データ解析における主要点の特性に関する研究

北海道大学大学院 工学研究科 システム情報工学専攻

清 水 信 夫

データ解析における主要点の特性に関する研究

北海道大学大学院　工学研究科　システム情報工学専攻

清水 信夫

目 次

図 目 次

第 1 章

はじめに

1.1 本研究の背景

ある地域において最適な施設の配置を見出す「最適施設配置問題」は、有史以来、さまざまな時代においてさまざまな形で考えられ続けてきた。しかも、高度情報化社会という言葉がすっかり定着してしまった感のある今日においては、日常生活において施設の数や種類が急速に増え、配置対象となる地域における状況もますます複雑化、多様化している。さらに、地理情報処理、VLSI 設計、パターン認識、コンピュータ・ビジョンなど、最適施設配置問題と密接に関連しているか、近い将来結びつく可能性の高い分野が急速に増えつつあり、これらの問題を効率的に解く重要性はますます高まっている。その重要性に呼応して、最適施設配置問題を解くための方法も数理的手法を軸にいろいろと開発され、「最適配置の理論 [71]」として体系付けられるようになってきた。

最適施設配置問題の解決にあたっては、組み合わせ構造や幾何学的図形の効率的な処理が大きな重要性をもつが、最近では、計算機の急速な進歩と普及によって、これまでの各種数理的手法によるアプローチに加え、計算幾何学的手法が問題の解決にあたって広く採り入れられるようになり、計算アルゴリズムの改良が大幅に進められた。その結果、精度の高い解を従来よりも短い時間で導出できるようになっている。しかしながら、現在も進んでいる情報化や、計算機の能力の向上により、最適施設配置問題はますます多様化、複雑化、大規模化しており、そのような状況に応じた高精度かつ高速な解法の開発や、それ

を支える理論の構築は、現在においても重要な課題である。

最適施設配置問題においては、定式化された問題におけるボロノイ図を作成し、その利用により解を導出する方法がよく利用されている。k個の施設の配置を考えるにあたって、各々の施設から最も近い領域を幾何学的に示した図形を考え、それに基づいて配置におけるコストの最小化による最適化を図るものである。

一方、Flury[21] により提案された主要点 (Principal Points) は、確率分布における密度関数を k個の領域に分割する際の各領域の代表点のうち、それらの点からの 2 乗距離の期待値 (目的関数) が最小となるような k個の点として定義されている。数学的にはクラスター分析における k-means 法と同等のアルゴリズムにより導出されるが、Principal Points の概念は、確率分布を k個の代表点により表すという観点で、「最適配置の理論 [71]」の基礎理論となり得る。また、Principal Points の導出手法は、天気図の解析 (村木・大瀧・水田 [85][86])、層別逆回帰法 (Sliced Inverse Regression) の改良 (大瀧・藤越 [70]) などにも応用されており、さらに主座標曲線 (Principal Curves) や関数データ解析 (Functional Data Analysis) への拡張 (水田 [82][84]) なども試みられている。一方、最近では、確率分布から生成された標本に基づいて Principal Points を推定する研究や、2 乗距離以外の距離を最小とした場合の Principal Points を求める研究も行われている (浅野・水田・佐藤 [65][66])。以下では、k個の主要点を k-Principal Points と表す。

期待値に関して対称な 1 変量確率分布における k-Principal Points は期待値に関して点対称となる (対称性をもつ) ことが予想される。この予想が正しい場合は、分布の対称性を利用して導出に必要な領域を縮小化することにより k-Principal Points の導出に要する時間を短縮することが可能になるが、常に正しいわけではない。このような確率分布において、Flury[21] は、期待値に関して点対称な 2 点からの 2 乗距離を極小にする必要条件および十分条件を導出し、さまざまな分布における 2-Principal Points の値を数値計算により求めている。これらの結果には、2-Principal Points が非対称となる例も含まれている。

対称な 1 変量確率分布における 2-Principal Points の対称性に関しては、密度関数の性質に着目した研究も行われている。Tarpey[61] は、密度関数が対称かつ強単峰 (strongly unimodal) であれば対称性が成立することを示している。さらに、山本・篠崎 [87][88] は対称性に関する十分条件を導出し、強単峰分布以外にも条件をみたす分布が存在することを

述べている。

しかし、対称な1変量確率分布において、3点以上の場合における Principal Points の具体的な値については、Flury[21] が1変量標準正規分布における k-Principal Points ($3 \leq k \leq 5$) の値を数値計算により示しているにとどまっており、さらなる考察が必要である。

また、$k \geq 3$ の場合における k-Principal Points の対称性について、Li & Flury[34] は、Chow[7] により示されている、区間 $[0, 1]$ において連続な関数の区分的多項式による最小2乗近似が一意に定まる十分条件を確率分布の分位関数に適用して、対称な1変量分布における k-Principal Points の対称性に関する定理を発表した。しかし、この定理は誤りであり、定理の訂正の他、新たな条件の導出についても研究が必要である。

一方、2変量が互いに独立な正規分布における k-Principal Points の配置について、Flury[21] は一方の分散が他方よりも大きい場合に k-Principal Points($k \leq 5$) が一直線上に並ぶ配置となる場合があることを計算機シミュレーションにより示している。しかし、分布の分散共分散行列と Principal Points の配置との間の関係や、Principal Points の数が多い場合の配置については未解明の部分が多く残されている。

1.2 本研究の目的

前節で述べた背景に基づき、本論文では、対称な1変量確率分布における k-Principal Points の対称性、および2変量が互いに独立な正規分布における k-Principal Points の配置に関する特徴的性質を、理論的およびシミュレーションを含む計算幾何学的立場から論じる。

まず、対称な1変量確率分布が与えられた場合において、期待値に関して対称な3点が目的関数の極小値を与えるときの必要条件を導出する。そして、さまざまな分布について、期待値に関して対称な3点が必要条件をみたすかどうか考察し、3-Principal Points の値を計算機シミュレーションを用いて求め、3-Principal Points の値および対称性について議論する。

次に、密度関数に Chow[7] の定理および Trushkin[63] の定理を適用する。これにより、k-Principal Points の対称性に関する定理および十分条件を導出し、対称性が成立する確率分布族の拡張を行い、さまざまな分布に関して対称性の検証を行う。

また、2 変量が互いに独立な正規分布において、さまざまな分散共分散行列における k-Principal Points($k \leq 5$) の配置を計算機シミュレーションを用いて求めることにより、Flury[21] により提起された問題である、k個の点が一直線上に並ぶ配置となる分散共分散行列の境界値を求め、分散共分散行列と配置との関係について検討する。さらに、2 変量標準正規分布における k-Principal Points の配置をシミュレーションにより求める。その上で、得られた配置のうち最適配置の候補となり得る

- k個の点の配置が原点を中心とする正 k角形となる場合
- k個の点の配置が原点を中心とする正 $(k-1)$ 角形＋原点となる場合

については 2 重積分を用いて目的関数を最小化する k個の点の値を導出し、目的関数値および k個の点の値の理論的根拠を与える。

1.3 本論文の構成

本論文は以下の 6 章から構成される。

第 1 章では、本論文の背景および Principal Points のもつ意味について触れ、本研究の目的について述べる。

第 2 章では、k-Principal Points に関する従来の研究について詳述する。まず、Principal Points の定義を示した上で、対称な 1 変量分布、多変量楕円分布、2 変量正規分布がそれぞれ与えられた場合における k-Principal Points について、理論的な考察や計算機シミュレーションにより得られている値を示す。また、本研究で使用している k-means 法を援用した k-Principal Points の導出アルゴリズムを示し、さらに、対称な 1 変量分布における k-Principal Points の対称性に関する十分条件について述べる。その他、Principal Points の概念の応用例として、天気図の解析 (村木・大瀧・水田 [85][86]) における代表的な天気図パターンの抽出について紹介する。

第 3 章では、対称な 1 変量分布における 3-Principal Points について論じる。期待値に関して対称な 3 点が目的関数の極小値を与えるときの必要条件を理論的に求め、種々の分布において対称な 3 点が必要条件を満たすかどうかを考察し、条件を満たさない場合におい

ては計算機シミュレーションを用いて、各分布における 3-Principal Points を求める。さらに、対称性の有無についても考察する。

第 4 章では、対称な 1 変量分布における k-Principal Points の対称性に関して、Chow[7] の定理に基づく新しい定理を導出し、Li & Flury[34] における k-Principal Points の対称性に関する定理の誤りを指摘する。また、Trushkin[63] の定理を密度関数に適用することにより新しい十分条件を導出し、k-Principal Points の対称性が成立する確率分布族を拡張する。

第 5 章では、2 変量正規分布が与えられたときの分散共分散行列が対角行列 $\mathrm{diag}(\sigma^2, 1)$ となる場合の k-Principal Points の配置について論じる。計算機シミュレーションを用いることにより、k-Principal Points が一直線上に並ぶ配置となるσを求め、2 変量標準正規分布 ($\sigma = 1$) における k-Principal Points($k \leq 12$) の配置に関してもシミュレーションを行い、得られた配置を検証するとともに、k-Principal Points の候補となり得るいくつかの配置について理論的考察を行う。

最後に第 6 章では、本論文における結論を述べる。また、今後の課題として、対称な 1 変量確率分布のうち t 分布、Johnson's S_u分布、混合正規分布など、k個の主要点の対称性に関する十分条件を満たさない分布の存在について触れ、これらの分布における主要点の対称性の有無に関する条件の導出の必要性について言及する。さらに、多変量分布における k-Principal Points の配置や最適配置問題への拡張、クラスター分析および主成分分析などの各種多変量解析への応用の可能性についても述べる。

第 2 章

主要点に関する従来の研究

本章では、Flury[21] により提唱された主要点 (Principal Points) の定義を第 2.1節で述べ、さらに対称な確率分布が与えられた場合における Principal Points に関する従来の研究内容を紹介する。

第 2.2節では、対称な 1 変量確率分布が与えられた場合において、2-Principal Points を理論的に考察し、種々の分布において求められた 2 点の値の対称性について検証した結果を紹介する。また、k-Principal Points($k \geq 3$) の対称性に関する研究として、標準正規分布における値を数値計算により示し、さらに密度関数の性質に着目することにより得られた各種定理について紹介する。

第 2.3節では、多変量楕円分布における k-Principal Points の値に関して行われた理論的考察について紹介する。

第 2.4節では、2 変量正規分布が与えられた場合に関し、分散共分散行列を変化させたときの k-Principal Points の配置の変化について計算機シミュレーションにより得られた結果を紹介する。また、本研究で使用した、k-means 法を援用した k-Principal Points の導出アルゴリズムについても紹介する。

最後に第 2.5節では、k-Principal Points の概念を導入したデータ解析の例として、天気図の解析 (村木・大瀧・水田 [85][86]) に関する研究を紹介する。

なお、対称な 1 変量確率分布において、2-Principal Points が非対称であるときには、2 組の 2-Principal Points が存在する。従って、Principal Points が対称である状態を Principal

Pointsが一意である (unique) と呼ぶ論文もあるが (Li & Flury[34], Tarpey[61], 山本・篠崎[87][88])、本論文では「Principal Pointsが対称である」という表現を使用する。

2.1 主要点の定義

この節では、Flury[21] による主要点 (Principal Points) の定義を述べる。

以下では、k個の p 次元空間の点の集まり $\boldsymbol{y}_j \in R^p$ $(1 \leq j \leq k)$ と $\boldsymbol{x} \in R^p$ との距離をユークリッド距離

$$d(\boldsymbol{x}|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k) = \min_{1 \leq h \leq k} \{(\boldsymbol{x} - \boldsymbol{y}_h)'(\boldsymbol{x} - \boldsymbol{y}_h)\}^{1/2} \tag{2.1.1}$$

で定義する。その上で、Principal Pointsは以下のように定義される。

定義 (Principal Points) (Flury[21])

$\boldsymbol{\xi}_j \in R^p$ $(1 \leq j \leq k)$ が確率分布 Fに従う確率変数 Xの k-Principal Pointsであるとは、

$$E_F\{d^2(X|\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_k)\} = \min_{\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\}} E\{d^2(X|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k)\} \tag{2.1.2}$$

が成立することである。

以下では、Fに関する期待値 $E_F(\cdot)$ を簡単のために $E(\cdot)$ と記す。また、

$$M(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k) = E\{d^2(X|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k)\} \tag{2.1.3}$$

とし、これを目的関数と呼ぶ。さらに、

$$P_F(k) = \min_{\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\}} M(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k) \tag{2.1.4}$$

とおき、場合により $P_F(k)$ を $P_X(k)$ と表記する。

2.2 対称な 1 変量確率分布の場合

対称な 1 変量確率分布が与えられている場合において、Flury[21] は 2-Principal Pointsに関する理論的考察を行い、期待値に関して対称な 2 点が目的関数を極小とする条件および期待値に関して非対称な 2-Principal Pointsの例を示している。水田 [83] は、さらにこの

結果を一般化し、種々の分布における2-Principal Pointsの対称性に関して考察を行っている。Flury[21]はさらに、標準正規分布におけるk-Principal Pointsの値を数値計算により示している。

一方、密度関数の性質に着目することによって、k-Principal Pointsの対称性に関する様々な十分条件が導出されている(Li & Flury[34], Tarpey[61], 山本・篠崎 [87][88])。

本節では、以上の研究について紹介する。

2.2.1 2個の主要点に関する理論的考察

連続な確率変数Xの分布関数を$F(x)$、密度関数を$f(x)$とし、$f(x)=f(-x)$かつ$E(X^2)=\sigma^2<\infty$が成り立つものとする。ここでクラスターの中心点を2個とするとき、関数

$$E\{d^2(X|y_1,y_2)\}=\int_{-\infty}^{\infty}\min\{(x-y_1)^2,(x-y_2)^2\}f(x)dx \tag{2.2.5}$$

を$y_1,y_2\in R$、$y_1<y_2$のもとで最小化したい。このとき、$y_1=c-h$、$y_2=c+h$とおけば、$c\in R$、$h\geq 0$として$E\{d^2(X|y_1,y_2)\}$は

$$\begin{aligned} H(c,h) &= \int_{-\infty}^{c}(x-c+h)^2f(x)dx+\int_{c}^{\infty}(x-c-h)^2f(x)dx \\ &= \sigma^2+c^2+h^2+2hG(c) \end{aligned} \tag{2.2.6}$$

と表せる。ただし、

$$G(c)=\int_{-\infty}^{c}(x-c)f(x)dx-\int_{c}^{\infty}(x-c)f(x)dx \tag{2.2.7}$$

とする。また、fが偶関数であることより、

$$G(c)=2\int_{-\infty}^{c}xf(x)dx+c(1-2F(c)) \tag{2.2.8}$$

となる。ここで、Hをc、hで偏微分すると、

$$G'(c)=1-2F(c)-2cf(c)+2cf(c)=1-2F(c) \tag{2.2.9}$$

であるから

$$\frac{\partial H}{\partial c} = 2c+2hG'(c) \tag{2.2.10}$$

$$= 2c + 2h(1 - 2F(c)) \tag{2.2.11}$$

$$\frac{\partial H}{\partial h} = 2h + 2G(c) \tag{2.2.12}$$

$$= 2h + 2c(1 - 2F(c)) + 4\int_{-\infty}^{c} xf(x)dx \tag{2.2.13}$$

となる。(2.2.10)、(2.2.12) が 0 に等しいとおくと、

$$c = G(c)G'(c) \tag{2.2.14}$$

が得られ、さらに (2.2.12) で $c = 0$ とおくと

$$h = -G(0) = 2\int_{0}^{\infty} xf(x)dx = E(|X|) \tag{2.2.15}$$

となる。また、(2.2.9) に $c = 0$ を代入すると $G'(0) = 0$ であり、$c = 0$ において (2.2.14) は成立する。さらに、

$$\frac{\partial^2 H}{\partial c^2} = 2(1 - 2hf(c)) \tag{2.2.16}$$

$$\frac{\partial^2 H}{\partial c \partial h} = \frac{\partial^2 H}{\partial h \partial c} = 2(1 - 2F(c)) \tag{2.2.17}$$

$$\frac{\partial^2 H}{\partial h^2} = 2 \tag{2.2.18}$$

より、2 階偏微分行列は

$$D(c,h) = \begin{bmatrix} \dfrac{\partial^2 H}{\partial c^2} & \dfrac{\partial^2 H}{\partial c \partial h} \\ \dfrac{\partial^2 H}{\partial h \partial c} & \dfrac{\partial^2 H}{\partial h^2} \end{bmatrix} = 2\begin{bmatrix} 1 - 2hf(c) & 1 - 2F(c) \\ 1 - 2F(c) & 1 \end{bmatrix} \tag{2.2.19}$$

であり、$D(0, E(|X|))$ は、$1 - 2f(0)E(|X|) > 0$ のとき正定値となる。

以上より、次の定理 1 が成り立つ。

定理 1. (Flury[21])

連続な 1 変量確率変数 Xの密度関数 $f(x)$ が平均$\mu = E(X)$ に関して対称で、かつ 2 次のモーメントが有限であるとき、

$$y_1 = \mu - E(|X - \mu|), \quad y_2 = \mu + E(|X - \mu|) \tag{2.2.20}$$

が $E\{d^2(X|y_1, y_2)\}$ の極小値をもたらすための十分条件は $f(\mu)E(|X - \mu|) < \frac{1}{2}$であり、その必要条件は $f(\mu)E(|X - \mu|) \leq \frac{1}{2}$である。

すなわち、$f(\mu)E(|X-\mu|) > \frac{1}{2}$の場合には、2-Principal Points は期待値に関して非対称となる。

また、水田 [83] は、目的関数 $M(y_1, y_2)$ が極小となる十分条件について考察し、定理 1 の結果を含む定理を導出している。一般性を失うことなく$\mu = 0$ とし、$c = \frac{y_1 + y_2}{2}$および $h = \frac{y_2 - y_1}{2}$とおく。さらに $M(y_1, y_2) = H(c, h)$、

$$G(c) = 2\int_{-\infty}^{c} xf(x)dx + c(1 - 2F(c)) \tag{2.2.21}$$

とおいたときに$\frac{\partial H}{\partial c} = 2c + 2hG'(c) = 0$ および$\frac{\partial H}{\partial h} = 2h + 2G(c) = 0$ を c、h について解くと $h = -G(c)$ および $2c - 2G(c)G'(c) = 0$ となる。ここで

$$p(c) = G(c)G'(c) - c \tag{2.2.22}$$

とおくと、$p(c) = 0$ かつ $p'(c) < 0$ をみたす c のうち $H(c, -G(c))$ を最小にする c に対応する 2 点の座標 $c \mp G(c)$ および $-c \pm G(-c)$ が 2-Principal Points となる。

また、$p(0) = 0$ かつ $p'(0) < 0$ であるとき、すべての $c > 0$ において $p(c) \neq 0$ ならば $p(c)$ の連続性より $p(c) = G(c)G'(c) - c < 0$ が成り立つ。ここで (2.2.21) 式より

$$G'(c) = 1 - 2F(c) \tag{2.2.23}$$

であるから、$p(c) = G(c)G'(c) - c < 0$ は (2.2.21) 式および (2.2.23) 式より

$$\begin{aligned}
& G(c)G'(c) - c \\
= \;& 2(1 - 2F(c))\int_{-\infty}^{c} xf(x)dx + c(1 - 2F(c))^2 - c \\
= \;& -4cF(c)(1 - F(c)) + 2(2F(c) - 1)\int_{c}^{\infty} xf(x)dx \\
= \;& -4cF(c)(1 - F(c)) + 2(2F(c) - 1)\left\{\int_{c}^{\infty}(x - c)f(x)dx + \int_{c}^{\infty} cf(x)dx\right\} \\
= \;& -2c(1 - F(c)) + 2(2F(c) - 1)\int_{c}^{\infty}(x - c)f(x)dx \qquad (2.2.24) \\
< \;& 0
\end{aligned}$$

と表せる。これより

$$\frac{c}{2F(c) - 1} > \frac{\int_{c}^{\infty}(x - c)f(x)dx}{1 - F(c)} \tag{2.2.25}$$

がすべての $c>0$ で成り立つならば、2-Principal Points は一意に定まり、原点に関して対称となる (山本・篠崎 [87][88])。

2.2.2 種々の分布における 2 個の主要点の値

以下では、期待値に関して対称な種々の分布に関する 2-Principal Points の値を示す。ただし、$f(x)$ は $x=\mu$ で連続であると仮定する。

2.2.2.1 標準正規分布

確率変数 X が標準正規分布に従うとき、密度関数は $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$ と書ける。$f(0)=\frac{1}{\sqrt{2\pi}}$, $E(|X|)=\frac{2}{\sqrt{2\pi}}$ より $f(0)E(|X|)=\frac{1}{\pi}\simeq 0.318<\frac{1}{2}$ であるから、2 点 $\pm\sqrt{\frac{2}{\pi}}$ は 2-Principal Points の候補となる。

ここで、$\int xf(x)dx=-f(x)$ および $f'(c)=-cf(c)$ より $G(c)=-2f(c)+c(1-2F(c))$、$G'(c)=1-2F(c)$ となるので、

$$\begin{aligned} p(c) &= 2(1-2F(c))\int_{-\infty}^{c} xf(x)dx+c(1-2F(c))^2-c \\ &= 2(2F(c)-1)f(c)+4cF(c)(F(c)-1) \end{aligned} \tag{2.2.26}$$

と書ける。これを c で微分すると、

$$\begin{aligned} p'(c) &= 2(2F(c)-1)f'(c)+4(f(c))^2+4F(c)(F(c)-1)+4cf(c)(2F(c)-1) \\ &= 4(f(c))^2+2cf(c)(2F(c)-1)+4F(c)(F(c)-1) \end{aligned} \tag{2.2.27}$$

$$\begin{aligned} p''(c) &= 8f(c)f'(c)+2f(c)(2F(c)-1)+2cf'(c)(2F(c)-1)+4c(f(c))^2+4f(c)(2F(c)-1) \\ &= 2f(c)\{(2F(c)-1)(3-c^2)-2cf(c)\} \end{aligned} \tag{2.2.28}$$

となる。ここで、$r(c)=(2F(c)-1)(3-c^2)-2cf(c)$ とおくと、

$$\begin{aligned} r'(c) &= 2f(c)(3-c^2)-2c(2F(c)-1)-2f(c)-2cf'(c) \\ &= 4f(c)-2c(2F(c)-1) \end{aligned} \tag{2.2.29}$$

$$\begin{aligned} r''(c) &= 4f'(c)-4cf(c)-2(2F(c)-1) \\ &= -8cf(c)-2(2F(c)-1) \end{aligned} \tag{2.2.30}$$

であるから、$c > 0$ において $r''(c) < 0$ となる。従って、$r'(c)$ は $c > 0$ で単調減少し、さらに $r'(0) = 4f(0) > 0$, $\lim_{c\to\infty} r'(c) = -\infty$ となるので、$r'(c_0) = 0$ をみたす c_0がただ 1 つ存在し、

$$\begin{cases} r'(c) \geq 0 & (0 \leq c \leq c_0) \\ r'(c) < 0 & (c_0 < c) \end{cases}$$

である。

ここで $r(0) = 0$ であるから、$r(c_0) > 0$ となる。また、$\lim_{c\to\infty} r(c) = -\infty$ となるので、$r(c_1) = 0$ すなわち $p''(c_1) = 0$ をみたす $c_1 > c_0$がただ 1 つ存在し、

$$\begin{cases} p''(c) \geq 0 & (0 \leq c \leq c_1) \\ p''(c) < 0 & (c_1 < c) \end{cases}$$

である。よって、$p'(c)$ は $0 \leq c \leq c_1$において単調増加、$c_1 < c$ において単調減少する。このとき $p'(0) = \dfrac{1}{\pi} - 1 < 0$, $\lim_{c\to\infty} p'(c) = 0$ となるので、$p'(c)$ の連続性より $p'(c_1) > 0$ である。これより $p'(c_2) = 0$ をみたす $0 < c_2 < c_1$がただ 1 つ存在し、

$$\begin{cases} p'(c) \leq 0 & (0 \leq c \leq c_2) \\ p'(c) > 0 & (c_2 < c) \end{cases}$$

となるので、$p(c)$ は $0 \leq c \leq c_2$で単調減少、$c_2 < c$ で単調増加する。さらに $p(0) = 0$, $\lim_{c\to\infty} p(c) = 0$ となるので、$c > 0$ において $p(c) < 0$ となる。

以上の結果より、$\pm\sqrt{\dfrac{2}{\pi}}$は 2-Principal Points である (図 2.1)。すなわち、$X \sim N(\mu, \sigma^2)$ かつ $k = 2$ ならば、2-Principal Points は$\mu \pm \sigma\sqrt{\dfrac{2}{\pi}}$ である (Cox[9]、Flury[21])。

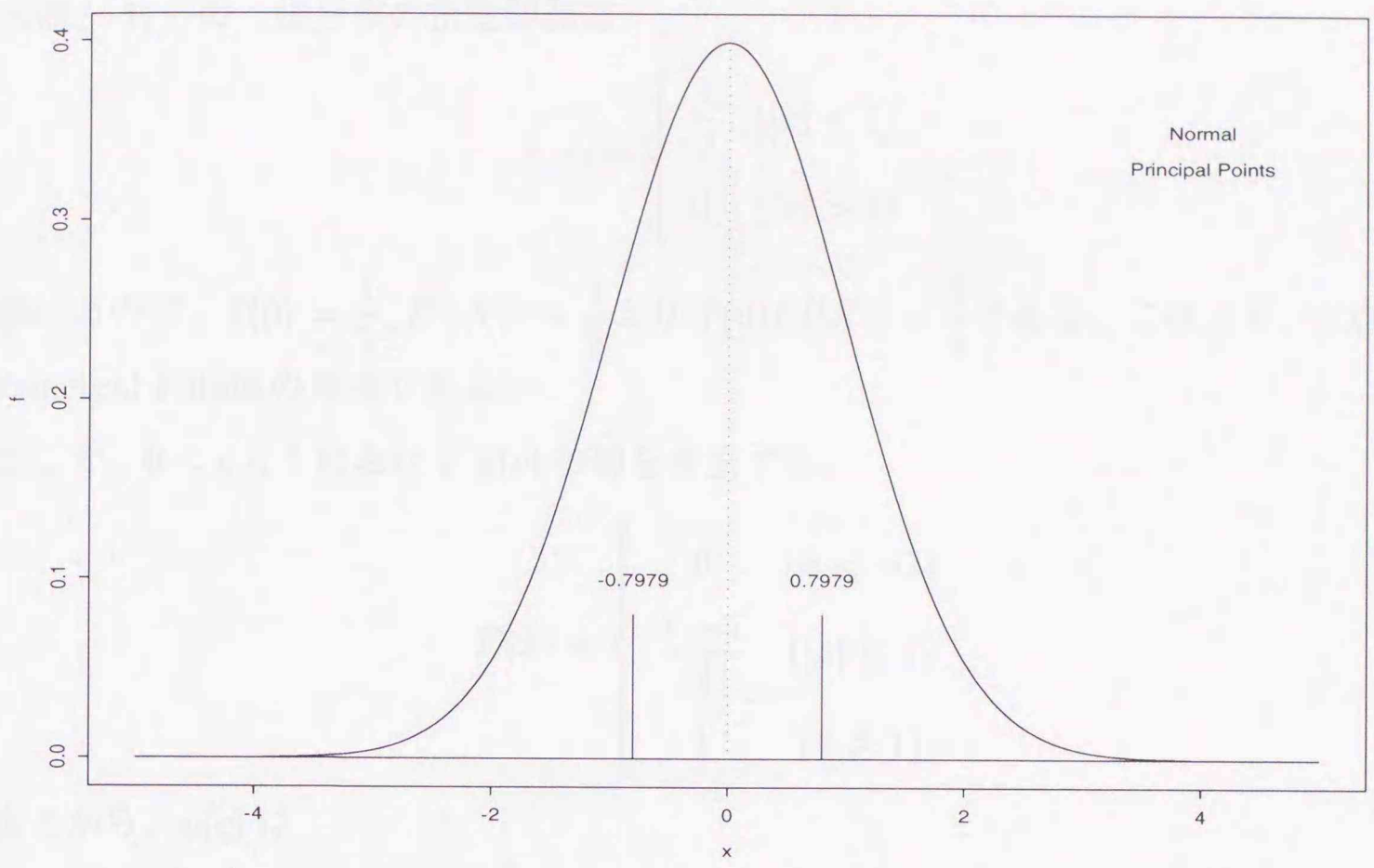


図 2.1: 標準正規分布における 2 個の主要点 (Flury[21])

図 2.2: 一様分布における 2 個の主要点 ($\mu = 0, \theta = 1$) (Flury[21])

2.2.2.2　一様分布

区間 $[-1,1]$ の一様分布の密度関数は

$$f(x) = \begin{cases} \dfrac{1}{2} & (|x| \leq 1) \\ 0 & (|x| > 1) \end{cases}$$

と書けるので、$f(0) = \dfrac{1}{2}$, $E(|X|) = \dfrac{1}{2}$より $f(0)E(|X|) = \dfrac{1}{4}$である。これより、2 点$\pm\dfrac{1}{2}$は 2-Principal Points の候補である。

ここで、$0 \leq c < 1$ における $p(c)$ の値を考察する。

$$F(x) = \begin{cases} 0 & (x < -1) \\ \dfrac{x+1}{2} & (|x| \leq 1) \\ 1 & (x > 1) \end{cases}$$

であるから、$p(c)$ は

$$\begin{aligned} p(c) &= 2(1-2F(c))\int_{-\infty}^{c} xf(x)dx + c(1-2F(c))^2 - c \\ &= -2c\int_{-1}^{c} \frac{1}{2}xdx + c^3 - c \\ &= c^3 - 2c^2 - c + 2 \end{aligned} \tag{2.2.31}$$

となる。ここで $p'(c) = 3c^2 - 4c - 1$ より、$0 \leq c < \dfrac{2+\sqrt{7}}{3}$のとき $p'(c) < 0$、$c < \dfrac{2+\sqrt{7}}{3}$のとき $p'(c) > 0$ となるので、$p(c)$ は $c = \dfrac{2+\sqrt{7}}{3}$で極小値をとる。さらに $p(0) = 0$, $p(1) = 0$ であるから、$p(c)$ の連続性より $0 < c < 1$ においては $p(c) < 0$ である。これより、区間 $[-1,1]$ の一様分布における 2-Principal Points は$\pm\dfrac{1}{2}$である。同様に、区間 $[\mu-\theta, \mu+\theta]$ $(\theta > 0)$ の一様分布における 2-Principal Points は$\mu \pm \dfrac{1}{2}\theta$である (Flury[21]、Krzanowski & Lai[33]、Marriott[37]、図 2.2)。

2.2.2.3　ロジスティック分布

ロジスティック分布の密度関数は

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

で与えられ、分布関数は、

$$F(x) = \frac{1}{1+e^{-x}}$$

となり、平均 0、分散$\frac{\pi^2}{3}$である。

$f(0) = \frac{1}{4}, E(|X|) = \log 2$より、定理 1 の条件式の値を求めると、$f(0)E(|X|) \simeq 0.346 < \frac{1}{2}$となり、$\pm E(|X|)$は 2-Principal Points の候補となる。

また、

$$\begin{aligned} G(c) &= 2\int_{-\infty}^{c} xf(x)dx + c(1-2F(c)) \\ &= 2\left\{\frac{ce^c}{1+e^c} - \log(1+e^c)\right\} + c\left(1 - \frac{2}{1+e^{-c}}\right) \\ &= c - 2\log(1+e^c) \end{aligned} \tag{2.2.32}$$

および

$$G'(c) = 1 - 2F(c) = \frac{1-e^c}{e^c+1} \tag{2.2.33}$$

より、

$$p(c) = \frac{-2ce^c - 2(1-e^c)\log(1+e^c)}{e^c+1} \tag{2.2.34}$$

となる。$t = e^c - 1$とおくと、

$$\begin{aligned} (e^c+1)p(c) &= -2ce^c - 2(1-e^c)\log(1+e^c) \\ &= -2(t+1)\log(t+1) + 2t\log(t+2) \end{aligned}$$

となる。$c > 0$より$e^c > 1$なので、$t > 0$となる。ここで右辺を$r(t)$とおくと

$$\begin{aligned} r(0) &= -2\log 1 + 0 = 0 \\ r'(t) &= -2\log(t+1) - 2 + 2\log(t+2) + \frac{2t}{t+2} \\ r''(t) &= \frac{2t}{(1+t)(2+t)^2} \end{aligned}$$

において$r''(t) > 0, \lim_{t\to\infty} r'(t) = 0$が成り立つので、$r'(t)$は$t > 0$において狭義単調増加で、$t \to \infty$のとき 0 である。従って、$t > 0$において$r'(t) < 0$が成り立つ。すなわち、$r(t)$は$t > 0$において狭義単調減少である。また、$r(0) = 0$より$r(t) < 0$となる。

以上により、$c > 0$において、$p(c) \neq 0$となり、ロジスティック分布における 2-Principal Points は$\pm E(|X|)$となる。これ以外には、$H(c,h)$は極小値もとらない (水田 [83])。

2.2.2.4　両側指数分布

両側指数分布 (ラプラス分布、二重指数分布) の密度関数として、

$$f(x) = \frac{1}{2}e^{-|x|}$$

を考える。$f(0) = \frac{1}{2}$, $E(|X|) = 1$ より、$f(0)E(|X|) = \frac{1}{2}$となる。この値では、定理 1 をそのまま適用すると $\pm E(|X|)$ は 2-Principal Points ではないことになる。しかし、$H(c,h)$ のヘシアンが非負定値であることを示しているので 2-Principal Points である可能性は十分残っている。

$G(c) = e^c + c$, $G'(c) = 1 - e^c$より、$p(c) = (1 - c - e^c)e^c$となる。ここで、$r(c) = 1 - c - e^c$ とおくと、$r'(c) = -1 - e^c < 0$, $r(0) = 0$ より、$c > 0$ のときに $r(c) < 0$ となる。すなわち $c \neq 0$ では 2-Principal Points とはならない (水田 [83])。

2.2.2.5　t 分布

自由度 n の t 分布の密度関数は

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

と表され、平均 0、分散$\frac{n}{n-2}$ $(n > 2)$ である。

ここで

$$f(0) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}, \qquad E(|X|) = \frac{2\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\frac{n}{n-1} \tag{2.2.35}$$

より、

$$f(0)E(|X|) = \frac{2}{\pi(n-1)}\left\{\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\right\}^2 \tag{2.2.36}$$

となる。$n = 3$ のとき $f(0)E(|X|) = \frac{4}{\pi^2} < \frac{1}{2}$、$n = 4$ のとき $f(0)E(|X|) = \frac{3}{8} < \frac{1}{2}$が成り立ち、これらの結果と数学的帰納法により、$n > 2$ のとき $f(0)E(|X|) < \frac{1}{2}$となる。従って、$\pm E(|X|)$ は 2-Principal Points の候補となる。

ここで、$\int xf(x)dx = -\dfrac{n+x^2}{n-1}f(x)$ および $f'(c) = -\dfrac{(n+1)c}{n+c^2}f(c)$ より

$$\begin{aligned} p(c) &= 2(1-2F(c))\int_{-\infty}^{c} xf(x)dx + c(1-2F(c))^2 - c \\ &= \frac{2(n+c^2)(2F(c)-1)}{n-1}f(c) + 4cF(c)(F(c)-1) \end{aligned} \tag{2.2.37}$$

と書ける。これを c で微分すると、

$$\begin{aligned} p'(c) &= \frac{4cf(c)(2F(c)-1) + 4(n+c^2)(f(c))^2 + 2(n+c^2)f'(c)(2F(c)-1)}{n-1} \\ &\quad +4F(c)(F(c)-1) + 4cf(c)(2F(c)-1) \\ &= \frac{4(n+c^2)}{n-1}(f(c))^2 + 2cf(c)(2F(c)-1) + 4F(c)(F(c)-1) \end{aligned} \tag{2.2.38}$$

$$\begin{aligned} p''(c) &= \frac{8c(f(c))^2 + 4(n+c^2)f(c)f'(c)}{n-1} \\ &\quad +2f(c)(2F(c)-1) + 4c(f(c))^2 + 2cf'(c)(2F(c)-1) + 4f(c)(2F(c)-1) \\ &= \frac{2\{3n-(n-2)c^2\}(2F(c)-1)f(c)}{n+c^2} \end{aligned} \tag{2.2.39}$$

となる。これより

$$\begin{cases} p''(c) > 0 & (0 < c < \sqrt{\frac{3n}{n-2}}) \\ p''(c) < 0 & (\sqrt{\frac{3n}{n-2}} < c) \end{cases}$$

であるから、$p'(c)$ は $c = \sqrt{\dfrac{3n}{n-2}}$ で極大値をとる。ここで $p'(0) = \dfrac{4}{\pi(n-1)}\left\{\dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\right\}^2 + 1 < 0$, $\lim_{c\to\infty} p(c) = 0$ なので、$p'(c)$ の連続性より $p'(\sqrt{\frac{3n}{n-2}}) > 0$ である。これより $p'(c_0) = 0$ をみたす $0 < c_0 < \sqrt{\dfrac{3n}{n-2}}$ なる c_0 がただ 1 つ存在し、

$$\begin{cases} p'(c) \le 0 & (0 \le c \le c_0) \\ p'(c) > 0 & (c_0 < c) \end{cases}$$

となるので、$p(c)$ は $0 \le c \le c_0$ で単調減少、$c_0 < c$ で単調増加する。さらに $p(0) = 0$, $\lim_{c\to\infty} p(c) = 0$ である。従って、$c > 0$ において $p(c) < 0$ である。

以上の結果より、$\pm E(|X|)$ は 2-Principal Points となる (山本・篠崎 [87][88])。

2.2.2.6 Johnson's S_u分布

平均が 0 の Johnson's S_u分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi}\sqrt{x^2+1}} \exp\left[-\frac{1}{2}\left\{\log(x+\sqrt{x^2+1})\right\}^2\right]$$

