



HOKKAIDO UNIVERSITY

Title	多次元データの図形表示について
Author(s)	新保, 勝; 宮腰, 政明
Citation	北海道大學工學部研究報告, 96, 65-71
Issue Date	1979-11-30
Doc URL	https://hdl.handle.net/2115/41587
Type	departmental bulletin paper
File Information	96_65-72.pdf



多次元データの図形表示について

新保 勝* 宮腰 政明

(昭和54年6月30日受理)

On a Graphical Representation of Multidimensional Data

Masaru SHIMBO and Masaaki MIYAKOSHI

(Received June 30, 1979)

Abstract

A reduction technique of plotting multidimensional data on a plane is proposed with reference to Fourier series expansion technique by D. F. Andrews, where the data represent the sampled components of the frequency spectrum. Based on the duality between time and frequency, the same components are regarded here as sampled values of a curve in time domain and the plotted points are connected by applying the wellknown *sampling theorem*. This enables us to read the components of multidimensional data more easily on the plane as well as to cluster the corresponding curves into several groups interactively. The trajectories of project vectors in these reduction techniques are also shown.

1. 序 論

2次元以上の多次元のデータを、それらのデータの持っている情報を出来るだけ損失する事なしに、元の次元より、より低次元の空間に表示する事は、元の多次元データに対する大局的把握を可能にする点で注目される。この様な多次元を低減する方法は、種々考えられているが、その主な手法は、主成分分析、Sammon の非線形写像法¹⁾、Andrews の高次元データプロット法^{2,3)}および吉田のスパイラル・マッピング⁴⁾である。Sammon の方法は、一種の多次元尺度構成法であって、データ間の距離関係に注目して、この距離関係を可能な限り保存するように低次元での点の配置を決める方法である。これに対して、Andrews の方法は、元の多次元空間内を動径ベクトルで掃引して、その動径ベクトルへの各点(データ)の射影値をプロットすることにより、多次元の点を平面上の一つの曲線で表わすという方法である。吉田の方法も Andrews の方法と基本的に同様であるが、動径ベクトルが超球面上を螺旋的に動く点が異っている。本論文では、Andrews による方法で平面にプロットされた曲線が多次元データの情報をプロットされる曲線の周波数スペクトル成分という形式で、データの座標成分を取り込んでいる点に注目し、時間と周波数の双対性を考慮して、新たに、ある時間曲線の標本値として座標成分を与え、標本値間を標本化定理によって補間した曲線により多次元のデータを平面上の曲線として表わす方法を考察する。

* 情報工学専攻 情報処理工学講座

2. 手法の概説

Andrews の方法は、与えられた p 次元のデータ $\mathbf{X}=(X_1, X_2, \dots, X_p)$ を動径ベクトル $\mathbf{t}(t)=(1/\sqrt{2}, \sin t, \cos t, \dots, \frac{\sin}{\cos}[p/2]t)$ に射影して、パラメータ t の値に対する射影値 $f_{\mathbf{X}}(t)$ をプロットするものである。したがって、 $f_{\mathbf{X}}(t)$ は

$$f_{\mathbf{X}}(t)=\frac{X_1}{\sqrt{2}}+X_2 \sin t+X_3 \cos t+\dots+X_p \frac{\sin}{\cos}\left[\frac{p}{2}\right] t \quad -\pi \leq t \leq \pi \quad (1)$$

と表わされる。ただし、 p 次元目に対する座標成分 X_p については、 p が偶数ならば正弦項、奇数ならば余弦項をとる。また、 $[\]$ はガウス記号である。Andrews の方法は、データの平均、距離を保存し、分散は t の値に対して可変である。特に点のユークリッド距離 $\|\mathbf{X}-\mathbf{Y}\|$ については、

$$\|f_{\mathbf{X}}(t)-f_{\mathbf{Y}}(t)\|_{L_2}^2=\int_{-\pi}^{\pi}|f_{\mathbf{X}}(t)-f_{\mathbf{Y}}(t)|^2 dt=\pi\|\mathbf{X}-\mathbf{Y}\|^2 \quad (2)$$

が成り立つ。(2) は、視覚的に人間が2つの曲線の近さを $\|\ \|_{L_2}$ で測るとすると、ユークリッド距離の近い点は、曲線も近いし、逆も言える事を意味する。

