統計数理 (1993)
第41巻 第1号 1-11

# Principal Curvesと数量化III類を用いた質的データの1次元構造の抽出*

北海道大学** 水田正弘
統計数理研究所 馬場康維

(1993年5月 受付)

## 1. はじめに

数量化III類は,林 (1956) によって質的データのパターン分析を目的として提案された手法で,個体と特性によるデータ行列に関して,反応パターンの近いもの同士が隣合うように,行と列をそれぞれ並べ替える手法である.言い換えると,各個体と各特性ごとに1次元尺度を構成することによって順序づけを行う手法と見なせる.また,1次元尺度で順序づけが無理な場合でも2次元以上の尺度を考察することによりデータの有する構造を把握することが可能になる場合が多い.特に,ある種の典型的なデータ (上三角行列のデータや完全尺度構造を有するデータなど) に対しては,個体または特性の散布図が多次元空間内の曲線 (2次,3次,... 曲線など) の上に存在することが示されている (駒澤 (1982), 岩坪 (1987)).すなわち,本来,1次元構造を有するデータであってもそれが「曲がった1次元」であるため,多次元の尺度によって表現されるということである.数量化III類の基本に戻って,個体と特性に順序を付けるためには,多次元空間に存在する曲線に沿った順序で個体や特性を並べ替えることが考えられる.

多次元空間内にある曲線を扱うための手法の一つに,Hastie and Stuetzle (1989) によるPrincipal Curvesがある.この手法により,多次元空間内に存在する点に当てはまる曲線を見いだすことができる.さらに,得られた曲線とデータ点との最短距離および最短距離を実現する曲線上の点が得られる.そこで,本論では,Principal Curvesを数量化III類による布置に適用することにより,データの有する1次元構造を把握する方法について考察する (水田・馬場 (1990)).また,いくつかの数値例について本手法を適用し,得られた結果を散布図および,各変量の全ての組合せを描く対散布図により示す.

## 2. Principal Curvesについて

多変量のデータに曲線を当てはめることは,データ解析において最も基本的な問題の一つである.変量が説明変量と目的変量に分けることができる場合には,回帰直線やその拡張である様々な回帰曲線のための手法が利用できる.しかし,各変量が同等な場合には,これらの手法は利用できない.主成分分析や一般化主成分分析を利用することによって特定の曲線群 (例えば,直線,2次曲線など) で当てはめることは可能であるが,一般的な曲線で当てはめをするこ

* 本研究の一部は,統計数理研究所の「非線形構造を持つ多次元データの解析法の数理的研究」(統計数理研究所共同研究 (1-共研-10, 2-共研-14, 3-共研-4)) によりなされた.
** 工学部:〒060 札幌市北区北13条西8.

とはできない．そこで，Hastie and Stuetzle (1989) は，このようなデータに対して，曲線を当てはめるための理論として，Principal Curves の概念を提案するとともに，コンピュータによるアルゴリズムを示した．以下で，その内容を要約する．

理論的な Principal Curves の定義は，データ点ではなくデータの $p$ 次元の密度関数 $h(\boldsymbol{x})$ ($\boldsymbol{x}\in R^p$) が既知である場合に与えられている．$\boldsymbol{f}(\lambda)\in R^p(\lambda$: パラメータ) によって $p$ 次元空間内の曲線を表す．$p$ 次元空間の任意の点 $\boldsymbol{x}$ に対して，その点と最も近い曲線上の点を表すパラメータ $\lambda$ を $\lambda_f(\boldsymbol{x})$ とする．すなわち，

$$\lambda_f(\boldsymbol{x})=\sup_{\lambda}\{\lambda\,|\,\|\boldsymbol{x}-\boldsymbol{f}(\lambda)\|=\inf_{\mu}\|\boldsymbol{x}-\boldsymbol{f}(\mu)\|\}$$

である．ここで，$\lambda$ について sup をとるのは，ある点と最も近い曲線上の点が複数存在する可能性があることによる．

**定義**．曲線 $\boldsymbol{f}(\lambda)$ が，密度関数 $h$ の Principal Curve であるとは，

$$E_h(\boldsymbol{X}\,|\,\lambda_f(\boldsymbol{X})=\lambda)=\boldsymbol{f}(\lambda)\qquad(\text{for a.e. }\lambda)$$

