# 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

Handling Missing Values in Fuzzy $c$-Varieties with Least Square Criterion

日本ファジィ学会誌，13，6，680-688（2001）

本多　克宏
杉浦　伸和
市橋　秀友
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
〒 599-8531　大阪府堺市学園町１－１

荒木　昭一
九津見　洋
松下電器産業 (株)　先端技術研究所
〒 619-0237　京都府相楽郡精華町光台３－４

Katsuhiro HONDA
Nobukazu SUGIURA
Hidetomo ICHIHASHI
Graduate School of Engineering, Electrical Engineering and Information Science, Industrial Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan

Shoichi ARAKI
Hiroshi KUTSUMI
Advanced Technology Research Laboratories, Matsushita Electric Industrial Co.,Ltd.
3-4 Hikaridai, Seika-cho, Souraku-gun, Kyoto 619-0237, Japan

## 要　約

Bezdek らの Fuzzy $c$-Varieties（FCV）法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ことにより線形のクラスタリングを行う手法で，局所的な主成分分析とみなすことができる．しかし，欠測値を含む実データの分析に際しては，データ行列を直接扱うことができない．本論文では，FCV 法の目的関数をデータ行列の最小 2 乗近似の立場から再検討することにより定義し，その一般化として，欠測値を含むデータから局所的な主成分を抽出しながら超楕円体状のクラスターを形成する手法を提案する．まず，FCV 法の目的関数が主成分分析で用いられる最小 2 乗基準にメンバシップを導入したものに等しいことを議論し，クラスターごとの局所的な主成分分析法として定式化を行う．次に，定義された最小 2 乗基準に基づく目的関数を，欠測値を含むデータを扱えるように拡張することにより，不完全データから局所的な主成分を抽出しながら線形クラスタリングを行う手法を提案する．その際，欠測値を含むデータに適切なメンバシップを割り当てられない場合があることから，クラスター中心からの距離を考慮する FCM 法の目的関数を重み付きで付加することにより，超楕円体状のクラスターを形成する．数値例では，感性評価データへの適用を通して提案手法の有効性を示す．



## キーワード

ファジィクラスタリング，欠測値，FCV 法，最小 2 乗基準

## Abstract

Fuzzy $c$-Varieties (FCV) clustering proposed by Bezdek *et al.* is a linear clustering method whose prototypes are linear varieties and can be regarded as a technique for extracting local principal components. In spite of its usefulness, the FCV algorithm cannot deal with an incomplete data set including missing values without elimination or imputation of data. In this paper, we propose a method for partitioning an incomplete data set including missing values into several fuzzy clusters using local principal components. First, FCV clustering is defined as the technique for the extraction of local principal components based on the minimization of the least square criterion, which performs the lower rank approximation of the data matrix. While the objective function of FCV clustering is based on the minimization of the distances between data points and prototypical linear varieties, the same objective function can be derived from the least square criterion under a certain condition. Second, a new technique for dealing with incomplete data sets is proposed by extending the method to extract local principal components. Numerical example shows the characteristic properties of our method.



## Keywords

Fuzzy Clustering, Missing Value, Fuzzy $c$-Varieties, Least Square Criterion

# 1　はじめに

多くの項目の間の関連性を統計的に分析し，現象を要約して簡潔な表現を与えたり，現象の背後に潜む構造を浮き彫りにするための主成分分析は，複雑な現象を解明するための有力な方法として幅広く利用されている [1]．しかし，実世界で収集されたデータを分析する際には，調査すべき項目に対する無回答や測定機器の故障・測定ミスなどの影響で，データ中に観測されなかった部分，すなわち欠測値（missing value）が含まれる場合が少なくない．このような不完全データを用いる場合には，欠測値を含むデータまたは属性を全て取り除いて分析を行う方法もあるが，データの持つ情報を切り捨ててしまうことになり，好ましくない．そこで，欠測値に対して適当なモデルを作って推定した値を代入した後に分析を施す方法がよく用いられている [2]．不完全データの補完法としては，おのおのの欠測値に対応する変量の平均値を推定値として代用する簡便な手法のほかに，正規分布などのデータ分布を仮定した上で，EM アルゴリズム [3] により尤度最大化の原理に基づいて推定された値を欠測値に代入する方法などが用いられることが多い．しかし，補完の際の誤差が得られる結果にも反映されてしまうという問題があるほか，多変量データの情報圧縮が目的である場合にそもそも欠測値の補完自体に意味があるかという議論もあることから，欠測値を補完することなしに，直接データ行列中の観測値のみを利用して主成分を抽出する手法が提案されている．Ruhe [4] や Wiberg [5] は特異値分解を用いた方法を提案し，Shum ら [6] はその改良手法を多面体物体のモデリングへ応用している．また，柴山 [7] は主成分分析をデータ行列の線形モデルによる最小 2 乗近似の立場 [8] [9] から再検討し，これを一般化することにより不完全データから主成分を抽出する手法を提案している．

