有限混合分布モデルの学習に関する研究

赤穂 昭太郎

この論文は 東京大学工学系研究科計数工学専攻 (2001 年 3 月 15 日授与) の博士論文を WWW で公開するために font 情報等を書き換え，誤植も適宜修正したものです.

修正履歴: 2002 年 1 月 7 日 公開

2003 年 7 月 22 日 誤植修正

目 次

第1章　序論

1.1　研究の背景と位置づけ

計算機や通信網の普及によって社会の情報化が急速に進みつつある．それにともなって世の中には膨大な情報が氾濫し，ともすれば人間を情報の洪水に巻き込んでしまいがちである．それを避けるためには，ユーザにとって必要な情報やデータの背後にある構造を適切に抽出する必要がある．そのための知的な情報処理手法が，データからの学習の問題として，人工知能，パターン認識，統計学などを中心とした学際領域で盛んに研究されるようになり，さらにはデータマイニングと呼ばれる一領域を成しつつある．

従来，人工知能とパターン認識は比較的最近までそれぞれが独立した研究分野として発展してきた．論理学に基礎をおく人工知能はどちらかというと記号化された理想的な環境を扱い，深くて演繹的な推論を研究してきた．一方，パターン認識をはじめとする確率・統計的研究は，より生のデータに近い所で，特徴抽出や判別，回帰など，比較的浅くて帰納的な推論についての研究を行ってきた．しかしながら，本当に現実世界の情報に対して人間が行っているような柔軟な情報処理を行うためには，それぞれの枠組みだけでは不十分であり，両者を統合した枠組みで研究を行っていく必要があるという認識が高まってきた [63, 72]．特に，対話システムや自律ロボットの研究では，外界から入って来る生データと内部的な記号的処理とを両方統合して扱うことが必要であり，情報統合の問題として研究が盛んに行われている [77, 64].

実世界は非常に複雑であり，それに対処するため情報処理システムには大量で多様な情報が与えられる．しかしながら，いくら大量の情報を与えられてもそれは複雑な世界の一部分を切り取った不完全な部分情報に過ぎない．また，情報自体があいまいで，誤りを含んでいたり不確実であることも多い．そのような状況下で

は，因果関係の記述や推論が計算論的爆発を引き起こすという，いわゆるフレーム問題と呼ばれる問題が起きる．この問題に対し，統計的なアプローチは二つの対処法を提供する．一つは不要な部分を確率的な変動として近似してしまうことによって情報の複雑度を減らし，それによって計算量を低減するというものであり，二つ目は情報のあいまい性を確率的なモデルとして表現することによって，単純な論理推論よりも柔軟な推論を行うというものである．

では，これらの複雑な世界のモデル化に適した統計モデルとはどのようなものであろうか． 一つの要件はモデルが任意に単純化したり複雑化したりできる柔軟性をもつことである．1980 年代にニューラルネットワークモデルが大きな脚光を浴びたのも，従来の線形の多変量解析モデルなどよりも複雑な表現が比較的単純なメカニズムで実現できたからだと考えられる． ニューラルネットワークの研究は，理論や応用の研究，あるいは統計学や統計物理学との交流などを通じてさまざまな派生モデルを生み出した．特に，複雑な情報処理のためのモデルとして注目されているのがグラフィカルモデルと呼ばれるモデルであり，ボルツマンマシンにはじまるマルコフランダム場モデルやベイジアンネットワークなどがこれにあたる．グラフィカルモデルはネットワークによる組合せ的な構造をもつことによって，記号的な情報とパターン的な情報を統合的に扱うことが可能になり，更に，因果関係や階層構造，情報の部分性といった処理対象の特性をうまく表現できる．またネットワークのノードやアークを増やしていくことにより，いくらでも複雑な構造を表現できる [54, 42, 56].

本研究では，グラフィカルモデルの中でも最も単純な形をした混合分布モデル (特に有限混合分布) を扱う．混合分布は統計学では古い歴史をもち，統計学におけるさまざまな知見の積み重ねがある [31, 86]．一方，単純な形をしているにも関わらず，モデルの非線形性などに関してボルツマンマシンやベイジアンネットワークの基本的な性質を受け継いでおり，混合分布の性質はより複雑なモデルの性質へと一般化できる可能性をもつと考えられる．そのため統計物理においても，混合分布を力学系としてとらえ，その緩和過程や相転移現象などの理論的な解析がなされてきた [75, 22, 23]．また，複雑な問題をモジュールの分担によって解決するという混合分布のもつ特質は，人工知能の研究においても重要である [79]． Jordan ら [43] がモジュール学習を混合分布の学習として定式化してから，混合分布はより広い研

究者の関心を集めるようになった．更に，クラスタリングを始めとして，混合分布を直接適用できるアプリケーションも数多く存在している [25].

さて，本研究では混合分布モデルの学習 (あるいは推定) の問題を扱う．人間は非常に柔軟な学習能力をもっており，複雑な対象をいとも簡単に学習することができる． 知的な情報処理を行う上で，学習は欠かせない要素である．学習に関しては，古くから心理学などで研究がなされてきたが，人工知能では，知能の計算機上での実現という観点から，特に計算論的な側面が重視され研究されてきた [77]．また，脳科学においても，脳の計算原理の解明を目指した計算論的神経科学の重要性が認知されつつある [46].

以下では，本研究で着目する汎化と学習アルゴリズムという二つの問題について述べる．計算論的に解釈すると，汎化はサンプル計算量に関する問題であり，学習アルゴリズムは，全体としてかかる実際の計算量を規定する問題とみなすことができる．

汎化とは，学習の結果与えられるサンプルデータの背景にある真の構造が抽出されることである． これは，学習における究極の目標であるが，実際には，有限個のサンプルからの学習によってその目標に対する近似値が得られるにすぎない．したがって，その近似値として得られたものが真の対象からどれだけ離れているかを知ることが学習において重要な問題となる．これは逆に，所望の近似精度を得るためにはどれだけのサンプルを収集しなければならないかというサンプル計算量の問題でもある．統計モデルの汎化能力の評価は，例題からの学習に起因する基本的な問題であり，統計学や情報理論などの枠組みからさまざまな研究がなされてきた [13, 87, 88, 73, 49, 15]．一般に，統計モデルの構造を複雑にすればするほど訓練データに対する当てはまり具合はよくなる． しかし逆に，過度に適合した統計モデルでは汎化能力は低くなってしまう．したがって，訓練サンプルに適合するだけではなく，できるだけモデルを単純化することによって，汎化能力を高く保つ必要がある．統計モデルの汎化能力は，正規分布に基礎をおく線形モデルについては古くからかなり詳しく調べられている． しかしながら，混合分布をはじめとするグラフィカルモデルは冗長性や特異性をもつので，従来考えられてきたのとは異なる振舞いを示すことがある． 近年，そのような冗長性のあるモデルに関する研究が盛んに行われるようになってきた [36, 29, 34, 90]．本研究では，あるクラスの正規

混合分布について汎化能力の性質を調べ，通常線形モデルで考えられていたのとは異なる振舞いがみられることを示す．

さて，汎化はサンプル計算量を規定するが，実際の学習は学習アルゴリズムによって行われるので，学習アルゴリズムを規定してはじめて全体でどれだけの計算量となるかが決まる．現実のアプリケーションでは学習の実時間性が要求されることも多い．計算機資源が豊富になってきたとはいえ，実世界の大量のデータを実時間で学習するためには，依然として計算量の少ないアルゴリズムを追求することに意味がある．更に，複雑な統計モデルを扱う際には，学習に反復アルゴリズムを用いる必要があるので，収束性なども重要な問題となる． グラフィカルモデルの学習アルゴリズムとして知られているのは，勾配法，Newton 法，EM (Expectation-Maximization) アルゴリズム，およびそれらの近似手法である． このうち，EM アルゴリズムは適用できる範囲に制限はあるが，安定した収束性と 1 回の繰り返しステップにかかる計算量の少なさにおいて，他の手法に優っている． EM アルゴリズムはもともと隠れマルコフモデル [24] などで個別に提案されていたものを Dempster らが一般的な形で定式化し [30]，混合分布の推定 [69] などに応用されてきた．また，理論的にもその幾何学的な構造が明らかになってきた [17, 80, 3]． すなわち，EM アルゴリズムは曲がった空間における推定問題をより平坦な空間での推定に帰着するための道具として解釈できる．例えば混合分布では，混合分布を構成する個々の要素分布の推定問題に帰着される．これは，全体としては困難な推定問題を，それぞれの構成要素の問題に帰着するという分割統治のアプローチに通じるものである．ただし，その要素分布が正規分布のような単純な分布では推定が簡単にできるが，そうでない場合はやはり難しい問題が残る．これを解決するためのアプローチとして，要素分布を再び混合分布でモデル化するという再帰的な手段を用いたのが Jordan らの階層的エキスパート混合モデルである [43]．しかしながら，このアプローチでは一般にパラメータ数が多くなるため汎化の点で不利であるという点と，既に要素分布として基本となる分布形が与えられている場合には用いることができない点で問題がある．そこで本研究では，既に要素分布の形が与えられていた場合に EM アルゴリズムを応用した学習アルゴリズムが適用できないかを考える．具体的には，任意に与えられた分布が位置・尺度パラメータをもつとき，正規混合分布の EM アルゴリズムを拡張した形の学習アルゴリズムが得られることを示す．

ところで，はじめに述べたように，実世界の情報は不確実性や不完全性をもっており，それが学習を困難にする一つの理由にもなっている．例えば，パターン認識の主な対象である画像や音声などのデータは，そこに記号的なラベルやタグが付与されている場合には比較的取扱いが易しくなるが，実際にはラベルづけされないデータが多く，(簡単に処理するために) 必要な情報が欠けているという意味で，これも一種の情報の不完全性とみることができる．そこで本研究では，そのような不完全性に対処するために，混合分布の適用を検討する． 記号的な情報が隠れているような新たな学習課題を設定し，混合分布によるモデル化と，EM アルゴリズムによる学習を行い，統計的手法の適用可能性について調べる．

以上で述べてきたように，本研究の目的は，混合分布モデルの学習における汎化能力の振舞いを数理的に明らかにするとともに，その学習アルゴリズムである EM アルゴリズムがどのような問題に適用可能なのかを探求することにある．

1.2 論文の構成

本論文は大まかに分けて，準備の部分である第 2, 3 章と，オリジナルの結果をまとめた第 4, 5, 6 章とからなる．

まず，第 2 章と第 3 章で本論文で扱う有限混合モデルとその学習アルゴリズムである EM アルゴリズムについて概説する． 従来なされてきた研究を概観するとともに，後の章で必要な基本事項をまとめる． 有限混合分布は単純な分布であるだけに，統計学などでは 19 世紀の終わりからさまざまな研究がなされてきた． 網羅的な内容については成書 [31, 86, 50, 51, 89] に譲り，本論文では学習，特に最尤推定に関わる事項を中心にまとめ，本論文で扱う問題点を明確化する．

第 4 章では，特殊な形の正規混合分布の汎化能力について述べる． 汎化能力を測るために用いられる尺度としてよく用いられるのは，訓練データに対する誤差がテストデータに対する誤差に対してもつバイアスである．AIC として知られている情報量規準もそのバイアスの期待値を評価したものである． 通常，バイアス値は統計モデルの可変なパラメータ数に比例し，モデルの複雑度が増すにしたがってバイアスも大きくなると考えられている． しかしながら，ここで扱うモデルではその傾向が破られる場合があることが示される．このモデルでは，要素分布の分散

がモデルの複雑度を制御するコントロールパラメータになっており，それを変化させることにより分岐現象を起こし，モデルパラメータの見かけの個数を変化させることができる． 分岐の振舞いは，適当な仮定のもとで分布の統計量に依存して異なった様相を示す． そのうちのある場合について，分岐の直後の汎化のバイアスを調べると，見かけのパラメータ数は増加するのにバイアスは逆に減少するという現象が見られる [11, 10].

第 5 章では，任意に与えられた分布の位置・尺度パラメータ (場合によっては回転パラメータも含む) の推定問題を扱い，正規混合分布の EM アルゴリズムを拡張した学習アルゴリズムを導く． 学習は 2 段階からなり，まず基本となる与えられた分布をあらかじめ正規混合分布を用いて十分な精度で近似しておく．次に与えられたデータに対して必要となるパラメータの推定を行う． 単純な正規混合分布では M ステップが指数分布族のパラメータ推定に帰着されるが，この問題の場合には必要な位置・尺度パラメータが混合分布の要素分布に共通して含まれており，曲がった部分空間 (曲指数分布族) での推定となってしまう． しかしながら，要素分布の形に制限を加えておくと，それが 2 次方程式の解として与えられることを示すことができる [6].

第 6 章では, 記号的な情報が隠れている不完全情報下での学習の例題として，複数情報源からの属性概念獲得の問題を考え，混合分布と EM アルゴリズムによる学習を実データに対して行った実験結果を示す．この問題では，音声と画像のペアのデータから，そこに隠された属性概念という構造を統計モデルを用いて学習する．この問題自体は発達心理学における概念形成のモデルと関連が深いが，工学的には近年盛んに研究されている電子秘書などのアプリケーション [38] における環境情報の学習の基礎となる研究である．このようなマンマシンインタフェースでは，ユーザに依存した情報 (好みや個人情報など) を反映させることによって，人間に優しく使いやすいものにすることができると考えられる [37, 9, 19]．しかしながら，それらの情報は多様であるのみならず，時間とともに変化する可能性もあるため，あらかじめ作りこむことは難しい． したがって，これらの情報を学習によって獲得させることが重要になるが，更に固有の問題として，カメラやマイクといった複数のチャネルを扱う必要があることが挙げられる．本研究はこういったユーザインタフェースの研究においても統計的手法が有効であることを検証する [4, 5, 8].

最後に，第 7 章では本研究の結論と今後の研究課題についてまとめる．

本論文での記号の使い方について少し説明を補っておく． まず，確率・統計学の文献では確率変数として大文字を，その実現値として小文字を用いることが多いが，本論文では基本的にすべて小文字で記す． また，確率分布は特に紛らわしくない場合を除き，すべて $p(x), p(y)$ のように p で表し，引数によりその関数を区別する． また，一般的な説明や定式化ではベクトルに成り得る変数やパラメータもボールド体のベクトル表現を特に用いない．

第2章　有限混合分布とその基本的性質

本章では，有限混合分布モデルとその基本的性質について従来行われてきた研究の概要と本論文で必要となる事項のまとめを行う．

2.1　定義

ランダム変数 x の有限個の確率密度関数 $f_1(x), \ldots, f_K(x)$ と，離散確率分布 $p_1, \ldots, p_K$ があるとき，p_k で $f_k(x)$ の重みつき線形和をとった関数

$$p(x) = \sum_{k=1}^{K} p_k f_k(x), \tag{2.1}$$

はまた確率密度関数になり，$p(x)$ を有限混合分布と呼ぶ [86]．一般に混合分布とは，K が無限個ある場合も含むが，本論文では特に断らない限り有限混合分布を単に混合分布と呼ぶ．

$f_k(x)$ および p_k は確率分布であるから，

$$f_k(x) \geq 0, \quad \int f_k(x)\,\mathrm{d}x = 1, \qquad k = 1, \ldots, K, \tag{2.2}$$

$$p_k \geq 0, \quad \sum_{k=1}^{K} p_k = 1, \qquad k = 1, \ldots, K, \tag{2.3}$$

を満たす．

$f_k(x)$ を要素分布と呼び，本論文では特にパラメータ θ_k をもつパラメトリックな分布

$$f_k(x) = f_k(x;\, \theta_k), \tag{2.4}$$

を考える．

例 1 (正規混合分布) 要素分布がすべて正規分布であるような混合分布を正規混合分布という.

$$p(\boldsymbol{x};\ p_k, \boldsymbol{\mu}_k, V_k;\ k = 1, \ldots, K) = \sum_{k=1}^{K} p_k \phi(\boldsymbol{x};\ \boldsymbol{\mu}_k, V_k), \tag{2.5}$$

ここで，$\phi(\boldsymbol{x};\ \boldsymbol{\mu}, V)$ は平均 $\boldsymbol{\mu}$, 分散共分散行列 V をもつ正規分布で，$\boldsymbol{x}$ の次元を d としたとき

$$\phi(\boldsymbol{x};\ \boldsymbol{\mu}, V) = \left(\frac{1}{2\pi}\right)^{d/2} |V|^{-1/2} \exp\left\{-\frac{1}{2} \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{V^{-1}}\right\} \tag{2.6}$$

で定義される． ただし，

$$\|\boldsymbol{x}\|^2_G = \boldsymbol{x}^{\mathrm{T}} G \boldsymbol{x}. \tag{2.7}$$

正規混合分布は，混合分布の中でも最も基本的なものの一つであり，本論文で扱うほとんどのモデルも正規混合分布に関連している.

また，第 6 章などで考える回帰の問題などでは入力 x から出力 y への条件付き分布 $p(y \mid x)$ をモデル化することが多い． この場合はランダム変数は y で，要素分布は $f_k(y \mid x;\ \theta_k)$ の形になる.

例 2 (線形回帰混合モデル) 要素分布が線形回帰モデルであるような条件付きの混合分布を線形回帰混合モデルという.

$$p(y \mid \boldsymbol{x};\ \boldsymbol{a}_k, b_k, \sigma_k^2;\ k = 1, \ldots, K) = \sum_{k=1}^{K} p_k f(y \mid \boldsymbol{x};\ \boldsymbol{a}_k, b_k, \sigma_k^2), \tag{2.8}$$

ここで，$f(y \mid \boldsymbol{x};\ \boldsymbol{a}, b, \sigma^2)$ は，$\boldsymbol{a}, b$ をパラメータに持ち，ノイズの分散が σ^2 であるような次の線形モデルである.

$$f(y \mid \boldsymbol{x};\ \boldsymbol{a}, b, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} - b)^2}{2\sigma^2}\right\}. \tag{2.9}$$

このモデルは第 6 章の多価関数の学習の際に用いる.

以下では，混合分布の特徴を，いろいろな観点からまとめてみる.

2.2 モジュール性

複雑な対象も，部分に分割すれば易しい問題に帰着できるという考え方は，分割統治として工学の広い分野で受け入れられている計算原理である． 混合分布の場

合，比較的単純な要素分布の重ね合わせによって複雑な分布を作ることでモジュール構造をもつ．混合分布では，要素分布 $f_k(x)$ が一般にオーバーラップしているので，柔軟な分割統治をしているとみなすことができる．

人工知能やニューロコンピューティングなどの分野で混合分布がモジュール構造をもつモデルとしてとらえられるようになったのは比較的最近で，Jordan ら [41, 43] が階層的エキスパートネットワークを提案してからである [22, 79].

データ解析においても，柔らかなクラスタリングとして混合分布による方法が広まっており，ファジークラスタリングとしても知られている [25]．また，判別分析で一部データのクラスラベルが欠損している場合も混合分布として定式化されることが多い．これらの応用では，モジュールは分類クラスとしての意味をもっている．

2.3 階層ベイズモデルとの関係

分布のパラメータを確率変量とみなして，その (事前) 分布を考えることによって，確率分布の分布というメタな階層構造を作ることができる．そのようにしてできる統計モデルを階層ベイズモデルと呼び，ロバストな推定を行ったり，複雑な対象に適合させるために用いられ，グラフィカルモデルとの関連から盛んに研究されている [74].

さて，混合分布で，要素分布のパラメータは固定し，重み確率 $p_1, \ldots, p_K$ に着目してみよう．ベイズ的には，要素分布を規定する確率変量 k に対する事前分布が $p(k) = p_k$ であるという階層ベイズモデル，

$$p(x, k) = p(x \mid k)\, p(k), \tag{2.10}$$

として解釈できる．そのため，p_k は (クラス k の) 事前確率と呼ばれることがある．

このような解釈に立つと，混合分布はクラスラベルと確率変量の 2 ノードの間に 1 本の有向アークがあるベイジアンネットワークの最も単純な場合とみなすことができる (ただしクラスラベルは観測されないノードである)．これにより，混合分布はベイジアンネットワークの性質を調べるための基本モデルとしての役割を果たすと考えられる．

よりベイズ的な考え方を進めると，要素分布のパラメータも変数とみてその事前分布をモデル化するという，より複雑な階層ベイズモデルを作ることができる [71] が，本論文ではそれについては扱わない．

2.4 パラメトリック性とノンパラメトリック性

混合分布では要素分布の数を変えることによって，パラメトリックな性質とノンパラメトリックな性質とを合わせ持っている．すなわち，要素分布の数を少なくすると，複雑な対象を少数のパラメータで記述するモデルになりパラメトリックモデルとして働く．一方，要素分布の数をサンプル数と同程度かそれ以上に増やして行くと個々のサンプルにフィットしたノンパラメトリックな性質が現われてくる．これは，データのもつ構造に対して大まかな視点と微細な視点とを自在に制御できるという，混合分布の柔軟性を表す性質であるといえる．

クラスタリングの例でいえば，最初全体を一つのクラスタとみなし，次第に細かく見ていき，最終的には個々のサンプルがそれぞれのクラスタとなる階層的クラスタリングと同様の振舞いである．汎化などの学習の側面からは，適当な個数のクラスタの数を選ばなければならないが，これに関しては第 4 章でより詳しく論じる．

2.5 RBF ネットワークとの関係

RBF ネットワークというのは，K 個の動径基底関数 (Radial Basis Function) と呼ばれる関数 $f_1(\boldsymbol{x}), \ldots, f_K(\boldsymbol{x})$ の線形和で定義されるモデル，

$$h(\boldsymbol{x}) = \sum_{k=1}^{K} w_k f_k(\boldsymbol{x}), \tag{2.11}$$

であり，入力 x に対する出力を学習する教師付き学習のモデルである [66, 27]．ここで動径基底関数というのは，中心 $\boldsymbol{\mu}_k$ と $\boldsymbol{x}$ との距離に依存して決まる関数のことである．典型的な例は正規分布の密度関数 $f_k(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{\mu}_k, V_k)$ で，

$$h(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \phi(\boldsymbol{x}; \boldsymbol{\mu}_k, V_k), \tag{2.12}$$

となる． これは，例 1 で示した正規混合分布に類似した形をしている． ただし，RBF ネットワークは関数を近似するのが目的であるから，w_k は確率値である必要はなく，すべての実数値を取り得る．従って，一旦動径基底関数を決めてしまえば，(最小自乗規準の下で) w_k は線形回帰係数として閉じた形で求められる． 一方，混合分布では，たとえ要素分布が固定されていたとしてもその係数である離散確率 p_k の推定は繰り返し演算を必要とする (3.4.3 参照).

一方，パラメータ $\boldsymbol{\mu}_k$ や V_k の推定は，RBF ネットワークの場合，EM アルゴリズムを用いることができないという点で混合分布よりも推定が難しい．そこで，準最適な方法として，入力 $\boldsymbol{x}$ の分布を (EM アルゴリズムなどを使って) 正規混合分布で推定し，それを動径基底関数として用いるという手法が用いられることがある．ただし，その方法では入出力関係の学習として最適化されているわけではないので，これを初期値として最急勾配法を用いて最適化を行う場合もある．

第3章　学習における汎化と EM アルゴリズム

本論文では，学習という用語を統計モデルのパラメータ推定の意味で用いる．混合分布の学習法としては，モーメント法など種々の方法が知られているが，最も一般性があるのは尤度を最大にする最尤推定法である．最尤推定は多くの場合計算が簡単であり，(正則条件のもとで) 漸近有効性をもつので，広く用いられている．

本章では，まず混合分布の最尤推定について述べ，汎化能力を評価するための尤度のバイアスに関する知見を整理する．次に，混合分布の学習アルゴリズムとして EM アルゴリズムを取り上げ，従来行われてきた研究の概要と本論文で必要とする事項のまとめを行う．

3.1　最尤推定

独立同分布に従うサンプル $X^N = x_{(1)}, \ldots, x_{(N)}$ が与えられたとき，モデル $p(x;\, \theta)$ の尤度関数の対数は，

$$\sum_{j=1}^{N} \log p(x_{(j)};\, \theta), \tag{3.1}$$

となり，これを最大にする θ が (サンプル X^N に対する) 最尤推定量である．これを一般化して，確率分布 $q(x)$ が与えられたとき，モデル $p(x;\, \theta)$ の (分布 $q(x)$ に対する) 最尤推定量は平均対数尤度

$$\int q(x) \log p(x;\, \theta)\, \mathrm{d}x, \tag{3.2}$$

を最大化する θ であるとする． サンプルに対する最尤推定量は $q(x)$ として経験分布

$$q(x) = \frac{1}{N}\sum_{j=1}^{N}\delta(x - x_{(j)}), \tag{3.3}$$

を選んだ場合に対応する． また，一般の分布 $q(x)$ に対する最尤推定量は，$q(x)$ から無限個のサンプルが得られた場合の最尤推定量の極限になっている． 本論文ではこの極限が存在するような正則条件を満たす $q(x)$ だけを考える．

情報幾何の言葉 [14, 16] でいえば，最尤推定は確率分布の空間において，対象の分布 $q(x)$ からモデルの空間への m-射影をとること，すなわち Kullback-Leibler ダイバージェンスの最小値をとるモデルを求めることと等価になっている． モデルが指数型分布族ならばその射影を求めることは易しいが，混合分布は指数型分布族ではない． 実際，混合分布の対数尤度は

$$\log\left\{\sum_{k=1}^{K} p_k f_k(x;\ \theta_k)\right\}, \tag{3.4}$$

と書け，パラメータに関して非線形な形をしている． したがって，混合分布の場合，EM アルゴリズムなどの反復法によって最尤推定量を求める必要がある．

3.2 汎化と竹内の情報量規準 (TIC)

3.2.1 汎化バイアス

学習の目的は汎化，すなわち，有限個の与えられた訓練サンプルにフィットするだけではなく，背後にある真の確率分布の構造を抽出することにある． 真の分布が既知ならば，それに対する尤度を最大にするパラメータを求めればよい．しかし実際には真の分布は未知だから，訓練サンプルだけを用いて真の分布の尤度を推定しなければならない． そのための方法として，ブートストラップなどのリサンプリングを用いる方法と，訓練サンプルに対する尤度と真の分布に対する尤度の統計量 (期待値など) を評価する方法がある．本論文で考えるのは後者の方法のうち，竹内の情報量規準 (TIC) と呼ばれている方法である．

TIC について説明する前に基本的な記号と用語を定義しておく．訓練サンプル $X^N = x_{(1)}, \ldots, x_{(N)}$ を発生する未知の確率分布を $q(x)$ とする．訓練サンプルに対

する対数尤度の $1/N$ を経験対数尤度と呼び,

$$R_{\rm emp}(\theta) = \frac{1}{N}\sum_{j=1}^{N} \log p(x_{(j)};\, \theta), \tag{3.5}$$

とおく．$R_{\rm emp}$ を最大にするパラメータが訓練サンプルに対する最尤推定量であり，$\theta = \theta_{\rm ML}$ とおく． 一方，対数尤度を $q(x)$ で期待値をとった値 (平均対数尤度) を真の対数尤度と呼び,

$$R_{\rm exp}(\theta) = {\rm E}_q\left[\, \log p(x;\, \theta)\, \right]. \tag{3.6}$$

とおく．ただし， ${\rm E}_q\left[\, \cdot \,\right]$ は

$${\rm E}_q\left[\, \cdot \,\right] = \int \cdot\, q(x)\, {\rm d}x, \tag{3.7}$$

で定義される．$R_{\rm exp}$ を最大にするパラメータ $\theta = \theta^*$ を真の最尤推定量と呼ぶことにする． 先に述べたように，本論文では θ^* が存在するような $q(x)$ だけを扱う．$R_{\rm emp}(\theta_{\rm ML}) - R_{\rm exp}(\theta_{\rm ML})$ を (最尤推定量の) 汎化バイアスと呼び，この値を評価することによって，経験尤度から真の尤度を推定し，モデル選択を行ったり，サンプル計算量を求めたりすることができる．

3.2.2 竹内の情報量規準 (TIC)

汎化バイアスの期待値は $R_{\rm emp}(\theta)$ を漸近展開して求めることができる [82, 57, 55, 58]. N が十分大きければ，汎化バイアスの期待値は漸近的に

$${\rm E}_N\left[\, R_{\rm emp}(\theta_{\rm ML}) - R_{\rm exp}(\theta_{\rm ML})\, \right] \simeq \frac{h_{\rm eff}}{N}, \tag{3.8}$$

で与えられる． ここで，${\rm E}_N\left[\ \right]$ はサイズ N のすべての訓練サンプル X^N に関する期待値,

$${\rm E}_N\left[\, \cdot \,\right] = \int \cdot \prod_{j=1}^{N} q(x_{(j)}) {\rm d}x_{(j)}, \tag{3.9}$$

をあらわす． また，$h_{\rm eff}$ は,

$$h_{\rm eff} = {\rm Tr}[H(\theta^*)^{-1} D(\theta^*)], \tag{3.10}$$

で定義される． ただし $H(\theta)$ と $D(\theta)$ は次で定義される行列である．

$$H(\theta) = -\mathrm{E}_q\left[\frac{\partial^2 \log p(x;\,\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\right], \tag{3.11}$$

$$D(\theta) = \mathrm{E}_q\left[\left(\frac{\partial \log p(x;\,\theta)}{\partial\theta}\right)\left(\frac{\partial \log p(x;\,\theta)}{\partial\theta}\right)^{\mathrm{T}}\right]. \tag{3.12}$$

h_{eff} を推定できれば，(3.8) を用いて，真の尤度 (のマイナス) を

$$-R_{\mathrm{exp}}(\theta_{\mathrm{ML}}) \simeq -R_{\mathrm{emp}}(\theta_{\mathrm{ML}}) + \frac{h_{\mathrm{eff}}}{N}, \tag{3.13}$$

と推定し，右辺を最小にするようなモデルを選べばよいことになる．この右辺を竹内の情報量規準 (TIC) と呼ぶ [82]．h_{eff} は Moody の有効パラメータ数と呼ばれることもある [55].

