum	38	13	56	73	20	7	207
AR	0.866	1.000	0.960	0.939	0.952	1.000	0.938

$r = 7$

	1	2	3	4	5	6	7	Sum
1	0	0	13	0	0	0	0	13
2	1	2	0	4	1	0	0	8
3	1	0	0	4	0	71	31	107
4	8	3	0	1	5	0	0	17
5	0	0	0	0	0	2	6	8
7	8	9	0	19	0	0	1	37
8	0	14	0	1	2	0	0	17
Sum	18	28	13	29	8	73	38	207
AR	0.942	0.967	1.000	0.977	1.000	0.939	0.866	0.941

表 9.10: 比率変数でのコンポーネント数の推定結果

初期分類法	$\hat{r}$	ICBoot	バイアスの標準誤差	$\hat{r}$	ICBoot_{M}	バイアスの標準誤差
GA	(2)	4603.8	0.1031	(2)	4608.9	0.0552
WA	(4)	4593.8	0.1162	(4)	4594.7	0.0729
WD	3	4599.4	0.1126	3	4599.2	0.0579
FX	7*	4569.8	0.1322	7*	4567.7	0.0961
MF	5	4589.6	0.1227	5	4589.2	0.0826
KM	6	4595.7	0.1407	6	4597.1	0.1080

表 9.11: 比率変数でのコンポーネント数の推定結果 (第 6,9 種群を取り除いた)

初期分類法	$\hat{r}$	ICBoot	バイアスの標準誤差	$\hat{r}$	ICBoot_{M}	バイアスの標準誤差
CL	5*	4215.0	0.1361	5*	4214.8	0.0933
GA	(3)	4236.4	0.1199	(3)	4239.1	0.0631
CD	(3)	4223.8	0.1208	(3)	4226.4	0.0659
WA	(5)*	4215.6	0.1475	(5)*	4216.3	0.1017
MD	(2)	4246.6	0.1025	(2)	4248.9	0.0376
WD	6	4233.3	0.1497	6	4231.5	0.1096
FX	6	4229.8	0.1377	6	4229.0	0.1000
MF	4	4229.4	0.1222	4	4230.3	0.0743
KM	5	4223.1	0.1251	6	4214.6	0.1082

WA と CL はパラメータの推定結果からほとんど同じ分類結果であった.

第 10 章

混合分布モデルによるLANDSAT画像データの分類

地球観測衛星 LANDSAT から送信される画像データの解析技術の進歩によって, 地表上の植生分布, 植物の活力度, 海洋, 河川, 湖沼の汚染状態, 農作物の作柄状態などが, 広範囲にわたって把握できるようになってきた. このような画像データに内在する情報を抽出するとき, "分類" は本質的な操作で, 実際に数多くの分類手法が提案されている.

現在広く利用されている LANDSAT 5号のセンサ TM (Thematic Mapper) の解像度 (分解能または瞬時視野のことで, 1画素の大きさ) は 30m×30m, また, MSS (Multispectral Scanner) は 83m×83m であり (宇宙開発事業団地球観測センター編集 (1990)), 日本のような複雑に入り組んだ土地利用状況等を分析するには決して解像度が高いとはいえない. つまり1つの画素中にいくつもの対象物が混入し, その画素を特定の対象物として分類・識別することには限界がある.

LANDSAT の画像データの特性は, ディスプレイ上に表示される画像の各画素の位置を表す座標と, いくつかの波長帯の観測値 (多重分光の輝度値) から構成される (観測値により作られる空間を "特徴空間 (feature space)" と言う). 従来の画像データの分類法には, この観測値のみを用いる手法や, これに座標を加えたものを用いる手法がある. これらの分類法は地表上の細かい部分を識別することが目的で, つまりクラスとして水田, 畑, 住宅地, 道路, 市街地などの特定の "対象物" に対する分類精度の向上を, 主な研究課題としてきた (ここで "クラス" とは, 利用者の目的によって意味付けられる似ているものの集まり, または画像データを何らかの方法で分類して得られる個々の等質な集合のこと). そして分類結果を画像表示する際には一つのクラスに一つの色を割り当てる方法が行われてきた.

このような従来の分類法や配色方法に対して, ここでは混合分布モデルと配色アルゴリ

ズムを用いた分類法を提案する．まず，特徴空間の構造が複数の分布による多変量混合分布からなっていると考える．ただし特徴空間の次元が高いので観測値に主成分分析を行い，次元を縮小した主成分スコアに対して混合分布モデルをあてはめる．そして，従来の分類法のような特定の対象物を識別するのではなく，画像全体の特徴をとらえるため，大まかな分類を行う．そのために，ほぼ“水域”（海，河川，湖沼），“植生”，“人工物”（建築物や市街地などの人工的に作られた建造物）のようなクラスを想定する．次に，推定した混合分布の情報を用いて分類結果を色彩情報として画像化する．これは，従来の分類法のように画素単位でクラスへの所属判定を行い，これに配色を行うことによって色彩画像を生成するのではなく，各コンポーネント分布の特徴を，ある種の平滑化したマクロな色彩イメージ情報として画像上に視覚化することである．

なお，実際の画像データの観測値には多変量正規分布より裾が重いデータが存在するため，混合分布モデルのコンポーネント分布として多変量正規分布と，これより裾の重い多変量 t 分布の 2 つを比較検討した．この結果，多変量正規分布にもとづく混合分布モデルのあてはまりが必ずしもよくない場合があり，この解決策としての，多変量 t 分布にもとづく混合分布モデルの有効性を検証することができた．

以上の内容に沿って，まず扱う画像データと具体的な分類手順について述べる．次に 3 種類の画像データの解析例をもちいて，分類結果を色彩画像表示する際の配色方法について述べる．

10.1 画像データの特徴

解析に用いる画像データは，LANDSAT 5 号の TM で観測された 7 つのバンド（観測波長帯）からなる多変量の観測値である．バンド 1 ～ 5，7 の分解能は 30m×30m，バンド 6 は 120m×120m である．各バンドの観測値は 0 ～ 255 の整数値からなり，バンド 1 ～ 3 は可視光線帯域，バンド 4，5，7 は近赤外線帯域，バンド 6 は熱赤外線帯域である．今回の解析では分解能が大きく異なるバンド 6 を除いた 6 つのバンドを解析対象とした．扱うデータとしては各バンドの観測値だけではなく，座標情報，時間的情報の利用も考えられるが，ここでは観測値のみを用いる．

一般に, 解析対象データは, その目的に応じたバンドの選択を行ったり, バンド間演算などにより次元を縮小することが行われる. 本来は6つのバンドを同時に観察したいが, 混合分布モデルを適用する上で次元が多いので, ここでは主成分分析により次元を縮小してデータの主な特徴を抽出することにする. そこで, 分解能の同じバンド（バンド1～5, 7）の観測値から得られる分散共分散行列にもとづく主成分分析によって次元縮小を行い, 正規化しない第1, 2主成分スコアを分類対象のデータとする. ここで, 主成分スコアの布置により構成される空間を "データ空間" と呼ぶ. 主成分スコアは各バンドの観測値の線形結合であるので, この操作は一種のバンド間演算と考えられる.

10.2 LANDSAT 画像データの解析手順

分析対象の各シーンについて以下の手順で解析を進める. 図10.1はこの手順の概略を模式的に示したものである.

ステップ1 [次元の縮小]

1シーンの画像の全画素上の観測値を用いて標本分散共分散行列を求め, 主成分分析により次元縮小を行う. 正規化しない第1, 2主成分スコアを求め, これを分類対象データとする.

ステップ2 [トレーニングデータの作成]

1シーンの全画素数は数万～数十万画素と膨大であるため, 1シーンあたりそれぞれのデータ空間から5% をランダム・サンプリングし, これを "トレーニング・データ" とする. 今回解析した画像データはいずれも $600 \times 800 = 480,000$ 画素であるから, トレーニング・データの標本数は, $N = 24,000$ 画素となる.

ステップ3 [コンポーネント分布の数の指定]

トレーニング・データの3次元ヒストグラムを描き（図10.5(a), 図10.6(a), 図10.6(b)）, これを観察して混合分布モデルのコンポーネント分布の数を指定する.

ステップ 4 [初期分類]

ステップ 3 で決めたコンポーネント分布の数にしたがって, トレーニング・データを分割型分類法（k-means 法, MacQueen (1967) など）で初期分類し, 各クラスターの統計量（平均ベクトル, 分散共分散行列, 混合比率）を計算する.

ステップ 5 [正規混合分布モデルのあてはめ]

ステップ 4 で求めた統計量を EM 法（次節で説明）の初期値として, 正規混合分布モデル（多変量正規分布の混合分布モデル）のパラメータ推定を行う.

ステップ 6 [t 混合分布モデルのあてはめ]

ステップ 5 で推定した正規混合分布モデルのパラメータの値を初期値として EM 法と *reweighting* 法（6.2 節で説明）を用いて t 混合分布モデル（多変量 t 分布の混合分布モデル）のパラメータ推定を行う.

ステップ 7 [事後確率による判別]

二つの混合分布モデルから推定されたパラメータを用いてそれぞれ判別ルール（事後確率による判別ルール）を構成し, サンプリングしたトレーニング・データも含めて, 1 シーンの全データの各コンポーネント分布に対する所属を決定する. この結果は配色ルールで用いる.

ステップ 8 [配色ルール]

ステップ 5, 6 のパラメータを用いて, 色彩画像表示のための配色ルールを構成する.

ステップ 9 [色彩散布図表示]

配色ルールにもとづき, 1 シーンの中のすべてのデータについて配色を行い, “色彩散布図” （配色されたデータ空間）を描く（図 10.8, 10.9, 10.10の (a), (c)）.

ステップ **10** [色彩画像表示]

色彩散布図の配色をもとに, "色彩画像" の表示を行う (配色された画像)

(図 10.8, 10.9, 10.10の (b), (d)) .

10.3 パラメータ推定アルゴリズム

前節の解析手順ステップ 5, 6 の混合分布モデルのあてはめるとき, そのパラメータ推定は 6.3 節のアルゴリズムで行う. ここで, 形状パラメータ ν_k の初期値は 4 とし, 収束条件は $\varepsilon = \delta = 10^{-7}$ とした. さらに, ステップ 6 の形状パラメータの推定は, 北川 (1993) の付録 A のプログラムを用いた.

10.4 分類結果の画像表示および配色方法

次に, 混合分布モデルによって推定されたデータ空間の特徴を効果的に表示する配色方法を提案する.

10.4.1 HSI 空間

まず, カラーモデルである HSI 空間について述べる. カラーディスプレイなどで画素上の色を特定する場合には, RGB(Red, Green, Blue) の数値を指定する方法が一般的であるが, これらの数値の組み合せから合成色の色調をコントロールするのは容易ではない. そこで感覚的に理解しやすい, マンセルの表色系で使われる 3 つの属性, 色相 h(hue) , 彩度 s(saturation) , 明度 i(intensity) を用いる. 通常用いる HSI 空間は六角錐を二つ張り合わせた双六角錐カラーモデルであるが, ここでは図 10.2に示す円柱座標系で表現したカラーモデルを用いる. これを "規格化された HSI 空間" と言う (高木・下田 (1991, 2.1.2.3 節)) . 各属性の定義域は $h \in [0,1)$, また $s, i \in [0,1]$ とする. この空間の中で色の合成を行い, HSI 空間から RGB 空間への変換を通してディスプレイに色彩表示する.

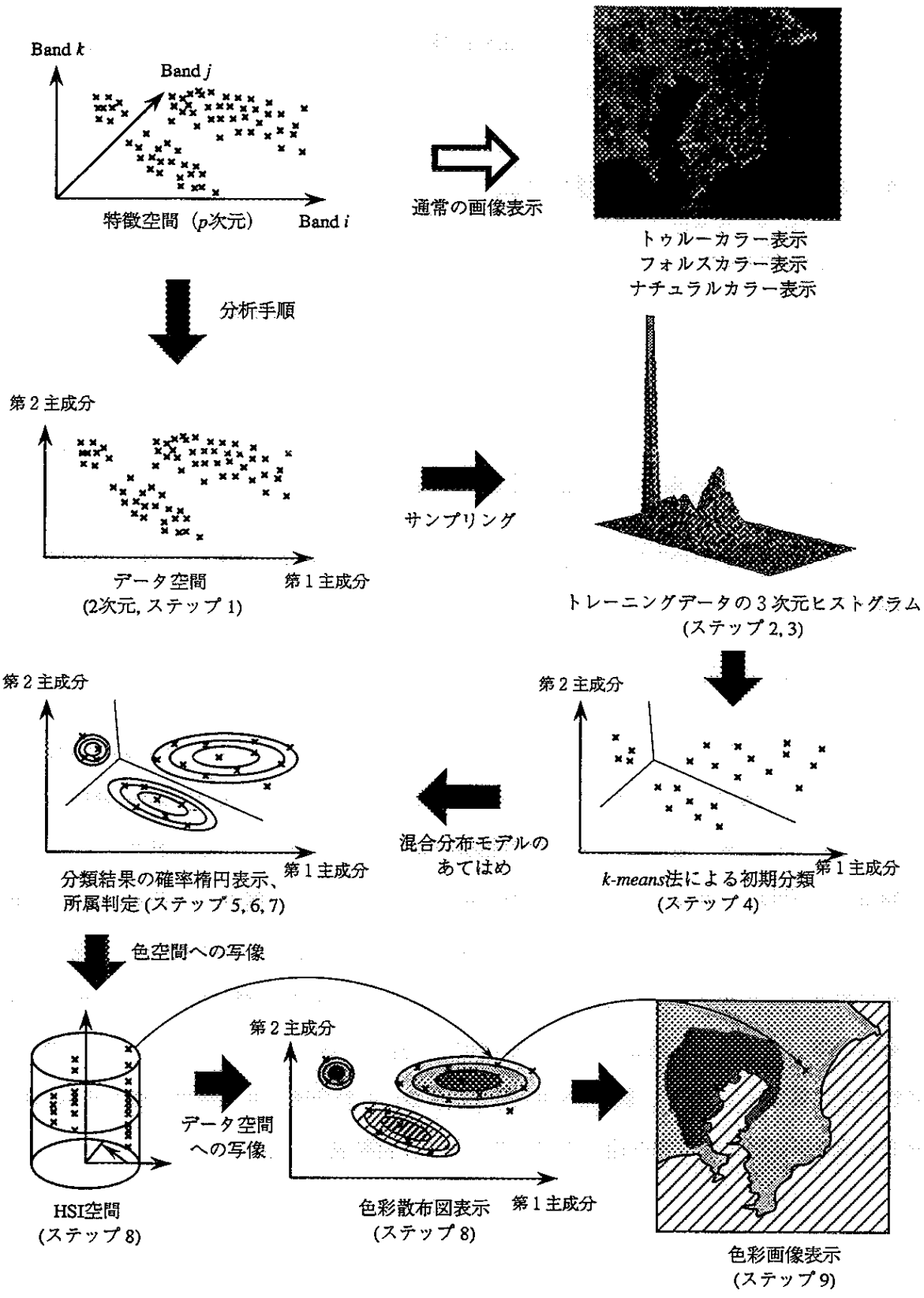

図 10.1: 解析手順の概略

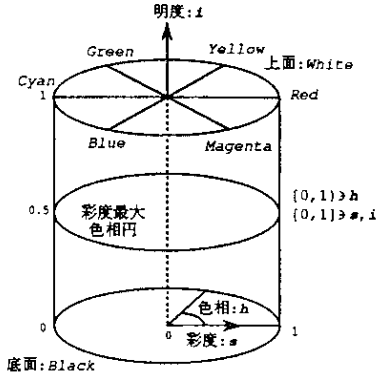

図 10.2: 規格化された HIS 空間

HSI 空間内の色彩の構造は次のとおりである. この空間内のある点 P が与えられたとき, "色相" はあらかじめ定めておいた基準位置との角度で表される. ここで角度は区間 $[0,1)$ で規格化されていて, たとえば Red=0 を基準位置にすると, Yellow=1/6, Green=1/3, Cyan=1/2, Blue=2/3, Magenta=5/6 という値をとる. "彩度" は点 P から HSI 空間の円柱の中心軸までの最短距離で表され, 軸に近いほど灰色に近くなる. "明度" は点 P から円柱の底面への距離になり, 中心軸に沿って底面から上面に向かって黒色から灰色, そして白色へと変化する. s=1, i=1/2 の位置にある円は彩度最大色相円である. なお, この HSI 空間から RGB 空間への変換方法は Foley and Van Dam (1982, chap.17), 高木・下田 (1991, 2.1.2.3 節) によるアルゴリズムを用いた.

10.4.2 配色アルゴリズム

推定した混合分布の各統計量などを用いて,

データ空間内の任意の点 $\boldsymbol{p}^{DS}$ を HSI 空間内の点 $\boldsymbol{p}^{HSI}$ に 1 対 1 写像する方法を示す. まずコンポーネント分布を従来の分類法におけるクラスと考え, 各コンポーネント分布に対して一定の色相と彩度を対応させる. それらの値は各コンポーネント分布の位置ベクトルと色相の基準位置に対応させるコンポーネント分布の位置ベクトルとの関係により定める. 明度は各コンポーネント分布の位置ベクトルと点 $\boldsymbol{p}^{DS}$ の距離により定める. 点 $\boldsymbol{p}^{DS}$ はすべてのコンポーネント分布の定義域にあるため, このまま各コンポーネント分布ごとに点 $\boldsymbol{p}^{DS}$ を HSI 空間上に対応させると, HSI 空間内でコンポーネント分布と同じ数の点 $\boldsymbol{p}^{HSI}$ が存在す

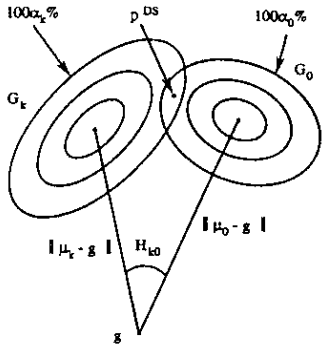

図 10.3: HIS 空間

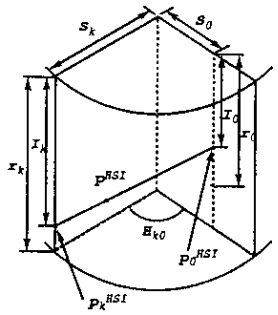

図 10.4: データ空間での分布の位置関係

るので, 点 $\boldsymbol{p}^{DS}$ と点 $\boldsymbol{p}^{HSI}$ の 1 対 1 の対応を, 次の写像アルゴリズムにより行う.

