支配的要因と独立な主成分を抽出する ファジィクラスタリング法*

(Fuzzy Clustering for Extracting Principal Components Independent of Dominant Factors)

呉　志賢†・本多　克宏†・小松　裕和†・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

Fuzzy c-Varieties（FCV 法）は，データをいくつかの線形多様体状に分類するもので，主成分分析とクラスター分析の同時分析法といえる．そのために，FCV 法は局所的な主成分を発見できる可能性を秘めているが，比較的自明な主成分を抽出することが多い．しばしば，支配的で自明な主成分よりも，データの下に隠された潜在的な主成分の方が知識として重要な場合がある．本研究では，獲得される主成分に影響を与える支配的要因を補助変数と考えることにより，それらと独立な主成分を抽出する方法を基にしたファジィクラスタリング法を提案する．数値実験として，あるスーパーマーケットの POS データと気象データに提案法を適用することにより，天候要因の販売に及ぼす影響を分析し，提案法の有効性を検証する．

(Fuzzy clustering algorithms are useful vehicles to search for structure in data sets by handling fuzzy clusters and have a lot of varieties. Fuzzy c-varieties (FCV) is one of those algorithms in which the prototypes are multi-dimensional linear varieties. The linear varieties are spanned by some local principal component vectors and the FCV clustering algorithm can be regarded as a simultaneous algorithm of fuzzy clustering and principal component analysis. Even though the FCV has the advantage of finding local principal components, they are sometimes strongly influenced by the dominant factors. To eliminate the influence, we propose a new method of fuzzy clustering which extracts local principal components independent of subsidiary variables. In the algorithm, subsidiary variables are regarded as dominant factors. A certain constraint which represents that principal components and subsidiary variables are uncorrelated is added to the objective function for the sake of realization of the proposed method. The solution algorithm is based on an iterative procedure through necessary conditions of optimality of Lagragian function. As numerical examples, first we apply the conventional FCV and the proposed method to an artificial data set and examine their performance. Next, we apply them to a POS (Point of Sales) transaction data set in order to discover associations among items without being influenced by the explicit dominant factors.)

keywords: Fuzzy clustering, principal component analysis, POS transaction.

1　はじめに

Bezdek らによって提案された Fuzzy c-Varieties (FCV) 法 [1] は，データをいくつかの線形多様体状に分類するクラスタリングの一手法である．また，その改良として，クラスターごとの分散共分散

*日本経営工学会論文誌, 52, 20-29 (2001)

行列の固有値をトレードオフパラメータに採用した Adaptive Fuzzy c-Elliptotypes (AFC) 法 [2] や，次元の異なる多様体の発見法である Fuzzy c-Varieties of Different Dimensionalities (FCD) 法 [3] などが提案されている．FCV 法のプロトタイプは線形多様体であるので，これらを主成分ベクトルと考えると，FCV 法はクラスタリングと主成分分析の同時分析とみなすことができる．FCV 法により得られた主成分は，クラスタリングを同時に行って計算されたものなので，データ集合の局所性を加味したものとなっている．しかし，局所的な主成分であっても支配的な要因がある場合，主成分はそれらの影響を多大に受けていることがしばしばある．データ集合を解析する上で，支配的ではあるが自明な主成分よりも，むしろ隠れた潜在的な主成分の方が知識として重要な場合もある．

補助変数とは無相関で，かつ，もとの変数の持つ情報をできるだけ多く取り込んだ主成分を得る方法として，補助変数に独立な主成分の抽出法がある [4]．本研究では，補助変数と独立な主成分の抽出法とファジィクラスタリングを同時に行う手法を提案する．補助変数をデータ集合における支配的要因とすると，提案手法により，支配的要因に独立な潜在的，かつ，局所的主成分を抽出することが可能となる．数値実験では，提案手法を人工的に作成したデータに適用することにより，有効性を検証する．また，スーパーマーケットの販売時点情報管理システムから収集される POS データに気象データを加えて分析することにより，従来の FCV 法との比較を行う．

2　クラスターと主成分の同時分析法

前述のとおり，FCV 法ではクラスタリングと主成分分析を同時に行うので，局所的主成分を抽出することができる．まず，以下のように記号を定義する．

与えられたデータ：$\boldsymbol{x}_k$
$\boldsymbol{x}_k = (x_{k1}, x_{k2}, \cdots, x_{kI})^{\mathrm{T}}, \quad k = 1, \cdots, n$
各クラスターに対するメンバシップ値：$\boldsymbol{u}_c$
$\boldsymbol{u}_c = (u_{c1}, u_{c2}, \cdots, u_{cn})^{\mathrm{T}}, \quad c = 1, \cdots, C$
各クラスターの中心：$\boldsymbol{v}_c$
$\boldsymbol{v}_c = (v_{c1}, v_{c2}, \cdots, v_{cI})^{\mathrm{T}}, \quad c = 1, \cdots, C$
クラスターごとの主成分ベクトル：$\boldsymbol{a}_{ci}$
$\boldsymbol{a}_{ci} = (a_{ci1}, a_{ci2}, \cdots, a_{ciI})^{\mathrm{T}},$
$c = 1, \cdots, C, \quad i = 1, \cdots, m$

