第II部

線形ファジィクラスタリング

第5章

異なる次元の線形構造発見のためのファジィクラスタリング

ファジィクラスタリングの最近の研究では，それぞれのクラスタとして何らかの構造を仮定し，クラスタリングによってデータ構造の解析への応用が数多くなされている。主な構造としては，クラスタとして回帰や線形多様体という線形構造モデルを用いるもの，2次曲線や，球形，楕円型クラスタといった非線形モデルを用いるものまで様々な提案がなされている。第II部では，多次元空間上に分布するデータのモデリングを行うために，多次元空間上のデータの部分線形構造をとらえたファジィクラスタリングを考える。部分構造モデルは，人間が把握しやすい部分的線形構造とし，大規模なデータからのいくつかの線形部分空間の発見を目的とする。

第5章ではまず，これまでに提案されている線形構造発見のためのファジィクラスタリング手法について述べる。次に節では，それぞれのクラスタ内データの分布に応じた適切な次元の線形多様体を得ることをねらった目的関数の導入する。またクラスタリングを行うにあたり，パラメータの決め方やクラスタリングをどのように進めるかについて，人工的なデータのクラスタリング例により考察を行う。さらに5.6節では，どのクラスタにも分類されないデータのために Dave によって提案されたノイズクラスタ [16, 17] を本手法にも導入し，そのクラスタリング結果についての検討を行う。

5.1　はじめに

ファジィクラスタリングの最近の研究では，それぞれのクラスタとして何らかの構造を仮定し，クラスタリングによってデータ構造の解析への応用が数多くなされている。

クラスタとして非線形モデルを用いるものとしては，2次曲線で表現されるクラスタを考慮した FCQS(fuzzy c-quadric shell) アルゴリズムや，球形や楕円形クラスタを考える FCS(fuzzy c-shell) や AFCS(adaptive fuzzy c-shell) アルゴリズム等が提案されている。

クラスタとして線形モデルを用いる手法としては，大きく回帰モデルによるものと線形多様体モデルによるものが存在する。

線形構造モデルのアプローチとして，まず Gustafson and Kessel[40] の方法がある。彼らのファジィ分散共分散を利用したクラスタリング手法は，クラスタの形状を行列であらわし，その行列を用いた距離で評価規範を計算する。しかし，これにはクラスタの形状を事前に与えなければならないという問題がある。

次にデータの線形構造発見のためのクラスタリングとして，Bezdek らの fuzzy c-varieties 法 [41] がある。これはクラスタを線形多様体 (linear variety) であると仮定し，その線形多様体クラスタとデータセットとの距離を fuzzy c-means 法の距離として用いる方法である。たとえば2次元のデータであれば，クラスタを直線と仮定し，データとクラスタとの距離として点から直線への（垂直）距離を用いる。この方法では線形多様体は無限の広がりを持つため，遠く離れたデータが同じクラスタに含まれてしまうという問題がある。この点を補うため，さらに Bezdek らは，fuzzy c-varieties 法と fuzzy c-means 法の評価規範を重み付き平均した評価規範を用いる fuzzy c-elliptotypes 法 [42] を提案している。この手法の問題点は，その重み付けを全体のクラスタで1つの値を用いることができるか，またその値をどう決めるかということである。この問題に対し Dave らは adaptive fuzzy c-elliptotypes 法 [43] で，この重み付けをそれぞれのクラスタの形状に応じて固有値を用い変化させている。この手法は，クラスタに含まれるデータの分布に応じてそのクラスタの形状が決定されるため，あらかじめクラスタの形状に関する仮定をおかなくてもよく，異なる形状のクラスタが存在するときかなりうまくいく。この手法の問題点としては，複数の異なる次元のクラスタ

の発見が難しいことである。第7章では，この手法のさらなる問題点について指摘し，第I部で述べた正則化の概念を導入した別の定式化によるクラスタリング手法の提案を行う。

さらに，線形モデルによるクラスタリングには，直接回帰モデルをクラスタとして用いるアプローチも研究されている。Hathaway らは fuzzy c-regression models 法 (FCRM)[36] でスイッチング回帰モデルによるクラスタリングを提案している。回帰モデルとデータセットとの距離を，目的変数の回帰モデルによる値との誤差によって定める手法である。また中森らはこの手法に Dave らのアイデア Dave[43] を取り入れた，Adaptive Fuzzy c-Regression Models 法 [44] を提案している。

それとは別に前件部がファジィ命題，後件部が線形式で構成されたような，if-then ルールで構成されるファジィ推論モデルの構築に関する研究 [22] が高木・菅野らによって提起されている。線形回帰モデルを用いたファジィクラスタリングは与えられたデータの解釈であり，ファジィ推論モデルはそのシステムの振る舞いの予測に用いられる。菅野・姜 [45] のファジィ推論モデルの同定方法は，入力空間の分割と線形モデルの同定を同時に行うことができ，この研究の後，菅野・田中 [46] や中森・領家 [47] らによって多くの同定手法が提案されている。

これらの研究ではしかし，多次元上の異なる次元の構造を同定する手法については研究されてこなかった。本章ではこのような問題に対する新しい目的関数によって，異なる次元の線形構造モデルを発見するためのクラスタリング手法を提案する。さらに，Dave[16, 17] によって提案されているノイズクラスタを導入し，はずれ値やばらつきの大きなデータに対する効果をいくつかの人工的なデータを用いた結果を示す。しかし，このクラスタリング結果はパラメータや初期値によって左右されやすい。そこで，簡単な例を用いてクラスタリングの進め方についてのアイデアについて述べることにする。

5.2 これまでの線形ファジィクラスタリング手法

いま，クラスタリングすべき個体の集合を $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ $(\subset R^p)$，クラスタの数を c，クラスタ i の中心を $\boldsymbol{v}_i$ とし，u_{ik} を個体 $\boldsymbol{x}_k$ がクラスタ i に属する帰属度とする。こ

のとき，fuzzy c-means 法は，2.2 節で見てきたように，u_{ik} が以下の制約

$$M_{fcm} = \left\{ (u_{ik}) \;\middle|\; \sum_{i=1}^{c} u_{ik} = 1,\ u_{ik} \in [0,1],\ k \leq n \right\}$$

のもと，$U = (u_{ik}), V = (\boldsymbol{v}_i)$ に関する

$$J_{fcm}(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^q \ \|\boldsymbol{v}_i - \boldsymbol{x}_k\|^2 \tag{5.1}$$

