音声言語情報処理　44－6
(2002. 12. 19)

[招待論文] パターン認識における特異モデルの役割について

渡辺　澄夫†

† 東京工業大学　精密工学研究所　〒226-8503 横浜市緑区長津田 4259
E-mail: †swatanab@pi.titech.ac.jp

あらまし　この論文では「特異モデル」の概念を紹介し、それがパターン認識において果たす役割について考察する。外界から直接には働きを定められていない部分（隠れた部分）を持つ学習モデルは、特異モデルである。特異モデルにおいては、パラメータの集合から確率分布の集合への対応が1対1でなく、統計的正則モデルの設計法は利用できない。特異モデルがパターン認識で役立つための学習理論と学習アルゴリズムについて述べる。
キーワード　特異モデル、隠れマルコフモデル、パターン認識

Singular Learning Machines in Pattern Recognition

Sumio WATANABE†

† PI Lab., Tokyo Institute of Technology　4259 Nagatsuda, Midori-ku, Yokohama 226-8503 Japan
E-mail: †swatanab@pi.titech.ac.jp

Abstract　A lot of learning machines such as hidden Markov models, layered neural networks, gaussian mixtures, reduced rank regressions, and Boltzmann machines are singular statistical models. The learning theory of regular statistical models can not be applied to them. In this paper we discuss the role of singular statistical models in pattern recognition.
Key words　Singular statistical models, hidden Markov models, pattern recognition.

1. まえがき

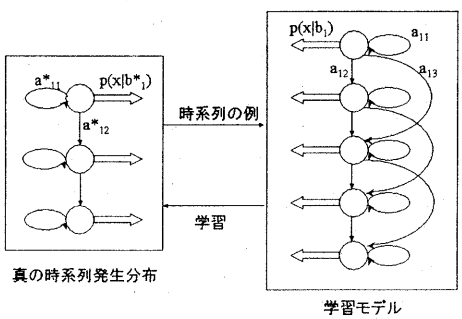

図1　隠れマルコフモデル（左：真）（右：学習）

この論文では、特異モデルの概念を説明し、パターン認識の中で果たす役割について考察する(注1)。

(注1)：この文章は,2002 年 12 月 19,20 日に東京大学で開催された電子情報通信学会・日本音響学会による第4回音声言語シンポジウムの講演の予稿である。

図1は、隠れマルコフモデルを描いたものである。隠れマルコフモデルは時系列 $x=(x_{(1)},x_{(2)},...,x_{(T)})$ を発生する確率モデルである。本論では簡単のため T は有限の定数であるとする。各 $x_{(t)}$ は、N 次元ユークリッド空間 R^N の元とする（離散集合の元を考える場合もあるが本論では扱わない）。このとき時系列 x は R^{NT} の元であると考えることができる。図1の左側の隠れマルコフモデルから、時系列 x が次の確率法則に従って発生しているとしよう。

(1) 時系列 x の背後に、M 個の状態の集合 s_1,s_2,...,s_M があって、状態間の時間推移はマルコフ確率過程に従っているとする。状態 s_i から s_j への推移確率は a_{ij}^* であるとする。また時刻 $t=1$ における状態は s_1 とする。

(2) 各時刻 t において、状態が s_i であるときに $x_{(t)}$ が得られる確率密度はパラメータ b_i^* により定まる $p(x_{(t)}|b_i^*)$ であるとする。この条件つき確率は、例えば、b_i^* をパラメータとする混合正規分布などによって決められているとする。

このようにして時系列 $x=(x_{(1)},x_{(2)},...,x_{(T)})$ を発生する確率モデルを考え、隠れマルコフモデルという。時系列 x の確率分布はパラメータ $w^*=(\{a_{ij}^*\},\{b_i^*\})$ が決まれば一意に定まるので、これを $p(x|w^*)$ と書くことにする。