が成立する場合をいう．

つまり，Principal Curve とは，曲線上の点 $\boldsymbol{f}(\lambda)$ に対して，その点に射影される $R^p$ の点の集合の期待値が $\boldsymbol{f}(\lambda)$ と一致するような曲線である（正確には，一致しない $\lambda$ の集合の測度が 0)．例えば，2 変量正規分布に対して，二つの主成分は共に Principal Curve になる．

Principal Curve を求めるには，以下のアルゴリズムを実行すればよい．

初期化：$\boldsymbol{f}^{(0)}(\lambda)=\bar{\boldsymbol{x}}+\boldsymbol{a}\lambda$ とおく．ただし，$\boldsymbol{a}$ は密度関数 $h$ の第 1 主成分とする．さらに，

$$\lambda^{(0)}(\boldsymbol{x})=\lambda_{f^{(0)}}(\boldsymbol{x})$$

とおく．以下，$j$ について繰り返す．

1. EXPECTATION STEP

$$\boldsymbol{f}^{(j)}(\lambda):=E(\boldsymbol{X}\,|\,\lambda_{f^{(j-1)}}(\boldsymbol{X})=\lambda)$$

2. PROJECTION STEP

$$\lambda^{(j)}(\boldsymbol{x}):=\lambda_{f^{(j)}}(\boldsymbol{x})\qquad(\text{for all }\boldsymbol{x}\in h)$$

さらに，$\lambda^{(j)}$ が曲線 $\boldsymbol{f}^{(j)}$ の弧長となるように変換する．

3. EVALUATION

$$\frac{|D^2(h,\boldsymbol{f}^{(j-1)})-D^2(h,\boldsymbol{f}^{(j)})|}{D^2(h,\boldsymbol{f}^{(j-1)})}$$

がある値（例えば，0.001）未満なら停止する．それ以上なら 2. に戻る．ただし，

$$D^2(h,\boldsymbol{f}^{(j)})=E_{\lambda^{(j)}}E_h\{\|\boldsymbol{X}-\boldsymbol{f}(\lambda^{(j)}(\boldsymbol{X}))\|^2\,|\,\lambda^{(j)}(\boldsymbol{X})\}$$

とする．

このアルゴリズムの EXPECTATION STEP では $\lambda_{f^{(j-1)}}(\boldsymbol{X})=\lambda$ を満たす $\boldsymbol{X}$ の集合の期待値をとり，その値によって $\boldsymbol{f}^{(j)}(\lambda)$ を更新する．PROJECTION STEP では，$R^p$ の各点を曲線 $\boldsymbol{f}^{(j)}(\lambda)$ に射影させることを意味する．

以上の定義およびアルゴリズムは，密度関数が与えられた場合のものである．しかし，実際の問題では，データ点が得られるだけであり，密度関数は未知である場合が大部分である．そこで Hastie and Stuetzle (1989) では，$n$ 個の $p$ 変量データ $x_{ik}(i=1, 2, \ldots, n, k=1, 2, \ldots, p)$ が与えられた場合のアルゴリズムも提案している．基本的なアルゴリズムの流れは同じであるが，曲線に制約があり，二つの STEP が少し異なる．

曲線 $\boldsymbol{f}(\lambda)$ は $n$ 個の点 $(\lambda_i, \boldsymbol{f}_i)$ によって決定される折れ線とする．PROJECTION STEP では，各データ点について，最も近い曲線（折れ線）上の点を探し，そのλの値を求める．EXPECTATION STEP では，各変量（各次元）$k$ ごとに $(\lambda_i, x_{ik})$ $(i=1, 2, \ldots, n, k=1, 2, \ldots, p)$ に注目し，$\lambda_i$ に対する $x_{ik}$ の平滑化を行い $\boldsymbol{f}^{(j)}(\lambda)$ を求める．Hastie and Stuetzle (1989) では，平滑化の手法として lowess などを利用している（大津 (1989)）．

## 3. 数量化 III 類への応用

数量化 III 類は，$n$ 個の個体が $M$ 個の特性について該当するかどうかというデータ行列，

$$D=\{\delta_{ij}\}$$

ただし，

$$\delta_{ij}=\begin{cases}1: \text{個体 } i \text{ が特性 } j \text{ に該当}\\ 0: \text{非該当}\end{cases}$$

