HOKKAIDO UNIVERSITY

| | |
|---|---|
| Title | QDFを用いた判別分析の有効性について : 3種の非正規分布のモデルを用いての考察 |
| Author(s) | 中西, 寛子; 水田, 正弘; 佐藤, 義治; 伊達, 惇; 河口, 至商 |
| Citation | 北海道大學工學部研究報告, 111: 75-81 |
| Issue Date | 1982-10-30 |
| Doc URL | http://hdl.handle.net/2115/41760 |
| Type | bulletin (article) |
| File Information | 111_75-82.pdf |

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

北海道大学工学部研究報告
第111号（昭和57年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 111 (1982)

# QDFを用いた判別分析の有効性について

## — 3種の非正規分布のモデルを用いての考察 —

中西寛子* 水田正弘* 佐藤義治* 伊達 惇** 河口至商*

(昭和57年6月30日受理)

# Robustness of Discriminant Analysis with QDF

## —— Study for Three Non-Normal Distribution Models ——

Hiroko NAKANISHI, Masahiro MIZUTA, Yoshiharu SATO,
Tsutomu DATE, and Michiaki KAWAGUCHI

(Received June 30, 1982)

## Abstract

Quadratic discriminant function (QDF) is commonly used for a two-group-discriminant-analysis. This method is optimal when both samples are from normal distribution with known means and variances. When QDF is applied to samples from non-normal distributions, the total probability of misclassification is larger than that of the optimal means (derived from Bayes discriminant rule). W.R. Clarke et al. showed that if non-normal distributions are highly skewed, QDF is sensitive, i.e. the total probability of misclassification of QDF is far from the optimal value.

In this paper, we check Clarke's study by use of 3-type non-normal distributions, and point out that the increase in misclassification depends on not only the skewness of distribution but also the mutual relation of skewness between the two distributions.

Moreover it is shown that this relation is more significant than the skewness itself. In order to investigate the robustness of QDF for non-normal distributions, a new measure is proposed.

## 1. はじめに

2群の判別分析において，各群の真の分布が既知であるとき，ベイズ決定方式を用いれば誤判別の確率の和を最小とする判別関数が得られる。特に各群の分布が正規分布である場合，その判別関数はQDF (Quadratic Discriminant Function) と呼ばれる二次関数となる。

一般に真の分布が未知であるとき，標本より平均と分散を推定し，QDFを用いた判別分析がよく行われる。すなわち，分布を正規分布であると仮定しQDFを用いる。分布が非正規分布で

* 情報工学専攻 情報数理工学第一講座
** 情報科学

ある場合，QDF を用いた判別分析は分布を正規分布と見なすため，真の分布が既知であるときに用いるベイズ決定方式による判別分析より誤判別の確率の和が増加する。この増加は，分布の非正規性や分布間の相互関係によって異なる。したがって，非正規分布のどの様な形に，また，分布のどの様な組み合わせに対して QDF が有効（ロバスト）であるかを調べる必要性が生じる。

この問題に対して，P. A. Lachenbruch et al.[1], W.R.Clarke et al.[2], B.Broffitt et al.[3] が研究を行っている。その結果として「一般に，QDF を用いた判別分析は非対称な分布に対して影響を受けやすい」ということが示されている。しかし，彼等の研究では，2群の組み合わせが歪度（後述）に関して非負のもののみ扱い，結果の妥当性が問われる。

本論文では，QDF による判別分析が非対称な非正規分布のどの様な組み合わせに対してどの程度，誤判別の確率が増すかという問題をより詳しく考察する。ただし，ここでは2群共に1次元データとし，非正規分布の形は単峰であると仮定する。また，QDF を用いたために増加する誤判別の確率を評価する量Hを提案する。

