FCV 法の一般化と最小絶対誤差に基づくロバストなシェルクラスタリング*

(Generalization of FCV and robust shell clustering based on least absolute deviations)

本多　克宏 † ・東江　伸浩 † ・市橋　秀友 †
† 大阪府立大学　大学院　工学研究科
Graduate School of Engineering,
Osaka Prefecture University;
1-1 Gakuen-cho, Sakai city, Osaka 599-8531, JAPAN

概 要

データが非線形な特異性を持っている場合への主成分分析の応用として，データの布置に最もよく合う非線形座標系を求める一般化主成分分析法が提案されている．一方，曲面や球面上に分布したデータの局所的な構造を抽出するシェルクラスタリング法は，画像認識などへの応用が研究されている．本研究では，メンバシップを考慮した一般化主成分分析法に最小絶対誤差に基づく定式化を導入することにより，ロバストなシェルクラスタリング法を提案する．拡張次元におけるマイナー成分分析と FCM 法の融合手法である提案法は，線形クラスタリング法である Bezdek らの Fuzzy c-Varieties（FCV）法のアルゴリズムを非線形構造の抽出に応用したものであり，線形多様体の抽出と球状のクラスタリングの優先度を決めるトレードオフパラメータを変化させることにより，初期分割に依存しないクラスタリング結果が得られる．数値実験では，従来の可能性的制約を用いたシェルクラスタリング法との比較を通して，最小絶対誤差に基づく手法の有効性を示す.

(Generalized Principal Component Analysis (Generalized PCA) is a useful extension of the PCA algorithm for estimating a suitable non-linear coordinate system when sample data points have non-linear distribution. The non-linear models derived by Generalized PCA is closely related to shell clustering that partitions data sets into several shell-shape fuzzy clusters by extracting local circles or ellipses as the prototypes of clusters. This paper proposes a robust shell clustering technique by generalizing a linear fuzzy clustering algorithm based on least absolute deviations. The proposed method is a hybrid technique of local minor component analysis and FCM-type fuzzy clustering in the enlarged data space and can be regarded as an application of Fuzzy c-Varieties (FCV) algorithm for capturing local non-linear singularities. The tuning of the trade-off parameter makes it possible to derive stable clustering results that are robust to the initial partitioning. Numerical example composed of a comparison with the possibilistic shell clustering method shows the characteristic properties of our method.)
keywords: Shell Clustering, Fuzzy c-Varieties, Least Absolute Deviations.

1　はじめに

線形ファジィクラスタリング手法である Bezdek らの Fuzzy c-Varieties(FCV) 法 [1, 2] は，クラスターのプロトタイプとして線形多様体を用いる方法であるが，同時にプロトタイプとなる線形多様体を張るベクトルとして局所的な主成分ベクトルが求まることから，クラスタリングと主成分分析の

*知能と情報（日本知能情報ファジィ学会誌）, 15, 6, 693-701 (2003)

同時適用法であるといえる [3]．局所的な主成分ベクトルの算出はクラスターごとにファジィ散布行列の固有値および対応する固有ベクトルを求める固有値問題に帰着され，最大固有値から順に必要とする数の固有値に対応する固有ベクトルを主成分ベクトルとして算出する．また，球状のクラスターが得られる Fuzzy c-Means(FCM) 法 [1] の目的関数との線形和を目的関数として用いることにより楕円体状のクラスターを得る Fuzzy c-Elliptotypes(FCE) 法 [2] も提案されている．

一方，データが非線形な特異性を持っているときへの主成分分析の応用として，データの布置に最もよく合う非線形座標系を求める一般化主成分分析法が提案されている [4, 5]．たとえば，2 変量 x_1, x_2 が与えられた場合に 2 次の座標系を求めたいとすると，

$$z = a_1^* x_1 + a_2^* x_2 + a_3^* x_1 x_2 + a_4^* {x_1}^2 + a_5^* {x_2}^2 \tag{1}$$

を考えて，z の分散が最大となるように係数が定められる．すなわち，

$$\begin{aligned} \boldsymbol{x}^* &= (x_1^*, x_2^*, x_3^*, x_4^*, x_5^*)^\top \\ &= (x_1, x_2, x_1 x_2, {x_1}^2, {x_2}^2)^\top \end{aligned} \tag{2}$$

とおいたときに，2 次の主成分分析は，$\boldsymbol{a}^{*\top}\boldsymbol{a}^* = 1$ のような正規化制約の下で，$z = \boldsymbol{a}^{*\top}\boldsymbol{x}^*$ の分散を最大にするような $\boldsymbol{a}^*$ を決定する問題として定式化され，$\boldsymbol{x}^*$ の分散共分散行列の最大固有値と対応する固有ベクトルを求める問題に帰着される．また，小さな固有値に対応する固有ベクトルを用いる場合は，次元数の非線形縮小化の方法としてとらえることもできる．すなわち，z の値がすべてのサンプルデータについて等しい値となっている場合には，(1) 式の方程式で表される曲線上にすべてのデータが存在すると考えられるので，2 次曲線上であたかも 1 次元データとして取り扱うことが可能となる．

そこで本論文では，FCV 法で用いるデータを任意の関数の 1 次結合からなる座標系に一般化することにより，データが有する局所的な構造をとらえる手法を提案する．クラスターごとにデータの布置に最もよく合う非線形座標系を求める問題は，データ集合の局所的な特徴をとらえた非線形関数を推定する問題であり，データの部分構造を考慮した関数適合の問題と考えられる．また，拡張された座標系においてファジィ散布行列の最小固有値とそれに対応する固有ベクトルを求める問題は，マイナー成分分析 [6, 7] の非線形座標系への一般化ととらえることもできる．さらに，提案法の目的関数はクラスターごとのマイナー成分ベクトルへの射影を用いて簡潔に表されていることから，クラスタリング基準としてデータ点と超平面との 2 乗距離の代わりに絶対値距離を用いることができ，最小絶対誤差に基づくロバストなデータ分割 [8] が可能である．