さて図1の右側は学習を行う隠れマルコフモデルであるが、最

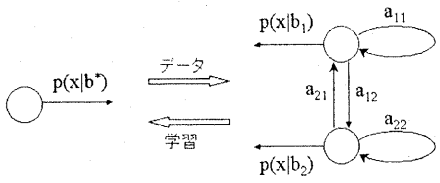

図 2　隠れマルコフモデルの冗長性

適化されるパラメータ w を持つので $p(x|w)$ と表記する。(図1の左側と右側は、厳密には構造が異なるので同じ表記を用いるのは望ましくないが、左側は、右側が特別なパラメータを持つ場合に相当するので、両方とも同じ表記を用いることにする)。隠れマルコフモデルは、時間的に変動する対象を認識する場合に極めて有用であることは周知の通りである。音声認識や動作認識などに広く利用され、実用上で使われているものも多い。

[注意点 1]　実世界の時系列が有限の大きさの隠れマルコフモデルから発生していることはめったに起こるものではない。多くの場合、真の分布は構造も法則も未知のものであるが、説明の都合上、しばらく図1のような場合を考える。真の分布が未知なものである場合に起こる問題は後で考察する。

さて、図1の右側の学習モデルのパラメータ $w = \{a_{ij}, b_i\}$ が取りうる値全体の集合を W と書くと、これは高次元のユークリッド空間 R^d (あるいはその部分集合) であると考えることができる。隠れマルコフモデル $p(x|w)$ は、パラメータの集合 $W \subset R^d$ から確率密度関数への写像 $w \mapsto p(\cdot|w)$ を定義している。一般に、この写像が1対1であるときには、すなわち、

$$p(x|w_1) = p(x|w_2) \quad (\forall x) \Longrightarrow w_1 = w_2$$

が成り立つときには、その学習モデル $p(x|w)$ は **特定可能** (識別可能) と呼ばれる。確率統計の教科書に紹介される確率モデルは特定可能であるものが多いが、私たちが情報システムを作るときに用いる学習モデルは特定可能でないものがほとんどである。例えば、隠れマルコフモデルは特定可能ではない。実際、図1の左側と右側のモデルが一致するパラメータの集合

$$W^* = \{w \in W \,;\, p(x|w^*) = p(x|w) \ (\forall x)\} \subset W$$

の要素は1個ではなく、多くの多様体の和集合で表されるものになっている (その集合は複雑な特異点も含んでいる)。このように、隠れマルコフモデルにおいては、考察している学習モデルよりも簡単な構造で表すことができるモデルに対応するパラメータは1個ではなく、次元を持つ広がりになっている。

[具体例 1]　真の分布と学習モデルとして図2の左側と右側を考える。学習モデルの状態推移確率は $\{a_{ij}, i = 1,2, j = 1,2\}$ である。確率の和が1であることから、条件 $a_{11} + a_{21} = 1$, $a_{12} + a_{22} = 1$ が必要である。状態から出力への条件つき確率 $p(x|b)$ は特定可能であると仮定する(注2)。もしも右側の学習モデルが

「$a_{12} = 0, b_1 = b^*$ かつ $\{a_{22}, a_{21}, b_2\}$ は自由」

あるいは「$b_1 = b_2 = b^*$ かつ $\{a_{i,j}\}$ は自由」

であるときには、図2の左側と右側は一致する。

[具体例 2]　特定可能な学習モデルの例をあげる。区間 $[0, 2\pi]$ に値をとる実数 x から実数 y への条件つき確率で表される確率モデル

$$y = \sum_{k=1}^{K} a_k \cos(kx) + \sum_{k=1}^{K} b_k \sin(kx) + \text{正規雑音} \qquad (1)$$

は特定可能である。このモデルが特定可能であることは、$\{\cos(kx), \sin(kx)\}$ が $[0, 2\pi]$ 上で一次独立であることから明らかであろう。

特定可能でない学習モデルの例としては、隠れマルコフモデルのほかに、**多層パーセプトロン、混合正規分布、ベイズネットワーク、ボルツマンマシン、縮小ランク回帰モデル**などがある。一方、特定可能な学習モデルとしては、**正規分布・指数分布** や **多項式回帰** などがある。学習モデルが特定可能であるかどうかは、学習アルゴリズムを考える場合に、どのような影響を及ぼすのだろうか。