HSI 空間への写像アルゴリズム

ステップ 1 [彩度の求め方]

第 k コンポーネント分布 G_k の彩度 S_k は, 混合比率を重みとする位置ベクトルの重心

$$\boldsymbol{g} = \sum_{k=1}^{r} \hat{\pi}_k \hat{\boldsymbol{\mu}}_k$$

と G_k の位置ベクトル $\hat{\boldsymbol{\mu}}_k$ との距離により決める (図 10.3) . すなわち,

$$S_k = \frac{\min\limits_{i\in\{1,\cdots,r\}} \|\hat{\boldsymbol{\mu}}_i - \boldsymbol{g}\|}{\|\hat{\boldsymbol{\mu}}_k - \boldsymbol{g}\|}$$

とする．これは重心 $\boldsymbol{g}$ に一番近いコンポーネント分布の彩度が最大になり，これから離れるほど灰色に近くなることを示す．

ステップ 2 [色相の求め方]

第 k コンポーネント分布 G_k の色相 H_{k0} は，$\hat{\boldsymbol{\mu}}_k$ と色相の基準位置に対応させるコンポーネント分布 G_0 の位置ベクトル $\hat{\boldsymbol{\mu}}_0$ が重心となす角度により決定する（図 10.3）．すなわち

$$H_{k0} = \cos^{-1} \frac{(\hat{\boldsymbol{\mu}}_k - \boldsymbol{g}, \hat{\boldsymbol{\mu}}_0 - \boldsymbol{g})}{\|\hat{\boldsymbol{\mu}}_k - \boldsymbol{g}\| \cdot \|\hat{\boldsymbol{\mu}}_0 - \boldsymbol{g}\|}$$

とする．G_0 の決め方は後で述べる．

ステップ 3 [明度の変化幅の求め方]

明度は各コンポーネント分布の位置ベクトルとデータの擬マハラノビス距離により決めるが，その変化幅は

$$r_k = \frac{\log|\widehat{\boldsymbol{V}}_k|}{\max\limits_{i\in\{1,\cdots,r\}} \log|\widehat{\boldsymbol{V}}_i|}, \quad r_k \in [0,1]$$

とする（図 10.4）．ここで $|\widehat{\boldsymbol{V}}_k|$ はコンポーネント分布 G_k の擬分散共分散行列の行列式の値である．

ステップ 4 [明度の求め方]

コンポーネント分布と点 $\boldsymbol{p}^{DS}$ の距離は，コンポーネント分布の確率楕円を考え，そのパーセント点により与える．これはコンポーネント分布が正規分布の場合はマハラノビス距離に対応する．いま点 $\boldsymbol{p}^{DS}$ がコンポーネント分布 G_k の確率楕円の $100\alpha_k$ G_k における点 $\boldsymbol{p}^{DS}$ の明度を，

$$I_k = 1 - \alpha_k r_k + \varepsilon_k$$

により与える（図 10.4）．ここで $\varepsilon_k(\geq 0)$ は点 $\boldsymbol{p}^{DS}$ がどのコンポーネント分布に所属するかを強調するためのものである．つまり，解析手順のステップ 7 のコンポーネント分布に対する所属判定の結果により，G_k に所属していれば $\varepsilon_k = constant(\ll 1)$，そうでなければ 0 とする．実際には

$constant = 0.005$ とした．ここで α_k は連続量であるが，ディスプレイの表示可能な色の数に制限があるため，実際は離散化したものを用いた．そこで，$\boldsymbol{p}^{DS}$ のパーセント点が $100\alpha_k\%$ と $100\alpha'_k\%$ の間（$\alpha_k > \alpha'_k$）にあるとき，$\boldsymbol{p}^{DS}$ は $100\alpha_k\%$ 点とする（図 10.3）．実際に離散化したパーセント点としては，0, 5, 10, 25, 50, 75, 90, 95, 97.5, 99% の 9 階調を用いた．

ステップ 5 [色の合成]

最後に，データ空間内のある点 $\boldsymbol{p}^{DS}$ と，HSI 空間の円柱座標内での点 $\boldsymbol{p}^{HSI}$ は次のように対応付けする．まず，ステップ 1, 2 によりコンポーネント分布 G_k の色相 H_{k0} と彩度 S_k の値が決まり，次に，任意の点 $\boldsymbol{p}^{DS}$ が与えられると，ステップ 3, 4 により分布 G_k に対する明度 I_k が決まる．その座標は $\boldsymbol{p}_k^{HSI} = (H_{k0}, S_k, I_k)$ である（図 10.3）．このとき円柱座標内での点 $\boldsymbol{p}^{HSI}$ の位置は，

$$\boldsymbol{p}^{HSI} = \sum_{j;\alpha_j<\beta} \frac{\alpha_j}{A} \boldsymbol{p}_j^{HSI}, \quad \text{ここで}, A = \sum_{\alpha_j<\beta} \alpha_j$$

で求める．この 2 つの和の記号は，各コンポーネント分布に対して点 $\boldsymbol{p}^{DS}$ のパーセント点が $100\beta\%$ 点より小さい場合のみについて和をとることを示す．つまり，$100\beta\%$ 点より大きい場合は，そのコンポーネント分布の影響がほとんどないと見なすことである．実際には $\beta = 0.99$ とした．

ここで，HSI 空間内で基準位置に対応させるコンポーネント分布 G_0 は，次のように定める．r 個あるコンポーネント分布から任意に 1 つを選んで色彩画像表示する．次に，たとえば植生に相当する領域を緑がかった色に配色する場合，植生に相当するコンポーネント分布を G_0 として HSI 空間で緑付近の色相を基準位置として指定する．

ここで示したアルゴリズムによりデータ空間内のすべてのデータは，色彩として変換され，そのデータに対応するもとの画像の画素上に置かれる．データ空間でコンポーネント分布が重なっている部分は，各コンポーネント分布からの距離を考慮した混色が行われる．つまり，画素上の色の情報から，データ空間内のそのデータのコンポーネント分布に対する所属の度合が，色調の変化として視覚的に観察できる．結果として，直感的に画像内の空間的分

布の特徴を把握することが可能になるところに, この色彩画像表示の利点がある.

10.5 解析例

解析対象地域は3シーンあり, 三浦半島, 横浜市, 千葉市～習志野市（それぞれ, 600 × 800 画素, 観測日 1986 年 8 月 6 日）である. これらのデータは幾何補正などの基本的な補正処理は施されている. また, 画像表示したときに真上が北になるような回転の処理が行われている.

2節で示した解析手順にしたがって解析した結果を, 図表の観察を中心に述べる.

10.5.1 データ空間の特徴

図 10.5(a) は三浦半島, 図 10.6(a) は横浜市, 図 10.6(b) は千葉市～習志野市のトレーニングデータの3次元ヒストグラムである. これらの3次元ヒストグラムの特徴として, 大きなピークが図 10.5(a) では3つあり, 図 10.6ではそれぞれ2つみられる. ただし, 図 10.6はCの位置に第3の小さなピークが確認できる. 以上のことから, 各シーンに対してコンポーネント分布の数を3として分類を行った.

10.5.2 混合分布モデルのあてはめ

表 10.1, 10.2は3シーンのトレーニングデータに正規混合分布モデルと t 混合分布モデルをあてはめた結果得られた数値である. 表 10.1は, 推定された対数尤度の値, AIC の値, そして EM 法の反復回数と所要計算時間である（計算に使用した機種は富士通 S-4/10 model 30）. 表 10.2は, 2つのモデルの各コンポーネント分布の混合比率と, t 混合分布モデルの形状パラメータの推定値, 第2主成分までの累積寄与率と第1, 2主成分の寄与率である.

図 10.5(b), (c) は三浦半島のトレーニングデータに, それぞれ正規混合分布モデルと t 混合分布モデルをあてはめ, 推定された混合分布の密度関数の立体グラフである. 図 10.5(a) の記号 A, B, C のピークが, 図 10.5(b) の A, B, C のコンポーネント分布に対応し, さらに図 10.7(a) の A, B, C の確率楕円にも対応する. また, 同様に図 10.5(a) の A, B, C は, 図 10.5(c) の A, B, C のコンポーネント分布に, 図 10.7(b) の A, B, C の確率楕円に対応する. このトレーニングデータは, 図 10.5(a) を見て分かるように, はっきり3つに分かれているた

め, 正規混合分布モデルによっても t 混合分布モデルによっても分類結果に大きな差はない. しかし, 図 10.5の 3 つの図を比較すると図 10.5(b) の正規モデルの B が低く推定されている. それに対して, 図 10.5(c) の t 混合分布モデルは B の高さを良くとらえている.

図 10.5は, 横浜市のトレーニングデータの 3 次元ヒストグラムである. これに正規混合分布モデルと t 混合分布モデルをあてはめた結果の各コンポーネント分布の確率楕円はそれぞれ図 10.7 (c) (d) である. この確率楕円は, 50, 90 ,95, 99 % である. この図 10.78(d) の B と C のコンポーネント分布の位置は, 図 10.7 (c) とは違う場所に推定されている. また, 図 10.7 (c) の正規混合分布モデルでの推定結果は, 図 10.6 (a) の A と B の大きな 2 つのピークを主にとらえている（この 2 つのコンポーネント分布の混合比率はあわせて 0.959 である）. これに対して図 10.7(d) の t 混合分布モデルは, 図 10.5の A, B, C が示すピークとほぼ同じ位置にコンポーネント分布が推定されている.

図 10.6(b) は, 千葉市～習志野市のトレーニングデータの 3 次元ヒストグラムである. このデータに正規混合分布モデルと t 混合分布モデルをあてはめた結果の各コンポーネント分布の確率楕円はそれぞれ図 10.7 (e) (f) である. この図 10.7 (e) の A のコンポーネント分布は, 図 10.7 (f) の A と同じ位置にあり, B のコンポーネント分布もほぼ同じ位置にあるが, C の位置が異なる. また, 図 10.7 (e) の正規混合分布モデルの B と C のコンポーネント分布は位置がはほぼ同じであるが, 図 10.7(f) に t 混合分布モデルでは C のコンポーネント分布が B の下の位置に推定されている. 図 10.6 (b) の C の部分に小さなピークがあるが, t 混合分布モデルはこれをとらえているようである.

10.5.3 色彩散布図と色彩画像表示

3 つのシーンの色彩画像はいずれも水域に相当するコンポーネント分布を G_0 として, 青に近い色相で表示してある.

図 10.8 (a), (c) は三浦半島の色彩散布図で, それぞれ正規混合分布モデルと t 混合分布モデルをあてはめた結果である. これらに対応する確率楕円が図 10.7(a), (b) である. これらのモデルの色彩画像表示はそれぞれ図 10.8 (b), (d) である. この色彩画像表示から, 図 10.5(a) と図 10.7(a), (b) の A, B, C は, それぞれ水域, 植生, 人工物に対応すると考えられる.

表 10.1: モデル推定に関する諸数値

シーン	モデル	対数尤度 $\widehat{\mathcal{L}}$	AIC[*1]	EM 法の反復回数	所要計算時間 (分)
三浦半島	正規混合分布モデル	−157237.5	314509.1	145	14
	t 混合分布モデル	−156890.3	313820.7	—[*2]	210
横浜市	正規混合分布モデル	−187931.0	375895.9	193	18
	t 混合分布モデル	−187144.1	374328.3	—[*2]	226
千葉市〜習志野市	正規混合分布モデル	−183540.3	367114.5	484	34
	t 混合分布モデル	−183065.5	366171.0	—[*2]	435

[*1] AIC= $-2\times$(最大対数尤度)$+2\times$(パラメータ数), [*2] EM 法と準ニュートン法を併用しているため 1 つの数値で示せない. 計算に使用した機種は, 富士通 S-4/10 model30.

同様に図 10.9 (a), (c) は横浜市の色彩散布図で, これらに対応するのが図 10.7(c), (d), その色彩画像表示は図 10.9 (b), (d) である. この色彩画像表示から, 図 10.6と図 10.7(d) の t 混合分布モデルの A, B, C は, それぞれほぼ水域, 植生, 人工物に対応すると考えられる. 図 10.7(c) 正規混合分布モデルの場合は A は水域に, B と C をあわせて陸域として対応がつく.

図 10.10 (a), (c) は千葉市〜習志野市の色彩散布図で, これらに対応するのが図 10.7(e), (f), その色彩画像表示は図 10.10 (b), (d) である. この色彩画像表示から, 図 10.6(b) と図 10.7(e) の t 混合分布モデルの A, B, C は, それぞれ水域, 植生, 人工物に対応すると考えられる. 図 10.7(e) の正規混合分布モデルの場合は, 横浜と同様に A は水域に, B と C をあわせて陸域に対応がつく.

表 10.2: 混合比率と形状パラメータの推定値と主成分寄与率

シーン		正規混合分布モデル $\hat{\pi}_k$	t 混合分布モデル $\hat{\pi}_k$	形状パラメータ $\hat{\nu}_k$	主成分累積寄与率 (pc1, pc2)
三浦半島	A	0.536	0.536	58.44	0.975
	B	0.269	0.237	4.50	
	C	0.196	0.227	∞	(0.914, 0.061)
横浜市	A	0.136	0.137	6.79	0.961
	B	0.823	0.639	3.50	
	C	0.041	0.224	∞	(0.876, 0.085)
千葉市	A	0.179	0.179	28.28	0.957
～	B	0.596	0.474	3.60	
習志野市	C	0.225	0.347	30.90	(0.888, 0.069)

記号 A, B, C は図 10.5, 10.6, 10.7のコンポーネント分布に対応する.

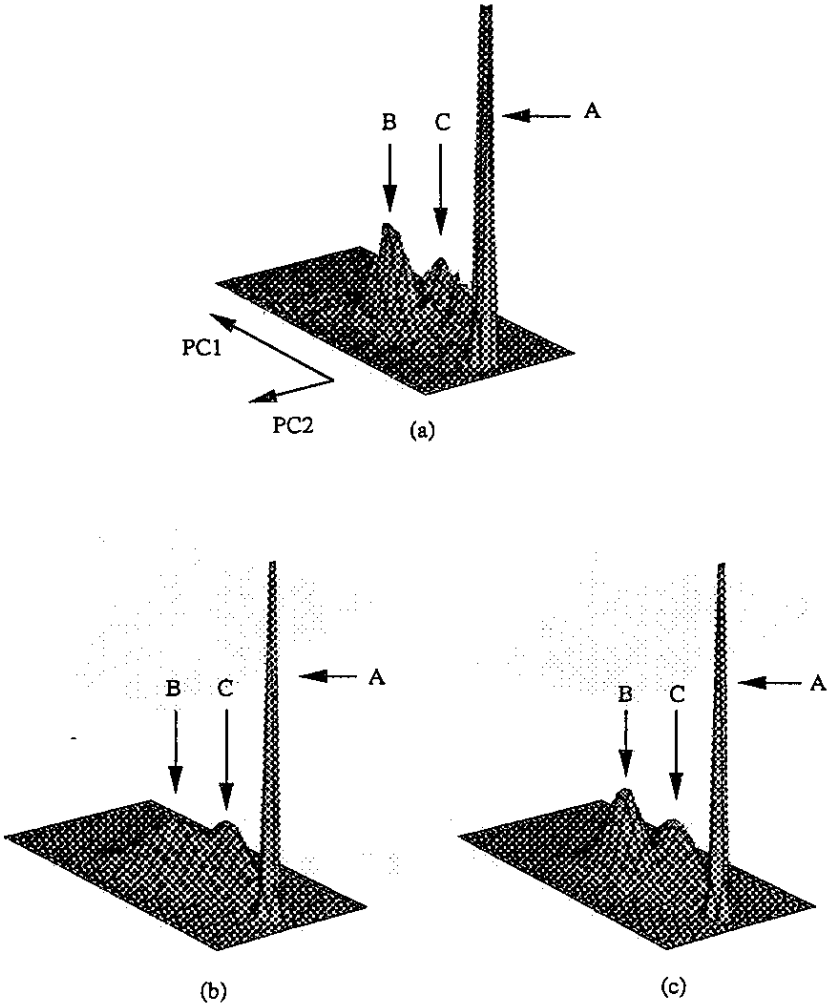

図 10.5: 三浦半島トレーニングデータの 3 次元ヒストグラムと混合分布モデルのあてはめ (a) トレーニングデータの 3 次元ヒストグラム, (b) 正規混合分布モデル, (c)t 混合分布モデル.

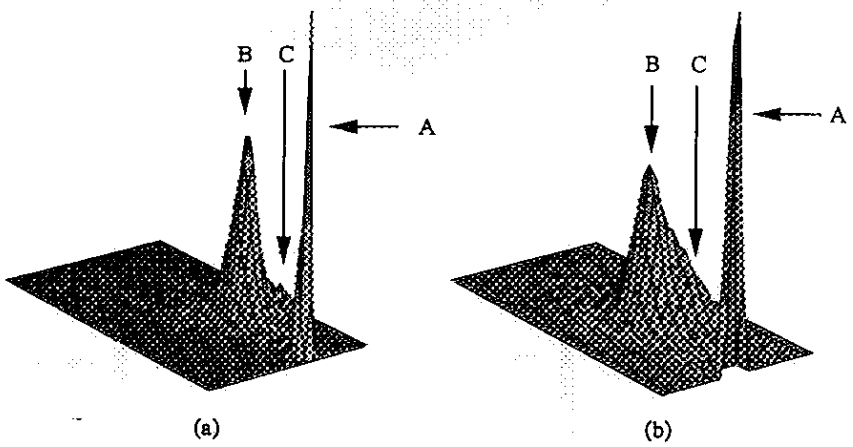

図 10.6: トレーニングデータの 3 次元ヒストグラム
(a) 横浜市, (b) 千葉市～習志野市.

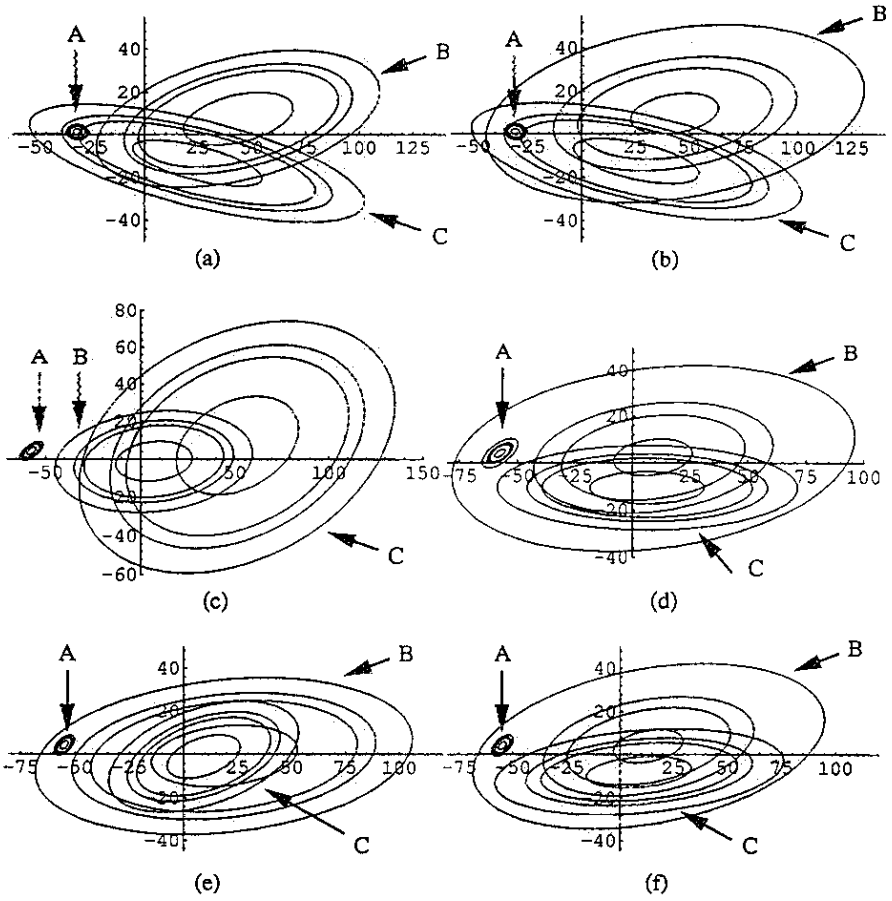

図 10.7: 混合分布モデルの確率楕円
(a) 三浦半島: 正規混合分布モデル, (b) 三浦半島: t 混合分布モデル, (c) 横浜市: 正規混合分布モデル, (d) 横浜市: t 混合分布モデル, (e) 千葉市～習志野市: 正規混合分布モデル, (f) 千葉市～習志野市: t 混合分布モデル. 確率楕円は, 50, 90 ,95, 99 % である.

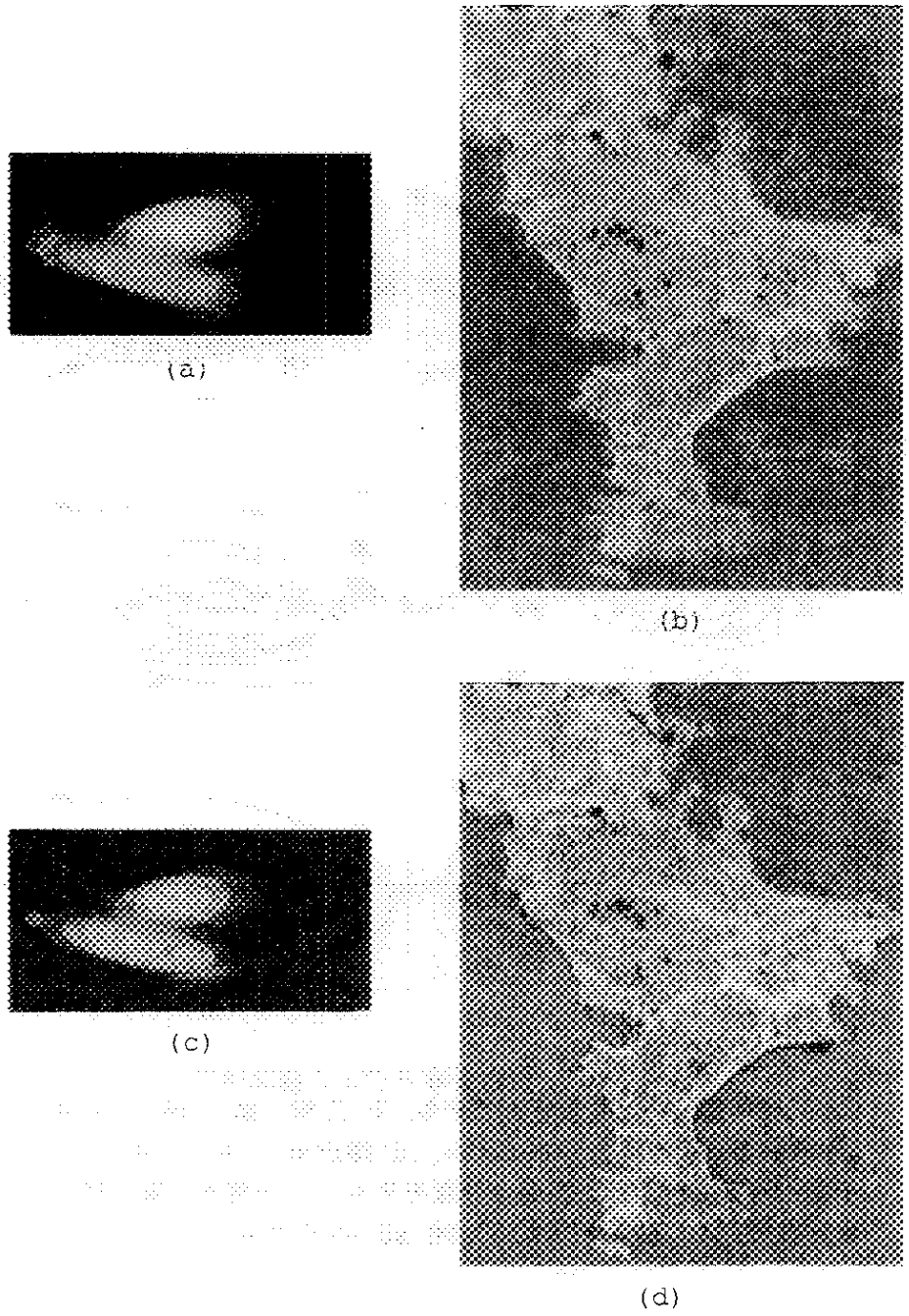

図 10.8: 三浦半島の色彩散布図と色彩画像
(a) 色彩散布図: 正規混合分布モデル, (b) 色彩画像: 正規混合分布モデル, (c) 色彩散布図: t 混合分布モデル, (d) 色彩画像: t 混合分布モデル.

