# 非線形クラスタリング手法の潜在構造モデルへの応用

加茂下 茜　(指導教員：吉田 裕亮)

## 1　はじめに

### 1.1　潜在構造モデルとは

世の中にある色々なものに対して, 人々は様々な感情をもつ．もし, このような好みを把握し, 傾向によって分類することができたなら，マーケティング戦略上重要な情報を得ることが可能となる．しかしこれらは存在が仮定できる概念であるのにもかかわらず, 直接的に測定することができないという特徴をもつ．

消費者の嗜好, 購買意欲などのように，直接的には観測できないが, 存在が仮定できる概念や背景因子のことを，潜在変数とよぶ．これを実際に観測することができる変数である顕在変数と関連付けてモデルに組み込むことで，複数個の変数間の複雑な関係を効率的に説明するモデル構築が可能となる．この種の統計モデルは，一般に潜在構造モデルとよばれる．

### 1.2　EM アルゴリズム

通常，潜在クラス分析法では，EM アルゴリズムを用いた推定法が一般的である．EM アルゴリズムは確率モデルのパラメータを最尤法に基づいて推定する反復法のひとつであり，期待値段階 (Expectation-step) と最大化段階 (Maximization-step) を交互にくりかえし計算を行う．EM アルゴリズムは，確率の定義域で安定した解が解け，かつ多様な潜在クラスモデルの制約条件に柔軟に対応できるとう利点を持つため，この種のモデルの推定法によく使用される．しかし反復法であるため，

・データ量が増えるほど推定すべきパラメータが増える
・データとモデルの適合度検定が必要なため，処理すべき内容が多い
・結果が可視化向きでない

などという不便な面もある．

## 2　非線形クラスタリング

多量のデータに内在する本質的な情報や構造を明らかにする手法の１つとして，クラスタリングがあげられる．クラスタリングは，分類すべきデータ間に定義された類似性や距離に基づいてグループ分けを行い，潜在する情報や構造を取り出す手法である．

実世界においては線形分離可能なデータよりも，線形分離不可能なデータの方が多く存在するため，クラスタリングの非線形化が必要となる．パターン認識の分野において，ニューラルネットワークや SVM などを応用した手法が提案されているが，これらはいずれも教師付き分類手法である．ここでは教師なし分類手法の非線形分離可能なアルゴリズムへの拡張が必要となる．そこで本研究では，内積に基づく類似度を利用し，カーネル法を用いることで，非線形クラスタリングを行い，潜在構造モデルへ応用することで従来用いられてきた手法との比較を行うことを目的とする．

## 3　カーネル PCA(カーネル主成分分析)

カーネル法とは，カーネル関数を用いて，入力特徴ベクトルを高次元特徴空間へ射影し、空間を拡張するという手法である．一般に，高次元特徴空間に写像すると，データの表現力が上がるため，線形分離が可能となる．しかし，このような射影によって得られる特徴空間の次元は非常に大きく，直接計算することは難しい．そこで，高次元特徴空間での内積を特定の条件を満たす入力空間上の関数で表現する．すなわち入力特徴ベクトルの組 $(x, y)$ に対して

$$K(x, y) =< \phi(x), \phi(y) >$$

とする．このような関数をカーネル関数という．カーネル法を用いることにより，線形的なクラスタリングを非線形に行うことが可能となる．また，PCA や SVM など，内積計算がメインのアルゴリズムとの併用は，特徴空間の明示的な計算を経由せず，計算量を減らすことができるため，相性がよい．

PCA とは多次元のデータから特徴的な指標を得るための方法で，共分散行列の固有値問題の解として得られることができる．これらの 2 つを組み合わせた方法を，一般的にカーネル PCA と呼ぶ．