もし $q(x)$ がモデル集合に入っていれば, $H(\theta^*)$ と $D(\theta^*)$ はいずれも Fisher 情報行列となり，h_{eff} は H および D の次元つまりパラメータ数に一致する． これはいわゆる赤池の情報量規準 (AIC) にほかならない [78]． ただし，第 4 章では，$q(x)$ がモデル集合に入っていない場合も扱うため，TIC をそのまま用いる必要がある．

さて，式 (3.8) はバイアスの平均的な振舞いを表すものであるが，実際に TIC をモデル選択に用いる際には 1 セットの訓練サンプルを用いてバイアスを評価する必要がある．そこで，平均を取らないバイアスの振舞いについて知られている式を書くと，

$$R_{\mathrm{emp}}(\theta_{\mathrm{ML}}) - R_{\mathrm{exp}}(\theta_{\mathrm{ML}}) \simeq \frac{h_{\mathrm{eff}}}{N} + \frac{U}{\sqrt{N}}, \tag{3.14}$$

となる． ここで，$U = \sqrt{N}\{R_{\mathrm{emp}}(\theta^*) - R_{\mathrm{exp}}(\theta^*)\}$ は平均 0, オーダー 1 の確率変数である．TIC をモデル選択に用いる際には U に起因するゆらぎの大きさに注意する必要がある [72]．今，二つのモデルを比較することを考えよう．この場合には U の値の差の振舞いが実質的にモデル選択に関係する．ここで，それぞれのモデルの U の値を U_1, U_2 とする．まず，比較するモデル間に階層関係がある場合，つまり，一方のモデルが他方のモデルのパラメータの一部を固定することによって制限されたモデルである場合には，$(U_1 - U_2)/\sqrt{N}$ は $1/N$ のオーダーとなり，これは h_{eff}/N と同じオーダーである．

しかしながら，モデル間にこのような階層関係がない場合には $(U_1 - U_2)/\sqrt{N}$ が $1/\sqrt{N}$ のオーダーとなるため，これは h_{eff}/N のオーダーよりも大きく，モデル選

択の信頼性が極めて低くなる可能性がある．第 4 章で扱うモデルでは厳密には階層性は保たれていないが，TIC を用いて導かれたバイアスの性質が実際のバイアスの性質にかなり合致していることが実験的に示される．

さて，TIC をはじめとする学習における汎化能力の理論では一般に，汎化バイアスは，統計モデルの複雑度とともに増えると考えられている．統計モデルを複雑にしていけばいくほど，訓練サンプルに対する尤度を増加させていくことができる．一方，それに伴って，バイアスの値も増加するので，情報量規準の中の各項が拮抗して最適なモデルが求まるわけである．このような考え方は「ケチの原理」あるいは「オッカムの剃刀」などと呼ばれ，訓練サンプルに適応するモデルの中で，できるだけ単純なモデルを選ぶという指針を与えている．しかしながら，第 4 章で示す例では，あるクラスの正規混合モデルに対してこの傾向が破られる場合がある．

3.3 冗長性と特異性

混合分布をはじめ多層のニューラルネットワークモデルやベイジアンネットワークといった階層的モデルでは，パラメータの冗長性などに起因する特異性が最尤推定に問題を生じる可能性があることが知られている [23, 36, 29, 34, 90].

正規混合分布などの混合分布では，冗長性は次の 3 つの場合に現われる．

1. クラスラベルについて対称な形をしているので，番号を入れ換えても分布は変化しない．ただし，この性質は局所的にはそれほど問題にはならない．

2. 全く同じパラメータをもつ要素分布 $f_k(x)$ と $f_l(x)$ があったときには $p_k + p_l$ が一定のもとで，p_k, p_l をどう動かしても分布は変化しない．

3. $p_k = 0$ の場合には，要素分布のパラメータは何であっても分布は変化しない．

これらの場合 (特に 2 と 3) には Fisher 情報行列が退化するという特異性により，最尤推定の漸近有効性が成り立たない．

更に，正規混合分布で，ある特定のサンプルに平均をおき，分散を 0 にする要素分布があると，尤度が無限大になる．これは明らかに無意味な解であるが，最尤解を求める際にはこういった解に収束しないように注意する必要がある．

第 4 章で考えるモデルではこれらの特異性はできるだけ排除したモデルを考えるが，それでも残る特異性によって通常のモデルとは異なる振舞いが見られる．

3.4 EM アルゴリズム

3.4.1 一般的な特徴

EM アルゴリズムは，観測できない隠れたパラメータが存在する時に最尤推定を行うための汎用手法であり，混合分布以外にも隠れマルコフモデルやグラフィカルモデルの学習に応用されている． EM アルゴリズムは Newton 法 (あるいは Fisher のスコアリング法) や勾配法と同様，反復法によって局所最適解を求めるアルゴリズムであるが，他の手法に比べて次のような長所をもつ [30, 69, 53, 84].

1. 尤度が単調に増加することが保証されており，アルゴリズムの振る舞いが安定している． 前節で述べたように，混合分布では尤度が無限大になる無意味な解が存在するので，アルゴリズムの安定性は重要である．

2. 速度に関しても収束の初期の段階では Newton 法と同程度の速さになることが知られている (ただし，最適解の近傍では 1 次収束なので種々の加速法が考案されている).

3. インプリメンテーションが簡単になることが多い．また，これと関係して 1 ステップに要する計算量が減らせる場合もある． Newton 法では尤度の Hessian を計算する必要があるが，混合分布などでは一般に複雑な形になり，多くの計算量を必要とする．

以下ではまず 3.4.2 で，EM アルゴリズムを Dempster らによってまとめられた一般的な形 [30] で説明し，尤度の収束性や収束の速さについて知られている事柄をまとめる． 次に混合分布に限定してアルゴリズムを導く．3.4.3 では，独立同分布に従う訓練サンプルが与えられるという通常学習において仮定される条件に特殊化し，更に，要素分布が互いに独立な場合を 3.4.4 で述べる．3.4.5 では，階層的な混合モデルに有用な重み付きの形でアルゴリズムをまとめる．続いて 3.4.6 では第 5

章でも用いる EM アルゴリズムの一般化について述べ，3.4.7 では直観的なイメージを与えるために，Amari[17] によって得られた幾何学的な解釈を述べる．

3.4.2 一般的な定式化

観測値 x が与えられたとき，確率モデル $p(x;\,\theta)$ の最尤推定量を求めることが問題である． 観測される変数 x は，背後にある完全な確率変数 y の不完全な観測値であるとする．すなわち y から x への多対一の写像 $x = x(y)$ が存在するとする．またその逆像を $Y(x)$ とする．具体的には y の一部の成分だけが x として観測されるという例が典型的であり，後で述べるように混合分布の場合もそれにあたる．

EM アルゴリズムは適当な初期値 $\theta^{(0)}$ から始め，以下の E と M の各ステップを繰り返すことによって解を更新する．(E と M を合わせて 1 ステップと数えたとき) t ステップ目のパラメータの値を $\theta^{(t)}$ とする．

1. E (Expectation) ステップ : 次で定義される完全データの対数尤度の条件付き期待値を計算する．

$$\begin{aligned} Q(\theta \mid x;\,\theta^{(t)}) &= \mathrm{E}_{\theta^{(t)}}\left[\log p(y;\,\theta) \mid x\right] \\ &= \int_{Y(x)} p(y \mid x;\,\theta^{(t)}) \log p(y;\,\theta)\,\mathrm{d}y. \end{aligned}$$

2. M (Maximization) ステップ : $Q(\theta \mid x;\,\theta^{(t)})$ を θ について最大化したものを $\theta^{(t+1)}$ とおく

特に，完全データの分布 $p(y;\,\theta)$ が指数分布族,

$$p(y;\,\theta) = \exp\{\sum_i F_i(y)\theta_i - \varphi(\theta) + C(y)\}, \tag{3.15}$$

ならば，Q を微分した式がパラメータについての線形方程式になるため，EM ステップは次のような簡単な形で書き表すことができる．

$$\eta_i^{(t+1)} = \mathrm{E}_{\theta^{(t)}}\left[F_i(y) \mid x\right]. \tag{3.16}$$

ただし，η_i は期待値パラメータ $\eta_i = \mathrm{E}_\theta\left[F_i(y)\right]$ である．η と θ は互いに双対座標系になっており，Legendre 変換により互いに変換することができる．指数分布族に

対してこのような簡単な形の式が得られるのは，3.4.7 で述べる幾何学的な意味づけに関係する．

さて，Dempster らは，EM アルゴリズムが x の尤度を単調に増加させることを示した．

定理 1 (Dempster[30]) EM アルゴリズムは x の尤度を単調に増加させる．

ここではその証明の概略を示しておく． まず，ベイズの定理から，完全確率変数 y と不完全確率変数 x の分布の間に次の関係が成り立つ．

$$p(y \mid x;\, \theta) = \frac{p(y;\, \theta)}{p(x;\, \theta)}. \tag{3.17}$$

両辺の対数を取ると，x の対数尤度は

$$\log p(x;\, \theta) = \log p(y;\, \theta) - \log p(y \mid x;\, \theta), \tag{3.18}$$

となる． そこで両辺を，パラメータ $\theta^{(t)}$ を持ち，観測された x についての y の条件付き分布で期待値を計算すると，

$$\log p(x;\, \theta) = \mathrm{E}_{\theta^{(t)}}\left[\log p(y;\, \theta) \mid x\right] - \mathrm{E}_{\theta^{(t)}}\left[\log p(y \mid x;\, \theta) \mid x\right], \tag{3.19}$$

となる． 右辺第 1 項は E ステップで計算される Q である．ここで，第 2 項を

$$H(\theta \mid x;\, \theta^{(t)}) = \int_{Y(x)} p(x \mid y;\, \theta^{(t)}) \log p(x \mid y;\, \theta)\, \mathrm{d}y, \tag{3.20}$$

とおき，$\theta = \theta^{(t+1)}$ とおくと，M ステップで Q を最大化するので，

$$Q(\theta^{(t+1)} \mid x;\, \theta^{(t)}) \geq Q(\theta^{(t)} \mid x;\, \theta^{(t)}), \tag{3.21}$$

を満たす． 一方 Jensen の不等式より H は $\theta = \theta^{(t)}$ で最大値を取るので，

$$H(\theta^{(t+1)} \mid x;\, \theta^{(t)}) \leq H(\theta^{(t)} \mid x;\, \theta^{(t)}), \tag{3.22}$$

を満たす． この両者から，

$$\log p(x;\, \theta^{(t+1)}) \geq \log p(x;\, \theta^{(t)}), \tag{3.23}$$

が成り立ち，x に関する尤度の単調性が示された． 収束性に関する厳密な議論は Wu[91] によって行われ，EM アルゴリズムが適当な正則条件のもとで局所最適解または鞍点に収束することが示されている．

次に，EM アルゴリズムの収束の速度について知られている結果を述べる． 真の解 θ^* の近傍での EM アルゴリズムの振舞いは次のように一次近似できる．

$$\theta^{(t+1)} - \theta^* = I_c^{-1}(\theta^*; x) I_m(\theta^*; x)(\theta^{(t)} - \theta^*). \tag{3.24}$$

ここで，$I_c(\theta; x)$ は完全変数の分布の条件付き Fisher 情報量

$$I_c(\theta; x) = -\mathrm{E}_\theta \left[\frac{\partial^2 \log p(y;\ \theta)}{\partial\theta\partial\theta^{\mathrm{T}}} \,\middle|\, x \right], \tag{3.25}$$

であり，$I_m(\theta, x)$ は条件付き分布の条件付き Fisher 情報量

$$I_m(\theta; x) = -\mathrm{E}_\theta \left[\frac{\partial^2 \log p(y \mid x;\ \theta)}{\partial\theta\partial\theta^{\mathrm{T}}} \,\middle|\, x \right], \tag{3.26}$$

である． $I_c^{-1} I_m$ は定性的には観測されない情報の割合を表している． したがって式 (3.24) は，x の 不完全性が増すに連れ収束が遅くなることを示している．

3.4.3 独立なサンプルが与えられた時の混合分布の学習

式 (2.1) の一般の混合分布の場合について EM アルゴリズムの具体的な形を示しておく． ただし，p_k の間には (総和が 1 という条件を除き) 関数的な依存関係がないと仮定する． モデルは

$$p(x;\ p_k, \theta_k;\ k = 1, \ldots, K) = \sum_{k=1}^{K} p_k f_k(x;\ \theta_k), \tag{3.27}$$

である． まず，独立な N 個の観測値 $x = x_{(1)}, \ldots, x_{(N)}$ が与えられたときの EM アルゴリズムを導く． 各観測値がどの要素分布から出てきたか (つまりクラスラベル) がわかれば要素分布ごとに推定を行えばよいので，それを隠れた変数 $k_{(j)}$ と考え，各サンプル $x_{(j)}$ に $k_{(j)}$ をつけ加えたものを完全変数 $y_{(j)} = (x_{(j)}, k_{(j)})$ とする． $y_{(j)}$ の分布は

$$p(x_{(j)}, k_{(j)};\ p_{k_{(j)}}, \theta_{k_{(j)}}) = p_{k_{(j)}} f_{k_{(j)}}(x_{(j)}; \theta_{(j)}), \tag{3.28}$$

となる． ここで，E ステップを計算するために必要な条件付き確率を

$$q^{(t)}(k \mid x_{(j)}) = \frac{p(x_{(j)}, k;\ p_k, \theta_k)}{\sum_{k'=1}^{K} p(x_{(j)}, k';\ p_{k'}, \theta_{k'})}, \tag{3.29}$$

と定義しておく．この値は各繰り返しステップでのパラメータの値から，計算可能な式である．E ステップにおける Q は (k は離散分布なので積分は総和になり)，サンプルの独立性などから，

$$Q(y \mid x;\ \theta^{(t)}) = \sum_{j=1}^{N} \sum_{k=1}^{K} q^{(t)}(k \mid x_{(j)}) \log\{p_k f_k(x_{(j)})\}, \tag{3.30}$$

となる． 続く M ステップでは，Q を最大化する． まず，p_k について考えると，Q に $\sum_k p_k = 1$ という条件に対応する Lagrange の未定係数を加えた関数を p_k で微分し 0 とおく．すると最終的に，

$$p_k^{(t+1)} = \frac{1}{N} \sum_{j=1}^{N} q^{(t)}(k \mid x_{(j)}), \tag{3.31}$$

が得られる． 一方，θ_k については，

$$\sum_{j=1}^{N} \sum_{k'=1}^{K} q^{(t)}(k \mid x_{(j)}) \frac{\partial \log f_{k'}(x_{(j)})}{\partial \theta_k} = 0, \tag{3.32}$$

となる． ちなみに，完全変数 $y_{(j)}$ の分布は p_k については指数分布族の形をしているので，式 (3.31) は 式 (3.16) からも導くことができる．

3.4.4 独立な要素分布の場合

前節での条件に加え，更に各 θ_k の間にも関数的な依存性がなければ，式 (3.32) は

$$\sum_{j=1}^{n} q^{(t)}(k \mid x_{(j)}) \frac{\partial \log f_k(x_{(j)})}{\partial \theta_k} = 0, \tag{3.33}$$

となり，これはサンプル $x_{(j)}$ を $q^{(t)}(k \mid x_{(j)})$ で重み付けた最尤方程式であり，各要素分布毎に重み付き最尤推定を行うことを意味している． ただし，第 5 章ではこれとは異なり，θ_k 間に依存関係がある場合を扱う．

例 3 (正規混合分布に対する EM アルゴリズム) 例 1 で取り上げた正規混合分布の EM ステップは重み付き平均と分散になる.

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{Np_k^{(t+1)}} \sum_{j=1}^{N} q^{(t)}(k \mid \boldsymbol{x}_{(j)}) \boldsymbol{x}_{(j)}, \tag{3.34}$$

$$V_k^{(t+1)} = \frac{1}{Np_k^{(t+1)}} \sum_{j=1}^{N} q^{(t)}(k \mid \boldsymbol{x}_{(j)}) (\boldsymbol{x}_{(j)} - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{x}_{(j)} - \boldsymbol{\mu}_k^{(t+1)})^{\mathrm{T}}. \tag{3.35}$$

ただし, $q^{(t)}(k \mid \boldsymbol{x}_{(j)})$ および $p_k^{(t+1)}$ はそれぞれ式 (3.29), (3.31) で与えられる.

3.4.5 サンプルに重みがある場合

式 (3.32) を見ると, 独立なサンプルからの学習は, $q(k \mid x_{(j)})$ で各サンプル $x_{(j)}$ を重み付けて学習しているとみなすことができる. したがって, もともとサンプル $x_{(j)}$ に $\rho_{(j)}$ という重みがついていたときの混合分布の学習法を定式化しておくと便利である. 具体的には次のような利点がある.

1. 混合分布の要素分布がまた混合分布である Jordan らの階層的エキスパートネットワークのような再帰的な適用が可能である.
2. データがヒストグラムの形で与えられている場合にはその頻度を $\rho_{(j)}$ として扱える.

まず, 一般的な混合分布の場合, クラス事前分布 p_i の学習式 (3.31) は

$$p_k^{(t+1)} = \left\langle q^{(t)}(k \mid x) \right\rangle_{\rho}, \tag{3.36}$$

となる. ただし, $\langle\, \rangle_{\rho}$ は重み付きのサンプル平均を表し, 例えば関数 $F(x)$ の平均は

$$\langle F(x) \rangle_{\rho} = \frac{\sum_{j=1}^{N} F(x_{(j)}) \rho_{(j)}}{\sum_{j=1}^{N} \rho_{(j)}}, \tag{3.37}$$

で与えられる. また, 要素分布の学習 (3.32) は

$$\left\langle \sum_{k'=1}^{K} q^{(t)}(k \mid x) \frac{\partial \log f_{k'}(x)}{\partial \theta_k} \right\rangle_{\rho} = 0, \tag{3.38}$$

となり，独立な要素分布の場合の式 (3.33) は

$$\left\langle q^{(t)}(k \mid x) \frac{\partial \log f_k(x)}{\partial \theta_k} \right\rangle_\rho = 0, \tag{3.39}$$

となる．

3.4.6 EM アルゴリズムの一般化

EM アルゴリズムは各ステップで尤度が単調増加することが保証されており，局所最適解または鞍点に収束することが知られている [91]. しかしながら，要素分布の学習の例で見たように，一般には M ステップにおける最適化問題が陽に解けるとは限らない．そこで制限を緩め，M ステップでは，

$$Q(\theta \mid x;\, \theta^{(t)}) \geq Q(\theta^{(t)} \mid x;\, \theta^{(t)}), \tag{3.40}$$

なる θ を見つければよいことにする． この場合でも尤度の単調増加性は保たれている． この方法を一般化 EM アルゴリズム (Generalized EM) と呼ぶ． ただし，局所最適解 (または鞍点) への収束は一般には成り立たなくなる．

Meng and Rubin[52] は一般化 EM アルゴリズムの一種として ECM (Expectation-Constrained Maximization) アルゴリズムと呼ばれる手法を提案した． これは，Q の最大化をすべての変数に対して同時に行うのではなく，いくつかの変数のまとまり毎に分けて最適化を行う手法であり，変数の同時最適化が難しい最適化問題においてしばしば用いられる coordinate descent 法 [93] の一種である． 一般に最適化問題では軸ごとに最適化を行っても，被最適化関数の単調増加性は維持されている． また，ECM アルゴリズムでは局所最適解 (または鞍点) への収束が示されている．第 5 章では，ECM アルゴリズムを用いて M ステップが陽な形で得られるようにする．

本論文では扱わないが，一般化 EM アルゴリズムがより有効な場合として，ベイズ推定への拡張について説明しておく．EM アルゴリズムは最尤推定を行うアルゴリズムであるが，ベイズ推定の枠組みにおける MAP (Maximum *a posteriori*) 推定にも適用可能である．θ の事前分布 $r(\theta)$ が与えられているとき，事後分布

$$p(\theta \mid x) \propto r(\theta)\, p(x \mid \theta), \tag{3.41}$$

を最大にする θ を MAP 解と呼ぶが，この場合には Q の代わりに，

$$\hat{Q}(\theta \mid x;\, \theta^{(t)}) = Q(\theta \mid x;\, \theta^{(t)}) + \log r(\theta), \tag{3.42}$$

を最大化すればよい． 一般にこの場合はより最適化が困難となるので一般化 EM アルゴリズムを適用することが多い．

さて，一般化 EM アルゴリズムは主に M ステップを一般化する試みであるが，サンプル数が増えたり，複雑なグラフィカルモデルに EM アルゴリズムを適用する際には E ステップが組合せ的に多くの計算量を要したり，場合によっては数値積分を必要とする場合がある． 本論文では用いていないが，いろいろな研究がされているので概略を述べておく． 広く用いられているのは E ステップにおける積分計算を Monte Carlo 法を用いてサンプリングする方法である．特に，Markov Chain を用いて Monte Carlo 法を行う Gibbs サンプリングなどの Markov Chain Monte Carlo (MCMC) 法は，実際の応用などで幅広く用いられている [35]． また，その更なる近似解法である変分法や平均場近似などの研究も盛んである [33].

3.4.7 EM アルゴリズムの幾何学的解釈

Amari[17] は EM アルゴリズムの幾何学的な意味を情報幾何学の観点から明らかにした． 一般に EM アルゴリズムがどういったことをしているのかについての直観的理解の助けとなると考えられるのでここにまとめておく．

最尤推定は，サンプル点からモデルへの m-射影を取ることと解釈できるが，一般にモデルの空間が曲がっているとその推定も難しくなる． このとき，サンプルを不完全データとみなし，完全変数の分布の空間でみたときにモデルが e-平坦 (あるいはそれに準ずる空間) だったとしたら，その空間で推定を行った方が有利になる． ただし，その場合にはサンプル点は完全変数の分布の空間の 1 点ではなく，観測されない自由度の分多様体となる． これをデータ多様体と呼ぶ．

ここで，データ多様体とモデル多様体を反復して，その間の Kullback-Leibler ダイバージェンスを最小にする点を求めるアルゴリズムを考えよう (図 3.1).