と表される。分散は$\dfrac{e^2-1}{2}$である。

ここで

$$f(0) = \frac{1}{\sqrt{2\pi}}, \quad E(|X|) = 2\sqrt{e}\int_0^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \ \ (\simeq 1.12556) \tag{2.2.40}$$

より、

$$f(0)E(|X|) = \sqrt{\frac{2e}{\pi}}\int_0^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \ \ (\simeq 0.449035) \tag{2.2.41}$$

となるので、$f(0)E(|X|) < \dfrac{1}{2}$である。従って、$\pm E(|X|)$ は 2-Principal Points の候補となる。

ここで、Φを標準正規分布関数とすると、

$$\begin{aligned}
G(c) &= 2\int_{-\infty}^{c} x f(x) dx + c(1-2F(c)) \\
&= 2\int_{-\infty}^{\log(c+\sqrt{c^2+1})} \frac{1}{\sqrt{2\pi}} \frac{e^u - e^{-u}}{2} e^{-\frac{1}{2}u^2} du - 2c\int_0^c f(x) dx \\
&= \sqrt{e}\left\{\int_{-\infty}^{\log(c+\sqrt{c^2+1})} \frac{1}{\sqrt{2\pi}} e^{-\frac{(u-1)^2}{2}} du - \int_{-\infty}^{\log(c+\sqrt{c^2+1})} \frac{1}{\sqrt{2\pi}} e^{-\frac{(u+1)^2}{2}} du\right\} \\
&\quad -2c\int_0^{\log(c+\sqrt{c^2+1})} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \\
&= \sqrt{e}\{\Phi(\log(c+\sqrt{c^2+1})-1) - \Phi(\log(c+\sqrt{c^2+1})+1)\} \\
&\quad +c\{1-2\Phi(\log(c+\sqrt{c^2+1}))\} \qquad (2.2.42)
\end{aligned}$$

$$G'(c) = 1-2F(c) = 1-2\Phi(\log(c+\sqrt{c^2+1})) \tag{2.2.43}$$

と書ける。これより $p(c)$ は

$$\begin{aligned}
p(c) &= G(c)G'(c) - c \\
&= \sqrt{e}\{1-2\Phi(\log(c+\sqrt{c^2+1}))\}\{\Phi(\log(c+\sqrt{c^2+1})-1) - \Phi(\log(c+\sqrt{c^2+1})+1)\} \\
&\quad +4c\Phi(\log(c+\sqrt{c^2+1}))\{\Phi(\log(c+\sqrt{c^2+1}))-1\} \qquad (2.2.44)
\end{aligned}$$

と表せる。ここで、$\log(c+\sqrt{c^2+1})=t$ とおくと、(2.2.44) 式は

$$\sqrt{e}\{1-2\Phi(t)\}\{\Phi(t-1)-\Phi(t+1)\}+2(e^t-e^{-t})\Phi(t)\{\Phi(t)-1\} \tag{2.2.45}$$

と表せる。(2.2.45) 式を $g(t)$ とおくと、t は c に関して単調増加するので、$t>0$ のとき $g(t)\neq 0$ であることは $c>0$ のとき $p(c)\neq 0$ であることと同値である。

ここで、Zを標準正規分布関数Φに従う確率変数として

$$r_1(t)=\{2\Phi(t)-1\}E(|Z|)-\frac{e^t-e^{-t}}{2} \tag{2.2.46}$$

とおくと、ϕをΦの密度関数として $r_1'(t)=2\phi(t)E(|Z|)-\dfrac{e^t+e^{-t}}{2}$ と書ける。$\phi(t)$ は $t\geq 0$ で狭義単調減少するので、任意の $t>0$ において $\phi(t)<\phi(0)$ が成立する。さらに $t>0$ において $\dfrac{e^t+e^{-t}}{2}>1$ であるから、

$$\begin{aligned} r_1'(t) &= 2\phi(t)E(|Z|)-\frac{e^t+e^{-t}}{2} \\ &< 2\phi(0)E(|Z|)-\frac{e^t+e^{-t}}{2} \\ &< 2\phi(0)E(|Z|)-1 \\ &< 0 \end{aligned}$$

となる。よって、$r_1(t)$ は $t\geq 0$ で単調減少する。さらに $r_1(0)=0$ であるので、任意の $t>0$ において常に $r_1(t)<0$ が成立する。これを変形すると、

$$E(|Z|)<\frac{e^t-e^{-t}}{2\{2\Phi(t)-1\}} \quad (^\forall t>0) \tag{2.2.47}$$

と書ける。

また、

$$r_2(t)=\{1-\Phi(t)\}E(|Z|)-\frac{\sqrt{e}\{\Phi(t+1)-\Phi(t-1)\}}{2}+\frac{e^t-e^{-t}}{2}\{1-\Phi(t)\} \tag{2.2.48}$$

とおくと、$\sqrt{e}\phi(t-1)=e^t\phi(t)$, $\sqrt{e}\phi(t+1)=e^{-t}\phi(t)$ より

$$\begin{aligned} r_2'(t) &= -E(|Z|)\phi(t)-\frac{\sqrt{e}\{\phi(t+1)-\phi(t-1)\}}{2}+\frac{e^t+e^{-t}}{2}\{1-\Phi(t)\}-\frac{e^t-e^{-t}}{2}\phi(t) \\ &= -E(|Z|)\phi(t)+\frac{e^t+e^{-t}}{2}\{1-\Phi(t)\} \end{aligned}$$

と表せる。ここで、

$$s(t) = -\frac{2E(|Z|)}{e^t + e^{-t}}\phi(t) + \{1 - \Phi(t)\} \tag{2.2.49}$$

とおくと、$\phi'(t) = -t\phi(t)$ より

$$\begin{aligned} s'(t) &= -\frac{2E(|Z|)}{e^t + e^{-t}}\phi'(t) + \frac{2(e^t - e^{-t})E(|Z|)}{(e^t + e^{-t})^2}\phi(t) - \phi(t) \\ &= \frac{\phi(t)}{e^t + e^{-t}}\left\{2tE(|Z|) + \frac{2(e^t - e^{-t})}{e^t + e^{-t}}E(|Z|) - (e^t + e^{-t})\right\} \end{aligned} \tag{2.2.50}$$

となる。さらに $v(t) = 2tE(|Z|) + \dfrac{2(e^t - e^{-t})}{e^t + e^{-t}}E(|Z|) - (e^t + e^{-t})$ とすると、

$$\begin{aligned} v'(t) &= \left\{2 + \frac{8}{(e^t + e^{-t})^2}\right\}E(|Z|) - (e^t - e^{-t}) \end{aligned} \tag{2.2.51}$$

$$\begin{aligned} v''(t) &= -\frac{16(e^t - e^{-t})E(|Z|)}{(e^t + e^{-t})^3} - (e^t + e^{-t}) \end{aligned} \tag{2.2.52}$$

と書けるので、$t \geq 0$ において $v''(t) < 0$ となり、$v'(t)$ は単調減少する。ここで $v'(0) = 4E(|Z|) > 0$, $\lim_{t\to\infty} v'(t) = -\infty$ だから $v'(t_0) = 0$ をみたす t_0がただ 1 つ存在し、

$$\begin{cases} v'(t) \geq 0 & (0 \leq t \leq t_0) \\ v'(t) < 0 & (t_0 < t) \end{cases}$$

である。これより $v(t)$ は $0 \leq t \leq t_0$で単調増加、$t_0 < t$で単調減少し、$t = t_0$で極大値をとる。ここで $v(0) = -2 < 0$, $\lim_{t\to\infty} v(t) = -\infty$ であり、また S 言語により $t_0 \simeq 0.97937$ であるから $v(t_0) \simeq -0.2742 < 0$ となる。これより、$t \geq 0$ において $v(t) < 0$ となるので $s'(t) < 0$ であり、ゆえに $s(t)$ は単調減少する。さらに $E(|Z|) = \dfrac{2}{\sqrt{2\pi}}$ より $s(0) = -E(|Z|)\phi(0) + \dfrac{1}{2} > 0$, $\lim_{t\to\infty} s(t) = 0$ なので $t \geq 0$ において $s(t) > 0$、すなわち $r_2'(t) > 0$ となる。

よって、$r_2(t)$ は $t \geq 0$ において単調増加し、さらに

$$r_2(0) = \frac{1}{2}E(|Z|) - \sqrt{e}\int_0^1 \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du \simeq -0.16384 < 0$$

かつ $\lim_{t\to\infty} r_2(t) = 0$ であるから、$t > 0$ において $r_2(t) < 0$ が成り立つ。これを変形すると

$$E(|Z|) > \frac{\sqrt{e}\{\Phi(t+1) - \Phi(t-1)\}}{2\{1 - \Phi(t)\}} - \frac{e^t - e^{-t}}{2} \quad ({}^\forall t > 0) \tag{2.2.53}$$

となる。

(2.2.47) 式および (2.2.53) 式により、すべての $t>0$ において

$$\frac{e^t-e^{-t}}{2\{2\Phi(t)-1\}} > \frac{\sqrt{e}\{\Phi(t+1)-\Phi(t-1)\}}{2\{1-\Phi(t)\}} - \frac{e^t-e^{-t}}{2}$$

が成り立つ。これより

$$\begin{aligned}
&\frac{(e^t-e^{-t})\Phi(t)}{2\Phi(t)-1} > \frac{\sqrt{e}\{\Phi(t+1)-\Phi(t-1)\}}{2\{1-\Phi(t)\}} \\
&\Leftrightarrow \quad 2(e^t-e^{-t})\Phi(t)\{1-\Phi(t)\} - \sqrt{e}\{2\Phi(t)-1\}\{\Phi(t+1)-\Phi(t-1)\} > 0 \\
&\Leftrightarrow \quad \sqrt{e}\{1-2\Phi(t)\}\{\Phi(t-1)-\Phi(t+1)\} + 2(e^t-e^{-t})\Phi(t)\{\Phi(t)-1\} < 0
\end{aligned} \tag{2.2.54}$$

と変形でき、(2.2.54) 式の左辺は (2.2.45) 式に等しい。これより、$c>0$ のとき $p(c)<0$ が常に成り立つ。また $p(0)=0,\ p'(0)=2\sqrt{\frac{2e}{\pi}}\int_0^1 \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du-1<0$ となるから、$c=0$ に対応する 2 点$\pm E(|X|)$ は 2-Principal Points となる (山本・篠崎 [87][88])。

2.2.2.7　混合正規分布

A　2 種類の分布からなる場合　平均がβ、分散がα^2 $(\alpha>0)$ である正規確率変数の分布関数を

$$N(x;\beta,\alpha^2)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\alpha}e^{-\frac{(t-\beta)^2}{2\alpha^2}}dt \tag{2.2.55}$$

と表す。ここで、確率変数 Xの分布関数が

$$F(x)=(1-\varepsilon)N(x;0,1^2)+\varepsilon N(x;0,\alpha^2)\quad(0\le\varepsilon\le1) \tag{2.2.56}$$

であるとき、

$$f(x)=F'(x)=(1-\varepsilon)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}+\varepsilon\frac{1}{\sqrt{2\pi}\alpha}e^{-\frac{x^2}{2\alpha^2}} \tag{2.2.57}$$

であるから

$$f(0)E(|X|)=\pi^{-1}\{1+\varepsilon(1-\varepsilon)(1-\alpha)^2/\alpha\} \tag{2.2.58}$$

である。すなわち、$c=0$ において$\varepsilon(1-\varepsilon)(\alpha+\alpha^{-1}-2)<\frac{\pi}{2}-1\simeq0.57$ であれば $h=E(|X|)$ は極小値となる。

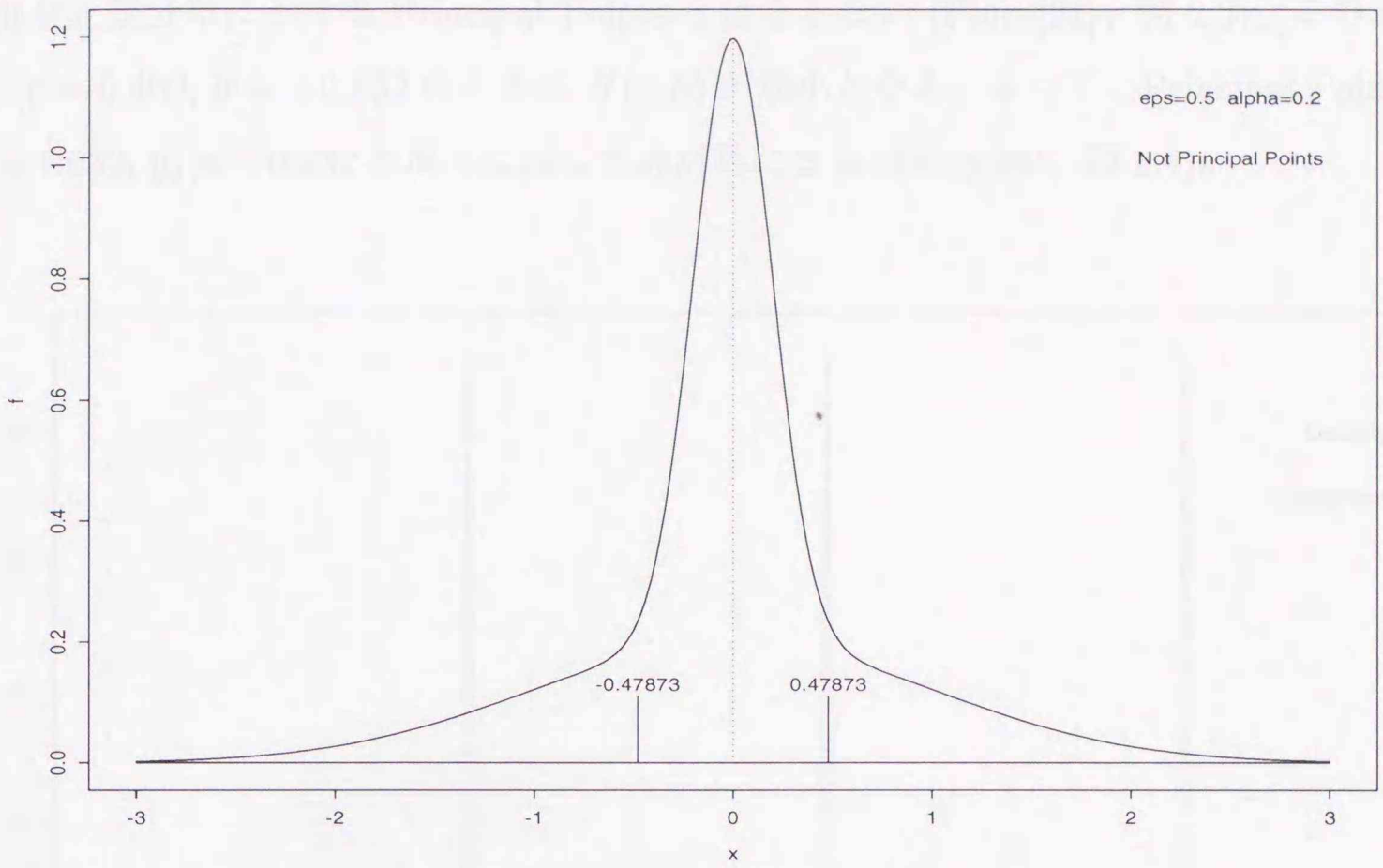


図 2.3: 混合正規分布における主要点とならない例 $(\varepsilon = 0.5, \alpha = 0.2)$ (Flury[21])

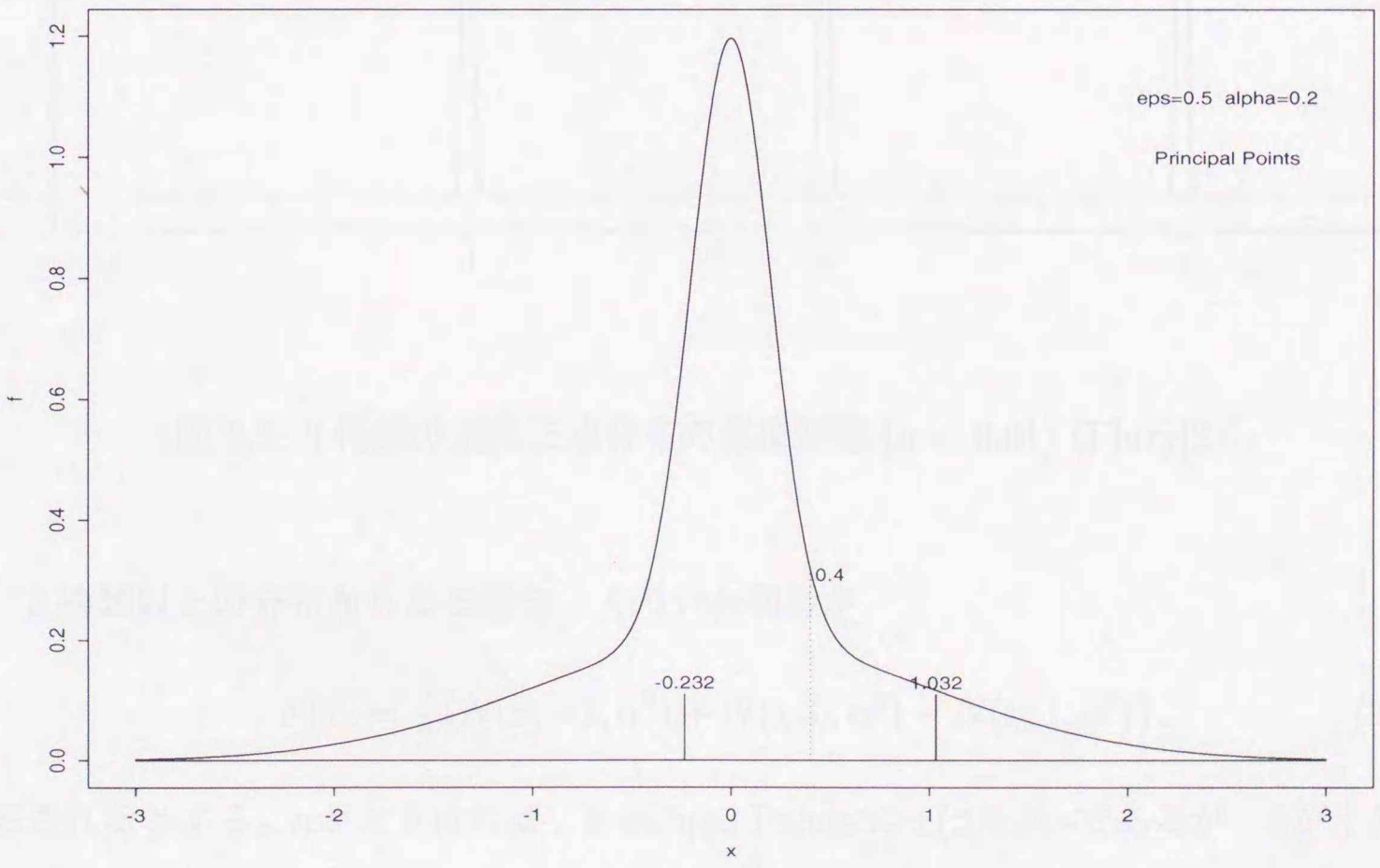


図 2.4: 混合正規分布における 2 個の主要点 $(\varepsilon = 0.5, \alpha = 0.2)$ (Flury[21])

例えば、$\varepsilon = \frac{1}{2}$, $\alpha = 5$ $(or\ \frac{1}{5})$ ならば、$f(0)E(|X|) > \frac{1}{2}$となり、$\pm E(|X|) = \pm 0.47873$ は混合正規分布における Principal Points とはならない (Flury[21]、図 2.3)。この場合では、$c = 0.400$, $h = \pm 0.632$ のときに $H(c,h)$ が最小となる。よって、Principal Points は、$y_1 = 1.032$, $y_2 = -0.232$ と原点に関して非対称になる (Flury[21]、図 2.4)。

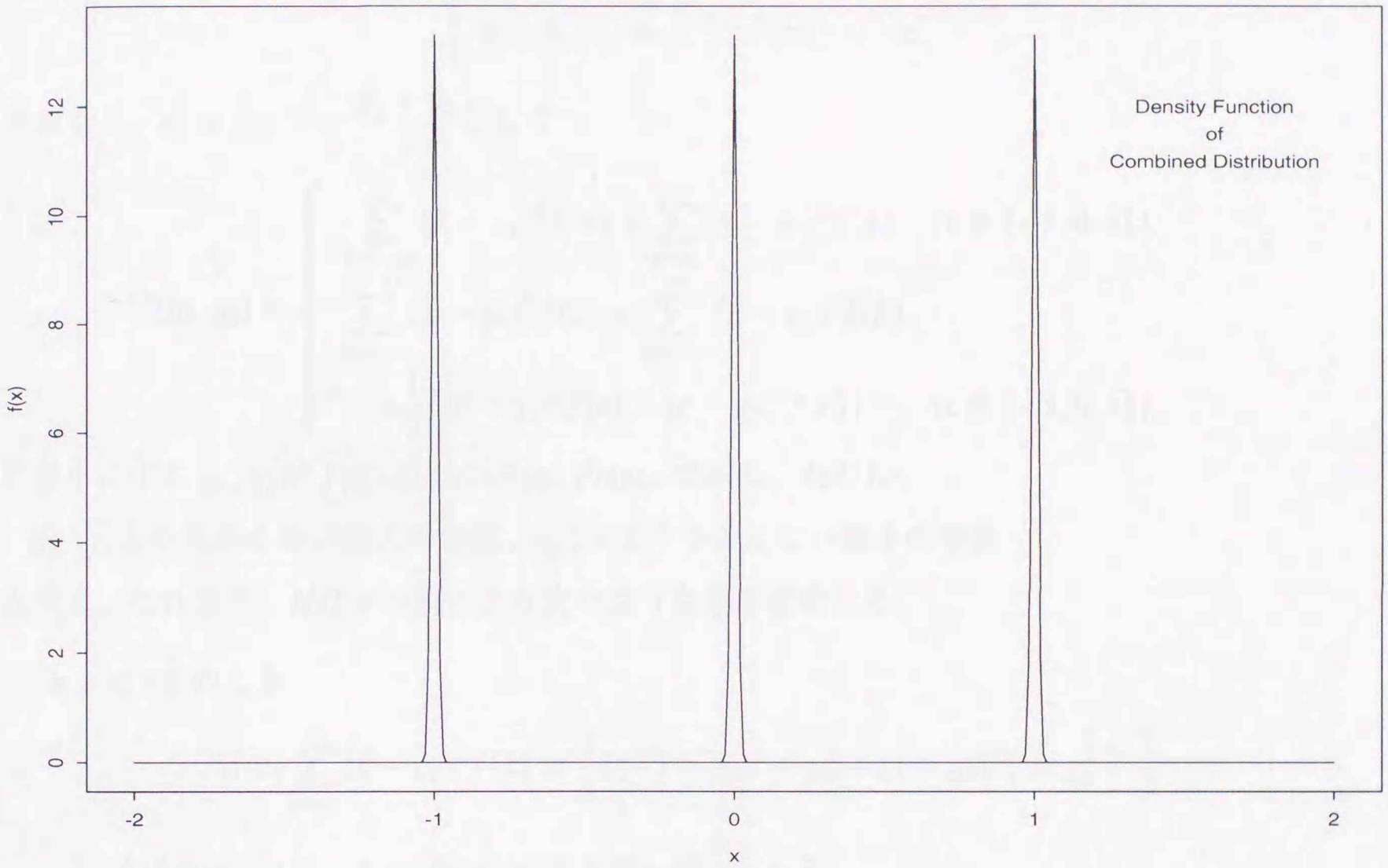


図 2.5: 3 種類の混合正規分布の密度関数 $(\alpha = 0.01)$ (Flury[21])

B　3 種類以上の分布からなる場合　Xの分布関数が

$$F(x) = \frac{1}{3}\{N(x;-1,\alpha^2) + N(x;0,\alpha^2) + N(x;1,\alpha^2)\} \tag{2.2.59}$$

で表されるとする。αが大きければ、Principal Points は$\pm(2/\pi)^{\frac{1}{2}}\alpha$であるが、$\alpha$が十分小さければ、$\alpha \to 0$ とすると $f(0)E(|X|) \to \infty$ となるため、Xの分布は $\mathrm{pr}(X_0 = j) = \frac{1}{3}$ $(j = -1, 0, 1)$ なる離散確率変数 X_0の分布に収束する (図 2.5)。これより、X_0の Principal

Points が $(y_1, y_2) = (-\frac{1}{2}, 1)$ または $(y_1, y_2) = (1, -\frac{1}{2})$ となることが容易に確かめられる (Flury[21])。

証明

離散確率変数 pr の確率関数を

$$f(k) = \begin{cases} \frac{1}{3} & (k = -1, 0, 1) \\ 0 & (k = -\infty, \cdots, -2; 2, \cdots, \infty) \end{cases}$$

とおくと、$y_1 < y_2$、$c = \frac{y_1 + y_2}{2}$として

$$M(y_1, y_2) = \begin{cases} \sum_{k=-\infty}^{c_f} (k - y_1)^2 f(k) + \sum_{k=c_c}^{\infty} (k - y_2)^2 f(k) & (c \notin \{-1, 0, 1\}) \\ \sum_{k=-\infty}^{c-1} (k - y_1)^2 f(k) + \sum_{k=c+1}^{\infty} (k - y_2)^2 f(k) & \\ \quad + \frac{1}{2}\{(c - y_1)^2 f(c) + (c - y_2)^2 f(c)\} & (c \in \{-1, 0, 1\}) \end{cases}$$

を最小にする y_1, y_2が $f(k)$ の Principal Points である。ただし、

c_f：c より大きくない最大の整数、c_c：c より小さくない最小の整数

とする。これより、Hは c の値により次のような最小値をとる。

- $c < -1$ のとき

$$M = \sum_{k=c_c}^{\infty} (k - y_2)^2 f(k) = \frac{1}{3}\{(-1 - y_2)^2 + y_2^2 + (1 - y_2)^2\} = y_2^2 + \frac{2}{3}$$

であるから、$(y_1, y_2) = (2c, 0)$ のとき最小値$\frac{2}{3}$をとる。

- $c = -1$ のとき

$y_1 = -2 - y_2$であることに注意すると

$$\begin{aligned} M &= \sum_{k=-\infty}^{-2} (k - y_1)^2 f(k) + \sum_{k=0}^{\infty} (k - y_2)^2 f(k) + \frac{1}{2}\{(-1 - y_1)^2 f(-1) + (-1 - y_2)^2 f(-1)\} \\ &= \frac{1}{3}\{y_2^2 + (1 - y_2)^2\} + \frac{1}{3}(y_2 + 1)^2 \\ &= y_2^2 + \frac{2}{3} \end{aligned}$$

であるから、$(y_1, y_2) = (-2, 0)$ のとき最小値$\frac{2}{3}$をとる。

- $-1 < c < 0$ のとき

$$\begin{aligned} M &= \sum_{k=-\infty}^{-1}(k-y_1)^2 f(k) + \sum_{k=0}^{\infty}(k-y_2)^2 f(k) \\ &= \frac{1}{3}(-1-y_1)^2 + \frac{1}{3}\{y_2^2 + (1-y_2)^2\} \\ &= \frac{1}{3}(y_1+1)^2 + \frac{2}{3}(y_2-\frac{1}{2})^2 + \frac{1}{6} \end{aligned}$$

であるから、$(y_1, y_2) = (-1, \frac{1}{2})$ のとき最小値$\frac{1}{6}$をとる。これは$-1 < c < 0$ をみたす。

- $c = 0$ のとき

$y_1 = -y_2$であることに注意すると

$$\begin{aligned} M &= \sum_{k=-\infty}^{-1}(k-y_1)^2 f(k) + \sum_{k=1}^{\infty}(k-y_2)^2 f(k) + \frac{1}{2}(y_1^2 f(0) + y_2^2 f(0)) \\ &= \frac{1}{3}(3y_2^2 - 4y_2^2 + 2) \\ &= (y_2 - \frac{2}{3})^2 + \frac{2}{9} \end{aligned}$$

であるから、$(y_1, y_2) = (-\frac{2}{3}, \frac{2}{3})$ のとき最小値$\frac{2}{9}$をとる。

- $0 < c < 1$ のとき

$$\begin{aligned} M &= \sum_{k=-\infty}^{0}(k-y_1)^2 f(k) + \sum_{k=1}^{\infty}(k-y_2)^2 f(k) \\ &= \frac{1}{3}\{(-1-y_1)^2 + y_1^2\} + \frac{1}{3}(1-y_2)^2 \\ &= \frac{2}{3}(y_1+\frac{1}{2})^2 + \frac{1}{3}(y_2-1)^2 + \frac{1}{6} \end{aligned}$$

であるから、$(y_1, y_2) = (-\frac{1}{2}, 1)$ のとき最小値$\frac{1}{6}$をとる。これは $0 < c < 1$ をみたす。

- $c = 1$ のとき

$y_2 = 2 - y_1$であることに注意すると

$$\begin{aligned} M &= \sum_{k=-\infty}^{0}(k-y_1)^2 f(k) + \sum_{k=2}^{\infty}(k-y_2)^2 f(k) + \frac{1}{2}\{(1-y_1)^2 f(1) + (1-y_2)^2 f(1)\} \\ &= \frac{1}{3}\{(-1-y_1)^2 + y_1^2\} + \frac{1}{3}(y_1-1)^2 \\ &= y_1^2 + \frac{2}{3} \end{aligned}$$

であるから、$(y_1, y_2) = (0, 2)$ のとき最小値$\frac{2}{3}$をとる。

- $1 < c$のとき

$$M = \sum_{k=-\infty}^{c_f} (k - y_1)^2 f(k) = \frac{1}{3}\{(1 - y_1)^2 + y_1^2 + (-1 - y_1)^2\} = y_1^2 + \frac{2}{3}$$

であるから、$(y_1, y_2) = (0, 2c)$ のとき最小値$\frac{2}{3}$をとる。

以上より、X_0の Principal Points は $(y_1, y_2) = (-\frac{1}{2}, 1)$ または $(y_1, y_2) = (1, -\frac{1}{2})$ となる。(証明終)

Flury[21] はさらに、4 種類の分布からなる混合正規分布関数

$$F(x) = \frac{1}{2}(1-\gamma)N(x; 0, 1^2) + \frac{1}{2}(1-\gamma)N(x; 0, 0.2^2) + \frac{1}{2}\gamma N(x; -\theta, \alpha^2) + \frac{1}{2}\gamma N(x; \theta, \alpha^2) \tag{2.2.60}$$

を考えている。ただし、γ、θ、αは正である。ここで、γが十分小さければ、Principal Points は 3 種類の分布からなる場合とほぼ同じとなり、0 に関して非対称となる。

2.2.2.8 まとめ

以上の例および定理 1 より、密度関数が平均に関して対称である分布のうち、標準正規分布や一様分布、ロジスティック分布、t 分布、Johnson's S_u分布については 2-Principal Points が期待値に関して対称となったが、混合正規分布においては 2-Principal Points が非対称となる場合があることがわかる。しかし、どのような場合に 2-Principal Points が対称(もしくは非対称) となるのかはこの時点では十分に考察されていない。

2.2.3 標準正規分布における k個の主要点

標準正規分布が与えられた場合において、Cox[9] は最小 2 乗距離規準により、$2 \leq k \leq 6$ の場合における最適なクラスター分割点の値を分布からの 100 個の観測値による数値計算により求めている (表 2.1)。一方、k-Principal Points $(1 \leq k \leq 5)$ の数値は表 2.2のようになる (Flury[21])。ただし、$k \geq 3$ の時は、数値積分及び繰り返し計算による目的関数の最小化を用いている。

表 2.1: 1 変量標準正規分布ϕにおける各クラスターの最適分割点の値 (Cox[9])

k	クラスター分割点	$P_\phi(k)$
2	0.0	0.3634
3	-0.612,0.612	0.1902
4	-0.980,0.0,0.980	0.1175
5	-1.230,-0.395,0.395,1.230	0.0799
6	-1.449,-0.660,0.0,0.660,1.449	0.0580

表 2.2: 1 変量標準正規分布ϕにおける k個の主要点 (Flury[21])

k	Principal Points	$P_\phi(k)$
1	0.0	1.0000
2	$-(2/\pi)^{1/2}, (2/\pi)^{1/2}$ $(\simeq \pm 0.79788)$	$1-2/\pi$ $(\simeq 0.3634)$
3	-1.227,0.0,1.227	0.1900
4	-1.507,-0.451,0.451,1.507	0.1170
5	-1.707,-0.754,0.0,0.754,1.707	0.0800

表2.1および表2.2より、$2 \leq k \leq 5$の場合においては、標準正規分布における k-Principal Points の中で各々隣接する 2 点の中点の値の集合が標準正規分布を k個の最適なクラスターに分割したときの点の値の集合にほぼ一致している。

ここでは、k-Principal Points がすべて平均に関して対称となっているが、平均に関して非対称となるような混合正規分布を作ることもできる。$k = 3$ の場合については第 3.2.3節で述べる。

2.2.4 k個の主要点の対称性に関する定理

第 2.2.1節から第 2.2.3節までの考察においては、k-Principal Points の対称性に関する直接的な定理は導出されていない。また、kが大きいときの対称性については数値計算による値からしか論じられていない。

そこで、新たなアプローチとして、密度関数の性質に着目した研究が行われている。Tarpey[61] は対称な 1 変量確率分布における 2-Principal Points の対称性に関する定理を初めて示した。山本他 [87][88] は Flury[21] および Tarpey[61] の研究を拡張し、2-Principal Points の対称性に関するより詳細な十分条件の導出を行っている。

一方、Li & Flury[34] は対称な 1 変量確率分布における k-Principal Points の対称性に関する定理を導出したが、この定理は一般には成立しない。

本小節では、密度関数の性質に着目した以上の研究およびそれらの研究に関連する内容について紹介する。

2.2.4.1 自己一致点

ここでは、Flury[22] に従い、自己一致点 (self-consistent points) という概念についての定義を述べる。

自己一致 (self-consistent) という概念は、Hastie & Stuetzle[29] における主曲線 (Principal Curves) の定義に関して考えられたものである。すなわち、多次元データのある種の「中心」を通る曲線上の各点において、その点が最も近いような点の集合の重心が曲線上の点と一致している状態を自己一致といい、自己一致する曲線を Principal Curves と定義している。

Principal Curves の場合と同様に、多変量確率分布に従う領域を k個に分割した際の各領域における代表点が各領域における重心と一致する状態を自己一致といい、自己一致する点を自己一致点と定義する。数学的には以下のように書ける。

定義 (自己一致点 (self-consistent points)) (Flury[22])

異なる k個の点$\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k$および p 変量確率変数 X上の任意の$\boldsymbol{x}$に関し、$i = 1, \cdots, k$において領域 $A_i = \{\boldsymbol{x}| \ |\boldsymbol{x} - \boldsymbol{y}_i| < |\boldsymbol{x} - \boldsymbol{y}_j| \ {}^\forall j \neq i\}$ を考える。このとき、$E[\boldsymbol{x}|\boldsymbol{x} \in A_i] = \boldsymbol{y}_i$ がすべての $i = 1, \cdots, k$で成り立てば、k個の点$\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k$を Xの自己一致点と呼ぶ。

自己一致点と Principal Points との関連については、以下の補題が成り立つ。

補題 1. (Flury[22])

Principal Points は自己一致点である。

証明

$\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_k$を p 変量確率分布 Fに従う p 変量確率変数 Xの k-Principal Points であるとする。ここで、$i = 1, \cdots, k$において領域 $A_i = \{\boldsymbol{x}| \ |\boldsymbol{x} - \boldsymbol{\xi}_i| < |\boldsymbol{x} - \boldsymbol{\xi}_j| \ {}^\forall j \neq i\}$ を考え、さらに$\pi_i = \Pr(X \in A_i)$ とすると、Principal Points の定義および (2.1.3) 式、(2.1.4) 式より

$$\begin{aligned} P_F(k) &= E\{d^2(X|\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_k)\} \\ &= \sum_{i=1}^{k} \pi_i E\{(X - \boldsymbol{\xi}_i)'(X - \boldsymbol{\xi}_i)|X \in A_i\} \\ &= \min_{\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\}} \left[\sum_{i=1}^{k} \pi_i E\{(X - \boldsymbol{y}_i)'(X - \boldsymbol{y}_i)|X \in A_i\}\right] \\ &= \sum_{i=1}^{k} \pi_i \min_{\boldsymbol{y}_i} [E\{(X - \boldsymbol{y}_i)'(X - \boldsymbol{y}_i)|X \in A_i\}] \end{aligned} \tag{2.2.61}$$

と書ける。ここで (2.2.61) 式における i 番目の項は$\boldsymbol{y}_i = E(X|X \in A_i)$ とすることにより最小化される。すなわち、Principal Points は条件つき期待値であり、自己一致点となる。(証明終)

Principal Points は必ず自己一致点になるが、逆は必ずしも成立しない。そこで、以下では自己一致点の一意性を検証することにより得られる、対称な 1 変量確率分布における k-Principal Points の対称性に関する種々の定理を示す。

2.2.4.2 Tarpey の定理

Tarpey[61] は、2-Principal Points の対称性に関する以下の定理を示している。

定理 2. (Tarpey[61])

1 変量確率分布 Fの密度関数 fが対称かつ強単峰であれば、2-Principal Points は対称となる。

証明

一般性を失うことなく fの期待値が原点であるとし、y_1、y_2を Fに従う確率変数 Xの自己一致点であるとする。このとき、y_1、y_2を区切り点 x の関数として

$$y_1(x) = \frac{\int_{-\infty}^{x} tf(t)dt}{\int_{-\infty}^{x} f(t)dt} \tag{2.2.62}$$

$$y_2(x) = \frac{\int_{x}^{\infty} tf(t)dt}{\int_{x}^{\infty} f(t)dt} \tag{2.2.63}$$