に対して，個体に対応する座標（尺度），

$$X=\{x_i\}=(x_1, x_2, \ldots, x_n),$$

および特性に対応する座標（尺度），

$$Y=\{y_j\}=(y_1, y_2, \ldots, y_M),$$

を，$X$ と $Y$ との相関係数，

$$r_{XY}=\frac{\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{M}\delta_{ij}x_iy_j-\left\{\frac{1}{N}\sum_{i=1}^{n}\delta_{i\cdot}x_i\right\}\left\{\frac{1}{N}\sum_{j=1}^{M}\delta_{\cdot j}y_j\right\}}{\sqrt{\frac{1}{N}\sum_{i=1}^{n}\delta_{i\cdot}x_i^2-\left\{\frac{1}{N}\sum_{i=1}^{n}\delta_{i\cdot}x_i\right\}^2}\sqrt{\frac{1}{N}\sum_{j=1}^{M}\delta_{\cdot j}y_j^2-\left\{\frac{1}{N}\sum_{j=1}^{M}\delta_{\cdot j}y_j\right\}^2}}$$

が最大となるように定める方法である．ここで，$\delta_{i\cdot}, \delta_{\cdot j}$ はそれぞれ $\delta_{ij}$ の $j$ に関する和，$i$ に関する和を表す．$N$ は $\delta_{ij}=1$ となる $(i, j)$ の総数である．実際の計算は，固有値問題に帰着され，最大固有値に対応する固有ベクトルから，$X$, $Y$ が求められる．一つの固有ベクトル，すなわち 1 次元の尺度では解釈づけが不明瞭な場合には，他の固有ベクトルを固有値の値が大きい順に利用し，多次元の尺度を求める．

個体および特性をそれぞれ，多次元空間内の点として表すことにより，データが有する構造を見いだすことができる．ガットマンは多次元の軸を，内容 (content)，強度 (intensity)，確信度 (closure)，複雑度 (involution) と名づけた．しかし，岩坪 (1987) によるいくつかの典型的なデータに関する考察によると，これらの軸そのものが意味のある軸ではなく，布置の全体としての並び方が重要になる．例えば，1 次元構造をもつ 2-way (1, 0) データ（表 1）に数量化 III 類を適用して 4 次元までの尺度を計算すると，個体の布置，$X=\{x_i\}=(x_1, x_2, \ldots, x_n)$, は全て，4 次元空間 $(x_1, x_2, x_3, x_4)$ における曲線，

表1. 1次元構造をもつ2-way (1, 0) データ (岩坪 (1987)).

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $y_7$ | $y_8$ | $y_9$ | $y_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_2$ | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_3$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_4$ | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| $x_5$ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| $x_6$ | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| $x_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| $x_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| $x_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| $x_{10}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

$$\begin{cases} x_2=\sqrt{2}(x_1^2-1), \\ x_3=-2x_1^3+3x_1, \\ x_4=-\sqrt{2}(2x_1^4-4x_1^2+1), \end{cases}$$

の上に存在する．また，特性の布置，$Y=\{y_j\}=(y_1, y_2, \ldots, y_M)$，も同じ曲線上に存在する．また，同じく岩坪により示されたように，円環構造をもつデータについても，数量化III類によって個体および特性が閉じた曲線の上に布置される．

すなわち，円環構造を持つデータを含めて，1次元構造を有するデータは，ある種の曲線の近くに個体および特性が布置されると考えられる．そこで，前章で紹介したPrincipal Curvesによって，数量化III類の結果から得られた布置に曲線を当てはめることができる．さらに，Principal Curvesにより，個体および特性にそれぞれ順番を付けることができるので，その順番に従って個体および特性を並べ替えることにより，個体間および特性間の関係を調べることができる．次章ではいくつかの数値例を使って解析例を示す．

## 4. 数 値 例

数量化III類の結果に対し，Principal Curvesで曲線を当てはめることにより，データの構造を解析することができる．本章では，二つの数値例を示す．はじめの例は，人工データによるもので，個体と特性を並べ替えてデータ行列を見やすい形にするために，本手法を利用した例である．第2の例は，実際のデータについて，数量化III類を適用した結果にPrincipal Curvesを求めて全体の構造を示したものである．

### 4.1 数量化III類の結果に対する適用例（その1）

人工的な2-wayデータ（表2）に対して，個体と特性の並べ替えについて検討する．このデータは表1に個体を一つ付け加えたものである．すなわち，$\delta_{10,3}=1$であるために，正確な意味での1次元構造は壊されている．このデータに対して数量化による最適スコアの大小により，個体および特性の並べ替えをし，さらにPrincipal Curvesに沿った順序で個体と特性の並べ替えをした．両手法による並べ替えの比較を以下で行う．