クラスターのプロトタイプが非線形関数により表されるファジィクラスタリング法としては，シェルクラスタリングと総称される手法がある．Fuzzy c-Shells (FCS) [9] や Adaptive Fuzzy c-Shells (AFCS) [10] は円形もしくは楕円形のクラスターを得るための方法として有効であるが，繰り返しごとに非線形方程式の組を解く必要があり，計算量の面で問題がある．Fuzzy c-Spherical Shells (FCSS) [11] では “代数的な” 距離を用いてプロトタイプを解析的に求めることにより，計算効率を向上させている．また，Krishnapuram らの Fuzzy c-Quadric Shells(FCQS) 法 [12, 13] は 2 次多項式を距離関数として用いることにより，楕円体や双曲線，放物線状のクラスターを得ることができる．さらに，可能性的アプローチにより，ノイズの影響を受けにくいデータ分割を得る研究もなされている [13]．提案法で得られるデータ分割はシェルクラスタリング手法の結果と密接な関係があり，それらの手法との比較・検討も行う．

2　線形クラスタリングとシェルクラスタリング

本節では，線形クラスタリングとシェルクラスタリングの概要を振り返る．m 次元の n 個の標本データ $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ が与えられたときに，n 個の標本データを C 個のクラスターに分割する問題を考える．Bezdek らの FCV 法 [1, 2] では第 c クラスターのプロトタイプとして互いに線形独立な単位ベクトル $\boldsymbol{a}_{cj}$ により張られる p 次元の線形多様体を用い，線形多様体とデータ点との 2 乗距離を分類尺度とすることにより，以下の目的関数の最小化が図られる．

$$\begin{aligned} L_{fcv} &= \sum_{c=1}^{C}\sum_{i=1}^{n} {u_{ci}}^{\theta} \Big\{ (\boldsymbol{x}_i - \boldsymbol{b}_c)^\top (\boldsymbol{x}_i - \boldsymbol{b}_c) \\ &\qquad - \sum_{j=1}^{p} \boldsymbol{a}_{cj}^\top R_{ci} \boldsymbol{a}_{cj} \Big\} \end{aligned} \tag{3}$$

$$R_{ci} = (\boldsymbol{x}_i - \boldsymbol{b}_c)(\boldsymbol{x}_i - \boldsymbol{b}_c)^\top \tag{4}$$

ただし，$\boldsymbol{b}_c = (b_{c1}, \cdots, b_{cm})$ は第 c クラスターの中心である．u_{ci} は第 i サンプルデータ $\boldsymbol{x}_i$ が第 c クラスターに属する度合を示すメンバシップであり，

$$\sum_{c=1}^{C} u_{ci} = 1 \quad ; i = 1, \cdots, n \tag{5}$$

なる制約条件を満たすものとする．また，メンバシップのべき乗はファジィ分割を得るために用いられたもので，指数 θ が大きくなるにつれておのおののデータの所属が明確でなくなり，あいまいなデータ分割が得られるようになる．(3) 式を最小にする $\boldsymbol{a}_{cj}$ を求める問題はファジィ散布行列の固有値問題に帰着され，p 個の大きい固有値に対応する固有ベクトルとして算出されることから，$\boldsymbol{a}_{cj}$ はメンバシップを考慮しながらクラスターごとに得られる局所的な主成分ベクトルと考えることができる．

一方，データが非線形な特異性を持っている場合への拡張として，クラスターのプロトタイプを非線形方程式で表すシェルクラスタリングが提案されている．クラスターのプロトタイプが解析的に求まる FCQS 法 [12, 13] は，楕円や双曲線あるいは放物線状のクラスターが得られるシェルクラスタリング法であり，第 c クラスターのプロトタイプとなる曲線の方程式は，

$$\boldsymbol{x}^\top A_c \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{v}_c + v_{c0} = 0 \tag{6}$$

で表される．ただし，A_c は対称行列であり，A_c, $\boldsymbol{v}_c$, v_{c0} がプロトタイプの形状を決定する変数である．データ点 $\boldsymbol{x}_i$ と第 c クラスターのプロトタイプとの “代数的な”2 乗距離 d_{ic}^2 は，

$$d_{ic}^2 = (\boldsymbol{x}_i{}^\top A_c \boldsymbol{x}_i + \boldsymbol{x}_i{}^\top \boldsymbol{v}_c + v_{c0})^2 \tag{7}$$

で定義され，最小化すべき目的関数は以下のように定義される．

$$L_{fcqs} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} d_{ic}^2 \tag{8}$$

ここで，

$$\begin{aligned} \boldsymbol{p}_c{}^\top &= (p_{c1}, \cdots, p_{cm}, p_{c(m+1)}, \cdots, p_{cr}, \\ &\quad p_{c(r+1)}, \cdots, p_{c(r+m)}, p_{cs}) \end{aligned} \tag{9}$$

$$\begin{aligned} \boldsymbol{q}^\top &= (x_1^2, \cdots, x_m^2, x_1 x_2, \cdots, x_{m-1} x_m, \\ &\quad x_1, \cdots, x_m, 1) \end{aligned} \tag{10}$$

$$s = \frac{m(m+1)}{2} + m + 1 = r + m + 1 \tag{11}$$

とおくと，

$$\boldsymbol{x}^\top A_c \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{v}_c + v_{c0} = \boldsymbol{p}_c{}^\top \boldsymbol{q} = 0 \tag{12}$$

となる．$\boldsymbol{p}_c$ の制約条件として，Krishnapuram らは，

$$||p_{c1}^2 + \cdots + p_{cm}^2 + \frac{1}{2} p_{c(m+1)}^2 + \cdots + \frac{1}{2} p_{cr}^2||^2 = 1 \tag{13}$$