## 2. 以下の議論における基本的知識

以下の議論を進めていく上で必要な事柄を述べる。

### 2.1 ベイズ決定方式

ベイズ決定方式を用いた判別分析は次のようになされる。

$f_1(x)$, $f_2(x)$ を各々群$G_1$，群$G_2$の確率密度関数とする。

$$B(x)=\ln(f_2(x)/f_1(x))$$

とおき，新しく観察された標本$x_0$に対して

$B(x_0)\leqq 0$ のとき $x_0 \in G_1$

$B(x_0)>0$ のとき $x_0 \in G_2$

と判別する。

### 2.2 QDF を用いた判別分析

$f_1(x)$, $f_2(x)$ が正規分布であるとき，$B(x)$ は次のように書き直される。

$$Q(x)=\{(x-\mu_2)^2\sigma_2^{-2}-(x-\mu_1)^2\sigma_1^{-2}+\ln(\sigma_2^2/\sigma_1^2)\}/2$$

ここで，$\mu_i$は群$G_i$の平均，$\sigma_i^2$は群$G_i$の分散である。一般に，群$G_1$から$n_1$個，群$G_2$から$n_2$個の標本が得られたとき$\mu_i$，$\sigma_i^2$は各々

$$\tilde{\mu}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij}$$

$$\tilde{\sigma}_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\tilde{\mu}_i)^2 \qquad (i=1,2)$$

と推定される。ただし，$x_{ij}$は群$G_i$から得られた$j$番目の標本の値である。

### 2.3 歪度（Skewness）

分布の対称性を測る尺度として古くからよく用いられる量として歪度がある。歪度$S$は次のように定義される。

$$S=\mu_3/(\sigma^2)^{3/2}$$

ここで，$\mu_3$は平均のまわりの3次のモーメント，$\sigma^2$は分散である。歪度は平均を中心とする分

布の対称性を見る量で，正規分布の場合は0である。歪度が正（または負）の場合，分布が左（または右）にずれていることを示す。

### 2.4 非正規分布のモデル

N. L. Johnson[4] は正規分布から非正規分布を作る変換方法に関して研究を行った。そのうちの2つの非正規分布：Lognormal分布（$S_L$），Inverse hyperbolic sine normal分布（$S_I$）は，歪度に関して様々な値をとり非正規分布のモデルとして非常に適している。以下にその変換方法を述べる。

平均$\mu^*$，分散$\sigma^{*2}$の正規分布の任意の標本値$x$に対して

（$S_L$）：$y=\exp x$

（$S_I$）：$y=\sinh x$

という変換を行い，新たに値$y$を標本値として見なす。これらの変換によって得られた非正規分布の確率密度関数$f$，平均$M$，分散$V$，歪度$S$は各々

$$(S_L)\quad \begin{cases} f_L=\dfrac{1}{\sqrt{2\pi}\,\sigma^*}\dfrac{1}{y}\exp\left\{-\dfrac{1}{2\sigma^{*2}}(\ln y-\mu^*)^2\right\}, & y>0 \\ \quad=0 & ,\ y\leqq 0 \\ M_L=\exp\left(\mu^*+\dfrac{1}{2}\sigma^{*2}\right) & \\ V_L=M_L(r-1) \qquad \text{ただし}\quad r=\exp(\sigma^{*2}) & \\ S_L=(r-1)^{1/2}(r+2) & \end{cases}$$

$$(S_I)\quad \begin{cases} f_I=\dfrac{1}{\sqrt{2\pi}\sigma^*}\dfrac{1}{\sqrt{y^2+1}}\exp\left\{-\dfrac{1}{2\sigma^{*2}}(\ln(y+\sqrt{y^2+1})-\mu^*)^2\right\} & \\ M_I=\dfrac{1}{2}\left\{\exp\left(\mu^*+\dfrac{\sigma^{*2}}{2}\right)-\exp\left(-\mu^*+\dfrac{\sigma^{*2}}{2}\right)\right\} & \\ V_I=\dfrac{1}{2}(r-1)(2M_I^2+r+1) \qquad \text{ただし}\quad r=\exp(\sigma^{*2}) & \\ S_I=\dfrac{M_I(r-1)^{\frac{1}{2}}\{4M_I^2(r+2)+3(r+1)^2\}}{\sqrt{2}\,(2M_I^2+r+1)^{3/2}} & \end{cases}$$

