北海道大学工学部研究報告
第106号（昭和56年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No.106 (1981)

母数空間における判別関数とそのロバスト性について

佐藤義治*・須川和明**・成田雅博***・林　義実****・河口至商*

（昭和56年6月30日受理）

On a Discriminant Function in the Parameter Space and its Robustness

Yoshiharu SATO, Kazuaki SUGAWA, Masahiro NARITA, Yoshimi HAYASHI
and Michiaki KAWAGUCHI

(Received June 30, 1981)

Abstract

In a family of statistical distributions, several measures for estimating the similarity of the distributions have been introduced.
C. R. Rao showed that, in a family of distributions with a parameter, the distance between the two distributions corresponding to the contiguous points in the parameter space was qiven by the quadratic form and introduced Riemmanian space as the model of the parameter space.

In this paper, for the problem of discrimination of distributions we introduce the criterion of discrimination by the metric of the parameter space and show that a sample is assigned to given distributions by the metric of their dual parameter spaces. Especially we give the method of discrimination of the exponential family and discuss the robustness of this method.

1．序　　論

統計的分布族において，分布間の類似性を評価する尺度は，以前から種々提案されている。これらの一つとして分布間の距離という概念を導入し、分布族を距離空間として扱うことが考えられる。

C. R. Rao[1]は母数（パラメータ）θを持つ分布族の中で，θが微小に変化したときの分布間の距離が，$d\theta$の二次形式$g_{ig}(\theta)d\theta^i d\theta^j$で与えられることを示し，母数空間のモデルとしてリーマン空間を提案した。その後，母数空間の幾何学的性質の研究は二，三見られるものの，統計学の立場からの研究，あるいは応用は，行なわれていなかったが，最近になり，B. Efron[2], A. P. Dawid[3], S. Amari[4]により，母数空間の幾何学的，統計学的意味がより明確にされている。

ここでは，分布間の判別の問題をとり上げ，判別の良さを母数空間の計量から導き，与えられ

* 情報工学専攻　情報数理工学第一講座
** 北大医学部　公衆衛生学講座
*** 北海道留萌高等学校
**** 釧路工業高等専門学校

た分布へのサンプルの割り当ては双対母数空間の計量によって行なうことを示す。特に指数型分布族における判別方式を求め、この判別方式のロバスト性について考察する。

2．指数型分布の母数空間

指数型分布族の確率密度関数は，確率測度 $\nu(x)$に関してつぎのような形で与えられる。

（2．1） $f(x, \theta)=\exp\{\theta^i x_i-\psi(\theta)\}$

ただし，すべての θ に対して $e^{\psi(\theta)}\equiv\int e^{\theta^i x_i}d\nu(x)<\infty$を満たすものとする。$\psi(\theta)$はパラメータ $\theta=(\theta^i)$, (i＝1，2，……，n) の convex 関数である。指数型分布族には多くの確率分が含まれている。たとえば，正規分布，ガンマ分布，指数分布，一様分布，多項分布，負の多項分布，ポアソン分布等である。

パラメータ $\theta=(\theta^i)$からなる空間 Θ を指数型分布族の母数空間と呼ぶ。一般に，パラメータを持つ分布に関して，そのパラメータ空間の計量及び接続はつぎのように与えられる。

（2．2） $l(x, \theta)=\log f(x, \theta)$

とおくと，基本計量テッソルは，

（2．3） $g_{if}(\theta)=E_\theta(\partial_i l\ \partial_j l)$ となり，線型接続係数を $\overset{\alpha}{\Gamma}_{ijk}$ と書くと，これは一般に，

（2．4） $\overset{\alpha}{\Gamma}_{ijk}=E_\theta[\partial_i\partial_j l\ \partial_k l]+\frac{1-\alpha}{2}E_\theta[\partial_i l\partial_i l\partial_k l]$ $(-1\leqq\alpha\leqq 1)$

で与えられる。ただし，E_θ は確率分布 $f(x, \theta)$ に関する期待値を表わし，∂_i・は θ_i に関する偏導関数を表わす。ここで，接続係数につくのは－1から1までの実数であり，$\alpha=0$ のときはユークリッド接続，すなわち $\overset{\alpha}{\Gamma}_{ijk}=[ij, k]$ なるクリストッフェルの第一種記号に一致する。また $\alpha=-1$，$\alpha=1$ の場合はそれぞれ mixture connection, Efron connection と呼ばれている。指数型分族（2．1）に関しては（2．3），（2．4）はさらに簡単な形で表わされる。

