x-means法とスパース因子分析を用いた美容品のマーケティング分析

鈴木　秀男

キーワード：データ解析，クラスター分析，L_1 ノルムによる正則化法

本稿は，鴫原 成美さんによる 2015 年度慶應義塾大学理工学部に提出した卒業論文をもとに加筆修正したものです．

1. はじめに

近年は，ID 付き POS データやスキャンパネルデータなど，大量で多様な顧客のデータが獲得できる環境にあり，それらを活用して顧客のニーズに対応した製品やサービスを提供するためのマーケティング分析を行うことの重要性が増しています．そのため，大量データの解析に対応した手法の活用が求められています．本研究では，ユーザーの購買履歴や性別や年齢などの属性に関するデータから，美容品の購入経験があるユーザーを x-means 法というクラスタリング手法 [1] によりいくつかのクラスターに分類し，より深く分析を行います．得られた各クラスターの構造をスパース因子分析により考察します．スパース因子分析とは，正則化法を因子分析モデルに応用した手法であり [2]，大規模データに対応した手法として期待されています．

2. 分析データ

本研究において，分析対象のデータは，株式会社ブレインパッドから提供されたもので，同社が展開するスマートフォンアプリ「ReceReco」[1]の購買履歴，ユーザーの属性のデータです．同アプリは，会員登録をしたユーザーがスマートフォンで撮影したレシートを読み込むと自動で家計簿をつけることのできるサービスです．

3. 顧客クラスタリングとターゲティング

年齢が 1〜99 歳と登録されているユーザーのうち，レシートを 20 枚以上登録している優良ユーザーに絞ります．そのうえで，今回美容品市場を対象としたので，美容品を購入したことのあるユーザーを分析対象としました．最初に，x-means 法というクラスタリング手法を用いて，男女別に顧客のクラスタリングを行います．x-means 法は，クラスター数を分析者が事前に決めてから分析を行う k-means 法の拡張で，情報量規準により分割が妥当と判断されるまで分割を繰り返す手法です [1]．すなわち，与えられたデータに対して、自動的に妥当なクラスター数を決定し、分類結果を与えてくれます．分類する際の変数は，全品目についての購入金額に対する美容品（美容健康費，美容品，コスメ，ネイル，エステ，ジム代，通院費，薬代，その他美容健康費）の購入金額の割合としました．その結果，クラスター数は女性 900 個，男性 117 個となりました．次に，LOF(Local Outlier Factor) による外れ値検出を行いました．LOF は，密度ベースの外れ値検出法です．ほかの点と比べて，ある点のまわりの密度が小さいほど，LOF のスコアは大きくなります．すなわち，スコアが大きい点は，外れ度合いが大きい点と言えます．本分析では，スコアの上位 20%を外れ値とするような閾値を設定しました．全サンプルに占めるクラスター内の観測点の割合が大きく，クラスター内の外れ値の割合が小さいクラスターを代表的なクラスターとして分析対象としました．その結果，男女各 4 個ずつ抽出しました．美容品の品目の割合と年齢について基礎集計をしたところ，たとえば，女性の各クラスターの特性は表 1 のようになりました．

すずき　ひでお
慶應義塾大学 理工学部管理工学科
〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
hsuzuki@ae.keio.ac.jp

[1] ReceReco では，ユーザーの属性データやレシートデータを，個人が特定できないようデータを加工したうえで，分析・販売できるよう利用規約が定義されています．

表 1　クラスター基礎集計結果（女性，$n = 6826$）

クラスター	年齢	美意識	各品目の割合
1	30 代前半から 40 代前半	強い	通院費＞薬代＞コスメ＞美容院 全体的に割合が高い
2	20 代後半から	中くらい	通院費＞薬代＞コスメ
3	40 代前半	弱い	美容健康品にお金をかけない
4	20 代前半から 30 代後半	強い	コスメが高い