一方，近年，ファジィクラスタリングと他の多変量解析手法を融合することにより，データの局所的な構造を考慮しながら特徴量を抽出する研究が盛んに行われている．クラスターのプロトタイプとして線形多様体を用いる Bezdek らの Fuzzy $c$-Varieties（FCV）法 [10] [11] は，所属度合いを表すメンバシップを考慮したファジィ散布行列の固有ベクトルとしてプロトタイプを張るベクトルを算出することから，局所的な構造をとらまえながらクラスターごとに局所的な主成分ベクトルを抽出する問題であると捕らえることができるので，主成分分析とクラスター分析の同時適用法であるといえ，複雑な分布形状を有するデータから局所的な特徴を抽出することができる．また，Bezdek らの Fuzzy $c$-Means（FCM）法 [10] の目的関数とその他の多変量解析手法の目的関数を組み合わせることにより，様々な局所的特徴量を抽出する試みもなされている [12] [13]．しかし，他の多変量解析と同様にファジィクラスタリングにおいても，欠測値を含むような実世界のデータを分析する場合には，どのように不完全データを取り扱うかという問題が生じる．FCM 法における欠測値の処理方法については，Miyamoto ら [14] が標準的な方法とエントロピー正則化 [15] を用いる方法の二つのバリエーションごとに考察を行っており，欠測値に重み付きの平均値を代入する手法のほか，欠測している属性を無視し，観測されたデータのみを用いて分析を行う手法を提案している．また，Timm ら [16] は標準的な FCM 法について，ノルム行列により距離の定義を変化させる場合も含めて同様の考察を行い，欠測している属性を無視する手法を用いる方が，メンバシップが全てのクラスターについて均等な値に近づく傾向があり，より曖昧なデータ分割が得られる実験結果を報告している．

そこで，本研究では，欠測値を含む不完全データから局所的な主成分を抽出しながら，データ集合を線形もしくは楕円体状のいくつかのクラスターに分割するファジィクラスタリング手法を提案する．まず，FCV 法が主成分分析と同様に最小 2 乗基準を用いて書き換えられることを議論し，その一般化により不完全データからの局所的な主成分の抽出法を定式化する．そして，Fuzzy $c$-Elliptotypes（FCE）法 [10] [11] が FCM 法と FCV 法の目的関数を重み付きで足し合わせることにより楕円体状のクラスターを得ているのに倣い，欠測値を含むデータを局所的な主成分を用いていくつかの楕円体状のクラスターに分割するクラスタリング法を提案する．クラスターの形状を楕円体状にすることにより，線形多様体からの距離だけでなくクラスター

中心からの距離も考慮したデータ分割が行われることになり，クラスター中心から遠く離れたデータ群を取り込むことを防ぎ，データの持つ局所的な特徴を抽出することができる．また，提案法には，主成分を抽出する際に固有値計算を用いず，FCM 法と同様の繰り返し計算のみで解が求まるというアルゴリズムの理解のしやすさの他に，因子分析と同様の解の不確定性という特性があり，プロクラステス直交回転法 [17] などを用いて主成分ベクトルを回転することにより，先見的知識を生かした分析への応用も可能である．数値例では，ワープロの便箋デザインに対する評価データの分析を通して提案法の有効性を示す．

## 2 欠測値を含むデータに対する主成分分析とクラスタリングの同時適用法

$m$ 次元の $n$ 個の標本データからなる（$n \times m$）データ行列 $X = (x_{ij})$ 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．ただし，データ行列 $X$ には欠測値が含まれるものとする．また，データ行列は適宜，変量 $j$ に係る $n$ 次元列ベクトル $\boldsymbol{x}_j$ を用いて $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_j, \cdots, \boldsymbol{x}_m)$，または標本 $i$ に係る $m$ 次元列ベクトル $\tilde{\boldsymbol{x}}_i$ を用いて $X = (\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_i, \cdots, \tilde{\boldsymbol{x}}_n)^{\mathrm{T}}$ と表す．（以降，太字はすべて列ベクトルを表し，行列の行方向の要素からなる列ベクトルは˜を付して表記することとする）

欠測値を含むデータに対してファジィクラスタリングを施す場合には，クラスターの形状にかかわらず適用可能な一般的な処理法は存在しない．データ集合を球状のクラスターに分割する FCM 法における欠測値の処理法としては，Miyamoto ら [14] がメンバシップのべき乗を用いる標準的な方法とエントロピー正則化 [15] を用いる方法の二つのバリエーションごとに，欠測値を無視して残りの座標で距離を定義する手法や欠測値に重み付きの平均値を代入する手法を提案している．また，Timm ら [16] も同様の考察を行い，欠測値を無視する手法の方が，代入する場合よりもより曖昧なメンバシップの割り当てを行う傾向があると報告している．

欠測値を無視することにより，エントロピー正則化を用いる FCM 法の目的関数は，次のように書き換えられる．

$$\begin{aligned} \psi &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \sum_{j=1}^{m} d_{ij}(x_{ij} - b_{cj})^2 \\ &\quad + \lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \log u_{ci} \end{aligned} \tag{1}$$

ここで，$\boldsymbol{b}_c = (b_{c1}, \cdots, b_{cm})^{\mathrm{T}}$ は第 $c$ クラスターの中心を表している．$d_{ij}$ はデータ行列 $X$ の $i$ 行 $j$ 列の要素 $x_{ij}$ が観測されているか否かを表す 2 値変数で，

$$d_{ij} = \begin{cases} 1 & ; x_{ij} \text{ is observed} \\ 0 & ; x_{ij} \text{ is missing} \end{cases} \tag{2}$$