ところで、(1) の曲線を周波数スペクトル成分 $A(w)$ についてみると、

$$A(0)=|X_1|, \quad A(2\pi)=\sqrt{X_2^2+X_3^2}, \quad A(4\pi)=\sqrt{X_4^2+X_5^2}, \dots$$

となる。ただし、 p が偶数の時、 $A(p\pi)$ は座標成分 X_p だけを含む。これより曲線 $f_{\mathbf{X}}(t)$ が点 $\mathbf{X}$ についての情報を $A(w)$ の成分の形で取り込んでいて、 $-\pi \leq t \leq \pi$ の間の曲線全体が点 $\mathbf{X}$ についての情報を担っている事がわかる。このことから、逆に周波数スペクトル $A(w)$ に対する座標成分の取り込まれ方を決める事により一つの動径ベクトルを決定する。

すなわち、周波数スペクトル $A(w)$ に p 次元の点 $\mathbf{X}=(X_1, X_2, \dots, X_p)$ が

$$A(w)=A(w; \mathbf{X}) \quad (3)$$

なる形で取り込まれるとすると、このときの動径ベクトル $\mathbf{t}$ への点 $\mathbf{X}$ のプロットされる曲線 $f_{\mathbf{X}}(t)$ は

$$f_{\mathbf{X}}(t)=F^{-1}(A(w; \mathbf{X})) \quad (4)$$

で与えられる。ここで F^{-1} は Fourier 逆変換である。

Andrews の方法は、 $A(w; \mathbf{X})$ の与えかたが、 $w_i=i\pi$ について

$$A(0)=|X_1|, \quad A(i\pi)=\sqrt{X_i^2+X_{i+1}^2} \quad i=2, 4, \dots$$

となっていて、 $i\pi$ 周波数成分を決めるのに2つの座標成分 X_i, X_{i+1} が寄与している。今、 $A(w; \mathbf{X})$ の与え方を種々変えて、Andrews の方法のように、 $i\pi$ 周波数成分を決定するのに2つの座標成分を用いずに、唯一つの座標成分によって決定されるようにも定められる。すなわち、 $w_i=2\pi(i-1)$ または $w_i=\pi(i-1)$ に対して

$$A(w_i; \mathbf{X})=|X_i| \quad i=1, 2, \dots, p \quad (5)$$

となるように、 $a_i^2+b_i^2=X_i^2$ を満すように a_i, b_i を決める方法を、下の3つの場合について考え、その場合の対応する動径ベクトルと、パラメータ t の区間を同時に示す。

(i) $a_i=b_i=X_i/\sqrt{2}$ のとき。

$$\mathbf{t}(t)=(1, \sin(t+\pi/4), \sin(2t+\pi/4), \dots, \sin\{(p-1)t+\pi/4\}), \quad -\pi \leq t \leq \pi \quad (6)$$

(ii) $a_i=X_i, b_i=0$ のとき。

$$\mathbf{t}(t)=(1, \cos t, \cos 2t, \dots, \cos(p-1)t), \quad 0 \leq t \leq \pi \quad (7)$$

(iii) $a_i=0, b_i=X_i$ のとき。

$$\mathbf{t}(t) = (\sin t, \sin 2t, \dots, \sin pt), \quad 0 \leq t \leq \pi \quad (8)$$

特に, $A(w; \mathbf{X})$ の与えかたとして, パラメータ T を与えると, $2WT=p$ となる W について,

$$(iv) \quad A(w; \mathbf{X}) = \frac{1}{2\sqrt{2\pi W}} \left\{ \left(\sum_{i=1}^p X_i \cos \frac{iw}{2W} \right)^2 + \left(\sum_{i=1}^p X_i \sin \frac{iw}{2W} \right)^2 \right\}^{1/2} \quad |w| \leq 2\pi W$$

$$= 0 \quad |w| > 2\pi W$$

と与えると, 動径ベクトル $\mathbf{t}$ は,

$$\mathbf{t}(t) = \left(\frac{\sin \pi(2Wt-1)}{\pi(2Wt-1)}, \frac{\sin \pi(2Wt-2)}{\pi(2Wt-2)}, \dots, \frac{\sin \pi(2Wt-p)}{\pi(2Wt-p)} \right) \quad 0 \leq t \leq T \quad (9)$$