1. (e-ステップ) モデル多様体からデータ多様体への e-射影をとる．
2. (m-ステップ) その点からモデル多様体への m-射影をとる．

このアルゴリズムは em アルゴリズムと呼ばれ，もともと Csiszár ら [28] によって提案され，Amari[17] で整理されたものである．

em アルゴリズムは必ずしも不完全変数の最尤推定量に収束するとは限らないが，多くの問題ではデータ多様体が m-平坦であることから，e-射影を取るのが幾何学的には自然であり，Amari[17] は em アルゴリズムが EM アルゴリズムに一致する条件を調べ，異なる場合でも両者が漸近的には等価となることを示した．指数分布族に対する EM アルゴリズムが簡単な形で得られるのも，完全変数の空間ではモデル多様体が平坦であるからであると解釈できる．

特に，分布の空間として関数空間を考えると，EM アルゴリズムと em アルゴリズムの一致が容易に導けるのでここではそれについて簡単に示しておく．関数空間は一般に多様体ではないが，em アルゴリズムは形式的に行うことができるので，以下では多様体として取り扱う．

定理 2 (Amari[17]) 分布の空間として関数空間をとって em アルゴリズムを構成すると，それは EM アルゴリズムに一致する．

これは以下のように示すことができる．まず，関数空間を考えているので，与えられたサンプル x に対応する分布は，x の分布の空間の上の $\delta(x)$ という点になる．これに対応する完全変数の分布の空間の中でのデータ多様体 D は，

$$p(y) = 0, \qquad y \notin Y(x), \tag{3.43}$$

を満たす $p(y)$ 全体の集合として表される．今，完全変数のモデルの空間を M とすると，D への e-射影は $p(y) \in D$ と $p(y;\, \theta^{(t)}) \in M$ との Kullback-Leibler ダイバージェンス，

$$K(p(y) \,\|\, p(y;\, \theta^{(t)})) = \int p(y) \log \frac{p(y)}{p(y;\, \theta^{(t)})} \, \mathrm{d}y, \tag{3.44}$$

を最小にする $p(y)$ として与えられ，

$$p(y) = \begin{cases} p(y;\, \theta^{(t)}) / p(x;\, \theta^{(t)}), & y \in Y(x), \\ 0, & y \notin Y(x), \end{cases} \tag{3.45}$$

となる．一方，この $p(y)$ から M への m-射影は，$p(y) \in D$ と $p(y;\, \theta) \in M$ との Kullback-Leibler ダイバージェンス $K(p(y) \,\|\, p(y;\, \theta))$ を最小にする θ として与えら

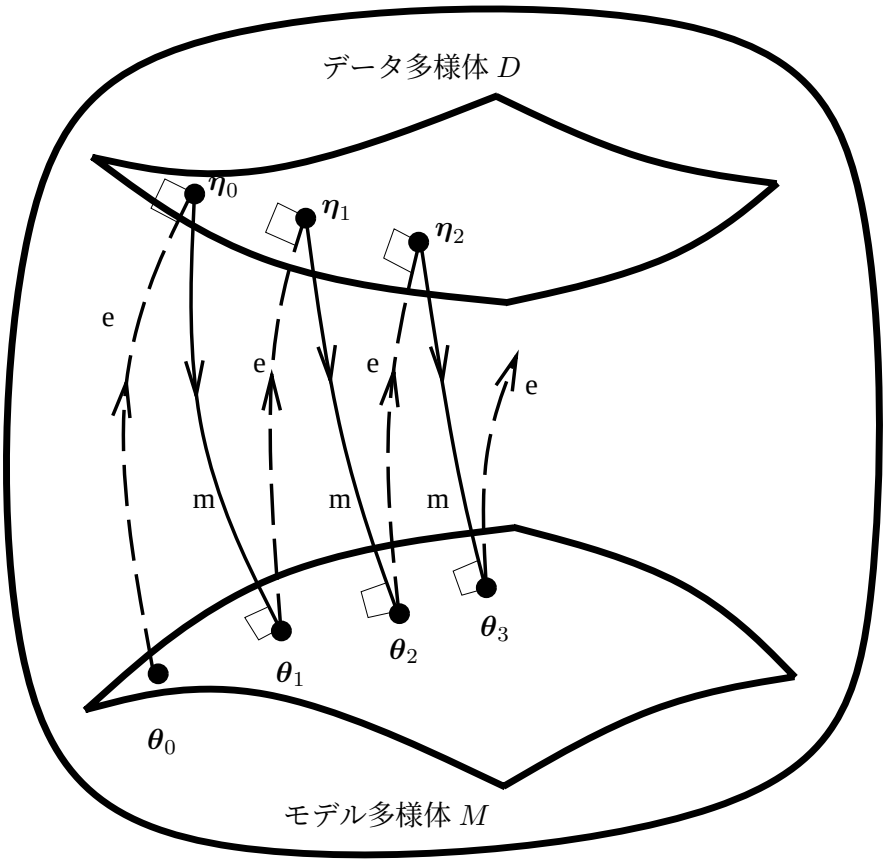

図 3.1: EM アルゴリズムの幾何学的イメージ

れる．これは式 (3.45) から,

$$\int_{Y(x)} \frac{p(y;\ \theta^{(t)})}{p(x;\ \theta^{(t)})} \log p(y;\ \theta)\, \mathrm{d}y, \tag{3.46}$$

を最大にする θ と同じであるが，上式はまさに EM アルゴリズムで最大化する $Q(y \mid x;\ \theta^{(t)})$ に他ならない.

第4章　正規混合分布の汎化バイアスの非単調性について

4.1　はじめに

3.2 節で述べたように，学習の目的は与えられた訓練サンプルにフィットするだけではなく，サンプルを生成する真の確率分布の構造を抽出することである． そのため，確率モデルの汎化能力に関する研究がさまざまに行われているが，一般に汎化バイアス (訓練サンプルに対する尤度と真の尤度との差) は，モデルの複雑度 (具体的には可変なパラメータの数) とともに増えると考えられている． 本章では，正規混合モデルに対してこの傾向が破られる場合があることを示す． ここで考えるのは統計物理モデルとして提案された Radial Basis Boltzmann Machine (RBBM) と呼ばれる，あるクラスの正規混合分布である [45]． 一般の正規混合分布では，要素分布の数を変化させることによってモデルの複雑度を変化させるが，RBBM では β という連続パラメータによって調節を行う．β の値が小さいときには RBBM の最尤解は一つの正規分布に退化しているが，β の値を次第に大きくしていくと，ある点で最尤解が相転移を起こして複数の正規分布に分岐する．β を更に大きくすると，分岐が再帰的に繰り返される．このように，β は混合モデルの「有効な」成分数を制御しているとみなせる． 本章では第一分岐点に着目して汎化バイアスを調べる．

本章の構成は以下の通りであり，そのうちオリジナルな結果は，4.3.2 以降である．まず，4.2 では RBBM を定義し，その分岐現象を説明する． 次に 4.3 で分岐の仕方に関する条件を明らかにする．その上で，4.4 では情報量規準を用いた汎化バイアスに関する解析結果を示し，分岐が 2 方向あるいは 3 方向のときは，見かけのパラメータ数が増えるにも関わらずバイアスが減少することを示す． 分岐する前の部分では経験尤度がほぼ一定値であることから，分岐後の最尤解は分岐する

直前よりも小さい誤差を達成できることを意味する．この結果は分岐点の近傍に関するものなので，厳密には正則条件等の考察が必要である． 4.5 ではこれらの点について吟味する．また，4.6 では理論的に得られた結果を数値的に確かめる．本章は，Akaho and Kappen[11, 10] に報告されている内容をより正確な形で詳しくまとめたものである．

4.2 Radial Basis Boltzmann Machine (RBBM)

重みの値がすべて等しく，すべての正規分布の分散共分散行列が等方的であるような正規混合モデルを考える．

$$p(\boldsymbol{x};\, W;\, \beta) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\beta}{\pi}}\exp(-\beta\,\|\boldsymbol{x}-\boldsymbol{w}_k\|^2). \tag{4.1}$$

ここで，W は可変なパラメータ $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K$ を表す．逆温度と呼ばれる制御パラメータ β および要素分布の数 K は学習の際は固定されている．統計モデルとしては，β は分散の逆数の 1/2 に等しいが，解析する上では分散よりも β を用いた方が式が単純になる．

このモデルは Rose ら [75] によって階層的クラスタリングの統計物理モデルとして導入されたものであるが，これを Radial Basis Boltzmann Machine (RBBM) と呼ぶことにする． RBBM はもともと Kappen[44, 45, 61] が，2 値出力の確率的動作を行うニューラルネットワークモデルであるボルツマンマシンを連続値も扱えるように拡張するものとして提案した． 式 (4.1) のモデルはその特殊な場合として定義されるが，本論文ではこのモデルのみを扱うので RBBM といえば式 (4.1) のモデルを指すことにする．

3.3 で述べたように，混合分布には冗長性や特異性がある． RBBM では，クラス事前分布を定数におくことによって，クラス事前分布に関する冗長性を排除する．また，分散を制御パラメータとして固定して考えることにより，尤度が無限大になるという無意味な局所解をなくすようにしている． 更に RBBM では分散共分散行列の等方性を仮定しているが，これは主に解析の簡単さのためであり，定性的には一般の分散共分散行列についても同様の性質が成り立っていると考える．

さて，β を変化させたときの最尤解の振舞いの例を図 4.1 に示す． 最尤解を求め

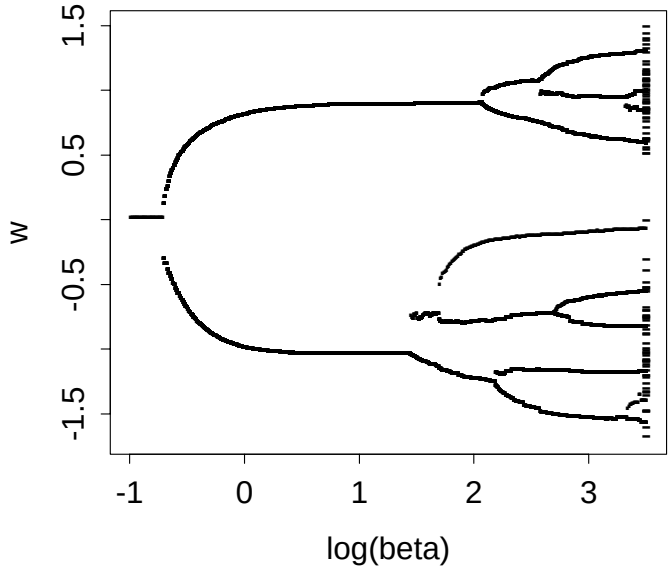

図 4.1: 分岐点での最尤解の例. 横軸: $\log(\beta)$; 縦軸: w および x. 点は各温度での最尤解右端の '–' が学習サンプルをあらわす

るのには EM アルゴリズムを用いた (したがって部分的には局所最適解に収束している可能性がある). 訓練サンプルは 1 次元上の二つの分布 $u[0.5, 1.5]$, $N[-1, 0.3^2]$ からそれぞれ 1/2 の等確率で 100 個生成した. ここで, $u[a, b]$ は $[a, b]$ 上の一様分布, $N[\mu, v]$ は平均 μ, 分散 v の正規分布を表す. 混合分布の要素数は $K = 100$ とした.

図からもわかるように, RBBM モデルは β を制御パラメータとする階層的クラスタリングと同様の構造が現われる. 小さな β では最尤解は $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \cdots = \boldsymbol{w}_K$ を満たしており, 全体が一つのクラスタとなっている. β を次第に大きくして行くと, ある β でそのクラスタは相転移を起こし, いくつかの部分に分岐する. β を更に大きい値にしていくと分岐が再帰的に起きる. 従って, 正規分布の要素数は温度を調節することによって制御できる. つまり, 要素分布の総数は K である

にもかかわらず，実際には β の値に応じて，より少ない数の要素分布が使われることになる．以上の理由から，以下の議論では K は十分大きいとしてよい．すると結局，このモデルでは複雑度は β のみに依存して制御される．

次の節に進む前に，RBBM モデルの背景について少し補足しておく．先にも書いた通り，このモデルはクラスタリングの統計物理モデルとして，またボルツマンマシンの拡張として提案されたものであり，いずれにしても統計物理的な背景を持つ．もともと 1980 年頃からニューラルネットワークの研究が盛んになったときに，Hopfield がスピングラスモデルとニューラルネットワークとの類似性を指摘し，その後 Geman and Geman が Ising モデルに基づいた画像モデルを提案して画像修復に適用したり，最近では符号化の問題との関連も研究されるなど，統計物理と情報処理の分野の距離が急速に縮まっていった [81, 62]．これらのモデルは，エネルギー関数の平衡状態を複数持ち，温度や秩序パラメータの変化によって相転移を起こすなど，統計力学的にも興味深い現象を示す．最尤推定の局所最適解が複数あり，それが温度を調節することにより相転移を起こす RBBM モデルも，まさにそのような流れで生まれて来たものである．

4.3 RBBM の分岐点

実質的なパラメータ数は分岐点で変化するので，分岐点での振舞いをより詳しく調べることにする．一般に分岐の振舞いは複雑であるが，最初の分岐点に関してはいくつかの結果を得ることができ，定性的には他の分岐点にも適用可能であると考える．

本節で考えるのは，真の分布 $q(\boldsymbol{x})$ が与えられている場合の真の最尤解の振舞いである (ただし，学習に EM アルゴリズムを用いる都合上本章の実験はすべて経験分布に対するものである)．$q(\boldsymbol{x})$ は必ずしも RBBM モデルに含まれている必要はないが，RBBM モデルの最尤解が存在するような正則条件を満たすものだけを考え，更に 4 次以下のすべてのモーメントの存在を仮定する．また，訓練サンプルを用いた学習の際の振舞いは，$q(\boldsymbol{x})$ を経験分布で置き換えて考えればよい．

4.3.1 分岐点より温度が高い場合

分岐点よりも温度が高い場合については Rose らの結果が知られている [75]．温度がある臨界点 $1/\beta_c$ よりも高いときには，すべての正規分布が一つに退化し，その解は $\boldsymbol{w}_k = \mathrm{E}_q[\boldsymbol{x}]$ となる．最初の分岐点は，

$$\beta = \beta_c = \frac{1}{2\lambda_1}, \tag{4.2}$$

で与えられる． ただし，λ_1 は $\boldsymbol{x}$ の分散共分散行列の最大固有値である． これは，$q(\boldsymbol{x})$ の第 1 主成分軸に沿った分散と RBBM の要素分布の分散 $(1/2\beta)$ とが一致したときに最初の分岐が起きることを示している．

この結果はそれ以降の分岐点に対して定性的には次のように働くと考えられる．分岐した各クラスタは β を大きくするともに次第に距離が離れていく．もし，クラスタ間の距離が十分に大きければ，サンプル点はより近いほうのクラスタの共分散行列のみに貢献するので，それぞれのクラスタに属するサンプル点の共分散行列の固有値が，それぞれのクラスタの分岐を決定すると考えられる．

4.3.2 分岐点より温度が低い場合

まず，以下の仮定をおく．

仮定 1 真の分布 $q(x)$ が 1 次元実空間 $\Re$ の上で定義され，対称な分布であると仮定する．また，RBBM における正規分布の要素数 K は偶数であるとする．

ここで，$\sigma^2 = \mathrm{E}_q\left[(x - \mathrm{E}_q[x])^2\right]$ および $s_4 = \mathrm{E}_q\left[(x - \mathrm{E}_q[x])^4\right]$ は $q(x)$ の 2 次および 4 次の (平均まわりの) モーメントとする． 4 次のキュムラントは $\kappa_4 = s_4 - 3(\sigma^2)^2$ と定義される (κ_4/s_4 は尖度と呼ばれる).

さて，上の仮定の下で，分岐点の振舞いを特徴付けることができる．

定理 3 仮定 1 の下で，十分 K が大きいとき，第 1 分岐点の振舞いは以下の 3 通りに分類される:

1. $\kappa_4 < 0$ のとき分岐は 2-way: 要素分布は 2 個のクラスタにわかれる．第 1 分

岐点 β_c の近傍での最尤解と β の関係は

$$\Delta\beta \simeq \frac{s_4}{6(\sigma^2)^4}(\Delta w_k)^2, \tag{4.3}$$

で与えられる． ただし $\Delta\beta = \beta - \beta_c > 0$, $\Delta w_k = w_k - \mathrm{E}_q[\,x\,]$.

2. $\kappa_4 > 0$ のとき分岐は 3-way: 要素分布は 3 個のクラスタにわかれる． 第 1 分岐点 β_c の近傍での最尤解は $w_k = 0$ または $\Delta\beta \propto (\Delta w_k)^2$.

3. $\kappa_4 = 0$ のとき分岐は K-way: 要素分布は K 個のクラスタにわかれる．個々の要素 Δw_k は 3 次展開まででの近似では決定条件が足りないが，第 1 分岐点の近傍での最尤解は,

$$\Delta\beta \simeq \frac{1}{2(\sigma^2)^2}\sigma_w^2, \tag{4.4}$$

を満たす． ここで $\sigma_w^2 = \dfrac{1}{K}\sum_k {\Delta w_k}^2$, $\Delta w_k = w_k - \mathrm{E}_q[\,x\,]$.

もし $q(x)$ が正規分布のときは κ_4 は 0 になる． したがって上記の条件は $q(x)$ の 4 次のキュムラントの意味での正規分布との類似性に関係している．2-way になるか 3-way になるかは κ_4 の符号，すなわち尖度に依存して決まるが，定性的には正規分布よりも尖った分布であれば 3-way に分岐し，尖り方が小さい分布では 2-way になる．3-way になる場合には $w_k = \mathrm{E}_q[\,x\,]$ を満たす要素分布の個数が $q(x)$ によって変化し，それに伴って $w_k \neq \mathrm{E}_q[\,x\,]$ となる w_k の具体的な値も変わるので定理の中では与えなかった． この値は後で述べる定理 4 に関係するので，具体的にはこれらの定理の証明を含めて付録 A.1 の中で示す．ちなみに，Akaho and Kappen[10] では 2-way になる場合だけを述べているが，本論文ではより精密な解析により，3-way になる場合が存在することを示した．

一方，式 (4.4) は $\kappa_4 = 0$ の場合 3 次までの漸近展開では解が一意に定まらないことを意味している．4 次以上の展開をすればこの自由度は減らせると考えられるが，このような解は非常に不安定である．むしろ決定条件が足りない式のままで次のように解釈するのが有用であろう．仮に無限にたくさんの要素分布があったとすると，逆温度 β の正規密度を $\phi(x;\,\beta)$ として，モデル $p(x;\,w) = (1/K)\sum_{k=1}^{K}\phi(x - w_k;\,\beta)$ は，たたみこみ

$$p(x;\,w) = \int r(w)\,\phi(x - w;\,\beta)\,\mathrm{d}w, \tag{4.5}$$

の形に書ける．ただし，$r(w)$ は w の分布を表す．ここで $q(x)$ が正規分布だとすると，分岐点より β が大きい場合には $r(w)$ も正規分布になり，その分散は式 (4.4) の関係を満たす．

さて，汎化バイアスの解析について述べる前に，分岐点の近傍での尤度の振舞いについて触れておく．

定理 4 仮定 1 の下で，RBBM の最尤解に対する尤度の β に関する傾きは，第 1 分岐点で 0 となる．

先に述べたように，この定理の証明は付録 A.1 で与える．この定理を経験尤度に対して適用し，次の節で述べる汎化バイアスに加えることにより，汎化能力の振舞いが評価できることになる．

4.4 汎化バイアスの非単調性

ここでは，RBBM の分岐点の前後での竹内の情報量規準 (TIC : 3.2.2 参照) において現われるバイアス項の振舞いを調べる．分岐点と分岐点の間では RBBM の有効な可変パラメータの数は一定であり，分岐のたびに増加する．TIC はモデルの複雑度を測っているので，β が増えれば TIC も増えると予想される．しかしながら，解析の結果，分岐点の前では TIC は β に関して線形に増加するが，分岐点の直後では κ_4 に依存して減少する場合があることを示す．前節で仮定したのと同様に，真の分布 $q(\boldsymbol{x})$ は，RBBM モデルの最尤解が存在するような正則条件を満たし，4 次以下のキュムラントがすべて存在するとする．

4.4.1 分岐点より温度が高い場合

一つのクラスタしかないときには ($\beta < \beta_c$)，有効パラメータ数 $h_{\rm eff}$ を陽に計算できる．以下の定理は汎化バイアスが β に比例して大きくなることを示している．β を制御パラメータとするモデルの $h_{\rm eff}$ を $h_{\rm eff}(\beta)$ と書く．

定理 5 $\beta < \beta_c$ のとき $h_{\rm eff}$ は

$$h_{\rm eff}(\beta) = 2\beta {\rm Tr}[V_{\boldsymbol{x}}], \tag{4.6}$$

で与えられる． ただし，$V_{\boldsymbol{x}}$ は $q(\boldsymbol{x})$ の共分散行列.

定理 5 の証明は付録 A.2 で与える．この定理により，分岐点より前では β が増加するとともにバイアスも増加することを示しており，従来から知られている汎化能力に関する一般的な結果に一致する.

4.4.2 分岐点より温度が低い場合

分岐した後は一般に複雑な振舞いを示すので，定理 3 の場合のうち，分岐が 2-way および 3-way の TIC を解析することにする．ここで，訓練サンプルに関する最尤解の要素分布の数に関する次の仮定を新たに置く.

仮定 2 真の分布に対する最尤解の分岐が 2-way または 3-way である場合を考える．このとき，真の最尤解と訓練サンプルに対する最尤解とで，分岐に含まれる要素分布の個数が等しいとする．すなわち，真の分布に対する最尤解の分岐が m-way ($m = 2$ または $m = 3$) であるとし，各分岐に含まれる要素分布の個数を $l_1, \ldots, l_m$ としたとき，訓練サンプルに対する最尤解も同様に m-way であり，各分岐に含まれる要素分布の個数もやはり $l_1, \ldots, l_m$ であるとする.

この仮定は，あくまで個数に関する仮定であり，それぞれの分岐におけるパラメータ w_k の値が等しい必要はない．この仮定の下で以下の定理が成り立つ.

定理 6 仮定 1 および仮定 2 のもとで，$\kappa_4 \neq 0$ かつ $s_4 \neq (\sigma^2)^2$ とする． このとき，$\lim_{\beta \to \beta_c} h_{\text{eff}}(\beta) = 1$ であり，

$$\lim_{\beta \downarrow \beta_c} \frac{\partial}{\partial \beta} h_{\text{eff}}(\beta) = -\infty \tag{4.7}$$

が成り立つ． $s_4 \neq (\sigma^2)^2$ という条件は 2 個の δ 関数の混合分布を除くすべての分布について成り立つ． $q(x) = (\delta(x-1) + \delta(x+1))/2$ のときは $\partial h_{\text{eff}}(\beta_c)/\partial \beta = -4$ となる.

定理 6 の証明は付録 A.3 で与える． 定理 6 は，分岐によってみかけのパラメータ数が増えても RBBM の TIC が減少することを主張している．K-way の分岐に関しては最尤解が一意的に求まらないので解析ができていない．これに関しては 4.6 で実験的に調べるが，実験結果ではその場合には非単調性は観察されなかった.

定理 4 で見たように，分岐点のまわりでは経験尤度はほぼ定数なので，2-way または 3-way の分岐の場合 RBBM は分岐点の直前よりも分岐点の直後の方が (TIC の意味で) 汎化がよいことを意味している． また，分岐点は汎化バイアスの極大点になっているので，それより前と後にそれぞれ局所最適解が存在している．

仮定 2 を置いたのは主に解析上の都合によるものである．この仮定は，K の数がサンプル数に比べて十分少ないときは仮定を満たす解が最尤解になる可能性が高い．しかしながら，K の数がサンプル数と同じかそれ以上のオーダーの場合には，サンプルの偏りに応じて非対称な解などが最尤解になり，仮定を満たすモデルは局所的な最適解になっている．仮定が破れる解は，訓練サンプルの連続な変化に対して (要素が無限個あったとしても) 不連続に現われるため，漸近展開による解析を困難にしている．更に，非対称な場合には最尤解の振舞いが明らかでないこともこの仮定を置く理由である．ただし，TIC は局所最適解に関しても成り立つ理論なので，(次節で述べる特異性に関する問題点は残るが) 定理 6 は局所最適解に関しても意味のある定理である．

TIC を用いた解析のもつもう一つの問題点は，3.2.2 で述べたゆらぎの問題である． 1 つの訓練データセットに対するバイアスの振舞いは特に階層的なモデルでない場合は深刻となる．RBBM の場合はモデルに制約を加えるという意味での階層性は持っていないので，大きなゆらぎをもつ可能性があるが，4.6 に示す実験結果では，1 つの訓練データセットでのバイアスの振舞いも，理論的な結果にある程度合致している．

さて，本章では β を制御パラメータとして固定したが，β も学習の対象とした場合について考えてみよう． この場合は，K を大きく取ると，それぞれのサンプルの上での δ 関数という無意味な解に収束してしまうので，K を制御する必要がある． $K = 1$ ならば，第 1 分岐点が解になり，K を大きくする毎にそれ以降の分岐点が解になる．つまり，β も学習するとすると，分岐点が最尤解になるのである．ところが，定理 6 から，そのような解は汎化能力が局所的には最も悪い解になっている可能性がある．このような場合には，β を少し小さい値にするか，K を一つ増やして分岐点の直後の解を選んだ方が汎化能力が優れていることになる．具体的にはクロスバリデーションやベイズ的な方法によって避けることができるであろう．

4.5 特異性に関する考察

前節で得られた結果は，特異点である分岐点の近傍における結果なので TIC の理論で仮定されている正則条件に関してのいくつかの問題点がある．そこで，本節ではそれらについて考察を行う．

TIC では，経験的な最尤推定量が局所的に正規分布をなすことを仮定している．しかしながら，分岐点の近傍ではこれが満たされないことがある．すなわち，訓練サンプルのゆらぎによって，(経験的な) 分岐点は変動し，これによって経験的な最尤推定量がなめらかでない変形を受ける．例えば，真の分布に対する分岐点より β が小さい場合を考えよう．このとき，真の分布に対しては 1 個の正規分布モデルが当てはまるが，訓練サンプルのゆらぎによって，その β が経験的な分岐点 β_c よりも大きくなることがある． すると，この場合は 2 個の正規分布モデルが当てはまってしまう．また，その逆に，真の分布が 2 個の正規分布モデルとなる β に対し，訓練サンプルによっては 1 個の正規分布モデルが当てはまる場合がある．

このようなことが起きるのは分岐点のゆらぎのオーダーの幅の範囲内であり，$O(1/\sqrt{N})$ である． 前節で得られた結果が理論的に適用可能なのは真の分岐点に対して $O(1/\sqrt{N})$ を除いた外側ということになる．この幅はサンプル数が増加すれば減少していく．

分岐点のゆらぎのオーダーの範囲内でバイアスがどうなっているかは，厳密には明らかではないが，近年行われている特異点に関する統計的な解析手法が応用できる可能性はある [36, 29, 34, 90]．本論文では，次節の計算機シミュレーションを用いた実験によって調べ，定性的には分岐点の極近傍でも理論的な結果が成り立つことを示す．

4.6 実験

本節では，定理 3 で与えた 4 次のキュムラントの 3 つの場合に対するシミュレーション結果を示す． いずれの場合も真の分布 $q(x)$ は平均 0.0, 分散 1.0 となるように設定する． このとき真の分布に対する分岐点は $\log \beta_c \simeq -0.693$ となり，訓練サンプルに関する最尤解はこの値のまわりで分岐を起こす． 訓練サンプルの数 (N)