本手法において，入力特徴ベクトル $x$，$y$ と定数 $\beta$ によって表される以下のようなカーネル

$$K(x, y) = \exp\{-\beta(1 - \frac{(x \cdot y)}{||x||||y||})\}$$

を用い，カーネル PCA を行うこととする．

## 4　手法の提案

本研究では，従来用いられてきた推定法である EM アルゴリズムと同様に，尤度に関連した推定法のひとつを提案する．潜在構造モデルにおいて，EM アルゴリズムは潜在確率を求めるための手法であるが，本手法では観測された応答パターンをバイナリデータに変換し，それらを座標として尤度に付随したある量に基づいて，空間に割り当て直し，カーネル PCA を援用する．その結果として 2 次元に布置することによりクラスタリングを行う．

### 4.1　対数尤度に付随した量

尤度とは，ある前提条件に従って結果が出現する場合に，逆に観察結果からみて前提条件がどの程度尤もらしいかを表す数値である．この値が大きくなれば，生起しやすいことになる．最尤法とは，このような尤度を最大とするような確率を，観測データから推定する方法である．

尤度の対数は，一般に対数尤度と呼ばれる．尤度を最大とすることは対数尤度を最大とすることと等価であるため，一般的に計算の行いやすい対数尤度を用いることが多い．本研究では，まずこの対数尤度に付随して，観測データであるバイナリ値を座標点に割り当て直す．

| No. | 1 | 2 | $\cdots$ | $j$ | $\cdots$ | $k$ | 度数 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | $\cdots$ | 1 | $\cdots$ | 1 | $N_1$ |
| 2 | 1 | 1 | $\cdots$ | 1 | $\cdots$ | 0 | $N_2$ |
| $\vdots$ | | | | | | | |
| $i$ | 1 | 1 | $\cdots$ | 0 | $\cdots$ | 1 | $N_i$ |
| $\vdots$ | | | | | | | |
| $n-1$ | 1 | 0 | $\cdots$ | 0 | $\cdots$ | 0 | $N_{n-1}$ |
| $n$ | 0 | 0 | $\cdots$ | 0 | $\cdots$ | 0 | $N_n$ |

図 1: 2 値型（バイナリデータ）で集計された，$k$ 個の評価項目に関する応答パターン表

図 1 のような，2 値型（バイナリデータ）で集計されたいくつかの評価項目に関する応答パターン表が与えられ，$n = 2^k$ 個のパターン $(a_{i1}, a_{i2}, \cdots, a_{ik})$, $a_{ij} \in \{0, 1\}$ のそれぞれについて，観測度数が $N_i$ であるとき，

$$p_j = P(a_{ij} = 1) = \frac{1}{N} \sum_{a_{ij}=1} N_i$$

となるような確率 $p_j$ が求められる．そこで $a_{ij} = 1$ のとき，$\alpha_{ij} = \dfrac{1}{\sqrt{|\log(p_j)|}}$，$a_{ij} = 0$ のとき $\alpha_{ij} = \dfrac{1}{\sqrt{|\log(1-p_j)|}}$ となるように第 $j$ 成分を取り直す．本提案手法では，0 または 1 の生起確率が大きいほど点 $x$ が原点から遠くなるように，このような量

$$\alpha_{ij} = \frac{1}{\sqrt{|\, a_{ij}\log(p_j) + (1-a_{ij})\log(1-p_j)\,|}}$$

を用いて割り当て直すことにする．

このように，対数尤度に付随して割り当て直した $k$ 次元空間の点に対して，さらに，非線形問題に対応可能なカーネル法を援用してクラスタの推定を行う．

## 5 数値実験

### 5.1 シミュレーション実験

ある居酒屋で適当な 1000 人に，今日飲んだお酒は何かについてのアンケートを行ったとする．この時，事前情報として，男性のビール，サワー，カクテルを飲む確率をそれぞれ 0.8，0.1，0.1 とし，女性の確率を 0.2，0.2，0.6 と仮定する．このような明らかに好みの傾向が違う男性を 650 人，女性を 350 人用意し，ランダムに組み合わせる．それにより，男性女性などの属性がわからない架空の観測データを作成する．この 2 値型の観測データに本提案手法を用い，推定を行う．これを，EM アルゴリズムによって推定された結果と比較，検討する．