を最小化する問題となるが，この目的関数を U と V について同時に最小化するのは困難である。そこで，最初に U を固定し V について，次に V を固定し U について $J_{km}(U,V)$ を最小化する2段階アルゴリズムを考える。[10]

CM 1. $\tilde{U}$ と $\tilde{V}$ の初期値を適当に与える。

CM 2. $\min_{V \in R^{pc}} J_{km}(\tilde{U}, V)$ の最適解を $\tilde{V}$ とする。

CM 3. $\min_{U \in M_{km}} J_{km}(U, \tilde{V})$ の最適解を $\tilde{U}$ とする。

CM 4. 最適解 $(\tilde{U}, \tilde{V})$ が収束していれば終了，そうでなければ CM 2 に戻る。

fuzzy c-varieties (FCV) 法[41]では，R^p における c 個の r 次元 $(0 \leq r < p)$ 線形多様体

$$V_i^r = \left\{ \boldsymbol{z} \in \boldsymbol{R}^p \ \middle|\ \boldsymbol{z} = \boldsymbol{v}_i + \sum_{j=1}^{r} t_{ij}\, \boldsymbol{e}_{ij}\ ,\ t_{ij} \in \boldsymbol{R} \right\} \tag{5.2}$$

を考える。ここで，$\boldsymbol{e}_{ij}\ \ (j = 1, 2, \cdots, r)$ をファジィ散布行列

$$S_i = \sum_{k=1}^{n} (u_{ik})^q \ (\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^{\mathrm{T}} \tag{5.3}$$

の固有値 $\lambda_{i1} \geq \lambda_{i2} \geq \cdots \geq \lambda_{ip} (\geq 0)$ に対応する固有ベクトルの最初の r 個を正規化したものとする。FCV 法では，評価規範として

$$J_{fcv}^r(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^q \ \ D_{ik}^r(\boldsymbol{v}_i) \tag{5.4}$$

を用いる。ただし，$D_{ik}^r(\boldsymbol{v}_i)$ は $\boldsymbol{x}_k$ と r 次元の線形多様体 V_i^r との2乗距離

$$\begin{aligned} D_{ik}^r(\boldsymbol{v}_i) &= \|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 - \sum_{j=1}^{r} |\langle \boldsymbol{x}_k - \boldsymbol{v}_i,\ \boldsymbol{e}_{ij} \rangle|^2 \\ &= \sum_{j=r+1}^{p} |\langle \boldsymbol{x}_k - \boldsymbol{v}_i,\ \boldsymbol{e}_{ij} \rangle|^2 \end{aligned} \tag{5.5}$$

である。ここで，$\langle \cdot, \cdot \rangle$ は内積を表す。

線形多様体 V_i^r は無限に広がっているので，2つ以上の離れたグループが同じくラスタに含まれることがある。このため fuzzy c-elliptotypes (FCE) 法[42]では，$D_{ik}^0(v_i)$（ユークリッド距離）と $D_{ik}^r(v_i)$ をパラメータ α を用いて加重平均することにより，距離とクラスタの線形性を同時に考慮した評価規範

$$J_{fce}^r(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^q \left[(1-\alpha)D_{ik}^0(v_i) + \alpha D_{ik}^r(v_i)\right] \tag{5.6}$$

を考える。

adaptive fuzzy c-elliptotypes (AFC) 法[43]では，α をクラスタリング過程においてクラスタごとに適応的に変化するパラメータ

$$\alpha_i = 1 - \frac{\lambda_{i,r+1}}{\lambda_{i1}}, \quad i = 1, 2, \cdots, c \tag{5.7}$$

とし，

$$J_{afc}^r(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^q \left[(1-\alpha_i)D_{ik}^0(v_i) + \alpha_i D_{ik}^r(v_i)\right] \tag{5.8}$$

を用いる。ただし，Dave[43]は2次元平面における直線の同定を扱っており，ここでは拡張して定式化している。

FCE法では，α をパラメータとして与えたのに対し，AFC法では，このパラメータをクラスタの形状に応じてクラスタごとに決定しようという発想である。

これによって，AFC法では，同じ次元の異なる形状の線形多様体に対応したクラスタリングを行うことができる。しかし，当時の Bezdek らの興味が画像認識にあり，2次元上のデータを主な対象として考えられていたため，多次元空間上の異なる次元の線形多様体を発見することはできない。

5.3　次元の異なる線形多様体発見法

AFC法では，データと r 次元線形多様体との2乗距離と，データとファジィ重心（0次元の線形多様体）との2乗距離を，データ分布を考慮して加重平均した規範を用いる。fuzzy

c-varieties of different dimensionalities (FVD) 法ではさらに，データとすべての次元の線形多様体との2乗距離をデータ分布を考慮して加重平均する規範を導入する。すなわち，つぎのような最小化すべき評価規範を考える。

$$J_{fvd}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^q \; E_{ik}(\boldsymbol{v}_i) \tag{5.9}$$

$$\text{(a)}\quad E_{ik}(\boldsymbol{v}_i) = \sum_{r=0}^{p-1} \beta_i^r \; \frac{D_{ik}^r(\boldsymbol{v}_i)}{p-r}$$

$$\text{(b)}\quad D_{ik}^r(\boldsymbol{v}_i) = \sum_{j=r+1}^{p} \| \langle \boldsymbol{x}_k - \boldsymbol{v}_i,\, \boldsymbol{e}_{ij} \rangle \|^2$$

$$\text{(c)}\quad \beta_i^r = \frac{\gamma_i^r}{\sum_{r=0}^{p-1} \gamma_i^r}$$

$$\text{(d)}\quad \gamma_i^r = \begin{cases} (\lambda_{im})^l, & r = 0 \\ (\lambda_{ir} - \lambda_{i,r+1})^l, & r > 0 \end{cases}$$

まず，(a),(b) について説明する。D_{ik}^r は p 次元空間における r 次元の線形多様体 V_i^r の直交補空間における2乗距離である。D_{ik}^r を直交補空間空間の次元数 $p-r$ で割ることにより，次元の増加による D_{ik}^r の自然増を押さえ，平等化を図っている。

$E_{ik}(\boldsymbol{v}_i)$ は，異なる次元のクラスタを見つけるために，D_{ik}^0 と特定の D_{ik}^r の2つだけでなく，すべての次元の線形多様体 V_i^r $(r = 0, 1, \cdots, p-1)$ を考え，データ $\boldsymbol{x}_k$ と V_i^r との2乗距離 D_{ik}^r に $1/(p-r)$ の重みづけをおこない，さらに以下で説明するクラスタの形状に応じた重み β_i^r をかけて加算する。