と定義すると、自己一致点の定義より x と $y_1(x)$ および $y_2(x)$ との間には

$$S(x) = 2x - y_1(x) - y_2(x) \tag{2.2.64}$$

とおいたときに $S(x) = 0$ という関係が成り立つ。

(2.2.64) 式より、$S(x) = 0$ ならば $-(x - y_1(x)) = x - y_2(x)$ であり、

$$S'(x) = 2 - \frac{(y_2(x) - x)f(x)}{F(x)(1 - F(x))} \tag{2.2.65}$$

となる。

ここで、$f_1(x) = \int_x^\infty (t-x)f(t)dt$ とおくと、$f_1(x)$ は 2 次の Polya 頻度関数 (PF_2関数) である (Karlin[32])。すなわち、任意の $u_1 < u_2$、$v_1 < v_2$に関して

$$\begin{vmatrix} f_1(u_1 - v_1) & f_1(u_1 - v_2) \\ f_1(u_2 - v_1) & f_1(u_2 - v_2) \end{vmatrix} \geq 0$$

が成り立つ (Karlin[32])。PF_2関数は log-concave、すなわち

$$\frac{d^2}{dx^2} \log f_1(x) = f_1(x)f_1''(x) - (f_1'(x))^2 \leq 0 \tag{2.2.66}$$

であり、等号が成り立つのは Fが両側指数分布 (二重指数分布、ラプラス分布) のときである (Karlin[32])。ここで $f_1'(x) = -(1 - F(x))$、$f_1''(x) = f(x)$ であるので、(2.2.66) 式は

$$\frac{f(x)(x - y_2(x))}{1 - F(x)} \leq 1 \tag{2.2.67}$$

と書ける。さらに、$x \geq 0$ においては $\dfrac{1}{F(x)} \leq 2$ なので、(2.2.67) 式より

$$2 - \frac{(y_2(x) - x)f(x)}{F(x)(1 - F(x))} \geq 0 \tag{2.2.68}$$

が成り立つ。

(2.2.64) 式および (2.2.68) より、$S(x) = 0$ なるすべての $x > 0$ において $S'(x) > 0$ となる。また $S(0) = 0$ かつ $S'(0) \geq 0$ である。ここで、$S(x_1) = 0$ なる $x_1 > 0$ が存在するならば、$S(x)$ の連続性より $S(x_2) = 0$ かつ $S'(x_2) < 0$ をみたす x_2が区間 $(0, x_1)$ 内に存在しなければならないが、そのような x_1および x_2は存在し得ない。よって、$x > 0$ のとき $S(x) \neq 0$ となる。また、$S(x)$ は原点に関して対称であるので、$x < 0$ のときにも $S(x) \neq 0$ が成立する。

以上より、$S(x) = 0$ をみたすのは $x = 0$ のときに限られ、それに伴い自己一致点 y_1、y_2は一意に定まる。さらに補題 1 より Principal Points は自己一致点となるので、一意に定まった 2 つの自己一致点は 2-Principal Points となり、期待値 (原点) に関して対称となる。(証明終)

定理 2 は、2-Principal Points の対称性に関する具体的な十分条件を示した最初の定理である。一方、山本・篠崎 [87][88] は、正規分布など、定理 2 をみたす各種強単峰分布に加え、t 分布なども (2.2.25) 式で示される十分条件をみたし、これらの分布における 2-Principal Points が対称であることを示している。

2.2.4.3 Li & Flury の定理

Li & Flury[34] は、対称な 1 変量確率分布における k-Principal Points の対称性に関して以下の定理を主張している。

定理 3. (Li & Flury[34])

対称な 1 変量確率分布において密度関数 fが $f \cdot f'' - 2(f')^2 < 0$ をみたすならば、あらゆる kにおいて k-Principal Points は期待値に関して対称となる。

証明

1 変量確率変数 Xにおける分布関数 Fの分位関数を $Q(u) = F^{-1}(u)$ とする。このとき、

$F(Q(u)) = u$ である。この両辺を u で微分すると

$$Q'(u)f(Q(u)) = 1 \tag{2.2.69}$$

となり、さらに (2.2.69) 式の両辺の対数をとると

$$\log Q'(u) + \log f(Q(u)) = 0$$

となる。ここで $g(u) = -\log f(Q(u)) = \log Q'(u)$ とおくと、

$$g'(u) = -\frac{f'(Q(u))Q'(u)}{f(Q(u))} = -f'(Q(u))(Q'(u))^2 \tag{2.2.70}$$

であり、かつ $Q''(x) = -\dfrac{f'(Q(x))Q'(x)^2}{f(Q(x))}$ より

$$\begin{aligned} g''(u) &= -\{f''(Q(u))(Q'(u))^3 + 2Q'(u)Q''(u)f'(Q(u))\} \\ &= -\left\{\frac{f''(Q(u))(Q'(u))^3 - 2Q'(u)f'(Q(u))(Q'(u))^2 f'(Q(u))}{f(Q(u))}\right\} \\ &= -\frac{f''(Q(u))f(Q(u)) - 2(f'(Q(u)))^2}{f(Q(u))^4} \\ &= \frac{2(f'(Q(u)))^2 - f(Q(u))f''(Q(u))}{f(Q(u))^4} \end{aligned}$$

となる。$g''(u) > 0$ ならば $g(u)$ は凹関数であるので、これより、$g''(u) > 0 \Leftrightarrow f \cdot f'' - 2(f')^2 < 0$ が成り立つ。(証明終)

fが 2 階微分可能であるとき

$$(\log f)'' = f \cdot f'' - (f')^2 \leq 0 \tag{2.2.71}$$

は fが強単峰であるための十分条件である。定理 3 における条件式をみたす確率分布族は、(2.2.71) 式をみたす確率分布族を包含する。従って、定理 3 は定理 2 の拡張となっている。しかし、実際には定理 3 は一般には成立しない。

そこで、以下では定理 3 の導出に利用された Chow[7] の定理を紹介し、この定理と Principal Points との関係について示す。

2.2.4.4 Chow の定理と主要点との関係

Chow[7] は、関数 gを区分的多項式により最小 2 乗近似する場合において、近似の一意性に関する以下の定理を示している。

定理 4. (Chow[7])

gが区間 $[0,1]$ において C^n-級かつ n 次導関数 $g^{(n)}$が $[0,1]$ において正であるとする。ここで、区間 $(0,1)$ において $\log g^{(n)}$が凹関数ならば、N個の異なる区切り点をもつ $(n-1)$ 次の区分的多項式による gの最小 2 乗近似はあらゆる自然数 Nに対して一意に定まる。

1 変量確率変数 Xにおける C^1-級分布関数 Fの分位関数を $Q(u) = F^{-1}(u)$ とする。ただし、Q'が区間 $[0,1]$ で連続かつ正であると仮定する。このとき、(2.2.69) 式より Q'が連続かつ正なので fも連続かつ正である。

ここで、Q を区間 $[0,1]$ において $(k-1)$ 個の区切り点$\nu_1 < \cdots < \nu_{k-1}$をもつ 0 次の区分的多項式 (区分的定数関数) により最小 2 乗近似したときの k個の関数値を $y_1 < \cdots < y_k$とすると、

$$2Q(\nu_i) - y_i - y_{i+1} = 0 \tag{2.2.72}$$

が一意に定まる十分条件は $\log Q'$が区間 $(0,1)$ において凹関数であることが Eubank[16] により示されている。このとき、

$$\nu_i = F(\frac{y_i + y_{i+1}}{2}) \tag{2.2.73}$$

であるので、最小 2 乗近似による誤差は

$$r(y_1, \cdots, y_k) = \sum_{i=1}^{k} \int_{F(\frac{y_{i-1}+y_i}{2})}^{F(\frac{y_i+y_{i+1}}{2})} (Q(u) - y_i)^2 du \tag{2.2.74}$$

と表される。ただし、$y_0 = Q(0)$、$y_{k+1} = Q(1)$ である (Eubank[16])。

(2.2.74) 式を y_i $(i = 1, \cdots, k)$ で偏微分すると

$$\frac{\partial r}{\partial y_i} = -2 \int_{F(\frac{y_{i-1}+y_i}{2})}^{F(\frac{y_i+y_{i+1}}{2})} (Q(u) - y_i) du \tag{2.2.75}$$

となる。ここで、$y_1, \cdots, y_k$は最小 2 乗近似における k個の関数値であるから$\frac{\partial r}{\partial y_i} = 0$ $(i =$

$1, \cdots, k)$ が成立する。この式を y_iについて解くと

$$y_i = \frac{\int_{F(\frac{y_{i-1}+y_i}{2})}^{F(\frac{y_i+y_{i+1}}{2})} Q(u)du}{F(\frac{y_i+y_{i+1}}{2}) - F(\frac{y_{i-1}+y_i}{2})} \tag{2.2.76}$$

と書ける。

ここで $Q(u) = x$ とおくと、$\frac{dx}{du} = Q'(u)$ なので、(2.2.69) 式は $\frac{dx}{du} f(x) = 1$ と表せる。これより (2.2.76) 式は

$$y_i = \frac{\int_{\frac{y_{i-1}+y_i}{2}}^{\frac{y_i+y_{i+1}}{2}} xf(x)dx}{\int_{\frac{y_{i-1}+y_i}{2}}^{\frac{y_i+y_{i+1}}{2}} f(x)dx} \quad (i = 1, \cdots, k) \tag{2.2.77}$$

となる。

ここで、異なる k個の点 $y_1, \cdots, y_k$および X上の任意の x に関し、$i = 1, \cdots, k$において区間 $A_i = \{x| \ |x - y_i| < |x - y_j| \ \ ^\forall j \neq i\}$ を考える。このとき、$E[x|x \in A_i] = y_i$がすべての $i = 1, \cdots, k$で成り立てば、k個の点 $y_1, \cdots, y_k$は Xの自己一致点である (Flury[22])。ここで、(2.2.72) 式において区切り点ν_iにおける分位関数の値 $Q(\nu_i)$ は y_iと y_{i+1}の中点となるので $A_i = (Q(\nu_{i-1}), Q(\nu_i))$ であり、さらに (2.2.77) 式より $i = 1, \cdots, k$で y_iは区間 $A_i = (Q(\nu_{i-1}), Q(\nu_i))$ における重心となるので、最小 2 乗近似の値 $y_1, \cdots, y_k$は自己一致点である。

ここで、異なる k個の点に関し、連続な確率分布に従う領域を各点からの最小 2 乗距離により k個に分割することを考える。このとき、各領域における代表点からの 2 乗距離の期待値の総和が最小となるような k個の点が常に存在することが Pärna[49] により証明されている。また、補題 1 より Principal Points は自己一致点となるので、一意に定まった自己一致点 $y_1, \cdots, y_k$は k-Principal Points となる。

2.2.4.5 Trushkin の定理と主要点との関係

Trushkin[63] は、平均 2 乗誤差規準により連続な 1 変量確率変数を k個の値で離散化する場合において、離散化の一意性に関する以下の定理を示している。なお、この定理は、Chow[7] とは独立に導出されたものである。

定理 5. (Trushkin[63])

連続な 1 変量確率変数 Xの密度関数 $f(x)$ に関し、区間 $I = \{x|f(x) > 0\}$ が連結であり、かつ fが Iにおいて連続であるとする。このとき、$\log f$が I において凹関数ならば、Xを k個の値に離散化する際に平均 2 乗誤差が最小となるような離散値の集合はあらゆる自然数 kに対して一意に定まる。

最小 2 乗距離により Xを離散化する k個の点を $y_1 < \cdots < y_k$とする。ここで $x_1 < \cdots < x_{k-1}$ を区間 Iにおける点とし、$x_0 = \inf\{x|x \in I\}$、$x_k = \sup\{x|x \in I\}$ と定義する。さらに、$i = 1, \cdots, k$において区間 (x_{i-1}, x_i) が $(x_{i-1}, x_i) = \{x|\ |x - y_i| < |x - y_j|\ \ ^\forall j \neq i\}$ をみたすならば、$x_1, \cdots, x_{k-1}$および $y_1, \cdots, y_k$に関して

$$2x_i - y_i - y_{i+1} = 0 \tag{2.2.78}$$

が成り立つ。このとき、Xの平均 2 乗誤差は

$$H(y_1, \cdots, y_k) = \frac{y_i + y_{i+1}}{2}(x - y_i)^2 f(x)dx \tag{2.2.79}$$

と表される。

(2.2.79) 式を y_i $(i = 1, \cdots, k)$ で偏微分すると

$$\frac{\partial H}{\partial y_i} = -2\int_{\frac{y_{i-1}+y_i}{2}}^{\frac{y_i+y_{i+1}}{2}} (x - y_i)f(x)dx \tag{2.2.80}$$

となる。ここで、$y_1, \cdots, y_k$は最小 2 乗距離による Xの離散化の k個の値であるから、$\dfrac{\partial H}{\partial y_i} = 0$ $(i = 1, \cdots, k)$ が成立し、この式を y_iについて解くと (2.2.80) 式より

$$y_i = \frac{\displaystyle\int_{\frac{y_{i-1}+y_i}{2}}^{\frac{y_i+y_{i+1}}{2}} xf(x)dx}{\displaystyle\int_{\frac{y_{i-1}+y_i}{2}}^{\frac{y_i+y_{i+1}}{2}} f(x)dx} \tag{2.2.81}$$

と書ける。すなわち、$y_1, \cdots, y_k$が各点から最も近い領域における重心となるときに Xの平均 2 乗誤差が最小になることが Tarpey, Li & Flury[62] により示されている。このとき、$y_1, \cdots, y_k$は前節と同様に Xにおける k個の自己一致点 (Flury[22]) となり、一意に定まった $y_1, \cdots, y_k$は k-Principal Points となる。

2.3　多変量楕円分布の場合

本節では、多変量楕円分布における k-Principal Points に関して行われた理論的考察について述べる。

2.3.1　2 個の主要点に関する理論的考察

p 変量確率変数 Xが n 個の値からなる離散的変数であるならば、式 (2.1.3) は級内平方和である。もし、Xが平均ベクトル $\mathbf{0}$、分散共分散行列 $\mathrm{diag}(\sigma_1^2, \cdots, \sigma_p^2)$ をもつ多変量正規分布に漸近的に従い、$\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_p^2$とすると、分散共分散行列の固有値が$\sigma_i^2$ $(i = 1, \cdots, p)$ であることから、Hartigan は式 (2.1.3) が漸近的に期待値$\sum_{i=1}^{p} \sigma_i^2 - \frac{2\sigma_1^2}{\pi}$、分散$\frac{2}{n}\sum_{i=1}^{p} \sigma_i^4 - \frac{16\sigma_1^4}{\pi^2}$ の正規分布に従うと推測した [27][28]。この推測を確かめ、2 次のモーメントが有限な任意の多変量楕円分布に一般化させたのが定理 6 である。ただし、p 変量楕円分布 Xとは、$\boldsymbol{\mu} = E(X) \in R^p$, $\Sigma = V(X) = E[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})']$ として、密度関数を

$$f(\boldsymbol{x}) = g[(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})] \quad (\text{ただし} \int f(\boldsymbol{x})d\boldsymbol{x} = 1, g\text{: 実数関数})$$

で表すことができる分布である。

定理 6. (Flury[21])

p 変量確率ベクトル $X = (X_1, \cdots, X_p)'$が楕円分布に従い、平均ベクトル$\boldsymbol{\mu}$、分散共分散行列ψをもつとすると、Xの 2-Principal Points は$\boldsymbol{y}_1 = \boldsymbol{\mu} + \gamma_1\boldsymbol{\beta}$, $\boldsymbol{y}_2 = \boldsymbol{\mu} + \gamma_2\boldsymbol{\beta}$ となる。ただし、

$\boldsymbol{\beta} \in R^p$：$\psi$の最大固有値に対応する正規化固有ベクトル

γ_1, γ_2：1 変量確率変数$\boldsymbol{\beta}'(X - \boldsymbol{\mu})$ の 2-Principal Points

とする。

証明

一般性を失わずに$\boldsymbol{\mu} = \mathbf{0}$ とおくことができる。ここで、$\boldsymbol{c}_1$, $\boldsymbol{c}_2$を R^pにおける任意の点、$P_X(\boldsymbol{c}_1, \boldsymbol{c}_2) = E\{d^2(X|\boldsymbol{c}_1, \boldsymbol{c}_2)\}$ とする。

step 1.

(2 点$\boldsymbol{c}_1$, $\boldsymbol{c}_2$が、平均を通る方向ベクトル$\boldsymbol{c}_2 - \boldsymbol{c}_1$の直線上にあるときに P_Xが小さくなる)

$\boldsymbol{c}_1, \boldsymbol{c}_2$を異なる任意の 2 点として$\boldsymbol{a} = \dfrac{\boldsymbol{c}_2 - \boldsymbol{c}_1}{||\boldsymbol{c}_2 - \boldsymbol{c}_1||}$とおく。ただし、分母はノルムを表す。
ここで、$(\boldsymbol{a}, \boldsymbol{B})$が p 次の直交行列となるように$\boldsymbol{B}$を定め、

$$Y = \begin{bmatrix} Y_1 \\ \\ Y_2 \end{bmatrix} = (\boldsymbol{a}, \boldsymbol{B})'X, \quad \boldsymbol{v}_i = \begin{bmatrix} v_{i1} \\ \\ \boldsymbol{v}_{i2} \end{bmatrix} = (\boldsymbol{a}, \boldsymbol{B})'\boldsymbol{c}_i \quad (i = 1, 2)$$

とし、さらに

$$Y_1 = \boldsymbol{a}'X, \quad v_{i1} = \boldsymbol{a}'\boldsymbol{c}_i$$

$$Y_2 = \boldsymbol{B}'X, \quad \boldsymbol{v}_{i2} = \boldsymbol{B}'\boldsymbol{c}_i$$

とおくと、$(\boldsymbol{a}, \boldsymbol{B})$の直交性より$\boldsymbol{v}_{12} = \boldsymbol{v}_{22} = \boldsymbol{v}_{(2)}$と推測できるから、

$$\begin{aligned} E\{d^2(X|\boldsymbol{c}_1, \boldsymbol{c}_2)\} &= E\{d^2(Y|\boldsymbol{v}_1, \boldsymbol{v}_2)\} \\ &= E\{d^2(Y_1|v_{11}, v_{21})\} + E\{d^2(Y_2|\boldsymbol{v}_{12}, \boldsymbol{v}_{22})\} \\ &= E\{d^2(Y_1|v_{11}, v_{21})\} + E\{d^2(Y_2|\boldsymbol{v}_{(2)}, \boldsymbol{v}_{(2)})\} \\ &\geq E\{d^2(Y_1|v_{11}, v_{21})\} + E\{d^2(Y_2|\boldsymbol{0}, \boldsymbol{0})\} \end{aligned}$$

となる (等号は$\boldsymbol{v}_{(2)} = \boldsymbol{0}$ のとき)。これは、Xが従う分布によらず、一般的に成り立つ。

step 2.

(ある線形補空間上に Principal Points が存在していれば、それらは Xの周辺分布の点である)

Y_1は 1 変量楕円分布に従うから、分布は 0 に関して対称であり、期待値 $E\{d^2(Y_1|v_{11}, v_{21})\}$は、$v_{11}, v_{21}$が Y_1の 2-Principal Points となるとき最小となる。このときの点の座標値をξ_1, ξ_2として、

$$\boldsymbol{v}_i^* = \begin{bmatrix} \xi_i \\ \\ 0 \end{bmatrix} \quad (i = 1, 2)$$

とすると、$E\{d^2(X|\boldsymbol{v}_1, \boldsymbol{v}_2)\} \geq E\{d^2(X|\boldsymbol{v}_1^*, \boldsymbol{v}_2^*)\}$ である。これより、確率ベクトル Xに関し、$\boldsymbol{c}_i^* = [\boldsymbol{a}, \boldsymbol{B}]\boldsymbol{v}_i^* = \xi_i\boldsymbol{a}$ $(i = 1, 2)$ とすると、すべての$\boldsymbol{c}_1, \boldsymbol{c}_2$について

$$E\{d^2(X|\boldsymbol{c}_1^*, \boldsymbol{c}_2^*)\} \leq E\{d^2(X|\boldsymbol{c}_1, \boldsymbol{c}_2)\}$$

が成り立つ。

step 3.

(2-Principal Points を含む 1 次元の線形補空間はψの最初の固有ベクトルによって作られる補空間として定義される)

ここで、簡単な補題を述べる。

補題 2.

ある$\rho \in R$に関し、Y_2とρY_1が同じ分布をもつような確率変数 Y_1, Y_2を定めると、すべての $k \geq 1$ に対して

$$\frac{P_{Y_1}(k)}{\mathrm{Var}(Y_1)} = \frac{P_{Y_2}(k)}{\mathrm{Var}(Y_2)}$$

が成り立つ。

ここで、$\boldsymbol{a}$を$\boldsymbol{a}'\boldsymbol{a} = 1$ なる任意のベクトルとし、$Y_{\boldsymbol{a}} = \boldsymbol{a}'X$とすると、$\mathrm{Var}(Y_{\boldsymbol{a}}) = \boldsymbol{a}'\psi\boldsymbol{a}$であるから、補題 2 より $P_{Y_{\boldsymbol{a}}}(2) = g\boldsymbol{a}'\psi\boldsymbol{a}$ $(g < 1)$ となる。$\xi_1\boldsymbol{a}, \xi_2\boldsymbol{a}$を $Y_{\boldsymbol{a}}$の 2-Principal Points とすると、

$$E\{d^2(X|\boldsymbol{a}\boldsymbol{\xi}_{1a}, \boldsymbol{a}\boldsymbol{\xi}_{2a})\} = \mathrm{tr}(\psi) - (1-g)\boldsymbol{a}'\psi\boldsymbol{a}$$

となる。これを最小化するには、$\boldsymbol{a}'\psi\boldsymbol{a}$を最大化すればよく、その解は、$\boldsymbol{a}$が$\psi$の最大固有値に対応する固有ベクトルになるときである。

ここで、以下の 2 つの系より、定理 6 を特殊な楕円分布に対して適用できる。

系 1. (Flury[21])

Xを p 変量正規分布 $N_p(\boldsymbol{\mu}, \psi)$ に従う確率変数とし、τ_1, $\boldsymbol{\beta}_1$をそれぞれψの最大固有値、最大固有ベクトルとする。ここで、τ_1の多重度が 1 であれば、Xの 2-Principal Points は $\boldsymbol{\mu} \pm (2\tau_1/\pi)^{\frac{1}{2}}\boldsymbol{\beta}_1$となる。多重度が r $(1 < r \leq p)$ であれば、Principal Points はτ_1の潜在空間中の中心$\boldsymbol{\mu}$、半径 $(2\tau_1/\pi)^{\frac{1}{2}}$の範囲における、$\boldsymbol{\mu}$に関して対称なあらゆる 2 点である。

系 2. (Flury[21])

ψを p 次正定値対称行列、$\boldsymbol{\mu} \in R^p$とする。確率変数 Xが集合

$$C = \{\boldsymbol{x} \in R^p : (\boldsymbol{x} - \boldsymbol{\mu})'\psi^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq 1\}$$

の内部で一様分布 Uに従い、τ_1, $\boldsymbol{\beta}_1$を系 1 と同様に定義して、$\boldsymbol{\beta}_1'\boldsymbol{\beta}_1 = 1$が成り立つとする。

ここでτ_1の多重度が1であれば、Xの2-Principal Pointsは

$$\boldsymbol{\mu} \pm \frac{2\sqrt{\tau_1}\Gamma\left(\frac{1}{2}p+1\right)}{(p+1)\Gamma\left(\frac{1}{2}p+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)}\boldsymbol{\beta}_1$$

である。

系2の証明

系1より、多変量正規分布の一般的な特性がわかる。系2を証明するために

$$f(\boldsymbol{\mu})E(|U-\boldsymbol{\mu}|) = \frac{2}{\pi}\frac{\Gamma^2\left(\frac{1}{2}p+1\right)}{(p+1)\Gamma^2\left(\frac{1}{2}p+\frac{1}{2}\right)} < \frac{1}{2}$$

をみたすすべてのpについて$f(\boldsymbol{\mu})E(|U-\boldsymbol{\mu}|)$を計算すると、$f(\boldsymbol{\mu})E(|U-\boldsymbol{\mu}|) \to \frac{1}{\pi}$ $(p \to \infty)$となる。これより、系2において与えられた2点は目的関数の極小値を与える。もっとも、この場合は2-Principal Pointsが一意であることは明らかと思われるが、Uの分布関数中の不完全なベータ積分のため、Principal Pointsの一意性はまだ証明されていない。

2.3.2 k個の主要点に関する理論的考察

任意の多変量分布のPrincipal Pointsに関する一般的な性質を述べるのは困難であり、実際には、単純な解析的結果が有用であれば、クラスターを見つけるための数値的アルゴリズムは不要と思われるが、以下の定理が証明可能である。

定理7. (Flury[21])

p変量確率ベクトルXが、平均$\boldsymbol{\mu}$、k-Principal Points $\boldsymbol{\xi}_1 \cdots, \boldsymbol{\xi}_k \in R^p$をもつとすると、$\mathrm{rank}(\boldsymbol{\xi}_1 - \boldsymbol{\mu} \cdots, \boldsymbol{\xi}_n - \boldsymbol{\mu}) < k$が成り立つ。

証明

$\boldsymbol{c}_1, \cdots, \boldsymbol{c}_k$を任意の点、$\boldsymbol{b}_i = \boldsymbol{c}_i - \boldsymbol{c}_k$ $(i = 1, \cdots, k)$とする。$m = \mathrm{rank}(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{k-1})$とすると、$m \leq k-1$は明らかである。また、$\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m$を$\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{k-1}$によって補われる線形補空間の直交基底とし、$A_1 = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m)$として$A = (A_1, A_2)$が$p$次の直交行列となるように$A_2$を定める。ここで、

$$\boldsymbol{d}_i = A'\boldsymbol{c}_i = \begin{bmatrix} A_1'\boldsymbol{c}_i \\ \\ A_2'\boldsymbol{c}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{d}_i^{(1)} \\ \\ \boldsymbol{d}_i^{(2)} \end{bmatrix} \quad (i = 1, \cdots, k)$$

とおく。なぜなら、A_1の各行によってできる補空間内のすべての$\boldsymbol{c}_i - \boldsymbol{c}_j$において、$\boldsymbol{c}_i - \boldsymbol{c}_j = A_i M_{ij}$ ($M_{ij} : m \times m$ 行列) となるからである。従って、

$$\boldsymbol{d}_i{}^{(2)} - \boldsymbol{d}_j{}^{(2)} = A_2(\boldsymbol{c}_i - \boldsymbol{c}_j) = A_2' A_1 M_{ij} = 0$$

がすべての (i, j) について成り立ち、$\boldsymbol{d}_i{}^{(2)}$はすべて等しくなるから、$\boldsymbol{d}_i{}^{(2)} = \boldsymbol{d}^{(2)}$ $(i = 1, \cdots, k)$ とおける。ここで、一般性を失わずに$\boldsymbol{\mu} = \mathbf{0}$ とおくことができ、

$$Y = \begin{bmatrix} Y_1 \\ \\ Y_2 \end{bmatrix} = A'X = \begin{bmatrix} A_1'X \\ \\ A_2'X \end{bmatrix}$$

とおくと、A は直交行列だから、

$$\begin{aligned} E\{d^2(X|\boldsymbol{c}_1, \cdots, \boldsymbol{c}_k)\} &= E\{d^2(Y|\boldsymbol{d}_1, \cdots, \boldsymbol{d}_k)\} \\ &= E\{d^2(Y_1|\boldsymbol{d}_1{}^{(1)}, \cdots, \boldsymbol{d}_k{}^{(1)})\} + E\{d^2(Y_2|\boldsymbol{d}^{(2)}, \cdots, \boldsymbol{d}^{(2)})\} \\ &\geq E\{d^2(Y_1|\boldsymbol{d}_1{}^{(1)}, \cdots, \boldsymbol{d}_k{}^{(1)})\} + E\{d^2(Y_2|\mathbf{0}, \cdots, \mathbf{0})\} \end{aligned}$$

となる (等号は$\boldsymbol{d}^{(2)} = \mathbf{0}$ のとき)。これより、$\boldsymbol{c}_i{}^*$を

$$\boldsymbol{c}_i{}^* = [A_1, A_2] \begin{bmatrix} \boldsymbol{d}_i{}^{(1)} \\ \\ \mathbf{0} \end{bmatrix} = A_1 \boldsymbol{d}_i{}^{(1)} = A_1 A_1' \boldsymbol{c}_i$$

と定義すると、

$$E\{d^2(X|\boldsymbol{c}_1^*, \cdots, \boldsymbol{c}_k{}^*)\} \leq E\{d^2(X|\boldsymbol{c}_1, \cdots, \boldsymbol{c}_k)\}$$

が成り立つ (等号は $A_1 A_1' \boldsymbol{c}_i = \boldsymbol{c}_i$ $(i = 1, \cdots, k)$ のとき)。なぜなら、

$$[\boldsymbol{c}_1^*, \cdots, \boldsymbol{c}_k{}^*] = A_1 A_1' [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_k]$$

かつ $\mathrm{rank}(A_1) = m \leq k - 1$ であるからである。(証明終)

この定理は、定理 6 の step 1 をより一般化したものであるが、$k > 2$ の場合における step 2 及び step 3 の一般化はまだ成功していない。

2.4　2変量正規分布の場合

本節では、2変量正規分布において分散共分散行列を変化させたときのk-Principal Pointsの配置の変化について、計算機シミュレーションにより得られている結果を示す。また、本研究で利用している、Principal Pointsの導出アルゴリズムについても述べる。

2.4.1　分散共分散行列とk個の主要点との関係

2次元のPrincipal Pointsは、座標系の平行移動と回転、相似変換に関して不変なので、一般性を失わずに2変量正規分布の平均ベクトルを$\mathbf{0}$、分散共分散行列を$\mathrm{diag}(\sigma^2, 1) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix}$とおくことができる。Flury[21]には、$\sigma = 1$、$\sigma = 1.5$、$\sigma = 3$の場合における$k$-Principal Pointsが、図2.6のような配置となることが示されている。ただし、$\sigma = 1.5$における$k = 5$の場合については、Flury[21]に誤りがあったため、正確な配置を計算して示した。また、$\sigma = 3$における$k = 4$及び$k = 5$の場合については、Flury[21]では求めていなかったため、新たに計算した。

図2.6より、$k = 3$におけるPrincipal Pointsは、σが小さい場合には(a)、(b)のように三角形を形成するが、σが大きくなると(c)のように一直線上に並ぶことがわかる。しかし、3-Principal Pointsの形が三角形から直線に変わるσの境界値は求められていない。

以上の結果より、Flury[21]は次のような問題を提示した。

問題 (a).

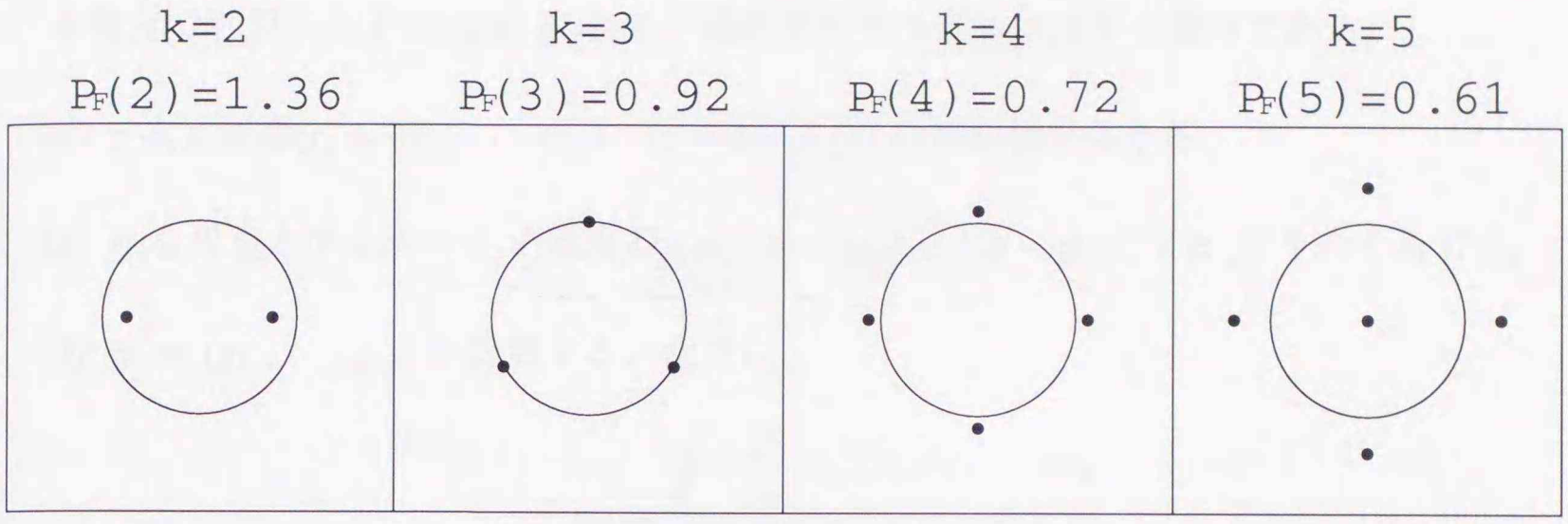
$k > 2$のとき、$\sigma_0(k)$の値はいくらになるか?

(ただし、$\sigma_0(k)$は$\sigma \geq \sigma_0(k)$なるσに関し2変量正規分布$N_2\{\mathbf{0}, \mathrm{diag}(\sigma^2, 1)\}$のPrincipal Pointsの第2座標がいずれも0となるような1より大きい数)

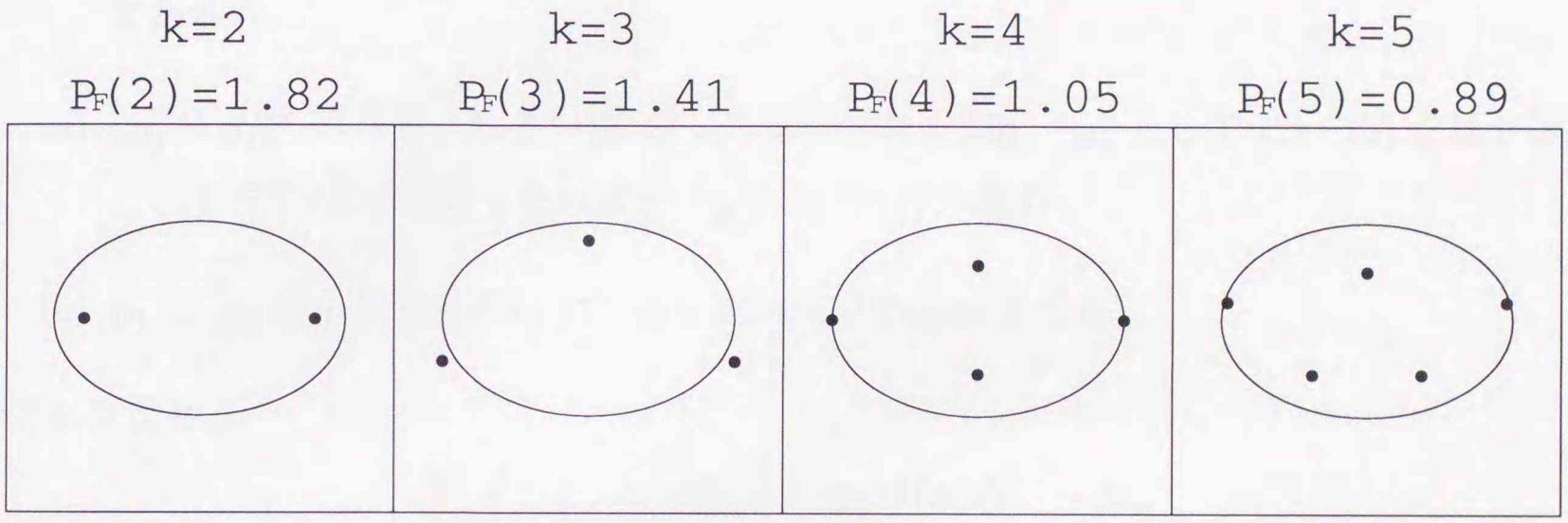
問題 (b).

$\sigma > 1$のとき、$N_2\{\mathbf{0}, \mathrm{diag}(\sigma^2, 1)\}$の$k$-Principal Pointsの第2座標がいずれも0であれば、kの最大値はいくらになるか?

