特異モデルにおける統計的推測
– 接錐によるアプローチ –

福水 健次　　栗木 哲
統計数理研究所*

平成 15 年 10 月 1 日

概 要

パラメトリックな統計モデルや学習機械を関数空間の中で考えたときに，滑らかでない点を持つようなモデルをここでは特異モデルと呼ぶ．多層パーセプトロン，ミクスチャモデルや ARMA，HMM などがその例となっている．このようなモデルでは，真のパラメータがその特異点にあるとき，推定や学習にさまざまな複雑で興味深い現象がみられる．本論文では，特異モデルに生じる統計的現象とその理論的解析について，接錐による見方を中心に述べる．

1　はじめに

ニューラルネットワークなどの機械学習，および統計学の分野では，有限次元のパラメータを持ったモデルを用意し，与えられたデータをよく説明するようなパラメータを求めて，それによる推論や予測を行うことが多い．統計学的にいうと，これはパラメトリックモデルを用いた統計的推論に他ならない．多くの問題においては，システムにはノイズなど確率的要素が含まれるため，パラメトリックモデルはデータを発生させる確率構造を確率密度関数族によってモデル化する．簡単な例として多層パーセプトロン

$$\psi(x;\theta) = \sum_{j=1}^{H} c_j \varphi(a_j^T x + b_j) + d \tag{1}$$

を考えよう．ここで $\theta = (a_j, b_j, c_j, d) \in \mathbb{R}^{3m+1}$ がパラメータであり，$\varphi(t)$ はロジスティック関数とする．データ $(x_1, y_1), \ldots, (x_n, y_n)$ が与えられたときに，$\psi(x_i, \theta)$ が y_i を近似するように θ を選ぶことがニューラルネットの学習であるが，この学習の手続きは統計モデルを用いて書くことができる．例えば，最小２乗誤差学習

$$\min_{\theta} \sum_{i=1}^{n} (y_i - \psi(x_i;\theta))^2$$

は，$\psi(x;\theta)$ に対してガウスノイズが加わるモデルを仮定し，x が与えられたときの y の条件付確率

$$f(y|x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y - \psi(x;\theta))^2\right\} \tag{2}$$

により統計モデルを導入すると，後述する最尤推定の枠組みで論じることができる．

* 〒 106-8569 東京都港区南麻布 4-6-7．{fukumizu,kuriki}@ism.ac.jp

一般にパラメトリックな統計モデル $\{f(x|\theta)\}$ は，確率密度関数全体の関数空間の中のある部分集合を成している．本論文は，関数空間に埋め込まれた集合を「統計モデル」ととらえ，統計モデルの中に滑らかでない点が存在するとき，統計的推論の結果がどのようになるかを論じる．確率密度関数 $f(x|\theta)$ は θ に関して滑らかな関数であることが多いが，このことは必ずしも，モデルが関数空間の中で滑らかな点だけを持つことを意味しない．実際，ミクスチャモデル，ARMA，HMM といった重要なモデルが，滑らかでない点＝特異点を持っている

甘利らの論文 [10] が，特異モデルに関するさまざまな話題に関するよいまとめになっているので，本論文は，最尤推定の特異点における挙動に話題を絞り，特に接錐を用いた尤度比の解析を中心に解説する．統計的推定の方法には最尤推定とは考えを異にするベイズ推定もあるが，ベイズ推定における特異点の影響は，本号の渡辺の解説 [13] およびそのグループの一連の研究を参考にしていただきたい．

2　最尤推定と推定量の挙動

ここでは最尤推定量の統計的挙動について基本的な事項を復習する．本論文の主な関心は，ここで述べる通常の理論が成り立たないケースであるが，そのような特異な場合の議論も通常の理論が基礎になっている．

測度空間 $(\mathcal{X}, \mathcal{B}, \mu)$ に対し，$\mathbb{R}^m$ の部分集合 Θ をパラメータ空間に持つパラメトリックモデル $S = \{f(x|\theta) \mid \theta \in \Theta\}$ を考える．$\theta_0 \in \Theta$ を真のパラメータとして固定し，真の確率分布 $f(x|\theta_0)\mu$ から発生した n 個の独立なサンプル $X_1, \ldots, X_n$ を用いて，θ_0 を推定したい．パラメータ推定の一般的手法として**最尤推定**がある．最尤推定は**対数尤度**

$$\sum_{i=1}^{n} \log f(X_i|\theta) \tag{3}$$

の最大値を達成するパラメータ θ を推定量として用いる．これを**最尤推定量**と呼び，$\widehat{\theta}_n$ で表す．最尤推定量はサンプルに依存する確率変数であり，その分布の性質を調べることは重要な課題である．最尤推定量の分布を調べるために，本論文では，分布に関する仮定はあまりおかずにサンプル数 n が非常に大きいと仮定して一般的性質を導く統計的漸近理論のアプローチを論じる．

いくつかの正則条件を仮定すると，n が無限大に近づくとき最尤推定量 $\widehat{\theta}_n$ は

$$\sqrt{n}\big(\widehat{\theta}_n - \theta_0\big) \quad \Longrightarrow \quad N(0, I^{-1}(\theta_0)) \tag{4}$$

と法則収束することが知られている．ここで，$N(\mu, \Sigma)$ は平均 μ 分散共分散行列 Σ の正規分布を表し，$I(\theta)$ は，その (a, b) 要素が

$$I(\theta)_{ab} = \int \frac{\partial \log f(x \mid \theta)}{\partial \theta^a} \frac{\partial \log f(x \mid \theta)}{\partial \theta^b} f(x \mid \theta) \mu(dx)$$

で与えられる $m \times m$ 行列で，**Fisher 情報行列**と呼ばれる．(4) 式は，$\sqrt{n}$ のスケール変換のあと，最尤推定量の分布が真のパラメータを中心とした正規分布に収束することを示しており，この性質を最尤推定量の**漸近正規性**という．最尤推定量を考えるとき**尤度比**

$$\max_{\theta \in \Theta} L_n(\theta) = L_n(\widehat{\theta}), \qquad L_n(\theta) = \sum_{i=1}^{n} \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \tag{5}$$