となる。式より明らかなように，両分布とも平均・分散・歪度のうち2つを固定すると残り1つが決まる。

### 2.5 QDFの有効性を測る評価量Hの提案

QDFを用いた判別分析と，真の分布が既知であるときのベイズ決定方式による判別分析との関係を調べるために次の評価量$H$を考える。

$$H=-P_Q\cdot\log_2 P_Q-(1-P_Q)\cdot\log_2(1-P_Q)$$
$$-\{-P_{OPT}\cdot\log_2 P_{OPT}-(1-P_{OPT})\cdot\log_2(1-P_{OPT})\}$$

ここで，$P_Q=(P_Q^{(1)}+P_Q^{(2)})/2$である。$P_Q^{(i)}$はQDFを用いたとき，群$G_i$から得られた標本が他群より観察されたと誤って判断される確率である。また，$P_{OPT}=(P_{OPT}^{(1)}+P_{OPT}^{(2)})/2$で，$P_{OPT}^{(i)}$はベイズ決定方式を用いたときの群$G_i$に対する誤判別の確率である。

この評価量$H$はエントロピーの概念を用いたもので，QDFを用いることによって真の分布が既知であるとき用いるベイズ決定方式よりもどの程度，判別におけるあいまいさが増したかという尺度となる。$H$の定義式からわかるように，値$H$が0に近いほどQDFが有効であることを示

している。

## 3. 非正規分布（$S_L$）（$S_I$）を用いた実験

2群の分布が非正規分布である場合，真の分布が既知であるときのベイズ決定方式による判別分析と，非正規分布を正規分布と見なして取り扱うQDFによる判別分析との間の関係を調べるための非正規分布（$S_L$）（$S_I$）をモデルとした実験のアルゴリズムを述べる。

Step 1.　群$G_1$の平均，分散，歪度を固定する。

Step 2.　群$G_2$の平均を固定する。ただし，群$G_1$の平均よりも大とする。

Step 3.　$r=1.0$ とし，群$G_2$の歪度を決める（このとき，分散も同時に決まる）。

Step 4.　条件Ⅰ，Ⅱ，Ⅲ（後述）を満たしているかどうかを調べる。満たしていない場合はStep 8に行く。また，歪度の値が大で条件Ⅲが満たされない場合はStopする。

Step 5.　2群の平均と分散よりQDFを作り判別を行う。そのときの誤判別の確率$P_Q$を計算する。

Step 6.　真の分布を用いてベイズ決定方式で判別を行う。そのときの誤判別の確率$P_{OPT}$を計算する。

Step 7.　$P_Q$と$P_{OPT}$より値$H$を計算する。

Step 8.　群$G_2$の歪度を変え（$r \leftarrow r+0.001$）Step 4にもどる。

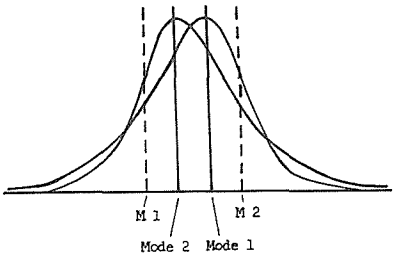
最後に縦軸に値$H$，横軸に歪度を取ってグラフを描く。これらの作業において経験的に判別分析が不可能，または，不適当と思われる状態を故意に除いた。すなわち，次の3つの条件を満たしている場合のみStep 5からStep 7の作業を行った。

条件Ⅰ.　群$G_1$のモードが群$G_2$のモードより小さい。

条件Ⅱ.　$|\mu_1-\mu_2|<L$，$L=0.5\times\sqrt{\min(\sigma_1^2,\sigma_2^2)}$。

条件Ⅲ.　$K\leqq 18$，$K=(\sigma_1^2+(\mu_1-\mu_2)^2)/2\sigma_2^2+(\sigma_2^2+(\mu_1-\mu_2)^2)/2\sigma_1^2-1$。

条件Ⅰでは図1のような平均とモードの関係が逆になる状態を排除した。条件Ⅱでは2つの群が近づきすぎることを排除している。また，条件Ⅲの値$K$はKullbackのダイバージェンス（ただし，この場合2群が正規分布$N(\mu_1,\sigma_1^2)$，$N(\mu_2,\sigma_2^2)$であることが仮定されている。詳しくは参考文献5）を参照のこと）であり，この条件によって2つの群が離れすぎることを禁じている。