(a)$\sigma = 1$



(b)$\sigma = 1.5$



(c)$\sigma = 3$

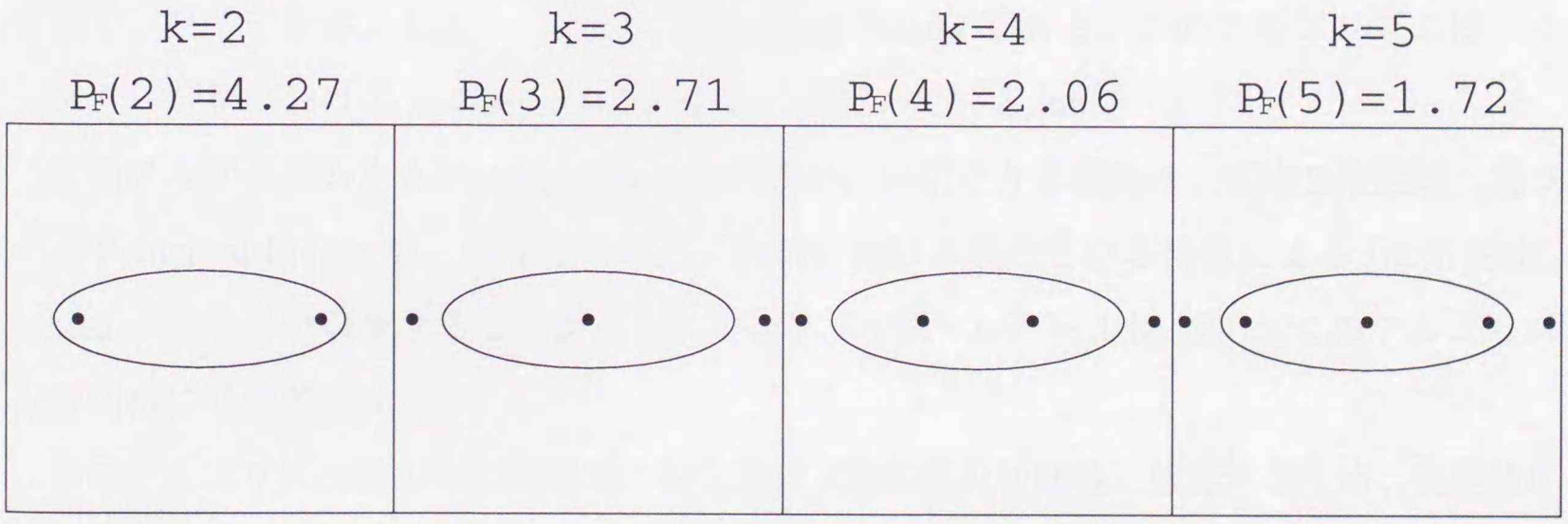


図 2.6: k個の主要点の配置図 $(k = 2, 3, 4, 5)$ (Flury[21])

2.4.2 導出アルゴリズム

本研究で利用したPrincipal Pointsの導出アルゴリズムは以下の通りである。

(1) p次元座標$\boldsymbol{y}_j = (y_{1j}, \cdots, y_{pj})$ $(j = 1, \cdots, k)$の初期値を与える。

(2) $\boldsymbol{y}_j$を母点とするボロノイ領域$D_j\{\boldsymbol{x}|\ ||\boldsymbol{x} - \boldsymbol{y}_j|| < ||\boldsymbol{x} - \boldsymbol{y}_i||\ \ ^\forall i \neq j\}$をつくる[71]。

(3) $\boldsymbol{g}_j = (g_{1j}, \cdots, g_{pj})$を計算する。ただし、

$$g_{ij} = \frac{\int \cdots \int_{D_j} x_i f(\boldsymbol{x}) dx_1 \cdots dx_p}{\int \cdots \int_{D_j} f(\boldsymbol{x}) dx_1 \cdots dx_p}$$

である。

(4) $\|\boldsymbol{g}_j - \boldsymbol{y}_j\|^2$のうち、しきい値$\varepsilon$以上のものがある間$\boldsymbol{g}_j \rightarrow \boldsymbol{y}_j$として(2)～(3)を繰り返し、すべて$\varepsilon$より小さくなれば$\boldsymbol{g}_j \rightarrow \boldsymbol{y}_j$として(5)へ進む。

(5) $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k$を$p$変量分布における$k$-Principal Pointsとする。

これにより、

$$\sum_{j=1}^{k} \int \cdots \int_{D_j} d^2(\boldsymbol{x}|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k) f(\boldsymbol{x}) dx_1 \cdots dx_p$$

の極小値を求めることができる。特に、求まった極小値が最小値となれば、その値は$P_F(k)$であり、$P_F(k)$を与える$\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_k$は$k$-Principal Pointsである。このアルゴリズムは、クラスター分析におけるk-means法の考え方に基づいている[68]。

このアルゴリズムをk-Principal Pointsの導出に利用できる根拠は、平均2乗距離に基づくk-Principal Pointsが、各々のボロノイ領域における重心となる性質による(水田[83])。k-means法は必ず収束することが知られており(大隅・ルバール他[69])、このアルゴリズムも同様に必ず停止する。

このアルゴリズムによる目的関数の収束値はk個の点の初期値に依存するため、初期値によっては収束値が最小値にならない場合も起こり得る。そのため、様々な初期値によるシミュレーションを行い、得られる収束値のうち最小となる場合におけるk個の点をk-Principal Pointsとする。

2.5　天気図の解析

本節では、Principal Points の概念を導入したデータ解析の例として、天気図の解析 (村木・大瀧・水田 [85][86]) についての研究を紹介する。

2.5.1　主要点の概念の導入の意義

この研究においては、1993 年から 1995 年の 7 月 1 日から 8 月 31 日までの毎日 21 時 (日本標準時) における極東地域 (日本付近) の地上および 500hPa 高層天気図 (各 186 パターン) を解析対象としている。図 2.7は解析対象とした天気図の例であり、各図中における太線内が解析範囲である。

1993 年から 1995 年にかけて、日本の夏期の天気は、それぞれ極端な特徴を有した。1993 年は全国的に記録的冷夏、翌年の 1994 年は全国各地で猛暑となり、そして 1995 年は北日本では冷夏、西日本では猛暑という地域格差が大きい年となった。これらの各年の夏期の天気図の特徴は、従来行われてきたように 7 月または 8 月の 1ヶ月平均天気図を求めて比較することで大まかには判断可能であるかも知れないが、日々刻々変化する天気図の特性から、平均天気図のみでは各夏の天気図の特徴を十分には捉え切れないおそれがある。そこ

(a) 地上天気図 [朝日新聞・中国新聞]　　(b)500hPa 高層天気図 [気象台資料]

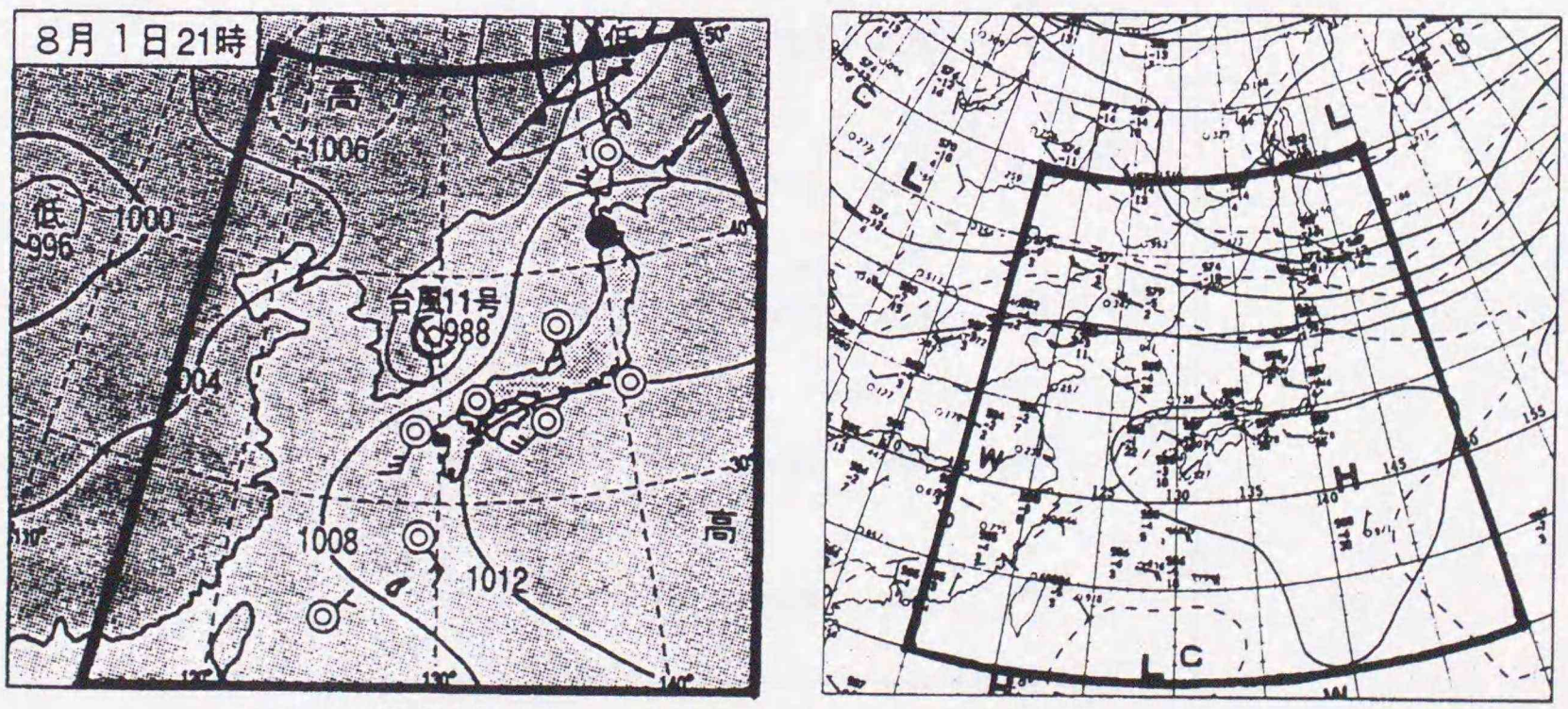


図 2.7: 解析対象とした資料天気図 (村木・大瀧・水田 [86])

で、このような各年において極端な特徴をもった天気図データを Principal Points の概念が導入された解析法により、複数の代表的な天気図パターンで要約することで、より詳細な各夏の天気図の特徴の記述を図る。

2.5.2 天気図パターンの数量化および主成分分析によるパターンの縮約

解析に Principal Points の概念を導入する前に、村木・大瀧・水田 [85][86] は以下の作業を行っている。

まず、天気図パターンの数量化を行うための評価地点を、図 2.8に示すように、地上天気図においては 40 箇所、高層天気図においては 42 箇所定め、それらの地点における海面更正済み気圧および 500hPa 高度を各天気図より読みとった。そして、気圧配置のパターンの特徴を記述する上で、日々の平均気圧値や年毎の平均気圧値の変化の影響を排除するため、各天気図の平均気圧が 0 となるような 40 次元ベクトル (高層天気図においては 42 次元ベクトル) に置き換えた。さらに、これらの 40 次元 (あるいは 42 次元) ベクトルからなる天気図パターンデータに対し、次元の縮小化およびベクトル成分の直交化を図るために

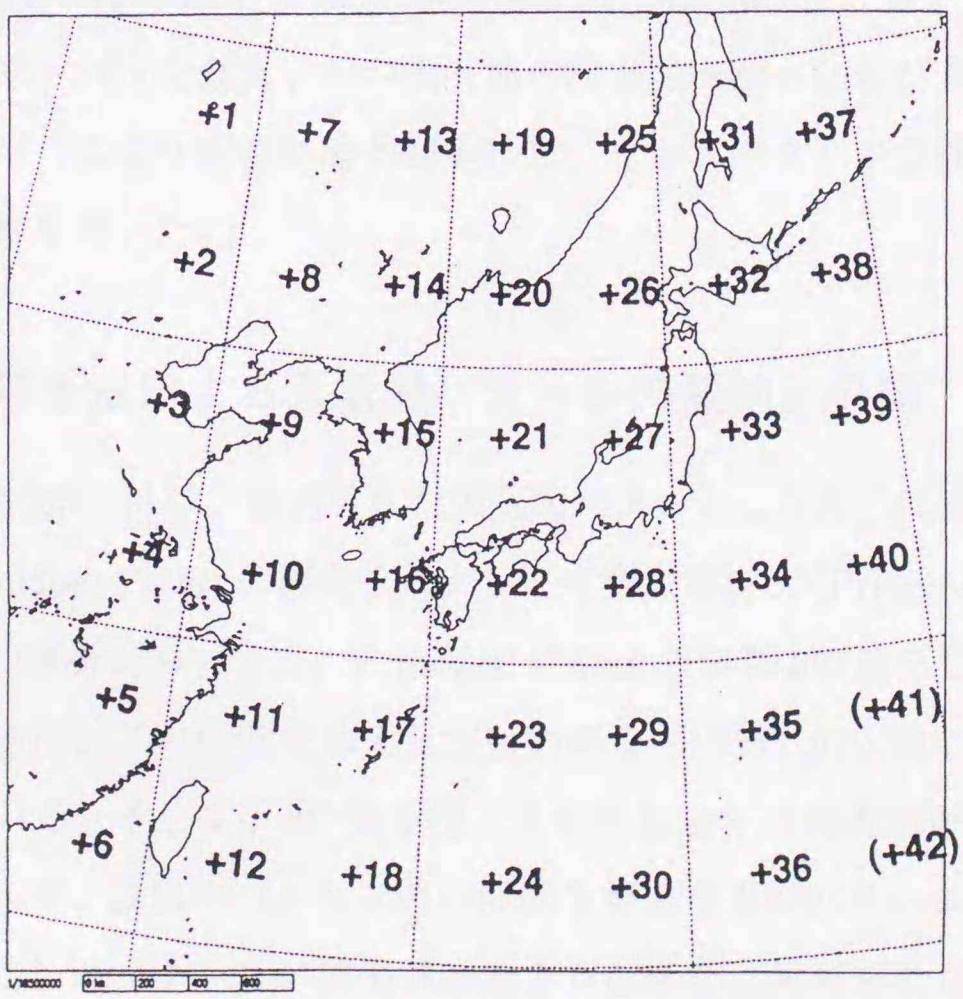


図 2.8: 天気図パターンの数量化評価地点 (村木・大瀧・水田 [86])

表 2.3: 主要主成分に関する固有値と寄与率 (村木・大瀧・水田 [86])

	(a) 地上天気図			(b)500hPa 高層天気図		
主成分	固有値	寄与率 (%)	累積寄与率 (%)	固有値	寄与率 (%)	累積寄与率 (%)
1st	171.3	31.4	31.4	213.6	36.5	36.5
2nd	96.7	17.7	49.1	112.6	19.3	55.8
3rd	56.5	10.3	59.4	79.0	13.5	69.3
4th	44.6	8.2	67.6	42.8	7.3	76.6
5th	27.2	5.0	72.6	33.2	5.7	82.3
6th	21.4	3.9	76.5	18.5	3.2	85.5
7th	18.7	3.4	79.9	17.1	2.9	88.4
8th	14.2	2.6	82.5	8.7	1.5	89.9
9th	12.7	2.3	84.8	8.0	1.4	91.3
10th	9.9	1.8	86.6	6.7	1.1	92.4

主成分分析を行った。その結果、表 2.3および図 2.9、図 2.10より、地上・高層各天気図は、それぞれ累積寄与率が 80%を超え、かつ固有値の得点分布が 0 に近づき始めている第 8 主成分までの主成分得点により要約可能と判定した。なお、ベクトル空間における距離としてユークリッド距離を用いた。

2.5.3 主要点解析法による天気図パターンの要約と分類

村木・大瀧・水田 [85][86] は、縮約された天気図パターンのうち、代表的なパターンを探索するために、Principal Points の概念を導入した主要点解析法 (Principal Points Analysis) を用いた。この種の解析においては、Principal Points の個数 kの値をどのように決めるかが大きな問題となるが、この研究では kを 1(いわゆる平均値) から順に大きくしていく方法により解析している。そして、kの値を増加させても全体の代表的なパターンの内容に大きな変化が見られず、既出のパターンの中間的な要素をもつパターンや、ある特定のパターンをより詳細にしたパターンが抽出されたような場合に解析を打ち切り、最後に解析を行った kの値より 1 つ少ない個数の代表的なパターンを解としている。この解析におい

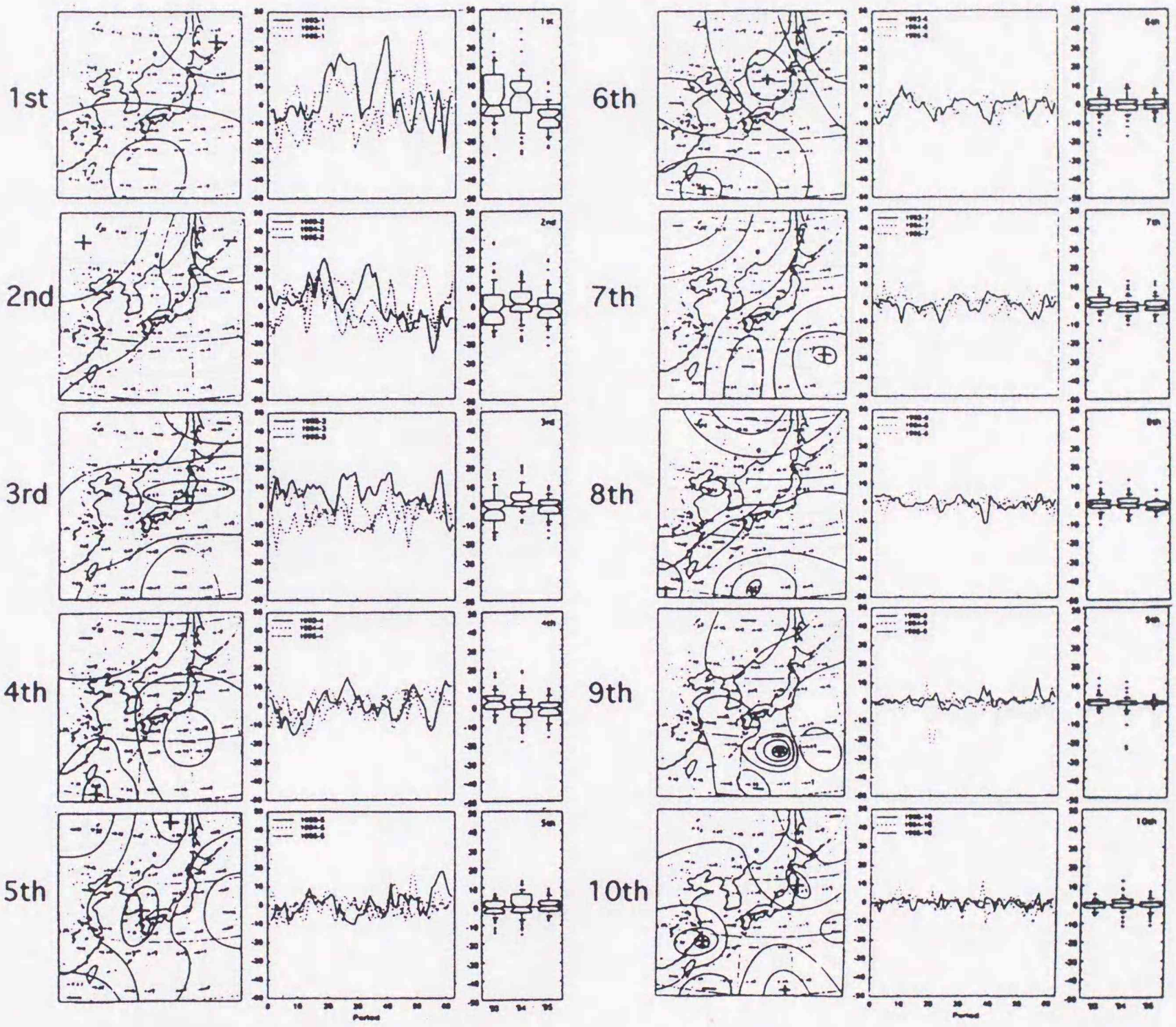


図 2.9: 地上天気図の主要主成分ベクトルのパターンおよび主成分得点の日々の動向と年別分布 (村木・大瀧・水田 [86])

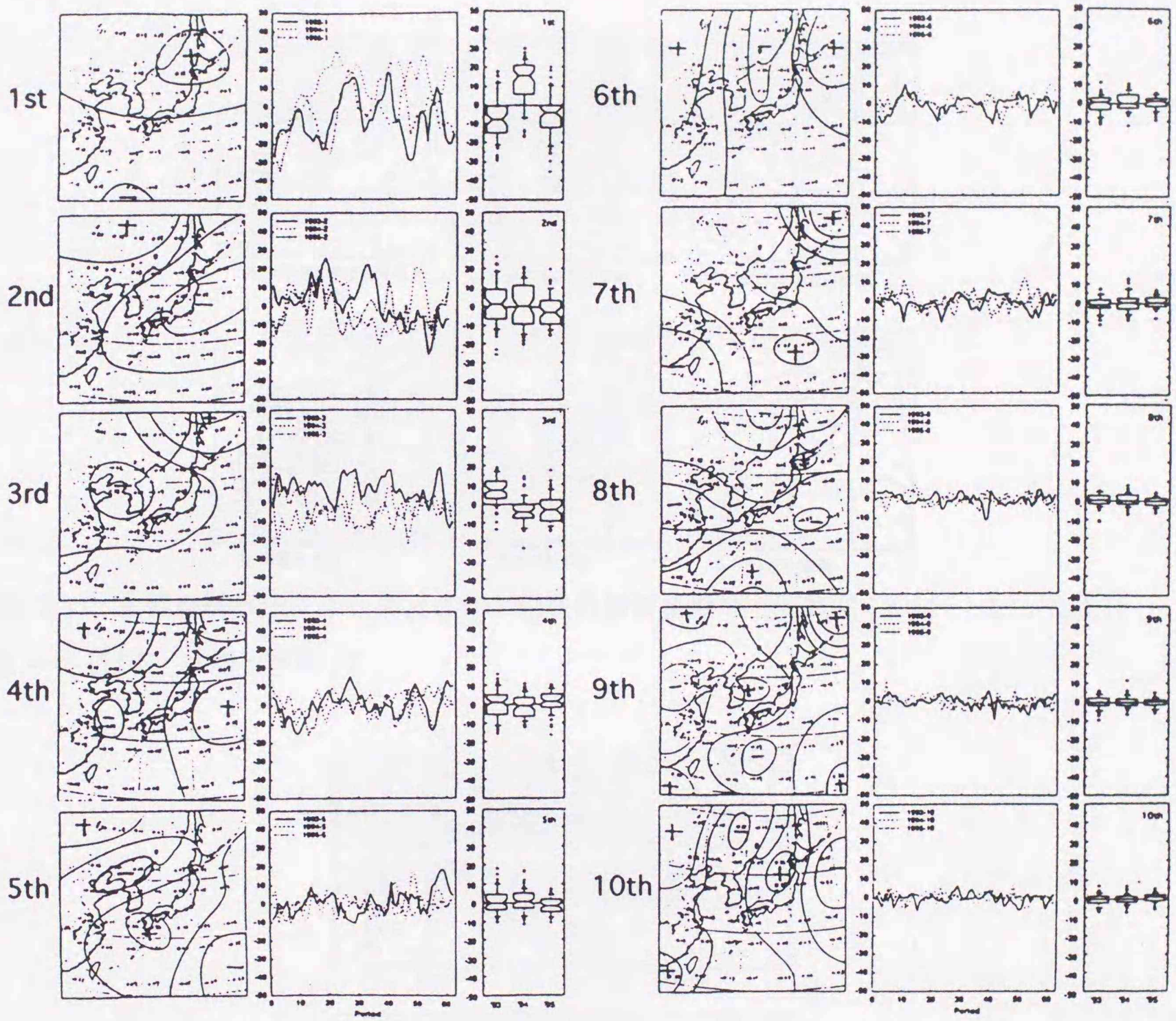


図 2.10: 500hPa 高層天気図の主要主成分ベクトルのパターンおよび主成分得点の日々の動向と年別分布 (村木・大瀧・水田 [86])

ては、地上天気図を図 2.11の A.~E. における 5 種類、高層天気図を図 2.12における 4 種類の代表的なパターンでそれぞれ特徴づけている。これにより各天気図をより詳細に分類、記述し、さらに台風の接近時など稀にしか出現しない特異な天気図パターンの抽出も行っている。

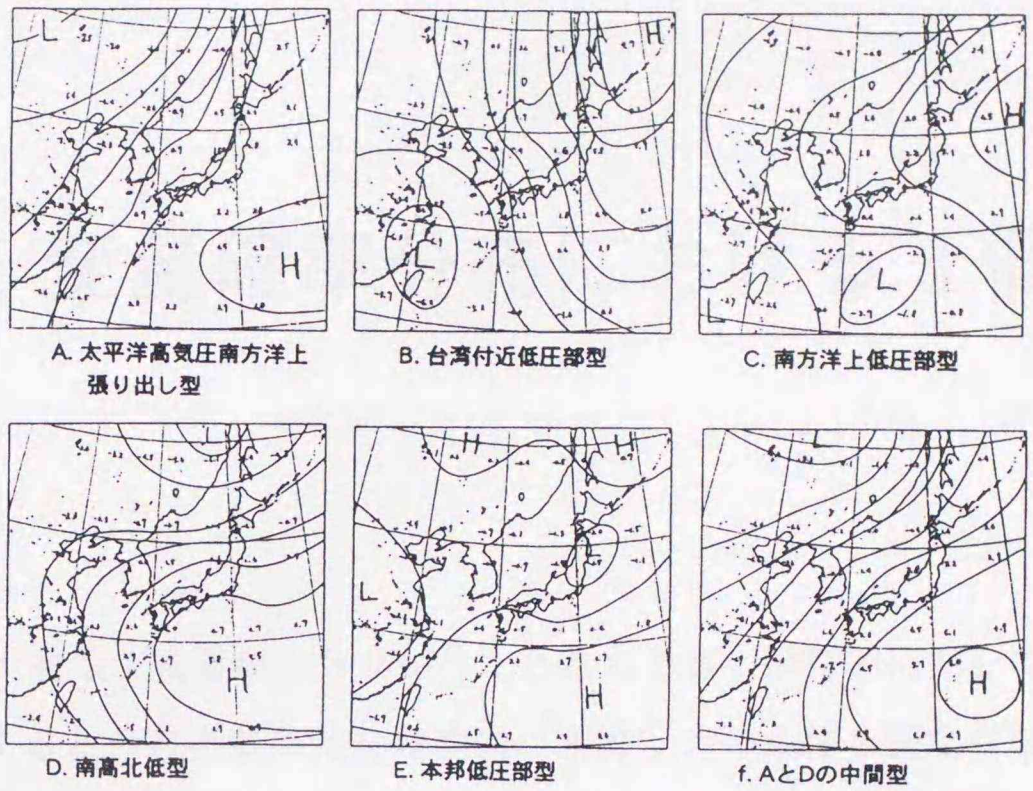


図 2.11: 主要点解析法により抽出された代表的極東夏期地上天気図 [A.~E. ; $k = 5$, f. ; $k = 6$] (村木・大瀧・水田 [86])

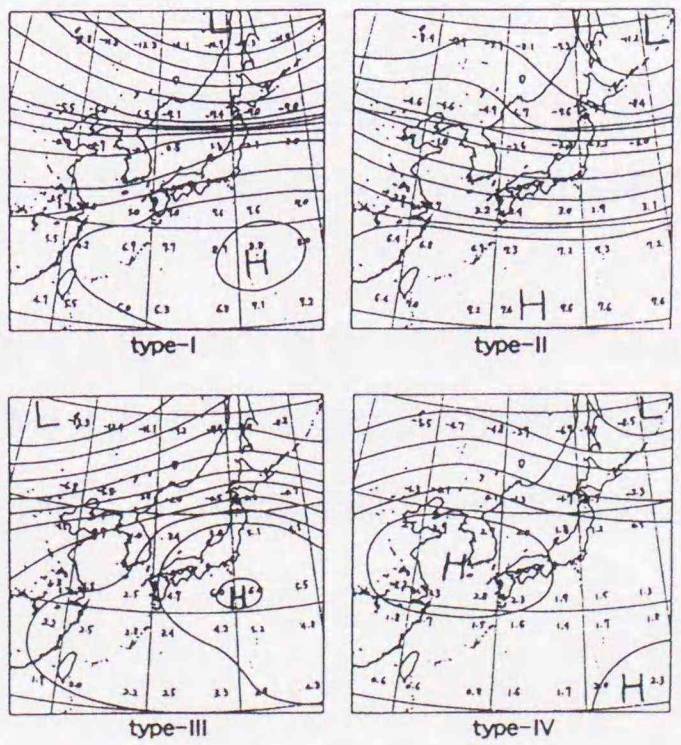


図 2.12: 主要点解析法により抽出された代表的極東夏期 500hPa 高層天気図 [(I)~(IV) ; $k = 4$] (村木・大瀧・水田 [86])

第 3 章

対称な 1 変量確率分布における 3 個の主要点

Flury[21] は、対称な 1 変量確率分布が与えられたとき、期待値に関して対称な 2 点が目的関数を極小にするときの必要条件を理論的に求め、その条件を満たさない場合において非対称な 2-Principal Points が存在する数値例を示した。しかし、3-Principal Points の値については、正規分布が与えられた場合に繰り返し計算によって得られた結果が、期待値に関して対称な値であったことが知られたにとどまっている。

本章は以下の 2 節からなる。

第 3.1節では、対称性を有する 1 変量分布が与えられた場合において、3-Principal Points に関する理論的考察を行い、期待値に関して対称な 3 点が目的関数の極小値を与える必要条件を導出する。

第 3.2節では、対称性を有する種々の 1 変量分布について得られる 3-Principal Points が第 3.1節で考察した条件を満たすかどうかを数学的に計算し、条件を満たさない場合においては計算機シミュレーションにより 3-Principal Points の値を求める。

3.1 理論的考察

本節では、正規分布が与えられた場合における 3-Principal Points について、期待値に関して対称という仮定がある場合とない場合それぞれについて数学的に考察を行い、仮定がある場合において求められた 3-Principal Points が Flury[21] において示された結果に一致することを述べる。また、対称な 1 変量分布が与えられた場合における 3-Principal Points に関連し、期待値に関して対称な 3 点が目的関数の極小値を与える必要条件を導出する。

3.1.1 正規分布の場合

1 変量正規分布においては、密度関数 $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$、分布関数 $F(x) = \int_{-\infty}^{x} f(t)dt$ である。ここで、$y_1 < y_2 < y_3$とすると、目的関数 $M(y_1, y_2, y_3)$ は

$$\begin{aligned} M(y_1, y_2, y_3) &= \int_{-\infty}^{\frac{y_1+y_2}{2}} (x - y_1)^2 f(x)dx + \int_{\frac{y_1+y_2}{2}}^{\frac{y_2+y_3}{2}} (x - y_2)^2 f(x)dx + \int_{\frac{y_2+y_3}{2}}^{\infty} (x - y_3)^2 f(x)dx \\ &= \sigma^2 + (y_1^2 - y_2^2) F\left(\frac{y_1 + y_2}{2}\right) + (y_2^2 - y_3^2) F\left(\frac{y_2 + y_3}{2}\right) + y_3^2 \\ &\quad + \frac{2\sigma}{\sqrt{2\pi}} \left\{ (y_1 - y_2) e^{-\frac{(y_1+y_2)^2}{8\sigma^2}} + (y_2 - y_3) e^{-\frac{(y_2+y_3)^2}{8\sigma^2}} \right\} \end{aligned} \tag{3.1.1}$$

である。Mを y_1, y_2, y_3でそれぞれ偏微分すると、

$$\begin{aligned} \frac{\partial M}{\partial y_1} &= 2y_1 F\left(\frac{y_1 + y_2}{2}\right) + \frac{y_1^2 - y_2^2}{2} f\left(\frac{y_1 + y_2}{2}\right) + \frac{2\sigma}{\sqrt{2\pi}} (1 - \frac{y_1^2 - y_2^2}{4\sigma^2}) e^{-\frac{(y_1+y_2)^2}{8\sigma^2}} \\ &= 2y_1 F\left(\frac{y_1 + y_2}{2}\right) + \frac{2\sigma}{\sqrt{2\pi}} e^{-\frac{(y_1+y_2)^2}{8\sigma^2}} \end{aligned} \tag{3.1.2}$$

$$\frac{\partial M}{\partial y_2} = 2y_2 \left\{ F\left(\frac{y_2 + y_3}{2}\right) - F\left(\frac{y_1 + y_2}{2}\right) \right\} + \frac{2\sigma}{\sqrt{2\pi}} \left\{ e^{-\frac{(y_2+y_3)^2}{8\sigma^2}} - e^{-\frac{(y_1+y_2)^2}{8\sigma^2}} \right\} \tag{3.1.3}$$

$$\frac{\partial M}{\partial y_3} = 2y_3 \left\{ 1 - F\left(\frac{y_2 + y_3}{2}\right) \right\} - \frac{2\sigma}{\sqrt{2\pi}} e^{-\frac{(y_2+y_3)^2}{8\sigma^2}} \tag{3.1.4}$$

となるから、(3.1.2)〜(3.1.4) がいずれも 0 であるならば、

$$(y_1 - y_2) F\left(\frac{y_1 + y_2}{2}\right) + (y_2 - y_3) F\left(\frac{y_2 + y_3}{2}\right) + y_3 = 0 \tag{3.1.5}$$

および

$$\frac{\sigma}{\sqrt{2\pi}} \left\{ \frac{y_2 - y_1}{y_1} e^{-\frac{(y_1+y_2)^2}{8\sigma^2}} + \frac{y_3 - y_2}{y_3} e^{-\frac{(y_2+y_3)^2}{8\sigma^2}} \right\} + y_2 = 0 \tag{3.1.6}$$

が成り立つ。

3.1.1.1 期待値に関する 3-Principal Points の対称性を仮定した場合

Principal Points が原点に関して対称であると仮定するならば、(3.1.5) より $y_2 = 0$, $y_1 = -y_3$ となる。ここで、

$$P(\sigma, y_3) = M(-y_3, 0, y_3) = \sigma^2 + 2y_3^2 \left\{1 - F\left(\frac{y_3}{2}\right)\right\} - \frac{4\sigma y_3}{\sqrt{2\pi}} e^{-\frac{y_3^2}{8\sigma^2}} \tag{3.1.7}$$

とすると、

$$F\left(\frac{y_3}{2}\right) = \int_{-\infty}^{\frac{y_3}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx = \int_{-\infty}^{\frac{y_3}{2\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi\left(\frac{y_3}{2\sigma}\right) \tag{3.1.8}$$

(Φ:1 変量標準正規分布関数) であることより

$$\begin{aligned}
\frac{\partial P}{\partial y_3} &= 4y_3 \left\{1 - F\left(\frac{y_3}{2}\right)\right\} - \frac{4\sigma}{\sqrt{2\pi}} e^{-\frac{y_3^2}{8\sigma^2}} \\
&= 4y_3 \left\{1 - \Phi\left(\frac{y_3}{2\sigma}\right)\right\} - \frac{4\sigma}{\sqrt{2\pi}} e^{-\frac{y_3^2}{8\sigma^2}} && (3.1.9) \\
\frac{\partial^2 P}{\partial y_3^2} &= 4\left\{1 - \Phi\left(\frac{y_3}{2\sigma}\right)\right\} - \frac{y_3}{\sqrt{2\pi}\sigma} e^{-\frac{y_3^2}{8\sigma^2}} && (3.1.10) \\
\frac{\partial^3 P}{\partial y_3^3} &= \frac{y_3^2 - 12\sigma^2}{4\sqrt{2\pi}\sigma^3} e^{-\frac{y_3^2}{8\sigma^2}} && (3.1.11)
\end{aligned}$$

であるから、$y_3 > 0$ かつ

$$\begin{cases}
\dfrac{\partial^3 P}{\partial y_3^3} < 0 & (0 < y_3 < 2\sqrt{3}\sigma) \\
\dfrac{\partial^3 P}{\partial y_3^3} > 0 & (y_3 > 2\sqrt{3}\sigma)
\end{cases}$$

より $\dfrac{\partial^2 P}{\partial y_3^2}$ は $y_3 = 2\sqrt{3}\sigma$ で極小値をとる。ここで、

$$\begin{aligned}
\lim_{y_3 \to +0} \frac{\partial^2 P}{\partial y_3^2} &= 2 > 0 \\
\left.\frac{\partial^2 P}{\partial y_3^2}\right|_{y_3 = 2\sqrt{3}\sigma} &= 4\{1 - \Phi(\sqrt{3})\} - \frac{2\sqrt{3}}{\sqrt{2\pi}} e^{-\frac{3}{2}} < 0 \\
\lim_{y_3 \to \infty} \frac{\partial^2 P}{\partial y_3^2} &= 0
\end{aligned}$$

であるから、

$$\left.\frac{\partial^2 P}{\partial y_3^2}\right|_{y_3 = z\sigma} = 4\left\{1 - \Phi\left(\frac{z}{2\sigma}\right)\right\} - \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{8}} = 0 \quad (0 < z < 2\sqrt{3}) \tag{3.1.12}$$

をみたす zがただ1つ存在する。そして$\dfrac{\partial^2 P}{\partial y_3^2} < 0 \quad (y_3 > 2\sqrt{3}\sigma)$ であるから$\dfrac{\partial P}{\partial y_3}$は $y_3 = z\sigma$ で極大値をとる。(3.1.12) を zについてS言語で計算すると、$z \simeq 2.3812$ であるから

$$\begin{aligned}
\lim_{y_3 \to +0} \frac{\partial P}{\partial y_3} &= -\frac{4\sigma}{\sqrt{2\pi}} < 0 \\
\left.\frac{\partial P}{\partial y_3}\right|_{y_3 = z\sigma} &= \frac{(z^2-4)\sigma}{\sqrt{2\pi}} e^{-\frac{z^2}{8}} > 0 \\
\lim_{y_3 \to \infty} \frac{\partial P}{\partial y_3} &= 0
\end{aligned}$$

となり、

$$\left.\frac{\partial P}{\partial y_3}\right|_{y_3 = \xi\sigma} = 4\xi\sigma \left\{1 - \Phi\left(\frac{\xi}{2\sigma}\right)\right\} - \frac{4\sigma}{\sqrt{2\pi}} e^{-\frac{\xi^2}{8}} = 0 \quad (0 < \xi < z) \tag{3.1.13}$$

をみたすξがただ1つ存在する。そして$\dfrac{\partial P}{\partial y_3} > 0 \ \ (y_3 > z\sigma)$ であるから、$P(\sigma, y_3)$ は $y_3 = \xi\sigma$ で最小値

$$P(\sigma, \xi\sigma) = \sigma^2 - \frac{2\sigma^2\xi}{\sqrt{2\pi}} e^{-\frac{\xi^2}{8}} = \sigma^2 \left(1 - \frac{2\xi}{\sqrt{2\pi}} e^{-\frac{\xi^2}{8}}\right) \tag{3.1.14}$$