が重要な役割を果たす．これは例えば尤度比検定を行うときに用いられ，漸近正規性が成り立つと，Taylor 展開による標準的な議論により

$$2L_n(\widehat{\theta}_n) \Longrightarrow \chi^2_m \tag{6}$$

（χ^2_m は自由度 m のカイ 2 乗分布）に法則収束することがわかる．

ここでは漸近正規性の正則条件についての精密な議論は行わないが，(4) 式を見るだけでも，いくつかの条件が必要であることに気づく．まず Fisher 情報行列は可逆でなければならない．また，正規分布を定義するためには，真のパラメータ θ_0 は Θ の内点になければならない．

以降で議論する特異モデルでは，以上述べた条件の多くが成り立たない．ある場合には，真のパラメータ θ_0 がパラメータ空間の境界上にあったり，他の例では Fisher 情報行列が逆行列を持たなかったりする．そのような場合には漸近分布が正規でなかったり，$\sqrt{n}$ のスケール変換では収束しなかったりするという，一見異常な現象が見られる．本論文では厳密な言葉の定義を行わず，このような状況を持つモデルのことを「特異モデル」と呼ぶことにする．

3　特異モデルと接錐

3.1　パラメータの識別不能性

特異モデルの典型的な例のひとつとして，パラメータの識別不能性を有限混合モデルの例によって説明する．**有限混合モデル**は，パラメータ a を持った確率密度関数 $p(x|a)$ に対して，

$$f(x|\theta) = \sum_{k=1}^{K} c_k p(x|a_k) \tag{7}$$

という密度関数族で定義される．ここで，$\{c_k\}$ は $\sum_{k=1}^{K} c_k = 1$ を満たす非負実数，モデルのパラメータは $\theta = (a_j, c_j)$ である．各コンポーネントが d 次元正規分布であれば，これは正規混合モデルと呼ばれ，クラスタリング手法としても用いられる．

さて，２個のコンポーネントを持つ１次元正規分布の混合モデルを考え，さらに簡単のため，分散はともに１に固定し，一方の平均パラメータは０であるとしよう．するとモデルは

$$f(x|\theta) = c\phi(x|\mu, 1) + (1-c)\phi(x|0, 1) \tag{8}$$

と表される．ここでパラメータ $\theta = (c, \mu)$ は $\Theta = [0,1] \times \mathbb{R}$ 内を動く．いま，真の確率分布が標準正規分布 $\phi(x|0,1)$ であると仮定し，(8) 式で定義されるモデル $\{f(x|\theta)\}$ の中で $\phi(x|0,1)$ がどのように表現されているか考えてみよう．単一の正規分布 $\phi(x|0,1)$ はコンポーネント数が１個の混合モデルとみなすことができるが，このように真の分布が設定したモデルサイズよりも小さいと仮定する状況は，モデルサイズの検定やモデル選択の問題に頻繁に出現する．今の場合，図 1 で示されるような１次元の連続集合

$$\Theta_0 = \{\theta \mid c = 0, \mu \in \mathbb{R}\} \cup \{\theta \mid \mu = 0, 0 \le c \le 1\}$$

を考えると，Θ_0 では全体が同一の分布 $\phi(x|0,1)$ を定めており，$\Theta - \Theta_0$ ではパラメータと確率分布が一対一に対応している．

この例のように，パラメータ空間 Θ の点 θ に対し，θ を含む１次元以上の Θ の部分多様体が存在し，その部分多様体上の任意の点が同一の確率分布を定めているとき，パラメータ θ は**（連続**

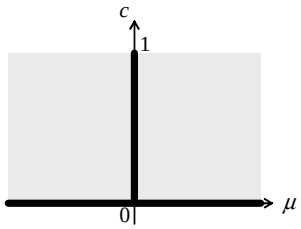

図 1: 正規混合モデル（(8) 式）における識別不能なパラメータ

的）識別不能であると呼ばれる．統計モデルが（連続的）識別不能なパラメータを持つと，漸近正規性は成立しない．同一の分布を定める方向への方向微分がゼロとなり，Fisher 情報行列が非可逆になるためである．ここでは正規分布の例を説明したが，全く同様の議論により，コンポーネントによらず，$(K+1)$ 個のコンポーネントを持つモデルのパラメータ空間の中で，K 個のコンポーネントで実現可能な密度関数を表すパラメータが識別不能になることも示される．また，多層パーセプトロンにおいても同様の識別不能性が生じることも容易に確認できよう．

3.2 パラメータ空間が境界を持つモデル

特異モデルのもうひとつの代表的な例は境界を持ったパラメータ空間である．これを単調回帰の例を通して説明する．

以下では最も簡単な回帰の例として，X_i が有限個の値 $\{1, 2, \ldots, H\}$ をとり，各 $k \in \{1, \ldots, H\}$ 上の平均パラメータ θ_k を用いて，サンプル $(X_1, Y_1), \ldots, (X_n, Y_n)$ が

$$X_i = k_i, \quad Y_i = \theta_{k_i} + Z_i, \quad (1 \leq i \leq n)$$

に従うモデルを考える．ここで $k_i \in \{1, \ldots, H\}$ は X_i が取る値であり，Z_i が平均 0 の独立な同一の正規分布に従うとし，さらに各 k に関して m 個のデータが採られたと仮定する．このとき，θ_k の最尤推定量は，k 上での Y のサンプル平均

$$\overline{Y}_{(k)} = \frac{1}{m} \sum_{i \in I_k} Y_i$$

となる．ここで $I_k = \{i \in \{1, \ldots, n\} \mid X_i = k\}$ である．

この回帰問題に対してさらにパラメータの単調性を仮定しよう．X の値が大きいほど，Y の値は大きくなりやすいことが分かっているとすると，この事前知識はパラメータ θ_k に対して

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_H \tag{9}$$

という制約の存在としてモデルに組み込むことが自然である．このような単調性を持った回帰は，例えば薬の効用を調べる問題などに現れる．X は投与した薬の量であり，1 単位から H 単位までがそれぞれ m 人の被験者に与えられ，薬の効用を示す値が Y_i として計測される．もし薬に全く効用がなかったとすると，$\theta_1 = \theta_2 = \cdots = \theta_H$ であるが，少しでも効用があると θ_k は k に対して単調に増加する．