はじめに，数量化III類により得られた，個体および特性の3次元の尺度を3次元空間内の点と見なして，それぞれPrincipal Curveを求めた．図1および図2に3次元空間を2次元ずつ表

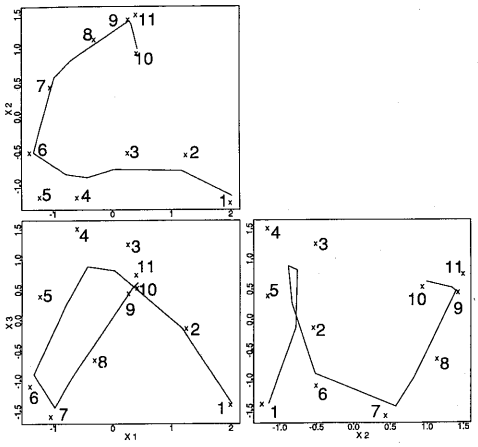

図 1.　個体の散布図と Principal Curves (データ 1).

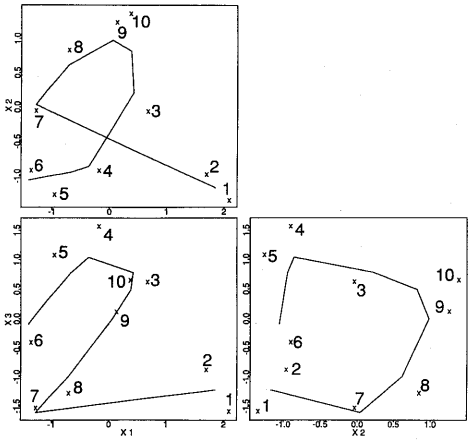

図 2.　特性の散布図と Principal Curves (データ 1).

表 2.　原データ (データ 1).

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $y_7$ | $y_8$ | $y_9$ | $y_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_2$ | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_3$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_4$ | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| $x_5$ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| $x_6$ | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| $x_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| $x_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| $x_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| $x_{10}$ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| $x_{11}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

表 3.　数量化のスコアによる並べ替え (データ 1).

| | $y_1$ | $y_2$ | $y_3$ | $y_{10}$ | $y_9$ | $y_4$ | $y_8$ | $y_5$ | $y_7$ | $y_6$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_2$ | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_{10}$ | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| $x_{11}$ | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_9$ | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| $x_3$ | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| $x_8$ | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| $x_4$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| $x_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| $x_5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| $x_6$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

表4. Principal Curvesによる並べ替え（データ1).

| | $y_1$ | $y_2$ | $y_7$ | $y_8$ | $y_9$ | $y_{10}$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_2$ | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| $x_3$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| $x_4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| $x_5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| $x_6$ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $x_7$ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_8$ | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| $x_9$ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| $x_{11}$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| $x_{10}$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |

示した散布図と Principal Curve を示す．データ数が10または11で少ないため，それほど滑らかな空間曲線とはなっていない．しかし，ここでの目的であるデータ点の順序づけには充分に利用できる．

数量化III類による1次元の尺度に従って，個体と特性を並べ替えたのが表3である．表4は，Principal Curveに沿った順序に従って個体および特性を並べ替えたものである．

数量化III類のアルゴリズムから表3が「個体の布置と特性の布置の相関係数が最大となる」のは当然であり，1が行列の対角部分の近くに分布している．それに比べて，表4は1が対角部分に集まるという性質はそれほど満たさないが，個体同士の関係および特性同士の関係を調べやすい．特に，1次元構造を乱している個体11を見つけ出すためには表4の方が役に立つ．

### 4.2 数量化III類の結果に対する適用例（その2）

前の（1, 0）データとほぼ同じ大きさでより複雑な構造を有する表5のデータ（岩坪(1987), p. 156）に対して同様な処理を試みる．このデータは個体が {1, 2, 3, 4}, {5, 6, 7}, {8, 9, 10} に分かれることが期待されるデータである．

数量化III類をこのデータに適用し，Principal Curveを求めた結果を図3，図4に示す．数量化III類による1次元の尺度に従って個体および特性を並べ替えた行列が表6，Principal Curveに沿った順序で並べ替えた行列が表7である．

表5. 原データ（岩坪（1987））（データ2).

