応用統計学 Vol. 27, No. 1 (1998)

研究論文

Principal Points の性質について

清水信夫

北海道大学 大学院工学研究科 水田正弘

佐藤義治

Some Properties of Principal Points

Nobuo SHIMIZU, Masahiro MIZUTA and Yoshiharu SATO

要 旨 Principal Points は，Flury (1990) により提案され，クラスター分析における k-means 法と同様の規準に基づいて，与えられた確率分布の密度関数を k 個の領域に分割する際に最適となるような各領域のある種の中心点として求められる．この Principal Points は，理論的に様々な興味深い性質をもつほか，応用的にもクラスター分析や最適配置の理論などと関連が深く，天気図の解析に応用する研究も行われている．

本論文では，対称性を有する確率分布が与えられた場合の Principal Points に関する理論的考察および計算機シミュレーションを行った．その結果，対称な1変量確率分布のうちロジスティック分布と両側指数分布に関しては 3-Principal Points が期待値に関し対称であることが示され，1変量混合正規分布については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，2変量が互いに独立な正規分布における分散共分散行列と k-Principal Points ($k \leq 5$) の関係について k-means 法と同様のアルゴリズムによる計算を行った結果，k-Principal Points の形が k 角形から直線に変わる分散共分散行列の境界値が求められた．さらに2変量標準正規分布における k-Principal Points ($k \leq 11$) の配置について考察した．

1. はじめに

今日，自然科学や社会科学のあらゆる分野で，複雑な現象を統計的に解析する必要性が増大し，多変量解析の各種の手法が広く用いられるようになってきた．その中で，クラスター分析は個体データをいくつかの群(クラスター)に分類する手法として出現し，発展してきた．最近では，計算機や各種計算機ソフトウェアが急激に進歩・充実し，分析に必要な複雑かつ大量の計算が高速に実行可能になり，分析結果を視覚化する手法も研究されている．

クラスター分析における代表的分析手法の1つに，k-means 法がある．k-means 法は，k 個の

Key words: cluster analysis, computational geometry, optimum location

データ点を各クラスターの重心の初期値として与えた上で，各データ点をデータ点からの距離が最も近い重心により作られるクラスターへ割り当て，さらに各クラスターの新たな重心の値に基づき個体データ点を再分類する作業を，クラスター間のデータ点の移動がなくなるまで繰り返す手法である．このアルゴリズムと同様の基準に基づいて，データ点の代わりに確率分布を与え，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ある種の中心点が，Flury (1990) により提案された Principal Points である．以下では，k 個の Principal Points を k-Principal Points とする．

Principal Points は，ある地域における最適な施設の配置を見出す「最適配置の理論（岡部・鈴木 (1992)）」への応用が可能であり，また，Principal Points を天気図の解析に応用する研究も行われている（村木他 (1996)）．

1 変量確率分布において，k-Principal Points が期待値に関して点対称となる場合の性質を「k-Principal Points の対称性」と呼ぶことにする．このとき，対称な 1 変量分布における k-Principal Points は対称性をもつことが予想される．しかし，Flury (1990) は確率分布の性質によっては期待値に関して非対称な 2-Principal Points が得られる場合があることを示した．また，2 変量確率分布において共分散が 0 である 2 変量正規分布では，一方の分散が他方よりも大きいと k-Principal Points ($k\leq 5$) が一直線上に並ぶ場合があることも示している．これらの性質は，主として数値計算によりいくつかのパラメータについての結果に基づいて考察されているが，さらに点が多い場合の値や対称性，配置など，より一般的な解析はあまりなされていない．

本論文では，対称性を有する連続な 1 変量確率分布が与えられた場合において，これまで研究されていない，3-Principal Points の対称性が成立する条件について，理論的考察および計算機シミュレーションの結果を報告する．また，2 変量確率分布において共分散が 0 である 2 変量正規分布の k-Principal Points ($k\leq 5$) を，様々な分散共分散行列について計算機シミュレーションにより考察し，k-Principal Points が一直線上に並ぶ分散共分散行列の値を求める．さらに，2 変量標準正規分布における k-Principal Points の配置に関するシミュレーションを行う．

