

データマイニング

第5回 クラスター分析

2023年春学期

宮津和弘

本日の講義・演習

日付	講義・演習内容
04/14/23	(1) イントロダクション
04/21/23	(2) ビジネスシミュレーション
04/28/23	(3) ID-POSデータ分析
05/12/23	(4) 対応分析
05/19/23	(5) クラスター分析
05/26/23	(6) 自己組織化マップ
06/02/23	(7) 線形判別分析
06/09/23	(8) 非線形判別分析
06/16/23	(9) ツリーモデル
06/23/23	(10) 集団学習
06/30/23	(11) サポートベクターマシン
07/04/23	(12) ネットワーク分析
07/14/23	(13) 共分散構造分析
07/21/23	(14) テキスト分析
07/28/23	(15) まとめ



本日の演習概要とポイント

- 因子分析の結果を用いたクラスター分析
- 階層的および非階層的クラスタリング
- ウォード法は階層構造を理解することが可能

因子分析の復習

観測されない潜在因子

例) 主要 5 教科テスト点数の背後に理系および文系の能力因子を仮定すると …

	数学	理科	英語	国語	社会	理系因子		文系因子	
						数理:平均	英国社:平均	数理:平均	英国社:平均
A	89	91	67	46	53	90.0	53.3		
B	57	69	80	85	91	63.0	85.3		
C	80	93	35	41	51	86.5	42.3		
D	41	61	53	45	55	51.0	37.7		
E	78	87	47	51	63	82.5	53.7		
D	53	66	81	73	86	59.5	80.0		
F	90	86	89	91	98	88.0	92.7		

因子分析モデル

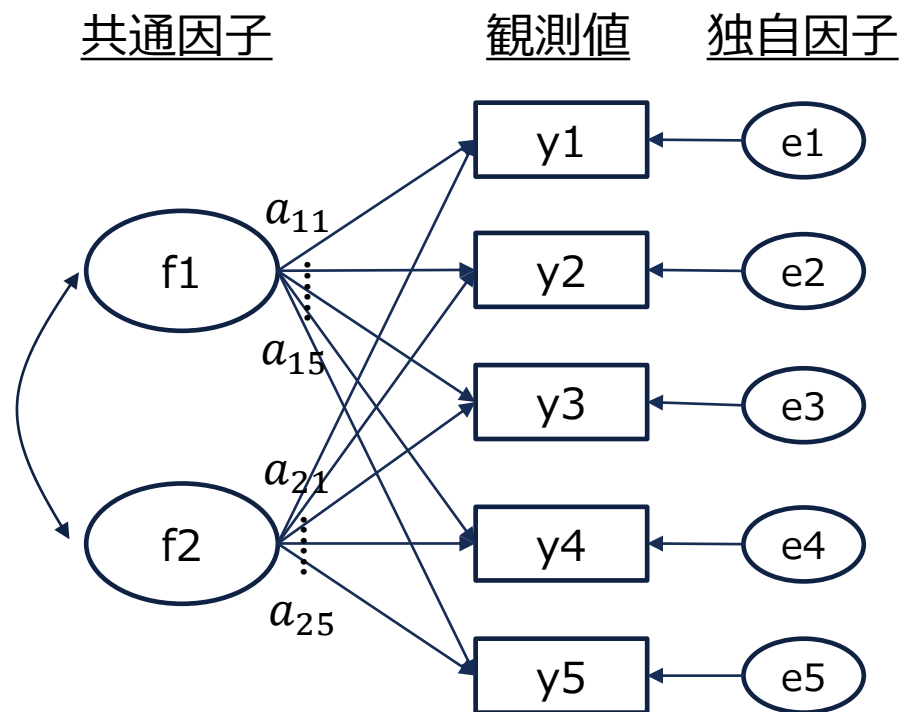
観測値の背後に共通因子と独自因子を仮定して、次元を減少する

$$\begin{pmatrix} y_{i,1} \\ y_{i,2} \\ y_{i,3} \\ y_{i,4} \\ y_{i,5} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \\ a_{14} & a_{24} \\ a_{15} & a_{25} \end{pmatrix} \begin{pmatrix} f_{i,1} \\ f_{i,2} \end{pmatrix} + \begin{pmatrix} e_{i,1} \\ e_{i,2} \\ e_{i,3} \\ e_{i,4} \\ e_{i,5} \end{pmatrix}, \quad i = 1, 2, 3, \dots, n$$

観測値 : $\mathbf{y}_i = \mathbf{A} \cdot \mathbf{f}_i + \mathbf{e}_i$

因子負荷量 因子得点 独自因子

因子分析モデルのパス図



共通・独自因子に関する仮定

- f_i は平均0、分散1
- f_i と f_j は独立
- f_i と e_i は独立
- $e_i \sim N(0, \sigma_i^2)$ 正規分布
- e_i と e_j は独立

ビジネス課題の例

某小売店では、11種類の煎餅を販売しています。ある日店長が、煎餅の売上を増加させるために、どのようなプロモーションを行えばよいか、コンサルタントに相談に来ました。今回は、来店客にアンケート調査を実施して、各製品の特徴を把握してから行いたいと言っています。