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $y_7$ | $y_8$ | $y_9$ | $y_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| $x_2$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| $x_3$ | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| $x_4$ | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| $x_5$ | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| $x_6$ | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| $x_7$ | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| $x_8$ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| $x_9$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| $x_{10}$ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |

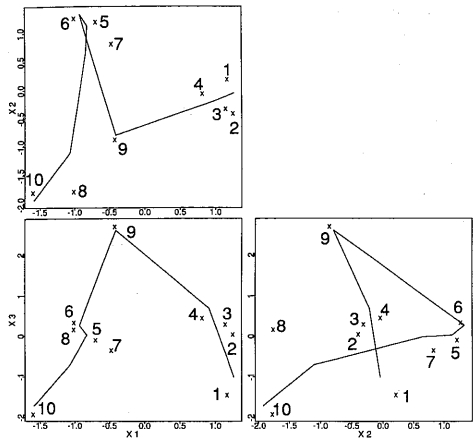

図 3.　個体の散布図と Principal Curves (データ 2).

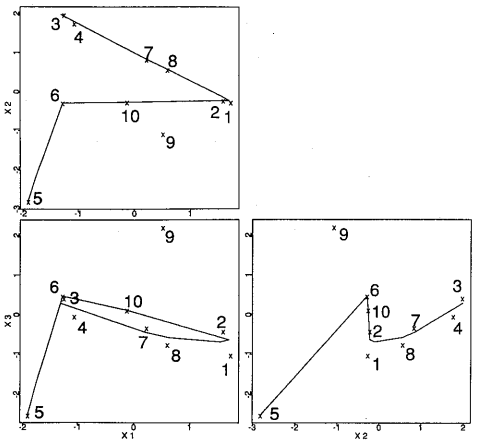

図 4.　特性の散布図と Principal Curves (データ 2).

表 6.　数量化のスコアによる並べ替え (データ 2).

| | $y_1$ | $y_2$ | $y_8$ | $y_9$ | $y_7$ | $y_{10}$ | $y_4$ | $y_3$ | $y_6$ | $y_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_2$ | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| $x_1$ | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| $x_3$ | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| $x_4$ | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| $x_9$ | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| $x_7$ | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| $x_5$ | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| $x_6$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| $x_8$ | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| $x_{10}$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

表 7.　Principal Curves による並べ替え (データ 2).

| | $y_3$ | $y_4$ | $y_7$ | $y_8$ | $y_1$ | $y_2$ | $y_{10}$ | $y_9$ | $y_6$ | $y_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| $x_2$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| $x_3$ | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| $x_4$ | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |
| $x_9$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| $x_6$ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| $x_5$ | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| $x_7$ | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| $x_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| $x_{10}$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |

表6から，個体を $\{1,2,3,4\}$ と $\{5,6,7,8,9,10\}$ に分けることを見いだせるが，$\{5,6,7\}$ および $\{8,9,10\}$ の集まりを見つけることは困難である．

表7から，$\{1,2,3,4\}$ と $\{5,6,7\}$ の集まりを見いだすことができる．しかし，$\{8,9,10\}$ の集まりを見いだすことは難しいと思われる．ただ，$\{8,10\}$ の類似性は認識しやすい．

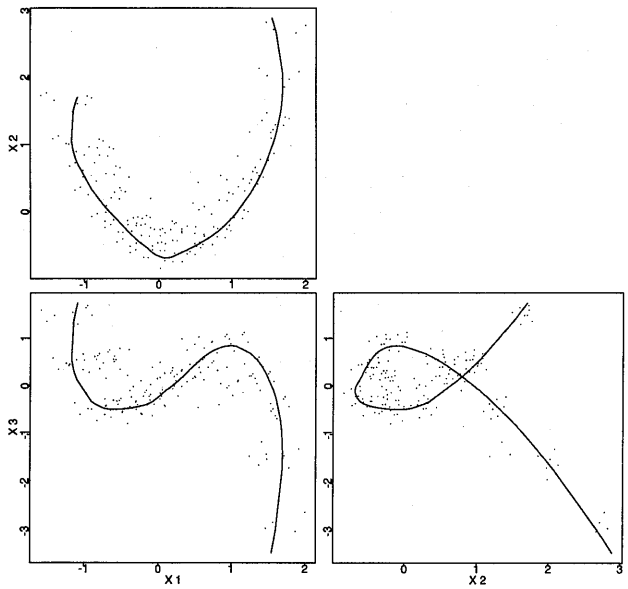


図5． 個体の散布図と Principal Curves.