と正規分布の要素数 (K) はともに 100 にした． 最尤解を求めるのには EM アルゴリズムを用いた． EM アルゴリズムの初期値には，各訓練サンプルに一つの正規分布をおくというものをとった．テストサンプルは 100,000 個を訓練サンプルと同じ分布から生成した．

4.6.1 $\kappa_4 < 0$ の場合

真の分布は 2 個の正規分布の混合分布にとった．

$$q(x) = \frac{1}{2}\sqrt{\frac{C_1}{\pi}}\left[\exp\left\{-C_1(x-C_2)^2\right\} + \exp\left\{-C_1(x+C_2)^2\right\}\right], \qquad (4.8)$$

ただし $C_1 = 12.5, C_2 = \sqrt{0.96}$ で，$q(x)$ の分散が 1.0 になるようにした． この分布の 4 次キュムラントは約 -1.84 である．温度を変えたときの最尤解の分布の例を図 4.2 に示す． これに対応する経験尤度は図 4.3 のように変化する．また，100,000 個のテストサンプルを用いて推定した真の尤度と経験尤度との差を汎化バイアスの推定値としてプロットしたのが図 4.4 である．定理 6 にあるようなバイアスの非単調性が観察される．

ただし，乱数の初期値によって生じる非対称性によって，仮定 2 から大きく外れた解が最尤解になる場合には，非単調性が観測されない場合もあった．

4.6.2 $\kappa_4 > 0$ の場合

真の分布は 3 個の正規分布の混合分布をとった．

$$\begin{aligned} q(x) \;=\; & \sqrt{\frac{C_3}{\pi}}\Big[0.8\exp\left\{-C_3x^2\right\} \\ & +0.1\exp\left\{-C_3(x-C_4)^2\right\} + 0.1\exp\left\{-C_3(x-C_4)^2\right\}\Big], \end{aligned} \qquad (4.9)$$

ただし，$C_3 = 50.0, C_4 = 2.22486$ で $q(x)$ の分散が 1.0 になるようにとった． この分布の 4 次のキュムラントは約 1.96 である．最尤解の例と経験尤度，テストサンプルを用いて推定した汎化バイアスをそれぞれ図 4.5，図 4.6，図 4.7 に示す．この場合も定理 6 の結果が定性的に観察される．また前節同様，仮定 2 から大きく外れた解が最尤解になる場合には，非単調性が観測されない場合もあった．

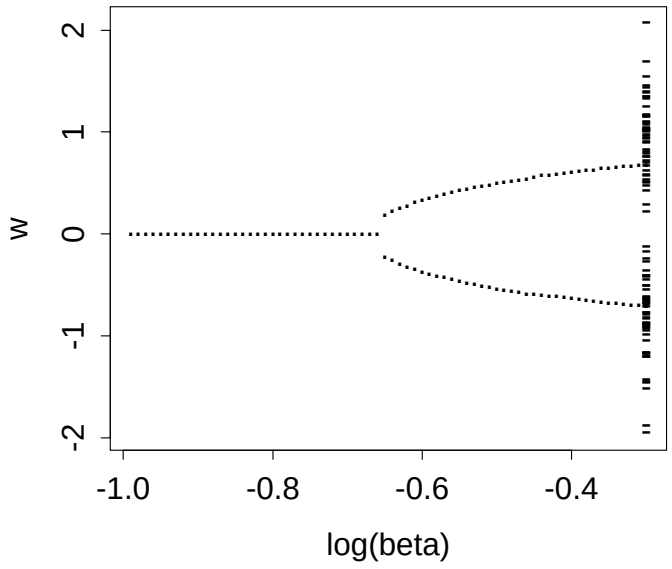

図 4.2: $\kappa_4 < 0$ のときの最尤解．横軸: $\log(\beta)$; 縦軸: w および x. 点は各温度での最尤解右端の ‘–’ が学習サンプルをあらわす

4.6.3 $\kappa_4 = 0$ の場合

真の分布は標準正規分布とした．

$$q(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}). \tag{4.10}$$

この場合，定理 3 で述べた理由により，EM アルゴリズムの収束は前の節の場合よりも非常に不安定になる． 最尤解の一例を温度の関数として示したのが図 4.8 である．理論的にはこの場合は要素分布が K 個に分かれるはずであるが，その性質は訓練サンプルの 4 次のキュムラントに依存しており，4 次のキュムラントが 0 から外れることが多いため実際には 2-way または 3-way の分岐を示す．経験尤度とバイアスの振舞いをそれぞれ図 4.9，図 4.10 に示す．キュムラントのゆらぎによって 2-way または 3-way に分岐しても，バイアスに関しては分岐点以降で減少する

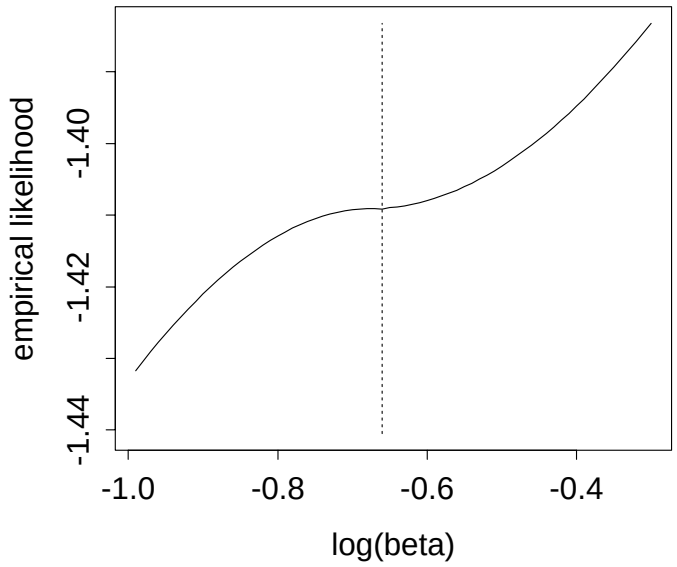

図 4.3: $\kappa_4 < 0$ のときの経験尤度．横軸: $\log(\beta)$; 縦軸: 経験尤度; 破線 : β_c の経験値

ような場合はほとんど観察されず，4 次キュムラントが 0 に近い場合には定理 6 の仮定が満たされにくい不安定な状態にある可能性がある．

4.7 本章のまとめ

正規混合分布の特殊なクラスである Radial Basis Boltzmann Machine と呼ばれるモデルの汎化バイアスの非単調な振舞いを示した． 高い温度 (つまり小さい分散) では汎化バイアスは β とともに線形に増加する． 一方，分岐点では 4 次のキュムラント κ_4 に依存した分岐の振舞いがみられた．

$\kappa_4 \neq 0$ のとき，汎化バイアスは分岐した直後では β の増加とともに減少する．バイアスが減少している間，可変なパラメータの数は増えている．これは通常 TIC が可変なパラメータ数を測っているという暗黙の了解を破っている． この現象は，

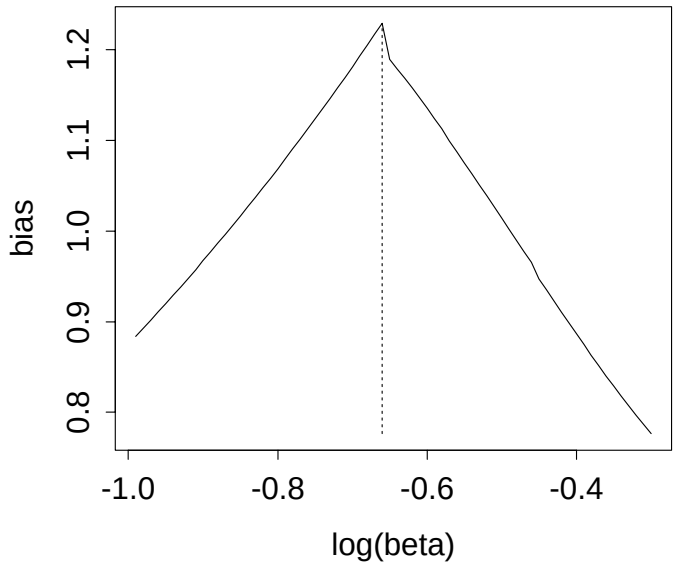

図 4.4: $\kappa_4 < 0$ のときのバイアスの推定値の振舞い．横軸: $\log(\beta)$; 縦軸: バイアスと訓練サンプルの数の積; 破線 : β_c の経験値

β の最適な選択に影響を与える． 最適な汎化能力が分岐点のあたりで達成されるときには，β を増やした方が経験尤度もバイアスも減らせるのである．

$\kappa_4 \simeq 0$ のとき，理論的には分岐が K-way であることを予測したが，数値的な不安定性などから，実験的には観察されなかった．

ここでの解析は 1 次元で対称な分布であったが，高次元でも定性的な性質が保たれていると考える．

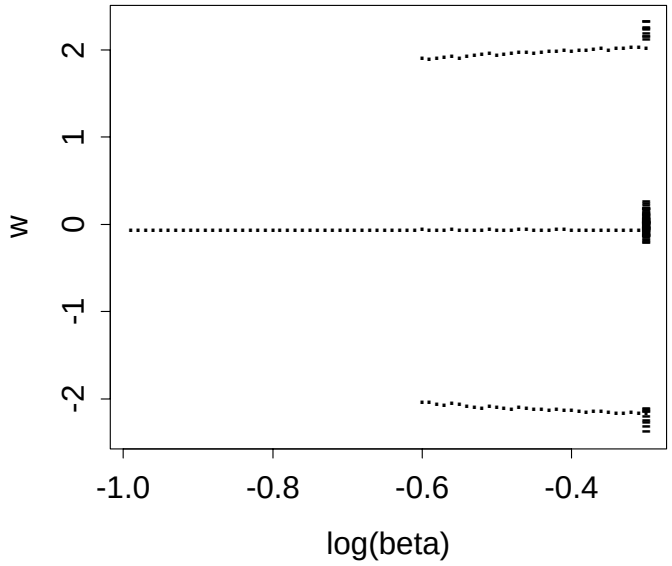

図 4.5: $\kappa_4 > 0$ のときの最尤解. 横軸: $\log(\beta)$; 縦軸: w および x. 点は各温度での最尤解右端の '–' が学習サンプルをあらわす

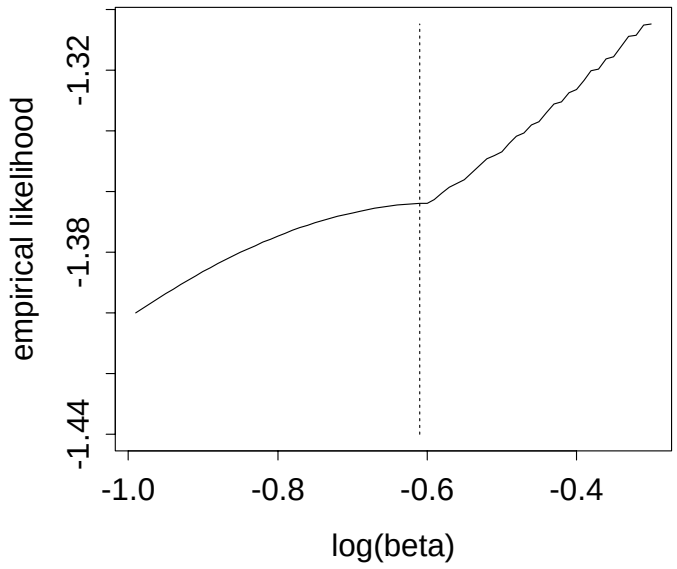

図 4.6: $\kappa_4 > 0$ のときの経験尤度．横軸: $\log(\beta)$; 縦軸: 経験尤度; 破線 : β_c の経験値

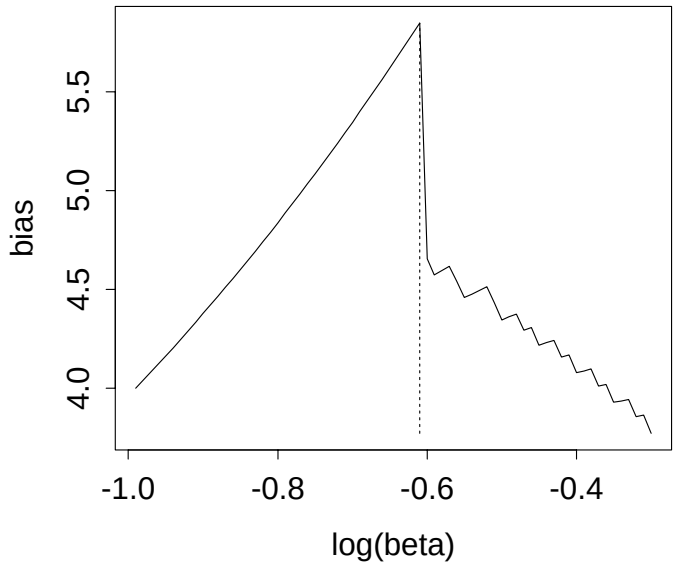

図 4.7: $\kappa_4 > 0$ のときのバイアスの推定値の振舞い．横軸: $\log(\beta)$; 縦軸: バイアスと訓練サンプルの数の積; 破線 : β_c の経験値

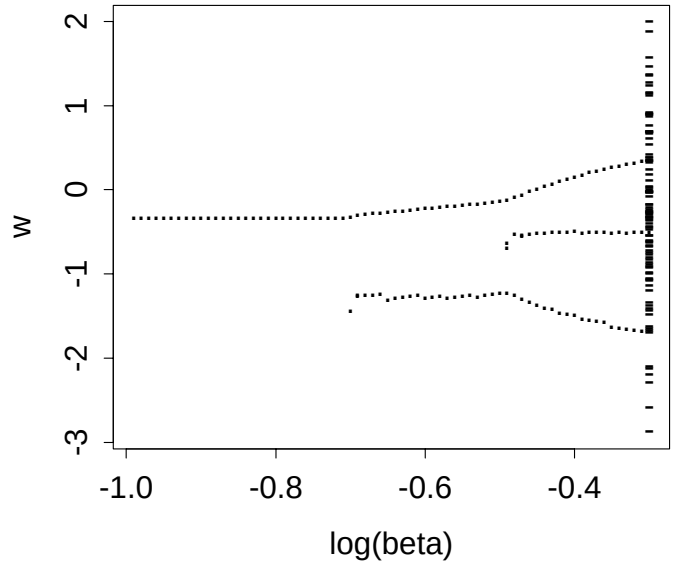

図 4.8: $\kappa_4 = 0$ のときの最尤解. 横軸: $\log(\beta)$; 縦軸: w および x. 点は各温度での最尤解右端の '$-$' が学習サンプルをあらわす

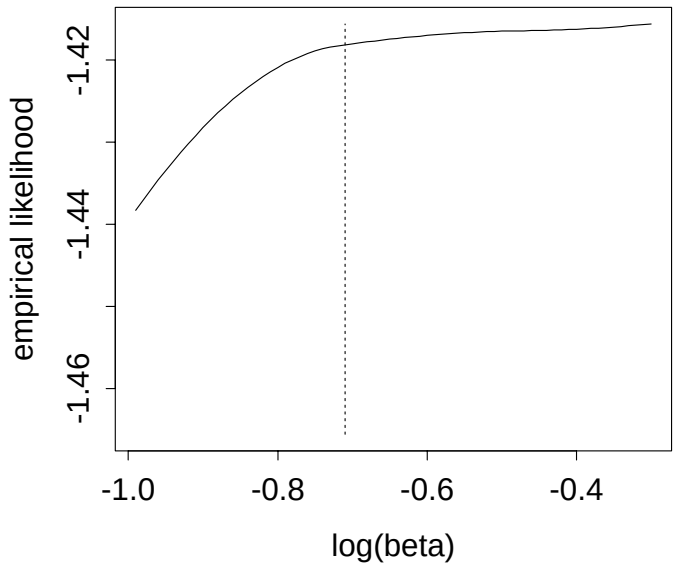

図 4.9: $\kappa_4 = 0$ のときの経験尤度．横軸: $\log(\beta)$; 縦軸: 経験尤度; 破線 : β_c の経験値

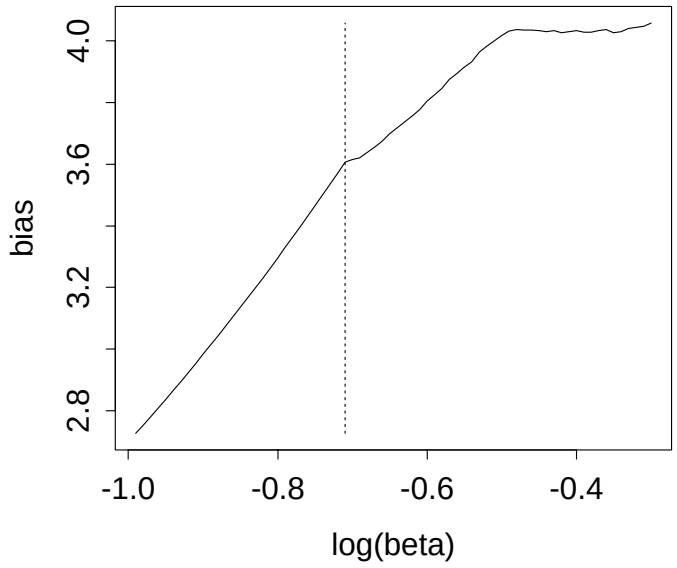

図 4.10: $\kappa_4 = 0$ のときのバイアスの推定値の振舞い横軸: $\log(\beta)$; 縦軸: バイアスと訓練サンプルの数の積; 破線 : β_c の経験値

第5章　確率分布の位置, 尺度, 回転パラメータの学習法

5.1　はじめに

本章では，任意の形の確率分布が与えられた時に，その位置・尺度等を調整してデータにもっともあてはまるようなパラメータを学習する手法について考える．位置や尺度パラメータをもつ確率分布に関する学習は，ロバスト推定 [39] やセミパラメトリック推定 [26] といった統計的推定の枠組の中でも基本的な問題である．また，空間上に分布するデータ点の集合にあらかじめ形のわかっている確率モデルを伸縮してあてはめる問題はテンプレートマッチングなど統計的パターン認識 [40] などにも幅広く応用されている．

しかしながら，このようなモデルでは，正規分布のような数少ない例外を除き，一般に位置や尺度パラメータの推定は容易ではない．これは，3 章で述べたように，幾何学的にはモデルの空間が曲がっていることに起因している．ロバスト推定では正規分布以外の分布が用いられるので，一般にパラメータ推定は Newton 法や勾配法で行われ，場合によっては安定性や速度の点で問題が生じることがある． さらに，テンプレートマッチングでは，複雑な形状を当てはめることが多いので勾配法は適用するのが困難で，全探索やランダム探索を行わざるを得ないこともある．

さて，3.4 で， EM アルゴリズムが複雑な確率モデルの学習を単純化できる場合があることを述べた．そこで，本章では混合分布に対する EM アルゴリズムを応用して，位置・尺度パラメータの学習アルゴリズムを導く．まず，あらかじめ与えられた確率分布を正規混合分布で近似しておく．これは正規混合分布のノンパラメトリック推定としての性質から任意に必要な精度で行うことが可能である [85]．その上で，与えられたサンプルに対し，その正規混合分布を伸縮して学習を行う．ただし，この場合位置や尺度のパラメータは正規混合分布の独立したパラメータでは

なく，非線形な部分多様体をなしているため，単純な正規混合分布のように閉じた形のアルゴリズムを求めることは困難である．そこで，近似を行う正規混合分布を特別なクラスに限定した上で，一般化 EM アルゴリズムの一種である ECM アルゴリズム (3.4.6 参照) を適用し，位置パラメータと尺度パラメータを順に最適化することにすると，推定アルゴリズムが 2 次方程式の解として閉じた形で得ることができることを示す．

EM アルゴリズムを用いると，尤度が単調に増加し，Newton 法などに比べて安定性が高く，単純な勾配法に比べて速度の点でも優れたアルゴリズムが得られる．また，もう一つの利点として，複数のモデルに対するデータあてはめに対して自然に拡張できることが挙げられる． つまり，位置や尺度のパラメータをもつ分布が複数個あったときに，それらの混合分布として全体の分布をモデル化し，階層的に EM アルゴリズムを適用することによって学習を行うことができるということである．これは，Jordan らによって提案された階層的エキスパートネットワーク [41, 43] の階層的なアプローチと類似のもので，基本的にはどちらも 3.4.5 に述べた重み付きの最尤推定を再帰的に行えばよい．しかしながら，彼らの手法が往々にしてパラメータ数過剰になりやすいのに対し，本論文での手法は位置や尺度といった本質的なパラメータのみに限定している点が異なる．

また，複数個のモデルがないような場合でも，モデル以外に存在するばらまきノイズを一様分布とみなして，モデルと一様分布の混合分布モデル (一種の汚染モデル [39]) の学習を行うと，ノイズに対してロバストな推定法が得られる．

以下では，まず最初に 2 種類の確率モデルを導入する． 一つは任意の次元の位置-尺度モデル [2] であり，もう一つは物体認識など応用上重要と思われる 2 次元のモデルで位置と尺度のほかに回転のパラメータを含むモデルである．次に，これらのモデルを適当なクラスに属する正規混合分布で近似し，ECM アルゴリズムを用いてパラメータ推定する方法について述べる．さらに，導かれたアルゴリズムが正しく動作することを示すために，複数モデルのあてはめなどを含めた簡単な実験を行う．

5.2 位置・尺度・回転パラメータ

あてはめの対象となるデータは d 次元実空間中の N 個の点集合 $\{\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(N)}\}$ として与えられるとする．モデルは同じ空間の上の確率分布 $p(\boldsymbol{x};\, \boldsymbol{\theta})$ として表現されており，データに最もよくあてはまるように位置や尺度のパラメータ $\boldsymbol{\theta}$ を求めることが問題である．

以下では，ECM アルゴリズムの閉じた形の更新式を得るために二つのモデルを考える． 一つは位置と尺度のパラメータだけをもつ場合でありこれを便宜上 Type I モデルと呼ぶことにする．もう一つは $d = 2$ に限定した上で，位置と尺度のほかに回転のパラメータを含めた場合であり，こちらを Type II モデルと呼ぶ．

以下ではまず，任意の形の確率分布が与えられたときに，位置や尺度，回転のパラメータがどのように入るかを Type I と Type II それぞれの場合について説明し，5.3 で，与えられた確率分布を適切な正規混合モデルで近似する．

5.2.1 Type I モデル

元となるモデルの確率分布が $f(\boldsymbol{x})$ で与えられたとしよう．各座標軸 x_i 方向に関する尺度パラメータ a_i，移動パラメータを b_i とする (ただし，$a_i > 0$)．すると，位置と尺度変換のパラメータをもつ確率モデルは

$$p_{\mathrm{I}}(\boldsymbol{x};\, A, \boldsymbol{b}) = |A| f(A\boldsymbol{x} + \boldsymbol{b}), \tag{5.1}$$

と書くことができる． ここで，行列 A は尺度に対応する

$$A = \begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_d \end{bmatrix}, \tag{5.2}$$

という対角行列で，$|A|$ は A の行列式 (この場合は対角要素の積) をあらわす．この分布は $f(\boldsymbol{x})$ を各軸方向に伸縮，移動したものである．すべての軸の尺度パラメータが同じ値であると制限をおけば，相似形の変換だけを扱うことになり，その場合もほとんど同じに扱うことができる (具体的には本論文で得られた a_i の算術平均をとればよい) が，本論文では各軸が独立に伸縮してよいとして定式化する．

5.2.2 Type II モデル

物体認識などへの応用を考えた場合には，対象はモデルに対して回転している場合もある．しかしながら，一般の次元では回転パラメータの推定問題は複雑になりすぎて，アルゴリズムを簡単な形であらわすことができない．ところが，問題を 2 次元に限ると回転パラメータを含めた場合についても Type I と同様な閉じた形のアルゴリズムを導くことができる．ただし，この場合には各軸の方向の尺度パラメータは同一 (相似変換) であると制限する必要がある．

$f(\boldsymbol{x})$ に位置と尺度，および回転の変換を施してできる確率モデルは

$$p_{\mathrm{II}}(\boldsymbol{x};\ H, \boldsymbol{b}) = |H| f(H\boldsymbol{x} + \boldsymbol{b}) \tag{5.3}$$

の形で書ける．ここで行列 H は回転と尺度に対応する行列，

$$H = \begin{bmatrix} h_1 & h_2 \\ -h_2 & h_1 \end{bmatrix} \tag{5.4}$$

である．

5.3 正規混合分布による近似

前節で説明した確率モデルは任意の形の分布 $f(\boldsymbol{x})$ に基づいたものであるが，このままの形では A あるいは H と $\boldsymbol{b}$ の推定を簡単に行うことはできない．そこで，$f(\boldsymbol{x})$ を適当な正規混合分布であらかじめ近似しておくことにする．$f(\boldsymbol{x})$ が適当な正則条件を満たせば，正規混合分布は $f(\boldsymbol{x})$ を任意に必要な精度で近似でき [85]，かつ，EM アルゴリズムに基づく単純な推定アルゴリズムを導くことができる．ただし，正規混合分布で近似したとしても A や H，$\boldsymbol{b}$ の推定は自明ではない (5.4 参照)．

5.3.1 Type I モデルの場合

正規混合分布は正規分布の重み付きの和として定義される分布であるが，閉じた形の EM アルゴリズムを得るためには一般的な形をした平均と共分散をもつ正規分布ではなく，対角行列を共分散としてもつ次のような K 個の正規分布 (楕円型正

規分布 [65]) の混合分布 $\hat{f}_{\mathrm{I}}$ で f を近似する必要がある．

$$\hat{f}_{\mathrm{I}}(\boldsymbol{x}) = \sum_{k=1}^{K} \xi_k \phi(\boldsymbol{x};\ \boldsymbol{\mu}_k, \Sigma_k). \tag{5.5}$$

ここで，ξ_k は非負の実数で $\sum_{k=1}^{K} \xi_k = 1$ を満たす．Σ_k は対角行列で，

$$\Sigma_k = \begin{bmatrix} \sigma_{j,1}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{j,d}^2 \end{bmatrix}. \tag{5.6}$$

また，$\phi(\boldsymbol{x};\ \boldsymbol{\mu}, \Sigma)$ は式 (2.6) で導入した平均 $\boldsymbol{\mu}$, 共分散行列 Σ の正規分布である．K を十分大きくとれば，任意の確率分布を必要な精度で近似することができる．実際には K を大きくとると，計算量が増加するので，計算量と精度との兼ね合いから、モデルの設計者が適切に決める必要がある．