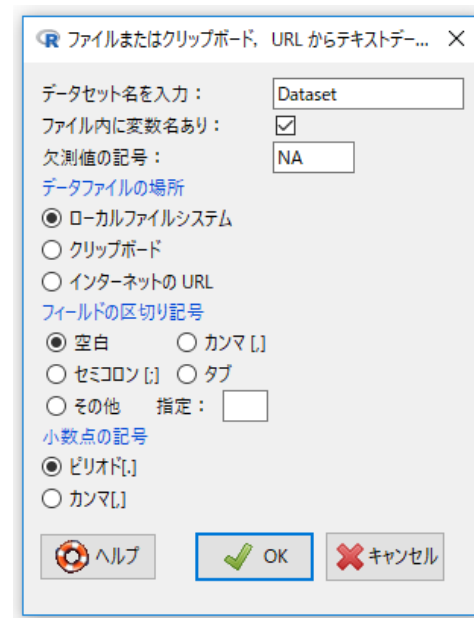
どのようなアンケート調査を実施して、アドバイスしますか？

- アンケートの質問項目とその質問の仮説を考えてください。



演習データの読み込み

- ① Rstudio起動する
- ② `> library(Rcmdr)` ※コマンドラインから Rコマンダー を起動する
- ③ 演習ファイル “senbei.txt” を読み込む
 - Rstudio `> Dataset<-read.table(“senbei.txt”)`
又は
 - Rコマンダー (データ) → (データインポート) → (テキストファイルまたはクリップボード...) →
✓ OKを選択して、senbei を指定する
- ④ 演習データが Dataset に読み込まれる



演習データの確認

Rstudioでデータ確認

※コマンドラインからDatasetデータを確認

```
> Dataset
製品名 味 パッケージデザイン 広告宣伝 素材栄養素 キャンペーンイベント
1 ハッピーターン 89 63 32 51 37
2 雪の宿 73 46 25 48 32
3 ぼたぼた焼き 65 45 17 32 21
4 黒豆せんべい 72 33 2 50 9
5 まがりせんべい 70 29 11 35 21
6 チーズアーモンド 71 20 10 38 24
7 手塩屋 71 38 3 38 13
8 ばかうけ 48 39 10 16 16
9 粒より小餅 49 26 17 27 30
10 田舎おかき 51 16 1 38 5
11 うまい！堅焼き 48 15 4 36 1
```

Rコマンドでデータ確認



	製品名	味	パッケージデザイン	広告宣伝	素材栄養素	キャンベ
1	ハッピーターン	89	63	32	51	
2	雪の宿	73	46	25	48	
3	ぼたぼた焼き	65	45	17	32	
4	黒豆せんべい	72	33	2	50	
5	まがりせんべい	70	29	11	35	
6	チーズアーモンド	71	20	10	38	
7	手塩屋	71	38	3	38	
8	ばかうけ	48	39	10	16	
9	粒より小餅	49	26	17	27	
10	田舎おかき	51	16	1	38	
11	うまい！堅焼き	48	15	4	36	

演習データについて

11種類の煎餅について、5つの項目に関して製品評価を実施した結果を示す。
各評価項目に関して、アンケート得点の平均値（100点満点）が示される。

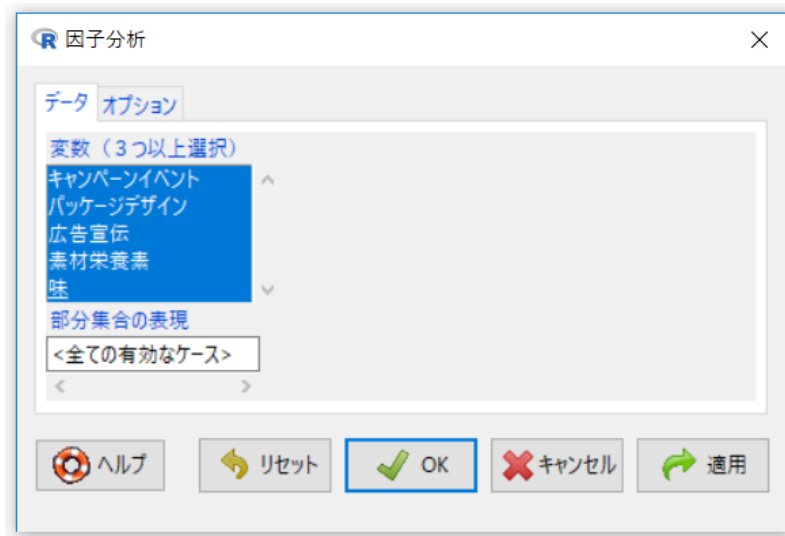
“senbei.txt”

製品名	味	パッケージデザイン	広告宣伝	素材栄養素	キャンペーンイベント
ハッピーターン	89	63	32	51	37
雪の宿	73	46	25	48	32
ばたばた焼き	65	45	17	32	21
黒豆せんべい	72	33	2	50	9
まがりせんべい	70	29	11	35	21
チーズアーモンド	71	20	10	38	24
手塩屋	71	38	3	38	13
ばかうけ	48	39	10	16	16
粒より小餅	49	26	17	27	30
田舎おかき	51	16	1	38	5
うまい！堅焼き	48	15	4	36	1