ここではさらに問題を簡単にして，パラメータ θ_k はすべて非負と仮定して，θ_k の最尤推定量を求めてみよう．このとき，例えば $H = 2$ に対するパラメータ空間 Θ は図 2 のようになる．

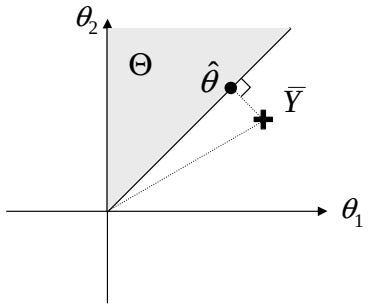

図 2: 制約のあるパラメータ空間

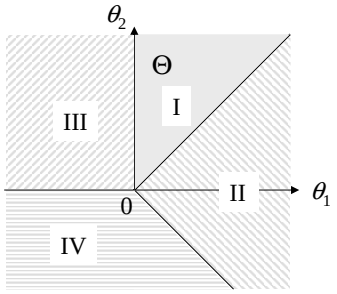

図 3: 単調回帰の最尤推定量を求めるためのパラメータ空間の分割

$\theta = (\theta_1, \ldots, \theta_H)$ の最尤推定量 $\widehat{\theta}$ は，容易に確かめられるように，制約のない場合の最尤推定量 $\overline{Y}_{(k)}$ を用いて

$$\min_{\theta \in \Theta} \sum_{k=1}^{H} \sum_{i \in I_k} (Y_i - \theta_k)^2 = \min_{\theta \in \Theta} \sum_{k=1}^{H} m(\overline{Y}_{(k)} - \theta_k)^2 + \text{Const.}$$

の解となる．したがって，最尤推定量 $\widehat{\theta}$ は，$\overline{Y} = (\overline{Y}_{(1)}, \ldots, \overline{Y}_{(H)})$ からユークリッド距離が最小になる Θ の点として与えられる（図 2）．

いま真のパラメータ θ_0 が $\theta_{0,1} = \cdots = \theta_{0,H} = 0$ を満たすと仮定しよう．すると，θ_0 はパラメータ空間 Θ の頂点に位置している（図 2）．$\overline{Y}$ は原点を中心とした正規分布に従うが，そこから最短距離にある Θ の点は，$\overline{Y}$ の位置によって特徴的に変化する．$H = 2$ の場合を考えると，図 3 のように $\mathbb{R}^2$ を 4 つの領域に分割したとき，$\overline{Y}$ が領域 I（パラメータ空間 Θ）にあれば $\widehat{\theta} = \overline{Y}$，領域 II,III にあればそれぞれにもっとも近い Θ の境界への射影が $\widehat{\theta}$ であり，領域 IV にあれば $\widehat{\theta} = 0$ である．この考察から，$\widehat{\theta}$ はあきらかに正規分布とは異なった分布を持っていることがわかる．この場合の最尤推定量に通常の漸近理論が適用できないことは，θ_0 のまわりでパラメータ空間がユークリッド的な開近傍を持っていないことからも窺い知れよう．

3.3 接錐

統計的推定の問題を考える際，推定量 $\widehat{\theta}$ 自身が重要なのではなく，確率密度関数 $f(x|\widehat{\theta})$ が重要なことが多い．最尤推定量の定義から，$f(x|\widehat{\theta})$ はパラメトリゼーションに依存せず定まるので，その挙動を調べる際には，パラメータ空間で考えるのではなく確率密度関数のなす関数空間で考える

のが自然である．特に最尤推定量の漸近挙動を知るには，真の確率密度関数の近傍の様子が重要な影響を持つ．上で述べた 2 つの例では，モデルは θ_0 に特異性を持っている．単調回帰では頂点が特異点にあたるので分かりやすいが，識別不能性も関数空間で考えれば，部分多様体 Θ_0 の部分だけで次元の縮退が起こることにより，関数空間内で考えたモデルの特異点に相当する．この点については [10] にもわかりやすい解説がある．

本論文では統計モデルの「特異点」を厳密に定義することは避け，そのかわりに,「接錐」を使って議論を進める．滑らかな多様体の局所的な性質は，その局所的な線形近似である接ベクトル空間によってよく記述された．特異点の近傍の様子を記述するには，接ベクトル空間の一般化である接錐を導入するのが便利である．接ベクトルと接錐を以下のように定義しよう．

統計モデル $S = \{f(x|\theta) \mid \theta \in \Theta\}$ とその中の一点 $f_0(x) = f(x|\theta_0)$ が与えられたとする．f_0 に関する 2 乗可積分関数全体を $L^2(f_0)$ で表すとき，$u \in L^2(f_0)$ が S の f_0 における**接ベクトル**であるとは，S 内の確率密度関数の列 f_n と，正数列 a_n があって

$$\log \frac{f_n}{f_0} \to 0, \qquad \frac{f_n - f_0}{a_n} \to u$$

が $L^2(f_0)$ の収束の意味で成り立つことをいう．S の f_0 における接ベクトル全体の集合は $L^2(f_0)$ 空間の中で閉錐をなしており，これを**接錐**という．

L^2 空間で述べたので多少わかりづらいかもしれないが，S 内の曲線がパラメータ θ によって $c(t) = (\theta^1(t), \ldots, \theta^m(t))$ で与えられ，$c(0) = \theta_0$ かつ

$$u(x) = \left. \frac{\partial \log f(x|\theta(c(t)))}{\partial t} \right|_{t=0} \tag{10}$$

であるとき，関数 u は緩やかな条件のもと上の定義の接ベクトルになることが示される．接ベクトルは**スコア関数**と呼ばれることもある．

特に，統計モデル S が，十分統計量 $a_1(x), \ldots, a_m(x)$ を持つ指数分布族

$$f(x|\theta) = \exp\Big\{ \sum_{j=1}^{m} a_j(x)\theta^j + b(x) - \psi(\theta) \Big\} \tag{11}$$