以下では，上記の手続きで近似したモデルを Type I モデルと呼ぶ．

5.3.2 Type II モデルの場合

Type II モデルで，アルゴリズムを閉じた形で導くためには，各軸の分散が等しい 2 次元正規分布 (等方正規分布) の和で近似する必要がある．

$$\hat{f}_{\mathrm{II}}(\boldsymbol{x}) = \sum_{k=1}^{K} \xi_k \phi(\boldsymbol{x};\ \boldsymbol{\mu}_k, \sigma_k^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}). \tag{5.7}$$

$\hat{f}_{\mathrm{II}}$ を用いても任意の確率分布 f を任意の精度で近似することはできるが，実際上 $\hat{f}_{\mathrm{I}}$ に比べて必要な精度を達成するために必要となる K の数は増える可能性がある．

最後に，Type I モデルおよび Type II モデルの特徴を表 5.1 にまとめる．これ以上に制限を緩めると，EM アルゴリズムが閉じた形では求まらなくなってしまう．

以下では，上記の手続きで近似したモデルを Type II モデルと呼ぶ．

5.4 学習アルゴリズム

前節で定義した Type I および Type II のモデルに対して 3.4.6 で導入した ECM アルゴリズムを適用する．これらのモデルに含まれる A, H, $\boldsymbol{b}$ といったパラメータは

表 5.1: Type I モデルと Type II モデルの特徴 (尺度に関しては Type I モデルの方が自由度が高い)

	Type I	Type II
位置/尺度	◎	○
回転	×	○
次元	任意	2
近似する混合モデルの要素	楕円型正規分布	等方正規分布

混合分布を構成するすべての正規分布に非線形な形で含まれており，そのままでは簡単な形で推定することは困難である．ここでは，尺度・回転パラメータと位置パラメータをそれぞれ順番に最適化することによって，閉じた形のアルゴリズムを導く．E ステップは 3.4.3 で説明した独立なサンプルが与えられたときの混合分布に対する EM アルゴリズムがそのまま適用され，M ステップでは交互最適化を行う．

5.4.1 Type I モデルの場合

モデルを構成するどの正規分布からデータが発生したかという値 k を隠れた変数としたときに，$(\boldsymbol{x}, k)$ の同時分布は

$$p_{\mathrm{I}}(\boldsymbol{x}, k;\ A, \boldsymbol{b}) = \frac{|A|\xi_k}{(2\pi)^{d/2}\sqrt{|\Sigma_k|}} \exp\left\{-\frac{1}{2}\left\|A\boldsymbol{x} + \boldsymbol{b} - \boldsymbol{\mu}_k\right\|^2_{\Sigma_k^{-1}}\right\}, \tag{5.8}$$

となる．まず，式 (3.29) に対応する次の条件つき確率を定義する．

$$\begin{aligned} q^{(t)}_{\mathrm{I}\,k}(\boldsymbol{x}) &= p_{\mathrm{I}}(k \mid \boldsymbol{x};\ A^{(t)}, \boldsymbol{b}^{(t)}) \\ &= \frac{p_{\mathrm{I}}(\boldsymbol{x}, k;\ A^{(t)}, \boldsymbol{b}^{(t)})}{\sum_{k'=1}^{K} p_{\mathrm{I}}(\boldsymbol{x}, k';\ A^{(t)}, \boldsymbol{b}^{(t)})}. \end{aligned} \tag{5.9}$$

この値は，各繰り返しステップでのパラメータの値から計算可能なものである．

さて，この場合 M ステップの最適化問題は陽に解くことができない．そこで ECM アルゴリズムを採用し，A と $\boldsymbol{b}$ のそれぞれを順番に最適化することにする．ECM アルゴリズムの各ステップの導出の詳細は 付録 B にまとめることにし，ここでは結果だけを述べる．

まず，$\boldsymbol{b} = \boldsymbol{b}^{(t)}$ に固定して, $A^{(t+1)}$ を最適化すると，

$$a_i^{(t+1)} = \frac{\sqrt{(Y_{\mathrm{I}\,i})^2 + 4X_{\mathrm{I}\,i}Z_{\mathrm{I}}} - Y_{\mathrm{I}\,i}}{2X_{\mathrm{I}\,i}}, \tag{5.10}$$

が得られる．ここで $X_{\mathrm{I}\,i}, Y_{\mathrm{I}\,i}, Z_{\mathrm{I}}$ はそれぞれ次のように計算される統計量である．

$$X_{\mathrm{I}\,i} = \left\langle x_i^2 \sum_{k=1}^{K} \frac{q_{\mathrm{I}\,k}^{(t)}(\boldsymbol{x})}{\sigma_{k,i}^2} \right\rangle_{\rho}, \tag{5.11}$$

$$Y_{\mathrm{I}\,i} = \left\langle x_i \sum_{k=1}^{K} \frac{(b_i^{(t)} - \mu_{k,i}) q_{\mathrm{I}\,k}^{(t)}(\boldsymbol{x})}{\sigma_{k,i}^2} \right\rangle_{\rho}, \tag{5.12}$$

$$Z_{\mathrm{I}} = \left\langle \sum_{k=1}^{K} q_{\mathrm{I}\,k}^{(t)}(\boldsymbol{x}) \right\rangle_{\rho}. \tag{5.13}$$

次に，得られた $A^{(t+1)}$ を用いて, $\boldsymbol{b}^{(t+1)}$ を求める．

$$b_i^{(t+1)} = \frac{U_{\mathrm{I}\,i}}{V_{\mathrm{I}\,i}}. \tag{5.14}$$

ここで $U_{\mathrm{I}\,i}, V_{\mathrm{I}\,i}$ は次のように計算される統計量である．

$$U_{\mathrm{I}\,i} = \left\langle \sum_{k=1}^{K} \frac{(\mu_{k,i} - a_i^{(t+1)} x_i) q_{\mathrm{I}\,k}^{(t)}(\boldsymbol{x})}{\sigma_{k,i}^2} \right\rangle_{\rho}, \tag{5.15}$$

$$V_{\mathrm{I}\,i} = \left\langle \sum_{k=1}^{K} \frac{q_{\mathrm{I}\,k}^{(t)}(\boldsymbol{x})}{\sigma_{k,i}^2} \right\rangle_{\rho}. \tag{5.16}$$

ここでは最初に $\boldsymbol{b}$ を固定したが， A を先にしてもよい．

5.4.2 Type II モデルの場合

Type II モデルでも同様に k を導入すると，$(\boldsymbol{x}, k)$ の同時分布は

$$\begin{aligned} p_{\mathrm{II}}(\boldsymbol{x}, k;\ H, \boldsymbol{b}) \quad = \quad & \frac{(h_1^2 + h_2^2)\xi_k}{(2\pi\sigma_k^2)} \exp\Big[-\frac{1}{2\sigma_k^2} \Big\{ (h_1 x_1 + h_2 x_2 - \mu_{k,1} + b_1)^2 \\ & + (-h_2 x_1 + h_1 x_2 - \mu_{k,2} + b_2)^2 \Big\} \Big], \end{aligned} \tag{5.17}$$

とあらわされる．

EM アルゴリズムを行うために必要な条件付き確率は

$$\begin{aligned} q^{(t)}_{\mathrm{II}\,k}(\boldsymbol{x}) &= p_{\mathrm{II}}(k \mid \boldsymbol{x};\ H^{(t)}, \boldsymbol{b}^{(t)}) \\ &= \frac{p_{\mathrm{II}}(\boldsymbol{x}, k;\ H^{(t)}, \boldsymbol{b}^{(t)})}{\sum_{k'=1}^{K} p_{\mathrm{II}}(\boldsymbol{x}, k';\ H^{(t)}, \boldsymbol{b}^{(t)})}, \end{aligned} \tag{5.18}$$

で計算することができる．

Type II の場合も H と $\boldsymbol{b}$ を同時に最適化することは困難であるので，まず，$\boldsymbol{b} = \boldsymbol{b}^{(t)}$ を固定し，$H^{(t+1)}$ を求める． すると，

$$h_1^{(t+1)} = \frac{Y_{\mathrm{II}\,1}\left(1 + \sqrt{1 + \dfrac{8 Z_{\mathrm{II}} X_{\mathrm{II}}}{Y_{\mathrm{II}\,1}^2 + Y_{\mathrm{II}\,2}^2}}\right)}{2 X_{\mathrm{II}}}, \tag{5.19}$$

$$h_2^{(t+1)} = \frac{Y_{\mathrm{II}\,2}}{Y_{\mathrm{II}\,1}} h_1^{(t+1)}, \tag{5.20}$$

が得られる．

ここで $X_{\mathrm{II}}, Y_{\mathrm{II}\,1}, Y_{\mathrm{II}\,2}, Z_{\mathrm{II}}$ はそれぞれ，次のように計算される統計量である．

$$X_{\mathrm{II}} = \left\langle (x_1^2 + x_2^2) \sum_{k=1}^{K} \frac{q^{(t)}_{\mathrm{II}\,k}(\boldsymbol{x})}{\sigma_k^2} \right\rangle_{\rho}, \tag{5.21}$$

$$Y_{\mathrm{II}\,1} = \left\langle \sum_{k=1}^{K} \frac{q^{(t)}_{\mathrm{II}\,k}(\boldsymbol{x})}{\sigma_k^2} \left\{ (\mu_{k,1} - b_1^{(t)}) x_1 + (\mu_{k,2} - b_2^{(t)}) x_2 \right\} \right\rangle_{\rho}, \tag{5.22}$$

$$Y_{\mathrm{II}\,2} = \left\langle \sum_{k=1}^{K} \frac{q^{(t)}_{\mathrm{II}\,k}(\boldsymbol{x})}{\sigma_k^2} \left\{ (\mu_{k,1} - b_1^{(t)}) x_2 - (\mu_{k,2} - b_2^{(t)}) x_1 \right\} \right\rangle_{\rho}, \tag{5.23}$$

$$Z_{\mathrm{II}} = \left\langle \sum_{k=1}^{K} q^{(t)}_{\mathrm{II}\,k}(\boldsymbol{x}) \right\rangle_{\rho}. \tag{5.24}$$

次に $H = H^{(t+1)}$ を固定して $\boldsymbol{b}^{(t+1)}$ を求める．

$$b_1^{(t+1)} = \frac{U_{\mathrm{II}\,1}}{V_{\mathrm{II}}}, \tag{5.25}$$

$$b_2^{(t+1)} = \frac{U_{\mathrm{II}\,2}}{V_{\mathrm{II}}}. \tag{5.26}$$

ここで $U_{\mathrm{II}\,1}, U_{\mathrm{II}\,2}, V_{\mathrm{II}}$ はそれぞれ次のように計算される統計量である．

$$U_{\mathrm{II}\,1} = \left\langle \sum_{k=1}^{K} \frac{q^{(t)}_{\mathrm{II}\,k}(\boldsymbol{x})}{\sigma_k^2} \left\{ \mu_{k,1} - (h_1^{(t+1)} x_1 + h_2^{(t+1)} x_2) \right\} \right\rangle_{\rho}, \tag{5.27}$$

$$U_{\mathrm{II}\,2} = \left\langle \sum_{k=1}^{K} \frac{q_{\mathrm{II}\,k}^{(t)}(\boldsymbol{x})}{\sigma_k^2} \left\{ \mu_{k,2} - (h_1^{(t+1)} x_2 - h_2^{(t+1)} x_1) \right\} \right\rangle_{\rho}, \tag{5.28}$$

$$V_{\mathrm{II}} = \left\langle \sum_{k=1}^{K} \frac{q_{\mathrm{II}\,k}^{(t)}(\boldsymbol{x})}{\sigma_k^2} \right\rangle_{\rho}. \tag{5.29}$$

5.5 実験

本論文で示したアルゴリズムが動作することを確認するための実験を人工データ，および，単純な実画像に対して行なった.

5.5.1 人工データ

Type I，Type II のそれぞれのモデルに対し，まず最初に単純に単独のモデルを用いて推定がうまく行えるかどうかを調べた． 続いて、複数個のモデル (ここでは 2 個の物体モデルとばらまきノイズの計 3 個のモデル) をあてはめる場合に提案手法がどのように振る舞うかを見るための実験を行った.

単独，複数それぞれの場合について，まず適当なモデルを人工的に作り，そのモデルから 1000 個のランダムサンプルを発生させる．そのサンプルをあてはめの対象データとし，サンプルを発生させたモデルとは異なる位置・尺度 (・回転) パラメータをもつモデルを初期解として本論文で述べたアルゴリズムを適用した.

Type I モデルの実験には 3 つの 2 次元楕円型正規分布の混合分布で定義されるモデルを用いた.

図 5.1 は Type I モデルを単独で用いた場合の結果である．点で示されているのがあてはめの対象データであり，破線で示されているのが初期解，実線で示されているのが ECM アルゴリズムを 10 ステップ繰り返した結果である．それぞれの楕円はモデルを構成する正規分布の $x^2/\sigma_x^2 + y^2/\sigma_y^2 = 2.0^2$ なる確率密度等高線であり，この楕円内の確率測度は約 0.86 である.

図 5.2 は 10 ステップまでの平均対数尤度の変化をプロットしたグラフで，破線がサンプルを発生させた分布 (真の分布) に関する平均対数尤度である． 有限サンプルによるバイアスを除けば，理想的にはアルゴリズムにより破線に近付く程望ましいことになる.

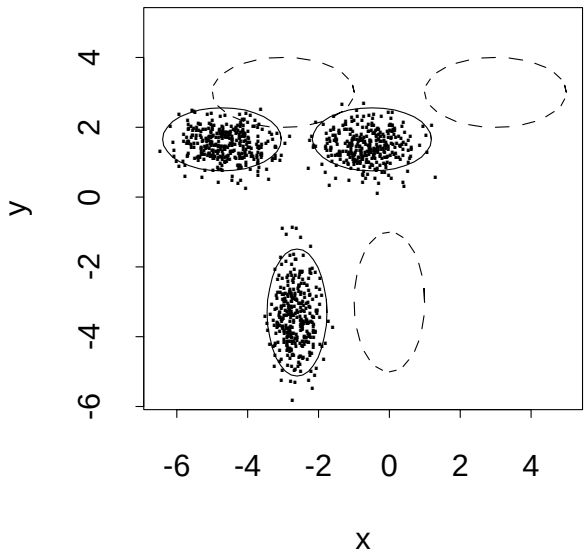

図 5.1: Type I モデル (単独) の実験結果

図 5.3 は単独で用いたのと同じモデルを 2 個組み合わせて (位置・尺度パラメータは異なる) モデルを作成し，ECM アルゴリズムを 20 ステップ繰り返した結果である． ノイズに対する影響を見るために，全体に一様分布によりばらまきノイズを加えた． モデルはそのノイズを吸収するための要素分布として，サンプルをすべて含む矩形上の一様分布を加えた．図 5.4 は平均対数尤度のプロットである．

Type II モデルの実験には 4 つの 2 次元等方正規分布の混合分布で定義されるモデルを用いた．

図 5.5，5.6 はこのモデルを単独であてはめたとき (10 ステップ) の結果および平均対数尤度のプロットであり，図 5.7，5.8 は複数 (2 個) の Type II モデルとばらまきノイズに対してあてはめたとき (30 ステップ) の結果および平均対数尤度のプロットである．

Type I モデルよりも Type II モデルの方が収束に時間がかかっており，複数物体の実験では単独であてはめるよりも時間がかかっている．

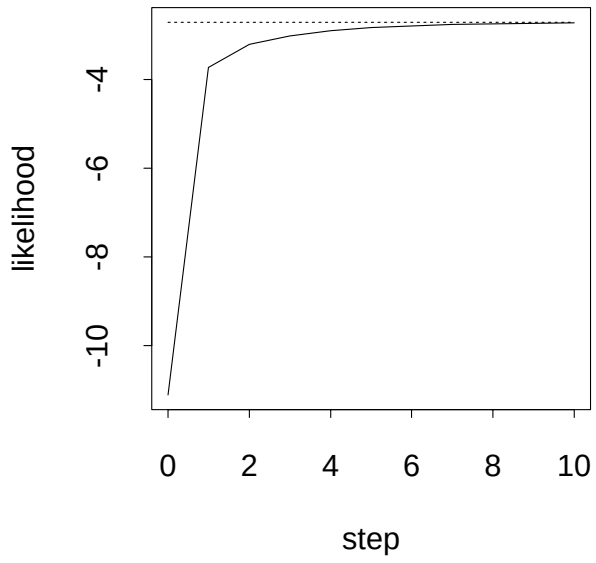

図 5.2: Type I モデルの平均対数尤度 (単独，点線:真のモデルの尤度)

5.5.2 実画像

ここでは，簡単な実画像データについて Type II モデルを用いた複数物体認識を行った結果を述べる．Type II モデルを用いた理由は，2 次元画像上の物体という対象の特性と，Type I モデルよりも推定が困難であることからより厳しい評価を行えると考えたからである．

画像取り込み

まず，図 5.9 の対象画像および，その構成要素のうち，物体 1(ステープラー) と物体 2(テープホルダー) をそれぞれ別々にデジタルカメラで撮影し (背景は同じ薄い灰色)，フルカラーで計算機に取り込んだ．物体画像については，物体を含んでいる部分の矩形領域の切り出しを行った．また，画像中の物体の色のついている部分はすべて青い色である．もう一つの構成要素 (ペン) はあてはめにおけるノイズとしての役割を果たす．

次に，対象画像と二つの物体画像をそれぞれ，128×96，64×32, 64×41 の大きさ

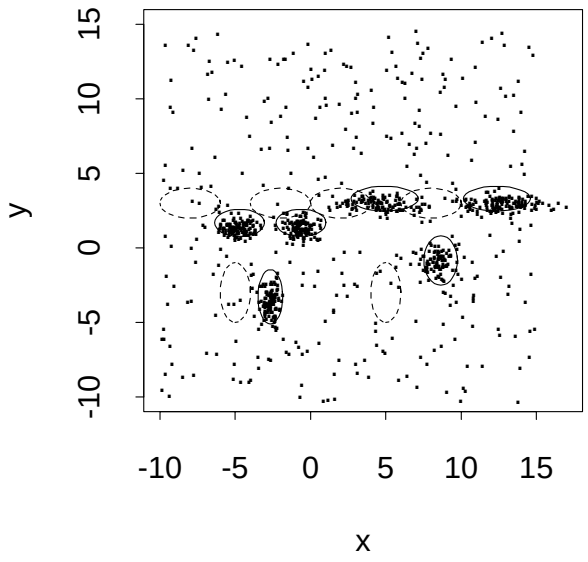

図 5.3: Type I モデル (複数) の実験結果

に解像度を落した.

特徴抽出

特徴抽出手法として，単純な色判別フィルタによる方法を行った．以下に具体的手順を示す.

対象画像の画素値データ (RGB 3 次元 × 画素数) を一つのクラス A とし，2 個の物体画像の画素値データをもう一つのクラス B として，2 値化のための判別分析を行った． 物体画像中，クラス B に判別された画素点集合を，本論文で提案するアルゴリズムを適用するための物体モデルを求めるのに用いた．それぞれ図 5.10, 5.11 上に点で示す.

また，対象画像中で，クラス B に判別された画素点集合を，アルゴリズムを適用する特徴点とした． それを図 5.12 上に点で示す.

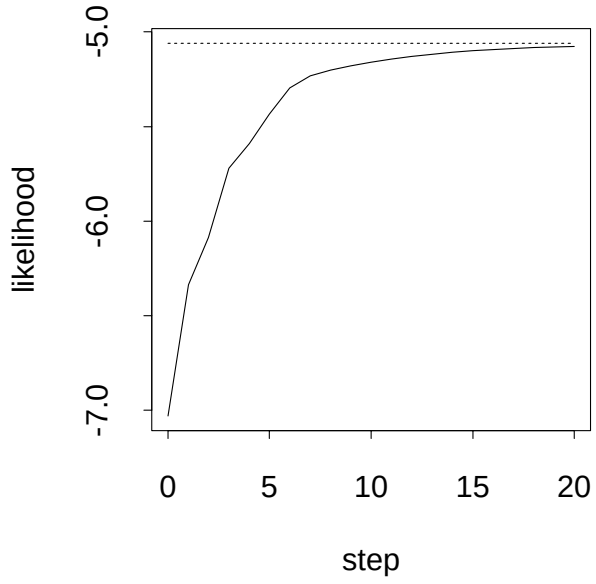

図 5.4: Type I モデルの平均対数尤度 (複数，点線:真のモデルの尤度)

物体モデルの学習

まず，等方的な正規分布の混合モデルを学習させ，物体モデルとした． 正規分布の個数は近似精度と計算量から勘案して 8 個とした．その結果を図 5.10, 5.11 上の点線の円で示す．学習は確実に収束させるため 30 ステップ繰り返した．

実験結果

作成した 2 個の物体モデルとノイズのための一様分布との 3 個の分布の混合モデルを対象画像から抽出した特徴点に対してアルゴリズムを適用した．

図 5.12 に初期解を示す．100 ステップ程度で概ね収束が完了し，図 5.13 に示されるような解に収束した．対数尤度は図 5.14 のように変化した．

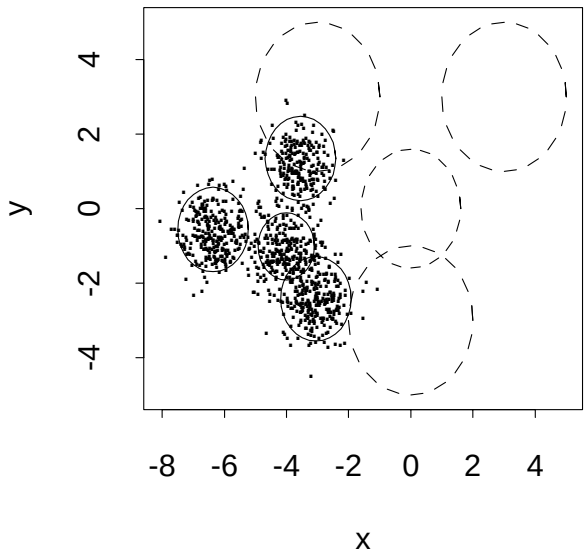

図 5.5: Type II モデル (単独) の実験結果

5.6 考察

提案手法は EM アルゴリズムに基づく手法であるため，実験においても初期解によっては (最適ではない) 局所解に収束することがあった．定性的には，Type I よりも Type II の方が，局所解に収束しやすく，複数モデルのあてはめの場合には，初期解とデータ点との重なりが少ない程，また，構成するモデルの数が増える程，真の解に収束する比率が下がる傾向にある．また，実画像においては 2 個の異なる物体モデルをあてはめたが，初期値として，それぞれの対象に誤った物体モデルを置いてしまうと，その間違った方に収束してしまう傾向がある (これは一面では欠損値や隠れに対する強さをも示すものである)．それらの問題に比べればノイズなどに対する影響は，比較的少なかった．

さて，本論文で提案したアルゴリズムの計算量は，EM アルゴリズムの 1 ステップあたり $O(dNK)$ で与えられる (単独モデルをあてはめる場合)．条件付き確率の計算が計算時間の主要な部分を占めている．より精度の高い推定を行なおうとすれば特徴点の数 N を大きくする必要があり，複雑なモデルを精度良くあてはめよ

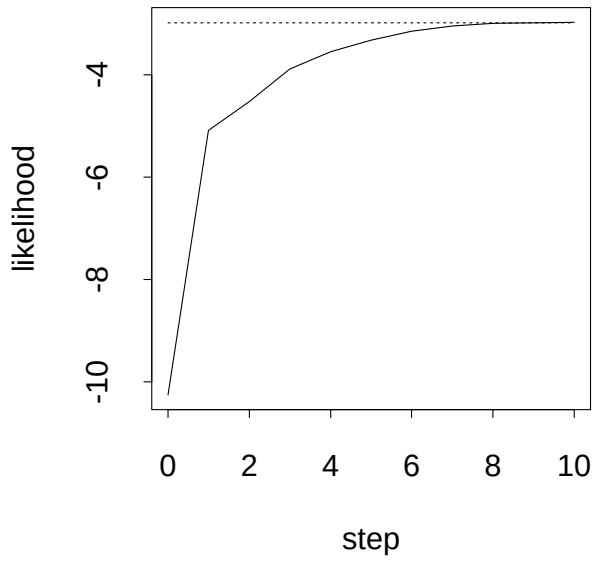

図 5.6: Type II モデルの平均対数尤度 (単独，点線:真のモデルの尤度)

うとすると K が大きくなる．しかしながら，初期解ではもともとあてはめの精度はよくないのだから，EM アルゴリズムの初期のステップではサンプルを間引いたり，モデルを少ない正規分布でおおまかに近似し，近似精度がよくなってきてからモデルを詳細化することによって全体の計算量を減らすことができる．また，モデルを複雑にする程局所解の数も増えると考えられるから，その観点からもそうした段階的詳細化は意味がある．

また，実験結果は，多くの場合，最初の数ステップでかなり収束し，最適解の近くでは非常に遅くなるという傾向を示している． これは，EM アルゴリズムは一次収束のアルゴリズムであるが，二次収束アルゴリズムの近似にもなっているという事実に合致するものである．最適解の近くでの収束性を改善させるためには，ある程度収束した時点で二次収束のアルゴリズムを併用することも考えられる．

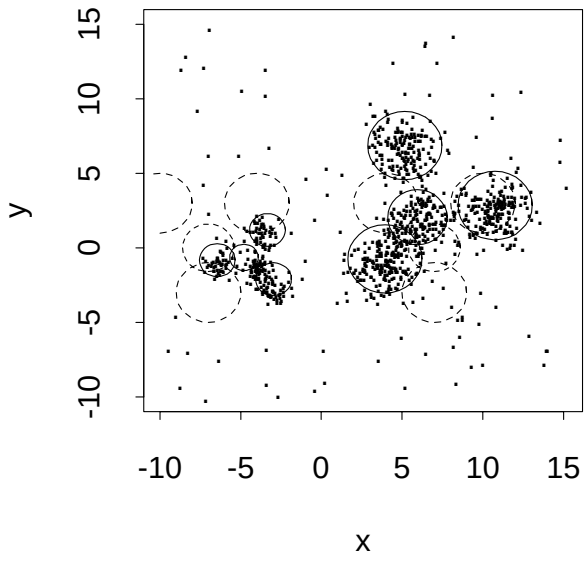

図 5.7: Type II モデル (複数) の実験結果

5.7 本章のまとめ

複雑な形の確率分布の位置・尺度 (・回転) パラメータの ECM アルゴリズムに基づく推定法を提案した．一般に, 独立でないパラメータをもつ混合分布では EM アルゴリズムの各ステップが複雑になるが，限定された形の正規混合分布を用いることと，ECM アルゴリズムの採用によって，2 次方程式を解く問題に帰着され，閉じた形のアルゴリズムが導かれた．

EM アルゴリズムを基礎としているため，比較的高速で安定した解が得られるという特徴をもつ一方で，局所最適解への収束，複数モデルのあてはめにおいて個数が決まっている必要があることなどの問題点も持っており，適用に際してはこれらの点に注意する必要がある．

応用面では，物体認識を想定した実験を行ったが，はじめにも述べたように分布の位置や尺度パラメータの推定は一般的な問題であり，物体認識以外の用途も今後探索していく必要がある．物体認識についても，手書き文字認識や 3D 物体認識では，位置や尺度，回転以外の変形を想定する必要がある．本論文ではそれを統計的

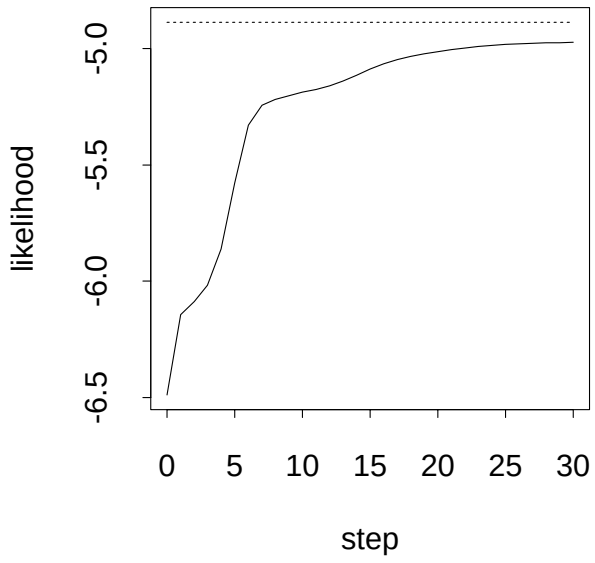

図 5.8: Type II モデルの平均対数尤度 (複数，点線:真のモデルの尤度)

なゆらぎとして扱うことによって無視することによって安定なアルゴリズムを得ることができた．一方陽にそれらのゆらぎを損失関数に入れたアルゴリズムも提案されているが [70]，勾配法を用いているので，安定性では本論文での手法に劣ると考えられる．

また，本論文では説明しなかったが，観測領域が限られていて観測できないデータがある場合でも，観測領域が軸に平行な矩形であれば，はみ出している部分のデータを欠測値とした EM アルゴリズムを閉じた形で適用することができることが著者によって示されている [1, 2].