（2．5） $g_{ij}(\theta)=\partial_i\partial_i\psi(\theta)$

（2．6） $\overset{\alpha}{\Gamma}_{ijk}=\frac{1-\alpha}{2}E_\theta[\partial_i l\partial_i l\partial_k l]=\frac{1-\alpha}{2}\partial_i\partial_j\partial_k\psi(\theta)$

指数型分布族（2．1）における母数空間 Θ に対して，標本 $x=(x_1, x_2, \cdots, x_n)$ からなる標本空間Sが考えられる。このとき，標本 x_i の期待値 η_iは，

（2．7） $\eta_i=E_\theta(x_i)=\partial_i\psi(\theta)$

で与えられる。ここで $\psi(\theta)$は θ の convex 関数であることに注意して，convex duality の理論[6]を用いると，Θ とSとは双対関係にあることがわかる。すなわち（2．1）の尤度関数

$$\log f(x, \theta)=\theta^i x_i-\psi(\theta)$$

において，

（2．8） $\phi(x)=\max\log f(x, \theta)=\max\{\theta^i x_i-\psi(\theta)\}$

なる関数 $\phi(x)$に対して

（2．9） $\phi(\eta)=\max\{\theta^i\eta_i-\psi(\theta)\}$

を対応させると，

（2．10） $\theta^i=\frac{\alpha\phi(\eta)}{\partial\eta_i}$

なる逆変換が得られ，関数 $\phi(\eta)$も convex 関数となることが示され，

（2．11） $\psi(\theta)=\max\{n_i\theta^i-\phi(\eta)\}$

なる関数が成り立つ。また $\phi(m)$と $\psi(\theta)$の間にはつぎの関係がある。

（2．12） $\psi(\theta)+\phi(\eta)=\theta^i\eta_i$

以上より，Θ とSとの間の変換はルジャンドル変換であることが示される。Amari[7]によりSの計

量および接続は Θ の計算からつぎのように与えられることが示されている。

（2．13）　$ds^2=g^{ij}(\eta)d\eta_i d\eta_j$

$$g^{ij}(\eta)=\frac{\partial^2\phi(\eta)}{\partial\eta_i\partial\eta_j}$$

$$\overset{\alpha}{\Gamma}{}^{ijk}=\frac{1+\alpha}{2}\ \frac{\partial^3}{\partial\eta_i\partial\eta_j\partial\eta_k}\ \phi(\eta)$$

また双対性において，Θ における α－接続に対してSにおける$-\alpha$－接続が対応する。

以上よりつぎの定理が得られる。

Theorem A (S. Amari[4])

Efron－接続は，パラメータ空間においては恒等的に $\Gamma_{ijk}=0$ である。一方，Mixture－接続は標本空間において恒第的に $\Gamma^{ijk}=0$ となる。2つの接続の間には双対性がある。

3．指数型分布族の判別

ここでは，2つの母集団 π_1，π_2の判別を考える。判別の問題として，つぎの2つがある。

問題1：N個の標本 x_1，x_2，……，x_N が与えられたとき，この標本は π_1からのものであるか，それとも π_2 からのものであるかを判定せよ。

問題2：1つの標本xが得られたとき，これが母集団 π_1 か π_2のどちらから生じたものか未知の場合，これをπ_1かπ_2 のいずれかへ判別せよ。

いま母集団 π_1，π_2の確率密度関数が同一の指数型分布族 $f(x,\theta)$ であり，その母数の値が異なるものとする。すなわち，

π_1：$f(x,\theta_1)$

π_2：$f(x,\theta_2)$　　$\theta_i=(\theta_i^1,\theta_i^2,\cdots\cdots,\theta_i^n)$

このとき問題1は，$f(x,\theta)$の母数空間 Θ においてつぎのような判定方式を考える。標本 $x_1, x_2,\cdots\cdots, x_N$ から得られる母数θの推定値を $\bar{\theta}$ とすると，母集団π_1, π_2は Θ における2点 θ_1，θ_2 で特徴づけられるから，Θ に与えられる計量 $g_{ij}(\theta)$ を用いて，$\hat{\theta}$と θ_1，θ_2の距離$D(\hat{\theta},\theta_1)$, $D(\hat{\theta},\ \theta_2)$ を用いて，