と定義する．また，$u_{ci}$ は第 $i$ 標本データが第 $c$ クラスターに属する度合いを表すメンバシップで，

$$\sum_{c=1}^{C} u_{ci} = 1 \quad ; i = 1, \cdots, n \tag{3}$$

を満たすものとする．(1) 式の第 2 項は，ファジィ分割を得るために標準的な方法のメンバシップのべき乗の代わりに用いたエントロピー項で，$\lambda$ が大きくなるにしたがってより曖昧なデータ分割が得られるようになる．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などの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．

Miyamoto らにより提案された (1) 式を目的関数とする手法は，球状のクラスターを得る場合には有効であるが，目的関数がデータ点とプロトタイプとなる線形多様体との距離であらわされる FCV 法 [10] [11] などには適用することができない．そこで，本論文では，欠測値を含むデータから局所的な主成分を抽出することにより，データ集合をいくつかの楕円体状のクラスターに分割する手法を提案する．

### 2.1 最小 2 乗基準を用いた局所的な主成分分析

提案法の定式化の前に，まず，FCV 法の目的関数を主成分分析で用いられる最小 2 乗近似 [8] [9] に基づいて再定義することにより，局所的な主成分分析として

定式化する．エントロピー正則化を用いる FCV 法の目的関数は，長さが 1 で互いに直交する $p$ 本のベクトル $\boldsymbol{a}_{cj}$ により張られる線形多様体をおのおののクラスターのプロトタイプとして用いることにより，以下のように定義される．

$$\begin{aligned} L_{fcv} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\Big\{(\tilde{\boldsymbol{x}}_i-\boldsymbol{b}_c)^{\mathrm{T}}(\tilde{\boldsymbol{x}}_i-\boldsymbol{b}_c) \\ &\quad -\sum_{j=1}^{p}\boldsymbol{a}_{cj}^{\mathrm{T}}R_{ci}\boldsymbol{a}_{cj}\Big\} \\ &\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} \end{aligned} \tag{4}$$

$$R_{ci} = (\tilde{\boldsymbol{x}}_i-\boldsymbol{b}_c)(\tilde{\boldsymbol{x}}_i-\boldsymbol{b}_c)^{\mathrm{T}} \tag{5}$$

ここで，$^{\mathrm{T}}$ は行列の転置である．(4) 式を最小とする $\boldsymbol{a}_{cj}$ を求める問題は，最適性の必要条件 $\partial L_{fcv}/\partial \boldsymbol{a}_{cj}=\boldsymbol{0}$ から，

$$\textstyle\sum_c \boldsymbol{a}_{cj} = \mu_{cj}\boldsymbol{a}_{cj} \tag{6}$$

なる固有値問題に帰着される．ただし，$\sum_c$ はファジィ散布行列であり，

$$\textstyle\sum_c = \sum_{i=1}^{n} u_{ci}R_{ci} \tag{7}$$

で表される．目的関数を最小とする $\boldsymbol{a}_{cj}$ は，ファジィ散布行列の固有値のうち最大のものから $p$ 個に対応する固有ベクトルとして算出することができるので，メンバシップを考慮しながら抽出されたクラスターごとの局所的な主成分ベクトルであるとみなされる．また，クラスター中心およびメンバシップの更新則も，同様に $\partial L_{fcv}/\partial \boldsymbol{b}_c=\boldsymbol{0}$ および $\partial L_{fcv}/\partial u_{ci}=0$ から求めることができ，$\boldsymbol{a}_{cj}$，$\boldsymbol{b}_c$ および $u_{ci}$ を収束するまで繰り返し更新することにより，線形多様体状のクラスターにデータを分割できる．

本論文では，最小 2 乗基準を用いる主成分分析の目的関数にメンバシップ $u_{ci}$ とそのエントロピー正則化を考慮することにより，局所的な主成分の抽出法の目的関数を以下のように定式化する．

$$\begin{aligned} \varphi &= \sum_{c=1}^{C}\mathrm{tr}\Big\{(X-Y_c)^{\mathrm{T}}U_c(X-Y_c)\Big\} \\ &\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} \end{aligned} \tag{8}$$

ただし，$U_c$ は $u_{c1},\cdots,u_{cn}$ を対角要素とする $(n\times n)$ 対角行列 $U_c=\mathrm{diag}(u_{c1},\cdots,u_{cn})$ であり，tr は行列の対角要素の和（トレース）を表す．$Y_c=(y_{cij})$ は第 $c$ クラスターにおけるデータ行列 $X$ の低階数近似行列であり，主成分得点からなる $(n\times p)$ の成分得点行列 $F_c=(\tilde{\boldsymbol{f}}_{c1},\cdots,\tilde{\boldsymbol{f}}_{cn})^{\mathrm{T}}$ と $p$ 本の局所的な主成分ベクトルを並べた $(m\times p)$ の主成分行列 $A_c=(\boldsymbol{a}_{c1},\cdots,\boldsymbol{a}_{cp})$ を用いて，以下のように表されるものとする．

$$Y_c = F_cA_c^{\mathrm{T}} + \boldsymbol{1}_n\boldsymbol{b}_c^{\mathrm{T}} \tag{9}$$