を用い，$r \times r$ 行列の固有値問題に帰着させてプロトタイプを求める方法を提案している．

また，ノイズが含まれるデータへの可能性的アプローチ法として，(5) 式のメンバシップの制約条件を取り除き，以下の目的関数を最小化する Possibilistic c-Quadric Shells (PCQS) 法 [13] が提案されている．

$$L_{pcqs} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} d_{ic}^2 + \sum_{c=1}^{C} \eta_c \sum_{i=1}^{n} (1 - u_{ci})^{\theta} \tag{14}$$

可能性的アプローチではメンバシップの制約として，おのおののメンバシップが [0,1] の値を持つことのみを用い，(5) 式を考慮しないことから，メンバシップはおのおのの標本データがクラスターに所属する可能性を表しているとみなされる．η_c は適当な正数であり，プロトタイプからの距離 d_{ic}^2 がいずれのクラスターについても η_c 以上のデータ点はどのクラスターにも属さないことになる．したがって，メンバシップの更新則は，以下のようになる．

$$u_{ci} = \frac{1}{1 + \left(\frac{d_{ic}^2}{\eta_c} \right)^{\frac{1}{\theta - 1}}} \tag{15}$$

可能性的クラスタリングでは，プロトタイプからどの程度離れていればノイズとみなすかをパラメータ η_c で決める必要があり，アルゴリズムが収束した後に，クラスターごとの d_{ic}^2 の平均値を η_c とおきなおして再度計算を行う手順などが提案されている．

ただし，FCQS 法や PCQS 法はクラスタリング結果が初期クラスターに大きく影響される傾向がある．Krishnapuram らは FCQS 法におけるメンバシップの初期値の決め方として，$\theta = 3.0$ とおいた FCM アルゴリズムで約 10 回更新した後に，クラスタリング基準としてマハラノビス距離を用いる Gastafson-Kessel(G-K) アルゴリズム [14] を $\theta = 2.0$ とおいて約 10 回更新し，さらに FCSS アルゴリズムにより約 5 回更新する方法が効果的であると報告している．また，PCQS 法の前処理としては FCQS 法を用いるべきであると述べている．しかし，これらの前処理の手順は煩雑であるばかりでなく，必ずしもすべてのデータ集合に有効であるとは限らない．

そこで本論文では，主成分分析で用いられる線形モデルの変量を非線形座標系へ一般化するという立場から FCV 法と同じ枠組みの中で局所的な非線形構造をとらえることにより，ロバストなモデルの推定に有効な最小絶対誤差に基づく定式化を行う．

3 FCV 法の一般化と最小絶対誤差に基づくロバストな非線形構造の抽出

本節では，FCV 法のアルゴリズムを一般化することにより，データ集合の局所的な非線形構造をとらえる手法を定式化する．まず，2 次元 ($m = 2$) のデータ $\boldsymbol{x}_i = (x_{i1}, x_{i2})^\top, i = 1, \cdots, n$ について，クラスターごとに合成変数

$$z_c = a_{c1}^* x_1 + a_{c2}^* x_2 + a_{c3}^* x_1 x_2 + a_{c4}^* {x_1}^2 + a_{c5}^* {x_2}^2 \tag{16}$$

のばらつきを最小化することを考える．ただし，係数ベクトル $\boldsymbol{a}_c^* = (a_{c1}^*, \cdots, a_{c5}^*)$ は $\boldsymbol{a}_c^{*\top} \boldsymbol{a}_c^* = 1$ なる制約を満たすものとする．このとき，(2) 式の $\boldsymbol{x}_i^*$ と 5 次元空間における第 c クラスターの中心 $\boldsymbol{b}_c^*$ を用いると，(8) 式の FCQS 法の目的関数は，以下のように書き換えられる．

$$\begin{aligned} L_{fcqs'} &= \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} (a_{c1}^* x_1 + a_{c2}^* x_2 + a_{c3}^* x_1 x_2 \\ &\quad + a_{c4}^* {x_1}^2 + a_{c5}^* {x_2}^2 - \boldsymbol{a}_c^{*\top} \boldsymbol{b}_c^*)^2 \\ &= \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} \left\{ \boldsymbol{a}_c^{*\top} (\boldsymbol{x}_i^* - \boldsymbol{b}_c^*)(\boldsymbol{x}_i^* - \boldsymbol{b}_c^*)^\top \boldsymbol{a}_c^* \right\} \\ &= \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} \boldsymbol{a}_c^{*\top} R_{ci}^* \boldsymbol{a}_c^* \end{aligned} \tag{17}$$

$$R_{ci}^* = (\boldsymbol{x}_i^* - \boldsymbol{b}_c^*)(\boldsymbol{x}_i^* - \boldsymbol{b}_c^*)^\top \tag{18}$$

すなわち，$-\boldsymbol{a}_c^{*\top} \boldsymbol{b}_c^*$ を FCQS 法で用いられている定数項と考えると，クラスター中心を考慮することにより，非線形な形状のクラスターを抽出するための目的関数が，FCV 法の目的関数に類似した形式で書き表されるようになる．ここで，5 次元空間での第 c クラスターにおける局所的な主成分ベクトルを主要なものから順に $\boldsymbol{a}_{c1}, \cdots, \boldsymbol{a}_{c5}$ とおくと，最小化するべきクラスタリング基準 $\boldsymbol{a}_c^{*\top} R_{ci}^* \boldsymbol{a}_c^*$ は，

$$\begin{aligned} \boldsymbol{a}_c^{*\top} R_{ci}^* \boldsymbol{a}_c^* &= |\boldsymbol{a}_c^{*\top} (\boldsymbol{x}_i^* - \boldsymbol{b}_c^*)|^2 \\ &= ||\boldsymbol{x}_i^* - \boldsymbol{b}_c^*||^2 - \sum_{j=1}^{4} |\boldsymbol{a}_{cj}^\top (\boldsymbol{x}_i^* - \boldsymbol{b}_c^*)|^2 \end{aligned} \tag{19}$$