ここで，I は与えられたデータの特性値の数（次元数），n はデータの個数，C はクラスターの数，m は求める主成分の個数を表す．また，データベクトル $\boldsymbol{x}_k$ の各成分は平均 0，分散 1 に基準化されているとする．メンバシップ値 $\boldsymbol{u}_c$ は各データがクラスターに属する度合いで，

$$\boldsymbol{u}_c = \{(u_{ck}) | \sum_{c=1}^{C} u_{ck} = 1,\ u_{ck} \in [0, 1]\}$$

である．

FCV 法は，データからクラスター中心までの 2 乗距離を最小にし，かつ，クラスター中心を通る線形多様体上に射影されたデータ点とクラスター中心からの 2 乗距離を最大にするという最適化問題を解くことによって実現される．そのために，ラグランジュの未定乗数法により，以下の評価規範（目的関数）L の最小化問題を考える．

$$\begin{aligned} L &= \sum_{k=1}^{n}\sum_{c=1}^{C} u_{ck}\Big\{(\boldsymbol{x}_k - \boldsymbol{v}_c)^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{v}_c) \\ &- \alpha\sum_{i=1}^{m} \boldsymbol{a}_{ci}^{\mathrm{T}} B_{ck} \boldsymbol{a}_{ci}\Big\} \\ &+ \beta\sum_{k=1}^{n}\sum_{c=1}^{C} u_{ck}\log u_{ck} \\ &+ \sum_{c=1}^{C}\sum_{i=1}^{m} \lambda_{ci}(\boldsymbol{a}_{ci}^{\mathrm{T}}\boldsymbol{a}_{ci} - 1) \end{aligned}$$

$$+ \quad \sum_{k=1}^{n} \gamma_k \Big(\sum_{c=1}^{C} u_{ck} - 1 \Big) \tag{1}$$

ここで,

$$B_{ck} = (\boldsymbol{x}_k - \boldsymbol{v}_c)(\boldsymbol{x}_k - \boldsymbol{v}_c)^{\mathrm{T}} \tag{2}$$

である．FCV 法での線形多様体 (variety) とは，クラスター中心を通る平面や直線に代表されるようなベクトルの張る線形空間を言う．

本研究では，従来の FCV 法におけるメンバシップ値のべき乗の代わりにエントロピー正則化 [5] [6] を採用している．このことにより，クラスター中心や線形多様体とデータ点が重なった場合であっても，従来のべき乗法で用いられる例外処理を必要とせず，アルゴリズムを簡便にすることができる．また，エントロピー正則化を用いた FCV 法では，データが遠くに離れるほどクラスターに属する度合いがクリスプになる．一方，従来法では，クラスターから遠くに離れたデータは属する度合いがあいまいになる [6].

目的関数 L の第 1 項目の $(\boldsymbol{x}_k - \boldsymbol{v}_c)^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{v}_c)$ は，データ $\boldsymbol{x}_k$ とクラスター中心 $\boldsymbol{v}_c$ との 2 乗距離を表し，これを最小化することでクラスタリングを行う．また，$\boldsymbol{a}_{ci}^{\mathrm{T}} B_{ck} \boldsymbol{a}_{ci}$ は,

$$\begin{aligned} \boldsymbol{a}_{ci}^{\mathrm{T}} B_{ck} \boldsymbol{a}_{ci} &= \boldsymbol{a}_{ci}^{\mathrm{T}} (\boldsymbol{x}_k - \boldsymbol{v}_c)(\boldsymbol{x}_k - \boldsymbol{v}_c)^{\mathrm{T}} \boldsymbol{a}_{ci} \\ &= |(\boldsymbol{x}_k - \boldsymbol{v}_c)^{\mathrm{T}} \boldsymbol{a}_{ci}|^2 \end{aligned} \tag{3}$$

のように表すことができ，$\boldsymbol{a}_{ci}$ により定まる主成分軸に射影されたデータとクラスター中心（平均）との 2 乗距離となる．したがって，$\boldsymbol{a}_{ci}^{\mathrm{T}} B_{ck} \boldsymbol{a}_{ci}$ を最大化することは主成分ベクトル $\boldsymbol{a}_{ci}$ を求めることになる．α は主成分分析に対する重み係数で，$\alpha = 0$ のときは Fuzzy c-Means 法 [7] を表すことになる．$\alpha = 1$ の場合は，第 1 項目の総和はデータ点から線形多様体までの距離の和に一致する．また，$\alpha = 1 - \lambda_{c\ m+1}/\lambda_{c1}$ とすると AFC 法 [2] を表すことになる．第 2 項目はファジィクラスターを得るためのエントロピー項で，β は重み係数である．第 3 項目は主成分ベクトルの長さを 1 にすることを表し，第 4 項目はメンバシップ値の和が 1 である制約を表している．λ_{ci}，γ_k はラグランジュ乗数である．