の部分モデルであったとしよう．ここで十分統計量は一次独立とすると，これにより指数分布族は $\mathbb{R}^m$ と同一視することができる．曲線 $c(t) = (\theta^1(t), \ldots, \theta^m(t))$ によって定まる接ベクトルは $\sum_{j=1}^{m} \frac{d\theta^j(0)}{dt} a_j(x)$ となるので，十分統計量による同一視のもと，接錐も $a_1(x), \ldots, a_m(x)$ の張る m 次元線形空間の部分集合として，幾何的に考えることが可能となる．

4 有限次元の特異モデル

4.1 凸錐モデルにおける漸近分布

まずはじめに，m 次元正規分布において平均 μ のみをパラメータに持つモデル

$$f(x|\mu) = \frac{1}{(2\pi)^{m/2}} \exp\Big\{ -\frac{1}{2}\|x - \mu\|^2 \Big\} \tag{12}$$

を考え，統計モデル S がその部分モデルの場合を考える．S は平均パラメータを与える $\mathbb{R}^m$ の部分集合 Θ によって定まる．サンプル $X_1, \ldots, X_n$ に対し，これらのサンプル平均を $\bar{X}$ とおくとき，S における最尤推定量 $\widehat{\theta}$ は簡単な計算により

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \|\bar{X} - \theta\|$$

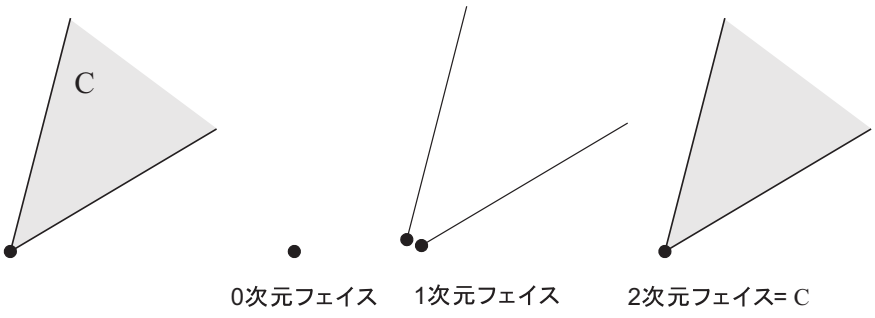

図 4: 凸集合のフェイス

であることがわかる．したがって最尤推定量はサンプル平均 $\bar{X}$ の Θ への最近点に一致する．これは 3.2 節の単調回帰の例と同様である．

いま，Θ が $\mathbb{R}^m$ の凸錐であり，真のパラメータ θ_0 が原点，すなわち m 次元標準正規分布 $N(0, I_m)$ であると仮定し，尤度比の漸近分布を以下で考えてみよう．凸錐とは，凸集合でありかつ錐（$x \in A$ なら任意の $r \geq 0$ に対し $rx \in A$ を満たす集合）である集合のことである．$\theta \in \Theta$ を $\theta = tu$（$t = \|\theta\|, u = \theta/\|\theta\|$）と分解して，$t$ に関する最大化を行うとわかるように，

$$L_n(\widehat{\theta}) = \frac{1}{2}\|\widehat{\theta}\|^2 \tag{13}$$

が得られる．

最尤推定量 $\widehat{\theta}$ の分布を記述するため，凸集合に関する用語を準備する．C を $\mathbb{R}^m$ 内の凸集合とする．$C^* = \{y | x^T y \leq 0 \text{ for all } x \in C\}$ で定まる凸錐を C の双対錐という．C の次元とは，C を含む最小のアフィン空間 $\mathrm{aff}(C)$ の次元のことをいう．また，C の相対的内点 $\mathrm{ri}(C)$ とは，$\mathrm{aff}(C)$ における相対位相に関する C の内点のことをいう．C の部分凸集合 F が C のフェイスであるとは，ある $\alpha \in (0,1)$ と $x_1, x_2 \in C$ に対して $x = \alpha x_1 + (1-\alpha)x_2 \in F$ であるなら $x_1, x_2 \in F$ であることをいう．C 自身も C のフェイスである．C の次元が m のとき，$m-1$ 次元以下のフェイスは C の相対的境界 $C - \mathrm{ri}(C)$ に含まれる．また，C が一次独立なベクトル $\{b_1, \ldots, b_L\}$ で張られる凸錐であるとき，C の k 次元フェイス（$0 \leq k \leq m$）は，$1 \leq \ell_1 < \cdots < \ell_k \leq m$ なる組に対し $F_{\ell_1,\ldots,\ell_k} = \{\sum_{i=1}^k \alpha_i b_{\ell_i} \mid \alpha_i \geq 0 \ (1 \leq i \leq k)\}$ の形の錐として与えられる．図 4 にフェイスの例を示した．

モデルが凸錐である場合の尤度比に関して次の結果が知られている．

定理 1 ([8]) $\Theta = C$ を凸錐とし，*(12)* 式の確率密度関数に対して，平均ベクトルが C に制約されたモデルを考える．真のパラメータを原点とすると，尤度比は

$$2L_n(\widehat{\theta}) \implies \sum_{k=0}^{m} q_k \chi_k^2 \tag{14}$$

（カイ *2* 乗分布の有限混合分布）と法則収束する．ここで χ_0^2 は *0* にマスを持つ一点分布である．特に C が有限個のベクトルで張られる凸錐であるとき，q_k は最尤推定量 $\widehat{\theta}$ が k 次元フェイスの相対的内点に含まれる確率を表す.