となる。この場合, プロットされる曲線 $f_{\mathbf{x}}(t)$ の $t=i/2W$ ($i=1, 2, \dots, p$) ごとの曲線の値は $f_{\mathbf{x}}(i/2W)=X_i$ となる。これは, (i), (ii), (iii) が (5) で示されるように, 周波数成分 w_i で $A(w_i; \mathbf{X})=|X_i|$ となった事に対応して, パラメータ領域 (時間領域) で, この事を示している。なお, Andrews の場合と同様に, (6), (7), (8), (9) の場合, 動径ベクトル $\mathbf{t}$ の各成分 $\phi_i(t)$ は各々の場合の区間に関して, 直交関数となっていて, 次の事が成り立つ。

$$\|f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)\|_{L_2} = \|f_{\mathbf{x}-\mathbf{y}}(t)\|_{L_2} = k \|\mathbf{X} - \mathbf{Y}\|$$

ただし, 正数 k は $\int_a^b \phi_i(t)\phi_j(t)dt = k^2 \delta_{ij}$ ($i, j=1, 2, \dots, p$) であり, 区間 $[a, b]$ は各場合の t の区間である。これは, 2つの曲線が近ければ, 対応する2つの点のユークリッド距離も近く, 逆も言える事を意味する。

3. 実験例と結果

$A(w; \mathbf{X})$ の与えかたによって呈示される曲線の相異を見るために, Andrews の論文²⁾に引用されている Ashton 他³⁾の人類学的データを引用する。データは, 人類2種 (英国人, オーストラリア原住民), 猿3種 (ゴリラ雄雌, オランウータン雄雌, チンパンジー雄雌), 人類化石2種 (パラントロプス・クランデンス, ピテクントロプス・ベキネンシス) について, 門歯, 犬歯, 小臼歯, 臼歯についての測定を行なったもので, 次元数は6であり, データ数は10である。大局的な見かたから言うと, 人類2種, 人類化石2種の4種と猿3種という様に2つの群に曲線群が分かれる事が期待される。データの数値は表-1にて示される。これらのデータについての各手法によって呈示される曲線 $f_{\mathbf{x}}(t)$ は図-1 から図-6 で示される。各図の説明は以下の通りである。

図-1 Andrews の方法 $\mathbf{t} = (1/\sqrt{2}, \sin t, \cos t, \sin 2t, \cos 2t, \sin 3t)$

$$-\pi \leq t \leq \pi, \quad A(0) = |X_1|, \quad A(i\pi) = \sqrt{X_i^2 + X_{i+1}^2} \quad i=2, 4, \dots$$

動径ベクトル $\mathbf{t}$ の軌跡を3次元の場合, 図-7 に示す。

図-2 (i) の方法 $\mathbf{t} = (1, \sin(t+\pi/4), \sin(2t+\pi/4), \dots, \sin(5t+\pi/4))$

$$-\pi \leq t \leq \pi, \quad A(2\pi(i-1)) = |X_i| \quad i=1, 2, \dots, 6$$

動径ベクトル $\mathbf{t}$ の軌跡を3次元の場合, 図-8 に示す。

図-3 (ii) の方法 $\mathbf{t} = (1, \cos t, \cos 2t, \dots, \cos 5t)$

$$0 \leq t \leq \pi, \quad A(\pi(i-1)) = |X_i| \quad i=1, 2, \dots, 6$$

図-4 (iii) の方法 $\mathbf{t} = (\sin t, \sin 2t, \dots, \sin 6t)$

$$0 \leq t \leq \pi, \quad A(\pi i) = |X_i| \quad i=1, 2, \dots, 6$$

図-5 (iv) の方法 $\mathbf{t} = \left(\frac{\sin \pi(0.6t-1)}{\pi(0.6t-1)}, \frac{\sin \pi(0.6t-2)}{\pi(0.6t-2)}, \frac{\sin \pi(0.6t-3)}{\pi(0.6t-3)} \right),$

$$\frac{\sin \pi(0.6t-4)}{\pi(0.6t-4)}, \frac{\sin \pi(0.6t-5)}{\pi(0.6t-5)}, \frac{\sin \pi(0.6t-6)}{\pi(0.6t-6)}\bigg)$$

$$0 \leq t \leq 10, \quad A(w; \mathbf{X}) = \frac{1}{0.6\sqrt{2\pi}} \left\{ \left(\sum_{i=1}^6 X_i \cos \frac{iw}{0.6} \right)^2 + \left(\sum_{i=1}^6 X_i \sin \frac{iw}{0.6} \right)^2 \right\}^{1/2}$$