となり，拡張したデータ次元においてプロトタイプを超平面とした線形クラスタリング，つまり局所的なマイナー成分分析のための基準であることが分かる．したがって，$\boldsymbol{a}_c^*$ は拡張したデータ次元における局所的なマイナー成分ベクトルであるといえる．

局所的なマイナー成分分析におけるロバストなモデル推定法として，本多ら [8] は最小絶対誤差に基づく定式化によりノイズの影響を無視した線形構造の抽出を行っている．そこで以下では FCE 法に倣い，FCM 法の目的関数にロバストなマイナー成分分析の目的関数を融合することにより，以

下の目的関数の最小化問題を考える．

$$
\begin{aligned}
L_{qfcea} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\Big\{\alpha(\boldsymbol{x}_i^*-\boldsymbol{b}_c^*)^\top(\boldsymbol{x}_i^*-\boldsymbol{b}_c^*) \\
&\quad +(1-\alpha)|\boldsymbol{a}_c^{*\top}(\boldsymbol{x}_i^*-\boldsymbol{b}_c^*)|\Big\} \\
&\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}
\end{aligned}
\tag{20}
$$

ただし，メンバシップは (5) 式の制約条件を満たすものとする．ここで，(20) 式の {} 内の第 1 項 $(\boldsymbol{x}_i^*-\boldsymbol{b}_c^*)^\top(\boldsymbol{x}_i^*-\boldsymbol{b}_c^*)$ は拡張された 5 次元空間における FCM 法の目的関数を表し，$\boldsymbol{x}_i^*$ は $\boldsymbol{x}_i$ の 2 変量に 2 乗および積を加えた (2) 式の 5 次元データである．$\boldsymbol{b}_c^*$ は 5 次元空間における第 c クラスターの中心である．第 2 項 $\boldsymbol{a}_c^{*\top}R_{ci}^*\boldsymbol{a}_c^*$ はクラスターごとに合成変数 z_c のばらつきを最小化するための項である．α はトレードオフパラメータであり，球状のクラスターを得る FCM クラスタリングと非線形構造の抽出の優先度を定める．Krishnapuram らの用いた前処理は，局所的な非線形構造の抽出においても，最適化の前段階では球状のクラスタリングをある程度考慮するべきであることを示唆していると考えられる．そこで本論文では，FCM クラスタリングはアルゴリズムの繰り返しの初期段階においてのみ考慮するものとし，α を繰り返しごとに減少させることにより 0 に収束させる．ファジィ分割を得るための手法としては，提案法では標準的な手法におけるメンバシップのべき乗の代わりに，エントロピー正則化 [15] を採用し，目的関数にエントロピー項を加えている．エントロピー正則化には，標準的な方法で必要となる例外処理が不要であるほか，クラスター中心の算出においてべき乗ではなくメンバシップそのものを用いて更新則が定義できるといった特長があることから，以下ではエントロピー正則化を用いるファジィクラスタリング法を中心に議論する．λ は標準的な手法における θ と同様にメンバシップのファジィ度を決める係数であり，大きいほどあいまいなデータ分割を与えるようになる．目的関数を最小化するためには，交互最小化の原理に基づいて局所的なマイナー成分ベクトル $\boldsymbol{a}_c^*$，クラスター中心 $\boldsymbol{b}_c^*$ およびメンバシップ u_{ci} の更新を繰り返すアルゴリズムを用いる．

メンバシップとクラスター中心を固定した際には，(20) 式の目的関数の最小化問題はクラスターごとに

$$
L_{qfcea}^{(c)} = \sum_{i=1}^{n} u_{ci}|\boldsymbol{a}_c^{*\top}(\boldsymbol{x}_i^*-\boldsymbol{b}_c^*)| \tag{21}
$$

を最小とする $\boldsymbol{a}_c^*$ を求める問題となる．ここで，中心 $\boldsymbol{b}_c^*$ を通り，(21) 式を最小とする局所的な超平面は，必ず超平面の次元数と等しい数のデータ点を通ることが示されている [8]．したがって，5 次元空間で局所的なマイナー成分ベクトルを求める場合には，n 個のサンプルデータ点のうちの 4 個の点とクラスター中心を通る超平面の中で，(20) 式を最小とする 4 個のサンプルデータ点の組合せを探索する組合せ最適化問題に帰着される．そして，選ばれた 4 個のサンプルデータ点とクラスター中心を通る超平面の法線ベクトルとしてマイナー成分ベクトルが得られる．

クラスター中心の更新の際には，最適なクラスター中心を厳密に求めることは容易ではない．そこで，本研究では目的関数の第 1 項目と第 2 項目を最小化するクラスター中心の重み付き平均値

$$
\boldsymbol{b}_c^* = \alpha\boldsymbol{b}_c^{*1}+(1-\alpha)\boldsymbol{b}_c^{*2} \tag{22}
$$

を新たなクラスター中心として用いることとする．(20) 式は $\alpha=1$ の時には FCM 法の目的関数に一致し，第 1 項を最小化するクラスター中心 $\boldsymbol{b}_c^{*1}$ はラグランジュ関数の最適性の必要条件から，

$$
\boldsymbol{b}_c^{*1} = \frac{\sum_{i=1}^{n} u_{ci}\boldsymbol{x}_i^*}{\sum_{i=1}^{n} u_{ci}} \tag{23}
$$