4. スパース因子分析による各クラスター構造の把握

各クラスターに属する顧客のデータについてスパース因子分析を行いました．スパース因子分析とは，L_1 ノルムによる正則化法を因子分析モデルに応用した手法です [2]．L_1 ノルムについて，たとえば，p 次元ベクトル $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ とすると，L_1 ノルムは，$\|\boldsymbol{x}\|_1 = |x_1| + |x_2| + \cdots + |x_p|$ となります．L_1 ノルムによる正則化法では，変数の数が膨大であっても，変数選択の際にいくつかのパラメータが正確に 0（その変数の影響が全くない）と推定することができるため，効率的に情報を取捨選択できます． 従来の因子分析では対数尤度関数 $l(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ を最大にするように因子負荷行列 $\boldsymbol{\Lambda} = (\lambda_{ij})$，独自分散行列 $\boldsymbol{\Psi}$ を求めるのに対して，スパース因子分析では式 (1) を最大にするように求めます．

$$l_\rho(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = l(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) - n \sum_{i=1}^{p} \sum_{j=1}^{m} \rho P(|\lambda_{ij}|) \qquad (1)$$

サンプル数は n，$\rho > 0$ は正則化パラメータです．パラメータ ρ は情報量規準の BIC [3, 4] を用いて適切な値を選択しました．変数は全品目についての購入金額に対する各品目の購入金額の割合としました．

男女各クラスターについてスパース因子分析を行いました．ここでは，女性クラスター 1 において抽出された因子の解釈について説明します．スパース因子分析による因子負荷量のうち，絶対値が大きいものを抽出して表 2 に示しています．たとえば，因子 1 のプラス方向については，家賃，洋服，コスメの因子負荷量の値が高いことから，家賃を多く支払い，洋服やコスメなど外見にお金をかける方向であると解釈できます．一方，因子 1 のマイナス方向については，食料品や食費にお金をかける方向であると解釈できます．因子 2 および因子 3 のプラス方向については，食費や日用品などの生活必需品にお金をかける方向であると解釈できます．因子 4 のマイナス方向については，ライフライン関連に多く払う方向であると考えられます．

表 2　女性クラスター 1 ($n = 6826$) における因子負荷量

	正		負	
因子 1	家賃	0.28341	食料品	-0.578548
	携帯電話	0.1865694	食費	-0.273109
	洋服	0.1782429	日用品	-0.232048
	飲み会	0.153651	消耗品	-0.196582
	電車	0.1494643	子供関連	-0.103728
	コスメ	0.1298761	被服費	-0.08646
	アクセサリー	0.1241202	薬代	-0.071998
因子 2	食費	0.2282923	昼食	-0.522181
	食料品	0.1704254	軽食	-0.451991
	家賃	0.127362	夕食	-0.443854
	日用品	0.1261782	朝食	-0.435973
	住宅ローン	0.075829	書籍	-0.164414
	携帯電話	0.0751116	漫画	-0.128617
	交際費	0.0747357	タクシー	-0.107146
因子 3	食費	0.6632337	食料品	-0.777886
	日用品	0.3422009	消耗品	-0.207548
	被服費	0.1947033	昼食	-0.104504
	交際費	0.1351311	コスメ	-0.054863
	教養娯楽費	0.1218587	軽食	-0.041301
	交通費	0.1156328	美容院	-0.034531
	美容健康費	0.0991056	薬代	-0.03209
	養育費	0.073033		
因子 4	昼食	0.1163464	電気料金	-0.691844
	洋服	0.1128362	水道料金	-0.575808
	軽食	0.0782413	ガス料金	-0.533572
	アクセサリー	0.0629589	インターネット	-0.214225
	食料品	0.0626487	携帯電話	-0.2024
	プレゼント代	0.0580628	固定電話	-0.171986
	下着	0.0575628	家賃	-0.122609

5. おわりに

本稿では，美容品の顧客クラスターの特徴を分析し，さらにスパース因子分析のマーケティングデータにおける説明力と解釈性について考察しました．その結果，美容品においては男女ともにライフステージによって特徴が大きく異なっていることがわかりました．本稿では詳しくは触れませんでしたが，スパース因子分析は従来の因子分析手法と比べて，解釈性には優れていることを確認しました．本研究を通じて，x-means 法やスパース因子分析は，大量データに基づくマーケティング分析手法として有効であることが示唆されました．

参考文献

[1] 石岡恒憲，“クラスター数を自動決定する k-means アルゴリズムの拡張について，” 応用統計，**29**, pp. 141–149, 2000.

[2] K. Hirose, “Sparse estimation via nonconcave penalized likelihood in factor analysis model,” *Statistics and Computing*, **25**, pp. 863–875, 2005.

[3] G. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, **6**, pp. 461–464, 1978.

[4] 小西貞則，北川源四郎，『情報量基準（シリーズ予測と発見の科学 2)』，朝倉書店，2004.