(c),(d) で定義されている β_i^r について説明する。β_i^r は (d) で定義されている γ_i^r に比例するように決める。γ_i^r は λ_{ir} と $\lambda_{i,r+1}$ の差によって定義され，この値が他の $\gamma_i^{r'}$ と比べて大きいということは，1番目から r 番目までの固有値にある程度の値があり，$r+1$ 番目以降

の固有値が r 番目の固有値よりかなり小さいことを意味しており，クラスタ i を r 次元の線形多様体として考えられる。このとき，評価規範の中で，データと r 次元の線形多様体との 2 乗距離に大きな重みを与える。これにより，r 次元の線形多様体を発見する可能性を高くする。ただし，β_i^0 については，最小固有値で定義する。

なお，(d) における l は線形度パラメータで，0 以上の実数である。l を大きくすることは線形性を強調されることを意味する。

本手法も AFC 法同様，各クラスタのファジィ散布行列の固有値を用いることによって，多次元空間上のデータからクラスタの形状を決定しようとするものである。

5.4 クラスタリング アルゴリズム

クラスタリングは，各パラメータ c, q, l を与え，目的関数を最小化するように変数 (U, V) を決定する過程によって行われる。第 I 部で見てきたように，クラスタリングは，U を仮定した V についての最適化と，V を仮定した U についての最適化によって行われる。

1. メンバシップ行列 $U = (u_{ik})$，ファジィ散布行列 S_i の固有値 $\{\lambda_{ij}\}$ 固有ベクトル $\{\boldsymbol{e}_{ij}\}$ が与えられたとする。このとき，V に関する目的関数は，

$$J_1(V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^q \; E_{ik}(\boldsymbol{v}_i). \tag{5.10}$$

となり，この目的関数が最適解となるための必要条件は，

$$\frac{\partial J_1}{\partial \boldsymbol{v}_i} = \sum_{k=1}^{n} (u_{ik})^q \left[\frac{\beta_i^0}{p} \frac{\partial D_{ik}^0(\boldsymbol{v}_i)}{\partial \boldsymbol{v}_i} + \sum_{r=1}^{p-1} \frac{\beta_i^r}{p-r} \frac{\partial D_{ik}^r(\boldsymbol{v}_i)}{\partial \boldsymbol{v}_i} \right] = \boldsymbol{0}. \tag{5.11}$$

となる。ここで，

$$\frac{\beta_i^0}{p} = a_i, \tag{5.12}$$

$$\sum_{r=1}^{p-1} \frac{\beta_i^r}{p-r} \left[I - \sum_{j=1}^{r} \boldsymbol{e}_{ij} \boldsymbol{e}_{ij}^{\mathrm{T}} \right] = A_i, \tag{5.13}$$

と置き，I を $p \times p$ 単位行列とすると，

$$(a_i I + A_i) \sum_{k=1}^{n} (u_{ik})^q \, (x_k - v_i) = 0 \tag{5.14}$$

と書ける。このとき，次の v_i はこの (5.14) 式を満たす。

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^q \, x_k}{\sum_{k=1}^{n} (u_{ik})^q}, \qquad i = 1, 2, \cdots, c \tag{5.15}$$

2. クラスタの中心 V と $\{\beta_i'\}$ が与えられたとする。このとき U に関する目的関数は，制約条件を付加した次のようなラグランジュ関数となる。

$$J_2(U) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^q \, E_{ik} + \sum_{k=1}^{n} \mu_k (\sum_{i=1}^{c} u_{ik} - 1), \tag{5.16}$$

ここで，$\mu_1, \mu_2, \cdots, \mu_n$ はラグランジュ乗数である。最適性の必要条件より，

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{E_{ik}}{E_{jk}} \right)^{\frac{1}{q-1}} \right]^{-1}. \tag{5.17}$$

を得る。ただし $E_{ik} = 0$ となる k が存在するときは，次式による。

$$u_{ik} = \begin{cases} \dfrac{1}{\#\{j \mid E_{jk} = 0\}}, & E_{ik} = 0 \\ 0, & E_{ik} > 0 \end{cases} \tag{5.18}$$

ここで，$\#\{\cdot\}$ は集合の要素数を表す。

クラスタリングのアルゴリズムを以下に記述する。

Step 1. $t = 0$ とし，パラメータ c, q, l, ε を定める。メンバシップ値の初期値 $U^{(0)} = \left(u_{ik}^{(0)} \right)$ を適当に与える。

Step 2. クラスタ i のファジィ重心を計算する。

$$v_i^{(t)} = \frac{\sum_{k=1}^{n} \left(u_{ik}^{(t)} \right)^q \, x_k}{\sum_{k=1}^{n} \left(u_{ik}^{(t)} \right)^q}, \quad i = 1, 2, \cdots, c$$

Step 3. クラスタ i のファジィ散布行列

$$S_i^{(t)} = \sum_{k=1}^{n} \left(u_{ik}^{(t)}\right)^q \left(\boldsymbol{x}_k - \boldsymbol{v}_i^{(t)}\right)\left(\boldsymbol{x}_k - \boldsymbol{v}_i^{(t)}\right)^{\top}$$

の固有値 $\left\{\lambda_{ij}^{(t)}\right\}$ と固有ベクトル $\left\{\boldsymbol{e}_{ij}^{(t)}\right\}$ を求める。ただし，$\|\boldsymbol{e}_{ij}^{(t)}\| = 1$ と正規化する。

Step 4. 次式により，データ $\boldsymbol{x}_k$ のクラスタ i へのメンバシップ値を更新する。

$$u_{ik}^{(t+1)} = \left[\sum_{j=1}^{c}\left(\frac{E_{ik}^{(t)}}{E_{jk}^{(t)}}\right)^{\frac{1}{q-1}}\right]^{-1}$$

$E_{ik}^{(t)} = 0$ となる i が存在する場合は次式による。

$$u_{ik}^{(t+1)} = \begin{cases} \dfrac{1}{\#\{j \mid E_{jk}^{(t)} = 0\}}, & E_{ik}^{(t)} = 0 \\ 0, & E_{ik}^{(t)} > 0 \end{cases}$$

ただし，$\#\{\cdot\}$ は集合の要素数を表す。

Step 5. 収束判定条件

$$\max_{i,k}\left\{\left|u_{ik}^{(t+1)} - u_{ik}^{(t)}\right|\right\} < \varepsilon$$