M 1　: Mean of Group 1
M 2　: Mean of Group 2
Mode 1 : Mode of Group 1
Mode 2 : Mode of Group 2

図1　条　件　Ⅰ

## 4. 実　験　結　果

前章で述べた実験を約20ケースについて行った。そのうちの代表的なもの5つを図で示す。図2，3は（$S_L$）に対して行ったもので，図4．5，6は（$S_I$）について行ったものである。各グラフにおいて歪度の値は約1000点動いている。図の上に示された値$M$，$V$，$S$は各々平均,分散,歪度である。これらの結果をまとめると次のようになる。

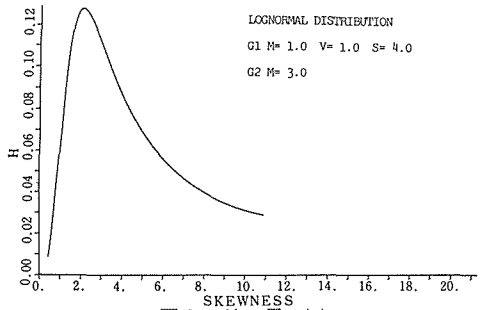


図2　結　果　(1)

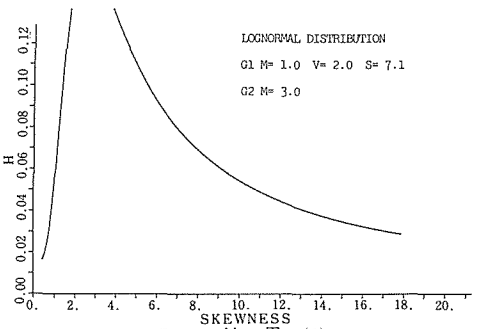


図3　結　果　(2)

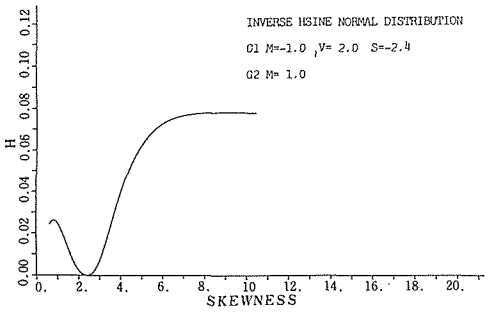


図4　結　果　(3)

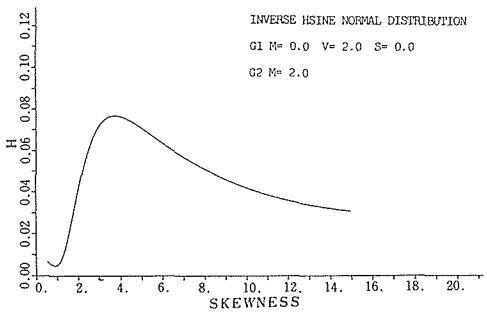


図5　結　果　(4)

(i)　歪度と値$H$との間の関係は単調でない。特に図4（群$G_1$の歪度の値が-2.4）では，群$G_2$の歪度が2.4で値$H$が0を示している。つまり，歪度そのものの値だけでなく2群の歪度に関する相互関係も値$H$に影響を与えている。

(ii)　($S_L$)のほうが($S_I$)より値$H$を大きくすることがある。これは，歪度の値が同じであっても実際に描くと($S_L$)と($S_I$)とではトガリ具合の異なった分布の形を示すためである。

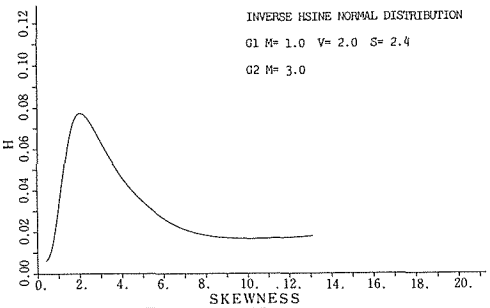


図6　結　果　(5)

(i)のことをより詳しく調べるために平均・分散・歪度が独立である三角分布を用いて実験を行った。次章にこの実験について述べる。また，(ii)に関しても($S_L$)と($S_I$)の分布の違いを調べることで値$H$に影響を与える歪度とその相互関係以外の要因がわかるであろう。