目的関数 L を最小にするような $\boldsymbol{v}_c$，$\boldsymbol{u}_c$ を求めるための最適性の必要条件 $\partial L/\partial v_{ci} = 0$ と $\partial L/\partial u_{ck} = 0$ から,

$$v_{ci} = \frac{\displaystyle\sum_{k=1}^{n} u_{ck} x_{ki}}{\displaystyle\sum_{k=1}^{n} u_{ck}} \tag{4}$$

$$u_{ck} = \frac{\exp A_{ck}}{\displaystyle\sum_{a=1}^{C} \exp A_{ak}} \tag{5}$$

$$\begin{aligned} A_{ak} = -\{ &\|\boldsymbol{x}_k - \boldsymbol{v}_a\|^2 \\ &-\alpha \sum_{i=1}^{m} \boldsymbol{a}_{ai}^{\mathrm{T}} B_{ak} \boldsymbol{a}_{ai} \} / \beta \end{aligned} \tag{6}$$

を得る．また，$\boldsymbol{a}_{ci}$ についての最適性の必要条件 $\partial L/\partial \boldsymbol{a}_{ci} = \boldsymbol{0}$ から,

$$\sum\nolimits_c \boldsymbol{a}_{ci} = \mu_{ci} \boldsymbol{a}_{ci} \tag{7}$$

$$\mu_{ci} = \frac{\lambda_{ci}}{\alpha} \tag{8}$$

を得る．ここで，$\sum_c$ はファジィ散布行列と呼ばれるもので,

$$\sum\nolimits_c = \sum_{k=1}^{n} u_{ck} B_{ck} \tag{9}$$

である．また，$\partial L/\partial\lambda_{ci}=0$ から，

$$\boldsymbol{a}_{ci}^{\mathrm{T}}\boldsymbol{a}_{ci}=1 \tag{10}$$

であるので，式 (7) より，

$$\boldsymbol{a}_{ci}^{\mathrm{T}}\sum_{c}\boldsymbol{a}_{ci}=\mu_{ci} \tag{11}$$

となる．したがって，$\boldsymbol{a}_{ci}$ は，式 (7) で表される固有値問題の固有値 μ_{ci} に対応する固有ベクトルであり，目的関数 L を最小にするような $\boldsymbol{a}_{ci}$ を求めるには，最大のものから m 個の固有値に対応する固有ベクトルをそれぞれ求めればよい ($\{\boldsymbol{a}_{ci} \mid \mu_{c1}\geq\cdots\geq\mu_{cm},\ c=1,\cdots,C,\ i=1,\cdots,m\}$)．FCV 法のアルゴリズムを以下に記述する．

Step 1 定数 $\alpha,\ \beta,\ C,\ m$ を定める．　メンバシップ u_{ck} の初期値を乱数により決定する．

Step 2 式 (4) より，クラスター中心ベクトル $\boldsymbol{v}_c$ を求める．

Step 3 式 (7) より主成分ベクトル $\boldsymbol{a}_{ci}$ を求める．

Step 4 式 (5) よりメンバシップ値 u_{ck} を求める．

Step 5 終了判定条件 $\max_{c,k}\{|u_{ck}^{NEW}-u_{ck}^{OLD}|\}<\varepsilon$ を満たせば終了．そうでなければ *Step 2* へ．

2.1　補助変数に独立な主成分を抽出するファジィクラスタリング法

分析の対象となるデータに主成分分析を適用する際に，ある変数の影響が非常に大きく働いており，自明な主成分しか求められないような場合に，支配的な変数の影響を取り除いた分析をすることにより，従来の主成分分析では求められない潜在的な主成分を得ることができる．ここで，影響を取り除くという意味は，データからその変数自体を取り除くことではなく，その変数と無相関な主成分を求めることである．本研究では，n 個の標本データを C 個のクラスターに分けるクラスター分析と，補助変数に独立な主成分の抽出を行う分析を組み合わせ，それらを同時に分析する手法を提案する．補助変数をデータの支配的要因とすると，提案手法により，支配的要因の影響を取り除いた潜在的，局所的主成分を抽出することが出来る．

p 個の補助変数 $\boldsymbol{x}_k$ で説明される q 種類の目的変数を $\boldsymbol{y}_k$ とし，これらのデータが n 個得られたとする．ただし，$p<q$ とする．

$$\left(X \,\middle|\, Y\right)=\left(\begin{array}{ccc|ccc} x_{11} & \cdots & x_{1p} & y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} & y_{n1} & \cdots & y_{nq} \end{array}\right) \tag{12}$$

$(\in R^{n\times(p+q)})$

X は補助変数，Y は目的変数であり，各変数は平均 0，分散 1 に基準化されているものとする．

ここで，n 個の標本データを C 個のクラスターに分割しながら，クラスターごとに変数 $\boldsymbol{y}_k$ の主成分を求めることを考える．ただし，主成分はすべて，補助変数 $\boldsymbol{x}_k$ と無相関となるように選ぶ．すなわち，クラスター c に属するデータに関して，変数 $\boldsymbol{y}_k$ の m 個の主成分 $\boldsymbol{z_c}$

$$\begin{aligned} z_{c1} &= a_{c11}y_1+a_{c12}y_2+\cdots+a_{c1q}y_q \\ &\vdots \\ z_{cm} &= a_{cm1}y_1+a_{cm2}y_2+\cdots+a_{cmq}y_q \end{aligned} \tag{13}$$

を求めたときに，クラスター内で補助変数との共分散が

$$\begin{aligned} cov[z_{ci},x_r] &= cov\left[\sum_{s=1}^{q}a_{cis}y_s,\ x_r\right]=0 \\ & i=1,\cdots,m,\ r=1,\cdots,p \end{aligned} \tag{14}$$