を満たせば終了。そうでなければ，$t = t+1$ として **Step 2** へ戻る。

以上の手法を Fuzzy c-Varieties of Different Dimensionalities (FVD) 法と呼ぶことにする。

5.5 数値例

この節ではクラスタリングがどのような過程で行われるかを，人工的なデータと簡単な実データによる例題によってを示す。

・クラスタリング例 1

最初の例は，クラスタリング過程の説明のための 3 次元上の人工的なデータである。図 5.1 はパラメータとして

$$c = 4,\ q = 1.8,\ l = 0.8,\ \varepsilon = 0.0001$$

を用いたときのクラスタリング結果を示している。各点はメンバシップ値が最も大きいクラスタに分類されている。図 5.2 に，この結果を得るまでの評価関数 J_{fvd} と

$$\delta = \max_{i,k} \left\{ \left| u_{ik}^{(t+1)} - u_{ik}^{(t)} \right| \right\}$$

のクラスタリング過程における値の変化する様子を示す。横軸はクラスタリングのステップ数を表し，左側の軸は δ の，右側の軸は J_{fvd} に対する座標軸をそれぞれ表している。

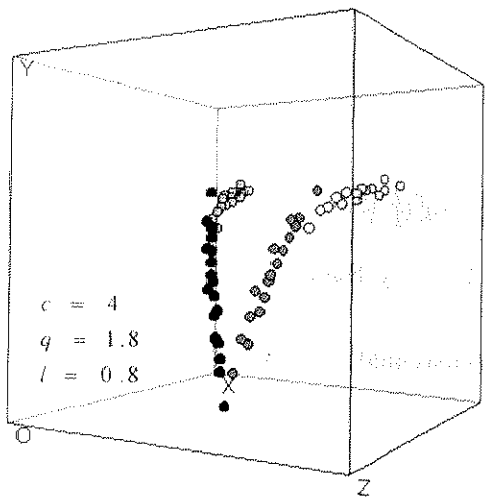

図 5.1: 人工的なデータに対するクラスタリング結果

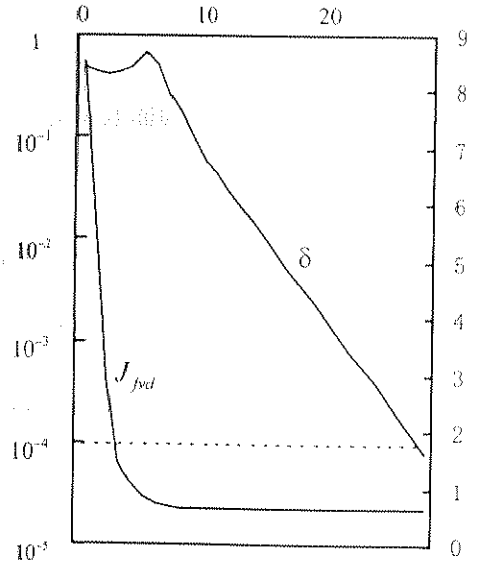

図 5.2: 図 5.1 のクラスタリング過程での J_{fvd} と δ の値の変化

クラスタリング結果は，4 つの直線状のクラスタとなり，J_{fvd} の値もよい値を示している。図 5.3，図 5.4 は，別の角度から見たクラスタリング結果を示している。3 次元までであれば，このようにクラスタリング結果の妥当性を視覚的に見て取ることができる。

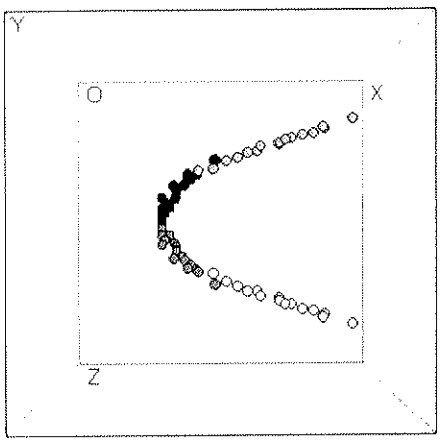

図 5.3: 図5.1を上から見た図

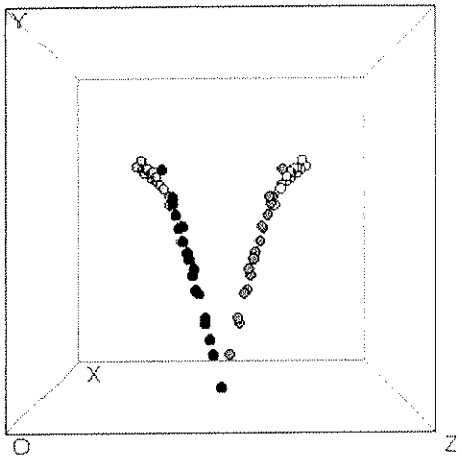

図 5.4: 図5.1を正面から見た図

一般にクラスタリング前にクラスタ数をいくつにするかはむつかしい問題であるが，クラスタ数 c を他の値にしたときの様子を見てみよう。クラスタ数としてそれぞれ $c=5$, $c=4$, $c=3$, $c=2$ としたときのクラスタリング結果を図5.5，図5.6，図5.7，図5.8にそれぞれ示す。クラスタリングは，$c=5$ によってはじめ，図5.5の結果を得た後，以下のようにしてクラスタ数を1つずつ減少させている。

- 各クラスタ C_i に対し，メンバシップ値の合計を計算する。

$$z_i = \sum_{k=1}^{n} u_{ik}, \quad i = 1, 2, \cdots, c \tag{5.19}$$

- 次に，この z_i のいちばん小さなクラスタ C_i のメンバシップを 0 とし，

$$u_{ik} = 0, \quad k = 1, 2, \cdots, n \tag{5.20}$$

- クラスタ数 c を $c-1$ としてクラスタリングを続ける。

このとき，特定のクラスタのメンバシップ値を 0 とする事で，メンバシップ値に関する制約が一時的に破られるが，次のクラスタリングの繰り返しによってこの制約は満たされるようになる。

このときの評価関数 J_{fud} の値の変化を図 5.9 に示す。評価関数 J_{fud} の値が $c=4$ と $c=3$ のときとの間で大きく変化しており，図 5.5 から図 5.8 の結果を見てわかるようにクラスタ数として $c=4$ が適当であることがわかる。