証明は略すが，C が有限個のベクトルで張られる凸錐であるときは，フェイス F に対して

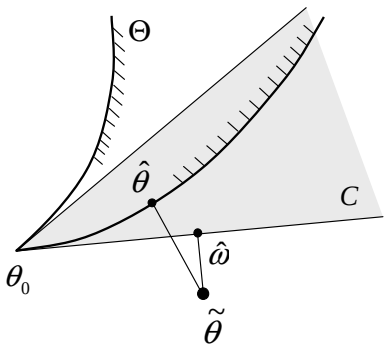

図 5: 接錐による近似

$F^\dagger = \mathrm{aff}(F)^\perp \cap C^*$ と書き，また集合のベクトル和を $\oplus$ で表すとき

$$\mathbb{R}^m = \bigsqcup_{F:\mathrm{Face\ of\ }C} \mathrm{ri}(F) \oplus F^\dagger$$

という形に分割可能（$\bigsqcup$ は排反和）となる事実を用いる．

3.2 の単調回帰の例では，図 3 がこの分割に対応している．0 次元フェイス（原点），1 次元フェイス（2 つの辺），2 次元フェイス（C 全体）に対して，$\mathrm{ri}(F) \oplus F^\dagger$ はそれぞれ，領域 IV，領域 II と III，領域 I となる．(14) 式の q_k は，k 次元フェイスに対応する領域の体積の比になり，$q_0 = 3/8$, $q_1 = 1/2$, $q_2 = 1/8$ である．したがって，尤度比の 2 倍は，この比率で自由度 k のカイ 2 乗分布を混合した分布に法則収束する．

以上では，正規分布の平均値の推定という簡単なモデルについて議論したが，接錐により局所的な近似を用いることにより，定理 1 の結論はもっと一般のモデルにも適用できる．そこで，(11) 式で定義される指数分布族を考え，モデル S がその部分モデルであると仮定しよう．真のパラメータ θ_0 が S に含まれているとし，S における最尤推定量を $\widehat{\theta}$，もとの指数分布族における最尤推定量を $\tilde{\theta}$ と書くことにする．指数分布族においては通常の漸近正規性が成立するので，

$$\sqrt{n}(\tilde{\theta} - \theta_0) \implies N(0, I(\theta_0)^{-1})$$

と法則収束する．ここで $I(\theta_0)^{1/2}$ を掛けてパラメータ θ を変換することにより，はじめから $I(\theta_0) = I_m$ としてよい．すると，n が十分大きいとき，$\tilde{\theta}$ は $\sqrt{n}$ のスケール変換の後，θ_0 のまわりに m 次元標準正規分布に従って分布すると考えてよい．さらに，図 5 にあるように，モデル S を定義する Θ の θ_0 における接錐を C とおく．このとき，$\tilde{\theta}$ からユークリッド距離の意味で最も近い C の点を $\widehat{\omega}$ とおくと，$\widehat{\omega}$ と $\widehat{\theta}$ の分布は $n \to \infty$ のときに一致することが示される．もし，接錐 C が凸であるならば．定理 1 の主張が適用可能となり，最尤推定量 $\widehat{\theta}$ の漸近分布はカイ 2 乗分布の混合分布となる．

4.2 一般の場合のアプローチ

上のケースでは接錐が凸であることが重要であったが，一般には接錐は凸とは限らない．接錐が凸でない場合，尤度比の漸近分布に関して一般的な結果を得るのは難しいがいくつかのアプローチが知られている．そのひとつが以下に紹介する「チューブ法」である．上の議論からわかるように，有限次元の指数分布族の部分モデルに対して最尤推定量の漸近分布を知るには，正規分布

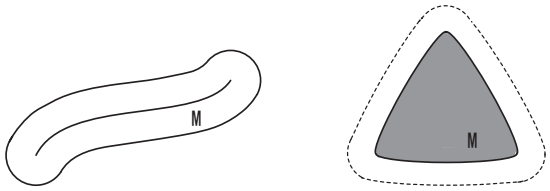

図 6: チューブ M_c

$N(0, I_m)$ から発生するサンプル Z から閉錐 C への最近点 Y の分布を求めればよい．錐 C を，原点からの距離 r と単位球面上の部分集合 $\eta \in M = \{\eta \in C \mid \|\eta\| = 1\}$ でパラメトライズし，$C = \{r\eta \in \mathbb{R}^m \mid r \geq 0, \eta \in M\}$ と表そう．すると簡単な議論により，尤度比

$$L_n(\widehat{\theta}) = \frac{1}{2}\|Y\|^2 = \frac{1}{2}\Big(\max\Big\{0, \max_{\eta \in M} \eta^T Z\Big\}\Big)^2 \tag{15}$$

が得られる．これは

$$\max_{\eta \in M} \eta^T Z \tag{16}$$

の単調増加関数であるので (16) 式の分布を求めれば尤度比の分布がわかることになる．いま $U = Z/\|Z\|$ とおくと，$Z \sim N(0, I_m)$ により，U と $\|Z\|$ は独立であり，U は球面上の一様分布，$\|Z\|^2$ は自由度 m のカイ 2 乗分布に従う．したがって $\max_{\eta \in M} \eta^T Z$ の分布関数に関して

$$\begin{aligned}
\mathrm{Prob}\Big(\max_{\eta \in M} \eta^T Z \geq a\Big) &= F(a) \quad (\text{とおく}) \\
&= \mathrm{Prob}\Big(\max_{\eta \in M} \eta^T U \geq \tfrac{a}{\|Z\|}\Big) \\
&= \int_{a^2}^{\infty} \mathrm{Prob}\Big(\max_{\eta \in M} \eta^T U \geq \tfrac{a}{\sqrt{\xi}}\Big) g_m(\xi) d\xi
\end{aligned}$$

を得る．ここで $g_m(\xi)$ は自由度 m のカイ 2 乗分布の確率密度関数である．$\{u \in S^{m-1} \mid \max_{\eta \in M} \eta^T u \geq \cos(c)\}$ で定まる集合 M_c を M のまわりの半径 c のチューブと呼ぶ（図 6）．

このとき，$\max_{\eta \in M} \eta^T u \geq b$ は $u \in M_{\cos^{-1}(b)}$ と同値であるから，$m-1$ 次元球面の体積を Ω_m とすると，結局

$$F(a) = \frac{1}{\Omega_m} \int_{a^2}^{\infty} \mathrm{Vol}(M_{\cos^{-1}(a/\sqrt{\xi})}) g_m(\xi) d\xi \tag{17}$$

を得る．したがって，体積 $\mathrm{Vol}(M_c)$ が得られれば $\max_{\eta \in M} \eta^T Z$ の分布 $F(a)$ が得られる．(実はその逆も言うことができる．(17) 式右辺の積分は，簡単な変数変換によってラプラス変換の形に書くことができる．ラプラス変換の一意性から，$\mathrm{Vol}(M_c)$ と $F(a)$ は関数として 1 対 1 である.)