図 5.9: 対象画像 (実際はカラー画像)

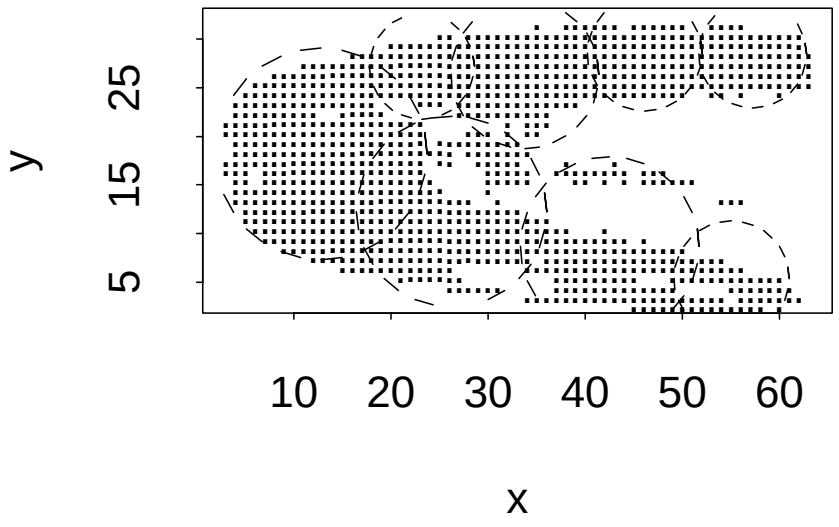

図 5.10: 物体画像 1 (ステープラー) の 2 値化データと，作成した Type II モデル (8 個の正規分布の混合分布)

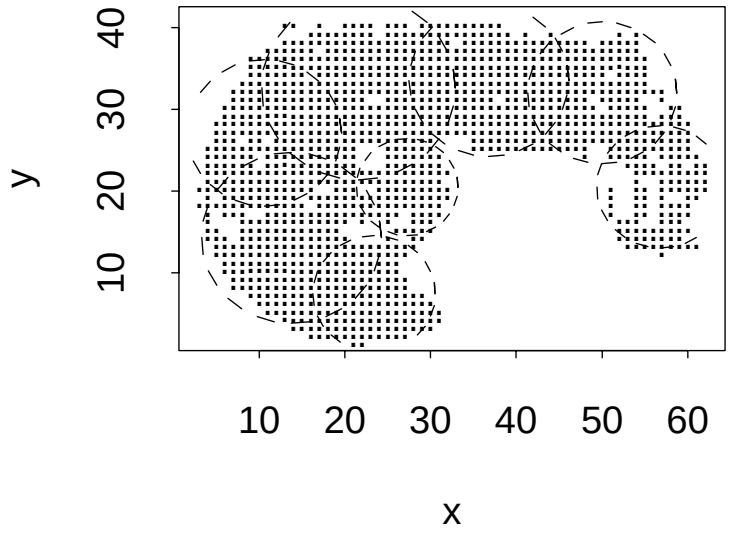

図 5.11: 物体画像 2 (テープホルダー) の 2 値化データと，作成した Type II モデル (8 個の正規分布の混合分布)

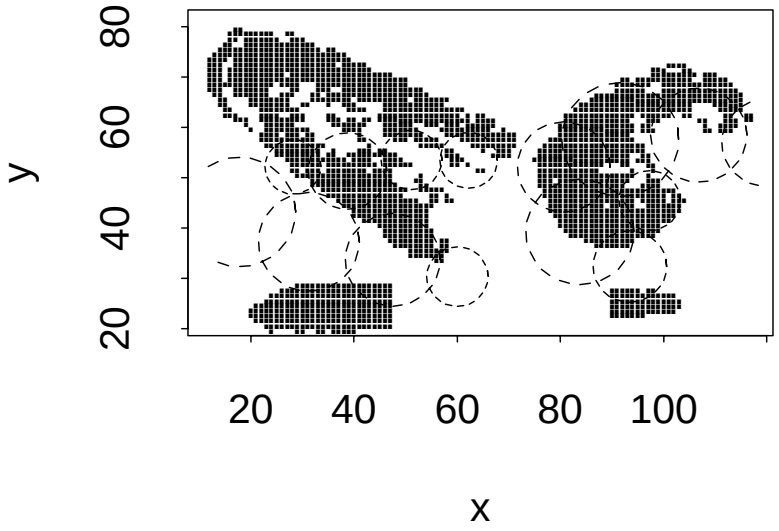

図 5.12: 対象画像の 2 値化データと，EM アルゴリズムの初期解

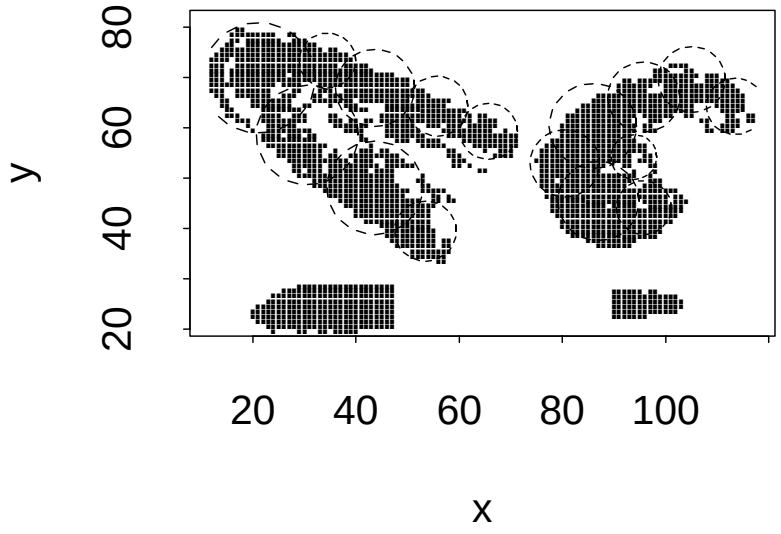

図 5.13: 実画像データに対する Type II モデルの 100 ステップ後の解

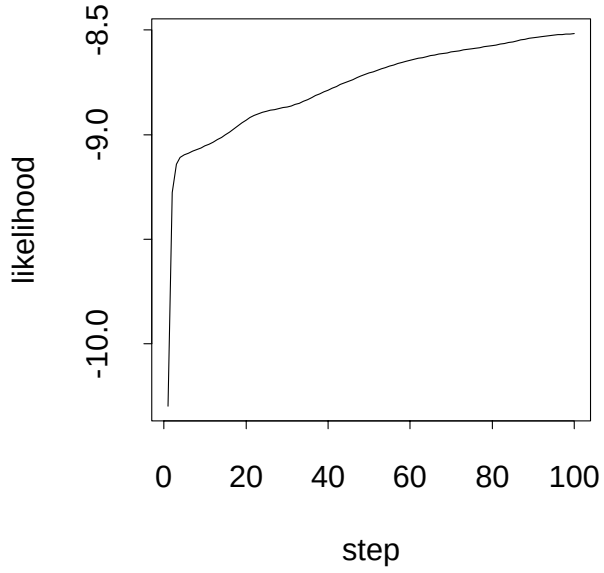

図 5.14: 実画像データに対する Type II モデルの対数尤度

第6章　複数情報源からの属性概念獲得

6.1　はじめに

自然なヒューマンインタフェースを実現するための技術として，多数のチャネルを用意して柔軟なコミュニケーションを図るというマルチモーダル対話の研究が盛んになってきている [48, 38, 9].

柔軟なシステムに必要とされる条件の一つとして，学習機能を持っていることが挙げられる． 実世界の対象は非常に多様であるので，実世界に関する知識をすべてシステムに埋め込んでおくことは不可能である．そこで，システムが人間との対話を通じて環境に関する知識を学習することが必要となる．

従来，対話システムにおける学習の重要性は認識されつつも，対象の多様性ゆえに学習自体が困難であったために，実現された例は少ない．また，非音声的な対話システムでは，記号的な情報はパターン的な情報に隠れていることが多く，従来の人工知能的なアプローチでは扱いにくい面があった．そこで，本章では画像と音声といった複数の情報源からのデータをもとにシステムに属性概念を学習させるという課題について統計的なアプローチから考察する．

まず準備として，物体の画像をシステムに見せて，その名称を音声で教えるという課題を考えよう． これは二つのパターンの間の直接的な (一対一の) 関係づけである． これに対して例えば次のような方法が考えられる．

さまざまな物体に対して，物体の画像とその物体の名称を表す発話の対を与えるということを，一定量の学習サンプルについて繰り返し行い，画像と音声のパターン間の関係付けを学習させておく． 次に，新しい画像をシステムに示して，その名称をたずねると，それまで学習した関係付けの結果にもとづいて，その画像にもっとも関連の高い音声パターンを再生することによって，その名称を答える． またある物体の名称を音声でたずねると，その物体の画像を想起して見せてくれる．

この方法で注目すべき点は，画像や音声に対する明示的なパターン認識を行っていないことである．現実のデータでは，記号的な情報がパターン情報に隠れていて陽には与えられないことが多い．したがって，明示的認識を行わない学習では，あらかじめカテゴリーを定めることが難しいような対象に対しても，システムと対話する人間の発話を利用することによって，明示的なカテゴリーの定義を避けることが可能となる．またシステムに対して提示するものが，画像と音声のパターンの対だけなので，個々のパターンに対して記号化したラベル付けをして学習させる場合に比べて，その手間が軽減される．

本論文では，一対一の対応でなく，一対多の対応の例として，物体の属性概念を学習させることについて考察する．物体には，名称のほかに，色や大きさといった複数の属性が存在するので，画像に対する複数の属性を自動的に分類して獲得するという問題が考えられる．すなわち，物体のある画像に対して，複数の属性のうちの１つ (例えばその物体の色) を人間が音声で教えるということを続け，システムに学習サンプルを自動的に分類させる．

学習の目的は，新たに示された物体の画像に対してその物体の複数の属性を，属性を表す音声の組によって答えることである．表 6.1 に，その例を示す．

画像のある属性を教えているときに，どの属性について教示しているかの情報がシステムには与えられないため，通常の個別な属性についてのパターン認識よりも難しい問題になる．この場合，属性概念という記号的な情報は個々のパターンだけからは抽出不可能で，複数のモダリティの情報を統合し，パターン集合全体に埋め込まれた構造を抽出する必要があるため，より複雑な学習課題となる．

この課題はまた，人間の発達に関する発達心理学の観点からも興味ある課題である． 人間の子供は，大人とのコミュニケーションを通じて同様の課題を解いている．実際にはもっと複雑で，発達途中で属性概念そのものを教えられたりするが，それ以前に単純な画像と音声のペアだけの教示段階でもある程度自己組織的に属性概念を形成できることが知られている [32].

統計的には，属性概念の学習は音声側のデータに対するクラスタリングである．しかしながら，そのクラスタ化の基準はペアとして与えられる画像によって規定される点で通常のクラスタリングとは異なる一種の制約付きのクラスタリングとみなせる．また，画像や音声はそれぞれ多変量であるから，それらから共通に含まれ

表 6.1: 考える課題の例

学習データの例:

画像 X	音声 Y
[白いコップ]	“しろい”
[青いペン]	“ぺん”
...	...
[赤い本]	“あかい”

学習後:

画像 X		音声 Y
[白いペン]	→	“しろい” + “ぺん”
[青い本]	→	“あおい” + “ほん”
...		...
[赤いコップ]	→	“あかい” + “こっぷ”

る特徴量をうまく取り出して低次元化することが重要となる．そこで本論文では，これらの問題に対して，正準相関分析による次元圧縮と混合分布による属性のモデル化を組み合わせた学習法を提案し，実験によって有効性を検討する．

視聴覚情報の統合化にもとづく概念獲得の研究としては，参考文献 [59, 60] がある． しかしながら，これらの研究では画像の属性についてあらかじめ記号化されており，また DP マッチングによる音声同士の類似区間の抽出や文法レベルの学習に関するものであり，本論文の対象とは，異なるものである．

6.2 複数情報源からの属性概念獲得

前節で述べた問題を以下のように定式化する．

二つの情報源 X, Y からそれぞれデータ $\boldsymbol{x}, \boldsymbol{y}$ が与えられるとしよう． 我々の問題では $\boldsymbol{x}$ は画像，$\boldsymbol{y}$ は音声である．$\boldsymbol{x}$ は K 種類の属性をもっており，$\boldsymbol{y}$ はその属

性のいずれかを表すものとする．属性の数 K は既知とするが，$\boldsymbol{y}$ が $\boldsymbol{x}$ のどの属性を表わしているかは未知とする．

後で詳しく述べるように，属性としては色，形，大きさの 3 種類 ($K = 3$)，あるいはそのうちの 2 種類 ($K = 2$) を考える．

与えられた学習サンプルをもとに属性の分類を行って $\boldsymbol{x}$ と $\boldsymbol{y}$ との対応を学習し，新たに与えられた $\boldsymbol{x}$ に対する K 個の属性 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$ を出力することが目的である．

6.3 正準相関分析による次元圧縮

$\boldsymbol{x}$ と $\boldsymbol{y}$ の対応を学習するにはそれぞれの次元が大きすぎる (音声の場合には音声の長さとともに次元の大きさも変化してしまう) ので，その前処理として次元を減らす処理をする必要がある．まず第一にすべきことはそれぞれ生データをある固定した次元の特徴量に変換することである．どのような特徴量をとればよいかという問題は情報源の種類に大きく依存するのでこれに関しては実験の学習データの節で述べる．

特徴量に変換された後もそれぞれのデータはまだかなり大きい次元であることが普通である．大きい次元同士の対応関係をとるのは汎化や計算量の観点から難しい．また，特徴量はそれぞれの情報源ごとに独立して設計されたものであり，情報源の間の対応を考える際には不要な情報も多く含まれている．

そのために正準相関分析 [68, 18, 92] によって，複数の情報源の対応づけに必要な情報だけを取り出す．正準相関分析とは，$\boldsymbol{x}$ と $\boldsymbol{y}$ を最も相関係数が大きくなるようにそれぞれを共通の空間 (正準空間) に写像する手法である．一般にこの条件だけではスケールや回転の自由度が存在し，写像が一意に定まらないので移された空間での分散共分散行列が単位行列になるように決める．また，今回の実験では写像としては最も簡単な線形変換を採用したが，非線形への拡張等も研究されている [20]．線形変換の場合は，変換行列が固有値問題の解で与えられる．ここではその具体的な形を示しておく．

一般性を失うことなく $\boldsymbol{x}$ および $\boldsymbol{y}$ のサンプル平均は 0 であるとする．そうでないときは各サンプルから平均ベクトルを引いておけばよい．このとき，$\boldsymbol{x}$ および $\boldsymbol{y}$

から正準空間の第 i 成分への射影

$$u_i = \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{x}, \qquad v_i = \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{y}, \tag{6.1}$$

は，次の一般化固有値問題の (大きい方から数えて) 第 i 番目の固有値に対応する固有ベクトルとして与えられる．

$$V_{xy} V_{yy}^{-1} V_{yx} \boldsymbol{a}_i = \rho_i^2 V_{xx} \boldsymbol{a}_i, \tag{6.2}$$

$$V_{yx} V_{xx}^{-1} V_{xy} \boldsymbol{b}_i = \rho_i^2 V_{yy} \boldsymbol{b}_i, \tag{6.3}$$

ただし，V_{zw} は $\boldsymbol{z}$ と $\boldsymbol{w}$ の共分散行列である．また，ρ_i は u_i と v_i の相関係数になっている．ちなみに，正準相関分析は情報源の結合正規性を仮定したとき，つまり $\boldsymbol{x}$ と $\boldsymbol{y}$ が同時正規分布しているとすると，最も相互情報量が高くなるような特徴量を抽出していることになっている．

6.4 混合分布と EM アルゴリズム

ここでは，前節の処理によって低次元化されたデータに対して一対多の対応を学習するための方式について考える． なお混乱がないと思われるので，$\boldsymbol{x}$，$\boldsymbol{y}$ を低次元化して得られたベクトルを以下でもやはり $\boldsymbol{x}, \boldsymbol{y}$ と書くことにする．

基本的には 3.4.3 で述べた独立な要素分布をもつ場合の条件付き混合分布によるモデル化を EM アルゴリズムで学習することになるが，本章で扱う問題に依存した初期化法によって収束性の改善を図る．

6.4.1 混合分布によるモデル化

$\boldsymbol{x}$ には K 個の属性が存在するが，そのうちの k 番目の属性だけに着目しよう．$\boldsymbol{x}$ とその k 番目の属性 $\boldsymbol{y}$ の対応が $f_k(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}_k)$ というパラメトリックな条件付き確率分布でモデル化できたとしよう．属性 k が確率 ξ_k でランダムに選ばれるとすれば $(\boldsymbol{y}, k)$ の分布は $\xi_k f_k(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}_k)$ となる．ただし，我々の問題では，どの属性が教えられているかは教えられない，すなわち k が未知であるのでこれは観測できない隠れた変数である． したがって観測されるデータの分布は (条件付き) 混合

分布モデル

$$f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}), \tag{6.4}$$

となる．f_k の取り方によっていろいろな形になり得るが，本稿ではとりあえず最も簡単なものとして第 2 章の例 2 で取り上げた線形回帰混合モデルを用いる．$\boldsymbol{y}$ の各成分は独立な線形モデルと仮定すると，

$$f_k(\boldsymbol{y} \mid \boldsymbol{x};\ \boldsymbol{\theta}_k;\ k = 1, \ldots, K) = \prod_{i=1}^{d} f(y_i \mid \boldsymbol{x};\ \boldsymbol{a}_{k,i}, b_{k,i}, \sigma_{k,i}^2), \tag{6.5}$$

ただし，

$$f(y \mid \boldsymbol{x};\ \boldsymbol{a}, b, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} - b)^2}{2\sigma^2} \right\}. \tag{6.6}$$

この場合には，EM アルゴリズムの各ステップは以下のように書ける．隠れ変数の条件付き分布を

$$q(k \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}^{(t)}) = \frac{f_k(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}{f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}, \tag{6.7}$$

とおく． ただし $\boldsymbol{\theta}^{(t)}$ は t ステップ目で得たパラメータである．すると $t+1$ ステップ目のパラメータは 3.4.3 で述べたアルゴリズムを条件付き分布に適用することにより，次のようになる．

- 重みの確率は

  $$\xi_k^{(t+1)} = \frac{1}{N} \sum_{(\boldsymbol{x}, \boldsymbol{y})} q(k \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}^{(t)}), \tag{6.8}$$

  となる． ここで N はサンプル数をあらわす．

- 線形写像のパラメータ $\boldsymbol{a}_{k,i}^{(t+1)}, b_{k,i}^{(t+1)}$ は，各サンプル $\boldsymbol{x}, \boldsymbol{y}$ を，$q(k \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}^{(t)})$ で重みをつけた線形回帰分析の解で与えられる．

- 誤差分散 ${\sigma_{k,i}^2}^{(t+1)}$ は各サンプルを $q(k \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}^{(t)})$ で重みをつけた線形回帰の自乗残差で与えられる．

6.4.2 パラメータの初期値

EM アルゴリズムは適当なパラメータの初期値から始めて，局所最適解に収束させるアルゴリズムであり，混合分布のように多数の局所最適解をもつ問題に対しては初期値の定め方が非常に重要となる．

我々の問題の場合，解が満たしている条件として次のようなものが考えられる．

> 同じ (あるいは類似の) $\boldsymbol{x}$ に対して異なる $\boldsymbol{y}$ が存在する場合にはその $\boldsymbol{y}$ は異なる属性である．

この特徴を反映させる方法として，評価基準

$$\sum_{i<j} \frac{(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}_i - \boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}_j)^2}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\| + \epsilon} \tag{6.9}$$

を最大化するような射影軸 $\boldsymbol{c}$ を求めて，その射影された軸上で属性の数だけ分割する． ただしここで i, j はサンプルの番号，ϵ は正の定数．これによって，大雑把ではあるが，類似の $\boldsymbol{x}$ に対して異なる $\boldsymbol{y}$ は別の属性として分類されるようになることが期待できる．

更に，この最大化問題は $\|\boldsymbol{c}\| = 1$ という条件のもとで固有値問題として陽に解ける． また，分母は任意の非線形な関数でも構わない．

上の式はサンプル数の自乗の項数を含むので，本稿での実験ではそのうちの適当な個数をランダムサンプリングで項数を減らして近似する (本稿で述べる実験ではサンプル数の 10 倍 (< サンプル数の自乗) の個数にした)．また，分割するといっても，非常に粗い解でしかないので分割された結果は 0 か 1 に割り振るのではなく，中間的な値で重みづけする．今回の実験では，属性数が 2 のときには，一方の属性のクラスには 0.6 で属し，他方のクラスには 0.4 で属するとした (3 属性のときは重みを 0.4, 0.3, 0.3 に配分した)．

6.5 収集データと特徴抽出

以上述べてきた学習を実際の画像と音声に対して実験を行った．ここでは収集したデータおよび特徴抽出法について説明する． 基本的には色のついた数字のブロックをカメラで撮影し，属性に関する発話をマイクから録音した．まず，画像と音声を別々に採集し，学習アルゴリズムのシミュレーションの際に組合せて学習セットを作ることにした．

特徴抽出は情報源の種類に依存してしまうが，本論文の枠組みでは，次元の大きさがそろっていることと，次元ができるだけ低く抑えられつつ必要な情報は落ちていないことが重要となる．

6.5.1 画像データ

本研究で使用した画像は，照明条件を固定した環境で，320 × 240 ピクセルのフルカラー画像をカメラから直接計算機に取り込むことによって入力した．

画像入力対象は，0 ～ 9 までの 10 種のプラスチック製の数字の模型とし，赤，青，黄，緑の 4 色，計 40 個を用意した．

画像の入力時には，模型の画像中での大きさを 3 段階に変えて撮影し (背景は黒で統一した)，大，中，小の 3 種類のバリエーションを加えた．

よって入力画像のクラス数は 120 (10 種 × 4 色 × 3 文字サイズ) となった．

また各クラスの画像は模型の位置と向きに若干の変化を加えた 29 枚とし，結果的に入力画像の総数は 3480 枚 (120 クラス × 29 枚) となった．

6.5.2 画像特徴量

画像の前処理: 前処理過程では，入力した各画像に対して RGB → YIQ 変換を施し，ここから Y（輝度）成分を捨てて，I 成分と Q 成分の画像を生成した．

次にこの I，Q 画像から，一辺の長さを元の画像の 1/1，1/4，1/16 に間引いた 3 種のピクセルサイズの画像を生成した．

よって前処理過程として，各入力画像から新たに 6 枚の画像が生成された．

入力画像の特徴ベクトルの構成: 前処理過程で生成した 6 枚の画像から 25 次元の局所高次自己相関特徴 [47] を算出し，これをその画像の特徴ベクトルとした．局所高次自己相関特徴は，近傍の画素値の積を画像全体に平均した特徴で，図 6.1 の 25 種類の 3 × 3 のパターンのそれぞれについて 1 で示されている画素の値をかけあわせて平均をとることにより得られる．例えば，No.1 は画像全体の平均画素値になり，No.2 は横のピクセルとの相関値，No.6 は隣り合った 3 つの画素の値の積の平均となる．

本論文での実験では，6 枚の画像より 25 次元ずつ特徴を算出することによって 150 次元 (25 次元 × 3 ピクセルサイズ × 2 色) の特徴ベクトルを構成し，これを元の入力画像に固有の特徴ベクトルとした．

したがって，全クラスの全画像に対しては，150 次元の特徴ベクトルを 3480 個

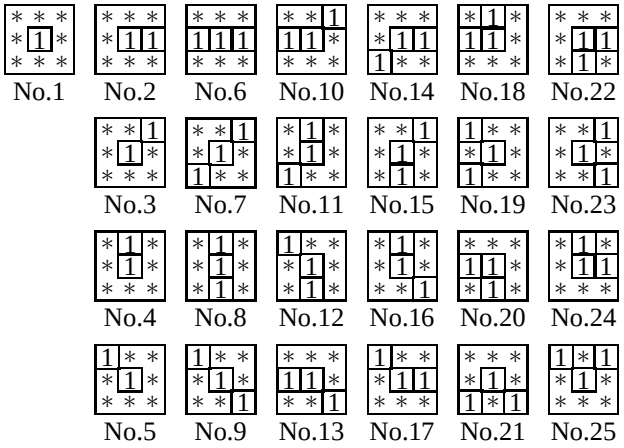

図 6.1: 局所高次自己相関特徴を計算するための 3×3 パターン.