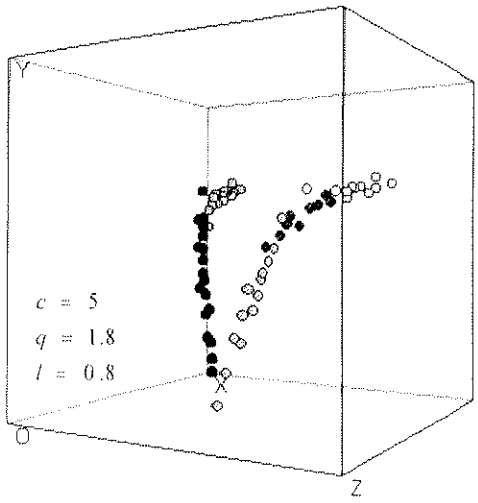

図 5.5: $c=5$ によるクラスタリング結果

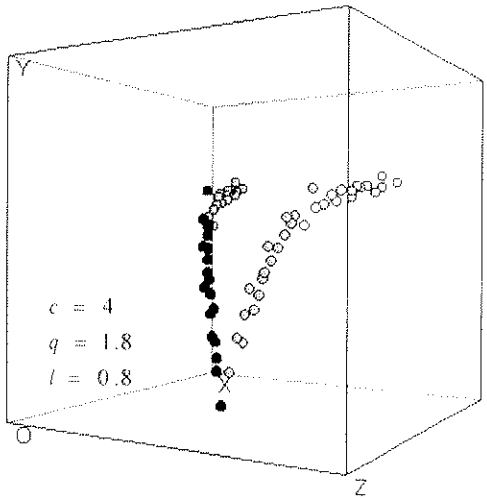

図 5.6: $c=4$ による図 5.1 と同じ結果

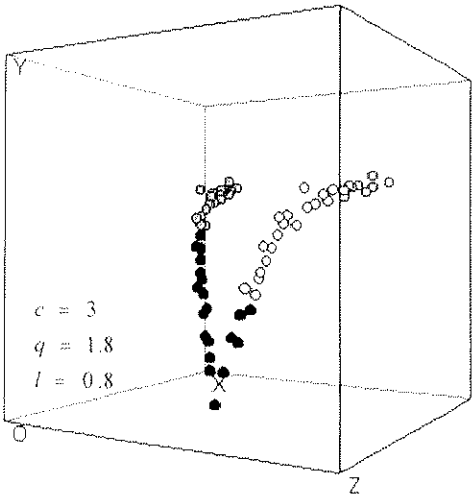

図 5.7: $c=3$ によるクラスタリング結果

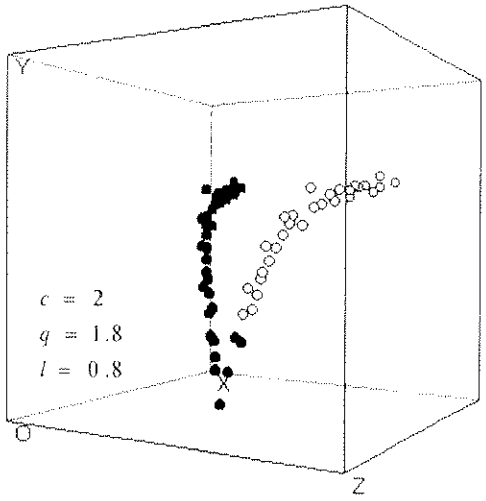

図 5.8: $c=2$ によるクラスタリング結果

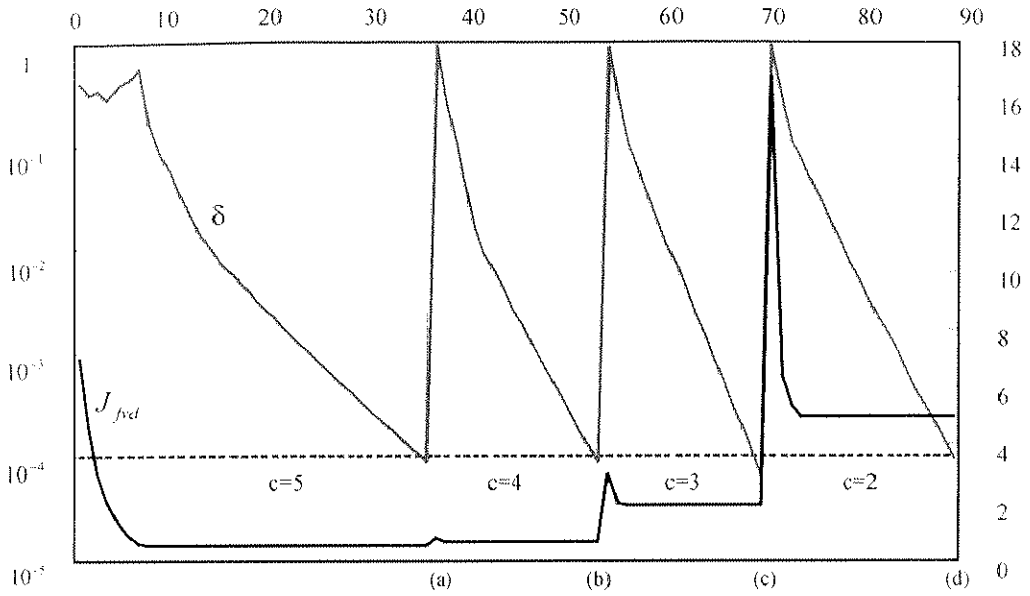

図 5.9: 図 5.5, 5.6, 5.7, 5.8 のクラスタリング過程を通しての J_{fvd} と δ の値の変化の様子

・クラスタリング例 2

次の例では，文部省が発表した 1950 年から 1996 年の間の中学 1 年から高校 3 年生の身体測定の男女別データである。変数は，身長，体重，座高が各年度の各学年の性別平均データを用いた。図 5.10 の X，Y，Z 軸がそれぞれ，身長，体重，座高を表している。

このデータのクラスタリング結果を図 5.11 に示す。クラスタ数は 2 で，各点はメンバシップ値の大きい値のクラスタに分類して別の色で示されている。この結果濃いマークのクラスタの各点はすべて男子のでーたであり，薄いグレーのクラスタの約 80% が女子のデータである。このようにデータ全体が性別によってほぼ 2 つのグループに分けられている。薄いグレーのクラスタに含まれる男子の約 20% は，男子と女子のクラスタの重なり合う部分にあり，体が小さいときにはこれらの値からでは区別できないことがわかる。データ全体としても各変数は強い相関を持っているが，クラスタリング結果は，男子と女子の各変数の相関に違った傾向があることをとらえている。これから体が大きくなるにしたがい，身長と座高が同じであれば，女子は男子よりも体重が一般的に重いということが観測でき，男子と女子はそれぞれ異なる 2 つの直線上に分布していることがわかる。