2. Principal Points とは

この節では，Flury (1990) に従い，Principal Points の定義及び本研究に関連した k-Principal Points に関する従来の研究を紹介する．また，本研究で使用した k-Principal Points の導出アルゴリズムについても示す．

2.1. 定　義

Flury (1990) に従い，Principal Points の定義を述べる．

以下では，$\boldsymbol{y}_j\in R^p(1\leq j\leq k)$ を k 個の p 次元空間の点とし，$\boldsymbol{x}\in R^p$ と $\{\boldsymbol{y}_j\}$ との距離を

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

で定義する．

定義 (Principal Points)

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

$$E_F\{d^2(X|\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_k)\} = \min_{\boldsymbol{y}_j \in R^p, 1 \le j \le k} E_F\{d^2(X|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k)\} \tag{2}$$

が成立することである.

すなわち, X と $\boldsymbol{y}=(\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_k)$ との距離の2乗の期待値が最小となる $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_k$ が X の k-Principal Points となる.

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,

$$P_F(k) = \min_{\boldsymbol{y}_j \in R^p, 1 \le j \le k} E\{d^2(X|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k)\} \tag{3}$$

とおく. さらに, $E\{d^2(X|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k)\}$ を目的関数と呼ぶ.

2.2. Principal Points に関する従来の研究

与えられた確率分布が期待値に関して対称な1変量分布である場合において, Flury (1990) は 2-Principal Points に関する理論的考察を行い, 次の定理が成り立つことを示している.

定理 1.

Xが期待値 $\mu=E(X)$ をとり, 密度関数 $f(x)$ が μ に関して対称で, 2次のモーメントが有限である連続な1変量確率変数のとき,

$$y_1=\mu-E(|X-\mu|),\ y_2=\mu+E(|X-\mu|) \tag{4}$$

が $E\{d^2(X|\, y_1, y_2)\}$ の極小値をもたらす必要十分条件は

$$f(\mu)E(|X-\mu|)<\frac{1}{2} \tag{5}$$

である[1].

定理1は, 確率変数 X の期待値に関して対称な2点 y_1, y_2 が目的関数 $E\{d^2(X|y_1, y_2)\}$ の極小値を与えるために X が満たすべき式を示している. 前節の定義より, y_1, y_2 が 2-Principal Points であるとき, 目的関数は最小となるから, $f(\mu)E(|X-\mu|)>\frac{1}{2}$ の場合, 及び (4) 式で表される2点における目的関数が極小であっても最小とならない場合には, 2-Principal Points は期待値に関して非対称となる.

2-Principal Points が X の期待値に関して非対称となる例を示す. X が混合正規分布 $F(x)=(1-\varepsilon)N(x\,;0,1^2)+\varepsilon N(x\,;0,\alpha^2)$ に従う場合, $\varepsilon=0.5, \alpha=0.2$ のときには, 期待値に関して対称な2点が条件 (4) より $y_1=-0.47873, y_2=0.47873$ と求められるが, $f(0)E(|X|)=0.573>\frac{1}{2}$ となり y_1, y_2 は定理1における条件 (5) を満たさない (図1(a)). この場合の 2-Principal Points は, 期待値 (原点) に関して非対称な2点 $y_1=-0.232, y_2=1.032$ となる (図1(b)).

また, Flury (1990) により, 標準正規分布における k-Principal Points が与えられた場合, Principal Points が2個以上のときの数値は, 計算機シミュレーションにより表1のように求め

---

[1] しかし, 定理1の証明を厳密に検討すると, 正確には「(4) が $E\{d^2(X|y_1, y_2)\}$ の極小値をもたらす必要条件は $f(\mu)E(|X-\mu|)<\frac{1}{2}$」であり,「(4) が $E\{d^2(X|y_1, y_2)\}$ の極小値をもたらす十分条件は $f(\mu)E(|X-\mu|)\le\frac{1}{2}$」であることがわかる.