という条件の下で主成分を求める．

補助変数に独立な主成分の抽出とファジィクラスタリングを同時に行うために，以下のラグランジュ関数の最小化問題を考える．

$$
\begin{aligned}
L &= \sum_{c=1}^{C}\sum_{k=1}^{n} u_{ck}\Big\{ ||\boldsymbol{x}_k - \boldsymbol{v}_c^x||^2 + ||\boldsymbol{y}_k - \boldsymbol{v}_c^y||^2 \\
&\quad -\alpha \sum_{i=1}^{m} \boldsymbol{a}_{ci}^{\mathrm{T}} B_{ck} \boldsymbol{a}_{ci} \Big\} \\
&\quad + \sum_{c=1}^{C}\sum_{i=1}^{m} \boldsymbol{\lambda}_{ci}^{\mathrm{T}} R_c \boldsymbol{a}_{ci} + \beta \sum_{k=1}^{n} u_{ck} \log u_{ck} \\
&\quad + \sum_{c=1}^{C}\sum_{i=1}^{m} \lambda_{ci0}(\boldsymbol{a}_{ci}^{\mathrm{T}} \boldsymbol{a}_{ci} - 1) \\
&\quad + \sum_{k=1}^{n} \gamma_k \Big(\sum_{c=1}^{C} u_{ck} - 1 \Big)
\end{aligned} \tag{15}
$$

ここで,

$$
B_{ck} = (\boldsymbol{y}_k - \boldsymbol{v}_c^y)(\boldsymbol{y}_k - \boldsymbol{v}_c^y)^{\mathrm{T}} \tag{16}
$$

$$
R_c = \{r_{crs}\} \tag{17}
$$

$$
\begin{aligned}
r_{crs} &= \sum_{k=1}^{n} u_{ck}(x_{kr} - v_{cr}^x)(y_{ks} - v_{cs}^y) \\
&\quad r = 1, \cdots, p,\ s = 1, \cdots, q
\end{aligned} \tag{18}
$$

である．$\boldsymbol{v}_c^x$, $\boldsymbol{v}_c^y$ ($\boldsymbol{v}_c = (\boldsymbol{v}_c^x,\ \boldsymbol{v}_c^y)$) は，それぞれ，補助変数 $\boldsymbol{x}_k$，目的変数 $\boldsymbol{y}_k$ のクラスター中心ベクトルを表す．

式 (15) の第 1 項目は，クラスタリングと主成分分析を表す項で，第 2 項目は，求まる主成分と補助変数との相関が 0 であるという条件を表している．第 3 項目はエントロピー項で，第 4，第 5 項目は，それぞれ，主成分ベクトルの長さとメンバシップ値の和が 1 である制約を表している．

ラグランジュ関数の最適性の必要条件 $\partial L/\partial \boldsymbol{a}_{ci} = \boldsymbol{0}$，$\partial L/\partial \boldsymbol{\lambda}_{ci} = \boldsymbol{0}$，$\partial L/\partial \lambda_{ci0} = 0$ から，

$$
2\alpha T_c \boldsymbol{a}_{ci} - 2\lambda_{ci0}\boldsymbol{a}_{ci} - R_c^{\mathrm{T}} \boldsymbol{\lambda}_{ci} = \boldsymbol{0} \tag{19}
$$

$$
R_c \boldsymbol{a}_{ci} = \boldsymbol{0} \tag{20}
$$

$$
\boldsymbol{a}_{ci}^{\mathrm{T}} \boldsymbol{a}_{ci} - 1 = 0 \tag{21}
$$

となる．さらに，式 (19)，(20)，(21) より，

$$
\mu_{ci0} = \boldsymbol{a}_{ci}^{\mathrm{T}} T_c \boldsymbol{a}_{ci} \tag{22}
$$

$$
\mu_{ci0} = \frac{\lambda_{ci0}}{\alpha} \tag{23}
$$

を得る．ただし，

$$
T_c = \sum_{k=1}^{n} u_{ck} B_{ck} \tag{24}
$$

である．したがって，式 (15) を最大化する主成分 $\boldsymbol{a}_{ci}$ を求める問題は，

$$
\alpha \left[I - R_c^{\mathrm{T}} (R_c R_c^{\mathrm{T}})^{-1} R_c \right] T_c \boldsymbol{a}_{ci} = \lambda_{ci0} \boldsymbol{a}_{ci} \tag{25}
$$

なる固有値問題に帰着し，解は右最大固有値（大きい方から順に m 個）に対応する固有ベクトル $\boldsymbol{a}_{ci}$ によって与えられる．

v_{cr}^x，v_{cs}^y，u_{ck} についても，ラグランジュ関数の最適性の必要条件 $\partial L/\partial v_{cr}^x = 0$，$\partial L/\partial v_{cs}^y = 0$，$\partial L/\partial u_{ck} = 0$ より，それぞれ，

$$v_{cr}^x = \frac{\sum_{k=1}^{n} u_{ck} x_{kr}}{\sum_{k=1}^{n} u_{ck}} \tag{26}$$

$$v_{cs}^y = \frac{\sum_{k=1}^{n} u_{ck} y_{ks}}{\sum_{k=1}^{n} u_{ck}} \tag{27}$$