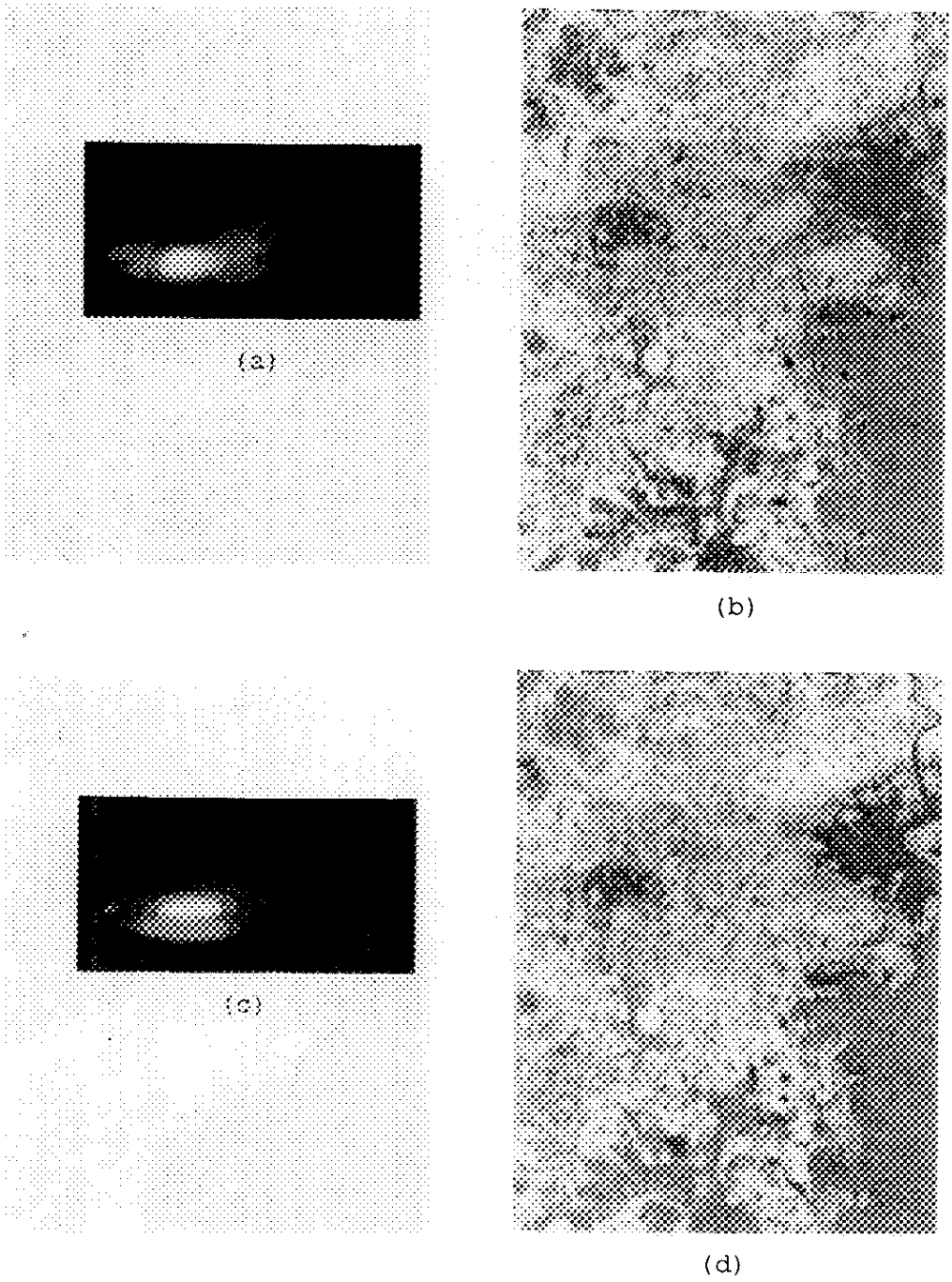

図 10.9: 横浜市の色彩散布図と色彩画像
(a) 色彩散布図: 正規混合分布モデル, (b) 色彩画像: 正規混合分布モデル, (c) 色彩散布図: t 混合分布モデル, (d) 色彩画像: t 混合分布モデル.

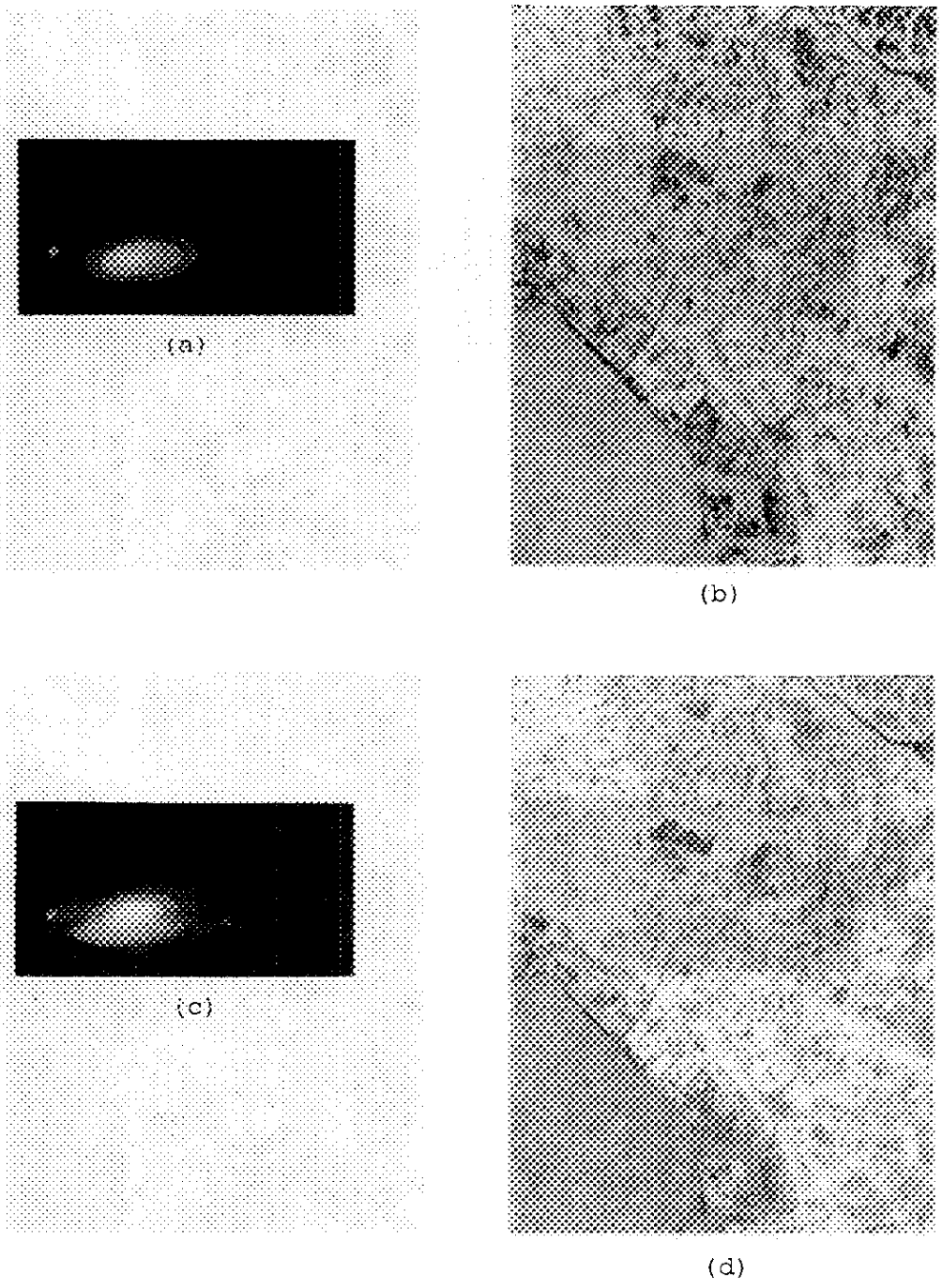

図 10.10: 千葉市～習志野市の色彩散布図と色彩画像
(a) 色彩散布図: 正規混合分布モデル, (b) 色彩画像: 正規混合分布モデル, (c) 色彩散布図: t 混合分布モデル, (d) 色彩画像: t 混合分布モデル.

10.6 考察

3シーンに共通して, t混合分布モデルをあてはめた場合, 人工物に相当するコンポーネント分布の形状パラメータ ν_k の値が3.5 ～ 4.5で推定され, 多変量 t分布を用いた効果が見られた. ν_k の値が小さいことは正規分布に比べて裾が重いことを示す. なお, ν_k が大きな値になると正規分布に近づくので（具体的には20 ～ 30程度以上）, t分布によるモデリングは正規分布を包含するモデルとして扱うことができる. この意味で ν_k は正規性からの乖離の程度を表すパラメータである. たとえば, 三浦半島, 横浜市のデータの人工物に相当するコンポーネント分布の形状パラメータの推定値はかなり大きな値（100以上）で, 正規分布とほとんど変わらない分布として推定された.

t混合分布モデルは水域, 植生, 人工物と各コンポーネント分布をある程度対応づけて推定ができた. 表10.1のAICの値を比較しても, いずれも正規混合分布モデルより t混合分布モデルの方の値が小さいので, モデル選択規準の観点からも t混合分布モデルの方がデータに対するあてはまりは良いと言える.

また, 水域に相当する分布の混合比率は, 正規混合分布モデルと t混合分布モデル双方の推定値がほぼ同じ値である. 水の分布が他の2つに比べて分散が非常に小さく, 他の分布から離れているため, このような結果になったと考えられる.

以上のことから, 多変量 t分布にもとづく混合分布モデルを用いることの利点は次のようになる. 解析に用いる画像データが正規分布より裾が重い場合がしばしばあり, t分布はその特徴をとらえるのに有効な分布と考えられる. このようなデータに対して複数の正規分布をコンポーネント分布とする混合分布モデルをあてはめる方法で対処したとき, 各コンポーネント分布の意味付けが困難になると同時に推定するパラメータの数が多くなるなどの問題が生じる (たとえば, 1コンポーネントに対してさらに正規混合分布モデルをあてはめる). これに対して, t分布を用いると1つのコンポーネント分布でとらえることができ, 推定するパラメータ数の節約ができる.

10.7 検討事項

第一の問題は計算処理時間である．EM 法自身の特性として収束が遅いことがあげられるが，初期値の与え方によっては，収束するまでの反復計算の回数を減せる可能性があるため，初期値の選定方法は十分検討の余地がある．たとえば，グランド・トゥルースの情報を用いてトレーニングデータを作ることや，データ空間の散布図からコンポーネント分布の核となる部分を指定し，そこから初期パラメータを求める方法 (McLachlan 1988) などが考えられる．今回の収束条件は $\varepsilon = \delta = 10^{-7}$ としたが，EM 法の反復計算のかなり早い段階でパラメータの収束値の付近に到達するので，ε と δ の値を 10^{-3} 程度に設定すると計算時間はかなり短縮できると考えられる．また，t 混合分布モデルでは，多変量 t 分布の形状パラメータ ν_k の推定に非線型最適化法を用いるため，正規混合分布モデルに比べて余計に時間を要する．表 10.1の計算時間を比較すると，t 混合分布モデルは正規混合分布モデルに比べ約 12 ～ 15 倍要している．形状パラメータの推定を伴うためこのような結果となったが，実用的な方法や数値計算上の改善が今後の課題である．

次に，トレーニングデータの標本数についても考えなければならない．標本数を増やすと計算時間の問題が生じ，少なくすると母集団（1 シーン内の全データ）の特性をどれくらい代表するか，どの程度信頼できる判別ルールを構成できるか等の問題が生じる．

10.8 おわりに

本章では LANDSAT 画像データに混合分布モデルをあてはめ，さらに混合分布の情報を色彩情報として表現する方法を提案した．この分類法の特徴をまとめると次のようになる．

(1) 主成分分析により次元縮小を行うことで，対象物の分布の様相を容易に観察することができる．また，水域，植生，人工物というクラスで大まかな分類を行うことにより，画像全体の特徴をとらえることができる．分離の良いクラスはこれを取り除いたうえで，再分類を行うなどの処理を行うことにより，二次分析への手掛りになる．

(2) ここで提案する配色方法は，推定した混合分布のパラメータの値や各コンポーネント分布に対する観測値の事後確率と確率密度に応じて配色を行う．つまり，特徴空間（デー

タ空間）の構造が画像上で視覚化され, 分類結果の画像の解釈が容易になる.

(3) 扱う画像データの画素数が比較的大きく（数十万程度）, グランド・トゥルース（地図や現地調査などにもとづく情報）による詳細な情報が入手困難な場合などに有効な手法と考えられる.

LANDSAT の画像データの画像表示方法としては, トゥルーカラー (true color) 表示, フォルスカラー (false color) 表示, ナチュラルカラー (natural color) 表示などの方法がある（詳しくは高木・下田 (1991, 2.1.2 節) を参照）. 基本的には各バンドの輝度値をカラーディスプレイの RGB に割り当て, 色の合成や特性 (バンド) の部分的な強調処理をしてバンドの特徴を見る方法である. ここで提案した方法はこれらの方法や従来の分類法による画像表示方法とは根本的に異なる点に特徴がある.

今回は第 2 主成分スコアまでの 2 次元のデータに対して混合分布モデルをあてはめ, 配色する方法をとったが, 3 次元以上のデータに対してもこの方法は適用可能である. 第 3 主成分以下を使った解析については別の機会に議論したい.

付録

付録 A

群平均法, 重心法, ウォード法の関係

A.1 クラスター間の距離

凝集型の階層的分類法の性質は距離の更新ルール（クラスター間の距離をどのように決めるか）に依存する．しかしその決め方には様々なものが存在する．ここでクラスター間距離として利用される代表的な指標を以下に要約しておく．

1. 最短距離（最近隣距離）

$$d_{ij} = \min\{d_{pq} | p \in C_i, q \in C_j\}$$

2. 最長距離（最遠隣距離）

$$d_{ij} = \max\{d_{pq} | p \in C_i, q \in C_j\}$$

3. 重心間距離

$$d_{ij}^2 = \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2 = \sum_k (\overline{x}_k^{(i)} - \overline{x}_k^{(j)})^2 \quad (k\text{は変量の添字})$$

4. 重みつき重心距離

$$d_{ij}^2 = \frac{n_i n_j}{n_i + n_j} \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2 = \frac{n_i n_j}{n_i + n_j} \sum_k (\overline{x}_k^{(i)} - \overline{x}_k^{(j)})^2$$

5. 対群距離

$$d_{pj} = \frac{1}{n_j} \sum_{q \in C_j}^{n_j} d_{pq} \quad (p \in C_i)$$

$$d_{qi} = \frac{1}{n_i} \sum_{p \in C_i}^{n_i} d_{pq} \quad (q \in C_j)$$

d_{pj} はクラスター C_i 内の個体 p とクラスター C_j 内のすべての個体間距離の平均に相当し, これにより個体とクラスターとの距離を定義している.

6. 群平均距離

$$d_{ij} = \frac{1}{n_i n_j} \sum_{p \in C_i}^{n_i} \sum_{q \in C_j}^{n_j} d_{pq} \quad (p \in C_i,\ q \in C_j)$$

7. 対群距離 (平方ユークリッド距離のとき)

個体間の距離として平方ユークリッド距離を用いるとき, 対群距離は次のように書きかえることができる.

$$d_{pj}^2 = S_j + \|\boldsymbol{x}_p - \overline{\boldsymbol{x}}^{(j)}\|^2 \quad (p \in C_i)$$

$$d_{qi}^2 = S_i + \|\boldsymbol{x}_q - \overline{\boldsymbol{x}}^{(i)}\|^2 \quad (q \in C_j)$$

S_i は C_i のクラスター内分散, $\boldsymbol{x}_p$ は個体 i の観測値ベクトルである. 平方ユークリッド距離による対群距離は, 分散および個体と平均ベクトル間の平方距離に分解できる.

8. 群平均距離 (平方ユークリッド距離のとき)

個体間の距離として平方ユークリッド距離を用いるとき, 群平均距離は次のように書きかえることができる.

$$\begin{aligned} d_{ij}^2 &= S_i + S_j + \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2 \\ &= (\text{クラスター内分散の和}) + (\text{クラスターの重心間距離}) \end{aligned}$$

A.2 群平均法の性質

クラスター間距離として群平均距離を用いるとき, 群平均法は次のようにして導かれる. いま, クラスター $C_t(= C_i \cup C_j)$ と C_k を併合したとき, 得られる距離を $d_{(ij)k}$ とする. ここ

でクラスターサイズは $n_t = n_i + n_j$ とすると, この距離は,

$$\begin{aligned} d_{tk} \equiv d_{(ij)k} &= \frac{1}{n_t n_k} \sum_{u \in C_t}^{n_t} \sum_{r \in C_k}^{n_k} d_{ur} \\ &= \frac{1}{n_t n_k} \sum_{r \in C_k}^{n_k} \{ \sum_{p \in C_i}^{n_i} d_{pr} + \sum_{q \in C_j}^{n_j} d_{qr} \} \\ &= \frac{n_i}{n_t} \frac{1}{n_i n_k} \sum_r \sum_p d_{pr} + \frac{n_j}{n_t} \frac{1}{n_j n_k} \sum_r \sum_q d_{qr} \\ &= \frac{n_i}{n_t} d_{ik} + \frac{n_j}{n_t} d_{jk} \\ &= \frac{n_i}{n_i + n_j} d_{ik} + \frac{n_j}{n_i + n_j} d_{jk} \end{aligned}$$

のようになる. ここで d_{ik} を平方ユークリッド距離としても今の式変形がなりたつので, 次のように書いてもよい.

$$d^2_{(ij)k} = \frac{n_i}{n_i + n_j} d^2_{ik} + \frac{n_j}{n_i + n_j} d^2_{jk}$$

以上のことから, 群平均距離によって更新される距離 $d_{(ij)k}$ は, クラスターサイズの比による d_{ik} と d_{jk} の比例配分である.