このとき、 $T=10$ としたので $W=0.3$ となる。

動径ベクトル $\mathbf{t}$ の軌跡を 3 次元の場合、図-9 に示す。

図-6 スパイラル・マッピング

$$\mathbf{t} = (\cos t, \sin t \cos 2t, \sin t \sin 2t \cos 4t, \sin t \sin 2t \sin 4t \cos 8t, \\ \sin t \sin 2t \sin 4t \sin 8t \cos 16t, \sin t \sin 2t \sin 4t \sin 8t \sin 16t)$$

$$0 \leq t \leq \pi/2$$

パラメータ $N=2$ とした。

図-1, 図-2は基本的に相似した曲線群が描かれており、曲線全体は1, 2, 9, 10と3, 4, 5, 6, 7, 8との2つの群に大別されうる。ただし、1, 2, 9, 10の曲線群の中で9は特異な動きをするように思われる。また曲線1, 2と曲線3, 4と曲線7, 8は互いに区間全体を通じて似た動きをする。図-3, 図-4についても図-1, 図-2の場合と同様な事が言えて、曲線は大別して1, 2, 9, 10と3から8までの2つの群に分かれる。図-5, 図-6は前述のプロット図と大分、印象が異なるが、図-5は、曲線の全体が区間を通じて横軸より下方にあるものと上方にあるものの2つの群に大別され、このとき曲線9は、やはり、1, 2, 10の群の中でも特異な動きをしている。図-5では縦軸と平行な直線と曲線の交点がその座標成分の座標値を示しているので曲線全体の動きから点の座標成分の符号を知る事が出来る。図-6も曲線はかなり複雑な動きをしていて、中間部ではその印象が強い。しかし、曲線はやはり全体の動きに注目すれば、1, 2, 9, 10と3から8までの2つの群に分かれうる。

このように、 p 次元のデータを平面上の曲線として表示する場合、 $A(w; \mathbf{X})$ に取り込まれる

表-1

Source	Tooth Type					
	1	2	3	4	5	6
1. British	-5.35	-7.07	-9.37	-4.28	-2.15	-2.93
2. Australian aboriginal	-3.93	-6.04	-8.87	-2.16	-0.50	-1.09
3. gorilla male	3.12	6.66	6.28	4.96	4.13	4.60
4. gorilla female	1.45	1.73	4.82	3.96	3.35	3.63
5. orang-outang male	2.83	5.10	5.11	2.70	1.21	1.49
6. orang-outang female	1.49	1.63	3.60	1.29	-0.17	0.05
7. chimpanzee male	0.38	3.82	3.46	-1.65	-2.32	-1.92
8. chimpanzee female	0.01	0.23	3.05	-2.25	-2.65	-2.15
9. <i>Paranthropus crassidens</i>	-4.52	-6.49	-7.79	3.45	4.91	3.72
10. <i>Pithecanthropus pekinensis</i>	-1.81	-2.94	-6.73	-0.36	-1.32	1.09

Tooth Types

- 1-permanent lower second incisor
- 2-permanent lower canine
- 3-permanent lower first molar
- 4-permanent upper second premolar
- 5-permanent lower first molar
- 6-permanent lower third molar