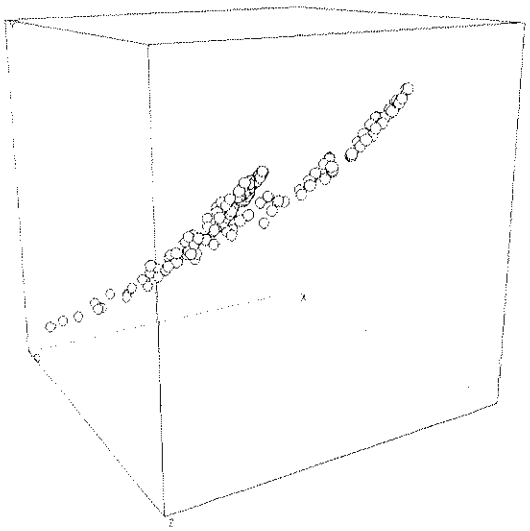

図 5.10: 中学生高校生の男女別身体測定データ

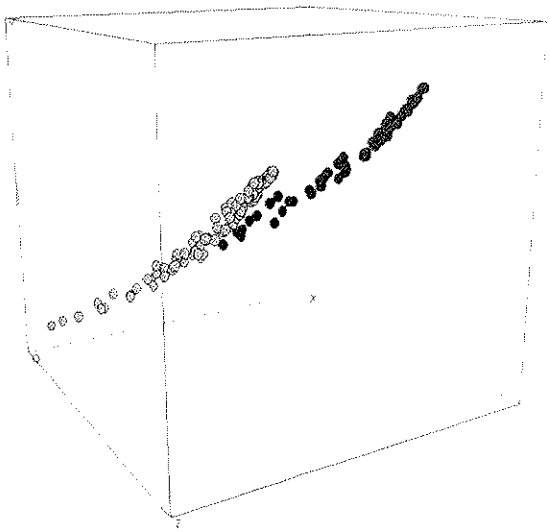

図 5.11: 身体測定データのクラスタリング結果

5.6　ノイズクラスタの導入

統計解析には，一般にはずれ値の扱いに関する問題がある。これはクラスタリングにおいても重要な問題である。Dave[16, 17] はこの問題を扱うための新しいクラスタリングの方法を提案している。一般のクラスタとは別に，はずれ値のためにノイズクラスタと呼ばれる特別なクラスタを準備してクラスタリングを行うのである。本章では Dave のこのアイデアを FVD 法にも適用し，分散の大きなデータの扱いに対しても有効であることを示す。

ノイズクラスタとは，通常のクラスタ以外に用意された特別なクラスタで，他のどのクラスタにも含まれない対象のためのものである。クラスタ $c-1$ までを通常のクラスタ，クラスタ c をノイズクラスタとする。このとき，ノイズクラスタ c と対象 x_k との距離 E_{ck} は，

すべての対象に対して等しく，

$$E_{ck} = \lambda \left[\frac{\sum_{i=1}^{c-1} \sum_{k=1}^{n} (E_{ik})^2}{n(c-1)} \right] \tag{5.21}$$

と定められる。ここで，λ は正の定数である。λ の値が小さいとほとんどの対象がノイズクラスタに含まれ，反対に λ の値が大きいとノイズクラスタに入る対象はほとんどなくなりノイズクラスタを用いないときと同じ結果となる。

式 (5.21) によって定められる距離 E_{ck} は，4 節のアルゴリズムにおいて， Step 4 の前に毎回計算される。

5.7 ノイズクラスタリングの例

クラスタリング例 3

図 5.12 はノイズクラスタを用いたクラスタリングのために準備された人為的なデータである。データは 3 つの部分で構成されている。図 5.13 に示されるように，2 つの互いに平行なの円盤状の部分 A, B と，A, B に直交する直線状の部分 C である。A, B はそれぞれ 100 個の点からなり，x 軸成分，y 軸成分は標準偏差 1.0 の，z 軸成分は標準偏差 0.1 の正規乱数によって生成されている。C も 100 個の点からなり，x, y, z 軸成分は，それぞれ 0.1, 0.1, 1.0 の標準偏差の正規乱数によって生成されている。

このデータに対して，ノイズクラスタを用いないときと，用いたときの結果を図 5.14, 5.15 に示す。図の中で，薄いグレーの点，濃いグレーの点は，各クラスタへのメンバシップ値が 0.5 以上となったことを示す。ただし，薄いグレーの点は，2 つのクラスタに用いられている。図中の白い点は，どのクラスタへのメンバシップ値が 0.5 以下であり，図 5.15 の中の黒い点は，ノイズクラスタにメンバシップ値 05. 以上で属していることを表す。