のように，各軸ごとにメンバシップの重み付きの平均値を計算することにより求められる．一方，$\alpha=0$ のときは目的関数は $\boldsymbol{a}_c^*$ に射影されたデータの絶対値距離の和であるので，$\boldsymbol{a}_c^*$ を基底ベクトルとする 1 次元空間に射影されたデータの l_1 距離の和であるととらえられる．したがって，$\boldsymbol{x}_1^*,\cdots,\boldsymbol{x}_n^*$ がマイナー成分ベクトルとの内積に応じて並べ替えられ，関数 $q(k), k=1,\cdots,n$ により添え字が

$$
\boldsymbol{a}_c^{*\top}\boldsymbol{x}_{q(1)}^* \le \boldsymbol{a}_c^{*\top}\boldsymbol{x}_{q(2)}^* \le \cdots \le \boldsymbol{a}_c^{*\top}\boldsymbol{x}_{q(n)}^* \tag{24}
$$

を満たすように変換されたとすると，最適なクラスター中心 $\boldsymbol{b}_c^{*2}$ を求める問題は，メンバシップを考慮しながら $\boldsymbol{a}_c^{*\top}\boldsymbol{x}_{q(k)}^*$ のメディアン（中央値）を探索する問題に等しい．$\boldsymbol{a}_c^*$ を一つの基底とする 5 次元の正規直交系において軸ごとにメンバシップの重み付きメディアンを求めるためには，以下の計算アルゴリズムが提案されている．

Algorithm BC

begin
  $\boldsymbol{b}_c^{*2} := \boldsymbol{0}$;
  ORTHONORMALIZING$(\boldsymbol{a}_c^*, W)$;
  $j := 0$;
  while $(j < 5)$ do begin
    $j = j + 1$;
    ORDERING$(\boldsymbol{w}_j, X)$;
    $S := -\sum_{i=1}^{n} u_{ci}$;
    $k := 0$;
    while $(S < 0)$ do begin
      $k := k + 1$;
      $S := S + 2u_{cq(k)}$;
    end;
    $\boldsymbol{b}_c^{*2} := \boldsymbol{b}_c^{*2} + (\boldsymbol{w}_j{}^{\top}\boldsymbol{x}_{q(k)}^*)\boldsymbol{w}_j$;
  end;
  output $\boldsymbol{b}_c^{*2}$;
end.

ただし，ORTHONORMALIZING$(\boldsymbol{a}_c^*, W)$ は $\boldsymbol{a}_c^*$ を第 1 基底ベクトルとする任意の正規直交系を張る基底ベクトル $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_5$ を作成するサブルーチンを，ORDERING$(\boldsymbol{w}_j, X)$ は，

$$\boldsymbol{w}_j{}^{\top}\boldsymbol{x}_{q(1)}^* \le \boldsymbol{w}_j{}^{\top}\boldsymbol{x}_{q(2)}^* \le \cdots \le \boldsymbol{w}_j{}^{\top}\boldsymbol{x}_{q(n)}^* \tag{25}$$

を満たすように添え字関数 $q(k)$ を定めるサブルーチンを表す．

ここで，(22) 式で求められるクラスター中心は，$0 < \alpha < 1$ の場合には必ずしも最適な値とはなっていないものの，トレードオフパラメータの減少により α が 0 に収束した後には，最適なクラスター中心となっている．

メンバシップの更新則は，ラグランジュ関数の最適性の必要条件から以下のように求まる．

$$u_{ci} = \frac{\exp(-\lambda^{-1}E_{ci})}{\sum_{l=1}^{C}\exp(-\lambda^{-1}E_{cl})} \tag{26}$$

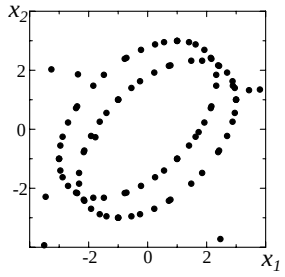
ただし，

$$E_{cl} = \alpha(\boldsymbol{x}_i^* - \boldsymbol{b}_c^*)^{\top}(\boldsymbol{x}_i^* - \boldsymbol{b}_c^*) + (1-\alpha)|\boldsymbol{a}_c^{*\top}(\boldsymbol{x}_i^* - \boldsymbol{b}_c^*)|$$

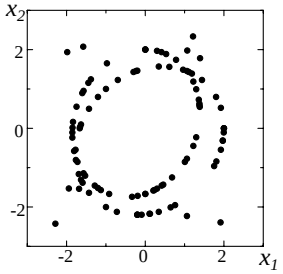
である．

以上の更新則を用いたアルゴリズムは，以下の通りである．

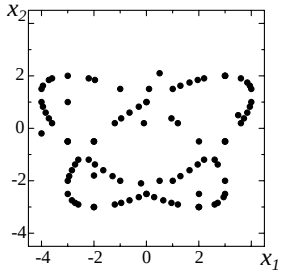
Step 1，拡張したデータ次元におけるデータ集合 $\boldsymbol{x}_i^*$，$i = 1, \cdots, n$ を作成し，制約条件に従うメンバシップ u_{ci}，$c = 1, \cdots, C$，$i = 1, \cdots, n$ の初期値を乱数を用いて定める．また，トレードオフパラメータ α の初期値およびファジィ度を調節する定数 λ を定める．



(a) データ集合 1



(b) データ集合 2



(c) データ集合 3

図 1: 実験に用いたデータ集合の散布図

Step 2，クラスターごとに中心 $\boldsymbol{b}_c^*$ の初期値を FCM 法の更新則，

$$\boldsymbol{b}_c^* = \frac{\sum_{i=1}^{n} u_{ci}\boldsymbol{x}_i^*}{\sum_{i=1}^{n} u_{ci}}$$