2. カルバック情報量

時系列 $x = (x_{(1)}, x_{(2)}, ..., x_{(T)})$ がユークリッド空間 R^{NT} の元であるとすると、x の密度関数は普通の意味で定義できる(注3)。ふたつの密度関数 $q(x)$ と $p(x)$ の **カルバック距離** を

$$D(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

と定義する。ここで $\int dx$ は NT 次元のユークリッド空間での積分を表している。カルバック距離には次の二つの性質がある。
(1) 任意の $q(x), p(x)$ について $D(q||p) \geqq 0$.
(2) $D(q||p) = 0$ ならば $q(x) = p(x) \quad (\forall x)$.
このことから、カルバック距離を用いて確率密度関数の集合に距離を導入することができる。(ただし $D(q||p) = D(p||q)$ は成り立つとは限らないので、数学的な意味での距離にはならない)。同様に学習モデル $p(x|w)$ において、パラメータの集合 W 上に次のように距離を入れることができる。w^* から w までの距離を

$$D(w^*||w) = D(\, p(\cdot|w^*) \,||\, p(\cdot||w)\,)$$

と定義する。特定可能なモデルでは「$D(w^*||w) = 0$ ならば $w^* = w$ 」が成り立つが、特定可能でない学習モデルでは集合

$$W^* = \{w \in W; D(w^*||w) = 0\}$$

(注2)：もしも $p(x|w)$ に混合正規分布や神経回路網が使われていれば、このモデルも特定可能でないので、事態はより複雑になる。

(注3)：ルベーグ測度に対して絶対連続な確率変数を考えている。任意の有限長の時系列の確率測度を作ることも可能であり、それが無矛盾であれば、無限の長さの時系列の確率測度に拡張することができるが、本論では有限次元の場合のみを考える。

は1点ではない。情報学で現れる多くの学習モデルでは $D(w^*||w)$ は多変数 w の解析関数（テーラー展開が絶対収束する関数）であることが多い。本論でも $D(w^*||w)$ は w の関数として解析的であるとする。一般に多変数 w の解析関数の零点全体の集合のことを**解析的集合** という。また w の多項式の零点全体の集合のことを **代数多様体** という。解析的集合や代数多様体は極めて複雑な特異点を持つことが知られている。隠れマルコフモデルや神経回路網も、この意味での特異点を非常に多く持っている。

[注意点 2] 任意の w^* について $D(w^*||w)=0$ となる w が $w=w^*$ だけであり、さらに (i,j) 成分

$$I_{ij}(w^*)=\frac{\partial^2}{\partial w_i \partial w_j} D(w^*||w)|_{w=w^*}$$

を持つ行列 $I(w^*)$ が任意の w^* において正定値であるとき、そのモデルを **統計的正則モデル** という（厳密には、もう少し付帯条件がつく）。行列 $I(w^*)$ を **フィッシャー情報行列** という。特異モデルという言葉は、まだ数学的に固定した定義が与えられていないが、本論では $\det I(w^*)=0$ となる w^* が存在するモデルは全て **特異モデル** と呼ぶことにする。

[注意点 3] もしも真の解の集合

$$W^*=\{w\in W; D(w^*||w)=0\}$$

が多様体ならば（W^* の任意の点の近傍で局所座標が取れるならば）、

W の局所座標として、W^* の接平面の座標と、それに直交する座標を取ることができ、数学的には比較的容易に議論を行うことができる。しかしながらほとんど全ての特定不能な学習モデルでは W^* は多様体にならない。つまり局所座標がとれない点が W^* の中に存在する。このような点を W^* の**特異点**という。読者は図2のような簡単な場合でさえ、特異点が存在することを確認せよ。

3. 学習理論とは

以上で特異モデルの概念を説明した。次に学習理論の概要を紹介する。

3.1 学習アルゴリズム

未知の確率密度関数 $q(x)$ に従う時系列 $x\in R^{NT}$ の独立な例として、$x_1,x_2,...,x_n$ が得られたとする。(x_i それぞれが R^{NT} の元である)。学習モデル $p(x|w)$ $(w\in W)$ が

$$D_n=(x_1,x_2,...,x_n)$$

を元にして、$q(x)$ を知ろうとするとき、どのくらいのことが分かるのだろうか。また、できるだけ正確に $q(x)$ についての情報を得るためには、どのようなアルゴリズムを用いるべきだろうか。学習理論は、このような疑問に答えようとするものである。まず、代表的な3つの学習アルゴリズムについて述べる。