$$u_{ck} = \frac{\exp A_{ck}}{\sum_{a=1}^{C} \exp A_{ak}} \tag{28}$$

$$\begin{aligned} A_{ak} &= \Big(\sum_{r=1}^{p} \big(x_{kr} - v_{ar}^x\big)^2 + \sum_{s=1}^{q} \big(y_{ks} - v_{as}^y\big)^2 \\ &- \alpha \sum_{s=1}^{q} \sum_{t=1}^{q} \sum_{i=1}^{m} \big(y_{ks} - v_{as}^y\big)\big(y_{kt} - v_{at}^y\big) \\ &\times a_{ais} a_{ait} \\ &+ \sum_{r=1}^{p} \sum_{s=1}^{q} \sum_{i=1}^{m} \big(x_{kr} - v_{ar}^x\big)\big(y_{ks} - v_{as}^y\big) \\ &\times a_{air} \lambda_{ais} \Big) / \beta \end{aligned} \tag{29}$$

となる．以下に補助変数に独立な主成分を抽出するファジィクラスタリング法のアルゴリズムを記述する．

Step 1 定数 α, β, C, m を定める． メンバシップ u_{ck} の初期値を乱数により決定する．

Step 2 式 (26)，(27) より，クラスター中心ベクトル $\boldsymbol{v}_c^x$, $\boldsymbol{v}_c^y$ を求める．

Step 3 式 (25) より主成分ベクトル $\boldsymbol{a}_{ci}$ を求める．

Step 4 式 (28) よりメンバシップ値 u_{ck} を求める．

Step 5 終了判定条件 $\max_{c,k}\{|u_{ck}^{NEW} - u_{ck}^{OLD}|\} < \varepsilon$ を満たせば終了．そうでなければ *Step 2* へ．

3 数値実験

提案法の有効性を調べるために，図 1 のような 2 次元の人工データを用いて実験を行った．図 1 の左側のグループと右側のグループにはそれぞれ 100 個のデータが含まれており，左側のグループには y_1 と y_2 の間に負の相関が，右側には正の相関がある．このデータに対して，補助変数 x_1 を考え，全てのデータに対して以下のような操作を加えた．x_1 は一様乱数とし，左側のグループに対しては，

$$y_1^* = y_1 + x_1 \tag{30}$$

$$y_2^* = y_2 + x_1 \tag{31}$$

右側のグループに対しては，

$$y_1^* = y_1 + x_1 \tag{32}$$

$$y_2^* = y_2 - x_1 \tag{33}$$

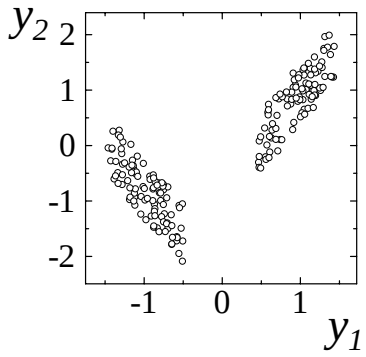

図 1: 変換前のデータ

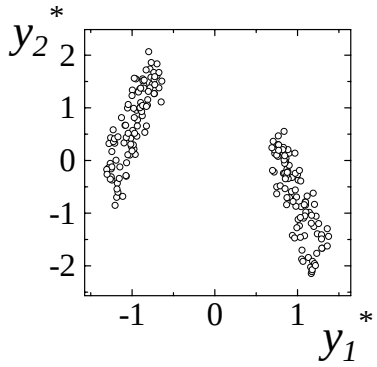

図 2: 変換後のデータ

とした．この操作は，補助変数を与えることによって，グループごとの y_1 と y_2 の相関関係を反転させることを意味している．変換後のデータの散布図を図 2 に示す．数値実験では，提案法を用いて，x_1 の影響を取り除いた主成分分析を行うことにより，y_1^*，y_2^* から y_1，y_2 間の相関関係をグループごとに取り出せるかどうかを調べた．FCV 法と提案法のパラメータは以下のように設定した．

クラスター数 C	: 2
求める主成分の数 m	: 1
エントロピー項の係数 β	: 0.08
主成分分析に対する重み α	: 0.4

また，目的変数を y_1^*，y_2^*，補助変数を x_1 として提案法を適用した．FCV 法を適用した結果得られた主成分と各変量のファジィ因子負荷量 [8] を表 1 と表 2 に示す．ファジィ因子負荷量は，主成分と各変量の関係を定量的に示す尺度であり，各変量と主成分の相関関係を表す．したがって，同一符号の因子負荷量を持つ変量同士は正の相関があり，異なる符号の因子負荷量を持つ変量同士は負の相関がある．

両手法とも，左右のグループでそれぞれクラスターを形成した．表 1 より，FCV 法では補助変数 x_1 の影響により逆転した相関関係を取り出していることがわかる．一方，提案法では，得られた主成分と変換後のデータ y_1^*，y_2^* との相関は小さくなっているが，変換前のデータの y_1，y_2 の相関関係がクラスターごとに正しく取り出せていることがわかる．つまり，提案法により，支配的要因 x_1 によって隠蔽されていた潜在的，局所的な相関関係を取り出すことができたといえる．なお，実験結果は局所最適解である可能性があるが，いくつか初期値を変えて実験を行ったところ，ほとんど差異が見られなかったので最適解が得られていると考えられる．