## 5.　三角分布を用いた実験と結果

三角分布に対して図7のように$a$，$b$，$\alpha$をおく。このときの三角分布の確率密度関数$f$，平均$M$，分散$V$，歪度$S$は

$$f=\frac{2}{ab}(x-\alpha) \qquad , \ \alpha \leqq x \leqq b+\alpha$$

$$=\frac{-2}{a-b}\left(\frac{x-\alpha-a}{a}\right), \quad b+\alpha<x\leqq a+\alpha$$

$$M=(a+b)/3+\alpha$$

$$V=(a^2-ab+b^2)/18$$

$$S=(a+b)(a-2b)(2a-b)/270V^{3/2}$$

となる。また，歪度は$-0.5657$～$0.5657$の範囲を動く。

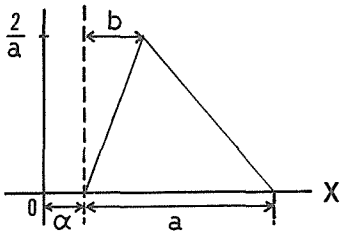


図7　三角分布

三角分布に対しても，($S_L$) 等と同様の方法で実験を行った。ただし，常に群$G_1$の平均は0とし，群$G_2$の平均は1とした。群$G_1$の分散・歪度，群$G_2$の分散の値に関して125種の組み合わせを考えた。そのうちのいくつかを図で示す（図8）。図の上に群$G_1$の分散・歪度，群$G_2$の分散を示した。また，参考として歪度が最小値をとるときの値$H$と，歪度が最大値をとるときの値$H$を示した。

これら125種の実験を詳しく調べ考察を行った。それをまとめると以下のようになる。

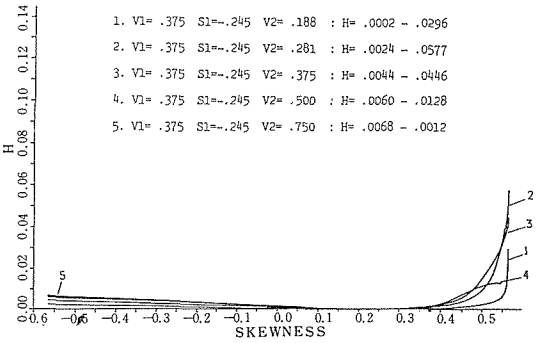


図8　三角分布における結果

(1)　群$G_1$の歪度と群$G_2$の歪度との和が0に近いとき値$H$は0に近い。これは，歪みが対称である場合，歪度の値に関係なく値$H$が0でありQDFが有効であることを示している。

(2)　群$G_2$の分散が非常に小さく，歪度が大きいとき値$H$は大きいことがある。これは，判別点の移動に非常に敏感であることを示している。

(3)　一方の分散が非常に大きいとき値$H$は0に近い。これは，真の分布が既知である状態においても誤判別の確率が大きく判別点の移動にあまり関係しないことを示している。

一般には(1)，(2)，(3)のことが言え，詳細に調べると2群の歪みの組み合わせ4種類に対して次のことが言える。

(4)　群$G_1$の歪度が0のとき群$G_2$の歪度が大きいほど値$H$が大きい。

(5)　群$G_1$の歪度が負，群$G_2$の歪度が正のとき一方の分散が大きいと値$H$は0に近い。また，群$G_2$の歪度が大きくかつ群$G_1$の歪度の絶対値が大きいときも値$H$は0に近い。その他の場合は値$H$は少し大きい。

(6)　群$G_1$の歪度と群$G_2$の歪度が共に正の場合，群$G_2$の歪度が小さいと値$H$は0に近い。特に群$G_1$の歪度が非常に大きいと値$H$も大きくなり，他の場合，値$H$は少し大きい。また，群$G_1$の歪度と群$G_2$の歪度が共に負の場合，両方の歪度の符号を変え，群$G_1$の歪度と群$G_2$の歪度を入れ換えると上記と同じことが言える。