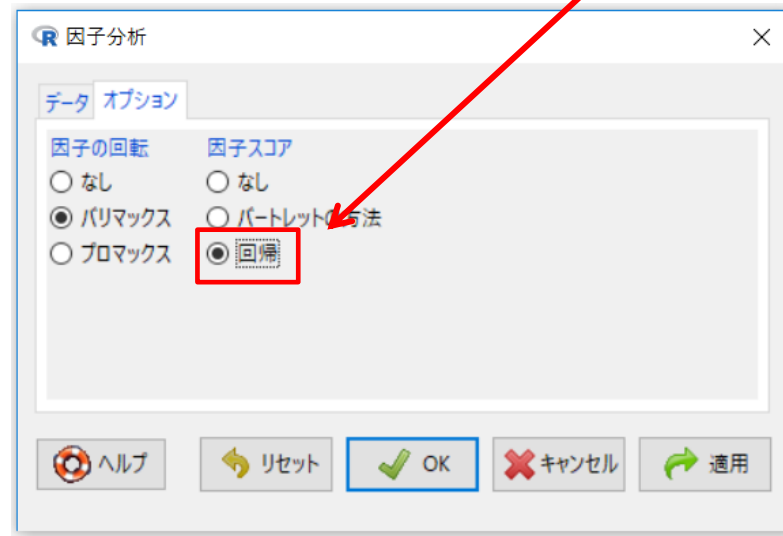
因子分析の推定

Rコマンダーから 【統計量】 → 【次元解析】 → 【因子分析】

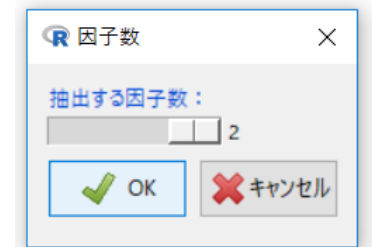
クラスタリングのための因子得点を
データ保存するため！



変数として5項目全てを選択



オプションはバリマックスと回転を選択
特に、「**回帰**」の指定は忘れずに！



因子数は 2 を設定

因子分析の推定結果

```
Rcmdr> local({
  Rcmdr+   .FA <- factanal(~キャンペーンイベント+パッケージデザイン+広告宣伝+素材栄養素+味,
  Rcmdr+                                   factors=2, rotation="varimax", scores="regression", data=Dataset)

  Rcmdr+   print(.FA)
  Rcmdr+   Dataset <- within(Dataset, {
  Rcmdr+     F2 <- .FA$scores[,2]
  Rcmdr+     F1 <- .FA$scores[,1]
  Rcmdr+   })
  Rcmdr+ })
```

```
Call:
factanal(x = ~キャンペーンイベント+パッケージデザイン+広告宣伝+素材栄養素+味,
  factors=2, data=Dataset, scores="regression", rotation = "varimax")
```

Uniquenesses:

キャンペーンイベント	パッケージデザイン	広告宣伝	素材栄養素	味
0.158	0.318	0.005	0.407	0.005

Loadings:

	Factor1	Factor2
キャンペーンイベント	0.886	0.237
パッケージデザイン	0.696	0.445
広告宣伝	0.985	0.158
素材栄養素	0.119	0.761
味	0.371	0.926

マーケティング因子 (Factor1) と 製品因子 (Factor2) が示されています。

	Factor1	Factor2
SS loadings	2.392	1.715
Proportion Var	0.478	0.343
Cumulative Var	0.478	0.821

各因子負荷の二乗平均

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 4.16 on 1 degree of freedom.
The p-value is 0.0415

因子分析では、因子（Factor1とFactor2）が表す意味が重要！ 各種要因から軸の意味を定義しないと知覚マップが表せない。

因子負荷量と因子得点の推定結果

$$y_k = a_{1,k}f_1 + a_{2,k}f_2 + e_k$$

$$k = 1, 2, 3, 4, 5$$

キャンペーンイベント
パッケージデザイン
広告宣伝
素材栄養素
味

	Factor1	Factor2
キャンペーンイベント	0.886	0.237
パッケージデザイン	0.696	0.445
広告宣伝	0.985	0.158
素材栄養素	0.119	0.761
味	0.371	0.926