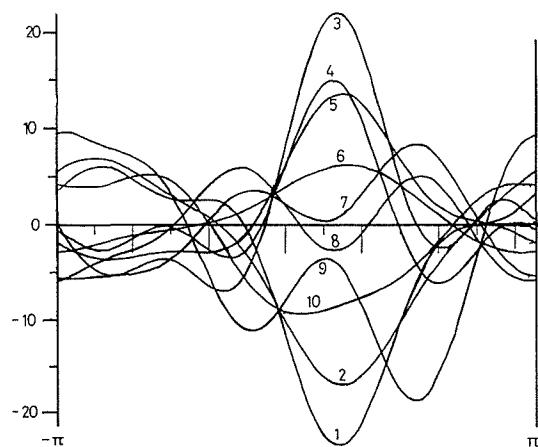


図-1 Andrews の方法によるプロット

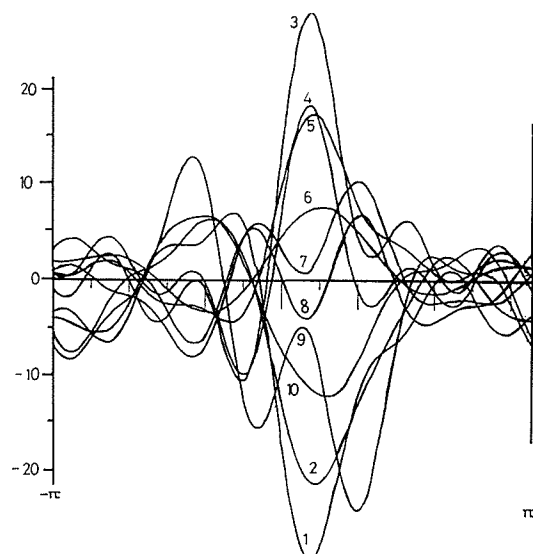


図-2 (i) の方法によるプロット

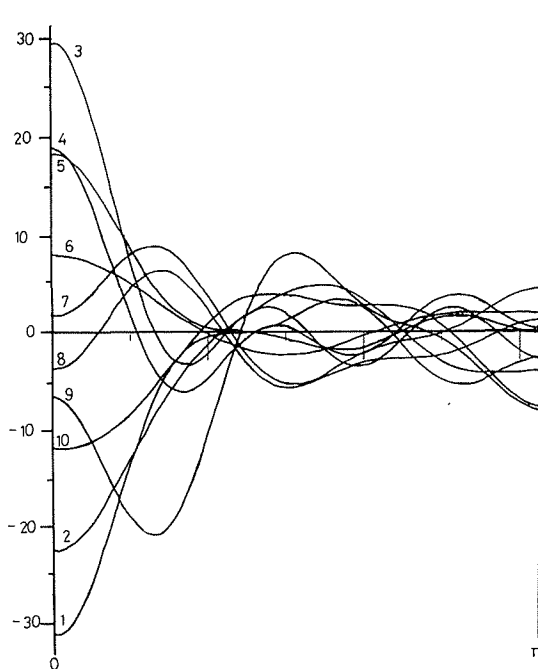


図-3 (ii) の方法によるプロット

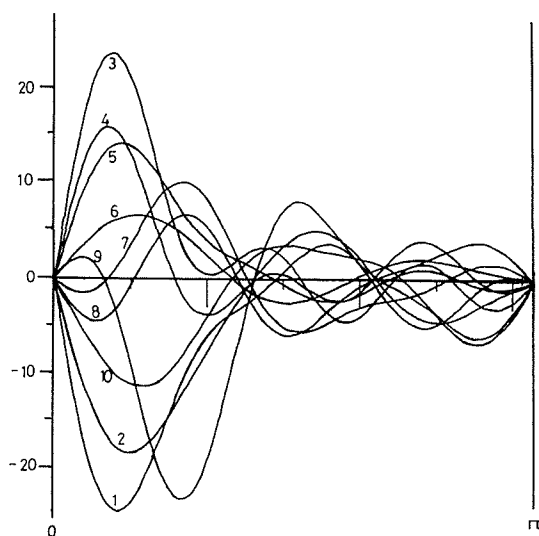


図-4 (iii) の方法によるプロット

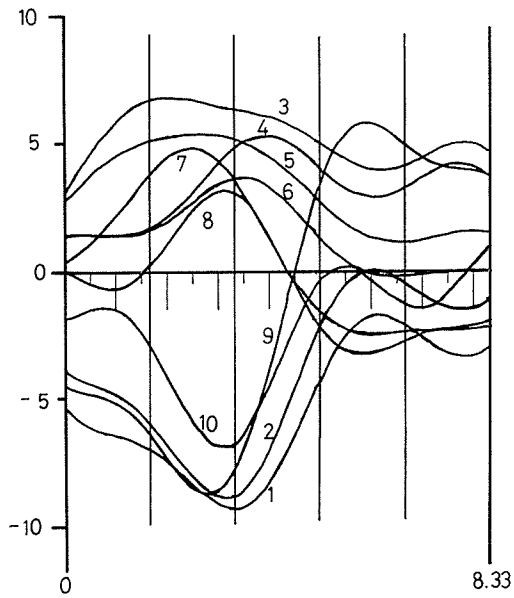


図-5 (iv) の方法によるプロット

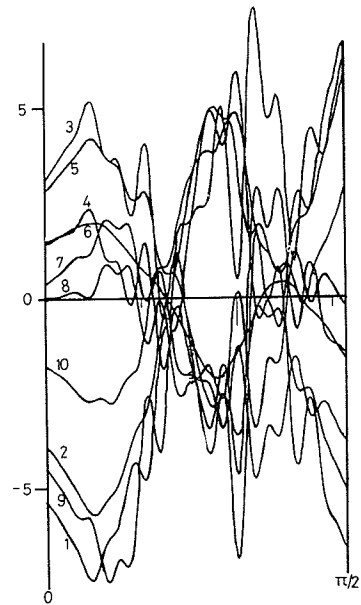


図-6 スパイラル・マッピングによるプロット

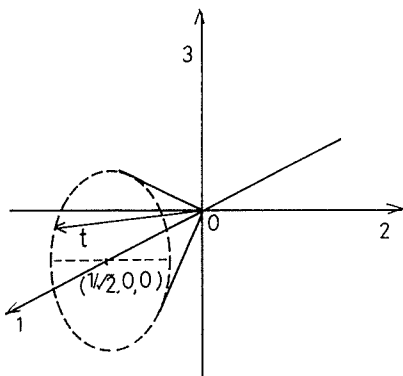


図-7 3次元での Andrews の動径ベクトルの軌跡

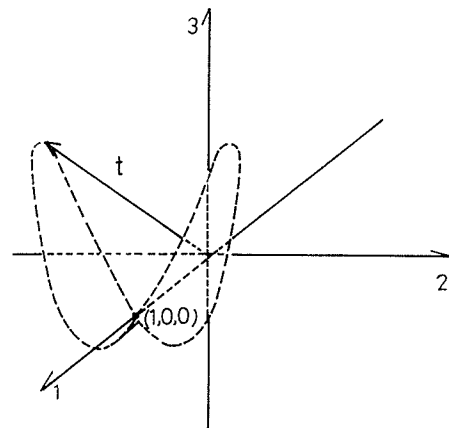


図-8 3次元での (i) の方法の動径ベクトルの軌跡

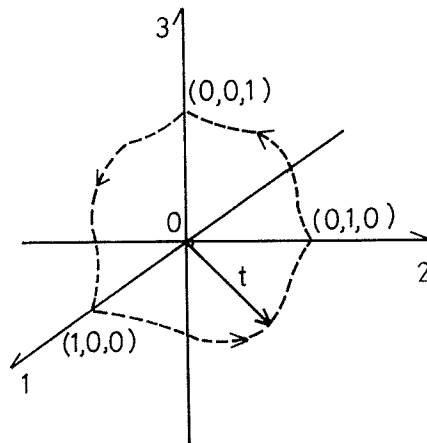


図-9 3次元での (iv) の方法の動径ベクトルの軌跡