ただし，$\boldsymbol{1}_n$ はすべての要素が 1 の $n$ 次元ベクトルである．

メンバシップを固定して考えた場合には，クラスターごとの局所的な主成分の抽出は，(8) 式の最小 2 乗基準が最小となる $F_c$，$A_c$ および $\boldsymbol{b}_c$ を求める問題に等しい．

(8) 式の最適性の必要条件 $\partial\varphi/\partial\boldsymbol{b}_c=\boldsymbol{0}$ から，クラスター中心 $\boldsymbol{b}_c$ は，

$$\boldsymbol{b}_c = (\boldsymbol{1}_n^{\mathrm{T}}U_c\boldsymbol{1}_n)^{-1}X^{\mathrm{T}}U_c\boldsymbol{1}_n \tag{10}$$

と求まる．これは FCV 法における $\boldsymbol{b}_c$ の更新則に等しい．更新後の $\boldsymbol{b}_c$ を (8) 式に代入すると，

$$\begin{aligned} \varphi &= \sum_{c=1}^{C}\Big\{\mathrm{tr}(X_c^{\mathrm{T}}U_cX_c) - 2\mathrm{tr}(X_c^{\mathrm{T}}U_cF_cA_c^{\mathrm{T}}) \\ &\quad +\mathrm{tr}(A_cF_c^{\mathrm{T}}U_cF_cA_c^{\mathrm{T}})\Big\} \\ &\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} \end{aligned} \tag{11}$$

となる．ただし，$X_c=X-\boldsymbol{1}_n\boldsymbol{b}_c^{\mathrm{T}}$ である．さらに，(11) 式について $F_c$ に関する最適性の必要条件 $\partial\varphi/\partial F_c=O$ を考慮することにより，

$$F_cA_c^{\mathrm{T}}A_c = X_cA_c \tag{12}$$

なる関係が導かれる．ここで，FCV 法と同様に，局所的な主成分ベクトル $\boldsymbol{a}_{cj}$ が互いに直交するという条件を付加すると，$A_c^{\mathrm{T}}A_c=I_p$ となることから，$F_c$ の最適解は $F_c=X_cA_c$ と求まる．したがって，局所的な主成分ベクトルを求めるための目的関数は，

$$\varphi = \sum_{c=1}^{C}\Big\{\mathrm{tr}(X_c^{\mathrm{T}}U_cX_c) - \mathrm{tr}(A_c^{\mathrm{T}}X_c^{\mathrm{T}}U_cX_cA_c)\Big\}$$

$$+\lambda\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci}$$

$$= L_{fcv} \tag{13}$$

となり (4) 式に等しくなるので，FCV 法と同様に $\boldsymbol{a}_{cj}$ は (6) 式の固有値問題の解として求められる．求まった $F_c$，$A_c$ および $\boldsymbol{b}_c$ を用いてメンバシップを更新する際には，(13) 式の関係から，FCV 法における更新則を用いることができる．以上のように，FCV 法の目的関数は，最小 2 乗基準を用いた局所的な主成分の抽出法として，(8) 式のように定式化することができる．

## 2.2 局所的な主成分を用いた欠測値を含むデータの楕円体状のクラスタリング

次に，FCV 法をデータ行列 $X$ に欠測値が含まれる場合に拡張することを考える．データ点とプロトタイプとなる線形多様体との距離で表される (4) 式の目的関数は，欠測値を含むデータ点と線形多様体との距離を定義できないため，不完全データに直接的に拡張することができない．そこで，(4) 式と等価な目的関数である (8) 式の最小 2 乗基準を，不完全データに拡張する．

欠測値を含むデータに対する主成分分析法として，柴山 [7] は，最小 2 乗基準を用いてデータ行列 $X$ の低階数近似行列 $Y = FA^{\mathrm{T}} + \mathbf{1}\boldsymbol{b}^{\mathrm{T}}$ を求める際に，目的関数，

$$\xi = \mathrm{tr}(E^{\mathrm{T}}E) \tag{14}$$

$$E = D\cdot(X - Y) \tag{15}$$

の最小化を考える方法を提案している．ただし，$D$ は (2) 式の 2 値変数を要素とする $(n \times m)$ 行列であり，$\cdot$ はアダマール積を表す．この方法は，$X$ の観測された部分に対してはそれと対応する $Y$ の要素がなるべく一致するようにするものの，欠測部分に関しては変化するに任せることを意味している．

本論文でも同様に，(8) 式の最小 2 乗基準に観測されているか否かを表す 2 値変数を導入することにより，欠測値を含むデータから局所的な主成分を抽出する．しかし，$X$ の欠測値に対応する $Y_c$ の要素を変化するに任せるということは，すべてのクラスターにおいて，欠測値を含むデータ点があたかも局所的な主成分ベクトルにより張られる線形多様体上に存在するか，あるいは線形多様体からの距離が最も近くなる点に存在するかのように，欠測値に対応する $Y_c$ の要素を推定することに他ならない．したがって，欠測値を含むデータ点のうちでプロトタイプ上に存在するとみなされるものについては，プロトタイプからの距離に基づいてメンバシップを定める際に，データとすべてのプロトタイプとの距離が 0 となることがあり得る．そこで，Bezdek らが FCV 法と FCM 法の目的関数を重み付きで足し合わせることにより，楕円体状のクラスターが得られる FCE 法 [10] [11] を定式化したのに倣い，(8) 式の最小 2 乗基準と (1) 式の目的関数を重み付きで足し合わせることにより，以下の目的関数を定式化する．

$$\begin{aligned} L &= \alpha\varphi + (1-\alpha)\psi \\ &= \sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\sum_{j=1}^{m}d_{ij}\Big\{\alpha(x_{ij} - \sum_{k=1}^{p}f_{cik}a_{cjk} \\ &\quad -b_{cj})^2 + (1-\alpha)(x_{ij} - b_{cj})^2\Big\} \\ &\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci} \end{aligned} \tag{16}$$