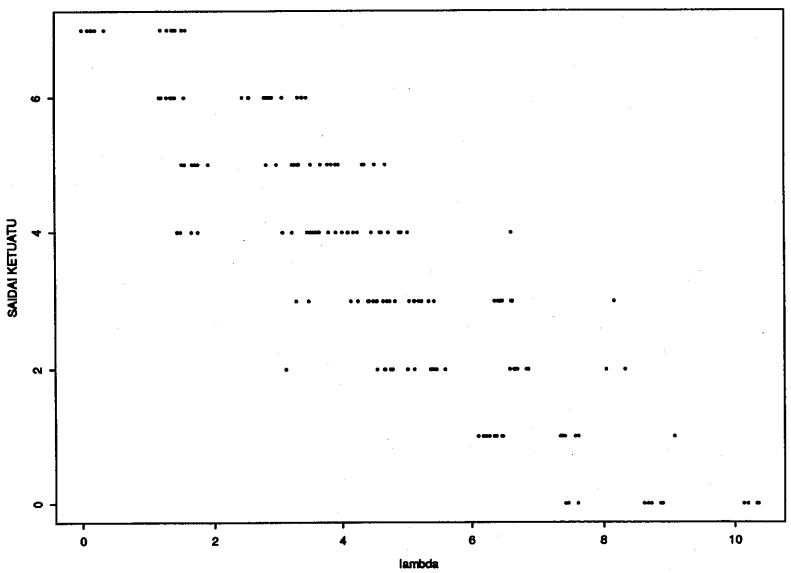


図6． パラメータ $\lambda$ と最大血圧．

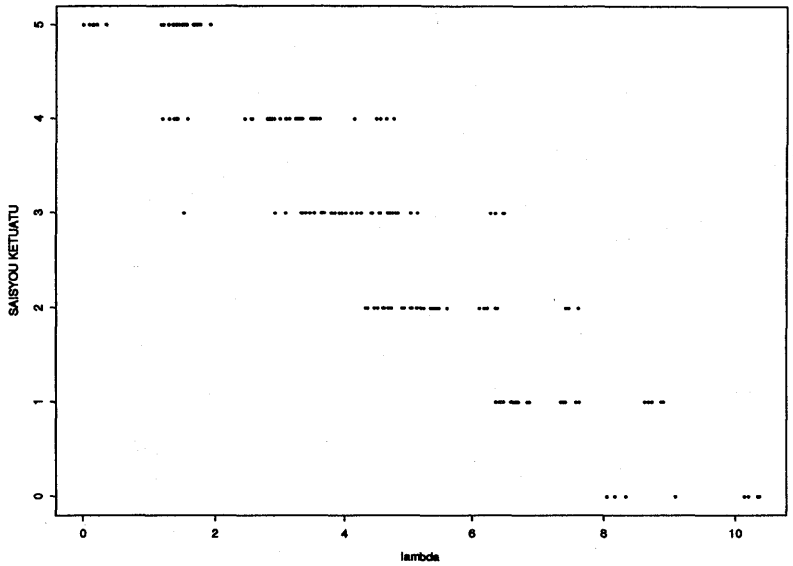


図 7. パラメータ $\lambda$ と最小血圧.

### 4.3 数量化 III 類の結果に対する適用例（実際のデータ）

Principal Curves の方法を実際のデータに適用した例を示す. 健康診断の結果の中から 3,335 人を選び, 年齢, 最大血圧, 最小血圧をカテゴリー化し, 数量化 III 類による分析を行った結果から 1,000 人分を抜き出し, それについて 1 軸-2 軸, 1 軸-3 軸等の個体スコアを対散布図としてプロットし, これに当てはまる Principal Curves を求めた（図 5). これにより, 3 次元空間内の点を当てはめる曲線（折れ線）が得られた.

ここで得られた Principal Curve のパラメータ $\lambda$ と原データにおける最大血圧の散布図（図 6）および $\lambda$ と最小血圧の散布図（図 7）を検討することにより, この曲線は血圧の高低による軸と解釈できることが分かる.