をとる。(3.1.13) をξについてS言語で計算すると、$\xi \simeq 1.224$ となり、3-Principal Points は $0, \pm 1.224\sigma$である。これより、

$$P(\sigma, 1.224\sigma) \simeq 0.19\sigma^2 \tag{3.1.15}$$

が目的関数の最小値となる。

3.1.1.2　期待値に関する 3-Principal Points の対称性を仮定しない場合

Principal Pointsが原点に関して対称であると仮定しないとき、(3.1.2)～(3.1.4) をいずれも0にする y_1、y_2、y_3を

$$(y_1, y_2, y_3) = (\xi_1\sigma, \xi_2\sigma, \xi_3\sigma) \quad (\xi_1 < \xi_2 < \xi_3) \tag{3.1.16}$$

とおくと、(3.1.2)～(3.1.4) 及び (3.1.6) より

$$F\left(\frac{\xi_1\sigma + \xi_2\sigma}{2}\right) = -\frac{1}{\sqrt{2\pi}\xi_1} e^{-\frac{(\xi_1+\xi_2)^2}{8}} \tag{3.1.17}$$

$$F\left(\frac{\xi_2\sigma + \xi_3\sigma}{2}\right) = 1 - \frac{1}{\sqrt{2\pi}\xi_3} e^{-\frac{(\xi_2+\xi_3)^2}{8}} \tag{3.1.18}$$

$$-\frac{\xi_2 - \xi_1}{\sqrt{2\pi}\xi_1} e^{-\frac{(\xi_1+\xi_2)^2}{8}} = \frac{\xi_3 - \xi_2}{\sqrt{2\pi}\xi_3} e^{-\frac{(\xi_2+\xi_3)^2}{8}} + \xi_2 \tag{3.1.19}$$

が成り立つ。ここで、2 階偏微分行列は

$$D(y_1, y_2, y_3) = \begin{bmatrix} \dfrac{\partial^2 M}{\partial y_1^2} & \dfrac{\partial^2 M}{\partial y_1 \partial y_2} & \dfrac{\partial^2 M}{\partial y_1 \partial y_3} \\ \dfrac{\partial^2 M}{\partial y_2 \partial y_1} & \dfrac{\partial^2 M}{\partial y_2^2} & \dfrac{\partial^2 M}{\partial y_2 \partial y_3} \\ \dfrac{\partial^2 M}{\partial y_3 \partial y_1} & \dfrac{\partial^2 M}{\partial y_3 \partial y_2} & \dfrac{\partial^2 M}{\partial y_3^2} \end{bmatrix} \tag{3.1.20}$$

となる。ただし、

$$\frac{\partial^2 M}{\partial y_1^2} = 2F\left(\frac{y_1+y_2}{2}\right) + \frac{y_1-y_2}{2} f\left(\frac{y_1+y_2}{2}\right) \tag{3.1.21}$$

$$\frac{\partial^2 M}{\partial y_1 \partial y_2} = \frac{\partial^2 M}{\partial y_2 \partial y_1} = \frac{y_1-y_2}{2} f\left(\frac{y_1+y_2}{2}\right) \tag{3.1.22}$$

$$\frac{\partial^2 M}{\partial y_1 \partial y_3} = \frac{\partial^2 M}{\partial y_3 \partial y_1} = 0 \tag{3.1.23}$$

$$\begin{aligned} \frac{\partial^2 M}{\partial y_2^2} &= 2F\left(\frac{y_2+y_3}{2}\right) + \frac{y_2-y_3}{2} f\left(\frac{y_2+y_3}{2}\right) \\ &\quad -2F\left(\frac{y_1+y_2}{2}\right) + \frac{y_1-y_2}{2} f\left(\frac{y_1+y_2}{2}\right) \end{aligned} \tag{3.1.24}$$

$$\frac{\partial^2 M}{\partial y_2 \partial y_3} = \frac{\partial^2 M}{\partial y_3 \partial y_2} = \frac{y_2-y_3}{2} f\left(\frac{y_2+y_3}{2}\right) \tag{3.1.25}$$

$$\frac{\partial^2 M}{\partial y_3^2} = 2 - 2F\left(\frac{y_2+y_3}{2}\right) + \frac{y_2-y_3}{2} f\left(\frac{y_2+y_3}{2}\right) \tag{3.1.26}$$

であり、さらに (3.1.16) を (3.1.20)～(3.1.26) に代入すると、(3.1.17)、(3.1.18) より

$$D(\xi_1\sigma, \xi_2\sigma, \xi_3\sigma) =$$

$$\begin{bmatrix} \left(-\dfrac{2}{\xi_1} - \dfrac{\xi_2-\xi_1}{2}\right)\phi\left(\dfrac{\xi_1+\xi_2}{2}\right) & -\dfrac{\xi_2-\xi_1}{2}\phi\left(\dfrac{\xi_1+\xi_2}{2}\right) & 0 \\ -\dfrac{\xi_2-\xi_1}{2}\phi\left(\dfrac{\xi_1+\xi_2}{2}\right) & 2 + \left(-\dfrac{2}{\xi_3} - \dfrac{\xi_3-\xi_2}{2}\right)\phi\left(\dfrac{\xi_2+\xi_3}{2}\right) & -\dfrac{\xi_3-\xi_2}{2}\phi\left(\dfrac{\xi_2+\xi_3}{2}\right) \\ & + \left(\dfrac{2}{\xi_1} - \dfrac{\xi_2-\xi_1}{2}\right)\phi\left(\dfrac{\xi_1+\xi_2}{2}\right) & \\ 0 & -\dfrac{\xi_3-\xi_2}{2}\phi\left(\dfrac{\xi_2+\xi_3}{2}\right) & \left(\dfrac{2}{\xi_3} - \dfrac{\xi_3-\xi_2}{2}\right)\phi\left(\dfrac{\xi_2+\xi_3}{2}\right) \end{bmatrix} \tag{3.1.27}$$

となる (Φ：1 変量標準正規分布関数、ϕ：Φの密度関数)。この行列が正定値である必要十分条件は

$$\left.\frac{\partial^2 M}{\partial y_1^2}\right|_{y_1=\xi_1\sigma} > 0 \tag{3.1.28}$$

$$\det D_{13}(\xi_1\sigma, \xi_3\sigma) > 0 \tag{3.1.29}$$

$$\det D(\xi_1\sigma, \xi_2\sigma, \xi_3\sigma) > 0 \tag{3.1.30}$$

が同時に成り立つことである。ただし、

$$D_{13}(y_1, y_3) = \begin{bmatrix} \dfrac{\partial^2 M}{\partial y_1^2} & \dfrac{\partial^2 M}{\partial y_1 \partial y_3} \\ \dfrac{\partial^2 M}{\partial y_3 \partial y_1} & \dfrac{\partial^2 M}{\partial y_3^2} \end{bmatrix} \tag{3.1.31}$$

である。これより、(3.1.28)、(3.1.29) はそれぞれ

$$\xi_1 < 0, \xi_1 < \xi_2 < \xi_1 - \frac{4}{\xi_1} \tag{3.1.32}$$

$$\xi_3 > 0, \xi_3 > \xi_2 > \xi_3 - \frac{4}{\xi_3} \tag{3.1.33}$$

と同値であり、このとき Mは極小値

$$\begin{aligned} M(\xi_1\sigma, \xi_2\sigma, \xi_3\sigma) &= \sigma^2 - \frac{\sigma^2(\xi_1^2 - \xi_2^2)}{\sqrt{2\pi}\xi_1} e^{-\frac{(\xi_1+\xi_2)^2}{8}} - \frac{\sigma^2(\xi_2^2 - \xi_3^2)}{\sqrt{2\pi}\xi_3} e^{-\frac{(\xi_2+\xi_3)^2}{8}} + \sigma^2\xi_2^2 \\ &\quad + \frac{2\sigma^2}{\sqrt{2\pi}} \left\{ (\xi_1 - \xi_2) e^{-\frac{(\xi_1+\xi_2)^2}{8}} + (\xi_2 - \xi_3) e^{-\frac{(\xi_2+\xi_3)^2}{8}} \right\} \\ &= \sigma^2 \left\{ 1 + \xi_2^2 + \frac{(\xi_2 - \xi_1)^2}{\sqrt{2\pi}\xi_1} e^{-\frac{(\xi_1+\xi_2)^2}{8}} - \frac{(\xi_3 - \xi_2)^2}{\sqrt{2\pi}\xi_3} e^{-\frac{(\xi_2+\xi_3)^2}{8}} \right\} \\ &= \sigma^2 \left\{ 1 + \xi_1\xi_2 - \frac{(\xi_3 - \xi_2)(\xi_3 - \xi_1)}{\sqrt{2\pi}\xi_3} e^{-\frac{(\xi_2+\xi_3)^2}{8}} \right\} \end{aligned} \tag{3.1.34}$$

をとる。特に、$(\xi_1, \xi_2, \xi_3) = (-\xi, 0, \xi)$ $(\xi \simeq 1.224)$ であれば、ξ_1, ξ_2, ξ_3は (3.1.30)～(3.1.33) を満たし、極小値は

$$M = \sigma^2 \left(1 - \frac{2\xi}{\sqrt{2\pi}} e^{-\frac{\xi^2}{8}} \right) \tag{3.1.35}$$

となる。これは、対称性を仮定した場合の結果に等しい。しかし、この値が最小であるかどうかを目的関数から理論的に調べることは困難である。

3.1.2 一般的な場合

以下では、一般性を失うことなく原点に関して対称な 1 変量分布における 3-Principal Points について詳しく考察する。

密度関数を $f(x)$、分布関数を $F(x)$ とし、$f(x)$ が絶対連続かつ常に正で、分散σ^2が有限であると仮定する。ここで、$f(x)$ は原点に関して対称だから $f(x) = f(-x)$ となる。

$M(y_1, y_2, y_3) = E(d^2(X \mid y_1, y_2, y_3))$ を考察する代わりに、

$$c_1 = \frac{y_1 + y_2}{2}, \quad c_2 = \frac{y_2 + y_3}{2}$$

$$h_1 = \frac{y_2 - y_1}{2}, \quad h_2 = \frac{y_3 - y_2}{2}$$

とおくと、

$$\begin{aligned}
M(y_1, y_2, y_3) &= E(d^2(X \mid y_1, y_2, y_3)) \\
&= \int_{-\infty}^{\infty} \min_{1 \le i \le 3} (x - y_i)^2 f(x) dx \\
&= \int_{-\infty}^{c_1} (x - c_1 + h_1)^2 f(x) dx + \int_{c_1}^{c_2} (x - c_1 - h_1)^2 f(x) dx + \int_{c_2}^{\infty} (x - c_2 - h_2)^2 f(x) dx \\
&= \sigma^2 - 4c_1 h_1 F(c_1) - 4c_2 h_2 F(c_2) + (c_2 + h_2)^2 + 4h_1 \int_{-\infty}^{c_1} x f(x) dx + 4h_2 \int_{-\infty}^{c_2} x f(x) dx \\
&= \sigma^2 + 2h_1 (G(c_1) - G(c_2)) + 2(c_2 - c_1) G(c_2) + (c_2 - c_1)^2 + (c_2 - h_1)^2
\end{aligned}$$

である。ただし、$h_2 = c_2 - c_1 - h_1$であり、

$$G(c) = 2 \int_{-\infty}^{c} x f(x) dx + c(1 - 2F(c))$$

とする。

ここで $M(y_1, y_2, y_3) = H(c_1, c_2, h_1)$ とすると、c_1, c_2, h_1が $H(c_1, c_2, h_1)$ の極小値となる十分条件は、$\dfrac{\partial H}{\partial c_1} = 0$, $\dfrac{\partial H}{\partial c_2} = 0$, $\dfrac{\partial H}{\partial h_1} = 0$ かつ Hのヘシアンが正定値である (ヘシアンが非負定値であることは、必要条件である)。

$H(c_1, c_2, h_1)$ の偏微分は、

$$\frac{\partial H}{\partial c_1} = 2h_1 G'(c_1) - 2G(c_2) - 2(c_2 - c_1) = 0 \tag{3.1.36}$$

$$\frac{\partial H}{\partial c_2} = 2(c_2 - c_1 - h_1) G'(c_2) + 2G(c_2) + 2(2c_2 - c_1 - h_1) = 0 \tag{3.1.37}$$

$$\frac{\partial H}{\partial h_1} = 2(G(c_1) - G(c_2)) - 2(c_2 - h_1) = 0 \tag{3.1.38}$$

となるので、(3.1.38) 式より

$$h_1 = -G(c_1) + G(c_2) + c_2 \tag{3.1.39}$$

が得られる。

Hのヘシアンが正定値である条件は、

$$\frac{\partial^2 H}{\partial c_1^2} = 2 + 2h_1 G''(c_1) > 0 \tag{3.1.40}$$

$$\frac{\partial^2 H}{\partial c_1^2}\frac{\partial^2 H}{\partial h_1^2} - \left(\frac{\partial^2 H}{\partial c_1 \partial h_1}\right)^2 = 4(1 + h_1 G''(c_1)) - 4(G'(c_1))^2 > 0 \tag{3.1.41}$$

$$\det D(c_1, c_2, h_1) > 0 \tag{3.1.42}$$

が同時に成り立つことである。ただし、

$$D(c_1, c_2, h_1) = \begin{bmatrix} \frac{\partial^2 H}{\partial c_1^2} & \frac{\partial^2 H}{\partial c_1 \partial c_2} & \frac{\partial^2 H}{\partial c_1 \partial h_1} \\ \frac{\partial^2 H}{\partial c_2 \partial c_1} & \frac{\partial^2 H}{\partial c_2^2} & \frac{\partial^2 H}{\partial c_2 \partial h_1} \\ \frac{\partial^2 H}{\partial h_1 \partial c_1} & \frac{\partial^2 H}{\partial h_1 \partial c_2} & \frac{\partial^2 H}{\partial h_1^2} \end{bmatrix}$$

である。ここで、

$$G'(c) = 1 - 2F(c)$$

$$G''(c) = -2f(c)$$

であり、(3.1.40) 式は、(3.1.41) 式が成立すれば必ず成立する。

ここで、3-Principal Points の対称性を仮定すると、$y_2 = 0, y_1 = -y_3 < 0$ である。従って、$c_1 = -c_2 < 0, h_1 = c_2$より

$$H(-c_2, c_2, c_2) = \sigma^2 + 4c_2 G(c_2) + 4c_2^2 \tag{3.1.43}$$

となる。(3.1.43) 式を $K(c_2)$ として c_2で微分すると

$$\begin{aligned} K'(c_2) &= 4G(c_2) + 4c_2 G'(c_2) + 8c_2 \\ &= 8\left\{\int_{-\infty}^{c_2} x f(x)dx + 2c_2(1 - F(c_2))\right\} \\ &= -8\int_{c_2}^{\infty} (x - 2c_2) f(x)dx \end{aligned} \tag{3.1.44}$$

となる。ここで $K'(t) = 0$ なる $t > 0$ が存在すれば、(3.1.43) 式は $c_2 = t$ で最小値をとり、(3.1.41) 式、(3.1.42) 式はそれぞれ

$$4(1 - 2tf(t)) - 4(1 - 2F(t))^2 > 0 \tag{3.1.45}$$

$$\det D(-t,t,t) > 0 \tag{3.1.46}$$

となる。(3.1.45) 式、(3.1.46) 式より

$$tf(t) < 2(1-F(t))(2F(t)-1) \tag{3.1.47}$$

である。

以上より、次の定理 8 が成り立つ。

定理 8.

密度関数 $f(x)$ が期待値$\mu = E(X)$ に関して対称、2 次のモーメントが有限である連続な 1 変量の確率変数 Xにおいて、$y_1 = \mu - 2t, y_2 = \mu, y_3 = \mu + 2t$ (ただし t は $\int_{\mu+t}^{\infty}(x-\mu-2t)f(x-\mu)dx = 0$ をみたす正の数) が $E\{d^2(X \mid y_1, y_2, y_3)\}$ の極小値をもたらすための十分条件は、

$$\frac{(\mu+t)f(\mu+t)}{(1-F(\mu+t))(2F(\mu+t)-1)} < 2 \tag{3.1.48}$$

であり、その必要条件は

$$\frac{(\mu+t)f(\mu+t)}{(1-F(\mu+t))(2F(\mu+t)-1)} \leq 2 \tag{3.1.49}$$

である。ただし、分布関数を $F(x)$ とする。

連続な確率変数 Xに関しては、確率分布を空集合とならない k個の領域に分割する方法が少なくとも 1 つ存在することが Pärna[49] により証明されている。従って、1 変量確率変数 Xの密度関数 $f(x)$ が期待値$\mu = E(X)$ に関して対称で、2 次のモーメントが有限であるとき、Xが (3.1.48) 式をみたさない場合は、期待値に関して非対称な 3-Principal Points が存在する。

3.2 種々の分布における値

本節では、ロジスティック分布、両側指数分布 (ラプラス分布、二重指数分布)、混合正規分布における 3-Principal Points $(y_1 < y_2 < y_3)$ について検討した結果を述べる。ただし、一般性を失わず $y_2 \leq 0$ と仮定する。

3.2.1 ロジスティック分布

ロジスティック分布の密度関数、分布関数は、それぞれ

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}, \qquad F(x) = \frac{1}{1+e^{-x}}$$

となり、$E(X) = 0$、$\sigma^2 = \dfrac{\pi^2}{3}$である。

3-Principal Pointsが原点に関して対称であると仮定した場合、

$$K'(t) = 8\left\{\frac{t}{1+e^t} - \log(1+e^{-t})\right\} = 0$$

をみたすtの値は$t \simeq 1.1446$となる。よって、

$$\frac{tf(t)}{(1-F(t))(2F(t)-1)} \simeq 1.679 < 2$$

となり、$(y_1, y_2, y_3) = (-2t, 0, 2t)$は3-Principal Pointsの候補となる。

3-Principal Pointsが原点に関して非対称であるとした場合、

$$\begin{aligned}
\frac{\partial H}{\partial c_1} &= 2h_1(G'(c_1)-1) + 2(c_1 - G(c_1)) \\
&= \frac{4}{1+e^{-c_1}}\left\{c_1 + (e^{-c_1}-1)\log(1+e^{c_1}) + 2\log(1+e^{-c_2})\right\} = 0 \qquad (3.2.50) \\
\frac{\partial H}{\partial c_2} &= 2(c_2 - c_1 - h_1)(G'(c_2)+1) + 2(c_2 + G(c_2)) \\
&= \frac{4}{1+e^{c_2}}\left\{c_2 + (1-e^{c_2})\log(1+e^{-c_2}) - 2\log(1+e^{c_1})\right\} = 0 \qquad (3.2.51)
\end{aligned}$$

である。(3.2.50)式より、

$$\begin{aligned}
c_2 &= -\log\left[\exp\{-\frac{c_1}{2} - \frac{1}{2}(e^{-c_1}-1)\log(1+e^{c_1})\} - 1\right] \\
&= J(c_1) \qquad (3.2.52)
\end{aligned}$$

と書くことができる。これを(3.2.51)式に代入して

$$\begin{aligned}
\frac{\partial H}{\partial c_2} &= \frac{4}{1+e^{J(c_1)}}\left[-2\log(1+e^{c_1}) + J(c_1) - \frac{1}{2}\{c_1 + (e^{-c_1}-1)\log(1+e^{c_1})\}(1-e^{J(c_1)})\right] \\
&= 0 \qquad (3.2.53)
\end{aligned}$$

となるような c_1、c_2を求めるとよい。S 言語を用いると (3.2.53) 式を満たす解は $(c_1, c_2) \simeq (-1.1446, 1.1446)$ となり、原点に関する対称性を仮定した場合と一致する。

以上より、ロジスティック分布における 3-Principal Points は対称であり、$(y_1, y_2, y_3) = (-2t, 0, 2t)$ (ただし $t \simeq 1.1446$) となる。

3.2.2 両側指数分布

両側指数分布 (ラプラス分布、二重指数分布) の密度関数は $f(x) = \dfrac{\lambda}{2} e^{-\lambda|x|}$ $(\lambda > 0)$ と表される。分布関数は

$$F(x) = \begin{cases} 1 - \dfrac{1}{2} e^{-\lambda x}, & (x \geq 0) \\ \dfrac{1}{2} e^{\lambda x}, & (x < 0) \end{cases}$$

となり、$E(X) = 0, \sigma^2 = \dfrac{2}{\lambda}$である。以下では、一般性を失わずに$\lambda = 1$とする。

3-Principal Points が原点に関して対称であると仮定した場合、

$$H(-c_2, c_2, c_2) = \sigma^2 - 4c_2 e^{-c_2} = K(c_2) \tag{3.2.54}$$

であるから、$K'(t) = 4e^{-t}(t-1) = 0$ をみたす t の値は $t = 1$ である。よって、

$$\frac{f(1)}{(1 - F(1))(2F(1) - 1)} = \frac{e}{e-1} < 2$$

となり、$(y_1, y_2, y_3) = (-2, 0, 2)$ は 3-Principal Points の候補となる。

3-Principal Points が原点に関して非対称であるとした場合、$c_2 \geq 0$ ならば

$$\begin{aligned} \frac{\partial H}{\partial c_1} &= 2h_1(G'(c_1) - 1) + 2(c_1 - G(c_1)) \\ &= 2e^{c_1}(1 - h_1) = 0 \end{aligned} \tag{3.2.55}$$

$$\begin{aligned} \frac{\partial H}{\partial c_2} &= 2(c_2 - c_1 - h_1)(G'(c_1) + 1) + 2(c_2 + G(c_2)) \\ &= 2e^{-c_2}(c_2 - c_1 - h_1 - 1) = 0 \end{aligned} \tag{3.2.56}$$

であるから、(3.2.55) 式、(3.2.56) 式より $h_1 = 1$, $c_1 = c_2 - 2$ となる。ここで $y_2 < 0$ より $0 \leq c_2 < 1$ であり、この範囲で

$$H(c_2 - 2, c_2, 1) = \sigma^2 - 2(e^{-c_2} - e^{c_2 - 2}) + (c_2 - 1)^2 \tag{3.2.57}$$

の最小値を求めればよい。(3.2.57) 式の右辺を $K(c_2)$ とおくと

$$K'(c_2) = 2(e^{-c_2} - e^{c_2-2}) + 2(c_2 - 1) \tag{3.2.58}$$

$$K''(c_2) = -2(e^{-c_2} + e^{c_2-2}) + 2 \tag{3.2.59}$$

$$K'''(c_2) = 2(e^{-c_2} - e^{c_2-2}) > 0 \tag{3.2.60}$$

であるから、(3.2.60) 式より $K''(c_2)$ は狭義単調増加となる。ここで

$$\begin{cases} K''(0) = -2e^{-2} < 0 \\ K''(1) = 2 - 4e^{-1} > 0 \end{cases}$$

だから、$K''(u) = 0$ をみたす $0 < u < 1$ なる u が唯一つ存在し、

$$\begin{cases} K''(c_2) \leq 0 \quad (0 \leq c_2 \leq u) \\ K''(c_2) > 0 \quad (u < c_2 < 1) \end{cases}$$

となる。さらに

$$\begin{cases} K'(0) = -2e^{-2} < 0 \\ K'(1) = 0 \end{cases}$$

となることから、$0 \leq c_2 < 1$ において $K'(c_2) < 0$ であり、$K(c_2)$ は単調減少する。従って、最小値は存在しない。

$c_2 < 0$ の場合は、(3.2.55) 式及び

$$\frac{\partial H}{\partial c_2} = 2(c_2 - c_1 - h_1)(G'(c_1) + 1) + 2(c_2 + G(c_2))$$

$$= 2(c_2 - c_1 - h_1 + 1)(2 - e^{c_2}) + 4(c_2 - 1) = 0 \tag{3.2.61}$$

が成り立つ必要がある。(3.2.55) 式より $h_1 = 1$ だから、(3.2.61) 式に代入すると

$$c_1 = c_2 - \frac{2(1 - c_2)}{2 - e^{c_2}} \tag{3.2.62}$$

となる。(3.2.62) 式と $h_1 = 1$ を $H(c_1, c_2, h_1)$ に代入した式を $K(c_2)$ とすると

$$K(c_2) = \sigma^2 + (c_2^2 - 1) + \frac{4(1 - c_2)(c_2 - e^{c_2})}{2 - e^{c_2}} + \left\{ \frac{2(1 - c_2)}{2 - e^{c_2}} \right\}^2 \tag{3.2.63}$$

である。これを c_2 で微分すると

$$K'(c_2) = \frac{8(1-c_2)^2 e^{c_2}}{(2-e^{c_2})^3} + \frac{4(c_2-1)(e^{2c_2}-c_2e^{c_2}+2)}{(2-e^{c_2})^2} + \frac{4(1-c_2)+2c_2e^{c_2}}{2-e^{c_2}} \quad (3.2.64)$$

となり、ここで $L(c_2) = e^{-c_2}(2-e^{c_2})^3 K'(c_2)$ とおくと

$$L(c_2) = 8c_2 + 4(c_2^2 - 2c_2 - 1)e^{c_2} + 2(2-c_2)e^{2c_2} \quad (3.2.65)$$

$$L'(c_2) = 8 + 4(c_2^2 - 3)e^{c_2} + 2(3-2c_2)e^{2c_2} \quad (3.2.66)$$

$$L''(c_2) = 4(c_2-1)e^{c_2}(c_2+3-2e^{c_2}) \quad (3.2.67)$$

であり、さらに $r(c_2) = c_2 + 3 - 2e^{c_2}$ とおくと

$$r'(c_2) = 1 - 2e^{c_2} \quad (3.2.68)$$

となる。従って、$c_2 \leq -\log 2$ のとき $r(c_2)$ は単調増加であり、

$$\begin{cases} r(-\log 2) = 2 - \log 2 > 0 \\ r(-3) = -2e^{-3} < 0 \end{cases}$$

より $r(c_2) = 0$ は $-\infty < c_2 < -\log 2$ において唯一の解 u をもつ。また $r(c_2)$ は $-\log 2 < c_2$ のとき単調減少し、$r(0) = 1 > 0$ であるから $-\log 2 < c_2 < 0$ では $r(c_2) > 0$ である。以上より

$$\begin{cases} L''(c_2) \geq 0 & (-\infty < c_2 \leq u) \\ L''(c_2) < 0 & (u < c_2 < 0) \end{cases}$$

が成立する。これより $L'(c_2)$ は $-\infty < c_2 \leq u$ で単調増加、$u < c_2 < 0$ で単調減少し、さらに

$$\begin{cases} \lim_{c_2 \to -\infty} L'(c_2) = 8 > 0 \\ L'(0) = 2 > 0 \end{cases}$$

であることから、$-\infty < c_2 < 0$ で $L'(c_2) > 0$ となり、$L(c_2)$ は単調増加である。ここで $L(0) = 0$ より $K'(c_2) < 0$ $(c_2 < 0)$ となるから $K(c_2)$ は単調減少となり、$c_2 < 0$ における最小値は存在しない。

以上より、両側指数分布における 3-Principal Points は $(y_1, y_2, y_3) = (-2, 0, 2)$ となり、これ以外では極小値も最小値もとらない。

3.2.3 混合正規分布

混合正規分布として、

$$F(x) = (1-\varepsilon)N(x;0,1^2) + \varepsilon N(x;0,\alpha^2) \tag{3.2.69}$$

を考えると、3-Principal Points が平均 (原点) に関して非対称となる場合がある。

1 例として、$\varepsilon = 0.88$, $\alpha = 0.231$ の場合がある。$t \simeq 0.375$ のとき $K'(t) = 0$ となるが、

$$\frac{tf(t)}{(1-F(t))(2F(t)-1)} \simeq 2.326 > 2$$

となり、$(y_1, y_2, y_3) = (-2t, 0, 2t)$ は 3-Principal Points とならない。この場合の 3-Principal Points は、k-means 法を援用したアルゴリズムにより計算すると $(-1.255, -0.113, 0.392)$ という非対称な値となる。

また、$K'(t) = 0$ をみたす t が複数存在することもある。1 例として、$\varepsilon = 0.97$, $\alpha = 0.12$ の場合、$K'(t) = 0$ をみたす t は $t \simeq 0.119, 0.206, 0.612$ であるが、

- $t \simeq 0.119$ のときは $\dfrac{tf(t)}{(1-F(t))(2F(t)-1)} \simeq 2.103 > 2$
- $t \simeq 0.206$ のときは $\dfrac{tf(t)}{(1-F(t))(2F(t)-1)} \simeq 3.097 > 2$

となりいずれも極小とはならない。しかし、$t \simeq 0.612$ のときに

$$\frac{tf(t)}{(1-F(t))(2F(t)-1)} \simeq 0.763 < 2$$

をみたすので $(y_1, y_2, y_3) = (-2t, 0, 2t)$ $(t \simeq 0.612)$ は極小値をとる。また、数値計算により 3-Principal Points となることが示される。

第 4 章

対称な 1 変量確率分布における k個の主要点の対称性に関する定理

第 3章では、期待値に関して対称な 3 点が対称な 1 変量確率分布における目的関数を極小にする十分条件を求め、種々の分布において計算機シミュレーションにより求められた 3-Principal Points の期待値に関する対称性について述べた。しかし、この条件は 3-Principal Points の対称性に関する直接的な十分条件にはなっていない。また、$k \geq 4$ のときには第 2章および第 3章におけるような、目的関数の最小化による十分条件の導出はあまり効果的ではない。

Tarpey[61] は、密度関数が対称かつ強単峰 (strongly unimodal) であれば 2-Principal Points が期待値に関して対称となることを示している。一方、Chow[7] は区間 $[0, 1]$ において連続な関数を区分的多項式を用いて最小 2 乗近似する際に、最適な近似が一意に定まる十分条件を示している。Li & Flury[34] はこの条件を確率分布の分位関数に適用して、2 以上のすべての自然数 kに関して成り立つ、k-Principal Points の対称性に関する定理を発表したが、この定理は誤りである。

本章では、Chow[7] の定理を分位関数 Q に適用した新しい定理を導出する。また、Li & Flury[34] における k-Principal Points の対称性に関する定理の誤りを指摘する。さらに、Trushkin[63] の定理を密度関数 fに適用した新しい定理を導出し、k-Principal Points が対称となる確率分布族を拡張する。

4.1 Chowの定理の適用

Li & Flury[34] による定理 3 の証明においては、$(\log Q'(u))'' > 0 \Rightarrow \log Q'(u)$ が凹関数としている点に誤りがあり、正しくは$\underline{(\log Q'(u))'' \leq 0} \Rightarrow \log Q'(u)$ が凹関数である。従って、Chow[7] の定理に基づく定理は以下のようになる。

定理 9.

対称性をもつ 1 変量分布において密度関数 fが正となる区間で 2 階微分可能であるとき、fが $f \cdot f'' - 2(f')^2 \geq 0$ をみたすならば、k-Principal Points はあらゆる自然数 kにおいて期待値に関する対称性をもつ。

Li & Flury(1995) においては $\log Q'(u)$ が 1 階微分または 2 階微分不能となる点が存在する場合に関しての考察がない。従って、$\log Q'(u)$ が凹関数であることについての考察としては不十分であり、定理 9 は fが正となる区間で 1 階微分または 2 階微分不能となる点が存在する場合においては適用できない。

$\log Q'(u)$ が 1 階微分または 2 階微分不能となる点が存在する場合に $\log Q'(u)$ が凹関数である十分条件は、$\lim_{u \to u_0 - 0} (\log Q'(u))' \geq \lim_{u \to u_0 + 0} (\log Q'(u))'$ が $\log Q'(u)$ の 1 階微分または 2 階微分不能な任意の u_0に関して成立し、さらに 2 階微分可能な任意の u に関して $(\log Q'(u))'' \leq 0$ が成立することである。これより、定理 9 は以下のように拡張可能である。

定理 10.

対称性をもつ 1 変量分布において密度関数 fが正となる区間で連続であるとき、

$$\lim_{x \to x_0 - 0} f'(x) \leq \lim_{x \to x_0 + 0} f'(x) \tag{4.1.1}$$

が fの 1 階微分または 2 階微分不能な任意の x_0で成立し、さらに 2 階微分可能な任意の x に関して

$$f(x)f''(x) - 2(f'(x))^2 \geq 0 \tag{4.1.2}$$

が成立するならば、k-Principal Points はあらゆる自然数 kにおいて期待値に関する対称性をもつ。

4.2 Trushkinの定理の適用

第4.1節と同様に、Trushkin[63]による定理5を対称性をもつ1変量分布の密度関数fに適用すると、あらゆる自然数kにおいてk-Principal Pointsが期待値に関する対称性をもつための十分条件は以下のように示される。

十分条件 (A)

対称性をもつ1変量分布において密度関数fが正となる区間で2階微分可能であるとき、常に$f \cdot f'' - (f')^2 \leq 0$が成立する。

また、$\log f(x)$が1階微分または2階微分不能となる点が存在する場合においても、第4.1節と同様の方法により、十分条件 (A) を以下のように拡張できる。

十分条件 (B)

対称性をもつ1変量分布において密度関数fが正となる区間で連続であるとき、

$$\lim_{x \to x_0 - 0} f'(x) \geq \lim_{x \to x_0 + 0} f'(x) \tag{4.2.3}$$

がfの1階微分または2階微分不能な任意のx_0で成立し、さらに2階微分可能な任意のxに関して

$$f(x)f''(x) - (f'(x))^2 \leq 0 \tag{4.2.4}$$

が成立する。

4.3 種々の分布への適用例

本節では、種々の対称な1変量確率分布に関して、第4.1節および第4.2節において導出した定理および十分条件を適用した例について述べる。

4.3.1 正規分布

平均μ, 分散σ^2の正規分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

で表される。これより f'および f''は

$$f'(x) = -\frac{x-\mu}{\sigma^2} f(x)$$

$$\begin{aligned} f''(x) &= \left(-\frac{x-\mu}{\sigma^2}\right)' f(x) + \left(-\frac{x-\mu}{\sigma^2}\right) f'(x) \\ &= \left\{\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\right\} f(x) \end{aligned}$$

と書けるので、すべての x において

$$\begin{aligned} f(x)f''(x) - (f'(x))^2 &= \left[\left\{\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\right\} - \left(-\frac{x-\mu}{\sigma^2}\right)^2\right] (f(x))^2 \\ &= -\frac{1}{\sigma^2}(f(x))^2 < 0 \end{aligned}$$

となり、第4.2節の十分条件 (A) をみたす。従って、正規分布における k-Principal Points は、あらゆる自然数 kに関して対称となる。

4.3.2 一様分布

一様分布の密度関数は

$$f(x) = \begin{cases} \dfrac{1}{2\theta}, & (\mu - \theta < x < \mu + \theta) \\ 0, & (\text{その他}) \end{cases}$$

と表すことができ、平均μ, 分散$\dfrac{\theta^2}{3}$である。

ここで $f(x) > 0$ となる区間 $(\mu-\theta, \mu+\theta)$ においては $f'(x) = 0$, $f''(x) = 0$ であるから、区間内のすべての x に関して $f(x)f''(x) - (f'(x))^2 = 0$ となる。これは第4.2節の十分条件をみたすので、一様分布における k-Principal Points は、あらゆる自然数 kに関して対称となる。

4.3.3 ロジスティック分布

ロジスティック分布の密度関数は

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

と表され、平均 0, 分散$\frac{\pi^2}{3}$である。

ここで、f'および f''は

$$\begin{aligned} f'(x) &= \frac{-e^{-x}}{(1+e^{-x})^2} + \frac{e^{-x}}{(1+e^{-x})^3}(2e^{-x}) \\ &= \left(1 - \frac{2}{1+e^{-x}}\right) f(x) \\ f''(x) &= \left(1 - \frac{2}{1+e^{-x}}\right) f'(x) + \left(1 - \frac{2}{1+e^{-x}}\right)' f(x) \\ &= \frac{1 - 4e^{-x} + e^{-2x}}{(1+e^{-x})^2} f(x) \end{aligned}$$

である。これより

$$\begin{aligned} f(x)f''(x) - (f'(x))^2 &= \left\{ \frac{1 - 4e^{-x} + e^{-2x}}{(1+e^{-x})^2} - \left(1 - \frac{2}{1+e^{-x}}\right)^2 \right\} (f(x))^2 \\ &= -\frac{2e^{-x}}{(1+e^{-x})^2}(f(x))^2 < 0 \end{aligned}$$

となり、第 4.2節の十分条件 (A) をみたす。従って、ロジスティック分布における k-Principal Points は、あらゆる自然数 kに関して対称となる。

4.3.4 両側指数分布

原点に関して対称な両側指数分布 (ラプラス分布、二重指数分布) の密度関数は

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x|} \quad (\lambda > 0)$$

と表される。分散は$\frac{2}{\lambda}$である。

$f(x)$ は $x \neq 0$ において微分可能であるので f'および f''は

$$f'(x) = \begin{cases} -\dfrac{\lambda^2}{2} e^{-\lambda x}, & (x > 0) \\ \dfrac{\lambda^2}{2} e^{\lambda x}, & (x < 0) \end{cases}, \qquad f''(x) = \frac{\lambda^3}{2} e^{-\lambda|x|}$$

となり、$f(x)f''(x) - (f'(x))^2 = 0\ (x \neq 0)$ である。

また、$x = 0$ のときには $\lim_{x \to -0} f'(x) = \dfrac{\lambda^2}{2}$, $\lim_{x \to +0} f'(x) = -\dfrac{\lambda^2}{2}$ であることから

$$\lim_{x \to -0} f'(x) \geq \lim_{x \to +0} f'(x)$$

が成り立つ。従って、密度関数 fは第 4.2節の十分条件 (B) をみたすので、両側指数分布における k-Principal Points は、あらゆる自然数 kに関して対称となる。