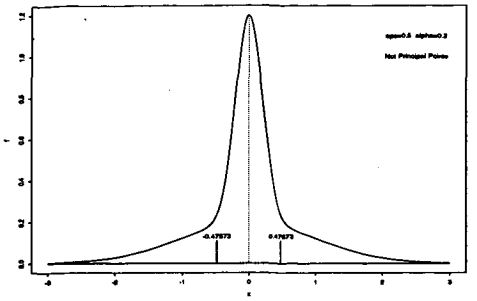

図 1(a). 混合正規分布における Principal Points とならない例 ($\varepsilon=0.5$, $\alpha=0.2$)

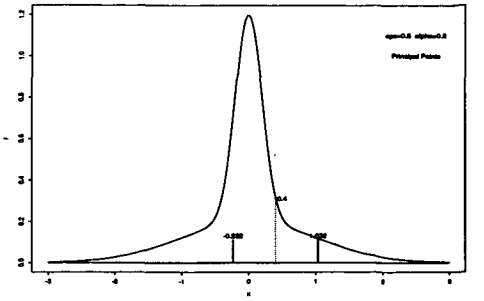

図 1(b). 混合正規分布における 2-Principal Points ($\varepsilon=0.5$, $\alpha=0.2$)

表 1. 1変量標準正規分布 ϕ における k-Principal Points (Flury (1990))

k	Principal Points	$P_\phi(k)$
1	0.0	1.0000
2	$-(2/\pi)^{1/2}$, $(2/\pi)^{1/2}$ ($\simeq\pm0.79788$)	$1-2/\pi$ ($\simeq0.3634$)
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

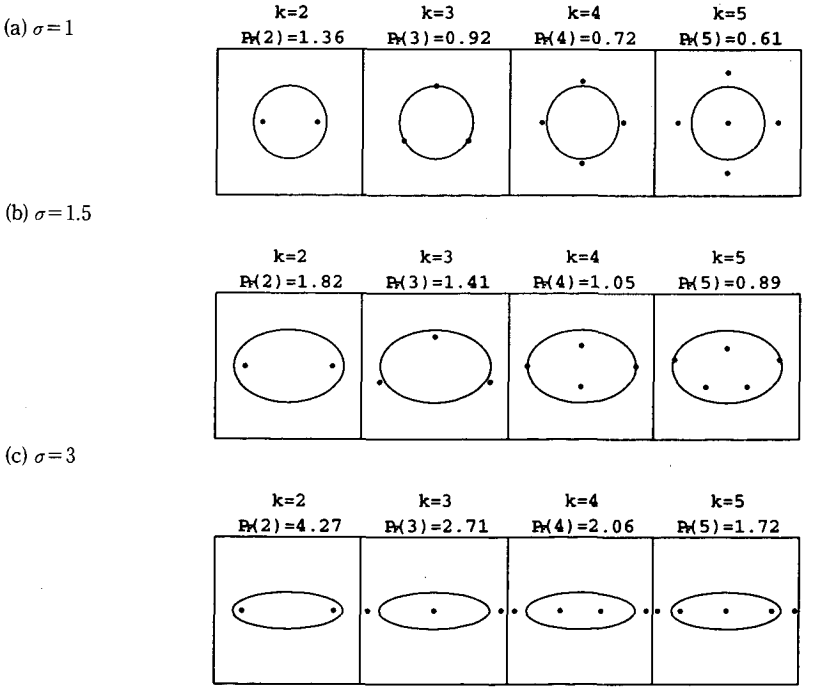

図 2. k-Principal Points の配置図 ($k=2, 3, 4, 5$) (Flury (1990) より一部修正, 追加)

られている．さらに，2変量正規分布において，共分散を0とし，2つの分散のうち1つを1に固定してもう1つの分散を変化させた時の k-Principal Points の形状の変化について，$\sigma=1$, $\sigma=1.5$, $\sigma=3$ の場合における k-Principal Points が，図2のような配置となることを示した．ただし，$\sigma=1.5$ における $k=5$ の場合については，Flury (1990) に誤りがあったため，正確な配置を計算して示している．