3.1.1 最尤推定法

パラメータ w を持つ学習モデル $p(x|w)$ が学習を行うとき

$$L(w)=\prod_{i=1}^{n} p(x_i|w)$$

を最大にするパラメータ w^{ML} が存在するならば、それを見つけだして、$p(x|w^{ML})$ が真の分布 $q(x)$ に近いだろうと考える方法を**最尤推定法**という。w^{ML} を**最尤推定量**という。

3.1.2 MAP 法

パラメータ空間 W 上の事前確率分布 $\varphi(w)$ を用意して、事後確率分布

$$p(w|D_n)=\frac{1}{Z(D_n)}\prod_{i=1}^{n} p(x_i|w)\,\varphi(w)$$

を考え（$Z(D_n)$ は正規化定数)、これを最大にするパラメータ w^{MAP} が存在するならばそれを見出して、$p(x|w^{MAP})$ が本当の確率分布 $q(x)$ に近いだろうと推測する方法を**事後確率最大化法 (MAP 法)** という。

3.1.3 ベイズ法

事前確率分布 $\varphi(w)$ を容易して、事後確率分布 $p(w|D_n)$ を用いてベイズ予測分布

$$p(x|D_n)=\int p(x|w)\,p(w|D_n)\,dw$$

を作り出し、これが本当の分布 $q(x)$ に近いと推測する方法をベイズ法という。

3.2 汎化誤差と学習誤差

学習アルゴリズムは上記の3個に限るわけではなく、これらの方法が改良された様々なバリエーションがある。与えられたデータ D_n を元に、ある学習アルゴリズムによって得られた学習結果を $p_n(x)$ と書く。学習理論では、まず最初に、汎化誤差と学習誤差とが検討される。汎化誤差は $q(x)$ から $p_n(x)$ までのカルバック距離によって定義される。

$$G(D_n)=D(q||p_n)$$

汎化誤差はパターン認識におけるテスト時の認識誤り率や時系列予測の予測誤差と対応する。学習理論の重要な目的のひとつは、汎化誤差を小さくする学習アルゴリズムを作りだすことである。

学習誤差は $q(x)$ から $p_n(x)$ までの経験カルバック距離によって定義される。

$$T(D_n)=\frac{1}{n}\sum_{i=1}^{n}\log\frac{q(x_i)}{p_n(x_i)}$$

学習誤差は学習時の認識誤り率や学習データの誤差と対応する。汎化誤差と学習誤差は、ともにデータ D_n に依存する。真の分布のエントロピーと経験エントロピーをそれぞれ

$$S=-\int q(x)\log q(x)dx$$

$$S_n=-\frac{1}{n}\sum_{i=1}^{n}\log q(x_i)$$

と定義すれば、定義から汎化誤差と平均対数尤度、学習誤差と

経験対数尤度の間には次が成立する。

$$G(D_n) = -S - \int q(x) \log p_n(x) dx$$

$$T(D_n) = -S_n - \frac{1}{n} \sum_{i=1}^{n} \log p_n(x_i)$$

S, S_n は学習モデルに依存しないので、$G(D_n)$ が小さいことと平均対数尤度が大きいことは等価であり、$T(D_n)$ が小さいことと経験対数尤度が大きいことは等価である。

3.3 統計的正則モデルでは

学習データ Dn は、データを得る度にばらつくものである。学習データの出方について平均を取る操作を E と書くことにする。統計的正則モデルの場合には、真の分布が学習モデルに含まれているならば、最尤法、MAP 法、ベイズ法のどの方法であっても、n が大きくなるときに

$$E[G(D_n)] = \frac{d}{2n} + o(\frac{1}{n})$$

$$E[T(D_n)] = -\frac{d}{2n} + o(\frac{1}{n})$$

が成り立つことが知られている(注4)。ここで d はパラメータ空間 W の次元である。このことから、平均対数尤度と経験対数尤度の間に

$$E[\int q(x) \log p_n(x) dx] = E[\frac{1}{n} \sum_{i=1}^{n} \log p_n(x_i)] - \frac{d}{n} + o(\frac{1}{n})$$