A.3 重心法の性質

重心間距離をクラスター間の距離とすると次のような式変形ができる. まずクラスター C_i, C_j, C_k, $C_{(ij)} = C_t$ 間の距離は,

$$d^2_{ik} = \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(k)}\|^2, \quad d^2_{jk} = \|\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(k)}\|^2, \quad d^2_{tk} = \|\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2, \quad d^2_{ij} = \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2$$

となる. このとき,

$$\begin{aligned} & \frac{n_i}{n_t} d^2_{ik} + \frac{n_j}{n_t} d^2_{jk} \\ = \; & \frac{n_i}{n_t} \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)} + \overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2 + \frac{n_j}{n_t} \|\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(t)} + \overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2 \\ = \; & \frac{n_i}{n_t} \{ \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + \|\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2 \} + 2 \frac{n_i}{n_t} (\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)})^T \cdot (\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}) \\ & + \frac{n_j}{n_t} \{ \|\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + \|\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2 \} + 2 \frac{n_j}{n_t} (\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(t)})^T \cdot (\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}) \\ = \; & \frac{n_i}{n_t} \{ \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + \|\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2 \} + \frac{n_j}{n_t} \{ \|\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + \|\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2 \} \\ = \; & \frac{n_i}{n_t} \|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + \frac{n_j}{n_t} \|\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + \frac{n_i + n_j}{n_t} \|\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2 \qquad (A) \end{aligned}$$

$(\frac{n_i}{n_t}(\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)})^T(\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}) + \frac{n_j}{n_t}(\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(t)})^T(\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(t)}) = 0$である)
ここで,

$$n_i\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + n_j\|\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}}^{(t)}\|^2 = \frac{n_i n_j}{n_t}\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2 \quad (n_t = n_i + n_j)$$

であるので,

$$\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2 = \frac{n_t}{n_j}\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + \frac{n_t}{n_i}\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}\|^2$$

とかける. そして次式の重みつき平方ユークリッド距離に代入すると, 次のようになる.

$$\begin{aligned}\frac{n_i n_j}{n_t^2}d_{ij}^2 &= \frac{n_i n_j}{n_t^2}\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2 \\ &= \frac{n_i}{n_t}\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}\|^2 + \frac{n_j}{n_t}\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(t)}\|^2\end{aligned}$$

(A) 式から上の式を差し引くと次の式が得られる.

$$d_{tk}^2 = \frac{n_i}{n_t}d_{ik}^2 + \frac{n_j}{n_t}d_{jk}^2 - \frac{n_i n_j}{n_t^2}d_{ij}^2$$

これで重心法の距離の更新式が得られた. ここで注意するのは, 重心法のクラスター間距離は平方ユークリッド距離であり, 群平均法ではユークリッド距離, 平方ユークリッド距離いずれでも良いことである.

A.4 ウォード法の性質

クラスター間距離として重みつき重心距離

$$d_{ij}^2 = \frac{n_i n_j}{n_i + n_j}\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2$$

を用いることを考える. この加重部分はクラスターサイズの調和平均になっている. ここで, 次の関係がなりたっている.

$$W_t = W_i + W_j + \frac{n_i n_j}{n_i + n_j}\|\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}}^{(j)}\|^2$$

ここで, W_i はクラスター C_i 内にある個体の平均からの平方和である. この重みつき重心間距離が小さいとき, 2 つのクラスター C_i, C_j の距離が近いとみることができる. この基準に従い, 重みつき重心間距離が小さいクラスターどうしを順次併合する階層的分類法が考えら

れる．これは Ward(1963) により提唱されたので，この階層的分類法はウォード法と呼ばれるようになった．

ここまでは C_i と C_j の併合を考えたが，さらに $C_t = C_i \cup C_j$ と C_k を併合したとすると $(n_t = n_i + n_j)$，次の関係がなりたつ．

$$W_{tk} = W_t + W_k + \frac{n_t n_k}{n_t + n_k} \|\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}}^{(k)}\|^2$$

ここで，重心間距離を用いると，

$$\boldsymbol{\Delta} W_{tk} = \frac{n_t n_k}{n_t + n_k} d_{tr}^2$$

となるので，

$$d_{tr}^2 = \frac{n_t + n_k}{n_t n_k} \boldsymbol{\Delta} W_{tk}$$

が得られる．この式を重心法の結果に代入してみると，次のような式が得られる．

$$\frac{n_t + n_k}{n_t n_k} \boldsymbol{\Delta} W_{tk} = \frac{n_i}{n_t} \cdot \frac{n_i + n_k}{n_i n_k} \boldsymbol{\Delta} W_{ik} + \frac{n_j}{n_t} \cdot \frac{n_j + n_k}{n_j n_k} \boldsymbol{\Delta} W_{jk} - \frac{n_i n_j}{n_t^2} \cdot \frac{n_i + n_j}{n_i n_j} \boldsymbol{\Delta} W_{ij}$$

したがって，

$$\boldsymbol{\Delta} W_{tk} = \frac{n_i + n_k}{n_t + n_k} \boldsymbol{\Delta} W_{ik} + \frac{n_j + n_k}{n_t + n_k} \boldsymbol{\Delta} W_{jk} - \frac{n_k}{n_t + n_k} \boldsymbol{\Delta} W_{ij}$$

となる．これで重みつき重心間距離を用いた，距離の更新式が得られた．これは Lance and Williams の組み合わせ的階層分類法の表現形式で，この整合は Wishart(1969b) が行った．

付録 B

混合分布モデルのパラメータの最尤推定量

B.1 多変量正規分布のパラメータの最尤推定量

ここでは多変量 (p 次元) 正規分布 $f(\boldsymbol{x}|\boldsymbol{\Phi})$ のパラメータの最尤推定量を考える．その方法は目的関数を対数尤度関数として，目的のパラメータで偏微分してゼロとおくことで得られる．ここで $\boldsymbol{X}_N = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ は観測データは p 次元の多変量正規分布にしたがっているとし，$\boldsymbol{\Phi} = \{\boldsymbol{\mu}, \boldsymbol{\theta}\}$ は多変量正規分布の平均ベクトルと分散共分散行列のパラメータとする．

まず準備として $\boldsymbol{V} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{V} = [v_{ij}])$ とおいて，二次形式を分散共分散行列で偏微分すると次のようになる．

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{V}}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{V}(\boldsymbol{x}_i - \boldsymbol{\mu}) &= \frac{\partial}{\partial \boldsymbol{V}}\mathrm{tr}[\boldsymbol{V}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T] \\ &= 2(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T - \mathrm{diag}[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T]. \\ &\quad (\boldsymbol{V}^T = \boldsymbol{V}\text{である})\end{aligned}$$

次に $\boldsymbol{V}$ の行列式の対数の v_{ij} 要素による偏微分は，

$$\begin{aligned}\frac{\partial}{\partial v_{ij}}\log|\boldsymbol{V}| &= \frac{1}{|\boldsymbol{V}|}\frac{\partial|\boldsymbol{V}|}{\partial v_{ij}} = \frac{1}{|\boldsymbol{V}|}\{2\boldsymbol{V}_{ij} - \delta(i,j)\boldsymbol{V}_{ij}\} \\ &\quad (\text{ここで，}\boldsymbol{V}^T = \boldsymbol{V},\ \boldsymbol{V}_{ij}\text{は}(i,j)\text{要素の}\boldsymbol{V}\text{の余因子})\end{aligned}$$

$\dfrac{\boldsymbol{V}_{ij}}{|\boldsymbol{V}|}$ は $\boldsymbol{V}^{-1}$ の (ij) 成分であるので，したがって，

$$\frac{\partial}{\partial \boldsymbol{V}}\log|\boldsymbol{V}| = 2\boldsymbol{V}^{-1} - \mathrm{diag}\boldsymbol{V}^{-1}$$

となる. 以上の準備の下で対数尤度関数は,

$$\begin{aligned}\ell(\boldsymbol{\Phi}|\boldsymbol{X}_N) &= \log[\prod_{i=1}^{N} f(\boldsymbol{x}_i|\boldsymbol{\Phi})] = \sum_{i=1}^{N} \log f(\boldsymbol{x}_i|\boldsymbol{\Phi}) \\ &= \sum_{i=1}^{N}\{-\frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{V}(\boldsymbol{x}_i-\boldsymbol{\mu})\} \\ &= -\frac{Np}{2}\log(2\pi) + \frac{N}{2}\log|\boldsymbol{V}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{V}(\boldsymbol{x}_i-\boldsymbol{\mu})\end{aligned}$$

となるので, まずこれを$\boldsymbol{V}$で偏微分してゼロとおくと次のようになる.

$$\frac{\partial \ell}{\partial \boldsymbol{V}} = \frac{N}{2}\frac{\partial}{\partial \boldsymbol{V}}\log|\boldsymbol{V}| - \frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{V}}[\mathrm{tr}\boldsymbol{V}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T] = \mathbf{o}.$$

これより次の式が得られる.

$$\begin{aligned}&\frac{N}{2}\{2\boldsymbol{V}^{-1} - \mathrm{diag}\boldsymbol{V}^{-1}\} - \frac{1}{2}\sum_{i=1}^{N}\{2(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T - \mathrm{diag}[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T]\} \\ =\ & N\boldsymbol{V}^{-1} - \frac{N}{2}\mathrm{diag}\boldsymbol{V}^{-1} - \frac{1}{2}\sum_{i=1}^{N}2(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T + \frac{1}{2}\sum_{i=1}^{N}\mathrm{diag}[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T] \\ =\ & N\{\boldsymbol{V}^{-1} - \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\} + \frac{N}{2}\{-\mathrm{diag}\boldsymbol{V}^{-1} + \frac{1}{N}\sum_{i=1}^{N}\mathrm{diag}[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T]\} \\ =\ & N\{\boldsymbol{V}^{-1} - \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\} - \frac{N}{2}\mathrm{diag}\{\boldsymbol{V}^{-1} - \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\} \\ =\ & \mathbf{o}\end{aligned}$$

$\boldsymbol{A} + \mathrm{diag}(\boldsymbol{A}) = \mathbf{o}$ の解は $\boldsymbol{A} = \mathbf{o}$ となることから, したがって,

$$\widehat{\boldsymbol{V}}^{-1} = \widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T$$

となる. 次に平均ベクトルで偏微分すると次のようになる.

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} = \frac{\partial}{\partial \boldsymbol{\mu}}[-\frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\{\boldsymbol{V}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\}] = \mathbf{o}.$$

したがって次式を得る.

$$\begin{aligned}\sum_{i=1}^{N} 2\boldsymbol{V}(\boldsymbol{x}_i-\boldsymbol{\mu}) &= \mathbf{o}, \\ \sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu}) &= \mathbf{o}, \\ \hat{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i.\end{aligned}$$

B.2 多変量正規混合分布モデルのパラメータの最尤推定量

観測データ $\boldsymbol{X}_N = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ が r 個のコンポーネント分布からなる多変量 (p 次元) 正規混合分布モデル

$$f(\boldsymbol{x}|\boldsymbol{\Phi}_r) = \sum_{k=1}^{r} \pi_k f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)$$

に従うものとして, 多変量正規混合分布モデルのパラメータの最尤推定量を求める. ここで, $\boldsymbol{\Phi}_r = \{\pi_1, \cdots, \pi_r, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_r\}$ である. このときこのモデルの対数尤度関数は次のようになる.

$$\begin{aligned}\ell(\boldsymbol{\Phi}_r|\boldsymbol{X}_N) &= \log[\prod_{i=1}^{N}\sum_{k=1}^{r}\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)] = \sum_{i=1}^{N}\log\{\sum_{k=1}^{r}\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)\} \\ &= \sum_{i=1}^{N}\log\{\sum_{k=1}^{r}\pi_k(2\pi)^{-p/2}|\boldsymbol{V}_k|^{1/2}\exp[-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{V}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)]\}\end{aligned}$$

ここで, $\boldsymbol{V}_k = \boldsymbol{\Sigma}_k^{-1}$ である. 準備として次のことを示しておく. まず, 観測値 $\boldsymbol{x}_i$ が第 k コンポーネント分布に所属する確率 (事後確率) は

$$P_{ki} = Pr(k|\boldsymbol{x}_i) = \frac{\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{f(\boldsymbol{x}_k|\boldsymbol{\Phi}_r)}$$

である. また,

$$\frac{\partial f}{\partial \boldsymbol{V}_k} = \frac{\partial}{\partial \boldsymbol{V}_k}\sum_{k=1}^{r}\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k) = \pi_k\frac{\partial f_k}{\partial \boldsymbol{V}_k},$$

$$\frac{1}{f_k}\frac{\partial f_k}{\partial \boldsymbol{V}_k} = \frac{\partial}{\partial \boldsymbol{V}_k}\log f_k.$$

まず, π_k の最尤推定量を求めるが, このパラメータには制約条件 ($\sum \pi_k = 1, \pi_k > 0$) があるので, ラグランジュ未定乗数法を使うことにする. ラグランジュ乗数 λ を使うと目的関数は次のようになる.

$$\mathcal{L} = \ell(\boldsymbol{\Phi}_r|\boldsymbol{X}_N) - \lambda(\sum_{k=1}^{r}\pi_k - 1)$$

とすると,

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{i=1}^{N}\frac{f_k}{f} - \lambda = 0 \tag{B.1}$$

ここでこの式の両辺に π_k を掛けて, $k=1,\cdots,r$ で和をとると,

$$\begin{aligned}\sum_{i=1}^{N}\sum_{k=1}^{r}\pi_k\frac{f_k}{f}-\lambda\sum_{k=1}^{r}\pi_k &= 0\\ \sum_{i=1}^{N}\frac{1}{f}\sum_{k=1}^{r}\pi_k f_k-\lambda &= 0\\ N-\lambda &= 0\end{aligned}$$

したがって, $\lambda=N$ となる. (B.1) 式に $\lambda=N$ を代入して, 両辺に π_k を掛けると,

$$\begin{aligned}\sum_{i=1}^{N}\frac{\pi_k f_k}{f}-N\pi_k &= 0\\ \sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)-N\pi_k &= 0\end{aligned}$$

したがって, π_k の最尤推定量

$$\hat{\pi}_k=\frac{1}{N}\sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)$$

を得る.

次に, 対数尤度関数を $\boldsymbol{V}_k$ で偏微分してゼロとおくと次のようになる.

$$\frac{\partial\ell}{\partial\boldsymbol{V}_k}=\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{V}_k}\log f=\sum_{i=1}^{N}\frac{1}{f}\frac{\partial f}{\partial\boldsymbol{V}_k}=\sum_{i=1}^{N}\frac{1}{f}\frac{\partial\pi_k f}{\partial\boldsymbol{V}_k}=\sum_{i=1}^{N}\frac{1}{f}\frac{\partial\pi_k f}{\partial\boldsymbol{V}_k}$$

$$=\sum_{i=1}^{N}\frac{Pr(k|\boldsymbol{x}_i)}{\pi_k f_k}\frac{\partial\pi_k f_k}{\partial\boldsymbol{V}_k}=\sum_{i=1}^{N}\frac{\pi_k}{f}\frac{\partial f_k}{\partial\boldsymbol{V}_k}=\sum_{i=1}^{N}\frac{Pr(k|\boldsymbol{x}_i)}{f_k}\frac{\partial f_k}{\partial\boldsymbol{V}_k}=\sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)\frac{\partial\log f_k}{\partial\boldsymbol{V}_k}$$

したがって,

$$\begin{aligned}&\sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)\frac{\partial}{\partial\boldsymbol{V}_k}\{-\frac{p}{2}\log(2\pi)+\frac{1}{2}\log|\boldsymbol{V}_k|-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{V}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\}\\ =&\sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)\{\frac{1}{2}[2\boldsymbol{V}_k^{-1}-\mathrm{diag}\boldsymbol{V}_k^{-1}]\\ &-\frac{1}{2}[2(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T-\mathrm{diag}[(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T]]\}\\ =&\sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)\{\boldsymbol{V}_k^{-1}-(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\}\\ &+\frac{1}{2}\sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)\{-\mathrm{diag}\boldsymbol{V}_k^{-1}+\mathrm{diag}[(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T]\}\\ =&\sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)\{\boldsymbol{V}_k^{-1}-(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\}\\ &-\frac{1}{2}\sum_{i=1}^{N}Pr(k|\boldsymbol{x}_i)\ \mathrm{diag}[\boldsymbol{V}_k^{-1}-(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T]\\ =&\ \mathbf{o}\end{aligned}$$

これより,

$$\begin{aligned}\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)\boldsymbol{V}_k^{-1} &= \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T,\\ \hat{\pi}_k N\boldsymbol{V}_k^{-1} &= \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T,\end{aligned}$$

したがって, 第 k コンポーネント分布の分散共分散行列の最尤推定量は, 次のようになる.

$$\widehat{\boldsymbol{V}}_k^{-1} = \widehat{\boldsymbol{\Sigma}}_k = \frac{1}{\hat{\pi}_k N}\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T.$$

次に, 平均ベクトルの最尤推定量は次のようになる.

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{\mu}_k}\log f = \sum_{i=1}^{N}\frac{1}{f}\frac{\partial f}{\partial \boldsymbol{\mu}_k} = \mathbf{o}$$

より,

$$\sum_{i=1}^{N}\frac{\pi_k}{f}\frac{\partial f_k}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)\frac{1}{f_k}\frac{\partial f_k}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)\frac{\partial \log f_k}{\partial \boldsymbol{\mu}_k} = \mathbf{o}.$$

したがって,

$$\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)\frac{\partial}{\partial \boldsymbol{\mu}_k}\{(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{V}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\} = \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)2\boldsymbol{V}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k) = \mathbf{o}$$

$$\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)(\boldsymbol{x}_i-\boldsymbol{\mu}_k) = \mathbf{o}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{\hat{\pi}_k N}\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i)\boldsymbol{x}_i.$$

B.3 楕円分布族のパラメータの最尤推定量

観測データ $\boldsymbol{X}_N = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ が楕円分布族

$$f(\boldsymbol{x}|\boldsymbol{\Phi}) = |\boldsymbol{V}|^{-1/2}h\{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

に従うとして, 楕円分布族のパラメータの最尤推定量を求める. ここで, $\boldsymbol{V}$ は擬分散共分散行列, $\boldsymbol{\mu}$ は位置ベクトル, $\boldsymbol{V}^{-1} = \boldsymbol{W}$, $s_i = (\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{W}(\boldsymbol{x}_i-\boldsymbol{\mu})$, $\boldsymbol{\Phi} = \{\boldsymbol{\mu}, \boldsymbol{V}\}$ とする. 対数尤度関数は,

$$\ell(\boldsymbol{X}_N|\boldsymbol{\Phi}) = \sum_{i=1}^{N}\log f(\boldsymbol{x}_i|\boldsymbol{\Phi}) = \sum_{i=1}^{N}\{\frac{1}{2}\log|\boldsymbol{W}| + \log h(s_i)\}$$

となる. ここで, $\boldsymbol{W}$で偏微分すると次のようになる.

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{W}} &= \frac{N}{2}\frac{\partial}{\partial \boldsymbol{W}}\log|\boldsymbol{W}| + \sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{W}}\log h(s_i) \\ &= \frac{N}{2}\frac{\partial}{\partial \boldsymbol{W}}\log|\boldsymbol{W}| + \sum_{i=1}^{N}\frac{\partial s_i}{\partial \boldsymbol{W}}\log h(s_i)\frac{\partial}{\partial s_i} \\ &= \frac{N}{2}\{2\boldsymbol{W}^{-1} - \mathrm{diag}\boldsymbol{W}^{-1}\} \\ &\quad + \sum_{i=1}^{N}\frac{\partial}{\partial s_i}\log h(s_i)\{2(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T - \mathrm{diag}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\}\end{aligned}$$

ここでこの式をゼロとおくと次のようになる.

$$\{2\boldsymbol{W}^{-1}-\mathrm{diag}\boldsymbol{W}^{-1}\}-\frac{1}{N}\sum_{i=1}^{N}\{-2\frac{\partial}{\partial s_i}\log h(s_i)\}\{2(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T-\mathrm{diag}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\} = \mathbf{o}$$

$w_i = -2\dfrac{\partial}{\partial s_i}\log h(s_i)$ とすると,

$$\{2\boldsymbol{W}^{-1} - \frac{2}{N}\sum_{i=1}^{N} w_i(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\} - \mathrm{diag}\{\boldsymbol{W}^{-1} - \frac{1}{N}\sum_{i=1}^{N} w_i(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T\} = \mathbf{o}$$

したがって, 擬分散共分散行列の最尤推定量は

$$\widehat{\boldsymbol{W}}^{-1} = \widehat{\boldsymbol{V}} = \frac{1}{N}\sum_{i=1}^{N} w_i(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T$$

となる. 次に $\boldsymbol{\mu}$ で偏微分すると次のようになる.

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\mu}} &= \sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{\mu}}\log h(s_i) = \sum_{i=1}^{N}\frac{\partial}{\partial s_i}\log h(s_i)\frac{\partial s_i}{\partial \boldsymbol{\mu}} \\ &= \sum_{i=1}^{N}\frac{\partial}{\partial s_i}\log h(s_i)(-2)\boldsymbol{V}(\boldsymbol{x}_i-\boldsymbol{\mu}) = \sum_{i=1}^{N} w_i\boldsymbol{V}(\boldsymbol{x}_i-\boldsymbol{\mu}) \\ &= \mathbf{o}\end{aligned}$$

これより,

$$\sum_{i=1}^{N} w_i(\boldsymbol{x}_i-\boldsymbol{\mu}) = \mathbf{o}$$

したがって, 位置ベクトルの最尤推定量は

$$\hat{\boldsymbol{\mu}} = \frac{1}{\sum_{i=1}^{N} w_i}\sum_{i=1}^{N} w_i\boldsymbol{x}_i$$

となる.