ここで，$\alpha$ は局所的な主成分分析と FCM 法の優先度を規定するトレードオフ係数で，0 のときは球状のクラスターが得られる (1) 式と等しく，1 に近づけるにしたがって楕円体状のクラスターが得られるようになる．クラスター中心からの距離の最小化を同時に考えることにより，欠測値を含むデータ点に対しても適切なメンバシップを割り振ることができる．ただし，唯一の解を得るために，

$$F_c^{\mathrm{T}}U_cF_c = I_p \quad ; \quad c = 1,\cdots,C \tag{17}$$

$$F_c^{\mathrm{T}}\mathbf{1}_n = O \quad ; \quad c = 1,\cdots,C \tag{18}$$

$$\sum_{c=1}^{C}u_{ci} = 1 \quad ; \quad i = 1,\cdots,n \tag{19}$$

および，$A_c^{\mathrm{T}}A_c$ が対角行列となる制約条件を付加する．

(16) 式の目的関数を最小とするパラメータを同定する際には，標本データに欠測値が含まれているために，FCV 法のように固有値問題に帰着させることができ

ず，交互最小 2 乗法に基づいて最適化を行う必要がある．本論文では，欠測値を含むデータの主成分分析における柴山 [18] の手法に倣い，目的関数を変形することにより，おのおののパラメータの更新則を求める．

まず，$A_c$ および $\boldsymbol{b}_c$ の更新則を求めるために，(16) 式の目的関数を以下のように書き換える．

$$\begin{aligned} L &= \sum_{c=1}^{C}\sum_{j=1}^{m}\Big\{\alpha(\boldsymbol{x}_j - F_c\tilde{\boldsymbol{a}}_{cj} - \mathbf{1}_n b_{cj})^{\mathrm{T}} U_c D_j \\ &\quad \times(\boldsymbol{x}_j - F_c\tilde{\boldsymbol{a}}_{cj} - \mathbf{1}_n b_{cj}) \\ &\quad +(1-\alpha)(\boldsymbol{x}_j - \mathbf{1}_n b_{cj})^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \mathbf{1}_n b_{cj})\Big\} \\ &\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} \end{aligned} \tag{20}$$

ただし，

$$\begin{aligned} A_c &= (\tilde{\boldsymbol{a}}_{c1},\cdots,\tilde{\boldsymbol{a}}_{cm})^{\mathrm{T}} \\ D_j &= \mathrm{diag}(d_{1j},\cdots,d_{nj}) \end{aligned}$$

である．最適性の必要条件 $\partial L/\partial\tilde{\boldsymbol{a}}_{cj} = \mathbf{0}$ および $\partial L/\partial b_{cj} = 0$ から，更新則は，

$$\tilde{\boldsymbol{a}}_{cj} = (F_c^{\mathrm{T}} U_c D_j F_c)^{-1} F_c^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \mathbf{1}_n b_{cj}) \tag{21}$$

$$b_{cj} = (\mathbf{1}_n^{\mathrm{T}} U_c D_j \mathbf{1}_n)^{-1} \mathbf{1}_n^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \alpha F_c\tilde{\boldsymbol{a}}_{cj}) \tag{22}$$

のように求まる．

同様に，$F_c$ および $u_{ci}$ についても，(16) 式を，

$$\begin{aligned} L &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\Big\{\alpha(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)^{\mathrm{T}}\tilde{D}_i \\ &\quad \times(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c) \\ &\quad +(1-\alpha)(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)^{\mathrm{T}} D_i (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)\Big\} \\ &\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} \end{aligned} \tag{23}$$

のように変形すると，$\partial L/\partial\tilde{\boldsymbol{f}}_{ci} = \mathbf{0}$ および $\partial L/\partial u_{ci} = 0$ から，

$$\tilde{\boldsymbol{f}}_{ci} = (A_c^{\mathrm{T}}\tilde{D}_i A_c)^{-1} A_c^{\mathrm{T}}\tilde{D}_i(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c), \tag{24}$$

$$\begin{aligned} u_{ci} &= \exp\Big\{-\Big(\alpha(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)^{\mathrm{T}} \\ &\quad \times\tilde{D}_i(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c) \\ &\quad +(1-\alpha)(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)^{\mathrm{T}}\tilde{D}_i(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)\Big)/\lambda - 1\Big\} \end{aligned} \tag{25}$$

が導かれる．ただし，

$$\tilde{D}_i = \mathrm{diag}(d_{i1},\cdots,d_{im})$$

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．

Step1 クラスターごとに $U_c, A_c, \boldsymbol{b}_c, F_c$ を乱数を用いて初期化し，(17)-(19) 式および $A_c^{\mathrm{T}}A_c$ が対角行列となる制約条件を満たすように基準化する．