図2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，次のような問題が提示されている．

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？

k-Principal Points に関する他の理論的考察としては，期待値に関して対称な1変量確率変数 X における 2-Principal Points が，密度関数が強単峰 (strongly unimodal) であるときに一意に定まる (期待値に関して対称となる) ことが Tarpey (1994) や Li and Flury (1995) により示されている．しかし，多変量確率変数の場合及び k が大きい場合については，目的関数が複雑になるため，Principal Points の配置や確率変数に何らかの強い制約条件がある場合を除き，考察は非常に困難となる．そこで，前述のような 2-Principal Points が期待値に関して非対称となる場合を含め，理論的に k-Principal Points の値を導出することが困難な場合には計算機シミュレーションを用いて求める．

最近では，Principal Points の概念を他の分野に応用する研究も考えられている．例えば，確率密度関数を人口，k個の点を施設と考えることにより，ある地域における最適な施設の配置を見出す「最適配置の理論 (岡部・鈴木 (1992))」への応用が可能となる．また，村木他 (1996) は，多数の天気図パターンの中から数枚の代表的なパターンを抽出して分類する際に，天気図パターン全体を確率密度関数，代表的なパターンを Principal Points とみなして解析を行っている．

2.3. 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[2]

$$D_j=\{\boldsymbol{x}\in R^p \ ; \|\boldsymbol{x}-\boldsymbol{y}_j\|<\|\boldsymbol{x}-\boldsymbol{y}_l\|\} \text{(ただし } l\neq j)$$

をつくる (岡部・鈴木 (1992))．

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

---

[2] 各 $\boldsymbol{x}\in R^p$ において，最も近い点が $\boldsymbol{y}_j$ となるような領域である．

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

である.

(4) $\|\boldsymbol{g}_j-\boldsymbol{y}_j\|^2$ のうち,しきい値 ε 以上のものがある間 $\boldsymbol{g}_j\to\boldsymbol{y}_j$ として (2)～(3) を繰り返し,すべて ε より小さくなれば $\boldsymbol{g}_j\to\boldsymbol{y}_j$ 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 $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_k$ 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.

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 (1995)),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 (1994)),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.

3. 対称性を有する1変量分布における 3-Principal Points

Flury (1990) は,対称性を有する1変量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.

この節では,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 が期待値に関して対称となる条件が存在するかどうかを理論的に考察する.また,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 が理論的に考察された条件を満たすかどうかを検討し,さらに条件を満たさない場合の 3-Principal Points の値について,計算機シミュレーションにより求めた値を示す.

3.1. 理論的考察

対称性を有する1変量分布における 3-Principal Points について考察するとき,3-Principal Points の期待値に関する対称性については,期待値を原点としても一般性を失わない.よって,以下では,原点に関して対称な1変量分布の場合について述べる.

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．

ここで $y_1<y_2<y_3$ とし，$M(y_1, y_2, y_3)=E(d^2(X|y_1, y_2, y_3))$ を考察する代わりに，

$$c_1=(y_1+y_2)/2, \quad c_2=(y_2+y_3)/2$$
$$h_1=(y_2-y_1)/2, \quad h_2=(y_3-y_2)/2$$

とおくと，

$$\begin{aligned}M(y_1, y_2, y_3)&=E(d^2(X|y_1, y_2, y_3))\\&=\int_{-\infty}^{\infty}\min_{1\le i\le 3}(x-y_i)^2 f(x)dx\\&=\int_{-\infty}^{c_1}(x-c_1+h_1)^2 f(x)dx+\int_{c_1}^{c_2}(x-c_1-h_1)^2 f(x)dx+\int_{c_2}^{\infty}(x-c_2-h_2)^2 f(x)dx\\&=\sigma^2-4c_1h_1F(c_1)-4c_2h_2F(c_2)+(c_2+h_2)^2+4h_1\int_{-\infty}^{c_1}xf(x)dx+4h_2\int_{-\infty}^{c_2}xf(x)dx\\&=\sigma^2+2h_1(G(c_1)-G(c_2))+2(c_2-c_1)G(c_2)+(c_2-c_1)^2+(c_2-h_1)^2\end{aligned}$$