B.4 楕円分布族の多変量混合分布モデルのパラメータの最尤推定量

観測データ $\boldsymbol{X}_N = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ が r 個の楕円分布族のコンポーネント分布からなる多変量 (p 次元) 混合分布モデル

$$f(\boldsymbol{x}|\boldsymbol{\Phi}_r) = \sum_{k=1}^{r} \pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k) = \sum_{k=1}^{r} \pi_k |\boldsymbol{W}_k|^{1/2} h_k((\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_i - \boldsymbol{\mu}_k))$$

に従うとする. ここで, $\boldsymbol{\Phi}_r = \{\pi_1, \cdots, \pi_r, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_r\}$, $\boldsymbol{V}_k^{-1} = \boldsymbol{W}_k$ である. このときこのモデルの対数尤度関数は次のようになる.

$$\ell(\boldsymbol{X}_N|\boldsymbol{\Phi}_r) = \sum_{i=1}^{N} \log f(\boldsymbol{x}_i|\boldsymbol{\Phi}_r) = \sum_{i=1}^{N} \log \sum_{k=1}^{r} \pi_k |\boldsymbol{V}_k|^{-1/2} h_k(s_{ki})$$

ここで, $\boldsymbol{V}_k$ は擬分散共分散行列, $\boldsymbol{\mu}_k$ は位置ベクトル, $s_{ki} = (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_i - \boldsymbol{\mu}_k)$ とする.

まず, π_k の最尤推定量は正規混合分布モデルと全く同じであるので省略する. 次に, 対数尤度関数を $\boldsymbol{W}_k$ で偏微分してゼロとおくと次のようになる.

$$\frac{\partial \ell}{\partial \boldsymbol{W}_k} = \sum_{i=1}^{N} \frac{\partial \log f}{\partial \boldsymbol{W}_k} = \sum_{i=1}^{N} \frac{1}{f} \frac{\partial f}{\partial \boldsymbol{W}_k} = \sum_{i=1}^{N} \frac{Pr(k|\boldsymbol{x}_i)}{f_k} \frac{\partial f_k}{\partial \boldsymbol{W}_k} = \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) \frac{\partial \log f_k}{\partial \boldsymbol{W}_k} = \mathbf{o}$$

したがって,

$$\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) \frac{\partial}{\partial \boldsymbol{W}_k} \{\log \pi_k + \frac{1}{2} \log |\boldsymbol{W}_k| + \log h_k(s_{ki})\} = \mathbf{o}$$

これより

$$\begin{aligned}
&\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) [\frac{1}{2} \{2\boldsymbol{W}_k^{-1} - \mathrm{diag}\boldsymbol{W}_k^{-1}\} \\
&\quad -\frac{1}{2} w_{ki} \{2(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T - \mathrm{diag}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\}] \\
= &\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) [\boldsymbol{W}_k^{-1} - w_{ki}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T] \\
&\quad -\frac{1}{2} \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) [\mathrm{diag}\boldsymbol{W}_k^{-1} - w_{ki}\mathrm{diag}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T] \\
= &\ \mathbf{o}
\end{aligned}$$

となる. ここで, $w_{ki} = -2\dfrac{\partial}{\partial s_{ki}} \log h_k(s_{ki})$ である. したがって, 擬分散共分散行列の最尤推定量は次のようになる.

$$\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) \boldsymbol{W}_k^{-1} = \hat{\pi}_k N \boldsymbol{V}_k = \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) w_{ki} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T$$

$$\widehat{\boldsymbol{V}}_k = \frac{1}{\hat{\pi}_k N} \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) w_{ki} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T.$$

次に $\boldsymbol{\mu}_k$ は次のようになる.

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu}_k} \log f(\boldsymbol{x}_i) = \sum_{i=1}^{N} \frac{\pi_k}{f} \frac{\partial f_k}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) \frac{\partial \log f_k}{\partial \boldsymbol{\mu}_k} = \mathbf{o}$$

これより,

$$\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) \frac{\partial}{\partial \boldsymbol{\mu}_k} \{\log \pi_k + \frac{1}{2} \log |\boldsymbol{W}_k| + \log h_k(s_{ki})\} = \mathbf{o}$$

$$\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) \frac{\partial}{\partial s_{ki}} \log h_k(s_{ki}) \frac{\partial s_{ki}}{\partial \boldsymbol{\mu}_k} = \mathbf{o}$$

$$\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) w_{ki} 2\boldsymbol{W}_k (\boldsymbol{x}_i - \boldsymbol{\mu}_k) = \mathbf{o}$$

$$\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) w_{ki} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) = \mathbf{o}$$

したがって, 位置ベクトルの最尤推定量は

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{\sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) w_{ki}} \sum_{i=1}^{N} Pr(k|\boldsymbol{x}_i) w_{ki} \boldsymbol{x}_i$$

となる.

付録 C

ブートストラップ法による情報量規準

ここでは小西 (1992, 1993a, b) を参考に, ブートストラップ法による情報量規準の構成について簡単にまとめる.

C.1 ブートストラップ法によるバイアス修正

情報量規準の基本的な考え方を簡単に整理する. 未知の確率分布関数 $G(\boldsymbol{x})$ からの大きさ N の無作為標本を $\boldsymbol{X}_N$ とする. 確率分布関数 $G(\boldsymbol{x})$ の密度関数を $g(\boldsymbol{x})$ とし, これに対して想定したモデルの密度関数を $f(\boldsymbol{x}|\boldsymbol{\theta})$ とする. モデルに含まれる未知パラメータ $\boldsymbol{\theta}(\in \boldsymbol{\Theta})$ は, p 次元ベクトルとする. このような設定の下で, 将来観測されるデータ z に対する (予測) 確率分布 $\hat{f}(z|\boldsymbol{X}_N)$ を構成したい. その 1 つの方法は, 想定したモデルの確率分布 $f(\cdot)$ に含まれるパラメータ $\boldsymbol{\theta}$ を何らかの方法で推定し, これを求めた推定値 $\hat{\boldsymbol{\theta}}$ で置き換えた $\hat{f}(z|\boldsymbol{X}_N) = \hat{f}(z|\hat{\boldsymbol{\theta}})$ を用いることである.

大きさ N の標本 $\boldsymbol{X}_N$ に基づいて推定された 1 つの予測確率分布 $\hat{f}(z|\boldsymbol{X}_N)$ と, この標本を生成した真の確率分布 $G(z)(g(z))$ との距離を Kullback-Leibler 情報量で測るとする. このとき, 標本 $\boldsymbol{X}_N$ によって推定される種々の予測確率分布の違いは, 平均対数尤度と呼ばれる項

$$\eta(G;\boldsymbol{X}_N) = \int g(z)\log\hat{f}(z|\boldsymbol{X}_N)dz = \int \log\hat{f}(z|\boldsymbol{X}_N)dG(z) \tag{C.1}$$

が関係する. 平均対数尤度は, 真の確率分布 G と予測確率分布の推定を通して標本 $\boldsymbol{X}_N$ に依存する未知の量である. ここで, 平均対数尤度の 1 つの推定量としては, (C.1) 式に含まれる

未知の確率分布 G を標本 $\boldsymbol{X}_N$ に基づく経験分布関数 $\hat{G}$ で置き換えた

$$\hat{\eta}(\hat{G};\boldsymbol{X}_N)=\int \log \hat{f}(z|\boldsymbol{X}_N)d\hat{G}(z)=\frac{1}{N}\sum_{i=1}^{N}\log \hat{f}(x_i|\boldsymbol{X}_N) \tag{C.2}$$

を用いることができる.

予測確率分布を構成するために用いたデータを再び利用して, 未知の確率分布 $G(z)$ を推定していることから, 推定量 (C.2) は見かけ上の推定量であるといえる. つまり, 対数尤度という 1 つの推定量で平均対数尤度を推定したときのバイアス

$$\mathrm{bias}(G)=\mathrm{E}_G[\hat{\eta}(\hat{G};\boldsymbol{X}_N)-\eta(G;\boldsymbol{X}_N)]$$

の補正が必要になる. したがって, このバイアスを何らかの方法で推定できれば対数尤度のバイアスを補正した 1 つの情報量規準

$$\mathrm{IC}(\hat{G};\boldsymbol{X}_N)=\frac{1}{N}\sum_{i=1}^{N}\log \hat{f}(x_i|\boldsymbol{X}_N)-\{\mathrm{bias}(G)\text{の推定量}\}$$

が得られる. 情報量規準 AIC は基本的にはこのように対数尤度を推定したときの漸近的なバイアスを補正した推定量として与えられた.

ブートストラップ法を適用するとバイアスの推定は

$$\begin{aligned}\widehat{\mathrm{bias}}(\hat{G}) &= \mathrm{E}_{\hat{G}}[\hat{\eta}(\hat{G}^*;\boldsymbol{X}_N^*)-\eta(\hat{G};\boldsymbol{X}_N^*)] \\ &= \mathrm{E}_{\hat{G}}[\frac{1}{N}\sum_{i=1}^{N}\log \hat{f}(x_i^*|\boldsymbol{X}_N^*)-\frac{1}{N}\sum_{i=1}^{N}\log \hat{f}(x_i|\boldsymbol{X}_N^*)]\end{aligned} \tag{C.3}$$

となる. ただし, $\boldsymbol{X}_N^*$ は大きさ N のブートストラップ標本, $\hat{G}^*$ はブートストラップ標本の各点に確率 $1/N$ をもつ経験分布関数とする. 期待値はモンテカルロ法によって数値的に計算される.

C.2 バイアス推定の変動減少法

対数尤度のバイアス推定をブートストラップ法を用いて求めるとき, その変動を減少させる方法がある. これは北川 (1991) により数値的にその良さが指摘され, その理論は小西 (1993a) により示されている. それは (C.3) 式におけるバイアス推定を以下のように補正し

たものである.

$$\begin{aligned}\widehat{\mathrm{bias}}_M(\hat{G}) &= \mathrm{E}_{\hat{G}}[\frac{1}{N}\sum_{i=1}^{N}\log\hat{f}(x_i^*|\boldsymbol{X}_N^*) - \frac{1}{N}\sum_{i=1}^{N}\log f(x_i^*|\boldsymbol{X}_N) \\ &+ \frac{1}{N}\sum_{i=1}^{N}\log f(x_i|\boldsymbol{X}_N) - \frac{1}{N}\sum_{i=1}^{N}\log\hat{f}(x_i|\boldsymbol{X}_N^*)]. \qquad \text{(C.4)}\end{aligned}$$

(C.3) 式の期待値の推定をブートストラップ法で行ったとき, その漸近分散の標本に関するオーダーは多くの場合 $O(1/N)$ とすることができ, また (C.4) 式の場合は $O(1/N^2)$ とできることが示されている. つまり, ブートストラップバイアス推定における変動を減少させることができることを示している. これはブートストラップ法を実行するにあたって, その回数を減らすことができることを意味し, 混合分布モデルにおける EM 法の収束の遅さを補うものである.

付録 D

キバハリアリの部位計測データ

第9章のデータ解析で用いたキバハリアリの生データを示す．表D.1には部位測定データを，表D.2には比率変換したデータを示す．

表 D.1: キバハリアリの部位測定データ

ID-SpeciesName	SPGroup Name	HW	HL	SL	ML	WL	PrW	HFL	PtW	PtL	PpW	PpL	GW
01aberrans	aberrans	59	58	40	47	84	40	60	20	25	34	26	50
02aberrans	aberrans	53	53	37	40	77	34	58	18	22	30	24	45
03aberrans	aberrans	63	61	41	48	89	40	65	22	25	38	28	55
01excavata	aberrans	57	55	37	44	82	39	56	20	25	33	22	48
02excavata	aberrans	53	52	35	41	75	35	50	17	22	30	21	43
01froggatti	aberrans	66	61	45	51	95	42	66	23	27	40	25	55
01maura	aberrans	69	67	43	50	90	44	69	25	27	40	26	59
02maura	aberrans	62	62	41	45	88	40	65	21	25	37	22	55
03maura	aberrans	65	63	40	50	–	–	–	–	–	38	–	56
01formosa	aberrans	62	60	41	47	87	40	65	20	24	34	25	48
02formosa	aberrans	66	65	41	49	91	43	68	23	27	40	27	55
01nobilis	aberrans	69	65	45	49	95	45	67	23	26	40	27	61
02nobilis	aberrans	55	55	40	42	89	42	62	22	22	40	26	59
03nobilis	aberrans	60	57	40	45	86	39	65	18	24	35	24	50
01callima	cephalotes	53	45	46	52	84	36	65	20	24	31	22	46
02callima	cephalotes	49	44	45	49	84	33	65	19	24	29	20	45
03callima	cephalotes	55	48	47	52	85	36	67	20	25	32	21	49
04callima	cephalotes	53	47	45	52	84	37	66	20	25	31	19	48
01hilli	cephalotes	55	50	50	54	92	38	60	22	25	35	24	50
02hilli	cephalotes	55	50	49	53	91	39	70	20	26	35	25	50
03hilli	cephalotes	68	58	55	59	–	49	80	29	32	47	30	67
04hilli	cephalotes	63	–	53	–	–	–	–	–	–	–	–	–
05hilli	cephalotes	56	–	50	–	–	–	–	–	–	–	–	–
01cephalotes	cephalotes	57	50	46	51	93	40	69	23	28	42	23	59
02cephalotes	cephalotes	51	–	46	–	–	–	–	–	–	–	–	–
03cephalotes	cephalotes	61	–	50	–	–	–	–	–	–	–	–	–
01arnoldi	gulosa	75	77	98	85	–	46	–	20	50	31	30	74
02arnoldi	gulosa	74	74	90	80	–	47	–	23	50	33	28	75
03arnoldi	gulosa	73	74	95	82	–	46	–	22	49	33	27	72
04arnoldi	gulosa	71	72	92	80	–	43	–	22	47	31	27	68
01auriventris	gulosa	70	66	65	68	–	46	–	22	33	35	26	69
02auriventris	gulosa	66	62	60	60	–	41	–	20	32	33	24	62
03auriventris	gulosa	70	67	65	66	–	44	–	21	32	33	26	66

キバハリアリの部位測定データ (つづき)

ID-SpeciesName	SPGroup Name	HW	HL	SL	ML	WL	PrW	HFL	PtW	PtL	PpW	PpL	GW
04auriventris	gulosa	64	60	59	61	–	40	–	18	31	30	24	57
05auriventris	gulosa	67	65	65	66	–	44	–	20	36	34	29	64
06auriventris	gulosa	75	71	66	70	–	51	–	24	40	38	30	71
01atrata	gulosa	70	73	90	80	–	44	–	21	51	32	28	68
02atrata	gulosa	69	71	90	76	–	46	–	23	50	34	32	69
03atrata	gulosa	62	64	83	72	–	36	–	17	44	27	27	57
04atrata	gulosa	74	77	94	85	–	47	–	24	55	35	35	73
05atrata	gulosa	51	55	70	61	–	34	–	16	36	26	25	55
06atrata	gulosa	83	84	103	99	–	52	–	25	60	40	37	85
01brevinoda	gulosa	65	64	67	67	–	41	–	24	37	36	27	67
02brevinoda	gulosa	46	48	49	49	91	30	72	18	29	–	–	–
03brevinoda	gulosa	66	66	68	75	–	44	–	25	37	37	28	67
01cardigaster	gulosa	54	57	55	62	–	36	–	16	23	26	24	56
01comata	gulosa	83	81	76	85	–	56	–	29	48	42	35	80
02comata	gulosa	74	76	73	80	–	52	–	27	43	39	31	71
01decipians	gulosa	82	76	85	87	–	55	–	28	48	42	33	78
02decipians	gulosa	72	72	80	82	–	47	–	25	45	38	28	73
03decipians	gulosa	90	86	90	93	–	58	–	30	50	48	37	87
01desertorum	gulosa	78	81	100	86	–	52	–	26	53	38	30	78
02desertorum	gulosa	70	74	90	79	–	45	–	24	50	34	28	74
03desertorum	gulosa	80	82	97	87	–	55	–	29	56	42	33	84
04desertorum	gulosa	88	87	–	–	–	54	–	29	65	42	36	85
05desertorum	gulosa	87	90	–	–	–	55	–	30	67	40	36	85
06desertorum	gulosa	66	70	90	79	–	41	–	21	47	33	30	67
07desertorum	gulosa	52	55	71	60	–	35	–	17	40	27	20	55
01dimidiata	gulosa	80	83	–	95	–	51	–	26	56	40	33	79
02dimidiata	gulosa	75	78	93	87	–	48	–	24	53	36	31	73
03dimidiata	gulosa	75	77	91	83	–	45	–	24	54	35	30	74
01esuriens	gulosa	58	59	64	57	–	40	88	25	31	39	30	64
02esuriens	gulosa	64	66	67	59	–	43	90	28	30	44	33	70
03esuriens	gulosa	55	58	60	50	–	38	63	25	30	39	30	61
04esuriens	gulosa	69	67	65	60	–	50	92	30	36	47	37	75
05esuriens	gulosa	53	53	57	48	–	40	75	23	28	35	27	58
01fasciata	gulosa	68	72	91	75	–	44	–	24	49	34	32	69
02fasciata	gulosa	72	77	93	84	–	46	–	25	53	35	33	73
03fasciata	gulosa	78	81	97	86	–	50	–	27	56	40	34	80
01minuscula	gulosa	68	69	75	70	–	42	–	21	38	34	30	62
01forceps	gulosa	95	90	–	–	–	52	–	26	55	40	35	82
02forceps	gulosa	84	81	98	100	–	45	–	24	48	35	33	72
03forceps	gulosa	72	68	86	79	–	40	–	20	40	31	25	68
01forficata	gulosa	69	68	67	67	–	45	–	28	44	40	30	75
02forficata	gulosa	60	62	67	65	–	40	–	22	39	34	25	65
01fulgida	gulosa	87	88	–	95	–	57	–	30	57	42	37	87
02fulgida	gulosa	79	80	93	80	–	52	–	27	49	38	33	78
03fulgida	gulosa	88	88	–	92	–	60	–	30	57	43	38	90
04fulgida	gulosa	85	85	–	–	–	–	–	–	–	42	–	89
05fulgida	gulosa	82	83	–	–	–	–	–	–	–	39	–	82
01gratiosa	gulosa	68	73	94	81	–	44	–	24	51	35	30	70
02gratiosa	gulosa	69	73	91	82	–	45	–	24	52	35	29	70
01gulosa	gulosa	77	79	88	82	–	52	–	25	47	40	34	73
01hirsuta	gulosa	79	78	93	92	–	52	–	28	50	40	31	75
02hirsuta	gulosa	62	63	76	74	–	40	–	24	43	34	27	65
03hirsuta	gulosa	59	61	75	68	–	40	–	21	39	29	26	57
01loginodis	gulosa	77	76	80	85	–	48	–	26	48	39	31	74
02loginodis	gulosa	55	55	57	58	–	35	–	20	35	29	23	58
03loginodis	gulosa	80	79	83	90	–	50	–	26	48	40	35	75
01midas	gulosa	64	60	60	63	–	39	92	21	33	35	23	63
02midas	gulosa	59	58	56	60	–	38	87	20	30	31	23	57
03midas	gulosa	58	57	58	55	–	38	–	20	–	–	–	–
04midas	gulosa	–	–	–	–	–	–	–	20	32	32	26	61

キバハリアリの部位測定データ (つづき)