という関係が導かれる。つまり、次の式の両辺の値は確率的にばらつくのであるが、その平均値は、ほぼ同じになる。

$$\int q(x) \log p_n(x) dx \approx \frac{1}{n} \sum_{i=1}^{n} \log p_n(x_i) - \frac{d}{n}$$

複数個の学習モデルの候補が与えられたときに、左辺の値を小さくするモデルを選ぶことを目的として、右辺の値を最小にするモデルを選ぶ方法が **AIC** である。AIC は、本来は交換できない「平均 E を取る操作」と「複数のモデルについての最小化の操作」の順番を入れ換えることによって得られるものであり、このため AIC は、サンプル数が幾ら大きくなっても真の分布を選ぶ確率が 1 に近づかないという問題点を有している（一致性を持たない)。AIC の持つこの問題点に対するひとつの解は、確率的複雑さ

$$F(D_n) = -\log Z(D_n) = -\log \int \prod_{i=1}^{n} p(X_i|w)\, \varphi(w)\, dw$$

を最小にするモデルを選ぶ方法である。ここで $Z(D_n)$ は、データ D_n が与えられたときの「学習モデル $p(x|w)$ と事前分布 $\varphi(w)$ の尤度」に等しいので、確率的複雑さの最小化は、「モデルと事前分布に関する最尤法」になっている。統計的正則モデルでは $n \to \infty$ において

$$E[F(D_n)] = nS + \frac{d}{2} \log n + O(1)$$

が成り立ち、この性質から **BIC** が導出される。こちらの規準は一致性を持っているが、汎化誤差を最小化するために最適な規準ではない。(AIC も汎化誤差最小化の最適な規準ではなく、汎化誤差を最小にする最適な規準は一般には存在しない。)

[注意点 4] 真の分布が有限の大きさの学習モデルに含まれているとき、サンプルの増大に伴って、1 に近づく確率で真の分布が選ばれるモデル選択アルゴリズムをモデル選択において一致性を持つ という。また、真の分布が有限の大きさの学習モデルに含まれている、いないに関わらず、最も汎化誤差の平均値を小さくするモデル選択アルゴリズムをモデル選択における有効性を持つ という。統計的正則モデルにおいても、両方の性質を持つモデル選択アルゴリズムは構成できない。特異モデルにおいては、一致性と有効性の相違は、統計的正則モデルのときよりも、ずっと大きくなる [11]。

4. 特異モデルの理論と方法

以下では、特異モデルについて、これまでに解明されていることを紹介する。

4.1 真の分布が学習モデルに含まれているとき

まず真の分布が学習モデルに含まれている場合に起こる現象を説明する。

4.1.1 最 尤 法

特異モデルでは、一般に最尤推定量の存在は保証されない。例えば、混合正規分布で各正規分布の分散まで推測する場合には、最尤推定量は存在しない。ニューラルネットや隠れマルコフモデルでは、最尤推定量が存在するかどうかわかっていない。

特異モデルでは、最尤推定の学習誤差は、統計的正則モデルよりも遥かに小さくなる。Hartigan は、混合正規分布で分散の推測をしない場合でも最尤推測について

$$\lim_{n \to \infty} E[T(D_n)] \times n = -\infty$$

になることを示している [4]。また萩原は、3 層パーセプトロンで、真の分布よりも中間ユニットが 2 個以上多いモデルでは学習誤差が

$$-c_1 \frac{\log n}{n} \leqq E[T(D_n)] \leqq -c_2 \frac{\log n}{n}$$

になることを示している $(c_1, c_2 > 0)$ [3]。

特異モデルの最尤推定における汎化誤差は解明されていないが、統計的正則モデルよりも遥かに大きくなると予想されている。すなわち

$$E[G(D_n)] \geqq c_3 \frac{\log n}{n}$$

と予想されている $(c_3 > 0)$。つまり特異モデルに最尤推測を適用すると学習誤差は小さくなるが、反対に予測は外れる。

最尤推定におけるこの発散は、パラメータが無限大になるときに、カルバック情報量の解析性が失われることが原因である。実際、パラメータの解析性が成り立つモデルならば、特異モデルであっても、学習誤差と汎化誤差について