(120 クラス × 29 枚) 算出して利用した.

6.5.3 音声データ

音声は画像の属性に対応する単語を 2 名の男性話者が静かな室内で，のべ 28 回づつ発声したものを録音した.

発音した単語は次の 17 語である.

あか，あお，きいろ，みどり，ぜろ，いち，に，さん，よん，ご，ろく，
なな，はち，きゅう，おおきい，ちゅうくらい，ちいさい

6.5.4 音声特徴量

録音した単語に対して前後の無音区間を切り，音声区間のみに対して，音声認識で一般的に用いられる周波数分析手法である，メルケプストラム分析を行った [67]. 抽出した次数は 1 次から 12 次までの全部で 12 次. 分析条件は 16 kHz サンプリング，分析フレーム長 25msec，フレーム周期 10msec，分析窓: ハミング窓，プリエンファシス係数 0.97 とした.

更に，特徴量の次元を一定にするために分析フレームを間引くことによって時間方向には 10 フレームとなるように正規化を行い，最終的に 120 次元 = 12(次)×10(フ

レーム) の特徴量を得た．つまり，最初にできたベクトル列は 12 ×(分析フレーム数 = n) となっているので，例えば n = 30 のときは第 1, 4, 7, 10, 13, 16, 19, 22, 25, 28 番目のフレームの分析結果だけを用いて，12 × 10 のベクトル列を作成した．

6.6 実験

画像特徴と音声特徴をそれぞれファイル化し以下の計算機実験を行った．

6.6.1 実験の手順

属性の種類が色と形の 2 種類の場合と，大きさも含めた 3 種類の場合について実験を行った．処理の流れは次のとおりである．

学習データ作成　画像と音声のデータからペアを作成する．画像データは，120 クラスあるが，それらの各クラスについて 29 個の採取データから画像特徴ベクトルを 10 個ランダムに抽出し，1200 個のデータセットを作成した． それぞれの画像に対して，その画像のすべての属性に対応する音声特徴ベクトルを採取した音声データの中からランダムに抽出し，対応づけてペアを作った．つまり，例えば 2 属性の実験では “赤の大きい 3” の画像について “あか”，“さん” の音声データを対応させて 2 つのペアを作った (3 属性の実験では “大きい” を含めて 3 つのペアを作った)． こうして 1200 × 属性数 だけのペアを学習データとして用意した．以下でこのようにして作った学習データセットを DataSet(1) と呼ぶことにする．

DataSet(1) は各クラスについてのすべての属性に対する教示を含んでいるが，必ずしもすべての属性に対する教示を含まない場合の性能を見るために，DataSet(1) から属性ごとにランダムに教示データを間引いた学習データセット，DataSet(0.8), DataSet(0.6), DataSet(0.4) を作成した．それぞれのデータセットは DataSet(1) から 20%，40%，60% のデータを間引いたものである． 具体的には，画像と音声の組合せで (120 × 属性数) 個あるクラスからランダムにクラスを 20%, 40%, 60% 選んで，それぞれのクラスに対応するデータを 10 個づつまとめて削除した．これらの学習データセットでは，例えば “赤の大きい 3” という画像についてのすべての教

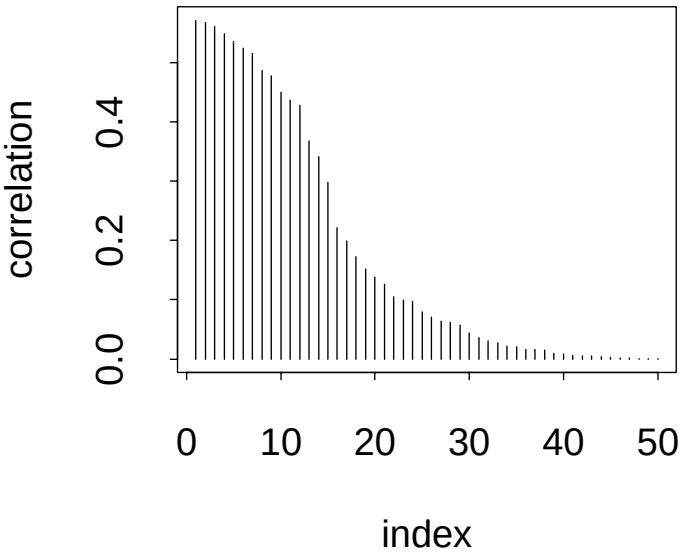

図 6.2: 画像と音声の正準空間での相関 (属性の数 3，DataSet(1)). 横軸： 次元の index(50 次元まで)，縦軸 : 各次元の相関の大きさ

示 (あか，おおきい，さん) が含まれているとは限らない.

正準相関分析　作成した学習データをもとに画像と音声の間で正準相関分析を行う． 正準相関分析は本質的な情報だけを残し余分な次元を落とすのが目的であるが，逆に次元を落としすぎると属性を分離できなくなってしまう可能性がある．最適な次数を決定するのは困難だが，本実験ではとりあえず，10 次元とした．図 6.2 は属性数を 3 にしたときの，学習サンプルしたときの正準空間での相関の大きさである． 横軸は次元のインデックスをあらわし，縦軸は各次元での相関の大きさを表す．10 次元というのは，相関がほぼ 0.5 以上になる部分に相当している．

EM アルゴリズムによる学習　正準空間の中で一対多の写像を 6.4 で述べた EM アルゴリズムを用いて学習する．

画像に対する属性の出力　学習サンプルを混合線形分布と EM アルゴリズムによって得られたパラメータにしたがって，属性数 (K 個) のクラスに分類しておく．具体的には，各サンプルに対して，属性の条件つき分布 $q(k \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta})$ を最大にするような k を割り当てる．

新たに，画像特徴が与えられたときには，画像特徴をまず正準空間に射影した後，EM アルゴリズムによって得られたパラメータにしたがって音声特徴の正準空間に 1 対多のマッピングを行ない，属性数だけの出力を得る． それぞれの出力に対して，その空間の中で最も近い音声特徴をあらかじめクラス分けしておいた学習サンプルの中から選ぶ (nearest neighbor 法)．音声データを蓄えておき，この特徴に対応する音声を再生させることもできるが，今回は音声再生はインプリメントを行なわず，選ばれた音声特徴の真のラベルを見て画像特徴にマッチしているかどうかの評価を行なった．

学習結果の評価　学習データセット DataSet(1), DataSet(0.8), DataSet(0.6) のそれぞれに対し，EM アルゴリズムによって学習した結果を，(1) 学習データセットに対して正解率を評価する (closed test)，(2) 学習に用いなかった画像データからランダムに抽出した 1000 個のテスト画像に対して正解率を評価する (open test)，の二通りの評価を行った． 全ての属性を正しく答えられたものの割合のほか，各属性ごとの正解率も算出した．また，EM アルゴリズムの局所最適解への収束の影響を避けるため，および，データセットによる偏りをなくすために，正解率の算出にあたっては，学習・テスト両方のデータセットを別々の乱数にしたがって 5 回生成し，それぞれに対して同じ実験を行なって，それらの結果の平均を取った．

なお，システムには属性名 (“いろ”，“かたち” など) は与えられていないため，内部では単に 1，2 という番号で表現されている． 番号と属性名の対応づけは最も自然なもの (属性分類の正解率が高くなるもの) を選んだ．

表 6.2: EM アルゴリズムの初期解での正解率 (単位は%．2 クラス， DataSet(1) の学習．括弧内は属性分類の正解率)

	Closed test	Open test
色	74.2 (78.4)	73.3 (77.5)
形	21.5 (26.2)	19.8 (24.9)
全属性	5.1	3.6

表 6.3: EM アルゴリズムの初期解での正解率 (単位は%．3 クラス， DataSet(1) の学習． 括弧内は属性分類の正解率)

	Closed test	Open test
色	31.1 (32.2)	27.8 (29.2)
形	8.8 (24.8)	9.4 (25.2)
大きさ	55.7 (58.3)	54.0 (57.7)
全属性	0.1	0.0

6.6.2 結果

まず，DataSet(1) に対して学習を行なった結果を示す.

表 6.2，6.3 はそれぞれ属性数 2，3 の場合の EM アルゴリズムを始める前の初期値の段階での正解率の表である．Closed test と open test のそれぞれに対し, 各属性ごとの正解率と全ての属性を正解した率を示している. 各属性の正解率の括弧内は, 属性の出力は必ずしも正解していなくても属性分類は正しいとしたときの正解率 (例えば赤に対して青と答えたものも含めたもの) をあらわす.

初期化を工夫することによって，ある程度の分類はできているが，色属性や大きさ属性に比べ，属性内のクラス数が最も多い形属性に関しては正解率が低い． また，全属性すべて正解するものはこの段階ではほとんどない.

EM アルゴリズムは属性数 2 の場合も属性数 3 の場合も，30 ステップ程度でどちらもある程度収束した． 30 ステップ後の正解率を表 6.4，6.5 に示す．形属性に

表 6.4: EM アルゴリズム 30 ステップ後の正解率 (単位は%．2 クラス， DataSet(1) の学習． 括弧内は属性分類の正解率)

	Closed test	Open test
色	98.1 (99.3)	98.1 (99.5)
形	76.3 (90.5)	74.2 (89.5)
全属性	74.5	72.6

表 6.5: EM アルゴリズム 30 ステップ後の正解率 (単位は%．3 クラス， DataSet(1) の学習． 括弧内は属性分類の正解率)

	Closed test	Open test
色	78.1 (79.5)	76.8 (78.3)
形	65.1 (92.4)	62.5 (91.8)
大きさ	84.3 (85.3)	84.7 (85.6)
全属性	47.9	45.2

関しては正解することが難しいことを示しているが，属性分類だけの正解率 (括弧内の数値) は高いので，分類そのものはある程度成功しているといえる．一方，色や大きさに関しては高い正解率を示している．

DataSet(0.8), DataSet(0.6), DataSet(0.4) については全属性の結果のみを表 6.6，6.7 に示す (比較のため，DataSet(1) の結果も載せた)．ある程度データを間引いても学習することは可能だが，間引きすぎると学習が困難になることを示している．

また，結果全体を通して 3 属性の学習のほうが 2 属性よりもかなり難しいことがわかる．

本論文で行なった実験結果については，正準相関分析の次元数の選択，および，正準相関分析や混合モデルの非線形への拡張，あるいは nearest neighbor 法を k-nearest neighbor 法に拡張するなどによって改善できる余地が存在するが，それらについては， 誤った教示が含まれた場合に対する頑健さの評価や実験結果のより詳細な分

表 6.6: 間引いたデータに対する学習の全属性の正解率 (単位は%．2 クラス． EM アルゴリズム 30 ステップ後)

	Closed test	Open test
DataSet(1)	74.5	72.6
DataSet(0.8)	58.0	53.1
DataSet(0.6)	56.4	46.3
DataSet(0.4)	38.9	26.6

表 6.7: 間引いたデータに対する学習の全属性の正解率 (単位は%．3 クラス． EM アルゴリズム 30 ステップ後)

	Closed test	Open test
DataSet(1)	47.9	45.2
DataSet(0.8)	49.5	47.0
DataSet(0.6)	20.3	15.5
DataSet(0.4)	2.7	2.7

析を含め，今後の課題として残されている．

6.7 本章のまとめ

物体の属性の学習を例として，マルチモーダル対話システムにおける学習の基本的な枠組みを提案した．ここで目指しているものは，記号レベルでの人間の関与をできるだけ減らし，人間とのインタフェースに現れるパターンレベルでの情報交換だけで，外界の構造を獲得するようなシステムである．

このような観点から，物体のある画像に対して，複数の属性のうちの 1 つを人間が音声で教えるという課題に対し，モダリティ間に共通に含まれる情報を抽出する正準相関分析と属性を隠れ変数とした混合分布によるモデル化と EM アルゴリズ

ムを組み合わせた統計的学習・認識の枠組みを示し，実験を行った．

以下では，本研究が提示した枠組をより一般化した形でまとめ，本研究の問題点を明らかにしておく．マルチモーダル情報源からのパターン的情報を統合して，そこに埋め込まれた構造を学習によって獲得するという課題は主に以下の 3 つの段階に分割できる．

各情報源における情報表現　音声信号，画像信号などパターン的なデータは往々にしてさまざまな変動や次元数の変化をもたらす． 本研究では，データの変動を最小限に抑えるとともに，正準相関分析などの多変量解析手法を適用するために，次元数の正規化を行っている．本研究では用いていないが，隠れマルコフモデルなどを用いた表現も次元数を一定にするために有力な表現法である．

また，モダリティによって提示されるパターンの長さやタイミングは異なるのが普通である．更に，どこからどこまでが一つのパターンかということも未知であることが多い． 本研究ではこの同期と切り出しの問題がすべて解かれているとして出発しているが，実際の場面への適用では重要な問題となる．

共通情報抽出　情報源毎の次元数が大きいので，そのままでは汎化などの観点から構造抽出がうまくいかない． そこで次元の圧縮が必要となり，特に，構造抽出に有効な情報として，両方のモダリティに共通して含まれている情報だけを取り出すということが行われる． 本研究では線形多変量解析手法である正準相関分析を用いたが，非線形への拡張を行ったり，相互情報量最大化といった一般的な尺度を用いた手法を試みる価値はある．

構造抽出　本研究で扱った属性概念獲得課題は，各モダリティ毎の処理では決して得られない構造抽出課題である． 現実には，そこまで複雑ではなくモダリティ間が互いに補完し合うような形で構造抽出が行われる場面が多いと考えられる． いずれにしても，パターン的な情報からシンボル的な構造の抽出は，いろいろな形のクラスタリングの過程であると見なせることが多い．この際，クラスタリングのためのモデルとして混合分布が有効に働く．ただし，より考えを進めると混合分布だけでは能力が不足してくる．例えば，本研究で扱った属性概念にしても，実際は均

一な構造ではなく，それ自身階層構造をなしている．これに対しては例えば階層的な混合分布の適用可能性を示しているし，因果関係などを扱う際にはグラフィカルモデルのような複雑なモデルを扱う必要がある．

以上で述べた 3 段階はすべて何らかの意味で「学習」に関係しており，しかもそれぞれの要素は完全に独立しているわけではなく，互いに関連している．学習という観点でもう一点重要なのは，本研究ではすべてデータを集めた上でのバッチ処理的な学習を行っているという点である．切り出しや同期の問題も関係するため，オンライン化が容易ではない．より実用的なマルチモーダル対話システムの学習への応用や，人間の発達モデルなどへの適用を行うためには，これらの問題を一つ一つ解決していく必要がある．

第7章　結 論

本章では，本論文を総括し，残された課題についてまとめる．

複雑な環境下における学習モデルとして組合せ構造を持つグラフィカルモデルが重要であり，その中でも最も基本的な有限混合分布モデルについてその学習に関して，理論と応用の側面から考察を行った．

第 2 章と第 3 章では混合分布および学習に関する総括的なまとめを行い，混合分布のもつ特殊性や一般的なモデルとの関連性を述べた． その中でも，汎化と学習アルゴリズムの問題が計算論的な学習理論の中核をなす重要な研究であることを示した．

第 4 章では，あるクラスの正規混合分布を取り上げ，そのモデルの複雑度が連続パラメータによって制御でき，その制御によって分岐現象が観測できることを示した．また，その分岐点について汎化能力を調べ，汎化のバイアスが通常考えられているようにモデルパラメータの数に対して単調に増加するのとは異なる振舞いを示すことを理論的に示した．実際に理論的に明らかになったのは，分岐点から少し離れた近傍であり，分岐点そのものの解析には理論的な困難が残されており，本研究で解決することはできなかった． 実験的にはそれほど病的な問題は起きなかったが今後の課題として残されている．

第 5 章では，確率分布の位置・尺度・回転パラメータの学習という一般的な問題に対し，正規混合分布を用いた ECM アルゴリズムを用いると各ステップが 2 次方程式の解として与えられることを示した．これは EM アルゴリズムが一般に難しい問題を簡単にすることができる可能性を示す例である．また，閉じた形のアルゴリズムにより尤度の単調性を保証し，安定したアルゴリズムを導くことができた．ただし，実験を行った物体認識などへの応用では位置・尺度・回転パラメータだけでは不十分なことも考えられ，その扱いは今後の課題として残されている．さらに，EM アルゴリズムは ICA (独立成分分析) や因子分析といった多変量解析手法への応

用も盛んになされている [21]．著者らは第 6 章の研究とも関連し，複数情報源に対する ICA についての研究も進めており [12, 7]，今後研究を進めて行く予定である．

第 6 章では，複数情報源からの属性概念獲得という，マルチモーダル対話システムや発達心理学において重要な課題に対して，混合分布による統計的な枠組みを設定し，その有効性を検討した．正準相関分析と混合分布は共に線形モデルに基づいたものを用いており，学習がバッチ処理で行われるなど，実用的には不十分な面もあるが，枠組み自体の有効性を示すには十分な結果が得られた． 今後は，6.7 でも述べたように，それぞれの要素技術を高度化し，更に課題自体もより現実的なものへと検討していく必要がある．MIT の Roy は最近，同様のマルチモーダル学習の枠組を単語獲得課題に適用し有効性を示しており [76]，今後もこれらの研究が人工知能や対話システムの研究において発展していくことが期待できる．

以上の結果から，それぞれ今後検討すべき課題は残されているものの，混合分布による統計的な枠組みが人工知能やパターン認識における複雑な学習課題に有効であることを示すという本論文の目的はある程度達成されたものと考える．

謝辞

本論文は筆者が通商産業省工業技術院電子技術総合研究所において行ってきた研究のうち，混合分布の学習に関する研究をまとめたものです．

本研究は電子技術総合研究所の多くの方々のご支援のおかげで行うことができました．特に，橋田浩一 情報科学部長をはじめ，田村浩一郎 現中京大学教授，諏訪基 現大阪工業技術研究所所長，中島秀之 現企画室長の歴代の情報科学部長には，終始多大の御支援と御理解を頂きました．

本研究の中でも数理的な側面に関しては，旧情報数理研究室，そして現在の情報ダイナミクスラボや学習統合基礎ラボのメンバーとの数理的な議論によって筆者の研究が深められました．大津展之 元情報数理研究室長 (現 知能情報部長)，梅山伸二 情報ダイナミクスラボリーダー，麻生英樹 学習統合基礎ラボリーダーをはじめ，栗田多喜夫氏，関田巌氏，田中勝氏，本村陽一氏，藤木淳氏，西森康則氏の旧情報数理研究室メンバーに謝意を表します．

さらにいくつかの章は通産省のプロジェクトであるリアルワールドコンピューティングプログラム (RWCP) におけるさまざまな方との共同研究によって生まれたものです．第 4 章は筆者が 1995 年にオランダに滞在した際に Nijmegen 大学 (RWCP SNN 研究所) の H. J. Kappen 氏と行った共同研究に基づき，学習統合基礎ラボの中で行われました．また，統計数理研究所の福水健次 助教授からも適切な御指摘を頂きました．第 5 章は，インターモーダル学習ラボの長谷川修氏，速水悟氏，吉村隆氏，麻生氏の協力のもと，音声や画像といった生のデータを扱った研究にまとめることができました．これらの方々に感謝致します．

また，甘利俊一先生 (現 理化学研究所) には筆者が修士在学中の指導教官として研究に対する姿勢や数理的な考え方に影響を与えて頂きました．ここに心から感謝の意を表します．

最後に，本論文をまとめるにあたって，東京大学工学部計数工学科 駒木文保 助教授に適切な御助言と御指導をいただきました．また，東京大学工学部 合原一幸 教授，同 岡部靖憲 教授，同 堀田武彦 講師，東京大学経済学部 竹村彰通 教授 には，本論文の完成に有益な御教示，御助言を頂きました．ここに深く感謝の意を表します．

参考文献

[1] 赤穂昭太郎. アテンション領域における EM アルゴリズム. 神経回路学会第 4 回全国大会講演論文集, 1994.

[2] S. Akaho. The EM algorithm for multiple object recognition. In *Proc. of Int. Conf. on Neural Networks (ICNN'95)*, pp. 2426–2431, 1995.

[3] 赤穂昭太郎. EM アルゴリズムの幾何学. 情報処理, Vol. 37, No. 1, pp. 43–51, 1996.

[4] 赤穂昭太郎, 長谷川修, 吉村隆, 麻生英樹, 速水悟. EM 法を用いた複数情報源からの概念獲得. 電子情報通信学会技術研究報告 PRMU96–91, 1996.

[5] 赤穂昭太郎, 速水悟, 長谷川修, 吉村隆, 麻生英樹. EM 法を用いた複数情報源からの概念獲得. 電子情報通信学会論文誌, Vol. J80-A, No. 9, pp. 1546–1553, 1997.

[6] 赤穂昭太郎. ECM 法を用いた確率分布の位置、尺度、回転パラメータの推定法. 電子情報通信学会論文誌, Vol. J82-D-II, No. 12, pp. 2240–2250, 1999.

[7] 赤穂昭太郎, 梅山伸二. マルチモーダル独立成分分析 —複数情報源からの共通特徴抽出法—. 電子情報通信学会論文誌, Vol. J83-A, No. 6, pp. 669–676, 2000.

[8] S. Akaho, S. Hayamizu, O. Hasegawa, T. Yoshimura, and H. Asoh. Multiple attribute learning with canonical correlation analysis and EM algorithm. Technical Report 97–8, Electrotechnical Laboratory, 1997.

[9] S. Akaho, Hayamizu, S., Hasegawa, O., Itou, K., Akiba, T., Asoh, H., Kurita, T., K. Sakaue K. Tanaka, and N. Otsu. Recent developments for multimodal interac-

tion by visual agent with spoken language. In *Proc. on Int. Conf. on Multimodal Interaction (ICMI'96)*, pp. 135–139, 1996.

[10] S. Akaho and H.J.Kappen. Nonmonotonic generalization bias of Gaussian mixture models. *Neural Computation*, Vol. 12, No. 6, pp. 1411–1428, 2000.

[11] S. Akaho and H. J. Kappen. Nonmonotonic generalization bias of Gaussian mixture models. Technical Report 98–22, Electrotechnical Laboratory, 1998.

[12] S. Akaho, Y. Kiuchi, and S. Umeyama. MICA: Multimodal independent component analysis. In *Proc. of Int. Joint Conf. on Neural Networks (IJCNN'99)*, pp. 927–932, 1999.

[13] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, Vol. 19, No. 6, pp. 716–723, 1974.

[14] S. Amari. *Differential Geometrical Methods in Statistics*. Lecture Notes in Statistics. Springer-Verlag, 1985.

[15] S. Amari. A universal theorem on learning curves. *Neural networks*, Vol. 6, pp. 161–166, 1993.

[16] 甘利俊一, 長岡浩司. 情報幾何の方法. 岩波講座 応用数学. 岩波書店, 1993.

[17] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, Vol. 8, No. 9, pp. 1379–1408, 1995.

[18] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 1984.

[19] H. Asoh, S. Akaho, O. Hasegawa, T. Yoshimura, and S. Hayamizu. Intermodal learning of multimodal interaction systems. In *Proc. of the Int. Workshop on Human Interface Technology (IWHIT '97)*, 1997.

[20] H. Asoh and O. Takechi. An approximation of nonlinear canonical correlation analysis by multilayer perceptrons. In *Artificial Neural Network (Proc. of ICANN'94)*, pp. 713–716, 1994.

[21] H. Attias. Independent factor analysis. *Neural Computation*, Vol. 11, pp. 803–851, 1999.

[22] N. Barkai, H. S. Seung, and H. Sompolinsky. Scaling Laws in Learning of Classification Tasks. *Physical Review Letters*, Vol. 70, pp. 3167–3170, 1993.

[23] N. Barkai and H. Sompolinsky. Statistical mechanics of the maximum-likelihood density estimation. *Physical Review E*, Vol. 50, No. 3, pp. 1766–1769, 1994.

[24] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, Vol. 41, No. 1, pp. 164–171, 1970.

[25] J. C. Bezdek. A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Trans. on PAMI*, Vol. 2, No. 1, pp. 1–8, 1980.

[26] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, 1993.

[27] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[28] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. In *Statistics and decisions*, pp. 205–237. Munich: Oldenburg Verlag, 1984. (Supplemental Issue, no. 1).

[29] D. Dacunha-Castelle and É. Gassiat. Testing in locally conic models, and application to mixture models. *ESAIM: Probability and Statistics*, Vol. 1, pp. 285–317, 1997.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of Royal Statistical Society Series B*, Vol. 39, pp. 1–38, 1977.

[31] B. S. Everitt and D. J. Hand. *Finite Mixture Distributions*. Chapman & Hall, 1981.

[32] D. H. Fisher, M. J. Pazzani, and P. Langley, editors. *Concept Formation: Knowledge and Experience in Unsupervised Learning*. Morgan Kaufmann, 1991.

[33] B. J. Frey. *Graphical Models for Machine Learning and Digital Communication*. MIT Press, 1998.

[34] K. Fukumizu. Special statistical properties of neural network learning. In *Proc. of Int. Sympo. on Nonlinear Theory and Its Applications (NOLTA'97)*, pp. 747–750, 1997.

[35] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

[36] K. Hagiwara, N. Toda, and S. Usui. On the problem of applying AIC to determine the structure of a layered feed-forward neural network. In *Proc. of Int. Conf. on Neural Networks (IJCNN'93)*, pp. 2263–2266, 1993.

[37] 長谷川修, 伊藤克亘, 麻生英樹, 赤穂昭太郎, 秋葉友良, 栗田多喜夫, 速水悟, 坂上勝彦, 田中, 大津展之. コンピュータとの対話におけるユーザの振舞の解析. 電子情報通信学会技術研究報告 PRU95–57, 1995.

[38] O. Hasegawa, K. Itou, T. Kurita, S. Hayamizu, K. Tanaka, K. Yamamoto, and N. Otsu. Active agent oriented multimodal interface system. In *Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI '95)*, pp. 82–87, 1995.

[39] P. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[40] 石井健一郎, 上田修功, 前田英作, 村瀬洋. わかりやすいパターン認識. オーム社, 1998.

[41] R. A. Jacobs and M. I. Jordan. A competitive modular connectionist architecture. In Lippman et al, editor, *Advances in Neural Information Processing Systems 3*, pp. 767–773. Morgan Kaufman, 1991.

[42] M. Jordan. *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

[43] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, Vol. 6, pp. 181–214, 1994.

[44] B. Kappen. Using Boltzmann Machines for probability estimation: A general framework for neural network learning. In S. Gielen et al, editor, *Proc. of Int. Conf. of Artificial Neural Networks (ICANN'93)*, pp. 521–526, 1993.