因子負荷量： $(a_{1,k}, a_{2,k})$

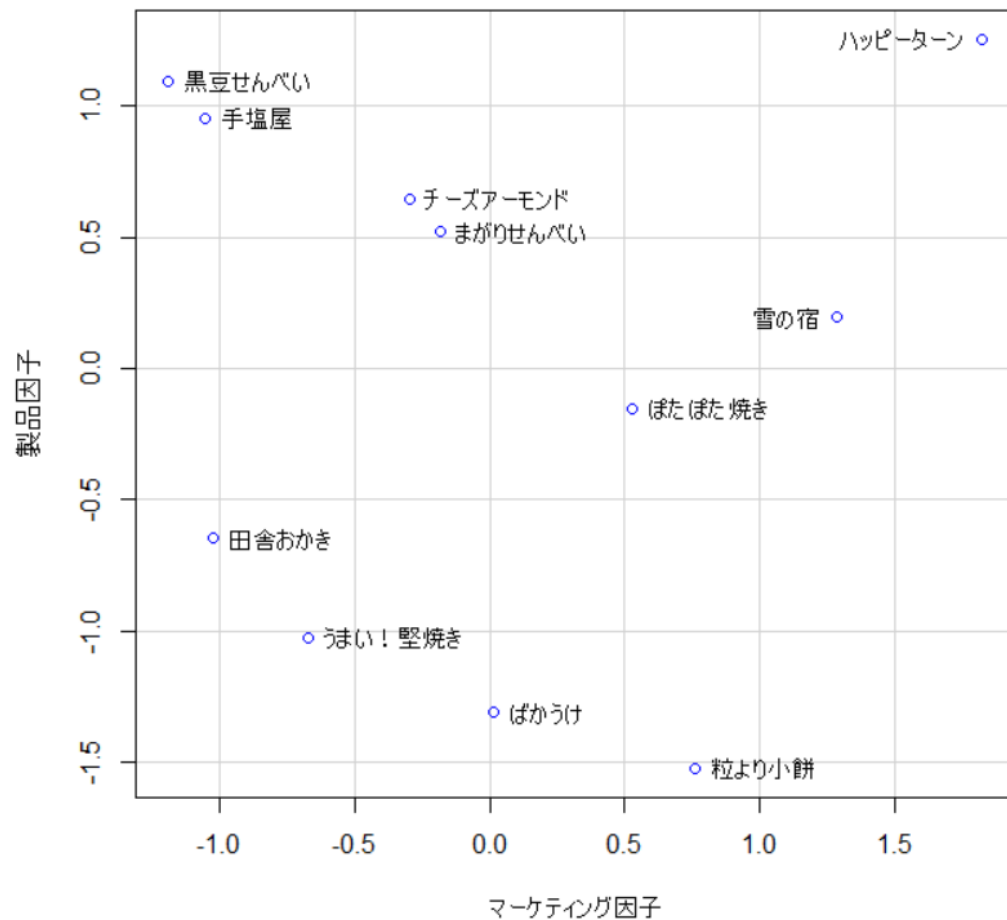
クラスタリング用データ

観測値： y_k

因子得点： (f_1, f_2)

製品名	味	パッケージデザイン	広告宣伝	素材栄養素	キャンペーンイベント	F1	F2
ハッピーターン	89	63	32	51	37	1.82050452	1.2510722
雪の宿	73	46	25	48	32	1.28176903	0.1942184
ぼたぼた焼き	65	45	17	32	21	0.52926245	-0.1552535
黒豆せんべい	72	33	2	50	9	-1.18662160	1.0960546
まがりせんべい	70	29	11	35	21	-0.17977160	0.5197866
チーズアーモンド	71	20	10	38	24	-0.29757433	0.6461495
手塩屋	71	38	3	38	13	-1.05010910	0.9528750
ばかうけ	48	39	10	16	16	0.01599978	-1.3103923
粒より小餅	49	26	17	27	30	0.76018030	-1.5198705
田舎おかき	51	16	1	38	5	-1.02021352	-0.6475204
うまい！堅焼き	48	15	4	36	1	-0.67342593	-1.0271194

知覚マップの表示





11種類の煎餅は、どのようにクラスタリングすべきか？

	Factor1	Factor2
キャンペーンイベント	0.886	0.237
パッケージデザイン	0.696	0.445
広告宣伝	0.985	0.158
素材栄養素	0.119	0.761
味	0.371	0.926


グラフは散布図から、F1とF2を指定する

ブランド名の表示

Rコマンドから 【データ】
→ 【アクティブデータセット】
→ 【ケース名の設定】



クラスター分析



クラスター分析

分類対象の集合を内的結合と外的分離が達成できるような部分集合に分割する多変量解析の手法の一つで、何らかの基準で同質性を判断する。類似度や距離を定義して、複数の部分集合に分ける。