$$0 \leqq E[G(D_n)] \leqq \frac{c_4}{n}$$

(注4)：最尤法、MAP 法、ベイズ法のどの場合でも、同じ結果になるのは、統計的正則モデルの時だけ成り立つ特別の性質である。

$$0 \geqq E[T(D_n)] \geqq -\frac{c_5}{n}$$

を証明することができる（$c_4, c_5 > 0$）[15]。

統計学や情報学において最尤法は基本的なものであり、最尤法の性質を解明することは理論的に重要な課題であるが、最尤法は、特異モデルにおいては、予測にもモデル選択にも仮説検定にも有用ではないように思われる。

[注意点 5] 特異モデルにおいて、実用に使うことができるパラメータは最尤推定量とはまったく別の何かである。例えば、ニューラルネットに最急降下法を適用し適度な回数で停止した場合、比較的良好な予測ができることが知られているが、そのときのパラメータは最尤推定量とはまったく異なる。また、混合正規分布では最尤推定量は存在しないので、「何か良好な性質を持つもの」をうまく見出す必要がある。そのために EM アルゴリズムにおける初期値の設定や途中のパラメータの人工的な操作が工夫されている。隠れマルコフモデルにおいても、初期値をうまく決める、EM アルゴリズムをうまく止める、幾つかのパラメータは人手で決めるなどのアドホックな工夫が用いられることが多い。

4.1.2 ベイズ法

ベイズ法は、特異モデルの学習法として、今のところ最も有力である。またその汎化誤差は数学的に解明されている。それは

$$E[G(D_n)] = \frac{\lambda}{n} + o(\frac{1}{n})$$

である。ここで定数 λ は、次のように定まる有理数である [12]。カルバック情報量 $D(w^*||w)$ と事前分布 $\varphi(w)$ から定義されるゼータ関数

$$\zeta(z) = \int D(w^*||w)^z \varphi(w) dw$$

は 1 変数 z の複素関数であるが、これは複素平面全体に解析接続できて有理型関数になり、その極は全て実数であり負の有理数である。$(-\lambda)$ は、その極の中で最も原点に近いものと一致する。$\varphi(w) > 0$ が成り立てば、λ は $d/2$ よりも遥かに小さい(注5)。ベイズ推測の学習誤差は解明されていないが、汎化誤差との対称性は一般に成立しないことが知られている。すなわち $E[T(D_n)] \neq -E[G(D_n)] + o(1/n)$ である。真の分布が特異点からずれている場合に生ずる現象については [17] をみよ。

ベイズ法の課題として、よく研究されているのは、(1) 事前分布の設計問題と、(2) 事後分布の実現アルゴリズムである。このうち、(1) 事前分布の設計については、確率的複雑さ $F(D_n)$ を小さくする方法が有用であり広く使われている [1]。(2) また事後分布の実現アルゴリズムとして中心的に研究されてきた MCMC 法は、従来は計算量が多いという課題を有していたが、計算機の発展に伴って実現が容易になり、現在では広く用いられている。ベイズ法は、最尤法やそのアドホックな改良よりも優れた予測結果を与えることが期待できるので、音声認識にお

(注5)：この定理の背後には、代数幾何 [5] や代数解析 [7] などの美しい一群の数学的構造があるが、本論の話題から外れるためにここでは省略する。

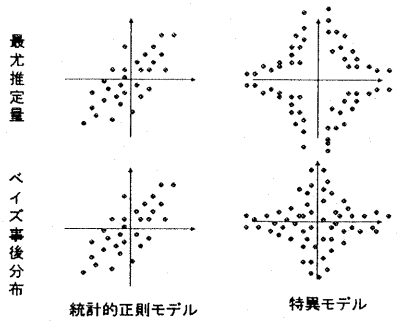

図 3 最尤推定量とベイズ事後分布

いても研究する価値があると思われる。隠れマルコフモデルと神経回路網の性能比較も、ベイズ法を適用するという条件下でなければ、意味のある比較とは言えないであろう。

[注意点 6] 特異モデルにおいては、サンプル数が増えても、ベイズ事後確率は正規分布に近づかないので、最尤推定量あるいは MAP 推定量を求めておいてから、その近傍で正規分布近似するという方法 [8] ではうまくいかない。また特異モデルでは、MAP 法とベイズ法はまったく異なる予測精度を持つ。MAP 法は最尤法よりは良い予測精度を持つが、ベイズ法には遠く及ばないので MAP 法をベイズ法の代用と考えることはできない。