である．ただし，$h_2=c_2-c_1-h_1$ であり，

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

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，$\frac{\partial H}{\partial c_1}=0, \frac{\partial H}{\partial c_2}=0, \frac{\partial H}{\partial h_1}=0$ かつ H のヘッシアンが正定値である(ヘッ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，

$$\frac{\partial H}{\partial c_1}=2h_1G'(c_1)-2G(c_2)-2(c_2-c_1)=0 \tag{6}$$

$$\frac{\partial H}{\partial c_2}=2(c_2-c_1-h_1)G'(c_2)+2G(c_2)+2(2c_2-c_1-h_1)=0 \tag{7}$$

$$\frac{\partial H}{\partial h_1}=2(G(c_1)-G(c_2))-2(c_2-h_1)=0 \tag{8}$$

となるので，(8) 式より

$$h_1=-G(c_1)+G(c_2)+c_2 \tag{9}$$

が得られる．

H のヘッシアンが正定値である条件は，

$$\frac{\partial^2 H}{\partial c_1^2}=2+2h_1G''(c_1)>0 \tag{10}$$

$$\frac{\partial^2 H}{\partial c_1^2}\frac{\partial^2 H}{\partial h_1^2}-\left(\frac{\partial^2 H}{\partial c_1\partial h_1}\right)^2=4(1+h_1G'(c_1))-4(G'(c_1))^2>0 \tag{11}$$

$$\det D(c_1, c_2, h_1)>0 \tag{12}$$

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，(10) 式は，(11) 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

$$H(-c_2, c_2, c_2)=\sigma^2+4c_2G(c_2)+4c_2^2 \tag{13}$$

となる．(13) 式を $K(c_2)$ として c_2 で微分すると

$$\begin{aligned} K'(c_2)&=4G(c_2)+4c_2G'(c_2)+8c_2 \\ &=8\left\{\int_{-\infty}^{c_2} xf(x)dx+2c_2(1-F(c_2))\right\} \\ &=-8\int_{c_2}^{\infty}(x-2c_2)f(x)dx \end{aligned} \tag{14}$$

となる．ここで $K'(t)=0$ なる $t>0$ が存在すれば，(13) 式は $c_2=t$ で最小値をとり，(11) 式，(12) 式はそれぞれ

$$4(1-2tf(t))-4(1-2F(t))^2>0 \tag{15}$$
$$\det D(-t, t, t)>0 \tag{16}$$

となる．(15) 式，(16) 式より

$$tf(t)<2(1-F(t))(2F(t)-1) \tag{17}$$

である．

以上より，次の定理 2 が成り立つ．

定理 2.

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，t を

$$\int_{\mu+t}^{\infty}(x-\mu-2t)f(x-\mu)dx=0 \tag{18}$$

をみたす正の数とするとき，

$$y_1=\mu-2t,\ y_2=\mu,\ y_3=\mu+2t \tag{19}$$

が $E\{d^2(X|\, y_1, y_2, y_3)\}$ の極小値をもたらす必要条件は，

$$\frac{(\mu+t)f(\mu+t)}{(1-F(\mu+t))(2F(\mu+t)-1)}<2 \tag{20}$$

であり，十分条件は，

$$\frac{(\mu+t)f(\mu+t)}{(1-F(\mu+t))(2F(\mu+t)-1)}\leq 2 \tag{21}$$

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 (1991) 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 (21) 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. 種々の分布について

以下では，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，$E(X)=0$, $\sigma^2=\frac{\pi^2}{3}$ 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，

$$\begin{aligned}\frac{\partial H}{\partial c_1}&=2h_1(G'(c_1)-1)+2(c_1-G(c_1))\\&=\frac{4}{1+e^{-c_1}}\{c_1+(e^{-c_1}-1)\log(1+e^{c_1})+2\log(1+e^{-c_2})\}=0\end{aligned} \tag{22}$$

$$\begin{aligned}\frac{\partial H}{\partial c_2}&=2(c_2-c_1-h_1)(G'(c_2)+1)+2(c_2+G(c_2))\\&=\frac{4}{1+e^{c_2}}\{c_2+(1-e^{c_2})\log(1+e^{-c_2})-2\log(1+e^{c_1})\}=0\end{aligned} \tag{23}$$