4.3.5　三角分布

原点に関して対称な三角分布の密度関数は $a > 0$ として

$$f(x) = \begin{cases} \dfrac{a - |x|}{a^2}, & (|x| < a) \\ 0, & (\text{その他}) \end{cases}$$

と表される。分散は$\dfrac{a^2}{6}$である。

$f(x) > 0$ となる区間 $(-a, a)$ においては $x \neq 0$ にで微分可能であるので、f'および f''は

$$f'(x) = \begin{cases} -\dfrac{1}{a^2}, & (0 < x < a) \\ \dfrac{1}{a^2}, & (-a < x < 0) \end{cases}, \qquad f''(x) = 0$$

となり、$f(x)f''(x) - (f'(x))^2 = -a^{-4} < 0\ (x \neq 0)$ である。

また、$x = 0$ のときには $\lim_{x \to -0} f'(x) = \dfrac{1}{a^2}$, $\lim_{x \to +0} f'(x) = -\dfrac{1}{a^2}$ であることから

$$\lim_{x \to -0} f'(x) \geq \lim_{x \to +0} f'(x)$$

が成り立つ。従って、密度関数 fは第 4.2節の十分条件 (B) をみたすので、三角分布における k-Principal Points は、あらゆる自然数 kに関して対称となる。

4.3.6　ベータ分布

原点に関して対称性をもつベータ分布の密度関数は $a > 0, \alpha > 0$ として

$$f(x) = \begin{cases} \dfrac{1}{2^{2\alpha-1} a B(\alpha, \alpha)} \left(1 - \dfrac{x^2}{a^2}\right)^{\alpha - 1}, & (|x| < a) \\ 0, & (\text{その他}) \end{cases} \tag{4.3.5}$$

と表される。ただし、$B(\cdot,\cdot)$ はベータ関数である。

$f(x)$ は$\alpha < 1$ のときに凸関数 (下に凸)、$\alpha > 1$ のときに凹関数 (上に凸) となる。また、$\alpha = 1$ のときには区間 $(-a, a)$ の一様分布となり、第 4.3.2節よりあらゆる自然数 kに関して k-Principal Points は対称となる。

$\alpha \neq 1$ のときには $|x| < a$ において f'および f''を

$$
\begin{aligned}
f'(x) &= c_0 \left(1 - \frac{x^2}{a^2}\right)^{\alpha-2} \left\{-\frac{2(\alpha-1)}{a^2}x\right\} \qquad (4.3.6)\\
f''(x) &= c_0 \left(1 - \frac{x^2}{a^2}\right)^{\alpha-3} \left\{\frac{4(\alpha-1)(\alpha-2)}{a^4}x^2\right\} - c_0 \left\{\frac{2(\alpha-1)}{a^2}\right\} \left(1 - \frac{x^2}{a^2}\right)^{\alpha-2}\\
&= \frac{2(\alpha-1)c_0}{a^2} \left(1 - \frac{x^2}{a^2}\right)^{\alpha-3} \left\{\frac{2(\alpha-2)}{a^2}x^2 - \left(1 - \frac{x^2}{a^2}\right)\right\}\\
&= \frac{2(\alpha-1)c_0}{a^2} \left(1 - \frac{x^2}{a^2}\right)^{\alpha-3} \left(\frac{2\alpha-3}{a^2}x^2 - 1\right) \qquad (4.3.7)
\end{aligned}
$$

と表せる。ただし $c_0 = \dfrac{1}{2^{2\alpha-1}aB(\alpha,\alpha)}$とする。

ここで、下に凸 $(0 < \alpha < 1)$ な場合には、$|x| < a$ において

$$
\begin{aligned}
f \cdot f'' - 2(f')^2 &= \frac{2(\alpha-1)c_0^2}{a^2} \left(1 - \frac{x^2}{a^2}\right)^{2\alpha-4} \left\{\frac{2\alpha-3}{a^2}x^2 - 1 - \frac{4(\alpha-1)}{a^2}x^2\right\}\\
&= \frac{2(1-\alpha)c_0^2}{a^2} \left(1 - \frac{x^2}{a^2}\right)^{2\alpha-4} \left\{\left(1 - \frac{x^2}{a^2}\right) + \frac{2\alpha}{a^2}x^2\right\}\\
&> \frac{2(1-\alpha)c_0^2}{a^2} \left(1 - \frac{x^2}{a^2}\right)^{2\alpha-3}\\
&> 0
\end{aligned}
$$

が成立し、密度関数 fは第 4.1節の定理 9 における十分条件をみたす。よって、あらゆる自然数 kに関して k-Principal Points は対称となる。

一方、上に凸 $(\alpha > 1)$ な場合には、$|x| < a$ において

$$
\begin{aligned}
f \cdot f'' - (f')^2 &= \frac{2(\alpha-1)c_0^2}{a^2} \left(1 - \frac{x^2}{a^2}\right)^{2\alpha-4} \left\{\frac{2\alpha-3}{a^2}x^2 - 1 - \frac{2(\alpha-1)}{a^2}x^2\right\}\\
&= -\frac{2(1-\alpha)c_0^2}{a^2} \left(1 - \frac{x^2}{a^2}\right)^{2\alpha-3}\\
&< 0
\end{aligned}
$$

が成立し、密度関数 f は第 4.2節の十分条件 (A) をみたす。よって、あらゆる自然数 k に関して k-Principal Points は対称となる。

以上の結果より、対称性をもつベータ分布における k-Principal Points は、あらゆる自然数 k に関して対称となることが任意の $\alpha > 0$ において示される。

4.3.7 t 分布

自由度 n の t 分布の密度関数は

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

と表され、平均 0, 分散 $\dfrac{n}{n-2}$ $(n>2)$ である。

ここで、f' および f'' は

$$\begin{aligned} f'(x) &= -\frac{n+1}{2}\frac{2x}{n}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+3}{2}} \\ &= -\frac{(n+1)x}{n+x^2}f(x) \\ f''(x) &= -\frac{(n+1)x}{n+x^2}f'(x) - \left(\frac{(n+1)x}{n+x^2}\right)' f(x) \\ &= \frac{(n+1)\{(n+2)x^2-n\}}{(n+x^2)^2}f(x) \end{aligned}$$

である。これより

$$\begin{aligned} f(x)f''(x) - (f'(x))^2 &= \left\{\frac{(n+1)\{(n+2)x^2-n\}}{(n+x^2)^2} - \left(-\frac{(n+1)x}{n+x^2}\right)^2\right\}(f(x))^2 \\ &= \frac{(n+1)(x^2-n)}{(n+x^2)^2}(f(x))^2 \end{aligned}$$

となり、$x > \sqrt{n}$ において $f(x)f''(x) - (f'(x))^2 > 0$ であるから、第 4.2節の十分条件 (A) をみたさない。また、

$$f(x)f''(x) - 2(f'(x))^2 = \left\{\frac{(n+1)\{(n+2)x^2-n\}}{(n+x^2)^2} - 2\left(-\frac{(n+1)x}{n+x^2}\right)^2\right\}(f(x))^2$$

$$= -\frac{(n+1)(x^2+n)}{(n+x^2)^2}(f(x))^2 < 0$$

であるから、第 4.1節の定理 9 における十分条件もみたさない。従って、t 分布における k-Principal Points は、$k \geq 3$ の場合には、本章における定理および十分条件からは対称性を判断することができない。

4.3.8 Johnson's S_u分布

平均が 0 の Johnson's S_u分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi}\sqrt{x^2+1}} \exp\left[-\frac{1}{2}\left\{\log(x+\sqrt{x^2+1})\right\}^2\right]$$

と表され、分散は$\dfrac{e^2-1}{2}$である。

ここで、f'および f''は

$$\begin{aligned} f'(x) &= \frac{-x}{\sqrt{2\pi}(x^2+1)^{\frac{3}{2}}} \exp\left[-\frac{1}{2}\left\{\log(x+\sqrt{x^2+1})\right\}^2\right] - \frac{\log(x+\sqrt{x^2+1})}{\sqrt{x^2+1}} f(x) \\ &= -\frac{x+\sqrt{x^2+1}\log(x+\sqrt{x^2+1})}{x^2+1} f(x) \\ f''(x) &= -\frac{x+\sqrt{x^2+1}\log(x+\sqrt{x^2+1})}{x^2+1} f'(x) - \left(\frac{x+\sqrt{x^2+1}\log(x+\sqrt{x^2+1})}{x^2+1}\right)' f(x) \\ &= \frac{\{x+\sqrt{x^2+1}\log(x+\sqrt{x^2+1})\}^2 + \{x\sqrt{x^2+1}\log(x+\sqrt{x^2+1}) - 2\}}{(x^2+1)^2} f(x) \end{aligned}$$

である。これより

$$f(x)f''(x) - (f'(x))^2 = \frac{x\sqrt{x^2+1}\log(x+\sqrt{x^2+1}) - 2}{(x^2+1)^2}(f(x))^2$$

となり、$f(x)f''(x) - (f'(x))^2 > 0$ となる x が存在するので第 4.2節の十分条件 (A) をみたさない。また、

$$f(x)f''(x) - 2(f'(x))^2 = \frac{\{x\sqrt{x^2+1}\log(x+\sqrt{x^2+1}) - 2\} - \{x+\sqrt{x^2+1}\log(x+\sqrt{x^2+1})\}^2}{(x^2+1)^2}(f(x))^2$$

より、$f(x)f''(x) - 2(f'(x))^2 < 0$ となる x が存在するので第 4.1節の定理 9 における十分条件もみたさない。従って、Johnson's S_u分布における k-Principal Points は、$k \geq 3$ の場合には、本章における定理および十分条件からは対称性を判断することができない。

4.3.9 まとめ

以上の例および定理9、十分条件(A)(B)より、対称な1変量分布のうち、これまで数値計算による値からしか k-Principal Points の対称性が考察されていなかった正規分布において、理論的に対称性が成立することが確かめられた。また、これまで3以上の Principal Points の対称性について考察されていなかった、分布においても、一様分布、ロジスティック分布、両側指数分布、三角分布、ベータ分布においては対称性が成立することがわかった。一方で、2-Principal Points の対称性が成り立つ分布のうち、t 分布、Johnson's S_u分布については、本章の定理および十分条件からは3以上の Principal Points の対称性が成立するかどうかは確認できなかった。

第 5 章

2 変量正規分布における k個の主要点の配置

Flury[21] は、2 変量が互いに独立な正規分布が与えられたとき、一方の分散を固定した上で他方の分散を変化させ、いくつかの場合における k-Principal Points ($k \leq 5$) の配置について数値計算で示した。特に、一方の分散と他方の分散の比が 3 のときに k-Principal Points($k \leq 5$) が一直線上に並ぶ配置が得られたことから、k-Principal Points が一直線上に並ぶ配置において

(a) kが一定かつ分散比が変化する場合における分散比の境界値

(b) 分散比が一定かつ kが変化する場合における kの最大値

に関する問題を提起した。しかしながら、これらの問題は未だ解明されていない。また、期待値に関して対称な分布が与えられた場合において、kが多い場合の k-Principal Points がどのような配置をとるかも未解明である。

本章では、2 変量正規分布の分散共分散行列 $\mathrm{diag}(\sigma^2, 1)$ におけるσの値と k-Principal Points について数値計算を行い、得られた結果を示す。また、2 変量標準正規分布が与えられた場合における k-Principal Points ($k \leq 12$) についても数値計算を行い、得られた配置について考察する。

5.1　分散共分散行列の値と k個の主要点との関係

この節では、様々なσについて、第2.4.2節で述べたPrincipal Pointsの導出アルゴリズムを用いてk-Principal Pointsを求めた。

各σの値に対する$P_F(k)(k=3,4,5)$の値を図5.1～図5.3に実線で示す。

これより、第2.4.1節の問題(a)において、

(i) $\sigma_0(3) \simeq 1.66$

(ii) $\sigma_0(4) \simeq 2.15$

(iii) $\sigma_0(5) \simeq 2.67$

と考えられる。また、これらの結果より、第2.4.1節の問題(b)において、

(i) $1 < \sigma < \sigma_0(3)$ ならば kの最大値は2

(ii) $\sigma_0(3) \leq \sigma < \sigma_0(4)$ ならば kの最大値は3

(iii) $\sigma_0(4) \leq \sigma < \sigma_0(5)$ ならば kの最大値は4

であることが確認できる。

5.2　2変量標準正規分布におけるk個の主要点の配置

前節までの議論は、Principal Pointsの数が5つ以下と少ない場合についてのものであるが、点の数をさらに増やした場合にどのような配置となるかを、2変量標準正規分布の場合において以下に示す。

5.2.1　計算機シミュレーションによる解

点の数が3以上の場合において、種々の初期値を与えてk-means法を援用したアルゴリズムにより計算機シミュレーションを行ったところ、$k=3$及び$k=4$の場合においては図5.4のような局所的最適配置が得られた。これらの配置のうち、各図の左端の太枠で囲った

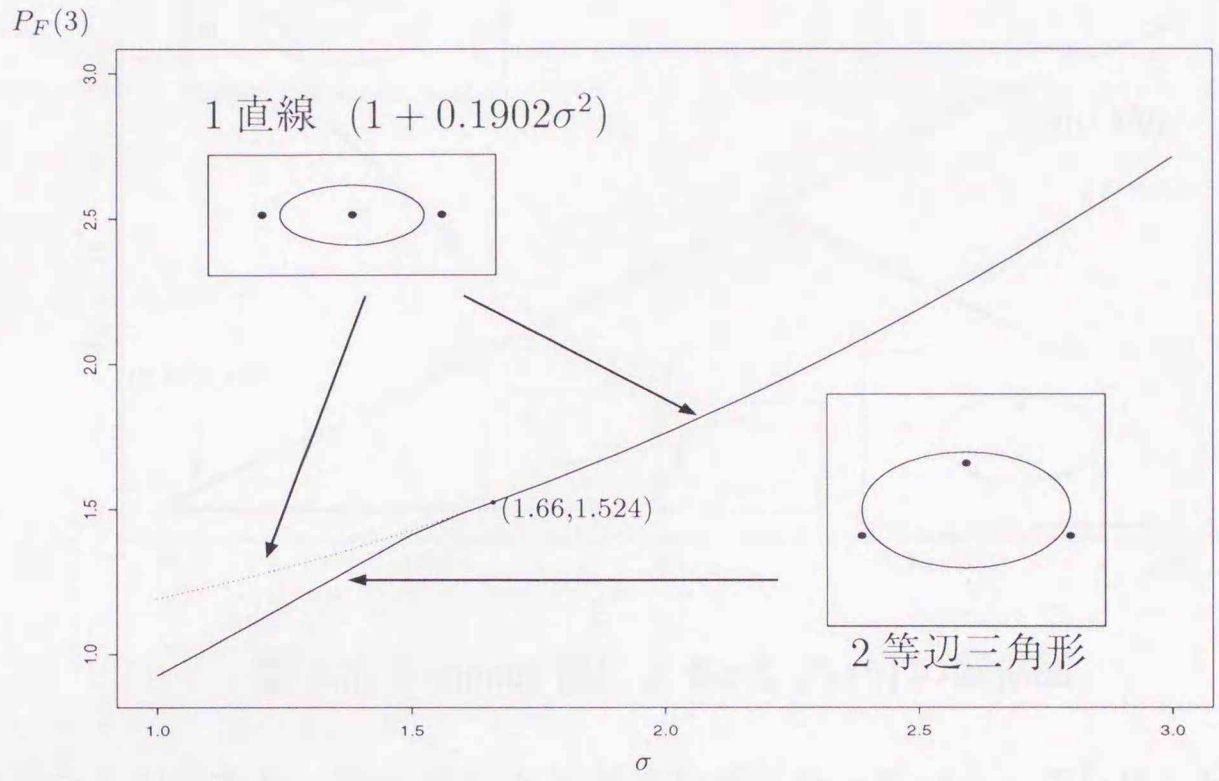


図 5.1: k-means 法によるσと $P_F(3)$ の関係図

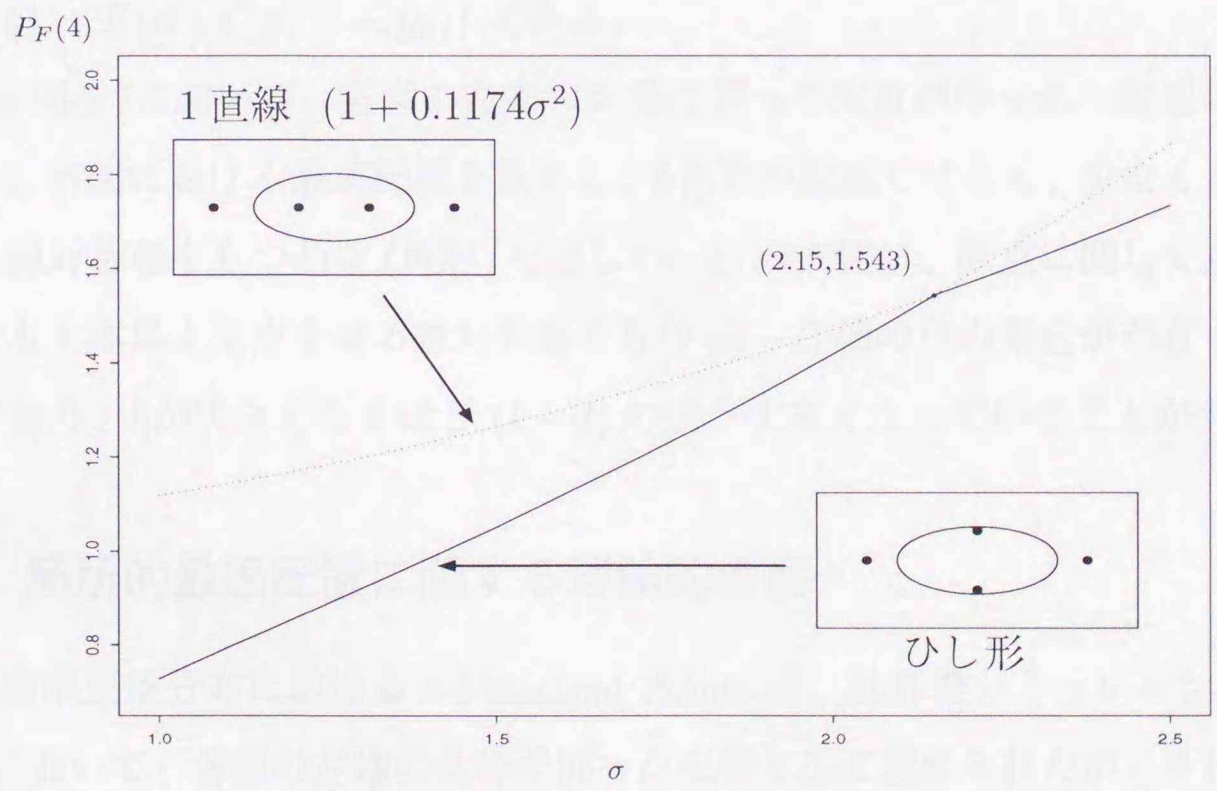


図 5.2: k-means 法によるσと $P_F(4)$ の関係図

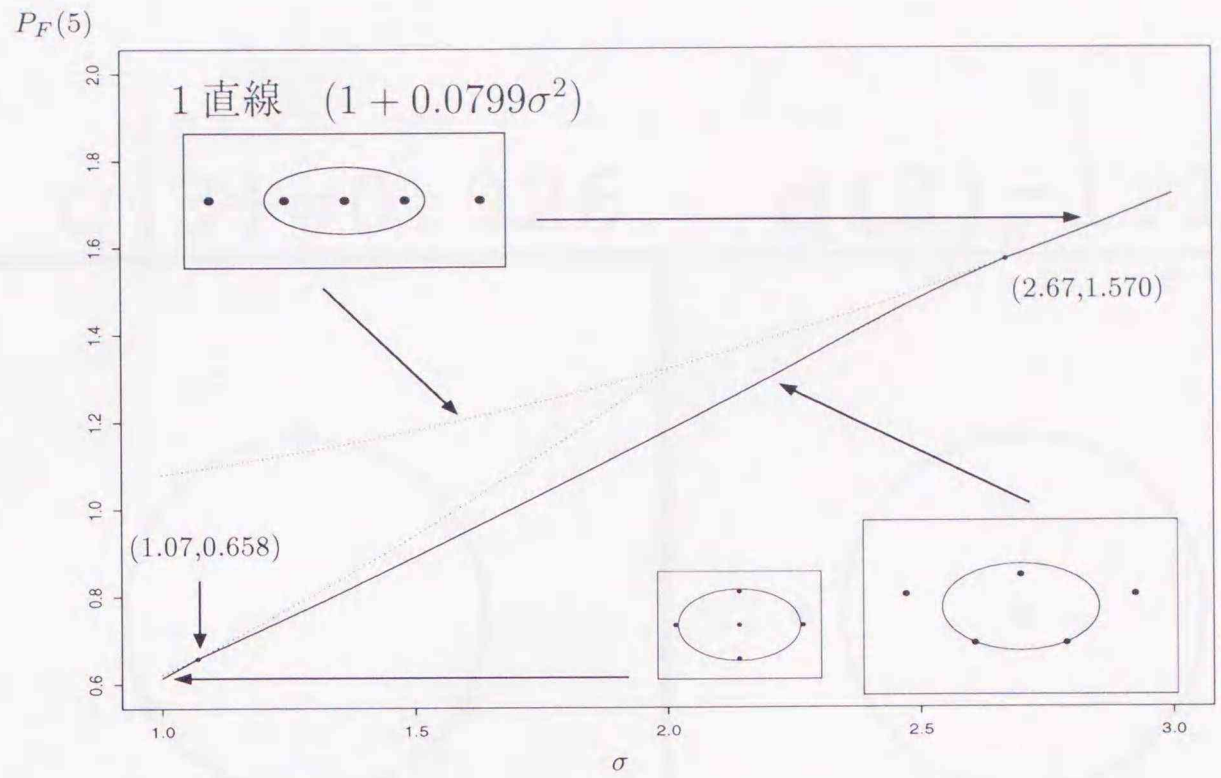


図 5.3: k-means 法によるσと $P_F(5)$ の関係図

配置が最適なものであり、Flury[21] と同様の結果となっている。さらに、kが 5 以上 12 以下の場合において、現時点では図 5.5～図 5.7のような局所的最適配置が得られている。ただし、$q(k) = E\{d^2(X|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k)\}$ である。

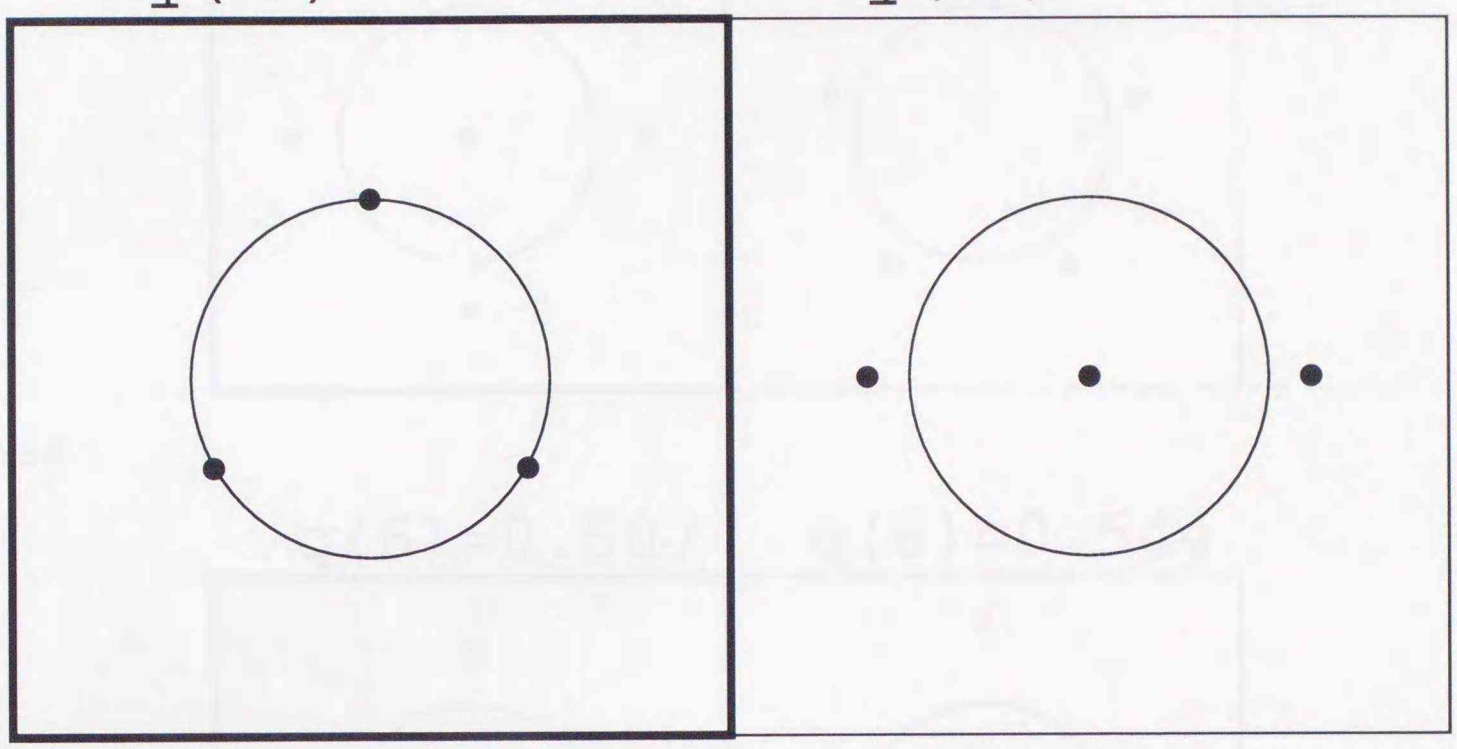
図 5.5～図 5.7において、各図の左端の太枠で囲った配置が得られた配置の中で最適なものである。各図における最適配置を見ると、k角形の配置ではなく、少なくとも 1 本以上原点を通る線対称軸をもつ凸な l角形 (ただし $l < k$) の内側に、原点に関して点対称もしくは少なくとも 1 本以上原点を通る線対称軸をもつ $(k-l)$ 個の点の集合が存在するような配置になっており、kが大きくなるほど $(k-l)$ の値が大きくなっていることがわかる。

5.2.2　局所的最適配置に関する理論的考察

2 変量標準正規分布における k-Principal Points は、計算機シミュレーションでは図 5.5～図 5.7において、各図の左端の太枠で囲った配置として求められたが、それ以外にも目的関数を極小にする配置がいくつか求められた。ここでは、それらの配置のうち、

- k個の点の配置が原点を中心とする正 k角形となる場合
- k個の点の配置が原点を中心とする正 $(k-1)$ 角形＋原点となる場合

(a)$k = 3$



(b)$k = 4$

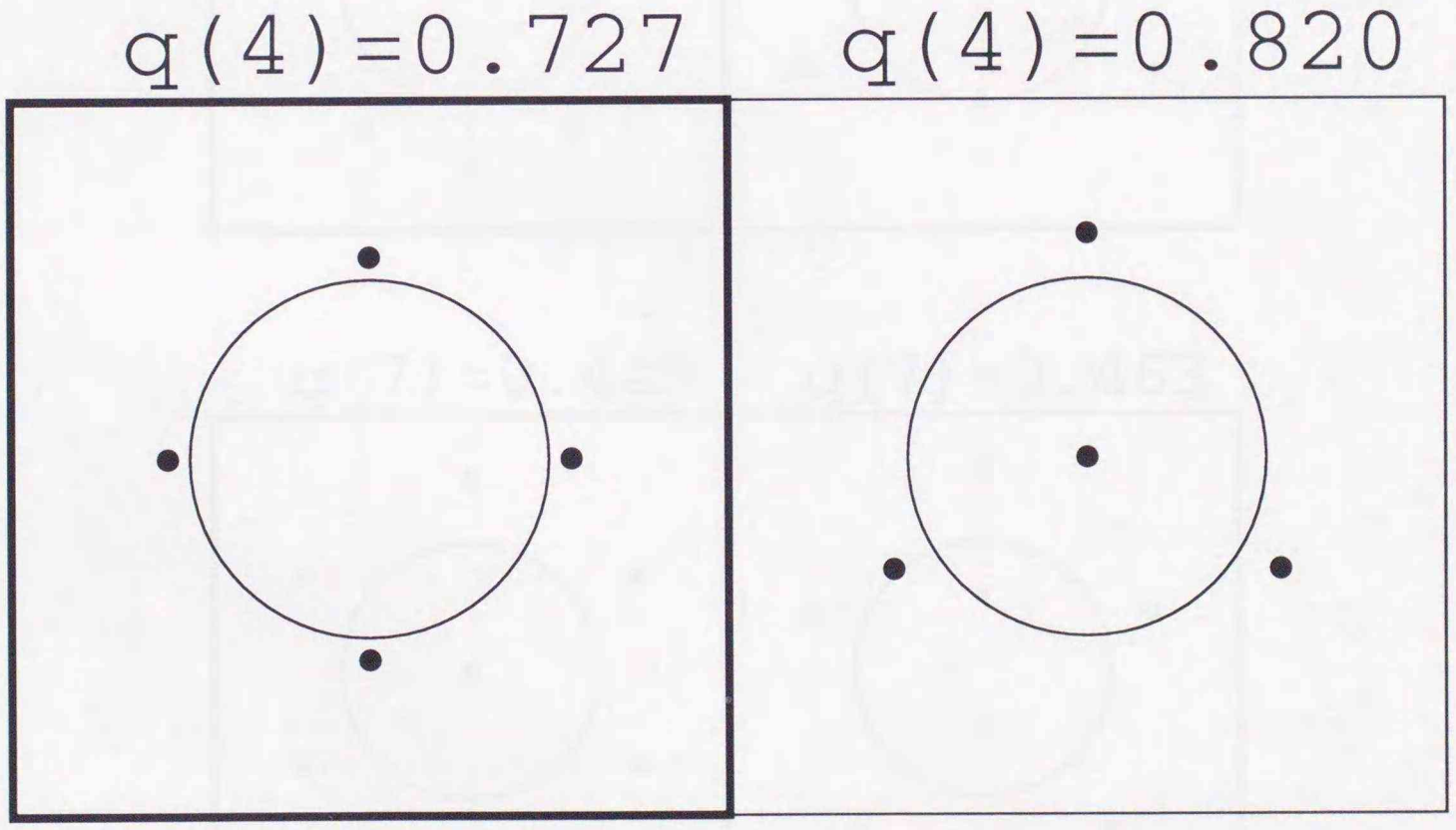
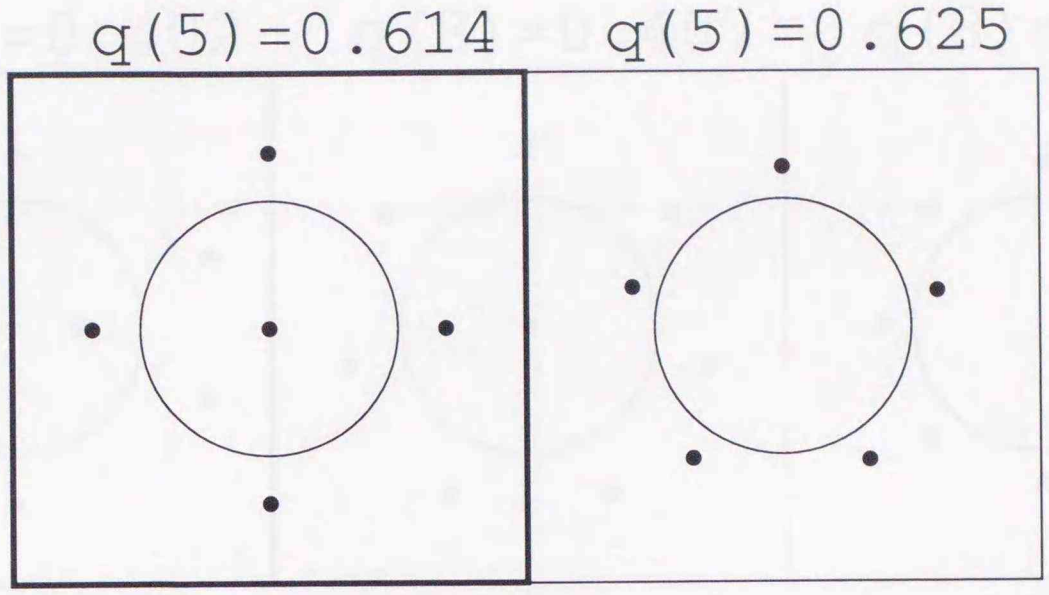
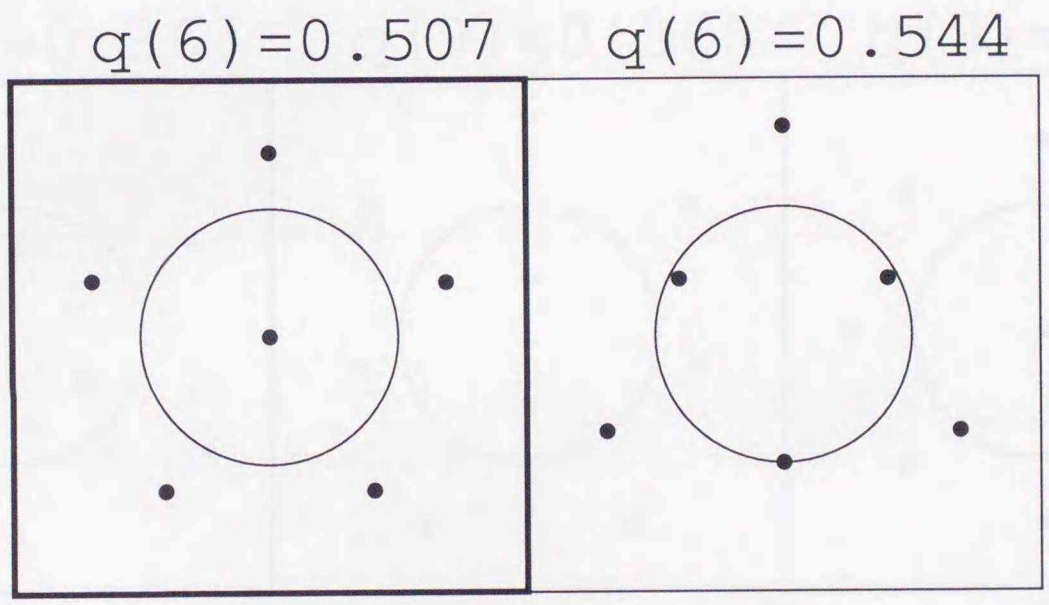