Step2 (21) 式を用いてクラスターごとに $A_c$ を更新し，$A_c^{\mathrm{T}}A_c$ が対角行列となるように基準化する．

Step3 (24) 式を用いてクラスターごとに $F_c$ を更新し，(17)，(18) 式の制約条件を満たすように基準化する．

Step4 (22) 式を用いてクラスターごとに $\boldsymbol{b}_c$ を更新する．

Step5 (25) 式を用いてクラスターごとに $U_c$ を更新し，(19) 式の制約条件を満たすように基準化する．

Step6 終了判定条件

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

を満たせば終了．それ以外は Step2 へ戻る．

ここで，得られた成分得点行列 $F_c$ および主成分行列 $A_c$ には，因子分析と同様の解の不確定性がある．今，$T$ を，

$$T^{\mathrm{T}}T = TT^{\mathrm{T}} = I_p \tag{26}$$

を満たす任意の正規直交行列とすると，

$$\begin{aligned} F_c^* &= F_c T \\ A_c^* &= A_c T \end{aligned}$$

と変形しても，$F_c^*$ および $A_c^*$ は (17)，(19) 式の制約条件を満たし，

$$\begin{aligned} Y_c &= F_c^* A_c^{*\mathrm{T}} + \mathbf{1}_n \boldsymbol{b}_c^{\mathrm{T}} \\ &= F_c^{\mathrm{T}} T T^{\mathrm{T}} A_c^{\mathrm{T}} + \mathbf{1}_n \boldsymbol{b}_c^{\mathrm{T}} \\ &= F_c A_c^{\mathrm{T}} + \mathbf{1}_n \boldsymbol{b}_c^{\mathrm{T}} \end{aligned} \tag{27}$$

から目的関数の値を変化させないので，$F_c$ および $A_c$ を任意の正規直交行列で回転できることが分かる．このように，提案法で求まる局所的な主成分ベクトルには回転の自由度があり，事前知識の検証などに応用することができる．

# 3　数値例

本章では，提案手法の有効性を検証するために，以下の数値例を示す．

## 3.1　感性評価データの分析

実世界で収集されたデータへの応用として，感性評価データの分析を行った．用いた感性データは 285 人の被験者が図 1 のワープロの便箋用背景デザイン 8 種類を評価したデータで，図 2 の評価尺度による SD（semantic differential）法により，それぞれのデザインが 7 段階評価されている．この 285 人分のデータについて，すべての被験者のデータからランダムに二つずつ値を欠測させることにより，データ行列に 25%の欠測値が含まれる標本データを作成して，提案法による分析を行った．ただし，$\alpha$ および $\lambda$ はそれぞれ 0.5, 2.0 とし，クラスター数は 2，クラスターごとの主成分ベクトルの数は 2 とした．すべてのデータが欠測値を含んでいるために，欠測値を含むデータを取り除く方法では，FCE 法などの従来法による分析ができない例となっている．クラスターごとに得られた局所的な主成分行列 $A_c$ を表 1 および表 2 に示す．表から，クラスター 1 ではデザイン 1 がデザイン 3，4 および 7 と似た傾向があるのに対して，クラスター 2 ではデザイン 8 がデザイン 3，4 および 7 と似た傾向があることが分かる．このように，285 人の被験者がデザイン 1 と 8 について相反する傾向を持つ 2 つのクラスターに分割されることが覗える．

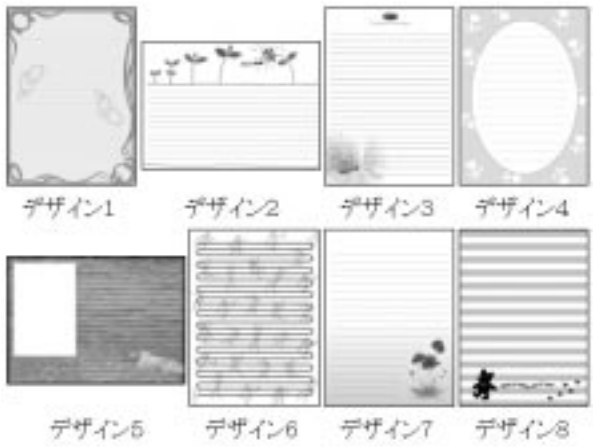


図 1: 8 種類の便箋用背景デザイン

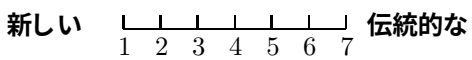


図 2: 7 段階の評価尺度

表 1: 局所的な主成分行列 (クラスター 1)

| デザイン | $A_1$ | |
|---|---|---|
| 1 | -2.02 | 12.88 |
| 2 | 7.37 | 4.03 |
| 3 | -10.92 | 4.52 |
| 4 | -0.98 | 11.58 |
| 5 | 9.46 | 1.30 |
| 6 | 7.33 | -6.17 |
| 7 | -10.92 | -1.38 |
| 8 | 9.48 | 7.89 |

そこで，得られた仮説を提案法の解の不確定性を利用して検証する．仮説を目標行列として表し，因子分析で用いられるプロクラステス直交回転法 [17] により変換した結果を表 3 および 4 に示す．回転後の主成分行列 $A_c^*$ から，デザイン 1 と 8 がクラスターごとに相反する傾向を持つことがよりはっきりと確認できる．このように，提案法の解の不確定性を用いることで，得られた結果に対するより詳細な分析が可能となる．