表 1: FCV 法によるファジィ因子負荷量

	ファジィ因子負荷量	
変数	左クラスター	右クラスター
y_1^*	0.946	0.939
y_2^*	0.961	-0.962
x_1	0.994	0.994

表 2: 提案法によるファジィ因子負荷量

	ファジィ因子負荷量	
変数	左クラスター	右クラスター
y_1^*	0.291	-0.315
y_2^*	-0.310	-0.318
x_1	0.000	0.000

4 POS データの解析

本章では，提案手法を用いて POS データの解析を行う．POS とは「point of sales」の略で店頭での販売情報を販売時点で把握するシステムを指す．そのシステムによって収集・蓄積されたデータが POS データである．販売金額や品目・品数などの情報からなる POS データを解析することにより，商品売れ行き予測などが行われている．

多くの店舗では，1 年間の営業日のうち定まった日や期間に来店客数が大幅に増減することがある．来店客数は年末年始，その店舗または競合店の特売日，競合店の定休日などの暦や他店舗を含めた販売状況に大きく左右される．しかし，日々の来店客数の変化は，これらの要因だけでは十分に説明することができず，来店客数を最終的に決定する要因に気象条件を無視できないことが指摘されている [9]．そこで，POS データに気象庁から配信されている天候データ（アメダスデータ）を加えて FCV 法と提案法をそれぞれ適用し，比較することによって，提案法の有効性を考察する．

4.1 POS データ

用いたデータは，ある 2 店舗のスーパーマーケットの POS データである．用いた変量を表 3 に示す．ただし，変量のうち，曜日，祝日の項目はダミー変数とし，各々あてはまるか否かで 1 または 0 の値を与えた．天気概況の項目は降水の状況により 0（晴れまたは曇り）から 3（大雨）の値を与えた．日配品とは，牛乳，たまご，うどん，豆腐など毎日配送され翌日か翌々日までに消費される食品で，工場で生産されるものを指す．来客数は日配品を購入して POS 端末を通った客の数である．日配品点数は POS 端末を通った客が購入した日配品の総数であり，例えばある客が日配品を 5 品購入すると 5 ポイント加算される．欠損している月日，および気象条件以外の社会的要因による影響を大きく受けていると思われるデータ（例えば 12 月 31 日）を除いた 333 日分について，各変量ごとに平均 0，分散 1 に基準化したものを用いた．

表 3: POS データの解析に用いた変量

変量　20 次元
祝日 金曜 土曜 日曜
平均気温 9 時気温 12 時気温 15 時気温 18 時気温
相対湿度 昼天気概況 夜天気概況 日降水量
9-12 時降水量 12-15 時降水量 15-18 時降水量
来客数（A 店）　日配品点数（A 店）
来客数（B 店）　日配品点数（B 店）

4.2 Fuzzy c-Varaieties 法の適用

まず，提案法との比較のために，FCV 法によって，POS データの解析を行った．パラメータを以下のように設定した．

クラスター数 C ：2
求める主成分の数 m ：3
エントロピー項の係数 β ：2.0
主成分分析に対する重み α ：0.5

表 4 に各クラスターのクラスター中心を示す．表 4 より，クラスター 1 は気温を表す変量の値が低

表 4: FCV 法により求まったクラスター中心

Item	Cluster 1	Cluster 2
祝日	-0.197	0.198
金曜	-0.022	-0.0221
土曜	0.000	-0.000
日曜	0.021	0.021
平均気温	-0.786	0.790
9 時気温	-0.775	0.779
12 時気温	-0.765	0.769
15 時気温	-0.766	0.770
18 時気温	-0.768	0.771
相対湿度	-0.344	0.345
昼天気概況	-0.177	0.178
夜天気概況	-0.170	0.171
日降水量	-0.105	0.105
9-12 降水量	-0.140	0.140
12-15 降水量	0.058	-0.059
15-18 降水量	0.033	-0.034
A 店来客数	-0.140	0.140
A 店日配品点数	-0.162	0.163
B 店来客数	-0.197	0.198
B 店日配品点数	-0.089	0.089

くクラスター 2 は気温を表す変量の値が高いことがわかる．つまり，それぞれのクラスターは主に気温を基準にしてクラスタリングされている．クラスター 1 を寒冷期のクラスター，クラスター 2 を温暖期のクラスターと呼ぶことにする．表 5，表 6 に各クラスターにおける主成分のファジィ因子負荷量を示す．ファジィ因子負荷量の絶対値が 0.4 以上のものに下線を引いている．寒冷期のクラスターでは，第 1 主成分に対応する因子負荷量から，2 店の来客数と日配品点数は，第 1 主成分と負の相関が大きいので，日中の天気が良く，降水量が少なければ両店の客数と売上が増えるといえる．同様に，第 2 主成分から土曜日であり，気温が高いほど両店の売上，客数が増えるといえる．第 3 主成分については，来客数，売上の因子負荷量が両店とも非常に小さいので来客数，売上に関する知見は得られない．温暖期のクラスターでは，寒冷期と同様に第 1 主成分から好天で雨が降らないほど，第 2 主成分から金曜でなく日曜であり，気温が低いほど，来客数および売上が増えることがわかる．第 3 主成分については，寒冷期と同じく来客数，売上の因子負荷量が両店とも小さいので特に来客数，売上に関する知見は得られない．FCV 法を適用した結果得られた知見は自明なものばかりであり，知識として重要とはいえない．