クラスタリング結果は，ノイズクラスタのあるなしに関わらずうまく 3 つのクラスタに分

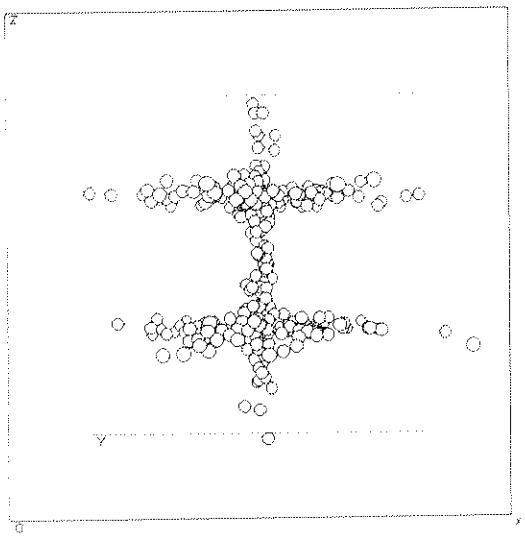

図 5.12: 3次元上の人工的なデータ

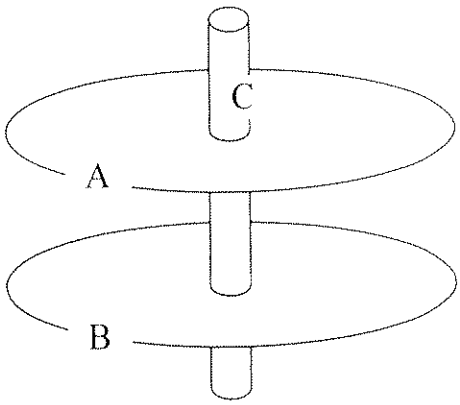

図 5.13: 図5.12のデータ分布の概念図

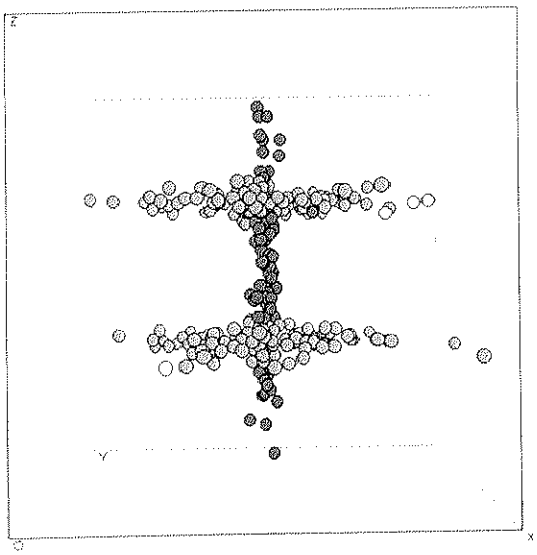

図 5.14: 図5.12へのノイズクラスタを用いないクラスタリング結果

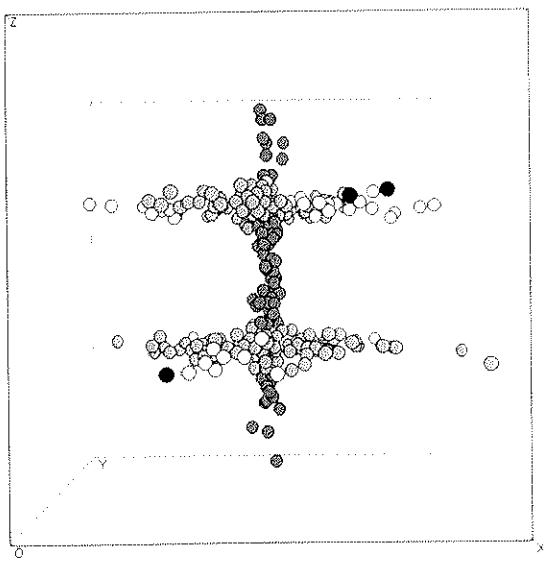

図 5.15: 図5.12へのノイズクラスタを用いたクラスタリング結果

類されていることが見てとれる。クラスタリングのメンバシップ値の初期値は乱数によって定められ，クラスタリングのパラメータとして $q = 1.8, l = 0.8, \varepsilon = 0.0001, \lambda = 0.8$ が用いられた。次の例 4, 例 5 でも同じ初期値とパラメータが用いられている。

クラスタリング例 4

次に図 5.16 は図 5.12 に，50 個のノイズデータを加えたデータである。50 個のノイズは x, y, z の各軸方向に標準偏差 1.0 の正規乱数によって生成された。図 5.17，図 5.18 はこのデータに対して，ノイズクラスタをそれぞれ用いないときと用いたときのクラスタリング結果である。メンバシップの初期値，クラスタリングパラメータは例 3 と同じものを用いている。

ノイズクラスタを用いないクラスタリング結果では，図 5.17 のようにノイズによる影響によって直線状のクラスタが大きく乱されていることがわかる。しかしノイズクラスタを導入したクラスタリング結果では，図 5.18 のようにこの影響が低く抑えられていることがわかる。

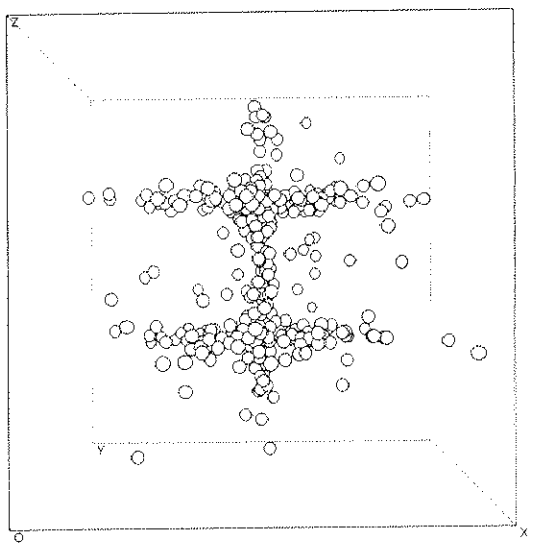

図 5.16: 図 5.12 にノイズを加えたデータ

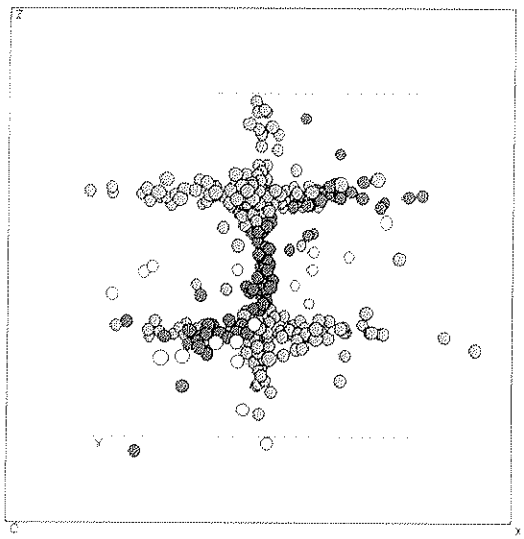

図 5.17: 図 5.16 へのノイズクラスタを用いないクラスタリング結果

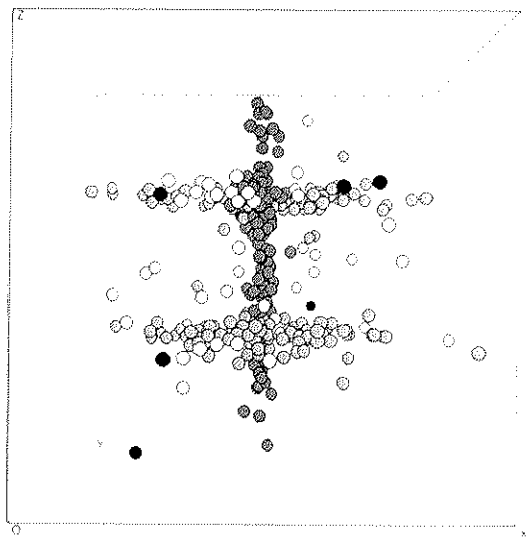

図 5.18: 図 5.16 へのノイズクラスタを用いたクラスタリング結果

クラスタリング例 5

図 5.19 は，図 5.12 のデータの円盤状の部分 A, B の z 軸方向の標準偏差を例 3 の 2 倍の 0.2 の正規乱数によって生成したデータである。図 5.20，図 5.21 はこのデータに対して，ノイズクラスタをそれぞれ用いないときと用いたときのクラスタリング結果である。