により定める．

Step 3，組み合わせ最適化問題を解くことより，クラスターごとに局所的なマイナー成分ベクトル $\boldsymbol{a}_c^*$ を更新する．

Step 4，クラスターごとに (22) 式を用いて中心 $\boldsymbol{b}_c^*$ を求める．

Step 5，メンバシップ u_{ci} を (26) 式により更新する．

Step 6，小さな正数 ϵ に対して，終了判定条件

$$\max_{c,i} \mid u_{ci}^{NEW} - u_{ci}^{OLD} \mid < \epsilon$$

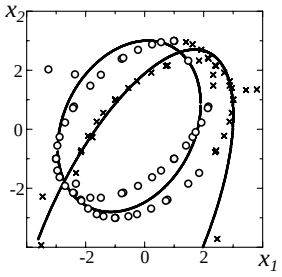
を満たせば終了．それ以外は α を適宜減少させ，**Step 3** へ戻る．

アルゴリズムの実行においては，トレードオフパラメータ α を適当なスケジュールにしたがって減少させ，0 に収束させる．これは，シェルクラスタリングにおける初期分割への依存度を減少させるために，分析の初期段階ではある程度球状のクラスタリングを考慮しながらメンバシップを割り当てるものの，分析が進むにつれてシェルクラスタリングを優先させ，α が 0 に収束した後はロバストな曲線の当てはめのみを考慮することを意味している．α の減少スケジュールを適当に定める必要があるものの，単一の目的関数の最小化に基づきながら，初期値依存を減少させることができる．

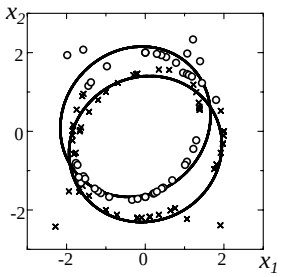
以上の議論は，m 次元データから q 次の座標系を得る問題に容易に一般化することができる．たとえば，2 次元データ $\boldsymbol{x}_i$ から 3 次の座標系を得ることが分析目的である場合ならば，拡張された観測データ $\boldsymbol{x}_i^*$ の次元は 9 次元となり，同様のアルゴリズムで任意の 3 次方程式を求めることができる．

4 数値実験

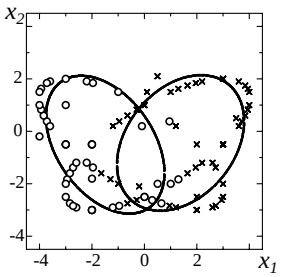
提案手法の有効性を検証するために，図 1 の 3 種類のデータを用いて数値実験を行った．用いたデータ集合はいずれも 100 個の標本データからなる 2 次元データで，二つの楕円を形成する 90 個の



(a) データ集合 1



(b) データ集合 2



(c) データ集合 3

図 2: FCQS 法によるクラスタリング結果の例

表 1: 最適解が得られた頻度の比較

	PCQS 法	提案法
データ集合 1	3	92
データ集合 2	76	97
データ集合 3	10	45

サンプルに 10 個のノイズデータが付加されている．これらのデータ集合について，標本データを二つのクラスター ($C = 2$) に分割しながら，クラスターごとに楕円をとらえる分析を行った．いすれのデータ集合においても二つの楕円が交差しており，FCM 法などのデータの塊を抽出する手法では構造の把握が困難な例となっている．

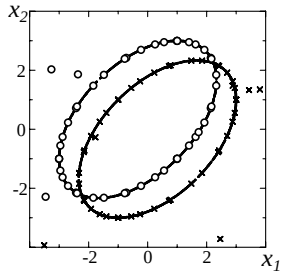
まず，距離の 2 乗をクラスタリング基準とする FCQS 法で分析した結果の例を，図 2 に示す．ただし，FCQS 法の前処理としては Krishnapuram らの提案した方法（$\theta = 3.0$ とおいた FCM アルゴリズムで 10 回更新した後に，G-K アルゴリズムを $\theta = 2.0$ とおいて 10 回更新し，さらに FCSS アルゴリズムにより 5 回更新する方法）を用い，メンバシップのファジィ度を制御するパラメータは $\theta = 2.0$ とした．図では，メンバシップが最大となるクラスターに分類することにより，データ集合は○と×で表される二つのクラスターに分けられ，黒い楕円で表されるプロトタイプが得られている．最小 2 乗法の原理に基づく FCQS 法では，ノイズの影響のために二つの楕円を正しくとらえられなかった．

そこで，ノイズの影響を受けにくい PCQS 法と提案法により分析を行い，クラスタリング結果の比較を行った．ただし，提案法では計算時間を短縮するために，マイナー成分ベクトルを求める組合せ最適化問題を解く際の組合せの探索範囲をメンバシップが 0.5 以上のものに限定して行った．また，ファジィ度を定める係数は $\lambda = 2.0$ とし，トレードオフパラメータ α は，

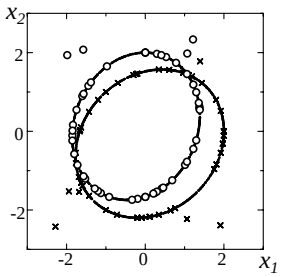
$$\alpha = \begin{cases} 1.0 \times 0.8^{t-1} & ; \alpha > 0.001 \\ 0 & ; \alpha \leq 0.001 \end{cases}$$

のスケジュールにしたがって減少させた．ただし，t はアルゴリズムの繰り返し回数を表す．一方，PCQS 法では $\theta = 2.0$ とし，FCQS 法を前処理として用いた．また，可能性的アプローチにおいてノイズの範囲を定めるパラメータは，

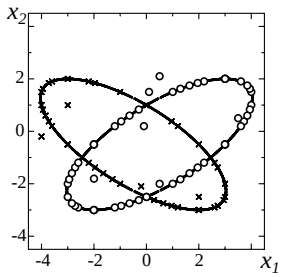
$$\eta_c = \frac{K}{\sum_{i=1}^{n} u_{ci}^{\theta}} \sum_{i=1}^{n} u_{ci}^{\theta} d_{ic}^2 \tag{27}$$