表 5: 寒冷期のファジィ因子負荷量

Item	1st PC	2nd PC	3rd PC
祝日	0.000	-0.003	-0.030
金曜	0.223	-0.383	0.142
土曜	-0.043	0.456	-0.131
日曜	-0.273	0.289	-0.239
平均気温	0.252	0.471	0.818
9 時気温	0.291	0.481	0.754
12 時気温	0.205	0.468	0.832
15 時気温	0.112	0.413	0.874
18 時気温	0.164	0.412	0.852
相対湿度	0.568	0.262	-0.312
昼天気概況	0.687	0.246	-0.358
夜天気概況	0.391	0.034	-0.178
日降水量	0.830	0.449	-0.181
9-12 降水量	0.426	0.040	-0.234
12-15 時降水量	0.698	0.435	-0.099
15-18 時降水量	0.641	0.387	-0.056
A 店来客数	-0.601	0.552	-0.002
A 店日配品点数	-0.577	0.687	-0.142
B 店来客数	-0.628	0.709	-0.028
B 店日配品点数	-0.554	0.712	-0.267

4.3 提案法の適用

平均気温を補助変数に設定する事によって気温の影響を取り除いた分析を行なった．パラメータは以下の通りである．

クラスター数 C : 2
求める主成分の数 m : 1
エントロピー項の係数 β : 3.0
主成分分析に対する重み α : 0.5

クラスターは FCV 法と同じく寒冷期と温暖期で形成された．それぞれのクラスターのファジィ因子負荷量を表 7 に示す．この実験では，提案法が潜在的な主成分を求めるため，寄与率が非常に低くなることを考慮し，主成分の数を 1 として考察を行なった．表 7 の寒冷期を見てみると，両店の来客数と売上の相関係数が大きいことから，金曜でなく日曜であれば，両店の来客数と売上が増えることがわかる．このことは，FCV 法では得られなかった知見である．温暖期では，相対湿度が低く，昼の天気が良く，降水量が少ないと両店の来客数と売上が増えることが分かる．天気に関する変量の内，夜の天気に関する相関係数のみ小さくなっていることに注目すると，温暖期においては，夜の天気はあまり来客数と売上に影響を及ぼさないことが分かる．

次に平均気温の代わりに，日降水量を補助変数にした．パラメータは前述の実験と同じものを用いた．クラスターは FCV 法と同じく寒冷期と温暖期で形成された．それぞれのクラスターのファジィ因子負荷量を表 8 に示す．まず，寒冷期について見ると，平均気温を補助変数にした時とほぼ同じ結果が得られた．次に温暖期について見てみると，金曜でなく，土日であれば両店の来客数と売上が増えることがわかる．土曜であればという条件が新たに発見できた．A 店と B 店の因子負荷量から，A 店の方が B 店よりも来客数も売上も下回っていることから，A 店よりも B 店のほうが曜日の影響を強く受けていることが分かる．

次に日降水量の代わりに曜日に関する変量，つまり金曜日，土曜日，日曜日を補助変数にした．パラメータは前述の実験と同じである．クラスターは同様に寒冷期と温暖期で形成された．それぞれの

表 6: 温暖期のファジィ因子負荷量

Item	1st PC	2nd PC	3rd PC
祝日	-0.080	0.117	-0.863
金曜	-0.227	-0.449	-0.101
土曜	0.087	0.338	0.142
日曜	0.332	0.428	0.097
平均気温	0.257	-0.404	0.653
9 時気温	0.255	-0.378	0.650
12 時気温	0.397	-0.443	0.598
15 時気温	0.381	-0.472	0.577
18 時気温	0.326	-0.444	0.608
相対湿度	-0.647	0.335	0.139
昼天気概況	-0.750	0.453	0.054
夜天気概況	-0.422	0.421	0.055
日降水量	-0.669	0.393	0.244
9-12 降水量	-0.693	0.260	0.289
12-15 時降水量	-0.522	0.182	0.223
15-18 時降水量	-0.306	0.258	0.078
A 店来客数	0.608	0.490	0.123
A 店日配品点数	0.630	0.608	0.136
B 店来客数	0.645	0.658	0.054
B 店日配品点数	0.562	0.740	-0.001

クラスターのファジィ因子負荷量を表 9 に示す．表 9 の寒冷期について見ると，日降水量，12-15 時降水量，15-18 時降水量が低いと A 店の来客数が増える事がわかる．さらに，上記の条件は B 店にはあまり影響を与えていないことがわかる．また，表 9 の温暖期を見ると，相対湿度が低く，昼と夜の天気が良く，日降水量，9-12 時降水量，12-15 時降水量，15-18 時降水量が低いと，A 店の来客数と売上が伸びる事がわかる．さらに，上記の条件は全て，天気に関するパラメータであるが，B 店の因子負荷量は来客数，売上共に小さくなっていることから，天気は B 店の来客数と売上にそれほど相関がないということがわかる．FCV 法と異なり，A 店と B 店の違いを明確にすることができた．