### 5.2 結果

| 評価項目 | クラス 1 | | クラス 2 | |
|---|---|---|---|---|
| | 構成率 0.407561 | | 構成率 0.592439 | |
| ビール | ○ 0.31 | × 0.69 | ○ 0.78 | × 0.22 |
| サワー | ○ 0.2 | × 0.8 | ○ 0.09 | × 0.91 |
| カクテル | ○ 0.61 | × 0.39 | ○ 0.05 | × 0.95 |
| | 女子 (と予想される) | | 男子 (と予想される) | |

図 2: EM アルゴリズムで推定された 2 クラスの結果

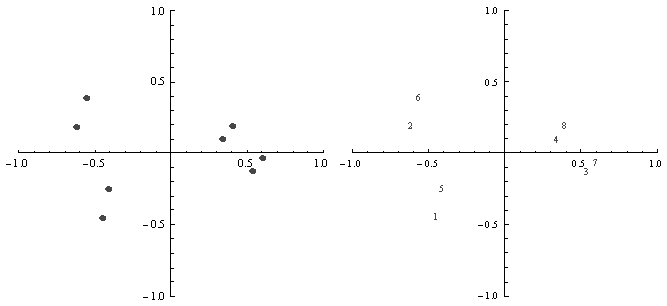

図 3: 本提案手法でカーネル PCA を行った結果 (左)
図 4: 図 3 を応答パターン番号 $j$ で布置しなおした結果 (右)

## 6 実データによる推定例

スーパーの店舗イメージに対する 5 つの評価項目に関しての応答パターンを，実データとして扱い，本手法を用いて推定を行った結果は図 6，7 で表される．図 6 は，データを 2 次元空間にプロットした概形であり．図 7 は，それらのうちそれぞれのデータ $x_i$ がどのグループに所属するのかを表したものである．EM アルゴリズムによって求められた潜在確率を，適合度検定などの組み合わせによって求めた最適なクラスタ数 (図 5) と比較すると，ある程度に見やすく布置されていることがわかる．

| クラス数 | 修正 $\overline{R^2}$ | AIC | 尤度比統計量 | ピアソンの $\chi^2$ |
|---|---|---|---|---|
| 1 | 0 | 7563.424 | 303.961 | 334.633 |
| 2 | 0.584 | 7343.534 | 97.334 | 96.692 |
| 3 | 0.854 | 7282.077 | 23.877 | 23.964 |
| 4 | 0.928 | 7276.937 | 6.737 | 6.589 |
| 5 | 0.836 | 7286.027 | 3.827 | 3.714 |

図 5: EM アルゴリズムの結果から適合度検定を行った結果

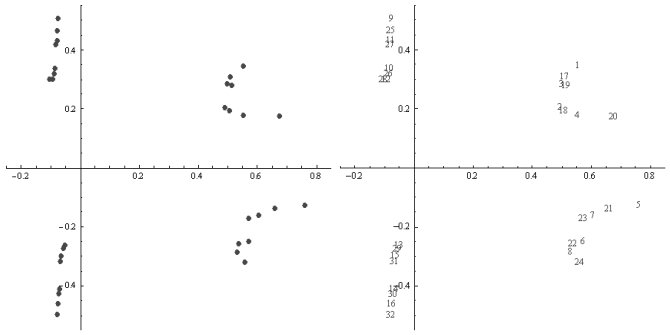

図 6: 本提案手法でカーネル PCA を行った結果 (左)
図 7: 図 6 を応答パターン番号 $j$ で布置しなおした結果 (右)

## 7 考察と今後の課題

本研究で提案した推定法により，従来より少ない手間でクラスタリングを行い，最適なクラスタ数を求めることや，データを可視化することなどが可能となった．しかし，EM アルゴリズムで求められた結果と厳密に比較すると，最適なクラス数は等しくなっているものの，データの属する割合やクラスにやや違いがみられた．

今後の課題としては，さらに精度の高いクラスタリングが行えるように，尤度に付随した，より最適な距離関数を見つけていくことが挙げられる．