■ 階層的クラスタリング

- 分類対象を分岐、または凝縮しながら階層的に分ける方法
- 分類されたクラスターは安定的である

■ 非階層的クラスタリング

- クラスター数を与件とし、統計的当てはまりが良くなるよう非階層的に分ける方法
- 分類されたクラスターは不安定なことがある

階層的クラスタリング

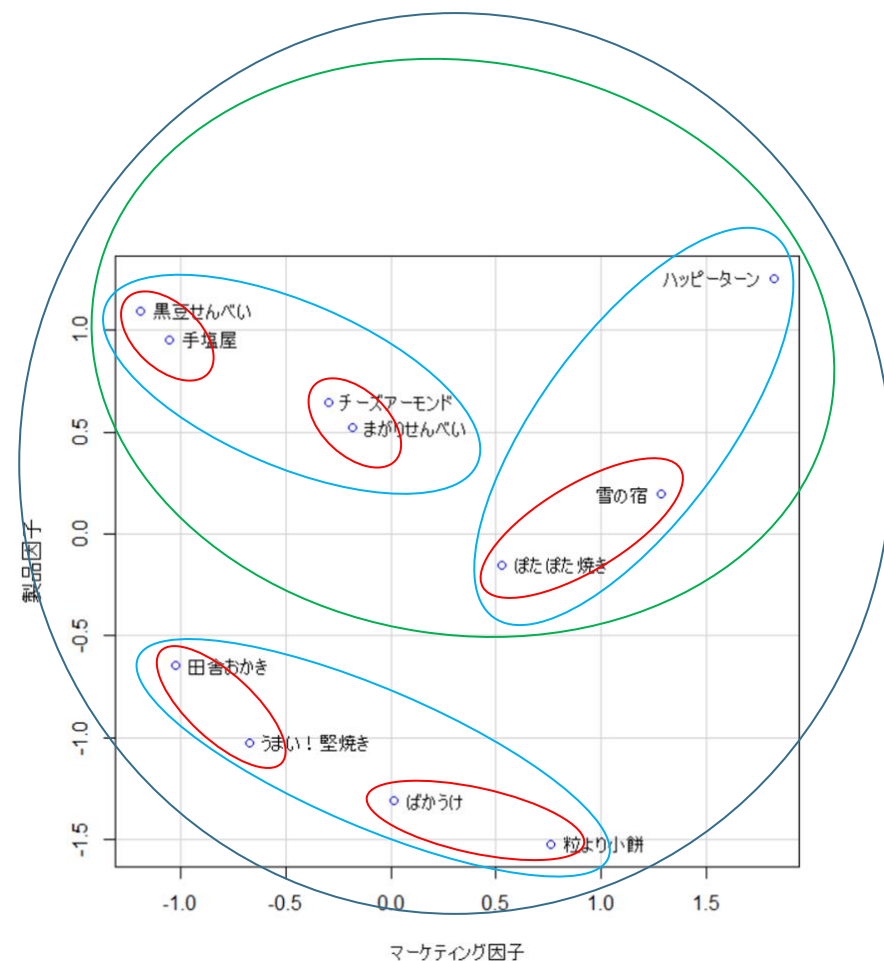
- ① データの中から互いの距離が最も近くなるペアを初期統合する
- ② クラスタ同士で互いの距離が最も近くなるもの同士で統合する
- ③ 上記を全体が一つのクラスタに統合されるまで繰り返す

【ワード法】 クラスタ内平方和の増分が少ないものを統合する（最小分散法）

$$d_{xc}^2 = \frac{n_x + n_a}{n_x + n_a + n_b} d_{xa}^2 + \frac{n_x + n_b}{n_x + n_a + n_b} d_{xb}^2 - \frac{n_x}{n_x + n_a + n_b} d_{ab}^2$$

$$\text{但し、} d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2, i, j = 1, 2, 3, \dots, n$$