5 おわりに

本研究では，補助変数に独立な主成分を抽出するファジィクラスタリング法を提案した．補助変数をデータ集合に大きな影響を与える支配的要因とすると，提案手法により，それらの影響を取り除いた潜在的，かつ，局所的な主成分を抽出することが出来る．数値実験では，本来の相関関係が隠蔽されたデータに提案法を適用し，隠された相関関係が取り出せることを示した．また，提案法と FCV 法を POS データに適用し，比較を行った．提案法により，従来の FCV 法では得られない新たな知識を発見できることがわかった．しかし，どの変数を補助変数に設定すべきかの具体的な選択基準がなく，補助変数の選択はユーザーに委ねられる．また，数値実験において採用されているパラメータは試行錯誤的に決定されたものである．主成分分析への重み α は形成されるクラスターに影響を与える．α を大きくするとクラスターは線形に広がり，小さくすると円形のクラスターが形成されやすくなる．エントロピー項の係数 β もやはりクラスターの形成に影響を与える．β が大きいとファジィなクラスターが，小さいとクリスプなクラスターが得られる．クラスタリングの結果によって抽出される主成分も異なってくる．本研究の数値実験では，妥当なクラスタリング結果が得られている値を採用している．しかし，クラスタリングの妥当性の判断にはかなりの主観を伴う．パラメータ設定などの種々の問題は今後の課題といえる．

表 7: 平均気温を補助変数としたときのファジィ因子負荷量

Item	寒冷期	温暖期
祝日	0.016	0.044
金曜	0.520	-0.160
土曜	-0.364	0.035
日曜	-0.422	0.352
平均気温	0.000	0.000
9 時気温	0.158	-0.203
12 時気温	0.115	-0.063
15 時気温	0.095	-0.019
18 時気温	0.123	-0.091
相対湿度	0.286	-0.634
昼天気概況	0.332	-0.720
夜天気概況	0.353	-0.306
日降水量	0.359	-0.757
9-12 降水量	0.311	-0.547
12-15 時降水量	0.312	-0.616
15-18 時降水量	0.333	-0.446
A 店来客数	-0.783	0.619
A 店日配品点数	-0.878	0.630
B 店来客数	-0.917	0.676
B 店日配品点数	-0.898	0.600

参考文献

[1] 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)

[2] R. N. Dave : "An Adaptive Fuzzy c-Elliptotype Clustering Algorithm", *Proc. NAFIPS 90*, Vol.1, pp.9-12 (1990)

[3] 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イモデル", 日本ファジイ学会誌, Vol.10, No.1, pp.142-149 (1998)

[4] 奥野 忠一，久米 均，芳賀 敏郎，吉澤 正 : 多変数解析法, 日科技連出版社 (1971)

[5] S. Miyamoto and M. Mukaidono : "Fuzzy c-means as a regularization and maximum entropy approach", *Proc. of the 7th International Fuzzy Systems Association World Congress*, Vol.II, pp.86-92 (1997)

[6] 宮本 定明，馬屋原 一孝，向殿 政男 : "ファジィc-平均法とエントロピー正則化法におけるファジィ分類関数", 日本ファジィ学会誌, Vol.10, No.3, pp.548-557 (1998)

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

[8] Y. Yabuuchi and J. Watada : "Fuzzy Principal Component Analysis and Its Application", *Biomedical Fuzzy and Human Sciences*, Vol.3, No.1, pp.83-92 (1997)

[9] 朝倉 正，赤津 邦夫，奥山 和彦 : 現代の気象テクノロジー 6, 経済活動と気象, 朝倉書店 (1992)

表 8: 日降水量を補助変数としたときのファジィ因子負荷量

Item	寒冷期	温暖期
祝日	0.020	-0.021
金曜	0.454	0.509
土曜	-0.356	-0.411
日曜	-0.450	-0.468
平均気温	-0.248	0.211
9 時気温	-0.235	0.177
12 時気温	-0.291	0.218
15 時気温	-0.295	0.273
18 時気温	-0.252	0.258
相対湿度	0.169	-0.199
昼天気概況	0.290	-0.266
夜天気概況	0.239	-0.328
日降水量	0.000	0.000
9-12 降水量	0.217	-0.045
12-15 時降水量	0.224	-0.091
15-18 時降水量	0.284	-0.223
A 店来客数	-0.812	-0.655
A 店日配品点数	-0.904	-0.779
B 店来客数	-0.952	-0.809
B 店日配品点数	-0.915	-0.871

表 9: 曜日を補助変数としたときのファジィ因子負荷量

Item	寒冷期	温暖期
祝日	0.000	0.000
金曜	0.000	0.000
土曜	0.000	0.000
日曜	0.000	0.000
平均気温	-0.198	-0.094
9 時気温	-0.280	-0.106
12 時気温	-0.196	0.000
15 時気温	-0.073	0.034
18 時気温	-0.119	-0.005
相対湿度	-0.169	-0.504
昼天気概況	-0.366	-0.665
夜天気概況	0.029	-0.453
日降水量	-0.622	-0.686
9-12 降水量	0.227	-0.524
12-15 時降水量	-0.882	-0.605
15-18 時降水量	-0.486	-0.673
A 店来客数	0.425	0.576
A 店日配品点数	0.333	0.493
B 店来客数	0.240	0.298
B 店日配品点数	0.191	0.189