[注意点 7] 座標不変な事前分布としてジェフリーズの事前分布が知られている。それは、フィッシャー情報行列 $I(w)$ を用いて

$$\varphi(w) = \frac{1}{C}(\det I(w))^{1/2}$$

と定義される。定義より、フィッシャー情報行列の行列式が 0 になる点、つまり特異点では、ジェフリーズの事前分布は 0 になる。この事前分布は統計的正則モデルにおけるリスクの Minimax 規準と関連してよく研究されているが、統計的正則モデルの場合でも、良い予測を与える事前分布ではない。特異モデルにジェフリーズ分布を適用すると、$\lambda \geqq d/2$ が成り立つ [13]。モデル選択のときには役立つ場合があるが、予測精度の点では適さない [6]。優れた予測精度を与える学習アルゴリズムは、座標不変性と両立しないようである。

4.2 最尤法とベイズ法では何が違うのか

上記で、特異モデルでは最尤法は適さず、ベイズ法は適するということを述べたが、それはなぜだろうか。ここでは例を用いて学習法の違いを説明してみよう。

[具体例 3] 統計的正則モデル

$$y = ax + b + \text{正規雑音}$$

を用いて、真の分布 $y = 0 +$ 雑音 を学習すると、最尤推定量の分布と事後確率から得られるパラメータの分布は図 3 のような正規分布になる（入力 x が従う分布によって形は変わる）。最尤推定量の分布も事後確率分布も実質的にはそれほど違わない。

[具体例 4] 特異モデル

$$y = a\tanh(bx) + \text{正規雑音}$$

を用いて、真の分布 $y = 0 +$ 雑音 を学習すると、最尤推定量の分布と事後確率から得られるパラメータの分布は図3のように異なる。最尤推測では、尤度関数を大きくするために、真のパラメータ集合 $\{a = 0\} \cup \{b = 0\}$ の近くにパラメータが来る確率は0になってしまう。一方、ベイズ法では、真のパラメータ集合の近くにパラメータが来る確率は大きくなる。

[注意点 8] 事後確率分布を「経験対数尤度に関するボルツマン分布」であると考えると、上記の違いは、特異点が持つエントロピーの問題であると説明することができる。

4.3 真の分布が学習モデルに含まれていないとき

実問題においては、真の分布が有限の大きさの学習モデルに完全に含まれていることは起こりにくい。その場合に、特異モデルが役立つかどうかについて考えよう。

隠れマルコフモデルにおいて、真の分布 $q(x)$ が完全にわかっているとして、状態の数 M を大きくしてゆくことによって近似しよう。

$$q(x) \cong p_M(x) \equiv p(x|\{a_{ij}^*, b_i^*; i, j = 1, 2, ..., M\}) \qquad (2)$$

完全な近似のためには M を無限に大きくとらなくてはならないが、M が大きくなるにつれて、上の近似の誤差（確率分布の関数近似誤差）は、急速に小さくなってゆく(注6)。

一方、学習時においては $q(x)$ は不明である。また学習例数 n は非常に多い場合であっても有限であることには変わりなく、学習データのばらつきのために、真の分布についてある程度の「解像度」までしか推測できない（統計的推定誤差のオーダーは n で定まる）。与えられた学習データのもとで、真の分布を最もよく捕まえることができるモデルは何であるか、という観点から M を定めようとすると、関数近似誤差と統計的推定誤差の和を小さくすることを考えなくてはならない。この問題は、学習理論において基本的なものであり、「関数近似誤差（バイアス）と統計的推測誤差（バリアンス）」の問題と呼ばれている。

特異モデルは関数近似誤差において統計的正則モデルよりも優れているが、ベイズ法を用いて学習すれば、ほとんど特定可能でない状態では統計的推定誤差においても優れている。音声や画像などのように高次元空間に複雑な広がりを持つ対象の認識では、真の分布を有限の大きさの単純なモデルで表現することはできないと考えられるから、そのような場合には「特異モデル＋ベイズ法」が有用であろう。