(a) データ集合 1

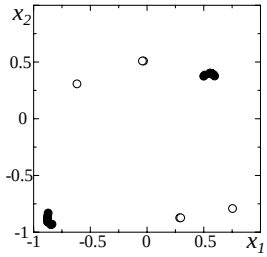


(b) データ集合 2

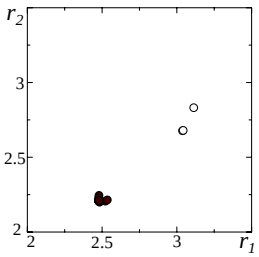


(c) データ集合 3

図 3: ロバストなクラスタリングの結果

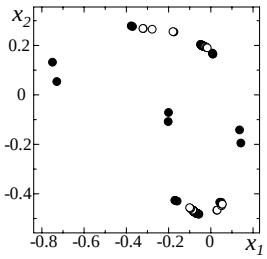


(a) 二つの円の中心

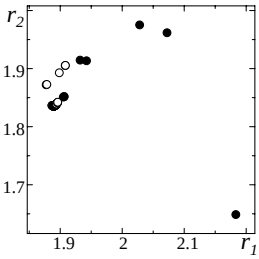


(b) 二つの円の半径

図 4: データ集合 1 の前処理の結果



(a) 二つの円の中心

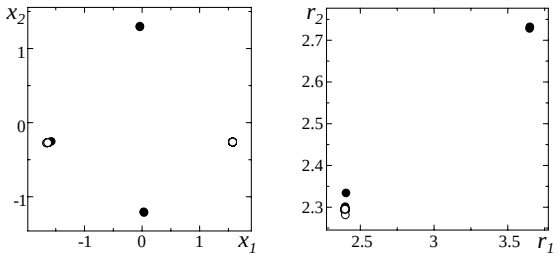


(b) 二つの円の半径

図 5: データ集合 2 の前処理の結果

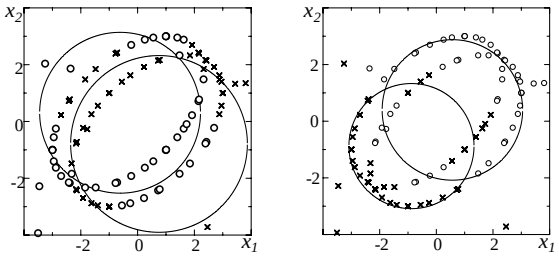
により繰り返しごとに与えることとし，定数 K は試行錯誤により，それぞれ，データ 1 では 3.5，データ 2 では 1.0，データ 3 では 7.5 とした．図 3 に二つの楕円を正しくとらえた分類結果を示す．さらに，表 1 に初期分割を変えながら行った 100 回の試行において，それぞれの手法で図 3 のクラスタリング結果が得られた回数を示す．表 1 においていずれのデータ集合についても提案法の方が最適解を与える頻度が高い．

ここで，PCQS 法の初期値への依存度に関して考察する．FCQS 法や PCQS 法は初期分割の影響を受けやすいという欠点があり，Krishnapuram らは FCM 法，G-K アルゴリズムおよび FCSS 法を順に適用する前処理を提案している．そこで，それらの前処理の結果と PCQS 法で得られる結果と



(a) 二つの円の中心　　(b) 二つの円の半径

図 6: データ集合 3 の前処理の結果



(a) 最適解の場合の例　　(b) 最適解でない場合の例
（L_{fcss}=575.8）　　（L_{fcss}=467.5）

図 7: データ集合 1 の前処理で得られるプロトタイプ

の関連を調べるために，FCSS 法でプロトタイプとして得られた二つの円の中心とそれらの半径を図 4〜6 に示す．図中，○が図 3 のプロトタイプが得られたときの前処理結果を，●が得られなかった場合の結果を表している．ただし，二つの円のうち，半径の大きなものをクラスター 1 のプロトタイプとし，それぞれの円の半径を r_1 および r_2 とおいた．図 4〜6 から，最適解を得るためには，前処理の結果が限られたパターンのいずれかで与えられなければならず，分析の成否が初期分割に大きく依存していることが分かる．そこで，初期値への依存度に顕著な差のあったデータ集合 1 について詳しく考察する．PCQS 法で最適な解が得られたときと得られなかったときのそれぞれについて，FCM 法の後に FCSS 法を用いて行う前処理で得られる初期分割の例を図 7 に示す．図 7 から，PCQS 法で最適な解を得るためには，前処理においてデータ集合の大まかな特徴をとらえておく必要があるといえる．しかし，おのおのの場合の FCSS 法における目的関数値

$$L_{fcss} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} (||\boldsymbol{x}_i - \boldsymbol{b}_c||^2 - r_c^2)^2 \tag{28}$$

を調べたところ，最適解が得られた場合が 575.8 であったのに対して，最適解でなかった場合の方が 467.5 と小さくなっており，これらの前処理の過程が必ずしもシェルクラスタリングの目的にあったものではないことが伺える．データ集合 2 および 3 についても，前処理が分析結果に大きく影響を与えていることが伺えるが，ほぼ等しい前処理結果ながら最適解を与える場合と与えない場合に分かれることもあり，前処理における分割のわずかな違いが結果に大きく反映されるといえよう．一方，提案手法では，前処理の過程を用いず，パラメータを徐々に変化させながら単一の目的関数の最適化を行っているため，分析の初期段階において好ましくない分割に陥ることが少なく，初期分割の影響を受けにくいと考えられる．このように，最小絶対誤差を用いる提案法は，FCQS 法や PCQS 法で問題となる前処理の影響が少ない手法であるといえる．

5　おわりに

本論文では，多次元に拡張したデータ集合を用いて一般化した線形クラスタリング手法とシェルクラスタリング手法との類似点を議論し，拡張したデータ空間での局所的なマイナー成分分析ととらえられる目的関数を最小絶対誤差に基づいて定式化することにより，ノイズの影響を無視しながらデータの非線形構造をとらえる手法を提案した．数値実験では，2 乗距離をクラスタリング基準としなが