ID-SpeciesName	SPGroup Name	HW	HL	SL	ML	WL	PrW	HFL	PtW	PtL	PpW	PpL	GW
01mjobergi	gulosa	70	83	99	–	–	43	–	25	55	40	33	79
01nigriceps	gulosa	75	79	98	90	–	50	–	26	55	36	35	75
02nigriceps	gulosa	74	80	97	87	–	49	–	25	53	35	33	74
01nigriscapa	gulosa	69	71	80	76	–	45	–	25	41	36	29	74
01pavida	gulosa	75	81	–	91	–	50	–	25	55	37	34	72
02pavida	gulosa	61	64	82	66	–	40	–	20	43	28	26	59
01picticeps	gulosa	68	66	72	64	–	43	–	24	36	36	25	71
02picticeps	gulosa	73	71	77	70	–	47	–	25	40	38	27	75
03picticeps	gulosa	69	68	74	66	–	45	–	24	38	38	25	73
01pulchra	gulosa	73	68	68	63	–	45	95	25	36	39	32	70
02pulchra	gulosa	68	65	65	62	–	45	92	23	32	38	30	69
03pulchra	gulosa	77	73	75	71	–	49	–	28	38	43	37	77
04pulchra	gulosa	63	60	64	59	–	42	90	22	30	35	30	63
05pulchra	gulosa	58	58	57	54	–	37	82	20	29	35	29	61
06pulchra	gulosa	74	70	73	68	–	45	–	25	38	42	35	74
01pyriformis	gulosa	63	63	69	73	–	40	–	22	41	32	30	65
02pyriformis	gulosa	67	70	71	75	–	45	–	25	45	35	31	70
03pyriformis	gulosa	76	75	80	84	–	50	–	29	47	39	34	71
04pyriformis	gulosa	60	60	67	62	–	39	–	20	38	30	27	58
01regularis	gulosa	58	59	61	65	–	38	–	20	35	30	25	60
02regularis	gulosa	59	60	63	63	–	38	–	20	36	30	26	58
03regularis	gulosa	66	65	68	67	–	42	–	22	40	34	30	65
04regularis	gulosa	67	66	70	68	–	44	–	23	40	35	28	66
01roWLandi	gulosa	58	63	60	69	–	40	–	18	36	30	25	60
02roWLandi	gulosa	65	67	64	75	–	44	–	20	38	34	28	65
03roWLandi	gulosa	66	68	65	77	–	45	–	20	40	33	29	65
01rubripes	gulosa	80	80	100	88	–	50	–	25	50	38	33	85
02rubripes	gulosa	75	74	99	80	–	46	–	20	49	32	34	70
03rubripes	gulosa	70	69	95	78	–	45	–	20	46	31	31	69
04rubripes	gulosa	70	71	94	77	–	43	–	20	47	30	30	67
01rufinodis	gulosa	67	70	85	72	–	44	–	20	45	30	26	68
02rufinodis	gulosa	41	47	52	41	80	27	80	14	29	20	19	–
03rufinodis	gulosa	44	48	59	43	–	30	82	13	–	20	–	50
04rufinodis	gulosa	80	80	96	82	–	52	–	26	51	39	32	84
01suttoni	gulosa	80	82	100	85	–	51	–	26	57	37	35	85
02suttoni	gulosa	88	88	–	95	–	56	–	33	60	45	42	93
03suttoni	gulosa	71	75	94	82	–	45	–	23	51	–	30	70
04suttoni	gulosa	–	–	–	–	–	–	–	–	–	35	–	80
01tarsata	gulosa	79	78	80	84	–	55	–	30	46	45	40	83
02tarsata	gulosa	–	–	–	–	–	51	–	30	45	45	38	80
03tarsata	gulosa	74	73	77	76	–	53	–	27	–	42	–	77
04tarsata	gulosa	65	64	73	64	–	46	–	25	40	36	30	67
05tarsata	gulosa	51	53	57	52	–	34	82	19	32	29	25	52
06tarsata	gulosa	90	83	–	–	–	–	–	32	50	47	37	85
01vindex	gulosa	75	76	98	86	–	49	–	26	54	39	31	80
02vindex	gulosa	77	78	99	86	–	49	–	24	53	36	31	74
01basirufa	gulosa	77	80	99	83	–	45	–	27	53	38	32	75
01simillima	gulosa	83	81	89	91	–	50	–	25	46	40	34	73
02simillima	gulosa	85	82	87	90	–	51	–	27	46	40	32	78
03simillima	gulosa	50	51	55	49	94	32	82	16	27	24	22	51
01analis	gulosa	63	62	73	70	–	42	–	22	38	31	27	65
02analis	gulosa	63	65	75	63	–	42	–	20	38	32	25	66
03analis	gulosa	61	60	74	68	–	41	–	20	36	29	26	61
04analis	gulosa	67	65	75	75	–	43	–	22	42	31	30	65
05analis	gulosa	63	64	77	71	–	41	–	22	39	31	27	65
06analis	gulosa	65	66	78	73	–	42	–	23	39	31	30	64
01gilberti	mandibularis	54	50	41	50	89	39	62	25	30	38	25	50
02gilberti	mandibularis	52	49	40	49	85	38	55	23	27	35	25	48
01luteiforceps	mandibularis	49	45	38	50	85	35	55	20	25	34	22	45
02luteiforceps	mandibularis	50	46	38	52	82	36	54	22	27	34	22	45

キバハリアリの部位測定データ (つづき)

ID-SpeciesName	SPGroup Name	HW	HL	SL	ML	WL	PrW	HFL	PtW	PtL	PpW	PpL	GW
01potteri	mandibularis	51	48	40	47	83	39	60	22	25	34	22	47
02potteri	mandibularis	51	48	39	46	85	38	59	23	27	36	23	47
01pilventris	mandibularis	59	54	49	55	95	41	66	25	30	42	30	55
02pilventris	mandibularis	64	58	50	59	–	48	73	28	34	47	34	61
03pilventris	mandibularis	42	39	33	40	56	30	45	19	21	28	21	40
01fulviculis	mandibularis	48	48	40	50	78	36	55	22	28	34	28	41
02fulviculis	mandibularis	50	48	40	50	80	37	57	24	30	35	28	45
01fulvipes	mandibularis	57	49	43	53	88	42	62	25	28	39	30	52
02fulvipes	mandibularis	58	50	44	56	89	–	62	25	27	39	30	53
01mandibularis	mandibularis	63	56	51	62	97	45	73	28	34	43	33	58
02mandibularis	mandibularis	57	50	48	59	92	42	64	23	29	38	30	50
03mandibularis	mandibularis	50	45	40	54	78	35	55	18	25	32	25	45
04mandibularis	mandibularis	61	54	50	63	94	44	69	25	31	40	30	53
01flammicollis	nigrocincta	46	49	55	53	94	30	64	15	33	22	20	43
02flammicollis	nigrocincta	55	56	65	59	–	35	93	19	37	27	25	50
01nigrocincta	nigrocincta	45	48	54	51	93	32	80	14	28	24	22	41
02nigrocincta	nigrocincta	54	55	60	57	–	37	92	19	32	31	26	50
03nigrocincta	nigrocincta	43	46	51	47	90	28	80	15	27	23	20	38
01petiolata	nigrocincta	44	45	–	49	85	30	–	15	26	23	20	40
02petiolata	nigrocincta	42	45	50	49	87	30	–	14	27	22	19	40
03petiolata	nigrocincta	45	45	50	49	–	30	72	–	–	21	–	40
04petiolata	nigrocincta	41	43	47	47	–	29	70	13	23	20	17	37
01fucosa	picta	42	45	34	41	69	27	54	18	25	25	18	40
02fucosa	picta	45	45	36	43	73	30	56	19	26	27	23	45
01picta	picta	44	47	37	43	78	30	59	18	22	28	20	43
02picta	picta	43	45	36	40	74	30	57	20	25	29	22	45
03picta	picta	45	47	36	42	75	32	50	20	–	30	–	44
04picta	picta	38	40	33	38	–	25	51	16	23	25	19	38
01apicalis	pilosula	44	44	45	43	83	32	62	15	23	27	17	48
01celaena	pilosula	42	39	35	40	64	29	53	16	19	26	18	38
01chasei	pilosula	60	51	50	52	96	45	81	27	30	45	28	67
02chasei	pilosula	51	46	48	50	85	38	74	20	25	32	20	49
01ludlowi	pilosula	64	58	50	55	96	48	80	29	28	47	30	66
02ludlowi	pilosula	49	46	45	47	84	37	70	22	23	34	23	50
03ludlowi	pilosula	57	51	50	52	92	40	75	25	28	40	26	58
01chrysogaster	pilosula	44	42	32	39	75	33	48	18	20	30	20	43
01cydista	pilosula	43	41	36	39	70	31	–	19	22	30	18	43
01dispar	pilosula	49	46	43	48	72	33	57	19	23	29	19	45
02dispar	pilosula	49	48	44	43	75	32	58	20	24	30	20	45
03dispar	pilosula	43	43	42	44	71	29	–	18	21	26	18	41
01elegans	pilosula	51	51	45	55	87	37	65	23	25	35	24	50
02elegans	pilosula	44	43	40	50	74	30	59	20	21	29	20	42
03elegans	pilosula	42	43	40	48	72	30	56	19	20	28	20	40
01harderi	pilosula	46	44	38	44	75	34	59	20	24	33	22	47
02harderi	pilosula	54	50	42	47	84	37	62	24	26	37	23	51
03harderi	pilosula	47	45	38	42	73	33	–	19	23	32	22	47
01michaelseni	pilosula	52	49	40	49	86	38	61	20	25	39	25	54
02michaelseni	pilosula	49	45	38	47	82	37	55	19	25	34	26	47
03michaelseni	pilosula	50	45	–	–	–	35	–	19	26	35	24	47
04michaelseni	pilosula	55	50	43	50	88	41	63	22	27	40	28	55
01occidentalis	pilosula	48	45	44	45	78	32	59	20	25	30	21	45
02occidentalis	pilosula	49	45	43	44	80	35	56	20	26	30	23	45
03occidentalis	pilosula	46	45	40	43	75	33	57	–	–	30	–	47
04occidentalis	pilosula	47	–	45	–	–	–	–	–	–	31	–	48
05occidentalis	pilosula	47	46	–	–	–	–	–	19	–	30	–	46
01opaca	pilosula	44	42	40	42	72	31	53	18	21	28	20	42
01rugosa	pilosula	55	50	45	49	90	40	65	25	27	42	28	55
02rugosa	pilosula	50	48	40	46	82	38	60	22	25	38	25	50
03rugosa	pilosula	43	40	35	40	73	34	50	18	22	31	21	43
01varians	pilosula	42	40	38	46	72	30	57	17	23	25	20	39

キバハリアリの部位測定データ (つづき)

ID-SpeciesName	SPGroup Name	HW	HL	SL	ML	WL	PrW	HFL	PtW	PtL	PpW	PpL	GW
02varians	pilosula	48	47	41	49	78	34	62	21	26	30	25	44
03varians	pilosula	43	43	38	45	75	30	57	18	24	27	23	40
04varians	pilosula	47	46	40	47	79	31	62	18	25	27	23	43
05varians	pilosula	45	45	39	48	80	32	60	19	25	30	24	44
06varians	pilosula	47	46	40	48	78	33	61	19	27	30	25	45
01pilosula	pilosula	–	–	–	–	79	35	60	20	25	30	21	46
02pilosula	pilosula	37	37	38	40	69	27	50	15	20	24	18	37
03pilosula	pilosula	48	44	42	45	80	35	57	20	25	30	24	49
04pilosula	pilosula	56	55	48	53	89	40	69	23	27	35	27	55
05pilosula	pilosula	58	55	49	55	92	42	70	23	30	36	26	56
01clarki	tepperi	44	43	40	49	72	30	51	20	21	29	20	40
02clarki	tepperi	49	47	43	52	78	34	55	23	25	34	23	49
01dixoni	tepperi	42	40	31	42	65	30	–	19	20	29	18	41
02dixoni	tepperi	46	41	34	42	70	33	43	22	23	30	21	44
03dixoni	tepperi	40	38	30	40	59	28	40	17	18	26	16	37
01swalei	tepperi	38	35	30	40	56	25	36	15	18	23	18	35
02swalei	tepperi	48	44	35	45	72	33	47	22	24	31	23	45
03swalei	tepperi	40	38	32	40	61	29	40	19	20	28	22	39
04swalei	tepperi	45	42	35	45	–	32	–	20	22	31	20	43
05swalei	tepperi	37	34	28	37	55	26	36	15	17	24	16	34
06swalei	tepperi	51	49	40	49	83	40	58	24	25	38	26	52
01tepperi	tepperi	42	40	36	46	65	29	47	15	21	25	20	37
02tepperi	tepperi	–	–	–	–	–	35	–	22	25	32	25	47
03tepperi	tepperi	50	49	46	50	82	35	60	22	25	35	24	48
04tepperi	tepperi	53	50	45	52	82	36	60	23	25	35	25	50
01testaceipes	tepperi	43	40	32	43	65	30	42	19	21	29	20	40
02testaceipes	tepperi	44	40	34	43	65	30	44	19	21	30	20	42
03testaceipes	tepperi	38	35	30	40	58	27	39	15	18	25	18	36
01dichospila	urens	34	33	23	26	62	25	39	15	18	24	15	34
02dichospila	urens	30	30	24	28	53	20	39	12	16	18	15	27
03dichospila	urens	26	27	21	25	48	18	34	10	14	16	13	24
01exigua	urens	23	24	19	23	41	15	29	10	14	16	10	22
01rubicunda	urens	25	23	19	24	41	15	29	10	14	17	10	25
01infima	urens	30	32	25	27	55	20	43	11	16	22	16	30
02infima	urens	30	33	25	28	54	20	43	11	17	22	15	29
03infima	urens	32	33	25	30	59	23	–	13	15	23	15	34
04infima	urens	34	35	26	31	65	24	–	17	20	28	20	38
05infima	urens	20	25	19	21	44	17	–	9	15	15	13	25
06infima	urens	25	25	19	24	–	–	–	–	–	17	–	25

表 D.2: キバハリアリの比率変数データ

ID-Species Name	SpG No.*	CI= HW/HL	SI= SL/HW	MI= ML/HL	LI=HFL /PrW	PpI= PpW/GW	PtW /PtL	PpW /PpL	PtW /GW	PrW /GW	HW /GW
01aberrans	1	101.72	67.80	81.03	150.00	68.00	80.00	130.77	40.00	80.00	118.00
02aberrans	1	100.00	69.81	75.47	170.59	66.67	81.82	125.00	40.00	75.56	117.78
03aberrans	1	103.28	65.08	78.69	162.50	69.09	88.00	135.71	40.00	72.73	114.55
01excavata	1	103.64	64.91	80.00	143.59	68.75	80.00	150.00	41.67	81.25	118.75
02excavata	1	101.92	66.04	78.85	142.86	69.77	77.27	142.86	39.53	81.40	123.26
01froggatti	1	108.20	68.18	83.61	157.14	72.73	85.19	160.00	41.82	76.36	120.00
01maura	1	102.99	62.32	74.63	156.82	67.80	92.59	153.85	42.37	74.58	116.95
02maura	1	100.00	66.13	72.58	162.50	67.27	84.00	168.18	38.18	72.73	112.73
03maura	1	103.17	61.54	79.37	–	67.86	–	–	–	–	116.07
01formosa	1	103.33	66.13	78.33	162.50	70.83	83.33	136.00	41.67	83.33	129.17
02formosa	1	101.54	62.12	75.38	158.14	72.73	85.19	148.15	41.82	78.18	120.00
01nobilis	1	106.15	65.22	75.38	148.89	65.57	88.46	148.15	37.70	73.77	113.11
02nobilis	1	100.00	72.73	76.36	147.62	67.80	100.00	153.85	37.29	71.19	93.22
03nobilis	1	105.26	66.67	78.95	166.67	70.00	75.00	145.83	36.00	78.00	120.00
01callima	2	117.78	86.79	115.56	180.56	67.39	83.33	140.91	43.48	78.26	115.22
02callima	2	111.36	91.84	111.36	196.97	64.44	79.17	145.00	42.22	73.33	108.89
03callima	2	114.58	85.45	108.33	186.11	65.31	80.00	152.38	40.82	73.47	112.24
04callima	2	112.77	84.91	110.64	178.38	64.58	80.00	163.16	41.67	77.08	110.42
01hilli	2	110.00	90.91	108.00	157.89	70.00	88.00	145.83	44.00	76.00	110.00
02hilli	2	110.00	89.09	106.00	179.49	70.00	76.92	140.00	40.00	78.00	110.00
03hilli	2	117.24	80.88	101.72	163.27	70.15	90.62	156.67	43.28	73.13	101.49
04hilli	2	–	84.13	–	–	–	–	–	–	–	–
05hilli	2	–	89.29	–	–	–	–	–	–	–	–
01cephalotes	2	114.00	80.70	102.00	172.50	71.19	82.14	182.61	38.98	67.80	96.61
02cephalotes	2	–	90.20	–	–	–	–	–	–	–	–
03cephalotes	2	–	81.97	–	–	–	–	–	–	–	–
01arnoldi	3	97.40	130.67	110.39	–	41.89	40.00	103.33	27.03	62.16	101.35
02arnoldi	3	100.00	121.62	108.11	–	44.00	46.00	117.86	30.67	62.67	98.67
03arnoldi	3	98.65	130.14	110.81	–	45.83	44.90	122.22	30.56	63.89	101.39
04arnoldi	3	98.61	129.58	111.11	–	45.59	46.81	114.81	32.35	63.24	104.41
01auriventris	3	106.06	92.86	103.03	–	50.72	66.67	134.62	31.88	66.67	101.45
02auriventris	3	106.45	90.91	96.77	–	53.23	62.50	137.50	32.26	66.13	106.45
03auriventris	3	104.48	92.86	98.51	–	50.00	65.62	126.92	31.82	66.67	106.06
04auriventris	3	106.67	92.19	101.67	–	52.63	58.06	125.00	31.58	70.18	112.28
05auriventris	3	103.08	97.01	101.54	–	53.12	55.56	117.24	31.25	68.75	104.69
06auriventris	3	105.63	88.00	98.59	–	53.52	60.00	126.67	33.80	71.83	105.63
01atrata	3	95.89	128.57	109.59	–	47.06	41.18	114.29	30.88	64.71	102.94
02atrata	3	97.18	130.43	107.04	–	49.28	46.00	106.25	33.33	66.67	100.00
03atrata	3	96.88	133.87	112.50	–	47.37	38.64	100.00	29.82	63.16	108.77
04atrata	3	96.10	127.03	110.39	–	47.95	43.64	100.00	32.88	64.38	101.37
05atrata	3	92.73	137.25	110.91	–	47.27	44.44	104.00	29.09	61.82	92.73
06atrata	3	98.81	124.10	117.86	–	47.06	41.67	108.11	29.41	61.18	97.65
01brevinoda	3	101.56	103.08	104.69	–	53.73	64.86	133.33	35.82	61.19	97.01
02brevinoda	3	95.83	106.52	102.08	240.00	–	62.07	–	–	–	–
03brevinoda	3	100.00	103.03	113.64	–	55.22	67.57	132.14	37.31	65.67	98.51
01cardigaster	3	94.74	101.85	108.77	–	46.43	69.57	108.33	28.57	64.29	96.43
01comata	3	102.47	91.57	104.94	–	52.50	60.42	120.00	36.25	70.00	103.75
02comata	3	97.37	98.65	105.26	–	54.93	62.79	125.81	38.03	73.24	104.23
01decipians	3	107.89	103.66	114.47	–	53.85	58.33	127.27	35.90	70.51	105.13
02decipians	3	100.00	111.11	113.89	–	52.05	55.56	135.71	34.25	64.38	98.63
03decipians	3	104.65	100.00	108.14	–	55.17	60.00	129.73	34.48	66.67	103.45
01desertorum	3	96.30	128.21	106.17	–	48.72	49.06	126.67	33.33	66.67	100.00
02desertorum	3	94.59	128.57	106.76	–	45.95	48.00	121.43	32.43	60.81	94.59
03desertorum	3	97.56	121.25	106.10	–	50.00	51.79	127.27	34.52	65.48	95.24
04desertorum	3	101.15	–	–	–	49.41	44.62	116.67	34.12	63.53	103.53
05desertorum	3	96.67	–	–	–	47.06	44.78	111.11	35.29	64.71	102.35
06desertorum	3	94.29	136.36	112.86	–	49.25	44.68	110.00	31.34	61.19	98.51
07desertorum	3	94.55	136.54	109.09	–	49.09	42.50	135.00	30.91	63.64	94.55
01dimidiata	3	96.39	–	114.46	–	50.63	46.43	121.21	32.91	64.56	101.27
02dimidiata	3	96.15	124.00	111.54	–	49.32	45.28	116.13	32.88	65.75	102.74

* SpG No. (Species Group Number) は表 9.2の種群番号を参照.