（3．1）　$D(\theta,\theta_1)\leqq D(\hat{\theta},\theta_1)$

ならば，標本は π_1 からのものであると判定する。このとき，$D(\hat{\theta},\theta_i)$ を母数定空間 Θ 上で θ_i と $\hat{\theta}$を結ぶいかなる道に沿って測るかという問題がある。ここではMixture　pathすなわちMixture接続による道に沿ってその長さを測るものとする。このとき道の方程式は弧長sをパラメータとして，一般に

（3．2）　$$\frac{d^2\theta^i}{ds^2}+\Gamma^i_{jk}\frac{d\theta^j}{ds}\frac{d\theta^k}{ds}=0$$

で表わされ，$\hat{\theta}$と θ_iの距離は

（3．3）　$$D(\theta_i,\bar{\theta})=\int\sqrt{g_{\alpha\beta}(\theta_i)\frac{d\theta_\alpha}{ds}\frac{d\theta^\beta}{ds}}\ ds$$

と表わすことができる。

一方問題2は Θ の双対空間すなわち標本空間Sで表現される。Sの座標系を標本 x_i の期待値

$\eta_i = E(x_i)$とすると，与えられた母集団π_1，π_2はS上の2点として（2．7）より

（3．4） $\eta^1{}_i = \partial_i \psi(\theta) \mid \theta = \theta_1$

（3．5） $\eta^2{}_i = \partial_i \psi(\theta) \mid \theta = \theta_2$

として表わされる。標本$x=(x_i)$が与えられたとき，S上でxと$\eta^i(i=1.2)$を結ぶ道をMixturepathの双対性から求められる道とすると，弧長を$\tilde{s}$としてその方程式は

（3．6） $$\frac{d^2\eta_i}{d\tilde{s}^2} - \tilde{\Gamma}^{jk}_i \frac{d\eta_j}{d\tilde{s}} \frac{d\eta_k}{d\tilde{s}} = 0$$

となる。ところが定理Aよりこの座標系η^iに関しては$\tilde{\Gamma}^k_{ij} 0$であるので（3．6）は

$$\frac{d^2\eta_i}{d\tilde{s}^2} = 0$$

と表わされる。したがって

$$\tilde{D}(\eta^i, x) \equiv g^{\alpha\beta}(\eta^i)(\eta^i_\alpha - x_\alpha)(\eta^i_\beta - x_\beta)$$

と表わすことができ，つぎのような判別方式が得られる

（3．7） $\tilde{D}(\eta^1, x) \leqq \tilde{D}(\eta^2, x)$

ならばχをπ_1に割りあてる。上式（3．7）の等号が成り立つχの関数が，通常，判別関数と呼ばれているものに相当する。例えば，正規分布，2項分布における判別関数を具体的に計算すると，つぎのようになる。2つの正規分布を$N(\mu_1, \sigma_1^2)$，$N(\mu_2, \sigma_2^2)$とすると

（3．8） $(\sigma_1 + \sigma_2)x - \sigma_2 \mu_1 - \sigma_1 \mu_2 = 0$

また，2つの2項分布を$P_1(x, p_1, N)$，$P_2(x, p_2, N)$とすると

（3．9） $$(\sqrt{p_1(1-p_1)} + \sqrt{p_2(1-p_2)})x - N(p_2\sqrt{p_1(1-p_1)} + p_1\sqrt{p_2(1-p_2)}) = 0$$

となる。

4．非正規性に対するロバスト性

統計学において，ロバスト性とは一般に，推定方式，決定方式が分布の形によらないという概念であり，分布形が未知のデータに対する推論では重要な性質である。しかしながら，ロバスト性の数量的表現は困難であり、種々の実験等により検証する必要がある。

ここでは2つの正規分布における判別関数（3．8）が非正規分布に対して，どの程度ロバスト性があるのかを考察する。確率分布の正規分布からのずれを歪度(Skewness)と尖度(kurtosis)によって示し，与えられた歪度Sと尖度Kをもつ分布をつぎようにして求める。ξを平均μ，分散σ^2，をもつ確率変数とし，その歪度をS，尖度をKとするとき，

$$x = (\xi - \mu)/\sigma$$

の確率密度関数は近似的に

（4．1） $$f(x) = \varphi(x) - \frac{S}{3!}\varphi^{(3)}(x) + \frac{K}{4!}\varphi^{(4)}(x) + \frac{10\,S^2}{6!}\varphi^{(6)}(x)$$