(7)　群$G_1$の歪度が正，群$G_2$の歪度が負のときは値$H$は0に近い。

値$H$に対する影響力は(1)，(2)，(3)の順で大きく(4)，(5)，(6)，(7)は同列である。これらのことを考慮してQDFの有効性を見る樹木表現（図9）を示してみた。ここで群$G_1$，群$G_2$の平均は各々0，1である。また，A，B，C，Dは値$H$を次のように分類したときの段階評価である。

A：$0.00\leqq H<0.04$

B：$0.04\leqq H<0.07$

C：$0.07\leqq H<0.10$

D：$0.10 \leqq H$

($S_L$)，($S_I$) の非正規分布でも同様の影響力に関するまとめ (1) ～ (7) が得られた。しかし，厳密には各非正規分布族に固有の特徴がありその点を考慮しなければならない。つまり，同様の樹木表現を行うには分散の大小，歪度の大小をその非正規分布族によって変えねばならない。

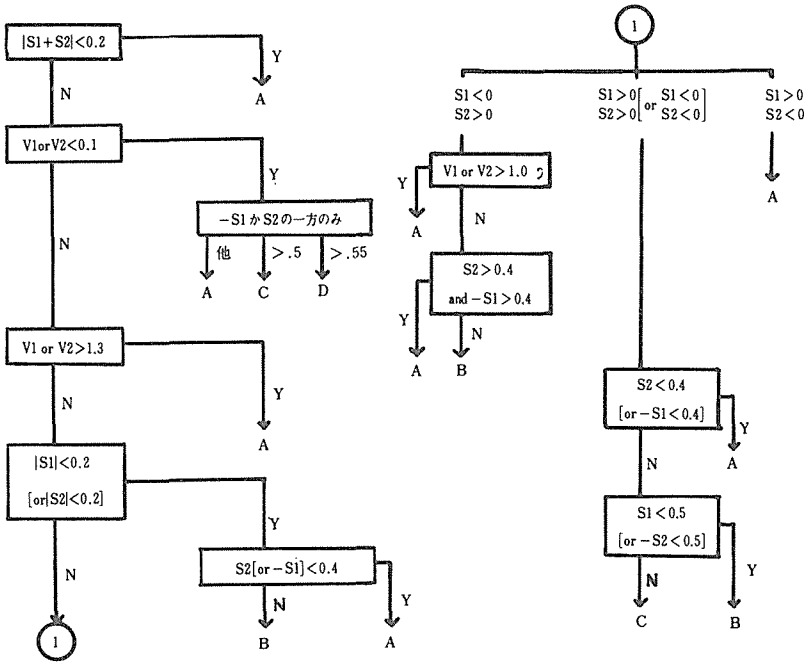



図9　三角分布の結果より得られた樹木表現

## 6. おわりに

QDFの有効性は単に歪度だけに関係するのでなく，分散にもまた2群の歪度の状態にも関係することを示した。三角分布に対する詳しいまとめは図9で示した樹木表現のようになる。($S_L$)，($S_I$) に関する樹木表現も同様に得られた。分散，歪度の値は異なるがそのおおまかな傾向は三角分布に対するものと同様である。

本論文では，非正規分布のモデルとして ($S_L$)，($S_I$)，三角分布を用いたが，これらが全ての非正規分布を代表しているとは言い難い。すなわち，ある非正規分布をモデルとして用いるとき，その分布の特徴をとらえておかねばならない。今回の場合も ($S_L$)，($S_I$)，三角分布の非正規性に対する細かい議論がなされるべきである。その上でより詳しい結果が得られるであろう。

## 参考文献

1) Lachenbruch, P. A., Sneeringer, C. and Revo, L. T.: *Commun. Statist.*, 1 (1973), 1, pp. 39-56.
2) Clarke, W. R., Lachenbruch, P. A. and Broffitt, B.: *Commun. Statist.*, A8 (1979), 13, pp. 1285-1301.
3) Broffitt, B., Clarke, W. R. and Lachenbruch. P. A.: *Commun. Statist.*, A9 (1980), 1, pp. 13-25.
4) Johnson, N. L.: *Biometrika*, 36 (1949), pp. 149-176.
5) Kullback, S.: Information Theory and Statistics, (1978), Dover Publications, Inc.