である．(22) 式より，

$$\begin{aligned}c_2&=-\log\left[\exp\{-\frac{c_1}{2}-\frac{1}{2}(e^{-c_1}-1)\log(1+e^{c_1})\}-1\right]\\&=J(c_1)\end{aligned} \tag{24}$$

とかくことができ，これを (23) 式に代入して

$$\frac{\partial H}{\partial c_2}=\frac{4}{1+e^{J(c_1)}}\left[-2\log(1+e^{c_1})+J(c_1)-\frac{1}{2}\{c_1+(e^{-c_1}-1)\log(1+e^{c_1})\}(1-e^{J(c_1)})\right]$$
$$=0 \tag{25}$$

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 (25) 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．

なお，この分布における 2-Principal Points は ± 0.692 となることが水田 (1995) により示されている．

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）の密度関数として，$f(x)=\frac{1}{2}e^{-|x|}$ を考える．分布関数は

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

であり，$E(X)=0, \sigma^2=2$ 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，

$$H(-c_2, c_2, c_2)=\sigma^2-4c_2e^{-c_2}=K(c_2) \tag{26}$$

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

$$\frac{\partial H}{\partial c_1}=2h_1(G'(c_1)-1)+2(c_1-G(c_1))$$
$$=2e^{c_1}(1-h_1)=0 \tag{27}$$
$$\frac{\partial H}{\partial c_2}=2(c_2-c_1-h_1)(G'(c_2)+1)+2(c_2+G(c_2))$$
$$=2e^{-c_2}(c_2-c_1-h_1-1)=0 \tag{28}$$

であるから，(27) 式，(28) 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

$$H(c_2-2, c_2, 1)=\sigma^2-2(e^{-c_2}+e^{c_2-2})+(c_2-1)^2$$
$$=K(c_2) \tag{29}$$

の最小値を求めればよい．(29) 式より

$$K'(c_2)=2(e^{-c_2}-e^{c_2-2})+2(c_2-1) \tag{30}$$
$$K''(c_2)=-2(e^{-c_2}+e^{c_2-2})+2 \tag{31}$$
$$K'''(c_2)=2(e^{-c_2}-e^{c_2-2})>0 \tag{32}$$

であるから，(32) 式より $K''(c_2)$ は狭義単調増加となる．ここで

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

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

$$K''(c_2)\leq 0 \quad (0\leq c_2\leq u)$$
$$K''(c_2)>0 \quad (u<c_2<1)$$

となる．さらに

$$K'(0)=-2e^{-2}<0$$
$$K'(1)=0$$

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，(27) 式及び

$$\frac{\partial H}{\partial c_2}=2(c_2-c_1-h_1)(G'(c_2)+1)+2(c_2+G(c_2))$$
$$=2(c_2-c_1-h_1+1)(2-e^{c_2})+4(c_2-1)=0 \tag{33}$$

が成り立つ必要がある．(27) 式より $h_1=1$ だから，(33) 式に代入すると

$$c_1=c_2-\frac{2(1-c_2)}{2-e^{c_2}} \tag{34}$$

となる．(34) 式と (9) 式，及び $h_1=1$ を $H(c_1, c_2, h_1)$ に代入した式を $K(c_2)$ とすると

$$K(c_2)=\sigma^2+(c_2^2-1)+\frac{4(1-c_2)(c_2-e^{c_2})}{2-e^{c_2}}+\left\{\frac{2(1-c_2)}{2-e^{c_2}}\right\}^2 \tag{35}$$

である．これを c_2 で微分すると

$$K'(c_2)=\frac{8(1-c_2)^2e^{c_2}}{(2-e^{c_2})^3}+\frac{4(c_2-1)(e^{2c_2}-c_2e^{c_2}+2)}{(2-e^{c_2})^2}+\frac{4(1-c_2)+2c_2e^{c_2}}{2-e^{c_2}} \tag{36}$$