らメンバシップに可能性的制約を考慮することによりロバストなシェルクラスタリングを行う PCQS 法と提案法との比較を通して，最小絶対誤差に基づく手法の方が初期分割の影響を受けにくいことを示した．ただし，提案法ではアルゴリズムの繰り返しごとに局所的なマイナー成分ベクトルを求める際に，組合せ最適化問題を解く必要がある．本論文の数値実験では，超平面上に存在するデータ点の組合せを求める際の探索範囲を，当該クラスターに所属するデータ点のみに限定して行うことにより計算時間の短縮を行ったが，Pentium III プロセッサを搭載したパーソナルコンピュータでアルゴリズムの収束に約 30 分を費やした．画像の処理などへの応用の際には，より多くのデータ点からなるデータ集合を取り扱う必要があると考えられることから，より効率的な探索方法の採用などが今後の課題といえる．しかし，データ数が比較的少ない場合でも，前処理を必要とせず，初期分割の影響を受けにくいという提案法の特色は，まばらに散布されたデータ集合の分析のほかに，大規模なデータ集合からランダムに抽出した部分集合を用いて PCQS 法の初期分割を得る際にも有効である．

参考文献

[1] J. C. Bezdek: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press (1981)

[2] J. C. Bezdek, C. Coray, R. Gunderson and J. Watson: Detection and Characterization of Cluster Substructure 2. Fuzzy c-Varieties and Convex Combinations Thereof; *SIAM J. Appl. Math.*, Vol.40, No.2, pp.358-372 (1981)

[3] 本多克宏, 杉浦伸和, 市橋秀友, 荒木昭一, 久津見洋: 最小 2 乗基準を用いた Fuzzy c-Varieties 法における欠測値の処理法; 日本ファジィ学会誌, Vol.13, No.6, pp.680-688 (2001)

[4] R. Gnanadesikan and M. B. Wilk: Data Analytic Methods in Multivariate Statistical Analysis; *Multivariate Analysis II (P. R. Krishnaiah, ed.)*, Academic Press, pp.593-638 (1969)

[5] R. Gnanadesikan: *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley & Sons (1977)

[6] E. Oja: Principal Components, Minor Components and Linear Neural Networks; *Neural Networks*, Vol.5, pp.927-935 (1992)

[7] F.-L. Luo, R. Unbehauen and A. Cichocki: A Minor Component Analysis Algorithm; *Neural Networks*, Vol.10, pp.291-297 (1997)

[8] 本多克宏, 東江伸浩, 市橋秀友: 最小絶対誤差に基づく線形ファジークラスタリング; 電子情報通信学会論文誌 (D-II), Vol.J86-D-II, No.1, pp.12-21 (2003)

[9] R. N. Dave: Fuzzy Shell-Clustering and Application to Circle Detection in Digital Images; *Int. J. Gen. syst.*, Vol.16, pp.343-355 (1990)

[10] R. N. Dave and K. Bhaswan: Adaptive Fuzzy C-Shells Clustering and Detection of Ellipses; *IEEE Trans. on Neural Networks*, Vol.3, No.5, pp.643-662 (1992)

[11] R. Krishnapuram, O. Nasraoui and H. Frigui: The Fuzzy C Spherical Shells Algorithms: A New Approach; *IEEE Trans. on Neural Networks*, Vol.3, No.5, pp.663-671 (1992)

[12] R. Krishnapuram, H. Frigui and O. Nasraoui: Quadric Shell Clustering Algorithms and the Detection of Second Degree Curves; *Pattern Recognition Lett.*, Vol.14, No.7, pp.545-552 (1993)

[13] R. Krishnapuram, H. Frigui and O. Nasraoui: Fuzzy and Possiblistic Shell Clustering Algorithms and their Application to Boundary Detection and Surface Approximation - PartI; *IEEE Trans. on Fuzzy systems*, Vol.3, No.1, pp.29-43 (1995)

[14] D. E. Gustafson and W. C. Kessel: Fuzzy Clustering with a Fuzzy Covariance Matrix; *Proc. of the IEEE Conf. Decision and Control*, Vol.2, pp.761-766 (1979)

[15] 宮本定明: クラスター分析入門, 森北出版 (1999)

［問い合わせ先］
〒 599-8531　大阪府堺市学園町 1-1
大阪府立大学大学院工学研究科
電気・情報系専攻経営工学分野
本多 克宏
TEL：072-254-9355
FAX：072-254-9915
E–mail: honda@ie.osakafu-u.ac.jp

著者略歴

本多　克宏　(ほんだ　かつひろ) [正会員]
1999 年大阪府立大学大学院工学研究科博士前期課程電気・情報系専攻修了．同年日本電信電話（株）入社，同年大阪府立大学工学部経営工学科助手，現在に至る．ニューラルネットワーク，ファジィクラスタリングの研究に従事．IEEE，日本ファジィ学会，システム制御情報学会，日本経営工学会の会員．

東江　伸浩　(とうごう　のぶひろ)
2003 年大阪府立大学大学院工学研究科博士前期課程電気・情報系専攻修了．同年（株）日本総合研究所入社．在学中はファジィクラスタリングの研究に従事．

市橋　秀友　(いちはし　ひでとも) [正会員]
1971 年大阪府立大学工学部経営工学科卒業．同年松下電器産業（株）入社，1981 年大阪府立大学工学部経営工学科助手，1987 年同講師，1989 年同助教授，1993 年同教授，現在に至る．工学博士．ファジィクラスタリングやニューラルネットワークなどでのデータ解析法，その知的システムや人間機械システムへの応用研究に従事．IEEE，日本ファジィ学会，システム制御情報学会，電子情報通信学会，日本経営工学会などの会員．