座標成分を指定する事によって種々の曲線群をうるし、これらの曲線群の人間の視覚による近さによって元の多次元での点の近さについての情報をえて、点どうしの近さを知る事が出来、多次元のデータ間の距離を計算する事なしに多次元データについての大局的なクラスターをとらせる事が出来る。データについて何らも事前の情報がない場合、全空間を動径ベクトルで掃引する事が望ましいが、Andrews が指摘²⁾しているように、動径ベクトルの掃引する領域が広くなる事と表示される曲線が単純である方が人間にとって見易いという事は、両立しないし、また、(3)、(4) 式によって曲線全体が座標成分について情報を含んでいても、人間の視覚によってその情報がどれだけ抽出されうるか分からないので、この様な手法は、あくまでもデータの前処理的な段階で使われるべきもののように思われる。

尚、計算並びにプロット図作成に際しては、北海道大学大学院情報工学専攻電子計算機システム ECLIPSE S-200 を使用した。

参 考 文 献

- 1) J. W. Sammon: IEEE Transactions on Computers, C-18, 5 (1969) 401.
- 2) D. F. Andrews: Biometrics, 28 (1972) 125.
- 3) D. F. Andrews: Discriminant Analysis and Applications (T. Cacoullous ed.), Academic Press (1972) 37.
- 4) 吉田秀逸: 北海道大学大学院工学研究科修士論文 (昭和 52 年).
- 5) E. H. Ashton et al.: Proc. Roy. Soc. London, B 146 (1957) 552.