キバハリアリの比率変数データ (つづき)

ID-Species Name	SpG No.	CI= HW/HL	SI= SL/HW	MI= ML/HL	LI=HFL /PrW	PpI= PpW/GW	PtW /PtL	PpW /PpL	PtW /GW	PrW /GW	HW /GW
03dimidiata	3	97.40	121.33	107.79	–	47.30	44.44	116.67	32.43	60.81	101.35
01esuriens	3	98.31	110.34	96.61	220.00	60.94	80.65	130.00	39.06	62.50	90.62
02esuriens	3	96.97	104.69	89.39	209.30	62.86	93.33	133.33	40.00	61.43	91.43
03esuriens	3	94.83	109.09	86.21	165.79	63.93	83.33	130.00	40.98	62.30	90.16
04esuriens	3	102.99	94.20	89.55	184.00	62.67	83.33	127.03	40.00	66.67	92.00
05esuriens	3	100.00	107.55	90.57	187.50	60.34	82.14	129.63	39.66	68.97	91.38
01fasciata	3	94.44	133.82	104.17	–	49.28	48.98	106.25	34.78	63.77	98.55
02fasciata	3	93.51	129.17	109.09	–	47.95	47.17	106.06	34.25	63.01	98.63
03fasciata	3	96.30	124.36	106.17	–	50.00	48.21	117.65	33.75	62.50	97.50
01minuscula	3	98.55	110.29	101.45	–	54.84	55.26	113.33	33.87	67.74	109.68
01forceps	3	105.56	–	–	–	48.78	47.27	114.29	31.71	63.41	115.85
02forceps	3	103.70	116.67	123.46	–	48.61	50.00	106.06	33.33	62.50	116.67
03forceps	3	105.88	119.44	116.18	–	45.59	50.00	124.00	29.41	58.82	105.88
01forficata	3	101.47	97.10	98.53	–	53.33	63.64	133.33	37.33	60.00	92.00
02forficata	3	96.77	111.67	104.84	–	52.31	56.41	136.00	33.85	61.54	92.31
01fulgida	3	98.86	–	107.95	–	48.28	52.63	113.51	34.48	65.52	100.00
02fulgida	3	98.75	117.72	100.00	–	48.72	55.10	115.15	34.62	66.67	101.28
03fulgida	3	100.00	–	104.55	–	47.78	52.63	113.16	33.33	66.67	97.78
04fulgida	3	100.00	–	–	–	47.19	–	–	–	–	95.51
05fulgida	3	98.80	–	–	–	47.56	–	–	–	–	100.00
01gratiosa	3	93.15	138.24	110.96	–	50.00	47.06	116.67	34.29	62.86	97.14
02gratiosa	3	94.52	131.88	112.33	–	50.00	46.15	120.69	34.29	64.29	98.57
01gulosa	3	97.47	114.29	103.80	–	54.79	53.19	117.65	34.25	71.23	105.48
01hirsuta	3	101.28	117.72	117.95	–	53.33	56.00	129.03	37.33	69.33	105.33
02hirsuta	3	98.41	122.58	117.46	–	52.31	55.81	125.93	36.92	61.54	95.38
03hirsuta	3	96.72	127.12	111.48	–	50.88	53.85	111.54	36.84	70.18	103.51
01loginodis	3	101.32	103.90	111.84	–	52.70	54.17	125.81	35.14	64.86	104.05
02loginodis	3	100.00	103.64	105.45	–	50.00	57.14	126.09	34.48	60.34	94.83
03loginodis	3	101.27	103.75	113.92	–	53.33	54.17	114.29	34.67	66.67	106.67
01midas	3	106.67	93.75	105.00	235.90	55.56	63.64	152.17	33.33	61.90	101.59
02midas	3	101.72	94.92	103.45	228.95	54.39	66.67	134.78	35.09	66.67	103.51
03midas	3	101.75	100.00	96.49	–	–	–	–	–	–	–
04midas	3	–	–	–	–	52.46	62.50	123.08	32.79	–	–
01mjobergi	3	84.34	141.43	–	–	50.63	45.45	121.21	31.65	54.43	88.61
01nigriceps	3	94.94	130.67	113.92	–	48.00	47.27	102.86	34.67	66.67	100.00
02nigriceps	3	92.50	131.08	108.75	–	47.30	47.17	106.06	33.78	66.22	100.00
01nigriscapa	3	97.18	115.94	107.04	–	48.65	60.98	124.14	33.78	60.81	93.24
01pavida	3	92.59	–	112.35	–	51.39	45.45	108.82	34.72	69.44	104.17
02pavida	3	95.31	134.43	103.12	–	47.46	46.51	107.69	33.90	67.80	103.39
01picticeps	3	103.03	105.88	96.97	–	50.70	66.67	144.00	33.80	60.56	95.77
02picticeps	3	102.82	105.48	98.59	–	50.67	62.50	140.74	33.33	62.67	97.33
03picticeps	3	101.47	107.25	97.06	–	52.05	63.16	152.00	32.88	61.64	94.52
01pulchra	3	107.35	93.15	92.65	211.11	55.71	69.44	121.88	35.71	64.29	104.29
02pulchra	3	104.62	95.59	95.38	204.44	55.07	71.88	126.67	33.33	65.22	98.55
03pulchra	3	105.48	97.40	97.26	–	55.84	73.68	116.22	36.36	63.64	100.00
04pulchra	3	105.00	101.59	98.33	214.29	55.56	73.33	116.67	34.92	66.67	100.00
05pulchra	3	100.00	98.28	93.10	221.62	57.38	68.97	120.69	32.79	60.66	95.08
06pulchra	3	105.71	98.65	97.14	–	56.76	65.79	120.00	33.78	60.81	100.00
01pyriformis	3	100.00	109.52	115.87	–	49.23	53.66	106.67	33.85	61.54	96.92
02pyriformis	3	95.71	105.97	107.14	–	50.00	55.56	112.90	35.71	64.29	95.71
03pyriformis	3	101.33	105.26	112.00	–	54.93	61.70	114.71	40.85	70.42	107.04
04pyriformis	3	100.00	111.67	103.33	–	51.72	52.63	111.11	34.48	67.24	103.45
01regularis	3	98.31	105.17	110.17	–	50.00	57.14	120.00	33.33	63.33	96.67
02regularis	3	98.33	106.78	105.00	–	51.72	55.56	115.38	34.48	65.52	101.72
03regularis	3	101.54	103.03	103.08	–	52.31	55.00	113.33	33.85	64.62	101.54
04regularis	3	101.52	104.48	103.03	–	53.03	57.50	125.00	34.85	66.67	101.52
01roWLandi	3	92.06	103.45	109.52	–	50.00	50.00	120.00	30.00	66.67	96.67
02roWLandi	3	97.01	98.46	111.94	–	52.31	52.63	121.43	30.77	67.69	100.00
03roWLandi	3	97.06	98.48	113.24	–	50.77	50.00	113.79	30.77	69.23	101.54
01rubripes	3	100.00	125.00	110.00	–	44.71	50.00	115.15	29.41	58.82	94.12

キバハリアリの比率変数データ (つづき)

ID-Species Name	SpG No.	CI= HW/HL	SI= SL/HW	MI= ML/HL	LI=HFL /PrW	PpI= PpW/GW	PtW /PtL	PpW /PpL	PtW /GW	PrW /GW	HW /GW
02rubripes	3	101.35	132.00	108.11	–	45.71	40.82	94.12	28.57	65.71	107.14
03rubripes	3	101.45	135.71	113.04	–	44.93	43.48	100.00	28.99	65.22	101.45
04rubripes	3	98.59	134.29	108.45	–	44.78	42.55	100.00	29.85	64.18	104.48
01rufinodis	3	95.71	126.87	102.86	–	44.12	44.44	115.38	29.41	64.71	98.53
02rufinodis	3	87.23	126.83	87.23	296.30	–	48.28	105.26	–	–	–
03rufinodis	3	91.67	134.09	89.58	273.33	40.00	–	–	26.00	60.00	88.00
04rufinodis	3	100.00	120.00	102.50	–	46.43	50.98	121.88	30.95	61.90	95.24
01suttoni	3	97.56	125.00	103.66	–	43.53	45.61	105.71	30.59	60.00	94.12
02suttoni	3	100.00	–	107.95	–	48.39	55.00	107.14	35.48	60.22	94.62
03suttoni	3	94.67	132.39	109.33	–	–	45.10	–	32.86	64.29	101.43
04suttoni	3	–	–	–	–	43.75	–	–	–	–	–
01tarsata	3	101.28	101.27	107.69	–	54.22	65.22	112.50	36.14	66.27	95.18
02tarsata	3	–	–	–	–	56.25	66.67	118.42	37.50	63.75	–
03tarsata	3	101.37	104.05	104.11	–	54.55	–	–	35.06	68.83	96.10
04tarsata	3	101.56	112.31	100.00	–	53.73	62.50	120.00	37.31	68.66	97.01
05tarsata	3	96.23	111.76	98.11	241.18	55.77	59.38	116.00	36.54	65.38	98.08
06tarsata	3	108.43	–	–	–	55.29	64.00	127.03	37.65	–	105.88
01vindex	3	98.68	130.67	113.16	–	48.75	48.15	125.81	32.50	61.25	93.75
02vindex	3	98.72	128.57	110.26	–	48.65	45.28	116.13	32.43	66.22	104.05
01basirufa	3	96.25	128.57	103.75	–	50.67	50.94	118.75	36.00	60.00	102.67
01simillima	3	102.47	107.23	112.35	–	54.79	54.35	117.65	34.25	68.49	113.70
02simillima	3	103.66	102.35	109.76	–	51.28	58.70	125.00	34.62	65.38	108.97
03simillima	3	98.04	110.00	96.08	256.25	47.06	59.26	109.09	31.37	62.75	98.04
01analis	3	101.61	115.87	112.90	–	47.69	57.89	114.81	33.85	64.62	96.92
02analis	3	96.92	119.05	96.92	–	48.48	52.63	128.00	30.30	63.64	95.45
03analis	3	101.67	121.31	113.33	–	47.54	55.56	111.54	32.79	67.21	100.00
04analis	3	103.08	111.94	115.38	–	47.69	52.38	103.33	33.85	66.15	103.08
05analis	3	98.44	122.22	110.94	–	47.69	56.41	114.81	33.85	63.08	96.92
06analis	3	98.48	120.00	110.61	–	48.44	58.97	103.33	35.94	65.62	101.56
01gilberti	4	108.00	75.93	100.00	158.97	76.00	83.33	152.00	50.00	78.00	108.00
02gilberti	4	106.12	76.92	100.00	144.74	72.92	85.19	140.00	47.92	79.17	108.33
01luteiforceps	4	108.89	77.55	111.11	157.14	75.56	80.00	154.55	44.44	77.78	108.89
02luteiforceps	4	108.70	76.00	113.04	150.00	75.56	81.48	154.55	48.89	80.00	111.11
01potteri	4	106.25	78.43	97.92	153.85	72.34	88.00	154.55	46.81	82.98	108.51
02potteri	4	106.25	76.47	95.83	155.26	76.60	85.19	156.52	48.94	80.85	108.51
01pilventris	4	109.26	83.05	101.85	160.98	76.36	83.33	140.00	45.45	74.55	107.27
02pilventris	4	110.34	78.12	101.72	152.08	77.05	82.35	138.24	45.90	78.69	104.92
03pilventris	4	107.69	78.57	102.56	150.00	70.00	90.48	133.33	47.50	75.00	105.00
01fulviculis	4	100.00	83.33	104.17	152.78	82.93	78.57	121.43	53.66	87.80	117.07
02fulviculis	4	104.17	80.00	104.17	154.05	77.78	80.00	125.00	53.33	82.22	111.11
01fulvipes	4	116.33	75.44	108.16	147.62	75.00	89.29	130.00	48.08	80.77	109.62
02fulvipes	4	116.00	75.86	112.00	–	73.58	92.59	130.00	47.17	–	109.43
01mandibularis	4	112.50	80.95	110.71	162.22	74.14	82.35	130.30	48.28	77.59	108.62
02mandibularis	4	114.00	84.21	118.00	152.38	76.00	79.31	126.67	46.00	84.00	114.00
03mandibularis	4	111.11	80.00	120.00	157.14	71.11	72.00	128.00	40.00	77.78	111.11
04mandibularis	4	112.96	81.97	116.67	156.82	75.47	80.65	133.33	47.17	83.02	115.09
01flammicollis	5	93.88	119.57	108.16	213.33	51.16	45.45	110.00	34.88	69.77	106.98
02flammicollis	5	98.21	118.18	105.36	265.71	54.00	51.35	108.00	38.00	70.00	110.00
01nigrocincta	5	93.75	120.00	106.25	250.00	58.54	50.00	109.09	34.15	78.05	109.76
02nigrocincta	5	98.18	111.11	103.64	248.65	62.00	59.38	119.23	38.00	74.00	108.00
03nigrocincta	5	93.48	118.60	102.17	285.71	60.53	55.56	115.00	39.47	73.68	113.16
01petiolata	5	97.78	–	108.89	–	57.50	57.69	115.00	37.50	75.00	110.00
02petiolata	5	93.33	119.05	108.89	–	55.00	51.85	115.79	35.00	75.00	105.00
03petiolata	5	100.00	111.11	108.89	240.00	52.50	–	–	–	75.00	112.50
04petiolata	5	95.35	114.63	109.30	241.38	54.05	56.52	117.65	35.14	78.38	110.81
01fucosa	6	93.33	80.95	91.11	200.00	62.50	72.00	138.89	45.00	67.50	105.00
02fucosa	6	100.00	80.00	95.56	186.67	60.00	73.08	117.39	42.22	66.67	100.00
01picta	6	93.62	84.09	91.49	196.67	65.12	81.82	140.00	41.86	69.77	102.33
02picta	6	95.56	83.72	88.89	190.00	64.44	80.00	131.82	44.44	66.67	95.56
03picta	6	95.74	80.00	89.36	156.25	68.18	–	–	45.45	72.73	102.27

キバハリアリの比率変数データ (つづき)

ID-Species Name	SpG No.	CI= HW/HL	SI= SL/HW	MI= ML/HL	LI=HFL /PrW	PpI= PpW/GW	PtW /PtL	PpW /PpL	PtW /GW	PrW /GW	HW /GW
04picta	6	95.00	86.84	95.00	204.00	65.79	69.57	131.58	42.11	65.79	100.00
01apicalis	7	100.00	102.27	97.73	193.75	56.25	65.22	158.82	31.25	66.67	91.67
01celaena	7	107.69	83.33	102.56	182.76	68.42	84.21	144.44	42.11	76.32	110.53
01chasei	7	117.65	83.33	101.96	180.00	67.16	90.00	160.71	40.30	67.16	89.55
02chasei	7	110.87	94.12	108.70	194.74	65.31	80.00	160.00	40.82	77.55	104.08
01ludlowi	7	110.34	78.12	94.83	166.67	71.21	103.57	156.67	43.94	72.73	96.97
02ludlowi	7	106.52	91.84	102.17	189.19	68.00	95.65	147.83	44.00	74.00	98.00
03ludlowi	7	111.76	87.72	101.96	187.50	68.97	89.29	153.85	43.10	68.97	98.28
01chrysogaster	7	104.76	72.73	92.86	145.45	69.77	90.00	150.00	41.86	76.74	102.33
01cydista	7	104.88	83.72	95.12	–	69.77	86.36	166.67	44.19	72.09	100.00
01dispar	7	106.52	87.76	104.35	172.73	64.44	82.61	152.63	42.22	73.33	108.89
02dispar	7	102.08	89.80	89.58	181.25	66.67	83.33	150.00	44.44	71.11	108.89
03dispar	7	100.00	97.67	102.33	–	63.41	85.71	144.44	43.90	70.73	104.88
01elegans	7	100.00	88.24	107.84	175.68	70.00	92.00	145.83	46.00	74.00	102.00
02elegans	7	102.33	90.91	116.28	196.67	69.05	95.24	145.00	47.62	71.43	104.76
03elegans	7	97.67	95.24	111.63	186.67	70.00	95.00	140.00	47.50	75.00	105.00
01harderi	7	104.55	82.61	100.00	173.53	70.21	83.33	150.00	42.55	72.34	97.87
02harderi	7	108.00	77.78	94.00	167.57	72.55	92.31	160.87	47.06	72.55	105.88
03harderi	7	104.44	80.85	93.33	–	68.09	82.61	145.45	40.43	70.21	100.00
01michaelseni	7	106.12	76.92	100.00	160.53	72.22	80.00	156.00	37.04	70.37	96.30
02michaelseni	7	108.89	77.55	104.44	148.65	72.34	76.00	130.77	40.43	78.72	104.26
03michaelseni	7	111.11	–	–	–	74.47	73.08	145.83	40.43	74.47	106.38
04michaelseni	7	110.00	78.18	100.00	153.66	72.73	81.48	142.86	40.00	74.55	100.00
01occidentalis	7	106.67	91.67	100.00	184.38	66.67	80.00	142.86	44.44	71.11	106.67
02occidentalis	7	108.89	87.76	97.78	160.00	66.67	76.92	130.43	44.44	77.78	108.89
03occidentalis	7	102.22	86.96	95.56	172.73	63.83	–	–	–	70.21	97.87
04occidentalis	7	–	95.74	–	–	64.58	–	–	–	–	97.92
05occidentalis	7	102.17	–	–	–	65.22	–	–	41.30	–	102.17
01opaca	7	104.76	90.91	100.00	170.97	66.67	85.71	140.00	42.86	73.81	104.76
01rugosa	7	110.00	81.82	98.00	162.50	76.36	92.59	150.00	45.45	72.73	100.00
02rugosa	7	104.17	80.00	95.83	157.89	76.00	88.00	152.00	44.00	76.00	100.00
03rugosa	7	107.50	81.40	100.00	147.06	72.09	81.82	147.62	41.86	79.07	100.00
01varians	7	105.00	90.48	115.00	190.00	64.10	73.91	125.00	43.59	76.92	107.69
02varians	7	102.13	85.42	104.26	182.35	68.18	80.77	120.00	47.73	77.27	109.09
03varians	7	100.00	88.37	104.65	190.00	67.50	75.00	117.39	45.00	75.00	107.50
04varians	7	102.17	85.11	102.17	200.00	62.79	72.00	117.39	41.86	72.09	109.30
05varians	7	100.00	86.67	106.67	187.50	68.18	76.00	125.00	43.18	72.73	102.27
06varians	7	102.17	85.11	104.35	184.85	66.67	70.37	120.00	42.22	73.33	104.44
01pilosula	7	–	–	–	171.43	65.22	80.00	142.86	43.48	76.09	–
02pilosula	7	100.00	102.70	108.11	185.19	64.86	75.00	133.33	40.54	72.97	100.00
03pilosula	7	109.09	87.50	102.27	162.86	61.22	80.00	125.00	40.82	71.43	97.96
04pilosula	7	101.82	85.71	96.36	172.50	63.64	85.19	129.63	41.82	72.73	101.82
05pilosula	7	105.45	84.48	100.00	166.67	64.29	76.67	138.46	41.07	75.00	103.57
01clarki	8	102.33	90.91	113.95	170.00	72.50	95.24	145.00	50.00	75.00	110.00
02clarki	8	104.26	87.76	110.64	161.76	69.39	92.00	147.83	46.94	69.39	100.00
01dixoni	8	105.00	73.81	105.00	–	70.73	95.00	161.11	46.34	73.17	102.44
02dixoni	8	112.20	73.91	102.44	130.30	68.18	95.65	142.86	50.00	75.00	104.55
03dixoni	8	105.26	75.00	105.26	142.86	70.27	94.44	162.50	45.95	75.68	108.11
01swalei	8	108.57	78.95	114.29	144.00	65.71	83.33	127.78	42.86	71.43	108.57
02swalei	8	109.09	72.92	102.27	142.42	68.89	91.67	134.78	48.89	73.33	106.67
03swalei	8	105.26	80.00	105.26	137.93	71.79	95.00	127.27	48.72	74.36	102.56
04swalei	8	107.14	77.78	107.14	–	72.09	90.91	155.00	46.51	74.42	104.65
05swalei	8	108.82	75.68	108.82	138.46	70.59	88.24	150.00	44.12	76.47	108.82
06swalei	8	104.08	78.43	100.00	145.00	73.08	96.00	146.15	46.15	76.92	98.08
01tepperi	8	105.00	85.71	115.00	162.07	67.57	71.43	125.00	40.54	78.38	113.51
02tepperi	8	–	–	–	–	68.09	88.00	128.00	46.81	74.47	–
03tepperi	8	102.04	92.00	102.04	171.43	72.92	88.00	145.83	45.83	72.92	104.17
04tepperi	8	106.00	84.91	104.00	166.67	70.00	92.00	140.00	46.00	72.00	106.00
01testaceipes	8	107.50	74.42	107.50	140.00	72.50	90.48	145.00	47.50	75.00	107.50
02testaceipes	8	110.00	77.27	107.50	146.67	71.43	90.48	150.00	45.24	71.43	104.76

キバハリアリの比率変数データ (つづき)

ID-Species Name	SpG No.	CI= HW/HL	SI= SL/HW	MI= ML/HL	LI=HFL /PrW	PpI= PpW/GW	PtW /PtL	PpW /PpL	PtW /GW	PrW /GW	HW /GW
03testaceipes	8	108.57	78.95	114.29	144.44	69.44	83.33	138.89	41.67	75.00	105.56
01dichospila	9	103.03	67.65	78.79	156.00	70.59	83.33	160.00	44.12	73.53	100.00
02dichospila	9	100.00	80.00	93.33	195.00	66.67	75.00	120.00	44.44	74.07	111.11
03dichospila	9	96.30	80.77	92.59	188.89	66.67	71.43	123.08	41.67	75.00	108.33
01exigua	9	95.83	82.61	95.83	193.33	72.73	71.43	160.00	45.45	68.18	104.55
01rubicunda	9	108.70	76.00	104.35	193.33	68.00	71.43	170.00	40.00	60.00	100.00
01infima	9	93.75	83.33	84.38	215.00	73.33	68.75	137.50	36.67	66.67	100.00
02infima	9	90.91	83.33	84.85	215.00	75.86	64.71	146.67	37.93	68.97	103.45
03infima	9	96.97	78.12	90.91	–	67.65	86.67	153.33	38.24	67.65	94.12
04infima	9	97.14	76.47	88.57	–	73.68	85.00	140.00	44.74	63.16	89.47
05infima	9	80.00	95.00	84.00	–	60.00	60.00	115.38	36.00	68.00	80.00
06infima	9	100.00	76.00	96.00	–	68.00	–	–	–	–	100.00

表目次

図目次

謝辞

本論文の第 II 部は, 日本大学大学院理工学研究科に在学中（博士前期課程数学専攻, 指導教官, 戸川隼人教授）に文部省統計数理研究所の受託学生となり, 修士論文として仕上げた内容に基づいています. そのとき現在の主任指導教官である大隅 昇教授に指導していただきました. 大隅, 戸川両教授との出会いが現在の研究の発端となっています.