しかしながら，一般の図形 M に対して M_c の体積を計算するのは困難である．次章ではチューブ M_c の体積計算を可能とするような正則条件と，その条件の下で得られる $\max_{\eta \in M} \eta^T Z$ の上側確率公式について説明する．

4.3 チューブ法

M を単位球面 S^{m-1} の閉部分集合とする．M のまわりの半径 c のチューブは，大円距離を dist で表すとき

$$M_c = \{u \in S^{m-1} \mid \mathrm{dist}(u, M) \leq c\}$$

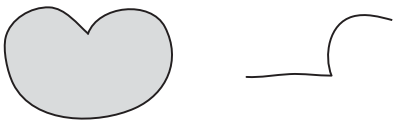

図 7: 臨界半径が 0 である集合

であった．いま M_c の各点 u に対して，$\mathrm{dist}(u, M) = \min_{v\in M} \mathrm{dist}(u, v)$ を達成する v が一意に定まるとする．このときの v を $p(u)$ とおく．$p(u) = v$ となるような $u \in M_c$ の全体を $N_{v,c}$ とおけば，M_c は

$$M_c = \bigsqcup_{v\in M} N_{v,c}$$

と排反分割される．各 v について $N_{v,c}$ の体積を求め，それを $v \in M$ の範囲について足し合わせる（積分する）ことによってチューブ M_c の体積を求めることができる．この方針によって，$c \in [0, c_0]$，ただし

$$c_0 = \sup\{c \geq 0 \mid \text{任意の } u \in M_c \text{ に対して } p(u) \text{ が一意に定まる }\},$$

の範囲で $\mathrm{Vol}(M_c)$ を評価することができる．c_0 は臨界半径 (critical radius) あるいはリーチ (reach) と呼ばれる．c_0 は $[0, \pi]$ の範囲の値を取りうるが，後の都合上 $c_0 > \pi/2$ のときは $c_0 = \pi/2$ と定義することにする．

以下では M に対する正則条件として，$c_0 > 0$ を仮定する．この条件を満たさない集合として，図 7 のような非凸な接錐を持つ集合がある．（ただし各点で凸の接錐を持つ集合であっても $c_0 = 0$ となることもある．）

M が正の臨界半径を持つとき，M のまわりのチューブの体積は一般に $c \in [0, c_0]$ の範囲で

$$\mathrm{Vol}(M_c) = \Omega_m \sum_{k=1}^{d+1} q_k \bar{B}_{\frac{k}{2},\frac{m-k}{2}}(\cos^2 c) \tag{18}$$

の形となる．ここで d は M の次元，

$$\bar{B}_{j,k}(b) = \frac{\Gamma(j+k)}{\Gamma(j)\,\Gamma(k)} \int_b^1 \xi^{j-1}(1-\xi)^{k-1}\, d\xi$$

は母数 (j, k) のベータ分布の上側確率である．その係数 q_k はワイルの幾何不変量と呼ばれる，M のみに依存する幾何量である．例えば M が線分 $[0, 1]$ あるいは円周 S^1 に同相な一次元多様体の場合は，$d = 1$,

$$q_2 = \frac{|M|}{2\pi}, \quad q_1 = \frac{\chi(M)}{2}$$

（$|M|$ は M の長さ，$\chi(M)$ は M のオイラー標数），M が区分的に滑らかな境界を持った 2 次元多様体の場合は，$d = 2$,

$$q_3 = \frac{|M|}{4\pi}, \quad q_2 = \frac{|\partial M|}{4\pi}, \quad q_1 = \frac{\chi(M)}{2} - \frac{|M|}{4\pi}$$

（$|M|$ は M の面積，$|\partial M|$ は M の境界の長さ，$\chi(M)$ は M のオイラー標数）である．ここでオイラー標数 (Euler-Poincaré characteristic) $\chi(\cdot)$ とは，ある位相空間 D が単体的複体 K と同相であるとき，

$$\chi(D) = \chi(K) = \sum_{F:\text{Face of } K} (-1)^{\dim F}$$

として定義される位相不変量である．$\chi([0,1])=1, \chi(S^1)=0, \chi(\emptyset)=0$ などが成り立つ．

チューブ体積公式 (18) を (17) の右辺に代入し，ξ に関する積分を行うと

$$\sum_{k=1}^{d+1} q_k \bar{G}_k(a^2) = \hat{F}(a) \quad (\text{とおく}) \tag{19}$$

となる．ただし $\bar{G}_k$ は自由度 k のカイ 2 乗分布の上側確率である．

チューブ体積公式 (18) は半径 c が小さいときにのみ有効なものであるので，それから導かれる $\hat{F}(a)$ は，当然 $F(a)$ とは異なるものである．しかしながら，積分 (17) の形を詳しく見ると，a が大きいときの $F(a)$ の積分には半径 c が小さいときの $\mathrm{Vol}(M_c)$ が寄与するため，a が大きいときには何らかの意味で $\hat{F}(a)$ は $F(a)$ を近似することが期待される．実際，以下が成り立つ．

定理 2 (Kuriki and Takemura [6]) $a \to \infty$ のとき，

$$|\hat{F}(a) - F(a)| = O(a^{m-2} e^{-\frac{1}{2}a^2(1+\tan^2 c_0)}).$$

$\hat{F}(a)$ の各項は $O(a^{k-2}e^{-a^2/2})$ であるので，正則条件 $c_0 > 0$ のもとで，$\hat{F}(a)$ の誤差はそれ自身の各項よりも指数的に小さいことが分かる．この意味で，定理 2 はチューブ法近似の正当性を示している．

M が，球面 S^{m-1} 上の大円を直線とみなしたときに凸（すなわち錐 C が凸）の場合は $c_0 = \pi/2$ となり，$\hat{F}(a) = F(a), a > 0,$ が成り立つ．また，全ての k について $q_k \geq 0$ であることも示される．このとき (19) 式はカイ 2 乗分布の有限混合分布の上側分布関数となり，定理 1 の結果と整合する．