図 5.4: 2 変量標準正規分布における k個の点の局所的最適配置図 $(k = 3, 4)$

(c)$k = 5$



(d)$k = 6$



(e)$k = 7$

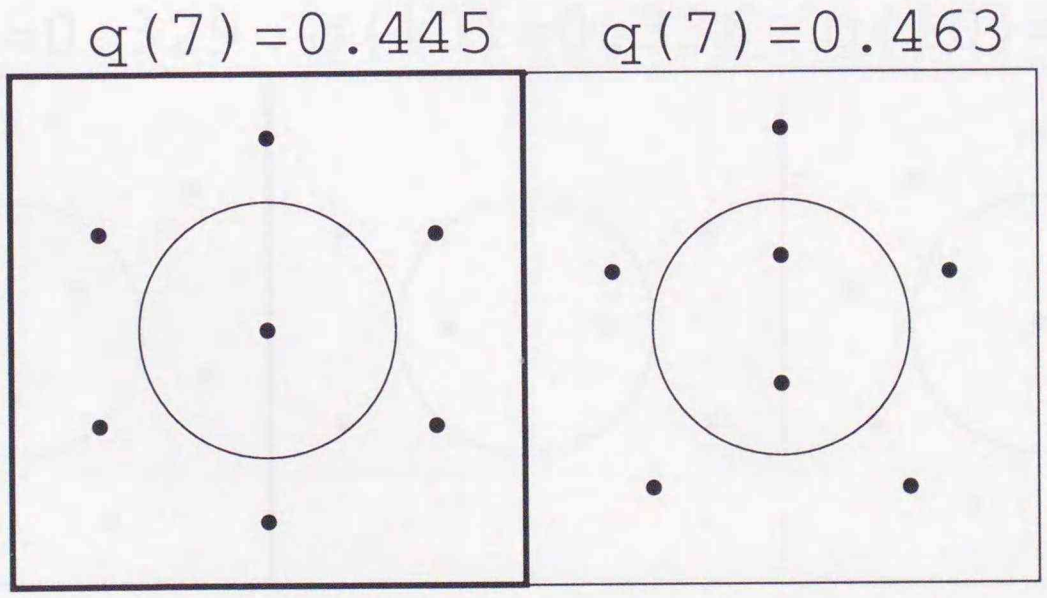
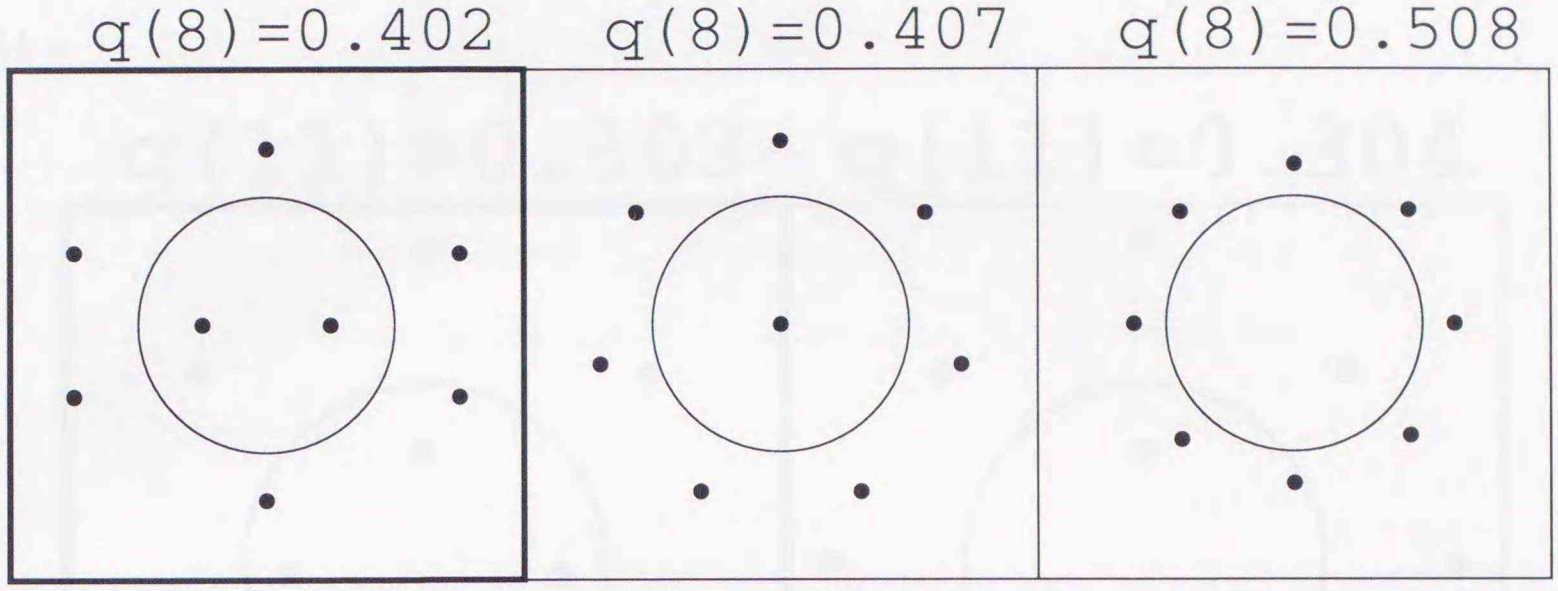
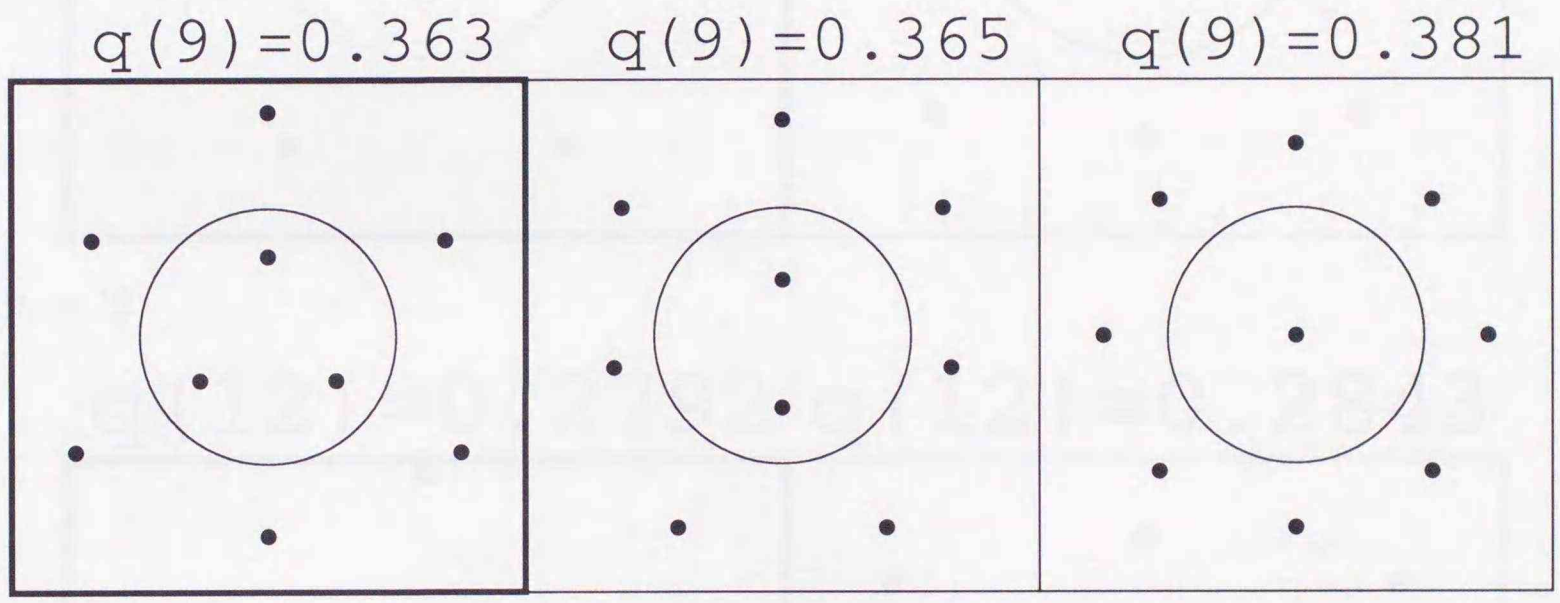


図 5.5: 2 変量標準正規分布における k個の点の局所的最適配置図 $(k = 5, 6, 7)$

(f)$k = 8$



(g)$k = 9$



(h)$k = 10$

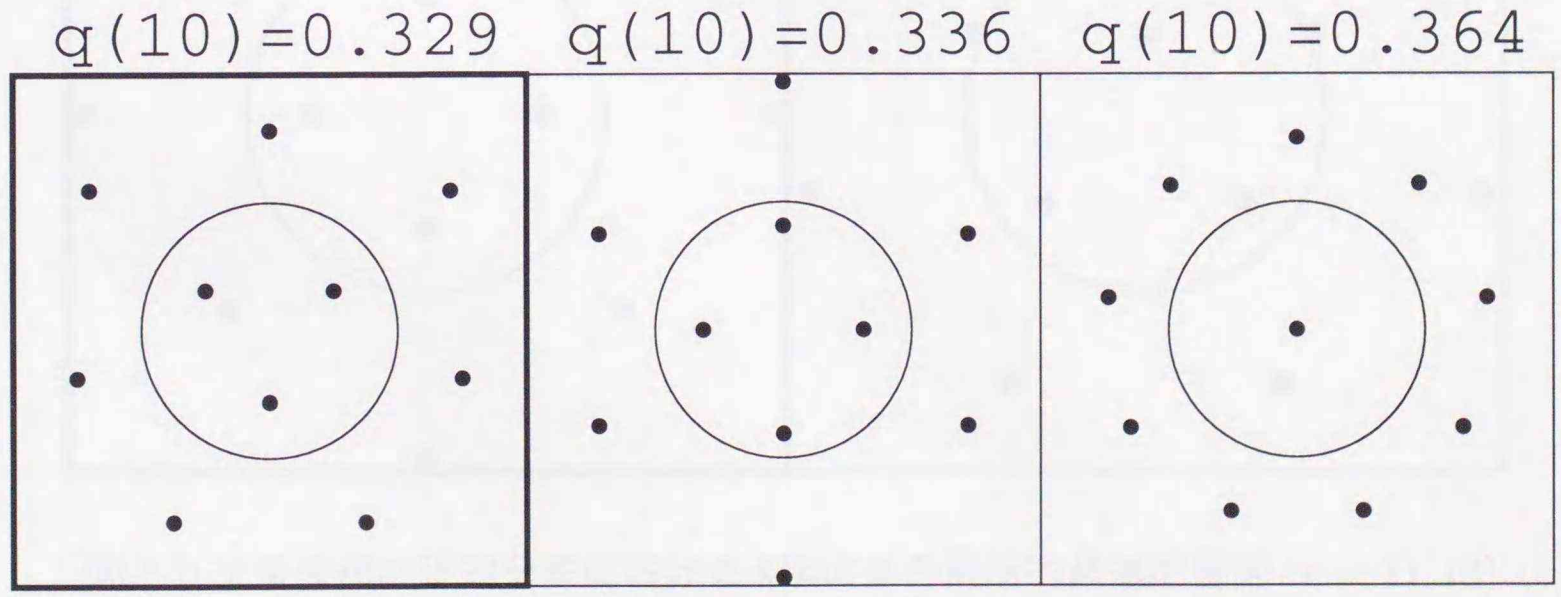
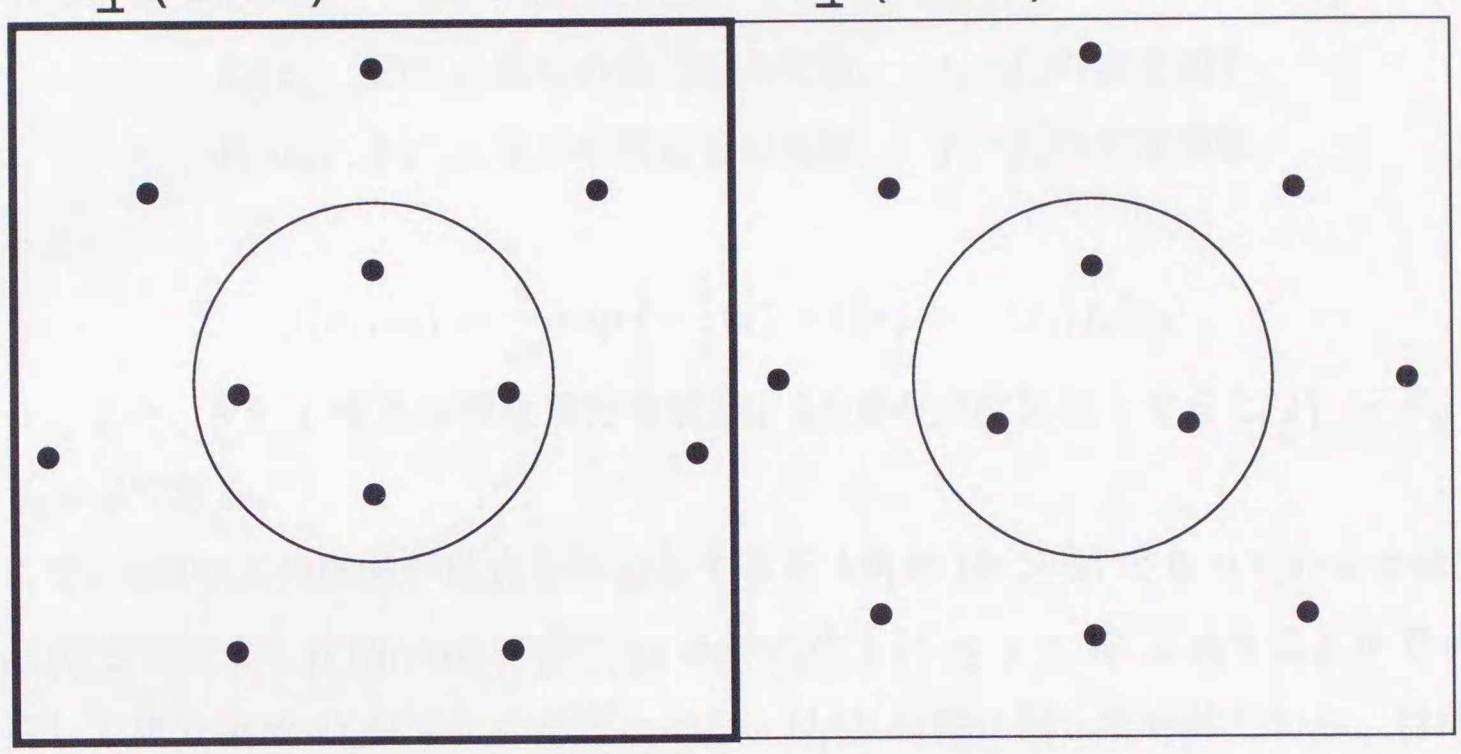


図 5.6: 2 変量標準正規分布における k個の点の局所的最適配置図 $(k = 8, 9, 10)$

(i)$k = 11$



(j)$k = 12$

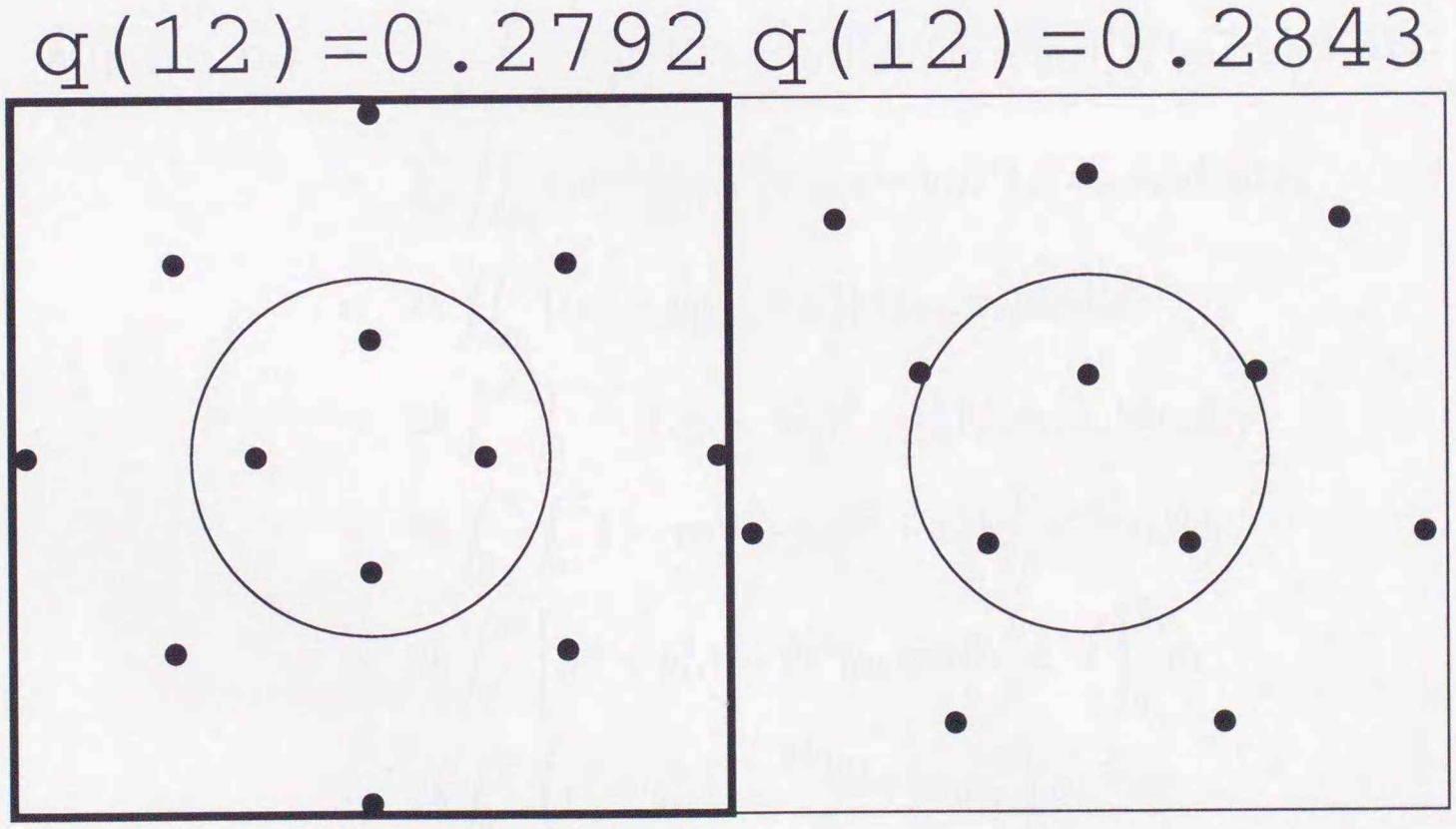


図 5.7: 2 変量標準正規分布における k個の点の局所的最適配置図 $(k = 11, 12)$

について理論的考察を行う。

5.2.2.1　原点を中心とする正 k 角形となる場合

2 変量標準正規分布の分布関数 F とその密度関数 f において

$$F_1(x_1) : F\text{の } x_1\text{成分の周辺分布関数}, \quad f_1 : F_1\text{の密度関数}$$
$$F_2(x_2) : F\text{の } x_2\text{成分の周辺分布関数}, \quad f_2 : F_2\text{の密度関数}$$

とすると、

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x_1^2 + x_2^2)\right\} = f_1(x_1) f_2(x_2)$$

である。また、Φ を 1 変量標準正規分布関数、ϕ を Φ の密度関数とすると $F_1 = F_2 = \Phi$, $f_1 = f_2 = \phi$ である。

ここで、k 個の点の配置が原点を中心とする正 k 角形 $(k \geq 3)$ となっていると仮定し、$\boldsymbol{p}_1(y_{11}, 0)$ とおくと、$\boldsymbol{p}_j(y_{11}\cos\frac{(j-1)2\pi}{k}, y_{11}\sin\frac{(j-1)2\pi}{k})$ $(1 \leq j \leq k)$ と表すことができ、各 $\boldsymbol{p}_j$ に関して積分領域 D_j が等しくなる。また、D_1 は x_1 軸に関して対称だから、目的関数 $M(\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k)$ は

$$\begin{aligned}
M(\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \min_{1\leq j\leq k}\{(x_1 - y_{j1})^2 + (x_2 - y_{j2})^2\} f(x_1, x_2) dx_2 dx_1 \\
&= \sum_{j=1}^{k} \iint_{D_j} \{(x_1 - y_{j1})^2 + (x_2 - y_{j2})^2\} f(x_1, x_2) dx_2 dx_1 \\
&= 2k \iint_{D_1'} \{(x_1 - y_{11})^2 + x_2^2\} f(x_1, x_2) dx_2 dx_1 \\
&= 2k \int_0^{\infty}\int_0^{m_k x_1} \{(x_1 - y_{11})^2 + x_2^2\} f(x_1, x_2) dx_2 dx_1 \\
&= 2k \int_0^{\infty}\int_0^{\frac{\pi}{k}} \{(r\cos\theta - y_{11})^2 + x_2^2\} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r d\theta dr \\
&= 2k \int_0^{\infty} \left[(r^3 + y_{11}^2 r - 2r^2 y_{11}\cos\theta)\frac{k}{\pi} e^{-\frac{r^2}{2}}\right]_0^{\frac{\pi}{k}} dr \\
&= 2k \int_0^{\infty} \left(r^3 + y_{11}^2 r - \frac{2ky_{11}}{\pi} r^2 \sin\frac{\pi}{k}\right) e^{-\frac{r^2}{2}} dr \\
&= y_{11}^2 - \frac{2k}{\sqrt{2\pi}} y_{11} \sin\frac{\pi}{k} + 2 \\
&= \left(y_{11} - \frac{k}{\sqrt{2\pi}}\sin\frac{\pi}{k}\right)^2 + 2 - \frac{k^2}{2\pi}\sin^2\frac{\pi}{k}
\end{aligned} \tag{5.2.1}$$

と書ける。ただし、

$$m_k = \tan\frac{\pi}{k}(k \geq 2)$$

$$\begin{cases} x_1 = r\cos\theta \\ x_2 = r\sin\theta \end{cases}$$

とおいた。また、$k=4$ のときの $D_j(1 \leq j \leq k), D_1'$の領域は図 5.8、図 5.9 のようになる。これより、式 (5.2.1) は $y_{11} = \dfrac{k}{\sqrt{2\pi}}\sin\dfrac{\pi}{k}$で最小値 $2 - \dfrac{k^2}{2\pi}\sin^2\dfrac{\pi}{k}$をとり、これは $q(k)$ となる。よって、各 $q(k)$ 及び y_{11}の値は表 5.1 のようになり、図 5.5〜図 5.7における $q(k)$ に一致する。

表 5.1: 2 変量標準正規分布における $q(k)$ 及び y_{11} (k個の点が正 k角形の場合)

k	$q(k)$	y_{11}
3	$2 - \dfrac{27}{8\pi}(\simeq 0.9257)$	$\sqrt{\dfrac{27}{8\pi}}(\simeq 1.0365)$
4	$2 - \dfrac{4}{\pi}(\simeq 0.7268)$	$\sqrt{\dfrac{4}{\pi}}(\simeq 1.1283)$
5	$2 - \dfrac{25(5-\sqrt{5})}{16\pi}(\simeq 0.6253)$	$\sqrt{\dfrac{25(5-\sqrt{5})}{16\pi}}(\simeq 1.1724)$
6	$2 - \dfrac{9}{2\pi}(\simeq 0.5676)$	$\sqrt{\dfrac{9}{2\pi}}(\simeq 1.1968)$
8	$2 - \dfrac{8(2-\sqrt{2})}{\pi}(\simeq 0.5083)$	$\sqrt{\dfrac{8(2-\sqrt{2})}{\pi}}(\simeq 1.2213)$

5.2.2.2　原点を中心とする正 $(k-1)$ 角形十原点となる場合

k個の点の配置が原点+原点を中心とする正 $(k-1)$ 角形 $(k \geq 4)$ となっていると仮定し、$\boldsymbol{p}_1(y_{11}, 0)$、$\boldsymbol{p}_k(0,0)$ とおくと、$\boldsymbol{p}_j(y_{11}\cos\frac{(j-1)2\pi}{k-1}, y_{11}\sin\frac{(j-1)2\pi}{k-1})$ $(1 \leq j \leq k-1)$ と表すことができ、$\boldsymbol{p}_1, \cdots, \boldsymbol{p}_{k-1}$に関して積分領域 $D_1, \cdots, D_{k-1}$が等しくなる。また、D_1, D_kは x_1軸に関して対称だから、目的関数 $M(\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k)$ は

$$M(\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \min_{1\leq j\leq k}\{(x_1 - y_{j1})^2 + (x_2 - y_{j2})^2\} f(x_1, x_2)dx_2dx_1$$

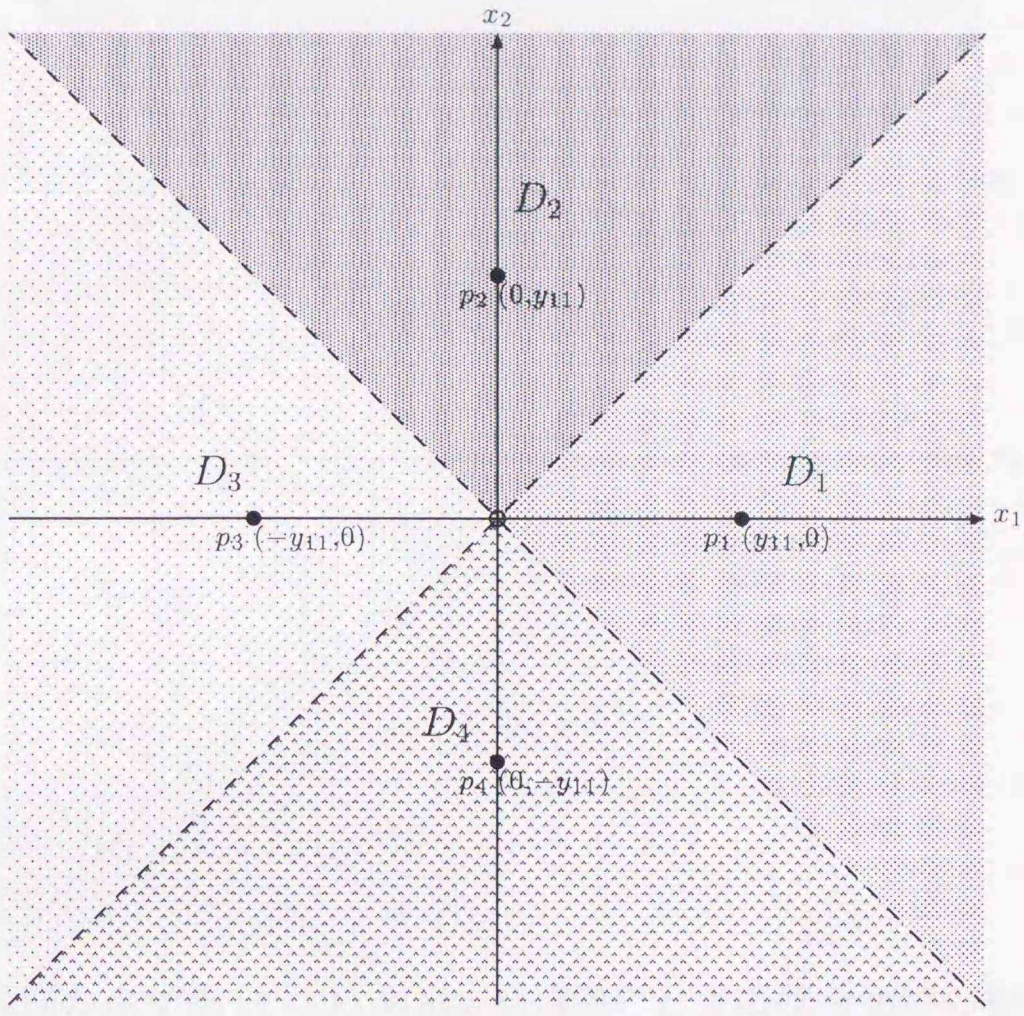


図 5.8: k個の点が正k角形のときの積分領域の概念図 (その 1)

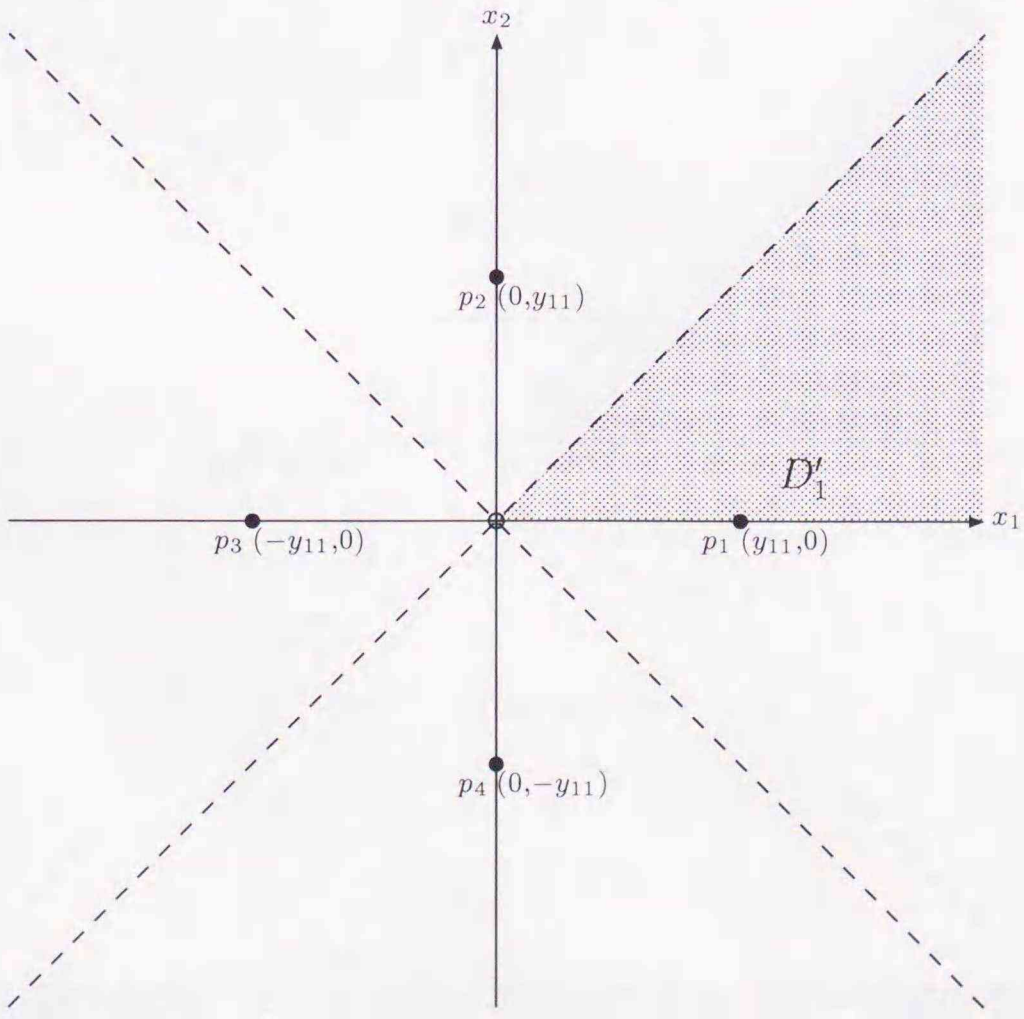


図 5.9: k個の点が正k角形のときの積分領域の概念図 (その 2)

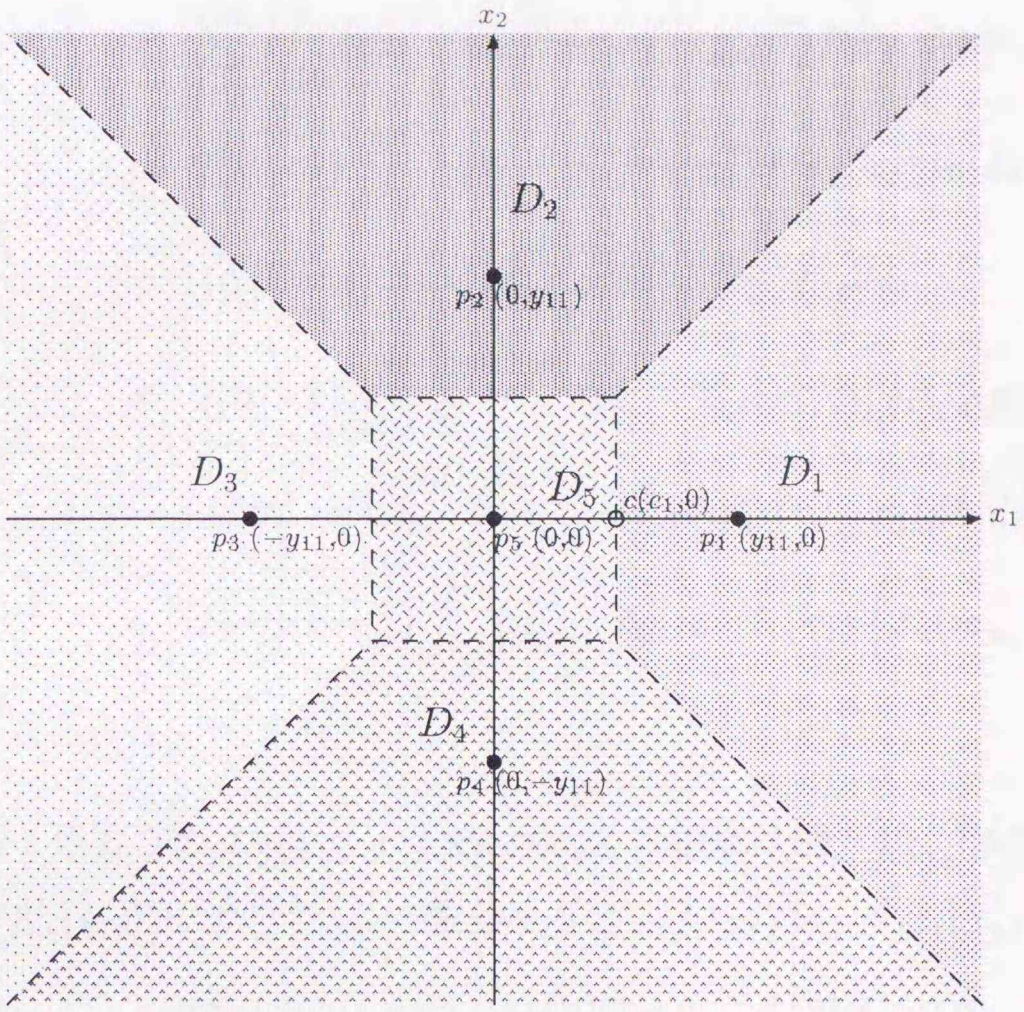


図 5.10: k個の点が正 $(k-1)$ 角形+期待値のときの積分領域の概念図 (その 1)

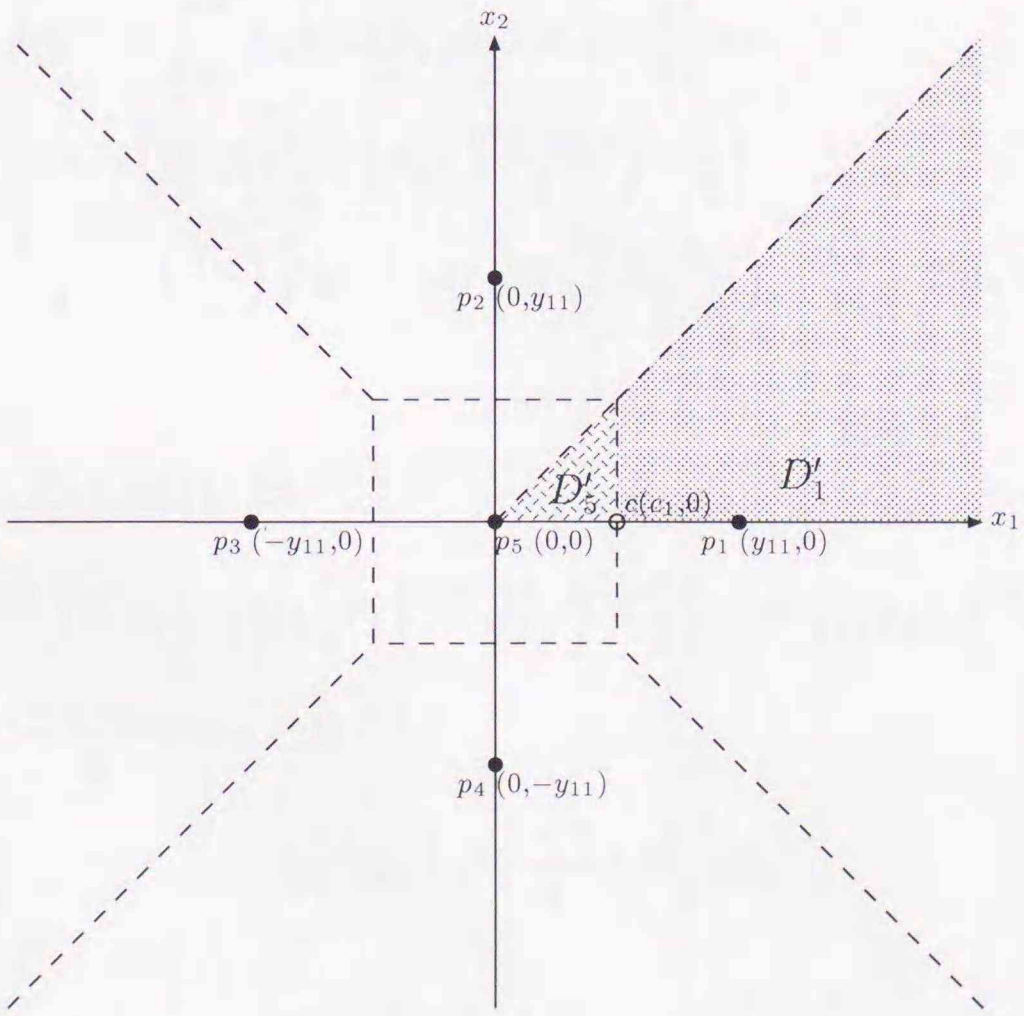


図 5.11: k個の点が正 $(k-1)$ 角形+期待値のときの積分領域の概念図 (その 2)

$$
\begin{aligned}
&= \sum_{j=1}^{k} \iint_{D_j} \{(x_1 - y_{j1})^2 + (x_2 - y_{j2})^2\} f(x_1, x_2) dx_2 dx_1 \\
&= 2(k-1) \iint_{D_1'} \{(x_1 - y_{11})^2 + x_2^2\} f(x_1, x_2) dx_2 dx_1 \\
&\quad +2(k-1) \iint_{D_k'} (x_1^2 + x_2^2) f(x_1, x_2) dx_2 dx_1 \\
&= 2(k-1) \int_{\frac{y_{11}}{2}}^{\infty} \int_{0}^{m_{k-1}x_1} (y_{11}^2 - 2y_{11}x_1) f_1(x_1) f_2(x_2) dx_2 dx_1 \\
&\quad +2(k-1) \int_{0}^{\infty} \int_{0}^{m_{k-1}x_1} (x_1^2 + x_2^2) f(x_1, x_2) dx_2 dx_1 \\
&= 2 + 2(k-1) \int_{\frac{y_{11}}{2}}^{\infty} (y_{11}^2 - 2y_{11}x_1) f_1(x_1) \left\{ F_2(m_{k-1}x_1) - \frac{1}{2} \right\} dx_1
\end{aligned}
\tag{5.2.2}
$$

と書ける。$k=5$ のときの $D_j(1 \leq j \leq k), D_1', D_k'$ の領域は図 5.10、図 5.11のようになる。

式 (5.2.2) を $K_k(y_{11})$ とおいて y_{11} で微分すると、$f_1'(x_1) = -x_1 f_1(x_1)$ より

$$
K_k'(y_{11}) = 4(k-1) \int_{\frac{y_{11}}{2}}^{\infty} (y_{11} - x_1) f_1(x_1) \left\{ F_2(m_{k-1}x_1) - \frac{1}{2} \right\} dx_1 \tag{5.2.3}
$$

$$
\begin{aligned}
K_k''(y_{11}) &= 4(k-1) \int_{\frac{y_{11}}{2}}^{\infty} f_1(x_1) \left\{ F_2(m_{k-1}x_1) - \frac{1}{2} \right\} dx_1 \\
&\quad -(k-1) y_{11} f_1(\frac{y_{11}}{2}) \left\{ F_2\left(\frac{m_{k-1}y_{11}}{2}\right) - \frac{1}{2} \right\}
\end{aligned}
\tag{5.2.4}
$$

$$
K_k^{(3)}(y_{11}) = \frac{k-1}{4} f_1\left(\frac{y_{11}}{2}\right) \left[(y_{11}^2 - 12) \left\{ F_2\left(\frac{m_{k-1}y_{11}}{2}\right) - \frac{1}{2} \right\} - 2m_{k-1}y_{11} f_2\left(\frac{m_{k-1}y_{11}}{2}\right) \right] \tag{5.2.5}
$$