表 2: 局所的な主成分行列 (クラスター 2)

| デザイン | $A_2$ | |
|---|---|---|
| 1 | 3.70 | 11.34 |
| 2 | 4.12 | 7.40 |
| 3 | 8.89 | 4.38 |
| 4 | 14.74 | 0.87 |
| 5 | 3.18 | 10.84 |
| 6 | 4.48 | 9.02 |
| 7 | 17.40 | -11.61 |
| 8 | 7.22 | 0.40 |

表 3: プロクラステス直交回転の結果 (クラスター 1)

| デザイン | 目標行列 | | $A_1^*$ | |
|---|---|---|---|---|
| 1 | 0 | 1 | 4.16 | 12.36 |
| 2 | 1 | 0 | 8.40 | 0.17 |
| 3 | 0 | 1 | -7.60 | 9.07 |
| 4 | 0 | 1 | 4.49 | 10.73 |
| 5 | 1 | 0 | 9.00 | -3.21 |
| 6 | 1 | 0 | 3.65 | -8.86 |
| 7 | 0 | 1 | -10.33 | 3.82 |
| 8 | 1 | 0 | 12.06 | 2.62 |

表 4: プロクラステス直交回転の結果 (クラスター 2)

| デザイン | 目標行列 | | $A_2^*$ | |
|---|---|---|---|---|
| 1 | 0 | 1 | 4.29 | 11.13 |
| 2 | 0 | 1 | 4.50 | 7.17 |
| 3 | 1 | 0 | 9.11 | 3.91 |
| 4 | 1 | 0 | 14.76 | 0.10 |
| 5 | 0 | 1 | 3.74 | 10.66 |
| 6 | 0 | 1 | 4.94 | 8.77 |
| 7 | 1 | 0 | 16.77 | -12.50 |
| 8 | 1 | 0 | 7.23 | 0.02 |

## 3.2 欠測値を含まない場合との比較

次に，得られたクラスタリング結果の妥当性を検証するために，欠測値を含まない場合との比較を行った．値を欠測させる前の欠測値を含まない感性評価データに対して，提案法を適用した．ただし，欠測値がある場合と同様に，$\alpha$ および $\lambda$ はそれぞれ 0.5，2.0 とし，クラスター数は 2，クラスターごとの主成分ベクトルの数は 2 とした．この結果は，エントロピー正則化を導入した FCE 法において，得られた解を (17)〜(19) 式の関係が満たされるように基準化したものに等しい．欠測値がある場合とない場合のクラスター中心の比較を表 5 に，欠測値がない場合のクラスターごとの局所的な主成分行列 $A_c$ を表 6 および 7 に示す．

表 5: クラスター中心の比較

| デザイン | $\boldsymbol{b}_c$ | | | |
|---|---|---|---|---|
| | 欠測値なし | | 欠測値あり | |
| | c=1 | c=2 | c=1 | c=2 |
| 1 | 3.67 | 3.14 | 3.74 | 3.14 |
| 2 | 2.74 | 2.80 | 2.75 | 2.82 |
| 3 | 4.26 | 4.12 | 4.51 | 4.08 |
| 4 | 3.61 | 4.10 | 3.64 | 4.19 |
| 5 | 2.82 | 2.73 | 3.03 | 2.59 |
| 6 | 2.93 | 3.18 | 3.24 | 2.92 |
| 7 | 5.00 | 4.22 | 5.00 | 4.36 |
| 8 | 2.84 | 2.99 | 3.00 | 2.76 |

表 6: 欠測値がない場合の主成分行列 (クラスター 1)

| デザイン | $A_1$ | |
|---|---|---|
| 1 | 13.94 | -0.30 |
| 2 | 2.05 | 6.90 |
| 3 | 10.68 | -4.59 |
| 4 | 11.66 | 3.26 |
| 5 | 4.32 | 10.53 |
| 6 | 0.12 | 10.47 |
| 7 | 5.20 | -8.51 |
| 8 | -0.18 | 8.62 |

表 5 から欠測値の有無にかかわらずほぼ等しいクラスター中心を持つことが，また表 6 および 7 から局所的な主成分ベクトルが類似した特徴を持つことがわかる．このように，提案法を用いることで，欠測値を含むデータに対しても欠測値を含まない場合と同様の妥当なクラスタリング結果が得られることが確認された．

表 7: 欠測値がない場合の主成分行列 (クラスター 2)

| デザイン | $A_2$ | |
|---|---|---|
| 1 | 3.61 | 7.83 |
| 2 | 5.98 | 7.10 |
| 3 | 6.37 | -4.53 |
| 4 | 11.15 | -6.99 |
| 5 | 3.51 | 9.73 |
| 6 | 4.84 | 5.19 |
| 7 | 12.05 | -4.36 |
| 8 | 9.04 | 3.25 |