図 5 から分かるように, 本データに関する数量化の結果は, 2 次曲線, 3 次曲線の近傍に散らばっていると見なせる. すなわち, 本データは表 1 のデータと類似の性質を有しており, 1 次元構造を持っていることが分かる. また, Principal Curves によって, 各データ点から最短距離を有する曲線上の点も同時に求まる. 曲線の各点には, 曲線上の一定点（任意であるが, 固定する）から曲線に沿って測った弧の長さによるパラメータが自然に定義できるので, 各データ点に 1 次元の値を付与することができる. このパラメータにより, 曲線関係の意味を一層詳細に検討することができる.

なお, 本データは昭和 59 年度 統計数理研究所 特別研究費による研究「疾病の早期発見と予防に関する統計的現象解析：担当 駒澤 勉」によるものである.

## 5. おわりに

質的データに対して数量化 III 類を適用することにより, 個体および特性をそれぞれ多次元空間内の点として布置することができる. その空間内の分布の状態を解析することにより, データの有する構造を把握することができる. データにある種の順序関係（1 次元構造）が存在す

る場合には，布置された点が空間内のある曲線の近くに分布している場合が多い．しかし，高次元空間内の曲線構造を散布図，または対散布図で見いだすことは，容易ではない．そこで，空間内の点にPrincipal Curvesなどを当てはめることにより，データの特徴を理解しやすくなる．

一般にPrincipal Curvesを求めるには計算時間がかかる．Principal CurvesのアルゴリズムにおけるPROJECTION STEPの改良および計算量の評価については山下 他（1992）が考察しているが，アルゴリズムの改良の余地は残されていると思われる．さらに，Principal Curvesを求めるためには，平滑化のアルゴリズム，端点の処理（開曲線で当てはめるか閉曲線にするかなど）を選択しなくてはならない．この適切な選択法については今後の課題としたい．

## 6. 謝　　辞

各種データの提供等でご協力いただいた統計数理研究所 駒澤 勉教授に感謝いたします．また，Principal Curvesのプログラムを作成してくれた北海道大学 工学研究科 修士課程2年の山下信之君（現 日立製作所）に感謝いたします．さらに，丁寧な査読を通して貴重なご指摘とご意見を頂戴した査読者各位に感謝申し上げます．

## 参 考 文 献

Hastie, T. and Stuetzle, W. (1989). Principal curves, *J. Amer. Statist. Assoc.*, **84**, 502-516.

林知己夫（1956）．数量化理論とその応用例（II），統数研彙報，**4**(2)，19-30.

岩坪秀一（1987）．『数量化法の基礎』，朝倉書店，東京.

駒澤 勉（1982）．『数量化理論とデータ処理』（林知己夫監修），朝倉書店，東京.

水田正弘，馬場康維（1990）．Principal Curvesについて，第58回日本統計学会講演報告集，244-246.

大津展之（1989）．パターン情報処理における平滑化と近似，応用統計学会第11回シンポジウム「平滑化とその周辺」予稿集，65-72.

山下信之，南 弘征，水田正弘，佐藤義治（1992）．プリンシパル曲線のアルゴリズムの改良とその計算量の評価，計算機統計学，**5**(1)，33-43.

# Exploratory Method to Find Out a One-dimensional Structure of Qualitative Data by Using Principal Curves and Hayashi's Quantification Method III

Masahiro Mizuta

(Department of Information Sciences, Faculty of Engineering, Hokkaido University)

Yasumasa Baba

(The Institute of Statistical Mathematics)

In this paper we propose a method for detecting a one-dimensional structure contained in qualitative data by applying Hayashi's Quantification Method III and Hastie & Stuetzle's Principal Curves.

Quantification Method III proposed by Hayashi is useful for grasping the structure of qualitative data. One of the fundamental concepts of the method is to represent the categories in a low dimensional space in such a way that the categories with similar patterns are located near with each other. In particular, it is known that categories which are essentially one-dimensional are represented as points along a polynomial curve, when their locations are given by the scores assigned to the categories with the method. Therefore it is important in such cases to find curves which represent the trend of the configurations of the assigned scores.

Principal curve method proposed as an extension of principal components analysis is useful for fitting curves to the points in multi-dimensional spaces. Thus, by applying it to the configuration obtained with the quantification method, we can arrange the categories in a one-dimensional scale.

Key words: Quantification Method III, Principal Curves, one-dimensional structure.