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

$$L(c_2)=8c_2+4(c_2^2-2c_2-1)e^{c_2}+2(2-c_2)e^{2c_2} \tag{37}$$
$$L'(c_2)=8+4(c_2^2-3)e^{c_2}+2(3-2c_2)e^{2c_2} \tag{38}$$
$$L''(c_2)=4(c_2-1)e^{c_2}(c_2+3-2e^{c_2}) \tag{39}$$

であり，さらに $r(c_2)=c_2+3-2e^{c_2}$ とおくと

$$r'(c_2)=1-2e^{c_2} \tag{40}$$

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，

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

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

$$L''(c_2)\geq 0 \quad (-\infty < c_2 \leq u)$$
$$L''(c_2)<0 \quad (u<c_2<0)$$

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

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

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．

なお，この分布における 2-Principal Points は ± 1 となることが水田 (1995) により示されている．

3.2.3. 混合正規分布

混合正規分布として，

$$F(x)=(1-\varepsilon)N(x;0, 1^2)+\varepsilon N(x;0, \alpha^2) \tag{41}$$

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．

一例として，$\varepsilon=0.88$, $\alpha=0.231$ 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 $\dfrac{tf(t)}{(1-F(t))(2F(t)-1)}\simeq 3.197>2$

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．

4. 2 変量正規分布における k-Principal Points

Flury (1990) は，共分散が 0 である 2 変量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．

この節では，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 ($k \leq 11$) 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．

4.1. 分散共分散行列の値と k-Principal Points

この節では，様々な σ について，2.2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3(a)～図3(c) に実線で示す．

これより，2.2節の問題 (a) において，

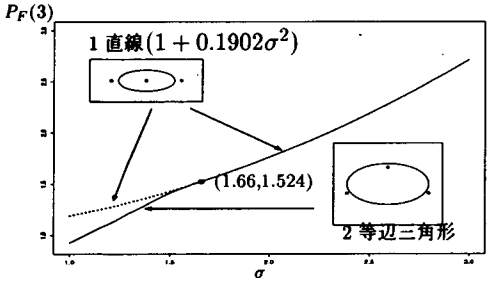

図3(a)．k-means 法による σ と $P_F(3)$ の関係図

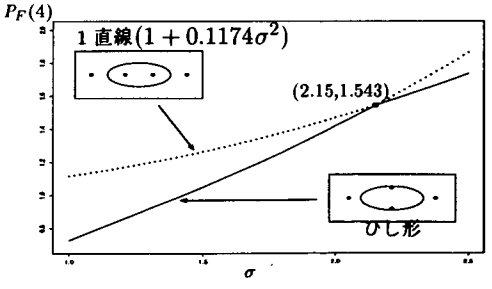

図3(b)．k-means 法による σ と $P_F(4)$ の関係図

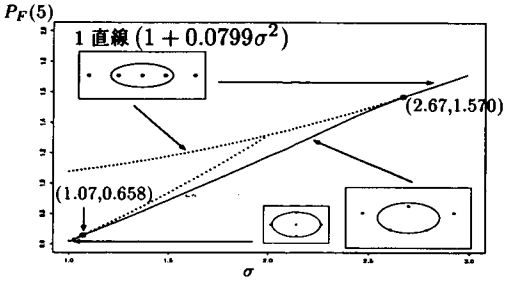

図3(c)．k-means 法による σ と $P_F(5)$ の関係図

(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，2.2 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．

4.2. k を大きくした場合の k-Principal Points

4.1 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．

4.2.1. 計算機シミュレーションによる k-Principal Points の配置

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 Flury (1990)