4.4 具体的な学習モデルについて

本論文では、音声認識の研究者の皆様に興味を持っていただくために、特異モデルの代表として隠れマルコフモデルを紹介したが、隠れマルコフモデルを特異モデルの観点から解析した研究はまだ存在しない。神経回路網については [14]、混合正規分布については [19]、縮小ランク回帰モデルについては [10]、ベイズネットワークについては [9] に、それぞれ研究がある。また特異モデルの選択の問題は [16] [6] で研究されている。

(注6)：特異モデルでは、隠れた部分を外界に適応させることにより、統計的正則モデルよりも、遥かに良い関数近似精度を持つ。

5. ま と め

隠れマルコフモデルを例として特異モデルの概念を紹介し、その特徴を説明した。本論文では、パターン認識への応用を中心的に考えた。数理的な基礎に興味のある方は参考文献を御覧ください。

この研究は科学研究費補助金 12680370 の援助を受けた。

文 献

[1] Akaike, H. (1980). Likelihood and Bayes procedure. In J.M. Bernald, *Bayesian Statistics*, (pp.143-166). Valencia, Spain: University Press.

[2] 甘利 俊一, 尾関 智子, 朴 慧暎 (2002) 階層モデルにおける学習と推論-特異構造を持つ統計モデル-. 電子情報通信学会誌,Vol.D-II-85J,No.5,pp.701-708.

[3] Hagiwara, K., Kuno, K., & Usui, S. (2000) On the problem in model selection of neural network regression in overrealizable scenario. *Proc. of Int. Joint Conf. on Neural Netork*, Italy, Como.

[4] Hartigan, J.A. (1985) A Failure of likelihood asymptotics for normal mixtures. *Proceedings of the Berkeley Conference in Honor of J.Neyman and J.Kiefer*, Vol.2, pp.807-810.

[5] Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, Vol.79, pp.109-326.

[6] 西上功一郎、渡辺澄夫 (2003) 特異な学習モデルの選択における事前分布の影響について. 電子情報通信学会誌, Vol.86J-D-II, No.1, to appear.

[7] Kashiwara, M.(1976) B-functions and holonomic systems. *Inventiones Mathematicae*, Vol.38, pp.33-53.

[8] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, Vol.6, No.2, 461-464.

[9] Rusakov, D, Geiger,D.(2002) Asymptotic model selection for naive Bayesian networks. *Proc. of UAI02.*

[10] 渡辺一帆、渡辺澄夫 (2003) 縮小ランク回帰モデルのベイズ汎化誤差について. 電子情報通信学会誌 Vol.86J-A,No.3. to appear.

[11] 渡辺澄夫, "データ学習アルゴリズム," 共立出版,2001.

[12] Watanabe, S.,(2001) Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13,(4), pp.899-933.

[13] Watanabe, S. (2001) Algebraic information geometry for learning machines with singularities. *Advances in Neural Information Processing Systems*, Vol.13, 329-336.

[14] Watanabe, S. (2001) Algebraic geometrical methods for hierarchical learning machines. International Journal of Neural Networks, Vol.14, No.8, 1049-1060.

[15] 渡辺澄夫 (2001) 代数的な特異点を持つ学習モデルの学習誤差と汎化誤差. 電子情報通信学会誌, Vol.J84-A, No.1, pp.99-108.

[16] 渡辺澄夫 (2001) 特異点を持つモデルと事前分布の代数幾何. 人工知能学会誌, Vol.16, No.2,pp.308-315.

[17] Watanabe,S. Amari, S. (2003) The effect of singularities when the true parameters do not lie on such singularities. *Advances in Neural Information Processing Systems*, Vol.15, to appear.

[18] 山崎啓介、渡辺澄夫 (2002) 特異点を持つ推論モデルの学習曲線の確率的計算法. 電子情報通信学会誌, Vol.J85-D-II, No.3,pp.363-372.

[19] Yamazaki, K. Watanabe, S. (2002) Resolution of singularities in mixture models and its upper bounds of the stochastic complexity.*ICONIP2002.*