が求められる。式 (5.2.5) において

$$
L_k\left(\frac{m_{k-1}y_{11}}{2}\right) = (y_{11}^2 - 12) \left\{ F_2\left(\frac{m_{k-1}y_{11}}{2}\right) - \frac{1}{2} \right\} - 2m_{k-1}y_{11} f_2\left(\frac{m_{k-1}y_{11}}{2}\right) \tag{5.2.6}
$$

とおき、さらに $\frac{m_{k-1}y_{11}}{2} = a$ とおくと

$$
K_k^{(3)}(y_{11}) = \frac{k-1}{4} f_1(\frac{y_{11}}{2}) L_k(a) \tag{5.2.7}
$$

であり

$$
L_k(a) = \left(\frac{4a^2}{m_{k-1}^2} - 12 \right) \left\{ F_2(a) - \frac{1}{2} \right\} - 4a f_2(a) \tag{5.2.8}
$$

となる。$\frac{k-1}{4}f_1(\frac{y_{11}}{2}) > 0$ だから、(5.2.6) の増減を調べるために (5.2.8) を a で微分すると

$$L'_k(a) = \frac{4}{m_{k-1}^2} f_2(a) \left\{ (1+m_{k-1}^2)a^2 - 4m_{k-1}^2 \right\} + \frac{8a}{m_{k-1}^2} \left\{ F_2(a) - \frac{1}{2} \right\} \tag{5.2.9}$$

$$L''_k(a) = \frac{4}{m_{k-1}^2} f_2(a) \left\{ -(1+m_{k-1}^2)a^3 + (4+6m_{k-1}^2)a \right\} + \frac{8}{m_{k-1}^2} \left\{ F_2(a) - \frac{1}{2} \right\} \tag{5.2.10}$$

$$L_k^{(3)}(a) = \frac{4}{m_{k-1}^2} f_2(a) \{ (1+m_{k-1}^2)a^4 - (7+9m_{k-1}^2)a^2 + 6(1+m_{k-1}^2) \} \tag{5.2.11}$$

が求められる。ここで式 (5.2.11) において $b = a^2$ とし、さらに

$$\begin{aligned} M_k(b) &= (1+m_{k-1}^2)b^2 - (7+9m_{k-1}^2)b + 6(1+m_{k-1}^2) \qquad (5.2.12) \\ &= (1+m_{k-1}^2) \left\{ b - \frac{(7+9m_{k-1}^2)}{2(1+m_{k-1}^2)} \right\}^2 - \frac{57m_{k-1}^4 + 78m_{k-1}^2 + 25}{4(1+m_{k-1}^2)} \end{aligned}$$

とおくと、$\min M_k(b) < 0$ であるから、$\min M_k(b) = 0$ は 2 つの実数解をもつ。実数解を α, β $(\alpha < \beta)$ とすると $\frac{7+9m_{k-1}^2}{2(1+m_{k-1}^2)} > 0$ かつ $\alpha\beta = 6$ より $0 < \alpha < \beta$ であり、$b > 0$ なので

$$\begin{cases} M_k(b) \geq 0 & (0 < b \leq \alpha, \beta \leq b) \\ M_k(b) < 0 & (\alpha < b < \beta) \end{cases}$$

が成り立つ。(5.2.12) より、(5.2.11) は

$$L_k^{(3)}(a) = \frac{4}{m_{k-1}^2} f_2(a) M_k(a^2) \tag{5.2.13}$$

と表すことができ、常に $\frac{4}{m_{k-1}^2} f_2(a) > 0$ だから

$$\begin{cases} L_k^{(3)}(a) \geq 0 & (0 < a \leq \sqrt{\alpha}, \sqrt{\beta} \leq a) \\ L_k^{(3)}(a) < 0 & (\sqrt{\alpha} < a < \sqrt{\beta}) \end{cases}$$

である。これより (5.2.10) は $a = \sqrt{\alpha}$ で極大値、$a = \sqrt{\beta}$ で極小値をとる。また、

$$\begin{cases} L''_k(0) = 0 \\ \lim_{a \to \infty} L''_k(a) = \frac{4}{m_{k-1}^2} > 0 \\ L''_k(\sqrt{\beta}) = \frac{8}{m_{k-1}^2} \left\{ F_2(\sqrt{\beta}) - \frac{1}{2} \right\} - 12\beta^{-1/2}(\beta-2)(1+m_{k-1}^{-2}) f_2(\sqrt{\beta}) > 0 \end{cases}$$

だから $L_k''(a)$ は常に正である。よって $L_k'(a)$ は単調増加であり、

$$\begin{cases} L_k'(0) &= -16f_2(0) < 0 \\ \lim\limits_{a\to\infty} L_k'(a) &= \infty \end{cases}$$

だから $L_k'(a) = 0$ をみたす a がただ1つ存在する。この値を a_0 とすると、

$$\begin{cases} L_k'(a) \leq 0 & (0 \leq a \leq a_0) \\ L_k'(a) > 0 & (a_0 < a) \end{cases}$$

であるから $L_k(a)$ は $a = a_0$ で極小値をとる。ここで

$$\begin{cases} L_k(0) &= 0 \\ \lim\limits_{a\to\infty} L_k(a) &= \infty \end{cases}$$

だから $L_k(0) > L_k(a_0)$ より $L_k(a_0) < 0$ となる。よって $L_k(a) = 0$ をみたす a は区間 (a_0, ∞) にただ1つ存在し、その値を a_1 とすると

$$\begin{cases} L_k(a) \leq 0 & (0 \leq a \leq a_1) \\ L_k(a) > 0 & (a_1 < a) \end{cases}$$

である。$\dfrac{m_{k-1}y_{11}}{2} = a$ だから

$$\begin{cases} L_k(\dfrac{m_{k-1}y_{11}}{2}) \leq 0 & (0 \leq a \leq \dfrac{2a_1}{m_{k-1}}) \\ L_k(\dfrac{m_{k-1}y_{11}}{2}) > 0 & (\dfrac{2a_1}{m_{k-1}} < a) \end{cases}$$

となり、(5.2.7) と $\dfrac{k-1}{4} f_1(\dfrac{y_{11}}{2}) > 0$ より $K_k''(y_{11})$ は $y_{11} = \dfrac{2a_1}{m_{k-1}}$ で極小値をとる。ここで

$$\begin{cases} K_k''(0) &= 4(k-1)\displaystyle\int_0^\infty f_1(x_1)\left\{F_2(m_{k-1}x_1) - \frac{1}{2}\right\} dx_1 > 0 \\ \lim\limits_{y_{11}\to\infty} K_k''(y_{11}) &= 0 \end{cases}$$

だから $K_k''(\dfrac{2a_1}{m_{k-1}}) < \lim\limits_{y_{11}\to\infty} K_k''(y_{11})$ より $K_k''(\dfrac{2a_1}{m_{k-1}}) < 0$ となる。よって $K_k''(y_{11}) = 0$ をみたす y_{11} が区間 $(0, \dfrac{2a_1}{m_{k-1}})$ にただ1つ存在し、その値を a_2 とすると

$$\begin{cases} K_k''(y_{11}) > 0 & (0 < y_{11} < a_2) \\ K_k''(y_{11}) \leq 0 & (a_2 \leq y_{11}) \end{cases}$$

である。これより (5.2.3) は $y_{11}=a_2$で極大値をとる。そして

$$\begin{cases} K_k'(0) &= -4(k-1)\displaystyle\int_0^\infty x_1 f_1(x_1)\left\{F_2(m_{k-1}x_1)-\frac{1}{2}\right\}dx_1 < 0 \\ \lim\limits_{y_{11}\to\infty} K_k'(y_{11}) &= 0 \end{cases}$$

であるから、$K_k'(a_2) > \lim\limits_{y_{11}\to\infty} K_k'(y_{11})$ より $K_k'(a_2)>0$ となる。これより $K_k'(y_{11})=0$ をみたす y_{11}が区間 $(0,a_2)$ でただ 1 つ存在し、その値を a_3とすると

$$\begin{cases} K_k'(y_{11}) < 0 & (0 < y_{11} < a_3) \\ K_k'(y_{11}) \geq 0 & (a_3 \leq y_{11}) \end{cases}$$

であるから、$K_k(y_{11})$ は $y_{11}=a_3$で極小値

$$\begin{aligned} K_k(a_3) &= 2+2(k-1)\int_{\frac{a_3}{2}}^\infty (a_3^2-2a_3x_1)f_1(x_1)\left\{F_2(m_{k-1}x_1)-\frac{1}{2}\right\}dx_1 \\ &= 2-2(k-1)a_3\int_{\frac{a_3}{2}}^\infty x_1 f_1(x_1)\left\{F_2(m_{k-1}x_1)-\frac{1}{2}\right\}dx_1 \\ &= 2-2(k-1)a_3\left\{\left[-f_1(x_1)\left\{F_2(m_{k-1}x_1)-\frac{1}{2}\right\}\right]_{\frac{a_3}{2}}^\infty + \int_{\frac{a_3}{2}}^\infty m_{k-1}f_1(x_1)f_2(m_{k-1}x_1)dx_1\right\} \\ &= 2-2(k-1)a_3\left[\phi(\frac{a_3}{2})\left\{\Phi\left(\frac{m_{k-1}a_3}{2}\right)-\frac{1}{2}\right\}+\int_{\frac{a_3}{2}}^\infty m_{k-1}\phi(\sqrt{1+m_{k-1}^2}x_1)dx_1\right] \\ &= 2-2(k-1)a_3\left[\phi(\frac{a_3}{2})\left\{\Phi\left(\frac{m_{k-1}a_3}{2}\right)-\frac{1}{2}\right\}+\frac{m_{k-1}}{\sqrt{1+m_{k-1}^2}}\Phi\left(\frac{\sqrt{1+m_{k-1}^2}a_3}{2}\right)\right] \end{aligned} \tag{5.2.14}$$

をとり、これが y_{11}に関する最小値 $q(k)$ となる。$K_k'(y_{11})=0$ をみたす y_{11}を計算すると、各 $q(k)$ 及び y_{11}の値は表 5.2のようになり、図 5.5〜図 5.7における $q(k)$ に一致する。

表 5.2: 2 変量標準正規分布における $q(k)$ 及び y_{11} (k個の点が原点+正 $(k-1)$ 角形の場合)

k	$q(k)$	y_{11}	k	$q(k)$	y_{11}
4	0.820	1.279	8	0.407	1.454
5	0.614	1.363	9	0.381	1.466
6	0.507	1.410	10	0.364	1.474
7	0.445	1.438	11	0.351	1.479

第 6 章

おわりに

本章では、本論文における研究についてまとめ、他分野との関連および今後に残された課題について述べる。

6.1 本研究のまとめ

本論文では、期待値に関して対称な 1 変量分布における k-Principal Points の対称性、および 2 変量正規分布における k-Principal Points の配置に関して得られた性質について述べた。

種々の対称な 1 変量分布においては、これまで 2-Principal Points の対称性 (一意性) に関してさまざまな研究が Flury[21] などにより行われていたが、3 個以上の Principal Points の対称性についてはこれまであまり考察が行われていなかった。そこで、期待値に関して対称な 3 点が目的関数を最小にするかどうかについて考察し、期待値に関して対称な 3 点が目的関数を極小にするための必要条件および十分条件を導出した。また、この条件を期待値に関して対称な種々の 1 変量分布に適用し、3-Principal Points の対称性が成り立つかどうかを調べた。そのうち、ロジスティック分布及び両側指数分布に関しては 3-Principal Points が期待値に関して対称となったが、混合正規分布については、2-Principal Points の場合と同様に、重みや分散の値によっては非対称な 3-Principal Points が存在することが確かめられた。また、計算機シミュレーションにより、非対称な 3-Principal Points の値を求

めた。しかし、以上に述べた研究においては、Principal Points が対称となるための十分条件は直接的には示すことはできなかった。

Tarpey[62] は、対称な 1 変量分布における 2-Principal Points の対称性 (一意性) に関する定理の十分条件を、密度関数の性質に着目して導出している。この定理を皮切りに、あらゆる kにおいて k-Principal Points の対称性が成立する確率分布族に関する条件についてさまざまな研究が行われた。その中で、Li & Flury[34] は、k-Principal Points の対称性が成り立つための十分条件を示し、対称性が成り立つ確率分布族を拡張した。しかしながら、この条件をみたすものの、Flury[21] の定理における必要条件をみたさない分布が存在するという矛盾が生じていた。

そこで、Li & Flury[34] の定理を再検証した結果、定理の十分条件の導出時において誤りがあることがわかった。そのため、条件の導出に利用された Chow[7] の定理を踏まえた上で、誤りを訂正した正しい定理を示した。また、Li & Flury[34] の定理における十分条件においては、密度関数が 1 階微分または 2 階微分不能となる点が存在する場合に関しての考察がなく、このような場合にも適用可能となるように定理を拡張した。一方で、Trushkin[63] における、平均 2 乗誤差規準により連続な 1 変量確率変数を最適量子化した値が k-Principal Points と同等となることから、Trushkin[63] の定理における十分条件をみたす確率分布族において k-Principal Points の対称性が成立することを示した。さらに、密度関数が 1 階微分または 2 階微分不能となる点が存在する場合においても適用可能となるように十分条件を拡張した。

これらの十分条件は、確率分布の密度関数 fの 1 階微分および 2 階微分、さらには fが微分不能な点における f'の左右の極限値の大小関係からのみ示されており、対称な 1 変量確率分布における k-Principal Points の対称性を判断する上で極めて便利な式である。これらの条件式を、種々の対称な 1 変量確率分布に対して適用し、正規分布など、数値計算によってしか 3 個以上の Principal Points の対称性が確認されていなかった分布や、これまで考察が行われていなかった分布において、これらの定理および十分条件をみたすものの例を示した。また、2-Principal Points が対称であるものの、密度関数がこれらの定理や十分条件をみたさない分布の例として、t 分布や Johnson's S_u分布などがあることも示した。

2 変量正規分布における分散共分散行列と k-Principal Points($3 \leq k \leq 5$) の配置との関

係については、Flury[21] により 3 種類の分散共分散行列についての数値計算例が示されていたが、本研究においては、分散共分散行列がさらにさまざまな値をとる場合において、より詳細に考察した。

2 変量正規分布における k-Principal Points は、分散共分散行列 $\mathrm{diag}(\sigma^2, 1)$ の値により、一直線に並ぶ場合と k角形を形成する場合がある。この問題について、種々の分散共分散行列の値を与え、k-Principal Points が第 2 変数軸に関して対称と仮定した上で、第 1 変数軸上に並ぶ場合及び k角形を形成する場合において k-means 法と同様のアルゴリズムによる計算を行い、目的関数の極小値を求めた。その結果、$3 \leq k \leq 5$ の場合において、k個の点の形が k角形から直線に変わるσの境界値がそれぞれ求まった。これは、Flury[21] が提起した第 2.4.1節の問題 (a) における$\sigma_0(k)$ の値となる。

次に、2 変量標準正規分布において k-Principal Points がどのような配置をとるかについても計算機シミュレーションを用いて考察を行った。その結果、点の数が多くなると、(最も外側に位置する l個の点 (ただし $l < k$)) + (その内側の $(k-l)$ 個の点) という配置が得られた。また、どの場合の最適配置についても、原点を通る直線のうち少なくとも 1 本が線対称軸となることがわかった。

さらに、k-Principal Points の配置およびいくつかの局所的最適配置のうち、理論的考察が比較的容易な

(i) k個の点の配置が原点を中心とする正 k角形となる場合

(ii) k個の点の配置が原点を中心とする正 $(k-1)$ 角形+原点となる場合

において、目的関数の値を 2 重積分により求め、目的関数および k個の点の値の数学的な裏付けを与えた。

6.2 他の研究分野への応用

本章では Principal Points と他分野との関連について触れ、本論文における研究成果の応用に関しても述べる。

6.2.1　最適施設配置問題との関連

確率分布の密度関数および Principal Points は、それぞれ最適施設配置問題において対象となる地域の人口分布および設置される施設の位置に対応づけが可能である。k-Principal Points は、各点から最も近い領域における重心としても求められ、最小 2 乗距離規準による目的関数を最小にすることが示されており、これは最適施設配置問題における最適解およびそれらにより形成されるボロノイ領域が利用者全体からの施設への平均距離に基づく配置コストを最小化することと対応する。

最適施設配置問題の解は、定式化された数学的問題を解くことにより求められるが、複雑な問題においては導出の過程で計算機シミュレーションが利用されることも多い。ここで、対象となる地域における人口分布が対称性をもつ場合において、最適な配置が常に対称となるかどうかは興味深い問題である。この問題を考える上では、k-Principal Points との前述のような関連より、本論文で述べられている、種々の対称な 1 変量分布における k-Principal Points の対称性に関する研究が基礎的理論として参考になる。すなわち、問題の対象となる 1 次元の地域における人口分布が対称性をもつ場合であっても、分布の形状によっては最適な配置が非対称となる場合もあり得る。

実際の最適施設配置問題においては、対象となる地域は 2 次元以上で表されることが多い。このような場合における最適配置に関する理論的な裏付けの第 1 歩として、2 変量正規分布における k-Principal Points の配置に関する考察を位置付けることが可能である。この考察により、対象となる 2 次元の地域における最適な配置の各種幾何学的性質を考えることができるほか、2 変量標準正規分布における k個の点の局所的最適配置およびその目的関数値を考えることで、理論的もしくは計算機的な解法により得られた施設の最適な配置が、地域の状況などさまざまな制約により採用できない場合が生じたときの代替配置を考える上での参考になると考えられる。

6.2.2　各種多変量解析およびデータ解析との関連

Principal Points の導出は、数学的にはクラスター分析における k-means 法と同等のアルゴリズムにより行われることなどから、クラスター分析および k-means 法とは密接な関連がある。k-Principal Points は、与えられた確率分布における密度関数を k個の領域に分

割する際に最適となるような各領域の重心として求められ、このときの各領域は k-means 法における各クラスターに相当する。しかしながら、k-means 法においては「データ点をいかに最適な k個のクラスターに分割するか」が主要な目的であるところが、「確率分布を k個の点で代表する」という立場である k-Principal Points との本質的な違いである。

クラスター分析においては、得られたクラスターの妥当性 (validity) をどのように判断するかが興味深いテーマの 1 つとなっているが、本論文で示されている、2 変量標準正規分布における k-Principal Points の配置は、2 変量標準正規分布が定義されている領域を大きな 1 つのクラスターとみなすとき、それをさらに k個のクラスターに分割する際の妥当性を判断する上で参考となる。

さらに、実際のデータ解析への応用として、第 2.5節に示されている天気図の解析 (村木・大瀧・水田 [85][86]) のように、主成分分析法に k-Principal Points の概念を導入した解析を組み合わせる手法が考えられているが、その他にも、層別逆回帰法 (Sliced Inverse Regression) の改良 (大瀧・藤越 [70]) などへの応用や、関数データ解析 (Functional Data Analysis) への拡張 (水田 [84]) などが試みられており、応用分野のさらなる拡大が期待できる。

6.3 今後の課題

期待値に関して対称な 1 変量分布における k-Principal Points の対称性については、密度関数の性質に着目した定理および十分条件により、対称性が成立する確率分布族が大幅に拡張された。しかし、2-Principal Points の対称性について確認されていても、k-Principal Points の対称性の有無が確認されていない分布が存在する (t 分布や Johnson's S_u分布など)。また、混合正規分布においては、同じ kにおいても、重みや分散の値によって対称性の有無が分かれる。$k = 2$ の場合においては、水田 [83] により対称性の有無に関する境界値が数値計算により示されているが、$k \geq 3$ の場合においては十分な考察が行われていない。また、同じ重みや分散をもつ場合において kの値の変化により対称性の有無が変わる場合があるかどうかもわかっていない。従って、これらの分布における対称性の有無に関して、密度関数における各種因子を考慮した新たな十分条件の導出は課題として残されている。

2 変量正規分布における k-Principal Points の配置に関しては、Flury[21] が提起した第 2.4.1節の問題 (a)、問題 (b) について $k \leq 5$ のときの解が求められたが、$k \geq 6$ のときの解については未解明である。さらに、2 変量標準正規分布においては、k-Principal Points となる配置が $k \leq 12$ の場合において計算機シミュレーションにより求められたが、点の数と配置の形状に関してどのような規則性があるのかについても課題といえる。また、各変量がそれぞれ独立かつ対称性をもつ各種多変量分布における k-Principal Points の配置への拡張も課題である。

参考文献

[1] Abaya, E. F. & Wise, G. L.(1984). Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, **44**, 183–189.

[2] Alvey, N. G., Banfield, C. F., Baxter, R. I., Gower, J. C., Krzanowski, W. J., Lane, P. W., Leech, P. K., Nelder, J. A., Payne, R. W., Phelps, K. W., Rogers, C. E., Ross, G. J. S., Simpson, H. R., Todd, A. D., Tunnicliffe-Wilson, G., Wedderburn, R. W. M., White, R. P. & Wilkinson, G. N.(1983). *Genstat : A General Statistical Program*, Oxford : Numerical Algorithm Group.

[3] Banfield, C. F. & Bassill, S.(1977). Algorithm AS 113 : A transfer algorithm for non-hierarchical classification. *Journal of the Royal Statistical Society. Series C, Applied statistics*, **26**, 206–210.

[4] Barrow, D. L., Chui, C. K., Smith, P. W. & Ward, J. D.(1978). Unicity of best mean approximation by second order splines with variable knots. *Mathematics of Computation*, **32**, 1131–1143.

[5] Bruce, J. D.(1964). On the optimum quantization of stationary signals. *IEEE International Convention Record*, **12**, 1, 118–125.

[6] Burchard, H. G. & Hale, D. F.(1975). Piecewise polynomial approximation on optimal meshes. *Journal of the Approximation theory*, **14**, 128–147.

[7] Chow, J.(1982). On the uniqueness of best $L_2[0,1]$ approximation by piecewise polynomials with variable breakpoints. *Mathematics of Computation*, **39**, 571–585.

[8] Cormack, R. M.(1971). Estimation of principal points. *Journal of the Royal Statistical Society. Series A, General*, **134**, 321–367.

[9] Cox, D. R.(1957). Note on grouping. *Journal of the American Statistical Association*, **52**, 543–547.

[10] Cuesta, J. A. & Matrán, C.(1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probability Theory and Related Fields*, **78**, 523–534.

[11] Dalenius, T.(1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, **33**, 203–213.

[12] Dalenius, T. & Gurney, M.(1951). The problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*, **34**, 133–148.

[13] Davis, P. J.(1975). Interpolation and Approximation. *Dover Publications*, **14**, New York.

[14] Day, N. E.(1969). Estimating the components of mixture of normal distributions. *Biometrika*, **56**, 463–474.

[15] Edwards, A. W. F. & Cavalli-Sforza, L. L.(1965). A method for cluster analysis. *Biometrics*, **21**, 362–375.

[16] Eubank, R. L.(1988). Optimal grouping, spacing, stratification, and piecewise constant approximation. *SIAM Review*, **30**, 3, 404–420.

[17] Fang, K., Kotz, S. & Ng, K.(1990). *Symmetric Multivariate and Related Distributions*, Chapman and Hall, New York.

[18] Fleischer, P. E.(1964). Sufficient conditions for achieving minimum distortion in a quantizer. *IEEE International Convention Record*, **12**, 1, 104–111.

[19] Flury, B.(1988). *Common Principal Components and Related Multivariate Models*, Wiley, New York.

[20] Flury, B. & Riedwyl, H.(1988). *Multivariate Statistics : a Practical Approach*, Chapman and Hall, London.

[21] Flury, B.(1990). Principal points. *Biometrika*, **77**, 1, 33–41.

[22] Flury, B.(1993). Estimation of principal points. *Journal of the Royal Statistical Society. Series C, Applied statistics*, **42**, 1, 139–151.

[23] Flury, B. & Tarpey, T.(1993). Representing a large collection of curves – a case for principal points. *American Statistical Association*, **47**, 304–306.

[24] Friedman, H. P. & Rubin, J.(1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, **62**, 1159–1178.

[25] Gordon, A. D.(1981). *Classification*, Chapman and Hall, London.

[26] Grime, J. P., Hunt, R. & Krzanowski, W. J.(1987). Evolutionary physiological ecology of plants. *Evolutionary Physiological Ecology*, Cambridge : Cambridge University Press.

[27] Hartigan, J. A.(1975). *Clustering Algorithms*, Wiley, New York.

[28] Hartigan, J. A.(1978). Asymptonic distributions for clustering criteria. *The Annals of Statistics*, **6**, 117–131.

[29] Hastie, T. & Stuetzle, W.(1989). Principal Curves. *Journal of the American Statistical Association*, **84**, 502–516.

[30] Herrndorf, N.(1983). Approximation of vector-valued random variables by constants. *Journal of Approximation Theory*, **37**, 175–181.

[31] Jones, M. C. & Rice, J. A.(1992). Displaying the important features of large collections of similar curves. *American Statistical Association*, **46**, 140–145.

[32] Karlin, S.(1968). *Total Positivity*, **1**, Stanford University Press, California.

[33] Krzanowski, W. J. & Lai, Y. T.(1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, **44**, 23–34.

[34] Li, L. & Flury, B.(1995). Uniqueness of principal points for univariate distributions. *Statistics and Probability Letters*, **25**, 323–327.

[35] Linde, Y., Buzo, A. & Gray, R. M.(1980). An algorithm for vector quantizer design. *IEEE Transactions on Communication Technology*, **COM-28**, 84–95.

[36] MacQueen, J.(1967). Some methods for classification and analysis of multivariate observations. *5th Berkeley Symposium on Mathematics, Statistics and Probability*, University of California Press, **1**, 281–298.

[37] Marriott, F. H. C.(1971). Practical problems in a method of cluster analysis. *Biometrics*, **27**, 501–514.

[38] Max, J.(1960). Quantization for minimum distortion. *IRE Transactions on Information Theory*, **IT-6**, 7–12.

[39] Menez, J., Boeri, F. & Esteban, D. J.(1979). Some methods for classification and analysis of multivariate observations. *1979 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 980–984.

[40] Milligan, G. W. & Cooper, M. C.(1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrica*, **50**, 159–179.

[41] Mosteller, F.(1946). On some useful 'ineffcient' statistics. *Annals of Mathematical Statistics*, **17**, 377–408.

[42] Muirhead, R. J.(1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

[43] Myers, R. H.(1986). *Classical and Modern Regression with Applications*, Duxbury, Boston.

[44] Ogawa, J.(1951). Contributions to the theory of systematic statistics. *Osaka Mathematical Journal*, **4**, 175–213.

[45] Ortega, J. M.(1972). *Numerical Analysis*, Academic Press, New York.

[46] Pärna, K.(1986). Strong consistency of k-means clustering criterion in separable metric spaces. *Acta et Commentationes Universitatis Tartuensis*, **733**, 86–96.

[47] Pärna, K.(1988). On the stability of k-means clustering in metric spaces. *Acta et Commentationes Universitatis Tartuensis*, **798**, 19–36.

[48] Pärna, K.(1990). On the existence and weak convergence of k-centers in Banach spaces. *Acta et Commentationes Universitatis Tartuensis*, **893**, 17–28.

[49] Pärna, K.(1991). Clustering in metric spaces : some existence and continuity results for k-centers. *Analyzing and Modeling Data and Knowledge*, Springer-Verlag, **86**, 85–91.

[50] Pearson, K.(1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**, 559–572.

[51] Pearson, K.(1920). On the probable errors of frequency constants. *Biometrica*, **13**, 113–132.

[52] Pollard, D.(1981). Strong consistency of k-means clustering. *The Annals of Statistics*, **9**, 135–140.

[53] Pratt, J. W.(1981). Concavity of the log likelihood. *Journal of the American Statistical Association*, **76**, 103–106.

[54] Ranga Rao, R.(1962). Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics*, **33**, 659–680.

[55] Rockafellar, R. T.(1970). *Convex Analysis*, Princeton, NJ : Princeton University.

[56] Rudin, W.(1973). *Functional Analysis*, McGraw-Hill, New York.

[57] Schwartz, J. T.(1969). *Nonlinear Functional Analysis*, Gordon and Breach, New York.

[58] Scott, A. J. & Symons, M. J.(1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387–398.

[59] Sharma, D. K.(1978). Design of absolutely optimal quantizers for a wide class of distortion measures. *IEEE Transactions on Information Theory*, **IT-24**, 693–702.

[60] Shimizu, N., Mizuta, M. & Sato, Y.(1997). Some Properties of Principal Points for Normal Distribution. *9th Korea and Japan Joint Conference of Statistics*, **9**, 257–262.

[61] Tarpey, T.(1994). Two principal points of symmetric, strongly unimodal distributions. *Statistics and Probability Letters*, **20**, 253–257.

[62] Tarpey, T., Li, L. & Flury, B.(1995). Principal points and self-consistent points of elliptical distributions. *The Annals of Statistics*, **23**, 1, 103–112.

[63] Trushkin, A. V.(1982). Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Transactions on Information Theory*, **IT-28**, 2, 187–198.

[64] Vassilyev, F. P.(1981). *Methods of Solution for Extremal Problems*, Nauka Publishers, Moscow.

[65] 浅野　亮, 水田正弘, 佐藤義治 (1998). 標本に基づく k-Principal Points の推定について. 第 66 回日本統計学会講演報告集, 266–267.

[66] 浅野　亮, 水田正弘, 佐藤義治 (1999). 様々な距離空間における主要点について. 日本行動計量学会第 27 回大会発表論文抄録集, 33–36.

[67] 伊理正夫, 腰塚武志 (1993). 計算幾何学と地理情報処理. 第 2 版, 共立出版.

[68] 大隅　昇 (1989). 統計的データ解析とソフトウェア. 放送大学教育振興会.

[69] 大隅　昇, L. ルバール, A. モリノウ, K. M. ワーウィック, 馬場康維 (1994). 記述式多変量解析法. 日科技連.

[70] 大瀧　慈, 藤越康祝 (1999). SIR-PPR アルゴリズムによる非線形回帰構造の探索. 科学研究費シンポジウム「非線形統計モデルとデータ解析」.

[71] 岡部篤行, 鈴木敦夫 (1992). 最適配置の数理. —シリーズ現代人の数理 3—, 朝倉書店.

[72] 清水信夫, 水田正弘, 佐藤義治 (1995). 2 変量正規分布における 3-Principal Points について. 情報処理北海道シンポジウム'95 講演論文集, 15–16.

[73] 清水信夫, 水田正弘, 佐藤義治 (1995). 3-Principal Points の性質について. 第 63 回日本統計学会講演報告集, 25–26.

[74] 清水信夫, 水田正弘, 佐藤義治 (1996). 最適配置における非対称性について. 情報処理北海道シンポジウム'96 講演論文集, 143–146.

[75] 清水信夫, 水田正弘, 佐藤義治 (1996). 対称な 1 変量分布における非対称な 3-Principal Points について. 第 10 回日本計算機統計学会大会論文集, 72–75.

[76] 清水信夫, 水田正弘, 佐藤義治 (1996). Principal Points が 3 以上の場合における配置について. 第 64 回日本統計学会講演報告集, 350–351.

[77] 清水信夫, 水田正弘, 佐藤義治 (1997). 2 変量標準正規分布における Principal Points について. 第 65 回日本統計学会講演報告集, 373–374.

[78] 清水信夫, 水田正弘, 佐藤義治 (1998). 単峰かつ対称な 1 変量分布における 3-Principal Points について. 第 66 回日本統計学会講演報告集, 264–265.

[79] 清水信夫, 水田正弘, 佐藤義治 (1998). Principal Points の性質について. 応用統計学, 第 27 巻第 1 号, 1–16.

[80] 清水信夫, 水田正弘, 佐藤義治 (1999). k-Principal Points の対称性に関する条件とその問題点について. 第 67 回日本統計学会講演報告集, 256–257.

[81] 清水信夫, 水田正弘, 佐藤義治. Principal Points の対称性に関する定理について. 計算機統計学, 第 12 巻第 2 号 (2000 年 4 月掲載予定).

[82] 水田正弘 (1994). Principal Points について. 第 62 回日本統計学会講演報告集, 260–261.

[83] 水田正弘 (1995). 対称分布における非対称な Principal Points について. 第 9 回日本計算機統計学会シンポジウム, 189–196.

[84] 水田正弘 (1999). 関数データ解析における主要点について. 第 67 回日本統計学会講演報告集, 355–356.

[85] 村木千恵, 大瀧　慈, 水田正弘 (1996). 極東における夏期天気図の分類. 第 64 回日本統計学会講演報告集, 227–229.

[86] 村木千恵, 大瀧　慈, 水田正弘 (1998). 主要点解析法による極東夏期天気図の解析. 応用統計学, 第 27 巻第 1 号, 17–31.

[87] 山本　渉, 篠崎信雄 (1997). 混合分布における 2 principal points の対称性. 第 65 回日本統計学会講演報告集, 375–376.

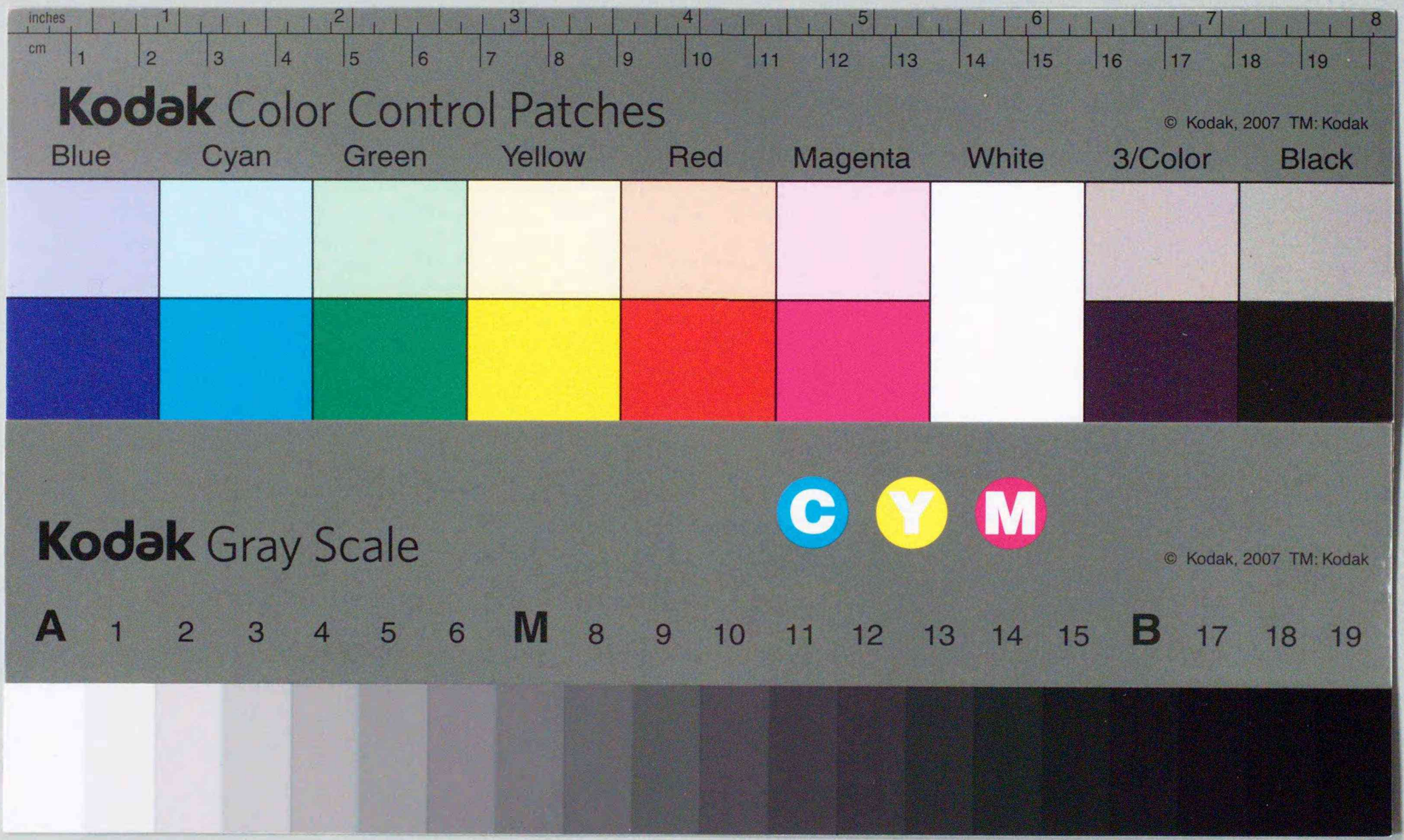
[88] 山本　渉, 篠崎信雄 (1999). 対称分布の混合における principal points について. 第 67 回日本統計学会講演報告集, 254–255.

謝辞

本研究を進めるにあたり、仔細にわたる御指導をいただいた北海道大学大学院工学研究科 システム情報工学専攻 数理情報工学講座情報解析学分野 佐藤 義治教授、北海道大学情報メディア教育研究総合センター 情報メディア科学基礎分野水田 正弘教授に深く感謝いたします。

本論文についてご討論いただき、貴重なご意見を頂いた北海道大学大学院工学研究科 システム情報工学専攻 新保 勝教授、伊達 淳教授に感謝いたします。また、分野雑誌会などを通じ本研究に際し有益なコメントをお寄せ下さった北海道大学大学院工学研究科 システム情報工学専攻 数理情報工学講座情報解析学分野 村井 哲也助教授、山本 義郎助手、北海道大学情報メディア教育研究総合センター 情報メディアシステム分野棟朝 雅晴助教授に感謝いたします。

本研究を行い本論文を作成するにあたって、北海道大学大学院工学研究科 システム情報工学専攻 数理情報工学講座情報解析学分野の学生諸氏からは有形無形の協力をいただきました。ここに感謝いたします。


inches
cm
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
Kodak Gray Scale
C
Y
M
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19