クラスタリング結果は例 4 のときと同様，ノイズクラスタを用いない結果では図 5.20 のように直線状のクラスタが大きく乱されているが，ノイズクラスタを用いた結果では図 5.21 のように分散の影響が低く抑えられている。

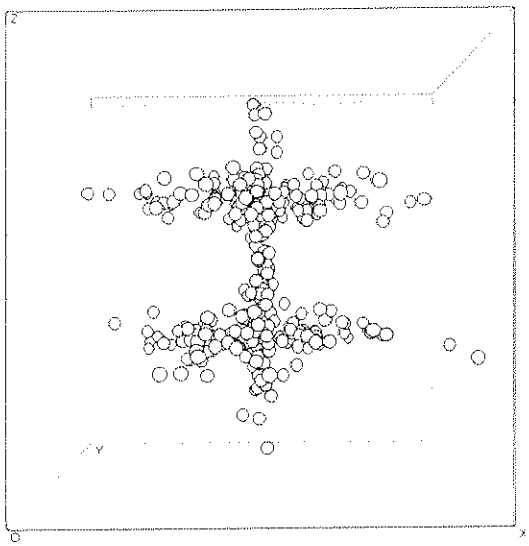

図 5.19: 図5.12 の分散を大きくしたデータ

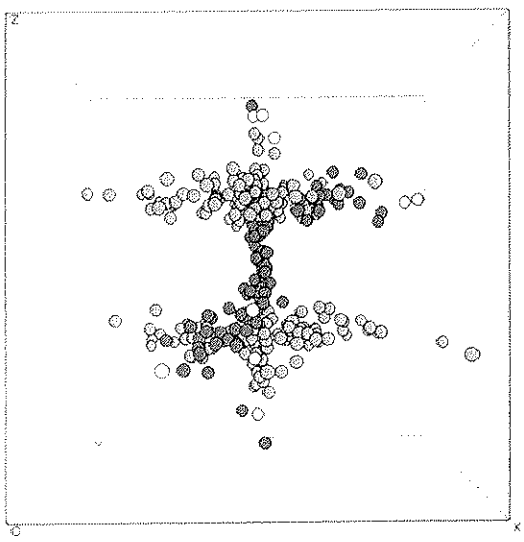

図 5.20: 図5.19 へのノイズクラスタを用いないクラスタリング結果

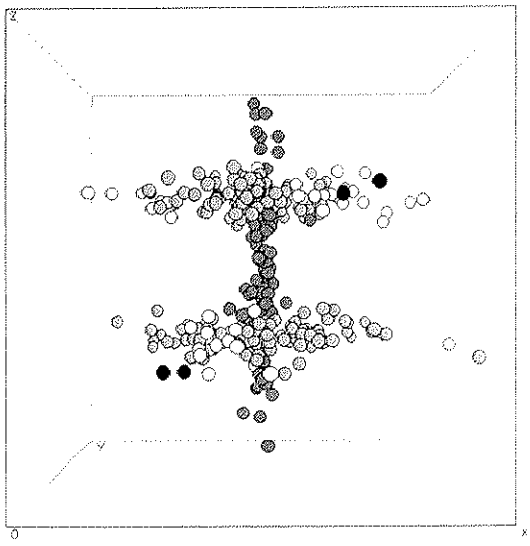

図 5.21: 図5.19 へのノイズクラスタを用いたクラスタリング結果

5.8　クラスタリングパラメータについて

クラスタリングを行うときには，多くのクラスタリングパラメータを与える必要がある。クラスタリング結果はこれらの値によって変化するために，クラスタリングはクラスタリング結果を観察しながら，何回も行う必要がある。ここでは，このときのクラスタリングパラメータについて簡単にまとめ，その決め方についていくつかの指針を示す。

1. クラスタ数 c

   クラスタ数については，これまでにも様々な提案がなされているが，ここでは別の観点からの提案をしたい。例1で行ったように，ある程度大きなクラスタ数からクラスタリングを行い，順々にクラスタを減少させていく繰り返しによる方法である。そしてその過程でメンバシップ値や目的関数の値の変化を検討する手法である。

2. スムージングパラメータ q

   このパラメータは分類からの見地では，1に近い値である方がよい。しかし，1に近いときは，メンバシップの初期値に依存したクラスタリング結果となりやすいく，クラスタの形状もあまり変化しなくなる。最初はある程度大きな値でクラスタリングを行い，段々と小さな値を試してゆくのがよい。クラスタリングの2段階アルゴリズムにおいて，自動的にこのパラメータを変化させることも今後の研究として考えることができる。

3. 線形度パラメータ l

   このパラメータはクラスタの線形性を定めるパラメータである。この値が0のときには，クラスタの形状はあいまいなものとなり，1に近づくとデータの分布に近いものになる。しかし，$l=1$ で最初からうまく形状をとらえた結果を得ることは難しく，クラスタリングを繰り返しながら，1に近づけてゆくようにする。

4. 停止条件パラメータ ε

   本クラスタリング手法では，メンバシップ値が単調に収束することは保証されていない。クラスタの形状が変化して大きく変化することもあるので，ε はある程度収束する範囲で十分小さな値とするべきである。

5. メンバシップの初期値 u_{ik}

本研究ではメンバシップの初期値は乱数によって与えられる。クラスタリング結果はある程度この初期値によって異なる結果となるため，異なる初期値を用いて何回かクラスタリングを行ってみる必要がある。もし安定したクラスタリング結果とならないときは，他のパラメータを調整してみることも必要である。

5.9 おわりに

この章では異なる次元のクラスタを発見するためのファジィクラスタリングのための新しい目的関数とそのアルゴリズムの提案を行った。さらにノイズクラスタを導入し，はずれ値やばらつきの大きなデータに対するクラスタリングを簡単な数値例に対して行い，有効性を示した。しかし本手法は万能というわけではない。何度もクラスタリングを行いながら，よい結果を探す必要がある。またクラスタリングの進め方やパラメータの決定方法についてのいくつかの提案を行った。