## 4 おわりに

本論文では，主成分分析とファジィクラスタリングの同時分析における欠測値の処理法を提案した．提案手法で用いた最小 2 乗基準による局所的な主成分分析のための目的関数は，データに欠測値が含まれない場合には FCV 法の目的関数と等価なものであることから，提案手法は FCV 法および FCE 法の欠測値を含む不完全データへの拡張手法であるといえる．また，パラメータを順次更新することにより解を求める繰り返しアルゴリズムは，固有値計算を必要とするアルゴリズムに比べて理解が容易であるという利点もある．ただし，提案法においても従来法と同様に，主成分分析の重み $\alpha$ や，ファジィ度を制御するパラメータ $\lambda$ を試行錯誤的に決定する必要がある．FCE 法において重みを変化させて適応的にクラスター形状を決定する手法の研究もいくつかなされている [19] [20] が，ファジィ散布行列の固有値を利用するため，提案法においては別のアプローチが必要となる．

提案法においておのおののクラスターで得られるデータ行列の近似行列 $Y_c$ には欠測値が含まれないことから，データ行列中の欠測値の推定値として近似行列の対応する要素を採用することができる．その際には，欠測値を含むデータはクラスター中心と局所的な主成分ベクトルによって張られる線形多様体上に存在すると仮定することに等しい．欠測値の推定法としての有効性の検証と，その応用が今後の課題である．

## 謝　辞

本研究の一部は文部科学省科学研究費補助金・基盤研究 (C)(13680375) の助成に基づいて行われたものであり，謝意を表する．

## 参考文献

[1] 田中　豊, 脇本和昌: 多変量統計解析法, 現代数学社 (1983)

[2] 竹内　啓: 統計学辞典, 東洋経済新報社 (1989)

[3] A.P. Dempster, N.M. Laird and D.B. Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm; *J. of the Royal Statistical Society*, Series B, vol. 39, pp. 1–38 (1977)

[4] A. Ruhe: Numerical Computation of Principal Components when Several Observations are Missing; *Tech Rep. UMINF-48-74, Dept. Information Processing*, Umea Univ. (1974)

[5] T. Wiberg: Computation of Principal Components when Data are Missing; *Proc. of Second Symp. Computational Statistics*, pp. 229–236 (1976)

[6] H. Shum, K. Ikeuchi and R. Reddy: Principal Component Analysis with Missing Data and its Application to Polyhedral Object Modeling; *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 9, pp. 854–867 (1995)

[7] 柴山　直: 欠測値を含む多変量データのための主成分分析的方法; 教育心理学研究, Vol. 40, No. 3, pp. 257–265 (1992)

[8] 丘本　正: 因子分析の基礎, 日科技連 (1986)

[9] 高根芳雄: 制約つき主成分分析法, 朝倉書店 (1995)

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

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

[12] 本多克宏, 山川あす香, 市橋秀友, 三好哲也, 奥山哲史: ファジィクラスタリングと回帰と主成分の同時分析法; システム制御情報学会論文誌, Vol. 13, No. 5, pp. 236–243 (2000)

[13] 山川あす香, 市橋秀友, 三好哲也: 正準相関係数を最大とするファジィc-Means クラスタリング法; 日本経営工学会論文誌, Vol. 51, pp. 17–26 (2000)

[14] S. Miyamoto, O. Takata and K. Umayahara: Handling Missing Values in Fuzzy c-Means;*Proc. of Third Asian Fuzzy Systems Symp.*, pp. 139–142 (1998)

[15] 宮本定明: クラスター分析入門, 森北出版 (1999)

[16] H. Timm and R. Kruse: Fuzzy Cluster Analysis with Missing Values; *Proc. of 17th International Conf. of the North American Fuzzy Information Processing Society*, pp. 242–246 (1998)

[17] 柳井晴夫: 多変量データ解析法, 朝倉書店 (1994)

[18] 柴山　直: 欠測値を含む多変量データのための主成分分析プログラム; 教育心理フォーラムレポート, FR-96-003 (1996)

[19] R. N. Dave: An Adaptive Fuzzy $c$-Elliptotype Clustering Algorithm; *Proc. of the North American Fuzzy Information Processing Society :Quater Century of Fuzziness*, Vol. 1, pp. 9–12 (1990)

[20] 馬屋原一孝, 中森義輝: 線形多様体クラスタリングと楕円形ファジィモデル; 日本ファジィ学会誌, Vol. 10, No. 1, pp. 142–149 (1998)

［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．

杉浦　伸和　(すぎうら　のぶかず)

2001 年大阪府立大学大学院工学研究科博士前期課程電気・情報系専攻入学，現在在学中．ファジィクラスタリング，ファジィモデリングに興味を持つ．

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．

荒木　昭一　(あらき　しょういち) [正会員]

1989 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．1993〜1997 年同社より奈良先端科学技術大学院大学情報科学研究科博士課程派遣留学．2001 年同社先端技術研究所主任研究員，現在に至る．博士（工学）．ユーザモデリング，パーソナライゼーションの研究に従事．1992 年電気関係学会関西支部連合大会奨励賞，1995 年電子情報通信学会学術奨励賞受賞．電子情報通信学会の会員．

九津見　洋　(くつみ　ひろし)

1994 年京都大学工学研究科精密工学専攻修士課程修了．同年松下電器産業（株）入社．中央研究所（現先端技術研究所）に勤務，現在に至る．非線形情報処理，感性情報処理，ユーザモデリングの研究に従事．1997 年電気関係学会関西支部連合大会奨励賞受賞．電子情報通信学会，計測自動制御学会，日本感性工学会の会員．