統計推測，特に多変量解析に現れるいろいろな検定問題において，幾何不変量 q_k と臨界半径 c_0 を具体的に評価することができる．統計的仮説検定の文脈では，検定の p 値がある程度小さい範囲，すなわち検定統計量の分布の上側裾確率が重要であるので，チューブ法が有用な局面は多い．また c_0 がある程度大きな値であるときは，実質的に $\hat{F}(a)$ と $F(a)$ の差異を無視できることが多い．チューブ法に関する最近の発展については [11], [6], [9], [1] およびそれらの引用文献を参照されたい．

5　無限次元の特異モデル

４章では，モデルが有限個の十分統計量を持っており，接錐はそれらが張る有限次元の関数空間内に含まれる場合を議論した．しかし，統計モデルの中には,（有限個のパラメータで定義されているにもかかわらず）接錐が有限次元の関数空間に入らないものもある．このような場合には，最尤推定量の挙動はさらに複雑となり，例えば尤度比の漸近オーダーが $O_p(1)$ よりも大きく，発散することも起こり得る．この尤度比の発散現象は，第２章で述べた正規混合モデルに対して，Hartigan [5] が最初にその証明を与えた．この章では，接錐が無限次元の空間を張るようなモデルについて論じ，特にニューラルネットなどの尤度比の漸近オーダーについて論じる．

5.1　局所錐型パラメトリゼーションと尤度比の発散

接錐の概念を強調するために，モデルの局所錐型パラメトリゼーションを導入しよう．$\mathbb{R}_+$ で非負実数全体を表す．統計モデル S と，S の中の確率密度関数 f_0 に対し，S の f_0 における局所錐型パラメトリゼーションとは，$\Theta_0 \subset \mathbb{R}^{m-1}$ と $\Theta \subset \Theta_0 \times \mathbb{R}_+$ による S のパラメトリゼーショ

ン $S = \{g(z|\alpha,\beta) \mid (\alpha,\beta) \in \Theta\}$ であって，以下の条件を満たすものをいう：(i) 任意の α に対し，$\sup\{\beta \mid \{\alpha\} \times [0,\beta) \in \Theta\} > 0$. (ii) 部分モデル $S_\alpha = \{g(z|\alpha,\beta) \mid (\alpha,\beta) \in \Theta\}$ は漸近片側正規性をみたす. (iii) S_α の $\beta = 0$ における Fisher 情報量は 1 に等しい.

S の f_0 における局所錐型パラメトリゼーションがあるとき，各 α を固定した β に関するスコア関数

$$t(z;\alpha) = \frac{\partial \log g(z|\alpha,0)}{\partial \beta} \tag{20}$$

は，接ベクトルであり，(iii) より $L^2(f_0)$ の中の単位ベクトルである．この単位接ベクトルを用いて，尤度比を一般的に書くことが可能である．まず，α を固定して部分モデル S_α に関して Taylor 展開にもとづく通常の漸近展開を行うと，仮定 (ii) より

$$\sup_\beta L_n(\alpha,\beta) = \frac{1}{2}\max\{0, U_n(\alpha)\}^2 + o_p(1) \tag{21}$$

ただし,

$$U_n(\alpha) = \frac{1}{\sqrt{n}}\sum_{i=1}^n t(Z_i;\alpha)$$

が得られる．$U_n(\alpha)$ は $n \to \infty$ のとき標準正規分布に従う．尤度比は

$$\sup_{\alpha,\beta} L_n(\alpha,\beta) = \frac{1}{2}\sup_{\alpha\in\Theta_0}\left\{\max\{0, U_n(\alpha)\}^2 + o_p(1)\right\} \tag{22}$$

で表される．これは（15）式の表現と非常に類似している．しかし，高次項 $o_p(1)$ は α に依存するので，これを $\sup_\alpha$ の外に出すことは一般にはできない．すなわち，尤度比の主要項が $\frac{1}{2}\sup_\alpha \max\{0, U_n(\alpha)\}^2$ であるとは一般には結論できない．この結論が成立するための条件は Dacunha-Castelle and Gassiat [2] で論じられている．

（22）式を使って尤度比の厳密な漸近分布を求めることは，一般には困難である．しかしながら，この表現は，尤度比の漸近オーダーが $O_p(1)$ より大きくなる現象を理解するのに役立つ．各 α に対して $U_n(\alpha)$ は正規分布に法則収束するが，それらを α 全体で定まる確率過程とみたとき，その確率過程がどれぐらいの最大値を取るのかを考えてみる．容易に推察できるように，各 $U_n(\alpha)$ がなるべくばらばらに（無相関に）分布するほどその最大値は大きくなる．ふたつの α_1, α_2 に対して，$U_n(\alpha)$ の極限分布であるガウス分布の相関

$$R(\alpha_1,\alpha_2) = E[t(X;\alpha_1)t(X;\alpha_2)]$$

は，接ベクトル $t(x;\alpha)$ によって次のような性質を持つ.

補題 3 単位接ベクトル $t(x;\alpha)$ に対し，ある系列 α_j $(j = 1,2,\ldots)$ が存在して，f_0 のもと $t(x;\alpha_j)$ が 0 に確率収束するならば，任意の $\varepsilon > 0$ に対し $\{\alpha_j\}$ の部分列 $\{\alpha_{j_k}\}$ が存在して任意の j,k に対し $|R(\alpha_{j_k},\alpha_{j_\ell})| \le \varepsilon$ とできる.

この補題は，$\{t(x;\alpha)\}$ のなかに殆ど直交する方向が無限個存在するための十分条件を与えている．$U_n(\alpha)$ の極限分布は正規分布であるから，それらがほとんど無相関であれば，その最大値は $N(0,1)$ に従う独立なサンプルの最大値に近いはずである．極値理論を用いると，$N(0,1)$ からの独立な m 個のサンプルの最大値は $m \to \infty$ のとき $\sqrt{2\log m}$ に収束することが知られている．このような考察を用いて，次の定理が得られる.