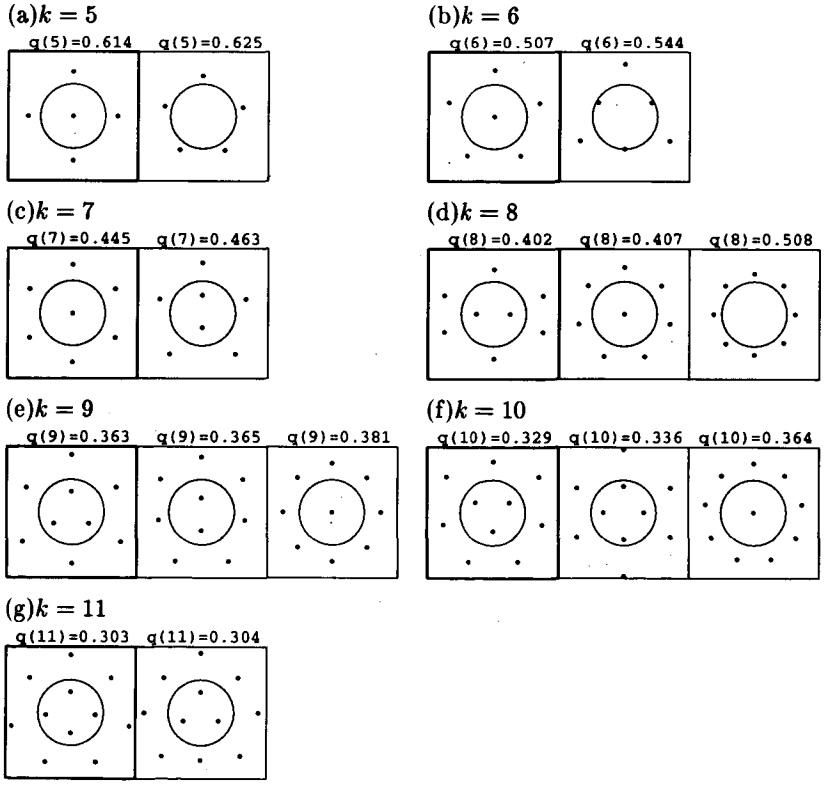

図 4. 2 変量標準正規分布における k 個の点の局所的最適配置図 ($k=5, 6, 7, 8, 9, 10, 11$)

と同様の配置が得られた．さらに，k が5以上11以下の場合において，現時点では図4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4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，最も外側に位置する l 個の点（ただし $l<k$）及びその内側の $(k-l)$ 個の点からなる配置になっており，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．また，どの最適配置も，原点を通る直線のうち少なくとも1本を線対称軸としてもつことがわかる．

5. おわりに

5.1. まとめ

本論文では，期待値に関して対称な種々の1変量分布及び2変量正規分布が与えられた場合の k-Principal Points について得られた性質を示した．

期待値に関して対称な1変量分布のうち，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，2変量正規分布における k-Principal Points は，2変量が互いに独立である正規分布の分散共分散行列 $\mathrm{diag}(\sigma^2, 1)$ 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で，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ところ，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2.2節の問題 (a) における $\sigma_0(k)$ の値となる．

さらに，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 $l<k$）及びその内側の $(k-l)$ 個の点からなる配置が得られた．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．これらの結果は，クラスター分析における妥当性 (validity) を研究する上での基礎資料になる．

5.2. 今後の課題

期待値に関して対称な1変量分布において，Principal Points が2個及び3個の場合については本論文及び Flury (1990) において研究がなされたが，4以上の Principal Points については詳細な研究が行われているとは言えず，今後の研究課題としたい．

また，2変量正規分布における k-Principal Points においても，Flury (1990) が提起した2.2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得られていない．さらに，2変量標準正規分布において，点の数と配置の形状に関するより一般的な規則性についても研究の余地があるものと思われる．これらの研究は，クラスター分析や最適配置の理論などとも関連があるので，応用も含めて研究を進める予定である．

参 考 文 献

Flury, B.A. (1990) : Principal points. *Biometrika* **77**, 1, 33-41.

Li, L. and Flury, B.A. (1995) : Uniqueness of principal points for univariate distributions. *Statistics and Probability Letters* **25**, 323-327.

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.

Tarpey, T. (1994) : Two principal points of symmetric, strongly unimodal distributions. *Statistics and Probability Letters* **20**, 253-257.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

(1997 年 3 月 18 日受付　1998 年 3 月 5 日最終修正)

著者連絡先 : 〒 060-8628 札幌市北区北 13 条西 8 丁目　北海道大学大学院工学研究科情報解析学分野
[redacted]