第 III, IV 部は, 総合研究大学院大学および文部省統計数理研究所の大隅 昇教授, 小西貞則教授（現在九州大学教授, 前統計数理研究所助教授）, 馬場康維助教授の懇切丁寧なご指導をいただきました. 大隅教授は主任指導教官であると同時にクラスター化法の基礎について教えていただき, 研究が順調に進められるよう積極的に支援していただきました. また, 本論文の全文について大変丁寧に読んで下さり, さらに推敲していただきました. 指導教官であった小西教授は多変量解析や情報量規準の理論的な部分や, 研究上の方向性を示していただき, 懇切丁寧なご指導をしていただきました. 指導教官の馬場助教授はデータ解析における細かいことや, 研究が円滑に進むよう様々なご助言をいただきました. 3 名の指導教官のご支援がなければ論文が完成することはなかったと思います. 心より感謝します.

さらに, 審査員である総合研究大学院大学および文部省統計数理研究所の駒沢 勉教授 (主査), 村上征勝教授, 大隅 昇教授, 馬場康維助教授, 九州大学の小西貞則教授, 大学入試センターの柳井晴夫教授にはご多忙ながら貴重な時間をさいて本論文の審査を引き受けて下さり, 多くの貴重なコメントをいただきました. とくに柳井教授からは第 II, IV 部について大変多くの適切なご助言をいただきました. 厚くお礼を申し上げます.

統計数理研究所の伊庭幸人助手, 金藤浩司助手, 総合研究大学院大学の学生の村田磨里子さん, 北門利英さん, 土屋高宏さんの方々とは貴重な議論していただきました. また, 総合研究大学院大学に在学中には統計数理研究所の教官の皆様, 技術課, 管理部の皆様, 総合研究大

学院大学数物科学研究科統計科学専攻の学生の皆様には大変お世話になりました．この場を借りて厚くお礼申し上げます．

最後に，ここまで見守ってくれた両親（中村 弘・嫩）に心から感謝します．

参考文献

参考文献

[1] Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models, *Journal of the Royal Statististical Society* Series B, **47**, 67-75.

[2] Aitkin, M., and Tunnicliffe Willison, G. (1980). Mixture models, outliers, and the EM algorithm, *Technometrics*, **22**, 325-331.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Proc. of 2nd International Symposium on Information Theory* (eds. B. Petrov and F. Csaki), 267-281, Akademiai Kiado, Budapest.

[4] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis, 2nd Edition*, Wiley, New York.

[5] Andrews, D. F. and Herzberg, A. M. (1985). *Data, a collection of problems from many fields for the student and research worker*, Springer-Verlag, New York.

[6] Banfield, D. B. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering, *Biometrics*, **49**, 803-821.

[7] Basford, K. E. and McLachlan, G. J. (1985). Estimation of allocation rates in a cluster analysis context, *Journal of the American Statistical Association*, **80**, 286-293.

[8] Batagelj, V. (1981). Note on ultrametric hierarchical clustering algorithms *Psychometrika*, **46**, 351-352.

[9] Bayne C. K., Beauchamp, J. J., Begovich, C. L. and Kane, V. E. (1980). Monte Carlo comparison of selected clustering procedures, *Pattern Recognition*, **12**, 51-62.

[10] Bock, H. H. (1974). Automatische Klassifikation, Vandenhoeck & Ruprecht.

[11] Bozdgan, H. (1992). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher matrix, In *Information and classification*, Proceedings of the 16th Annual Conference of the "Gesellschaft Klassifikation e.V." (eds. O. Opitz, B. Lausen and R. Klar), Springer-Verlag, pp40-54.

[12] Bozdgan, H. and Sclove, S. L. (1984). Multi-sample cluster analysis using Akaike's information criterion. *Annals of Institute of the Statistical Mathematics*, **36**, 163-180.

[13] Celeux, G. (1986). Validity test in cluster analysis using a probabilistic teacher algorithm, *Compstat*, **86**, 163-168, Springer Verlag.

[14] Celeux, G. (1987). Thèse de Doctorat d'État es Sciences Mathématiques, Université de Paris IX Dauphine. (この論文は未入手, Soromenho(1994) から引用)

[15] Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probablistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quarterly*, **2**, 73-82.

[16] Charlier, C. V. L. and Wicksell, S. D. (1924). On the dissection of frequency functions, *Arkiv för Mathematik, Astronomi och Fysik*, **BD. 18**, No 6.

[17] Cohen, A. C. (1967). Estimation in mixtures of two normal distributions, *Technometrics*, **9**, 15-28.

[18] Cormack, R. M. (1971). An review of classification (with Siscussion), *Journal of the Royal Statististical Society* Series A, **134**, 321-367.

[19] Day, N. E. (1969). Estimating the components of a mixture of normal distributions, *Biometrika*, **56**, 463-474.

[20] Dempster, A. P., Laird, N, M. and Rubin, D. B. (1977) . Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* Series B, **39**, 1-38.

[21] Dempster, A. P., Laird, N, M. and Rubin, D. B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed, *Multivariate Analysis V*, (ed. P. R. Krishnaiah), 35-57, North-Holland.

[22] Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society*, Series B, **56**, 363-375.

[23] DuBien, J. L. and Warde, W. D. (1979). A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms, *The Canadian Journal of Statistics*, **7**, 29-38.

[24] Duda, R. O. and Hart, P. E. (1973). *Pattern Classificaation and Scene Analysis*, New York: Wiley.

[25] Duran, B. S. and Odell, P. L. (1974). *Cluster Analysis: A survey*, Berlin: Springer-Verlag.

[26] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Statist.* **7**, 1-26.

[27] Everitt, B. S. (1980). *Cluster Analysis*, Seond edition, Wiley-Halsted, London.

[28] Everitt, B. S. (1981). A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions, *Multivar. Behav. Res.* **16**, 171-180.

[29] Everitt, B, S (1993). *Cluster Analysis*, Third edition, Wiley-Halsted, London.

[30] Everitt, B, S and Hand, D. J. (1981). *Finite Mixture Distributions*, Chapman and Hall, New York.

[31] Foley, J. D. and Van Dam, A. (1982). *Fundamentals of Interactive Computer Graphics*, Addison Wesley, Massachusetts.

[32] Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition*, Academic Press, New York.

[33] Furman, W. D. and Lindsay, B. G. (1994a). Testing for the number of components in a mixture of normal distributions using moment estimators, *Computational Statistics & Data Analysis*, **17**, 473-492.

[34] Furman, W. D. and Lindsay, B. G. (1994b). Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods, *Computational Statistics & Data Analysis*, **17**, 493-507.

[35] Fryer, J. G. and Robertson, C. A. (1972). A comparison of some methods for estimating mixed normal distributions, *Biometrika*, **59**, 639-648.

[36] 伏見 正則 (1989). 乱数, 東京大学出版, 東京.

[37] Ganesalingam, S. (1989). Classification and mixture approaches to clustering via maximum likelihood, *Applied Statistics* **38**, 455-466.

[38] Ganesalingam, S. and McLachlan, G. J. (1980). A Comparison of the mixture and classification approaches to cluster analysis, *Communications in Statistics —Theory and Methods* **A9**, 923-933.

[39] Gordon, A. D. (1981). *Classification: Methods for the Exploratory Analysis of Multivariate data*, Chapman and Hall, New York.

[40] Gordon, A. D. (1987). A review of hierarchical classification, *Journal of the Royal Statistical Society* Series A, **150**, 119-137.

[41] Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions, *Technometrics*, **8**, 431-444.

[42] Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family, *Journal of the American Statistical Association*, **64**, 1459-1471.

[43] Hartigan, J. A. (1975). *Clustering Algorithms*, Wiley, New York.

[44] Huber, P. J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics*, **35**, 73-101.

[45] Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy*, Wiley, London.

[46] Jolliffe, I. T. (1972). Discarding variables in aprincipal component analysis, I: artificial data, *Applied Statistics*, **21**, 160-173.

[47] Jolliffe, I. T. (1973). Discarding variables in aprincipal component analysis, II: real data, *Applied Statistics*, **22**, 21-31.

[48] 石黒 真木夫, 坂元 慶行, 北川 源四郎 (1992). ベイズモデルと EIC, 第 60 回日本統計学会講演予稿集, 264-266.

[49] 狩野 裕 (1992) 楕円分布と統計的推測, 科研費シンポジウム予稿.

[50] Kano, Y., Berkane, M. and Bentler, P. M. (1990). Covariance structure analysis with heterogeneous kurtosis parameters, *Biometrika*, **77**, 575-585.

[51] Kano, Y., Berkane, M. and Bentler, P. M. (1993). Statistical inference based on pseudo-maximum likelihood estimators in elliptical populations, *Journal of the American Statistical Association*, **88**, 135-143.

[52] 北川 源四郎 (1993). FORTRAN 77 時系列解析プログラミング, 岩波書店, 東京.

[53] 北川 源四郎, 石黒 真木夫, 坂元 慶行 (1992). EIC によるモデルの予測評価, 第 60 回日本統計学会講演予稿集, 258-260.

[54] 北川 源四郎, 石黒 真木夫, 坂元 慶行 (1993). 情報量規準 AIC と EIC, 電子情報通信学会, **92**, 503, 49-62.

[55] 小西 貞則 (1992). ブートストラップ法と予測誤差推定, 東京大学統計学輪講資料.

[56] 小西 貞則 (1993a). 予測誤差推定とブートストラップ法, 第 61 回日本統計学会講演予稿集, 308-309.

[57] 小西 貞則 (1993b). 予測誤差推定とブートストラップ法, Research Memorandum No.482, Technical Report of *The Institute of Statistical Mathematics.*

[58] 久保川 達也, 江口 真透, 竹村 彰通, 小西 貞則 (1993). 統計的推測理論の現状, 日本統計学会誌, **22**, (増刊号), 257-312.

[59] Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies, 1.Hierarchical systems, *The Computer Journal*, *9*, 373-380.

[60] Lance, G. N. and Williams, W. T. (1977). Hierarchical classificatory methods, *"Statistical Methods for Digital Computers"*, Volume 3 of *"Mathmatical Methods for Digital Computers,"* Enslein, K., Ralston, A. and Wilf, H. S. (edi.), Wiley, pp.267-268.

[61] Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using the t distributions, *J. Amer. Statist. Assoc.*, **84**, 881-896.

[62] Little, R. J. A. (1987). Robust estimation of the mean and covariance matrix from data with missing values, *Applied Statistics*, **37**, 23-38.

[63] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, In J. neyman (Ed.), *Proc. 5th Berkeley Symp.* **Vol. 1** (pp. 281-297), Berkeley.

[64] Mangin, B., Goffinet, B. and Elsen, J. M. (1993). Testing in normal mixture models with some information on the parameters, *Biometorical Journal*, **35**, 771-783.

[65] Mardia, K. V. Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, London.

[66] McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, *Appl. Statist.* **36**, 318-324.

[67] McLachlan, G. J. (1988). On the choice of starting values for the EM algorithm in fitting mixture models. *Statistician*, **37**, 417-425.

[68] McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley, New York.

[69] McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

[70] Mezzich, J. E. and Solomon, H. (1980). *Taxonomy and Behavioral Science — Comparative Performance of Grouping Methods*, Academic Press, New York.

[71] Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms, *Psychometrika*, **44**, 343-346.

[72] 三中 信宏 (1995). 系統分類学における定量的研究の動向：形態情報と分子情報の利用, 統計数理研究所 共同研究リポート 67, (6- 共研 A-57).

[73] 宮川 雅巳 (1987). EM アルゴリズムとその周辺（コメント付き）, 応用統計学, **16**, 1-21.

[74] 中村 永友 (1994). 混合分布の推定手続きとその応用, 計算機統計学会誌, (投稿中).

[75] Nakamura, N. and Ohsumi, N. (1994). Note on Lance and Willimas' flexible method and its extensions, *Psychometrika*, (投稿中).

[76] 中村 永友, 小西 貞則 (1994a). 多変量混合分布モデルのコンポーネント数の推定, Research Memorandum No. 507, Technical Report of *The Institute of Statistical Mathematics.*

[77] 中村 永友, 小西 貞則 (1994b). 多変量混合分布モデルのコンポーネント数の推定, 第62回日本統計学会講演予稿集, 184-185.

[78] 中村 永友, 小西 貞則, 大隅 昇 (1993). 混合分布モデルを用いた画像分類と色彩変換－LANDSAT 画像の解析－, 統計数理, **41**, 149-167.

[79] Nakamura, N. and Ohsumi, N. (1990). Space-distorting properties in agglomerative hierarchical clustering algorithms, Research Memorandum No.387, Technical Report of *The Institute of Statistical Mathematics.*

[80] 緒方 一夫 (1995). アリ類における分類学の現状：キバハリアリと亜科の分類を例に, 統計数理研究所 共同研究リポート 67, (6- 共研 A-57).

[81] Ogata, K. (1991). Ants of the genus *Myrmecia* (Fabricius): a review of the species groups and their phylogenetic relationships (Hymenoptera: Formicidae: Myrmecinae), *Systematic Entomology*, **16**, 353-381.

[82] Ogata, K. and Taylor, W. (1991). Ants of the genus *Myrmecia* Fabricius: a preliminary review and key to the named species (Hymenoptera: Formicidae: Myrmecinae), *Journal of Natural History*, **25**, 1623-1673.

[83] 大隅 昇 (1989). 統計的データ解析とソフトウェア, 放送大学教材 56495-1-8911.

[84] Ohsumi, N. and Nakamura, N. (1989). Space-distorting properties in agglomerative hierarchical clustering algorithms, DATA ANALYSIS 1989 LEARNING SYMBOLIC AND NUMERIC KNOWLEDGE, 47th ISI Satellite meeting session.

[85] Ohsumi, N and Nakamura, N. (1994). Comparison of hierarchical clustering algorithms based on space-distorting properties, In *Clustering and Classification*, World

Scientific (to be published).

[86] Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications, *Proc. 6th Berkeley Symp. on Math. Statist.* 1, (pp. 697-715), Berkeley.

[87] Pearson, K. (1894). Contributions to the mathematical theory of evolution *Philosophical Transactions of the Royal Society of London*, Series A **185**, 71-110.

[88] Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification, *Journal of the Royal Statistical Society* Series B, **10**, 15-203.

[89] Reaven, G. M., and Miller, R. G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis, *Diabetologia* **16**, 17-24.

[90] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, **26**, 195-239.

[91] Rissanen, J. (1978). Modeling by shortest data description, *Automatica*, **14**, 465-471.

[92] Rissanen, J. (1984). Universal coding, information, prediction and estimation, *IEEE Transactions on Information Theory*, **IT–30**, 629-636.

[93] 坂元 慶行, 石黒 真木夫, 北川 源四郎 (1992). ABIC 最小化法と EIC, 第 60 回日本統計学会講演予稿集, 261-263.

[94] Sclove, S. L. (1983). Application of the conditional population-mixture model to image segmentation, *IEEE Transactions on Pattern Analysis and Machine intelligence*, **PAMI-5**, 428-433.

[95] Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria, *Biometrics*, **27**, 387-397.

[96] Seber G. A. F. (1984). *Multivariate Observations.* Wiley, New York.

[97] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London, Chapman and Hall.

[98] Sneath, P. H. A and Sokal, R. R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* W.H. Freeman, San Francisco.

[99] Sokal, R. R. and Sneath, P. H. A (1963). *Principles of Numerical Taxonomy.* W.H. Freeman, San Francisco.

[100] Soromenho, G. (1994). Comparing approaches for testing the number of components in a finite mixture model, *Computational Statistics*, **9**, 65-78.

[101] Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures, *Biometrics*, **37**, 35-43.

[102] 高木 幹雄, 下田 陽久監修 (1991). 画像解析ハンドブック, 東京大学出版, 東京.

[103] 竹内 啓, 竹村彰通 (1986). 主成分分析法における変数選択の考え方, 第 14 回日本行動計量学会.

[104] Tan, W. Y. and Chang, W. C. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of mixture of two normal distributions, *Journal of American Statistical Association*, **67**, 702-708.

[105] Taylor, J. M. G. (1992). Properties of modelling the error distribution with an extra shape parameter, *Computational Statistics & Data Analysis*, **13**, 33-46.

[106] Thode, Jr. H. C., Finch, S. J. and Mendell, N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals, *Biometrics*, **44**, 1195-1201.

[107] Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

[108] Tukey J. W. (1960). A survey of sampling from contaminated distributions, In *Contributions to Probability and Statistics*, I. Olkin, *et al.* (Eds.). Stanford University Press, Stanford, pp. 448-485.

[109] Rubin, D. B. (1983). Iteratively reweighted least squares, In S. Kotz and N.L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, **4**, John Wiley, New York, pp. 272-275.

[110] Van Emden, M. H. (1971). *Analysis of Complexity*, Mathematical Center Tracts, **35**, Amsterdam.

[111] 宇宙開発事業団地球観測センター編集 (1990). 地球観測データ利用ハンドブック — ランドサット編・改定版—, (財) リモート・センシング技術センター, 東京.

[112] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of American Statistical Association*, **58**, 236-244.

[113] Windham, M. P. and Cutler, A. (1992). Information ratios for validating mixture analysis, *Journal of American Statistical Association*, **87**, 1188-1192.

[114] Wishart, D. (1969a). Mode analysis: a generalization of nearest neighbour which reduces chaining effects (with Discussion), In *Numerical Taxonomy*, A.J. Cole (ed.), Academic Press, London, pp. 282-311.

[115] Wishart, D. (1969b). An algoritym for hierarchical classification, *Biometrics*, **25**, 165-170.

[116] Wolfe, J. H. (1967). **NORMIX**; computational methods for estimating the parameters of multivariate normal mixtures of distributions. *Research Memorandum*, **SRM 68-2**, US Naval Personnel Research Activity, San Diego.

[117] Wolfe, J. H. (1969). Pattern clustering by multivariate mixture analysis. *Research Memorandum*, **SRM 69-17**, US Naval Personnel Research Activity, San Diego.

[118] Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis, *Multivariate Behavioral Research*, **5**, 329-350.

[119] Wolfe, J. H. (1971). A Monte Carlo study of sampling distribution on the likelihood ratio test for mixtures of multinormal distributions, *Technical Bulletin STB 72-2*, San Diego: U.S. Naval Personnel and Training Research Laboratory. (これは McLachlan and Basford (1988) より引用)

[120] Wong, M. A. (1985). A bootstrap testing procedure for investigating the number of subpopulations, *Journal of the Statistical Computation and Simulation*, **22**, 99-112.

[121] Wong, M. A. and Lane, T. (1983). A k-th nearest neighbour clustering procedure, *Journal of the Royal Statistical Society* Series B, **45**, 362-368.

[122] 矢島 敬二, 他 (1971). クラスター・アナリシス (1)-(4), オペレーションズ・リサーチ, **16**, 7 ～ 10.

[123] Yanai, H. (1980). A proposition of generalized method for forward selection of variables, *Behaviormetrika*, **7**, 95-107.