[45] B. Kappen. Deterministic learning rules for Boltzmann Machines. *Neural Networks*, Vol. 8, No. 4, pp. 537–548, 1995.

[46] 川人光男. 脳の計算理論. 産業図書, 1996.

[47] T. Kurita, N. Otsu, and T. Sato. A face recognition method using higher order local autocorrelation and multivariate analysis. In *Proc. of 11th International Conf. on Pattern Recognition (ICPR)*, Vol. II, pp. 213–216, The Hague, 1992.

[48] 黒川隆夫. ノンバーバルインタフェース. オーム社, 1994.

[49] E. Levin, N. Tishby, and S. A. Solla. A statistical approach to learning and generalization in layered neural networks. *Proc. of IEEE*, Vol. 78, No. 10, pp. 1568–1574, 1990.

[50] G. J. McLachlan and K. E. Basford. *Mixture models*. Mercel Dekker, 1987.

[51] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, 1997.

[52] X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: a general framework. *Biometrika*, Vol. 80, pp. 267–278, 1993.

[53] 宮川雅巳. EM アルゴリズムとその周辺. 応用統計学, Vol. 16, No. 1, pp. 1–19, 1987.

[54] 宮川雅巳. グラフィカルモデリング. 朝倉書店, 1997.

[55] J. Moody. The *effective* number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In J. Moody et al, editor, *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann, 1992.

[56] 本村陽一, 赤穂昭太郎, 麻生英樹. ベイジアンネット学習の知能システムへの応用. 計測と制御, pp. 468–473, 1999.

[57] N. Murata, S. Yoshizawa, and S. Amari. A criterion for determining the number of parameters in an artificial neural network model. In T. Kohonen et al, editor, *Artificial Neural Network (Proc. of ICANN'94)*, pp. 9–14. Elsevier, 1994.

[58] N. Murata, S. Yoshizawa, and S. Amari. Network information criterions — determining the number of parameters for an artificial neural network model. *IEEE Trans. on Neural Networks*, Vol. 5, pp. 865–872, 1994.

[59] 中川聖一, 中西宏文, 古部好計, 板橋光義. 視聴覚情報の統合化に基づく概念の獲得. 人工知能学会誌, Vol. 8, No. 4, pp. 499–508, 1993.

[60] 中川聖一, 升方幹雄. 視聴覚情報の統合化に基づく概念と文法の獲得システム. 人工知能学会誌, Vol. 10, No. 4, pp. 619–627, 1995.

[61] M. J. Nijman and H. J. Kappen. Symmetry breaking and training from incomplete data with radial basis Boltzmann machines. *International J. of Neural Systems*, Vol. 8, pp. 301–316, 1997.

[62] 西森秀稔. 統計力学と知識情報処理の関わり合い. 数理科学, Vol. 438, pp. 5–11, 1999.

[63] 大津展之. パターン認識における特徴抽出に関する数理的研究. 研究報告 818, 電子技術総合研究所, 1981.

[64] 大津展之. リアルワールドコンピューティング研究計画—実世界における柔軟な知能を目指して. 人工知能学会誌, Vol. 9, No. 3, pp. 358–364, 1994.

[65] J. K. Patel and C. B. Read. *Handbook of the Normal Distribution, Second Edition*. Marcel Dekker, 1996.

[66] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, Vol. 78, No. 9, pp. 1481–1497, 1990.

[67] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. PTR Prentice-Hall, 1995. 古井 貞熙 (訳): 音声認識の基礎, NTT アドバンステクノロジ, 1995.

[68] C. R. Rao. *Linear Statistical Inference and Its Applications, 2nd edition*. John Wiley & Sons, 1973. 奥野ほか (訳) : 統計的推測とその応用, 東京図書, 1977.

[69] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, Vol. 26, pp. 195–239, 1984.

[70] M. Revow, C. Williams, and G. Hinton. Using generative models for handwritten digit recognition. *IEEE Trans. on PAMI*, Vol. 18, pp. 592–606, 1996.

[71] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. of the Royal Statistical Society, Series B*, Vol. 59, No. 4, pp. 731–792, 1997.

[72] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, 1995.

[73] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, Vol. 14, pp. 1080–1100, 1986.

[74] C. R. Robert. *The Bayesian choice: a decision-theoretic motivation*. Springer-Verlag, 1994.

[75] K. Rose, E. Gurewitz, and G. Fox. Statistical mechanics of phase transitions in clustering. *Physical Review Letters*, Vol. 65, pp. 945–948, 1990.

[76] D. Roy. Integration of speech and vision using mutual information. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP2000)*, 2000.

[77] S. J. Russell and P. Norvig. *Artificial Intelligence, Modern Approach*. Prentice-Hall, 1995. 古川康一 (監訳): エージェントアプローチ 人工知能, 共立出版, 1997.

[78] 坂本慶行, 石黒真木夫, 北川源四郎. 情報量統計学. 共立出版, 1983.

[79] A.J.C. Sharkey, editor. *Combining Artificial Neural Nets, Ensemble and Modular Multi-Net Systems*. Springer-Verlag, 1998.

[80] H. Shimodaira. A new criterion for selecting models from partially observed data. In P. Cheeseman and R. W. Oldford, editors, *Selecting Models from Data: AI and Statistics IV*, chapter 3, pp. 21–29. Springer-Verlag, 1994.

[81] 篠本滋. 情報の統計力学. 丸善, 1992.

[82] 竹内啓. 情報統計量の分布とモデルの適切さの規準. 数理科学, Vol. 153, pp. 12–18, 1976.

[83] 竹内啓（編）. 統計学辞典. 東洋経済新報社, 1989.

[84] M. A. Tanner. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer-Verlag, 1993.

[85] R. A. Tapia and J. R. Thompson. *Nonparametric Probability Density Estimation*. Johns Hopkins Univ. Press, 1978.

[86] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distribution*. John Wiley & Sons, 1985.

[87] L. G. Valiant. A theory of learnable. *Comm. ACM*, Vol. 27, No. 11, pp. 1134–1142, 1984.

[88] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, 1984.

[89] 渡辺美智子, 山口和範 (編). EM アルゴリズムと不完全データの諸問題. 多賀出版, 2000.

[90] S. Watanabe. Algebraic analysis for singular statistical estimation. In *Lecture Notes in Computer Sciences*, Vol. 1720, pp. 39–50. Springer-Verlag, 1999.

[91] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, Vol. 11, pp. 95–103, 1983.

[92] 柳井晴夫, 高木広文. 多変量解析ハンドブック. 現代数学社, 1986.

[93] I. Ziskind and M. Wax. Maximum likelihood localization of multiple sources by alternating projection. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 10, pp. 1553–1560, 1988.

付 録A　第 4 章の付録

A.1　定理 3 および定理 4 の証明

一般性を失うことなく，$\mathrm{E}_q[\,x\,]=0$ としてよい．仮定より，正規分布の要素数は偶数なので，$K=2K_1$ とする．

真の分布は対称で正規分布の数は偶数だから，分岐は対称で

$$p(x;\,W;\,\beta)=\frac{1}{2K_1}\sum_{k=1}^{K_1}\sqrt{\frac{\beta}{\pi}}\,p_k(x;\,w_k;\,\beta), \tag{A.1}$$

と書ける．ただし

$$p_k(x;\,w_k;\,\beta)=\exp(-\beta(x-w_k)^2)+\exp(-\beta(x+w_k)^2). \tag{A.2}$$

ここで，分布の対数の微分を

$$L_k(x;\,W;\,\beta)=\frac{\partial}{\partial w_k}\log p(x;\,W;\,\beta)=\frac{\partial_k p_k(x;\,w_k;\,\beta)}{p(x;\,W;\,\beta)}, \tag{A.3}$$

とおく．真の分布 q に関する尤度は，真の最尤解 W^* で 0 になり，

$$R_{\exp}(W^*;\,\beta)=\mathrm{E}_q[\,L_k(x;\,W^*;\,\beta)\,]=0, \tag{A.4}$$

を満たす．この解は，分岐点より前では $W^*=0$ となり，分岐以後では W^* は一般に 0 ではなくなる．$\Delta\beta=\beta-\beta_c$ を用いて $\Delta w_k=w_k$ を表すために，$L_k(x;\,W,\beta)$ を W については 3 次まで，β については 1 次まで，$\beta=\beta_c$ と $W=0$ のまわりで展開する．

$$\begin{aligned}\mathrm{E}_q[\,L_k(x;\,W;\,\beta)\,] \;=\;& \frac{2}{K_1}\Delta\beta\Delta w_k-\frac{1}{2{K_1}^2}\sum_{j\neq k}(\frac{s_4}{(\sigma^2)^2}-1)\Delta w_k\Delta w_j^2\\ &+\frac{1}{6K_1(\sigma^2)^2}\left\{\frac{s_4}{(\sigma^2)^2}-3-\frac{3}{K_1}(\frac{s_4}{(\sigma^2)^2}-1)\right\}\Delta w_k^3\\ &+\text{higher order terms.}\end{aligned} \tag{A.5}$$

$\mathrm{E}_q\,[\,L_k(x;\,W;\,\beta)\,]=0$ とおき，高次の項を省略すると，K_1 個の $\Delta\beta$ と $\Delta w_k=w_k$ に関する方程式を得る．これらはそれぞれ，$f_k(W,\Delta\beta)\,w_k=0$ という形をしているので，$w_k=0$ または $f_k(W,\Delta\beta)=0$ を満たす．そこで，K_1 個の中で $f_k(W,\Delta\beta)=0$ を満たす方程式の個数を K_2 と置く．

まず，$s_4=3(\sigma^2)^2$ のとき，$f_k(W,\Delta\beta)=0$ はすべて

$$\Delta\beta=\frac{\sum_k\Delta w_k^2}{2K_1(\sigma^2)^2} \tag{A.6}$$

という単一の方程式に帰着され，定理 3 の第 3 の場合になる．したがって，高次の項を無視すると w_k が β から一意には定まらないことを意味する．

次に，$s_4\neq 3(\sigma^2)^2$ とすると，$f_k(W,\Delta\beta)=0$ は

$$\Delta\beta=\left\{\frac{s_4-3(\sigma^2)^2}{12(\sigma^2)^4}+\frac{K_2}{4K_1}(s_4-(\sigma^2)^2)\right\}\Delta w_k^2, \tag{A.7}$$

という方程式となり，0 でない w_k は符号を除きすべて同じ値 w になる．

このうち尤度を最大にする K_2 を求める．ある温度 $1/\beta$ での尤度 $R(W;\,\beta)$ を分岐点での尤度の周りで展開し整理すると，

$$\begin{aligned}R(W;\,\beta)\;\simeq\;& R(0;\,\beta_c)+\frac{1}{2}\frac{\partial^2R(0;\,\beta_c)}{\partial\beta^2}\Delta\beta^2\\&+\frac{1}{24}\sum_{k=1}^{K_1}\left(\frac{\partial^4R(0;\,\beta_c)}{\partial^4w_k^4}+\sum_{j\neq i}\frac{\partial^4R(0;\,\beta_c)}{\partial^2w_k\partial^2w_j}\right)\Delta w_k^4\\&+\text{higher order terms},\end{aligned} \tag{A.8}$$

となる． $\partial^2R(0;\,\beta_c)/\partial\beta^2$ 等に具体的な値を代入し，更に式 (A.7) を $\Delta\beta$ に入れると，

$$\begin{aligned}&R(W;\,\beta)=R(0;\,\beta_c)\\&\quad+\frac{3(\sigma^2)^2-s_4}{144(\sigma^2)^2}\left\{(s_4-(\sigma^2)^2)\left(\frac{K_2}{K_1}\right)^2-\frac{2}{3}(s_4-2(\sigma^2)^2)\frac{K_2}{K_1}+s_4-3(\sigma^2)^2\right\}w^4\\&\quad+\text{higher order terms},\end{aligned} \tag{A.9}$$

となる．

ここで，$s_4<3(\sigma^2)^2$ と仮定すると，一般に，$s_4\geq(\sigma^2)^2$ なので，右辺第 2 項は K_2/K_1 について下に凸になる．したがって，最大値は $K_2=K_1$ または $K_2=0$ の

ときに限られ，それぞれの場合を比較すると，$K_2 = K_1$ が最尤解となる． これは $w_k = 0$ となる解が一つもないことを意味しており，この場合分岐は 2-way になることが示された．

一方，$s_4 > 3(\sigma^2)^2$ と仮定すると，今度は上に凸になるので，極大点を調べると，

$$K_2 = \frac{s_4 - 2\sigma^4}{3(s_4 - \sigma^4)} K_1 < K_1 \tag{A.10}$$

となり，この場合分岐が 3-way になることが示された．

また，式 (A.9) から，$O(\Delta\beta) = O(w^2)$ の項は 0 になることがわかり，定理 4 も導かれる．

証明終

A.2 定理 5 の証明

温度が最初の分岐点より高いとき，すべての要素分布は一つの正規分布に退化する． したがって，この場合 1 個の要素分布だけからなるモデルを考えればよい．

1 個の正規分布に対する $D(W^*)$ と $H(W^*)$ は次のように計算できる．

$$D_{ij}(W^*) = 4\beta^2 V_{ij}, \tag{A.11}$$

$$H_{ij}(W^*) = 2\beta\delta_{ij}, \tag{A.12}$$

ここで V_{ij} は x_k と x_j の間の共分散行列であり，δ_{ij} は Kronecker の δ である．したがって TIC は

$$h_{\text{eff}}(\beta) = \mathrm{Tr}[H(W^*)^{-1} D(W^*)] = 2\beta \mathrm{Tr}[V_{\boldsymbol{x}}] \tag{A.13}$$

で与えられる．

証明終

A.3 定理 6 の証明

A.1 と同様に $\mathrm{E}_q[x] = 0$ を仮定する．定理 3 より分岐は 2-way または 3-way だが，まず，2-way の場合から示す．

仮定 2 より，このモデルの TIC は 2 個の正規分布の混合モデルに等しい．2 個の正規分布の混合分布は

$$p(x;\ w_1, w_2;\ \beta) = \frac{1}{2}\sqrt{\frac{\beta}{\pi}}\left[\exp\{-\beta(x-w_1)^2\} + \exp\{-\beta(x+w_2)^2\}\right]. \quad \text{(A.14)}$$

と書ける．$D(w_1, w_2)$ と $H(w_1, w_2)$ はその定義から計算できる．最尤解では $w_1 = w_2 = w$ だから，

$$D(w,w) = \begin{bmatrix} d_1 & d_3 \\ d_3 & d_1 \end{bmatrix}, \quad \text{(A.15)}$$

$$H(w,w) = \begin{bmatrix} d_2 & d_3 \\ d_3 & d_2 \end{bmatrix}, \quad \text{(A.16)}$$

となる． ここで，

$$d_1 = \mathrm{E}_q\left[4\beta^2(x-w)^2\frac{(p_1)^2}{p^2}\right], \quad \text{(A.17)}$$

$$d_2 = d_1 + \mathrm{E}_q\left[2\beta\frac{p_1}{p} - 4\beta^2(x-w)^2\frac{p_1}{p}\right], \quad \text{(A.18)}$$

$$d_3 = \mathrm{E}_q\left[-4\beta^2(x-w)(x+w)\frac{p_1 p_2}{p^2}\right], \quad \text{(A.19)}$$

$p_1 = \exp(-\beta(x-w)^2)$, $p_2 = \exp(-\beta(x+w)^2)$, $p = p_1 + p_2$. さて，

$$\hat{h}_{\text{eff}}(\beta, w) = \mathrm{Tr}[H(w,w)^{-1}D(w,w)], \quad \text{(A.20)}$$

とおくと，これは $w = w^*$ で $h_{\text{eff}}(\beta)$ になる．

$$\hat{h}_{\text{eff}}(\beta, w) = \mathrm{Tr}[H^{-1}D] = 2\frac{d_1 d_2 - d_3^2}{d_2^2 - d_3^2}, \quad \text{(A.21)}$$

を最初の分岐点 ($\beta = \beta_c, w^* = 0$) のまわりで β と w について展開すると，

$$\begin{aligned}\hat{h}_{\text{eff}}(\beta, w) &= \hat{h}_{\text{eff}}(\beta_c, 0) + \left\{\frac{\partial}{\partial\beta}\hat{h}_{\text{eff}}(\beta_c, 0)\right\}\Delta\beta \\ &\quad + \frac{1}{2}\left\{\frac{\partial^2}{\partial w^2}\hat{h}_{\text{eff}}(\beta_c, 0)\right\}\Delta w^2 \\ &\quad + \text{higher order terms}, \end{aligned} \quad \text{(A.22)}$$

が得られる．上式の第 2 項と第 3 項は式 (4.3) から，ともに $\Delta\beta$ のオーダーである．

d_1, d_2, d_3 をそれらの値で置き換えると，まず $\lim_{\beta\downarrow\beta_c} \hat{h}_{\rm eff}(\beta, 0) = 1$ が得られる．これと，定理 5 の 1 次元の場合を考えることにより，$\lim_{\beta\to\beta_c} \hat{h}_{\rm eff}(\beta, 0) = 1$ となる．また，第 2 項の係数は

$$\frac{\partial}{\partial\beta}\hat{h}_{\rm eff}(\beta_c, 0) = 2\sigma^2, \tag{A.23}$$

となり，第 3 項の係数は $\beta = \beta_c$ に置き換える前の形で

$$\frac{1}{2}\frac{\partial^2}{\partial w^2}\hat{h}_{\rm eff}(\beta, 0) = 4\beta(1 - 2\beta\sigma^2 - \frac{1 - 4\beta^2 s_4}{1 - 2\beta\sigma^2}). \tag{A.24}$$

となる．$s_4 \neq (\sigma^2)^2$ のとき，$\beta_c = 1/(2\sigma^2)$ を用いると，$s_4 > (\sigma^2)^2$ なので，β が β_c に右から収束していくとき，式 (A.24) は $-\infty$ に発散する．

$s_4 = (\sigma^2)^2$ となるのは $q(x)$ が $\delta(x)$ か $(\delta(x-a) + \delta(x+a))/2$ に等しいときである．前者では分岐は起きないので，後者のみを考えればよい．一般性を失うことなく $a = 1$ とすると，右微分は

$$\frac{\partial}{\partial\beta}\hat{h}_{\rm eff}(\beta_c, 0) = -4, \tag{A.25}$$

となる．

次に分岐が 3-way の場合についても同様にして証明できる．この場合はモデルが

$$\begin{aligned} p(x;\ w_1, w_2, w_3;\ \beta) \quad = \quad & \sqrt{\frac{\beta}{\pi}}\Bigg(a[\exp\{-\beta(x - w_1)^2\} + \exp\{-\beta(x + w_2)^2\}] \\ & +(1 - 2a)\exp\{-\beta(x - w_3)^2\}\Bigg), \end{aligned} \tag{A.26}$$

という 3 つの混合分布モデルで，定理の仮定より a は定数となる．また最尤解に置いては $w_1 = w_2 = w, w_3 = 0$ となる．2-way の場合と同様にして，$D(w_1, w_2, w_3)$ および $H(w_1, w_2, w_3)$ の値を計算すると，

$$D(w, w, 0) = \begin{bmatrix} d_1 & d_2 & d_3 \\ d_2 & d_1 & -d_3 \\ d_3 & -d_3 & d_4 \end{bmatrix}, \tag{A.27}$$

$$H(w, w, 0) = \begin{bmatrix} d_5 & d_2 & d_3 \\ d_2 & d_5 & -d_3 \\ d_3 & -d_3 & d_6 \end{bmatrix}, \tag{A.28}$$

という形に書け，$\hat{h}_{\mathrm{eff}}(\beta, 0)$ が

$$\hat{h}_{\mathrm{eff}}(\beta, 0) = \frac{2d_1d_3^2 - 2d_1d_5d_6 + d_2^2d_4 + 2d_2^2d_6 + 6d_2d_3^2 + 4d_3^2d_5 - d_4d_5^2}{(d_2 + d_5)(d_2d_6 + 2d_3^2 - d_5d_6)}, \quad \text{(A.29)}$$

となるので，分岐点の周りで漸近展開する．式が複雑になるので途中の式は省略するが，2-way の場合と同様に $\partial^2\hat{h}_{\mathrm{eff}}(\beta, 0)/\partial w^2$ の右微分係数が

$$-(2a^2 - 2a + 1)(s_4 - (\sigma^2)^2), \quad \text{(A.30)}$$

の符号に応じて ∞ または $-\infty$ になる．分岐が 3-way のときは $s_4 > 3(\sigma^2)^2$ だから，上式は常に負であり，3-way のときも右微係数が $-\infty$ になることが示された．

証明終

付 録B　第 5 章の付録

B.1　Type I モデルの EM 法の導出

E ステップにおける関数 Q は

$$Q_{\mathrm{I}}(A, \boldsymbol{b}) = \left\langle \sum_{k=1}^{K} q_{\mathrm{I}\,k}^{(t)}(\boldsymbol{x}) \log p_{\mathrm{I}}(\boldsymbol{x}, k;\ A, \boldsymbol{b}) \right\rangle_{\rho} \tag{B.1}$$

で与えられるので，これをまず，$\boldsymbol{b} = \boldsymbol{b}^{(t)}$ に固定して A を求める．

$$\frac{\partial}{\partial a_i} Q_{\mathrm{I}}(A, \boldsymbol{b}^{(t)}) = \left\langle \sum_{k=1}^{K} \left\{ \frac{1}{a_i} - \frac{x_i(a_i x_i + b_i^{(t)} - \mu_{k,i})}{\sigma_{k,i}^2} \right\} q_{\mathrm{I}\,k}^{(t)}(\boldsymbol{x}) \right\rangle_{\rho} = 0 \tag{B.2}$$

を整理すると

$$\frac{1}{a_i} Z_{\mathrm{I}} - a_i X_{\mathrm{I}\,i} - Y_{\mathrm{I}\,i} = 0 \tag{B.3}$$

となる．ただし，$X_{\mathrm{I}\,i}, Y_{\mathrm{I}\,i}, Z_{\mathrm{I}}$ はそれぞれ式 (5.11), (5.12), (5.13) で与えられる．

これは a_i に関する 2 次方程式で，二つの解が得られるが，そのうち $a_i > 0$ という条件を満たすものは式 (5.10) の場合に限る．

一方得られた $A^{(t+1)}$ を用いて, $\boldsymbol{b}^{(t+1)}$ を求めると,

$$\frac{\partial}{\partial b_i} Q_{\mathrm{I}}(A^{(t+1)}, \boldsymbol{b}) = -U_{\mathrm{I}\,i} + b_i V_{\mathrm{I}\,i} = 0 \tag{B.4}$$

から，(5.14) の更新式が得られる．

B.2　Type II モデルの EM 法の導出

E ステップにおける関数 Q は

$$Q_{\mathrm{II}}(H, \boldsymbol{b}) = \left\langle \sum_{k=1}^{K} q_{\mathrm{II}\,k}^{(t)}(\boldsymbol{x}) \log p_{\mathrm{II}}(\boldsymbol{x}, k;\ H, \boldsymbol{b}) \right\rangle_{\rho} \tag{B.5}$$

となる．まず，$\boldsymbol{b} = \boldsymbol{b}^{(t)}$ を固定し, $H^{(t+1)}$ を求める．

$$\frac{\partial}{\partial h_1} Q_{\mathrm{II}}(H, \boldsymbol{b}^{(t)}) = \Big\langle \sum_{k=1}^{K} \Big[\frac{2h_1}{h_1^2 + h_2^2} - \frac{q_{\mathrm{II}\,k}^{(t)}(\boldsymbol{x})}{\sigma_k^2} \Big\{ h_1(x_1^2 + x_2^2) + (b_1^{(t)} - \mu_{j,1})x_1 + (b_2^{(t)} - \mu_{j,2})x_2 \Big\} \Big] \Big\rangle_{\rho} = 0, \tag{B.6}$$

$$\frac{\partial}{\partial h_2} Q_{\mathrm{II}}(H, \boldsymbol{b}^{(t)}) = \Big\langle \sum_{k=1}^{K} \Big[\frac{2h_2}{h_1^2 + h_2^2} - \frac{q_{\mathrm{II}\,j}^{(t)}(\boldsymbol{x})}{\sigma_j^2} \{ h_2(x_1^2 + x_2^2) + (b_1^{(t)} - \mu_{j,1})x_2 - (b_2^{(t)} - \mu_{j,2})x_1 \Big\} \Big] \Big\rangle_{\rho} = 0, \tag{B.7}$$

を整理すると

$$\frac{2h_1}{h_1^2 + h_2^2} Z_{\mathrm{II}} = h_1 X_{\mathrm{II}} - Y_{\mathrm{II}\,1}, \tag{B.8}$$

$$\frac{2h_2}{h_1^2 + h_2^2} Z_{\mathrm{II}} = h_2 X_{\mathrm{II}} - Y_{\mathrm{II}\,2} \tag{B.9}$$

となる．ただし，$X_{\mathrm{II}}, Y_{\mathrm{II}\,1}, Y_{\mathrm{II}\,2}, Z_{\mathrm{II}}$ はそれぞれ式 (5.21), (5.22), (5.23), (5.24) で与えられる．第一式の各辺を第二式の各辺で割り，h_1 または h_2 を消去すると，

$$h_2 = \frac{Y_{\mathrm{II}\,2}}{Y_{\mathrm{II}\,1}} h_1 \tag{B.10}$$

なる関係が得られ，h_2 を消去すると連立方程式は 2 次方程式に帰着され, 二つの解が得られる．

得られた二つの解を $Q_{\mathrm{II}}(H, \boldsymbol{b}^{(t)})$ の式に代入して整理すると

$$\log\{(1 \pm C_1)^2\} - C_2(1 \mp C_1)^2 + \text{定数} \tag{B.11}$$

という形に書ける (複号同順, $C_1, C_2 > 0$)．各項を比較することにより尤度が大きいのは式 (5.19) のときであることが言える．

次に $H = H^{(t+1)}$ を固定して $\boldsymbol{b}^{(t+1)}$ を求める．

$$\frac{\partial}{\partial b_1} Q_{\mathrm{II}}(H^{(t+1)}, \boldsymbol{b}) = \Big\langle -\sum_{k=1}^{K} \frac{q_{\mathrm{II}\,j}^{(t)}(\boldsymbol{x})}{\sigma_j^2} (h_1^{(t+1)} x_1 + h_2^{(t+1)} x_2 + b_1 - \mu_{j,1}) \Big\rangle_{\rho} = 0 \tag{B.12}$$

という一次方程式を解けば $b_1^{(t+1)}$ に関する更新式 (5.25) が得られ, 同様にして $b_2^{(t+1)}$ に関する更新式 (5.26) が得られる.