と表わせられる。ただし，$\varphi(x)$は標準正規分布の密度関数$\varphi(x) = e \times P\{x^2/2\}/\sqrt{2\pi}$であり，また$H_\nu(x)$をエルミート多項式として$\varphi(\nu)$は次式により定義される。

$$\varphi^{(\nu)}(x) = (-1)^\nu H_\nu(x)\varphi(x)$$

ここで用いる尖度K，歪度Sは

$$K=E(x-\mu)^4/\sigma^4-3$$

$$S=E(x-\mu)^3/\sigma^3$$

と確立密度関数が（4．1）で与えられる2つの母集団π_1，π_2の密度関数をそれぞれ

$$f_1(x_i, \mu_1, \sigma_1, s_1, k_1),$$

$$f_2(x_i, \mu_2, \sigma_2, s_2, k_2)$$

と表わし，判別関数（3．8）を用いて判別したときの誤判別の確率を$\bar{P}$とする。一方，f_1，f_2を最尤法により，

$$V=\log(f_1/f_2)=0$$

を境界として判別したとき，すなわち

$v \geqq 0$　ならば　$x \in \pi_1$

$v < 0$　ならば　$x \in \pi_2$

と判別したときの誤判別の確率をPとする。ここでロバスト性の評価を$\bar{P}$とPとの比

（4．2）　　$R=P/\bar{P}$

によって行う。

数値実験として，π_1の確率密度関数を　$f(x_i, 0.0, 1.0, 0.5, 0.5)$と固定し，$\pi_2$に関する$\mu_2$，$s_2$，$k_2$のそれぞれつぎのように与える。

μ_2	σ_2^2	s_2	k_2
−2.0	2.0	−0.5	0.5～9.0
−2.0	2.0	−1.0	2.0～10.0
−1.0	0.5	−1.5	0.5～10.0
−1.0	0.5	−2.0	0.5～10.0

ただし，k_2は0.5きざみで与えている。π_2の各値の組に対して（4．2）のRの値を求めたものを，図4.1, 4.2に示す。図4.1と図4.2ではKの値の増加に対して逆の傾向を示しているが，これは図4.3(a)～(d)に示す密度関数のおおよその形から理解することができる。図4.3(a)～(d)はいずれも$k_2=6.0$の場合について作成したものである。図のx軸上に示すB点が（3．8）による判別点であり，A点が最尤法による判別点である。図4.3(a), (d)で示される場合には，k_2が増加するにつれてA点は負の方向(左の方向)に移動するため，Rの値は減少するが，(c), (d)の場合には，k_2が増加するにつれてA点は負の方向に移動するけれども，A点とB点の位置関係が，(a), (b)とは逆になっているためRの値は増加する。一方s_2が−0.5～1.0の場合にはk_2が変化してもRの値はほとんど0.9以上の値をとっておりこの範囲ではロバストといえよう。(Sの値はその絶対値で考えるべきである。ここではπ_1がπ_2より正の方向にあるため負の値について考察する）しかしs_2が−1.5, −2.0とその絶対値が大きくなると，尖度k_2の影響は無視できないことを示している。従来の研究には（5参照）kの値の影響はさほどないという結果も得られているが，ここでの判別関数に関しては，明らかな傾向が見られる。以上より結果として，Sの値は絶対値が1を超えると2つの密度関数の相対的位置関係がロバスト性に大きな影響をもち，特にkの値に制限がつけられる。