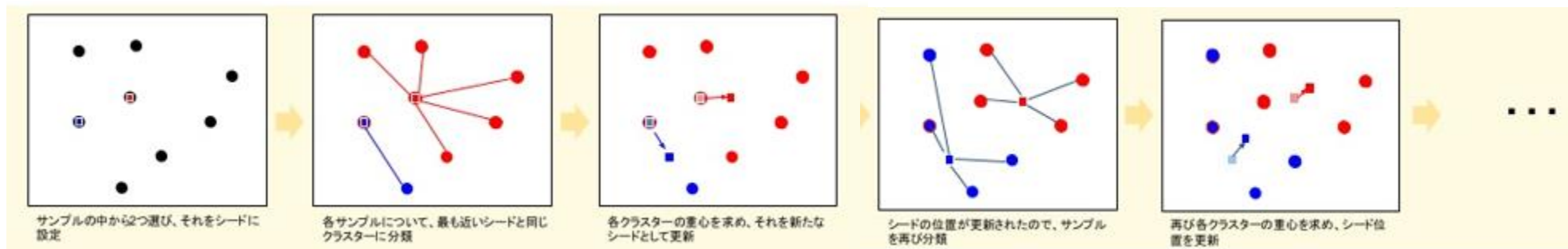
その他にも、重心法、メディアン法、最長距離法、最短距離法、群間平均法がある。



非階層的クラスタリング

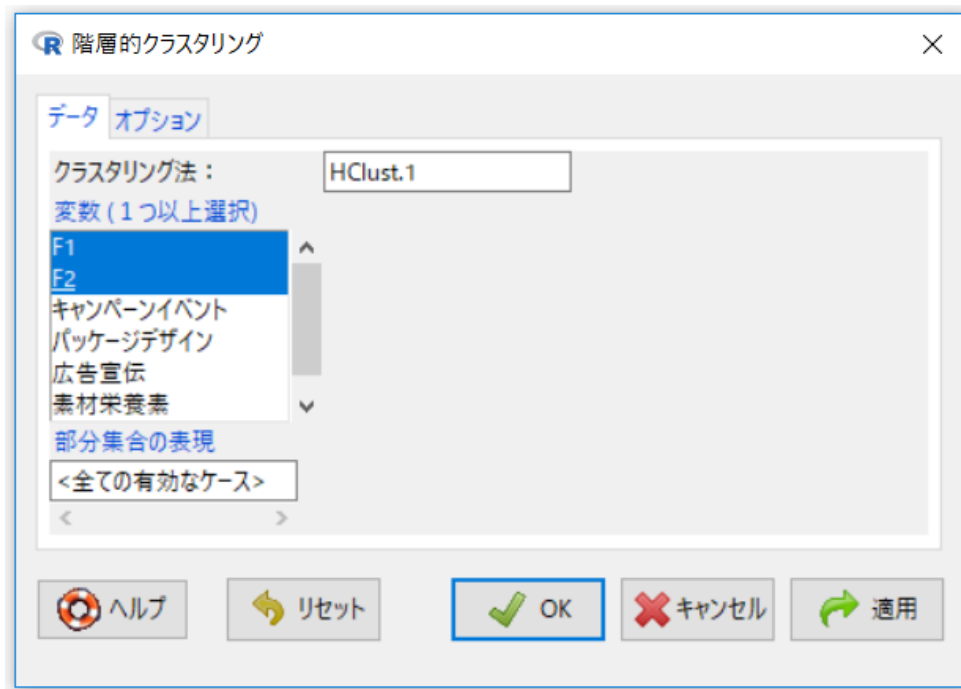
- 【k-means法】
- ① クラスタ数 (k) を決め、 k 個の適当な点（分類中心点）を与える
 - ② 各個体を最も近い分類中心点に割当て k 個の暫定クラスタを生成する
 - ③ 各クラスタの重心を求め、新たな分類中心を算出する
 - ④ 新しい分類中心に対して、分類中心の差が少なくなるまで②と③を繰り返す