定理 4 (Fukumizu [3]) 統計モデル S が真のパラメータ θ_0 を中心とする局所錐型パラメトリゼーション $\{g(x|\alpha,\beta)\}$ を持つとする．このとき（20）式で定まる単位接ベクトルの中に0に確率収束する系列が存在すれば，任意の $M>0$ に対し

$$\mathrm{Prob}\Big(\sup_{\alpha,\beta} L_n(\alpha,\beta) \leq M\Big) \to 0 \qquad (n\to\infty)$$

が成り立つ．

第4章で考えた（15）式の場合では，η は有限次元の球面の部分集合であったから，その自由度は m で抑えられる．しかし $\{t(x;\alpha)\}$ は関数空間の集合なので，その自由度（直交する方向）は無限大を取り得るのである．

尤度比が発散する例として多層パーセプトロンについて見ておこう．中間素子を1個だけ持つ簡単な多層パーセプトロンモデル

$$\psi(x;\theta) = b\varphi(ax+c) \tag{23}$$

を考える．ただし $\varphi(t)=1/(1+e^{-t})$ とする．統計モデルは（2）式のガウスノイズモデルにより定め，真の関数は定数0，すなわち Y_i はガウス分布に従う独立なサンプルだと仮定する．また X を発生させる分布 Q も仮定する（23）式のモデルは

$$\tilde{\psi}(x;\omega,\beta) = \beta v(x;\omega)$$

と書き直せる．ここで $\beta\geq 0$ ，$\omega=(a,c)\neq(0,0)$ であり，

$$v(x;\omega) = \frac{\varphi(ax+c)}{\|\varphi(ax+c)\|_{L^2(Q)}}$$

により定義される．このとき（20）式の接ベクトルを求めると，

$$t(x,y;\omega) = y\, v(x;\omega)$$

であることが容易に確かめられる．$v(x;\omega)$ は a,c を適当に選ぶと各点で0に収束させることが可能である．したがって定理4から尤度比は発散することが分かる．一般に，多層パーセプトロンにおいては，真の関数がモデルよりも少ない中間素子で実現可能であれば，この例と同様の理由により尤度比が発散することが知られている（[3]）．

5.2 尤度比のオーダー

尤度比が発散するとき，その n に関する漸近オーダーはどのようになるのであろうか？その一般的な答えは未解決である．ニューラルネットを含む非線形回帰問題に対しては以下のような一般的な上界が知られている．

定理 5 (Fukumizu and Hagiwara [4]) $\mathbb{R}^m$ から $\mathbb{R}$ への関数の族 $\mathcal{F}$ と，確率モデル $p(y|u)$ により定義される回帰モデル $\{p(Y|\psi(X)) \mid \psi\in\mathcal{F}\}$ を考える．独立なサンプルを与える真の関数 $\psi_0(x)$ が有界で，$\mathcal{F}$ の VC 次元が有限であると仮定する．このとき確率モデルに対する正則条件（ガウスノイズやバイナリ回帰はこれを満たす）のもと，以下が成り立つ．

$$\sup_{\psi\in\mathcal{F}} \sum_{i=1}^{n} \log\frac{p(Y_i|\psi(X_i))}{p(Y_i|\psi_0(X_i))} \;=\; O_p(\log n)$$

多層パーセプトロンのある種の場合には，$\log n$ の下界も知られており，ちょうど $\log n$ のオーダーが実現される（[3]）．

6　おわりに

特異モデルの最尤推定に関して，接錐による見方を中心としてやや駆け足で紹介した．しかしながら，現状得られている結果は必ずしも完全な体系を持っているわけではない．有限次元の場合でも，接錐が凸の場合にはかなり一般的で簡単な漸近分布が得られるが，凸でない場合に対しては現在さまざまな研究が行われている段階である．無限次元の場合にはさらに問題は複雑となり，未開拓の問題が多く残っている．特に $O_p(1)$ より大きなオーダーを持つ尤度比に関しては，厳密な漸近分布が知られているものはごく限られている（[7]）．本論文では最尤推定のみを論じたが，他の推定量，特に正則化項ないしはペナルティ項を導入した場合の推定量の漸近的挙動を調べることは非常に重要である．実際，正則化項をうまくいれると尤度比の発散などの複雑な現象が回避できることも知られている（[12]）．しかし，正則化項や正則化係数をどのように設定すべきかを理論的に詰めようとすると，やはり最尤推定量の挙動をすることが基本となるので，本論文で扱った問題に関する研究がさらに発展することが期待される．

参考文献

[1] R. Adler and J. E. Taylor. *Random Fields and Their Geometry*. Birkhäuser, Boston. (To appear, available at `http://iew3.technion.ac.il/~radler/publications.html`)

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

[3] K. Fukumizu. Likelihood ratio of unidentifiable models and multilayer neural networks. *The Annals of Stastistics*, 31(3):833-851, 2003.

[4] K.Fukumizu and K.Hagiwara. A general upper bound of likelihood ratio for regression. Research memorandum No.887, The Institute of Statistical Mathematics. 2003.

[5] J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, 807–810, 1985.

[6] S. Kuriki and A. Takemura. Tail probabilities of the maxima of multilinear forms and their applications. *The Annals of Statistics*, 29(2), 328-371, 2001.

[7] X. Liu and Y. Shao. Asymptotic distribution of the likelihood ratio test in a two-component normal mixture model. Technical report, Department of Statistics, Columbia University, 2001.

[8] A. Shapiro. Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review*, 56(1):49–62, 1988.

[9] A. Takemura and S. Kuriki. On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains. *The Annals of Applied Probability*, 12(2):768–796, 2002.

[10] 甘利俊一，尾関智子，朴慧暎．神経多様体の特異点と学習．日本神経回路学会誌，??(?), 2003.

[11] 栗木哲，竹村彰通．正規確率場の最大値の分布．統計数理, 47(1):201–221, 1999.

[12] 福水健次. 局所錐型モデルの漸近理論とそのニューラルネットへの応用. 第４回情報論的学習理論ワークショップ予稿集, 135–140, 2001.

[13] 渡邊澄夫. 特異モデルとベイズ学習. 日本神経回路学会誌, ??(?), 2003.