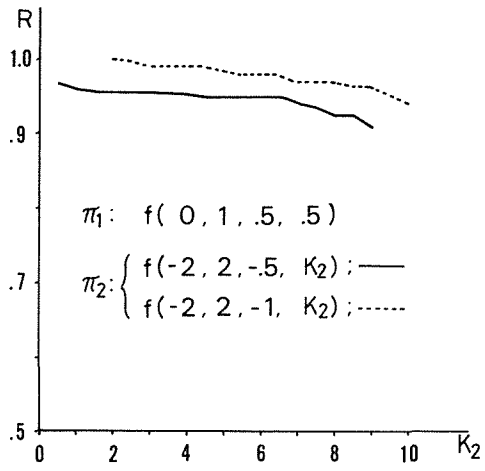

図 4.1　確率の比

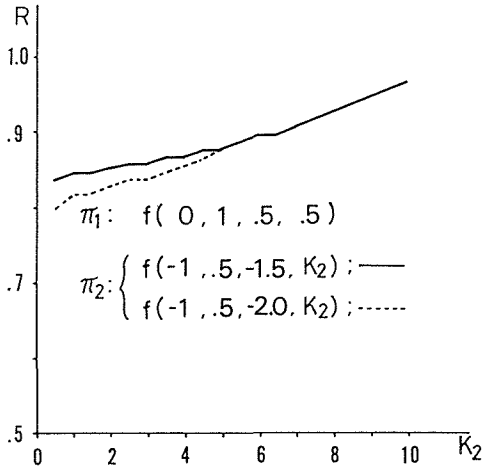

図 4.2　確率の比

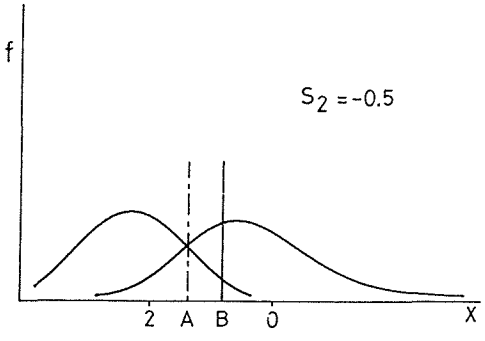

図 4.3　(a)

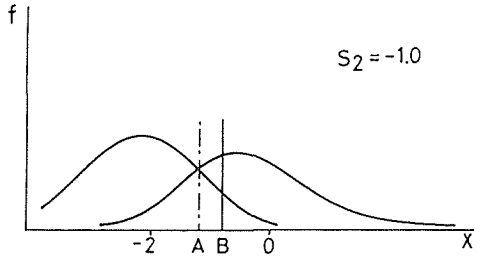

図 4.3　(b)

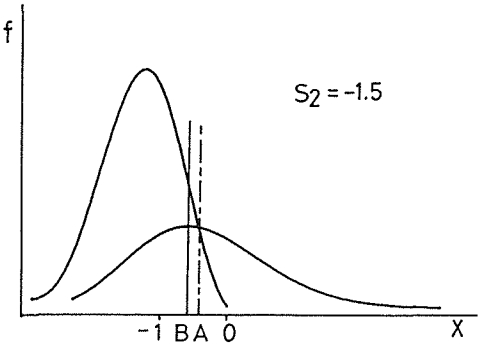

図 4.3　(c)

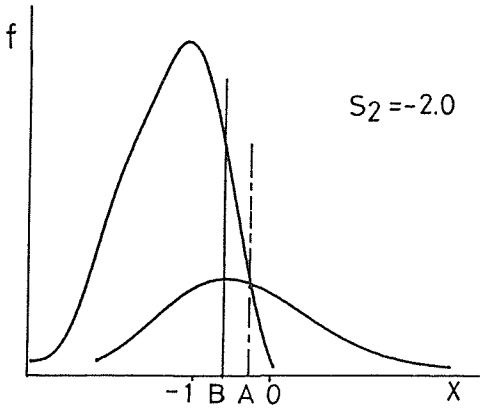

図 4.3　(d)

5．結　　論

母数空間の計量をもとに，２つの分布に関する判別の問題について考察したものである。ここで得られた判別関数は，対称な分布に関して，最尤法の場合と異なり，２つの誤判別の確率，すなわち，π_1からのものをπ_2と誤まる確率$P(2|1)$，およびπ_2からのものをπ_1と誤まる確率$P(1|2)$を等しくするという特徴をもっている。したがって，正規分布に関する判別の問題においては，実際の応用に際して，損失関数の決定等には特に有効である。またここで考察した正規分布に関する判別関数のロバスト性については，歪度，尖度を正規分布からのずれの尺度としたとき，歪度の絶対値が１以内であれば，尖度の値が多少変化しても十分ロバスト性があるといえるが，１より大きくなった場合，尖度の影響を考慮する必要がある。この場合さらにロバスト性を上げるためには損失関数を適当に選ぶことが有効であろう。一方ここで用いた密度関数（４．１）は近似関数であるため，確率の正確な値の計算が，任意のｓ，ｋに対して不可能である。したがってｓｓおよびｋの値の制御が容易な密度関数，あるいは正規分布からの変数変換等を検討する必要がある。

参　考　文　献

1） Rao, C.R. : Bull. Cal. Math. Soc. 37 (1945), pp.81－91.
2） Efron, B. : Ann. Statist. 3 (1975), pp,1189－1242.
3） Dawid, A.P. : Ann. Statist. 3 (1975), Discussion to Efron's paper.
4） Amari, S. : Post-RAAG Rep. 106 (1980), pp.1－53.
5） Clarke, W.R., Lachenburch, P.A. and Broffitt, B. : Communications Statist. A8 (1979), pp.1285－1301.
6） Barndorff-Nielsen,O. : Information and Exponential Families in Statistical Theory, (1978), John Wiley, New York.