(例) $k=2$ の場合

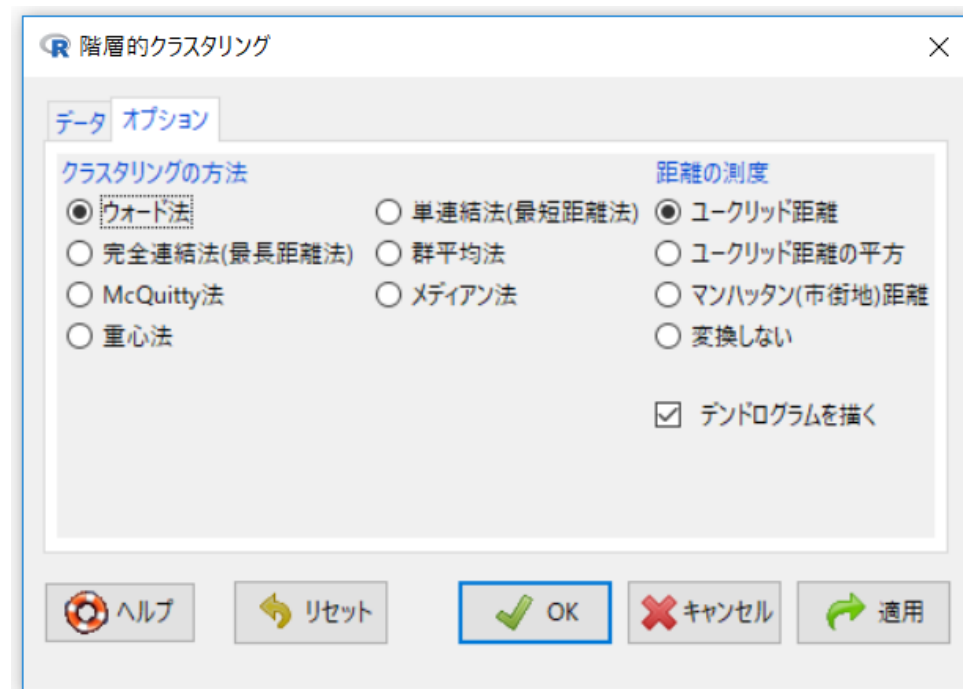


階層的クラスタリングの推定：ワード法

Rコマンダーから【統計量】→【次元解析】→【クラスター分析】→【階層的クラスター分析】



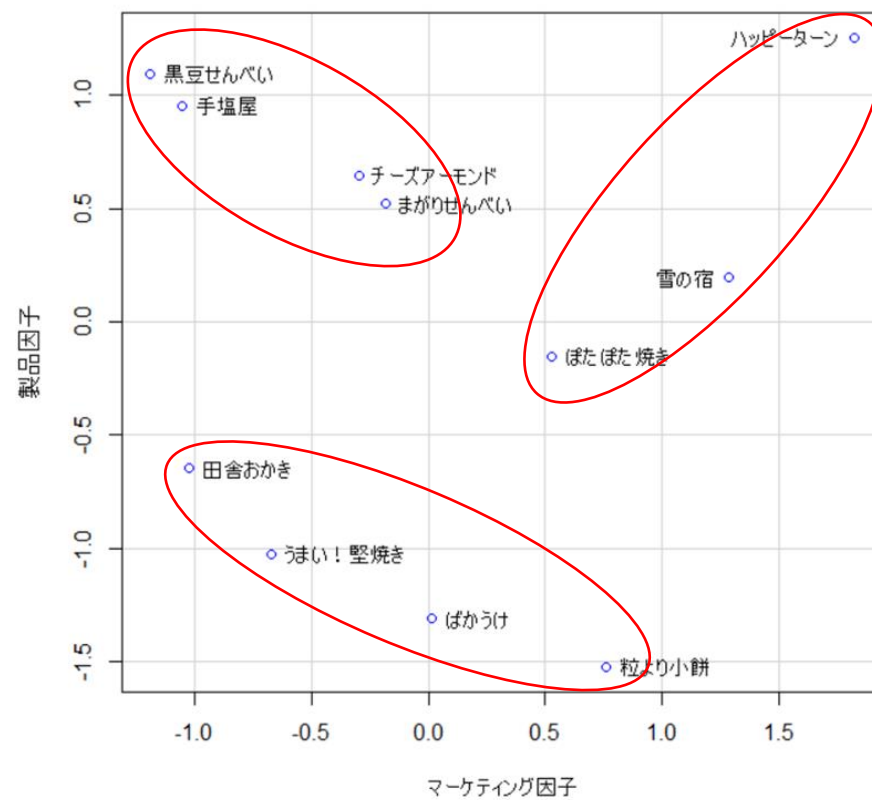
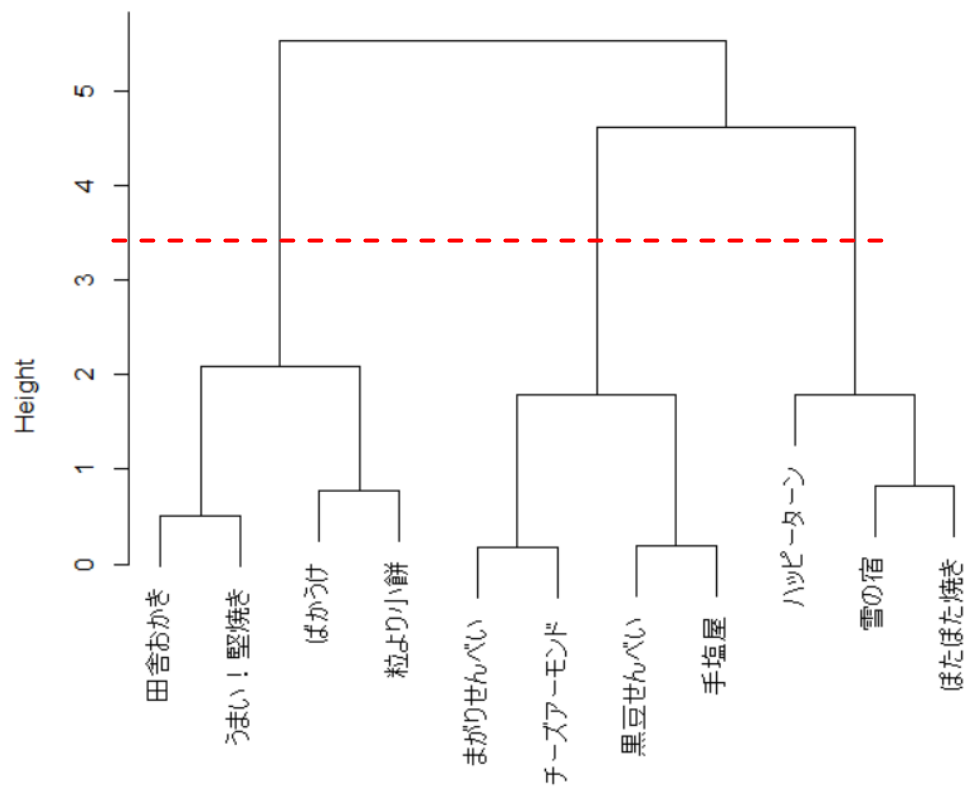
F1とF2（因子得点）を選択



ワード法とユークリッド距離を選択

3 セグメントの例：ワード法

Cluster Dendrogram for Solution HClust.1



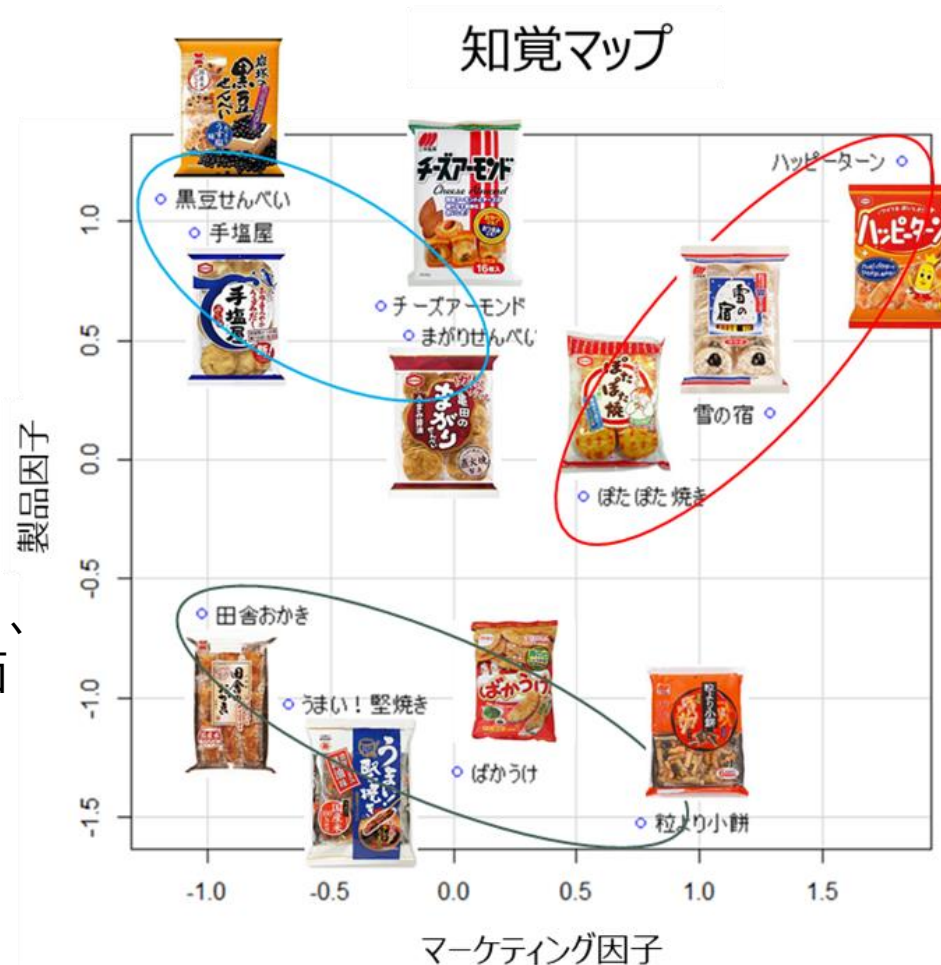
消費者の嗜好に関する分析

- 宣伝広告では印象がうすが、味や素材で他より評価が高い

昔からのベストセラーで認知度が高く、棚の端中段あたりに安定的に配置する。

- 宣伝広告の印象がうすく、味や素材でも他より評価が高くない

玄人好みの商品で、ユーザに関する詳細な調査を実施して、消費者理解を深める。（味の濃い派な中高年層？）



- 宣伝広告が印象的で、味や素材でも評価が他よりも高い

大衆受けする。宣伝広告の影響も大きいため、目に良くつく棚の目線の高さに配置する。

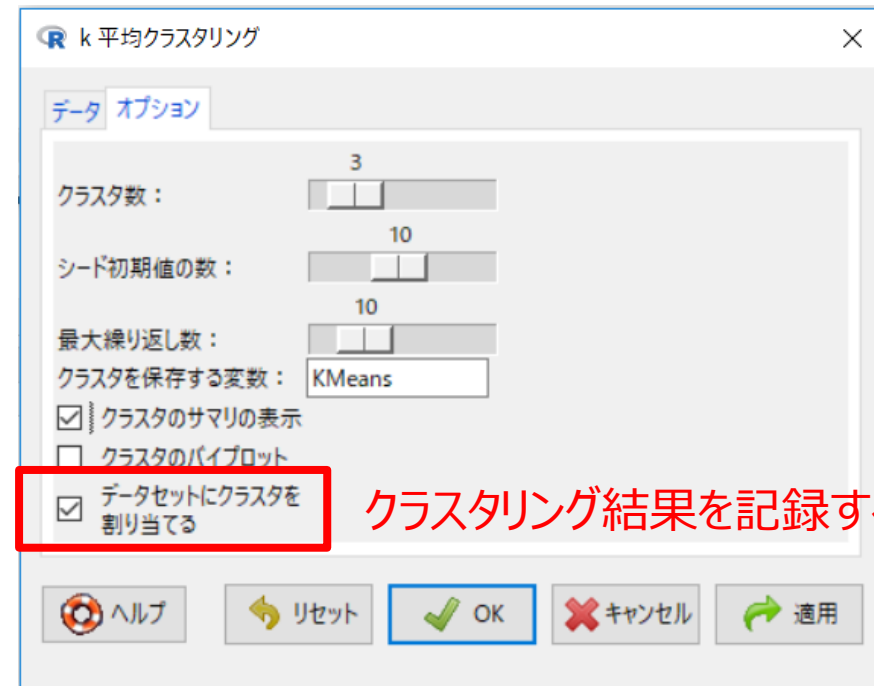
プロモーションやキャンペーンと合わせて山積みやPOPを立て、チラシにも載せる。

非階層的クラスタリングの推定：K平均法

Rコマンダーから【統計量】 → 【次元解析】 → 【クラスター分析】 → 【k-平均クラスター分析】



F1とF2（因子得点）を選択



クラスタリング結果を記録する！

ワード法とユークリッド距離を選択

K平均法のクラスタリング結果

3つの場合

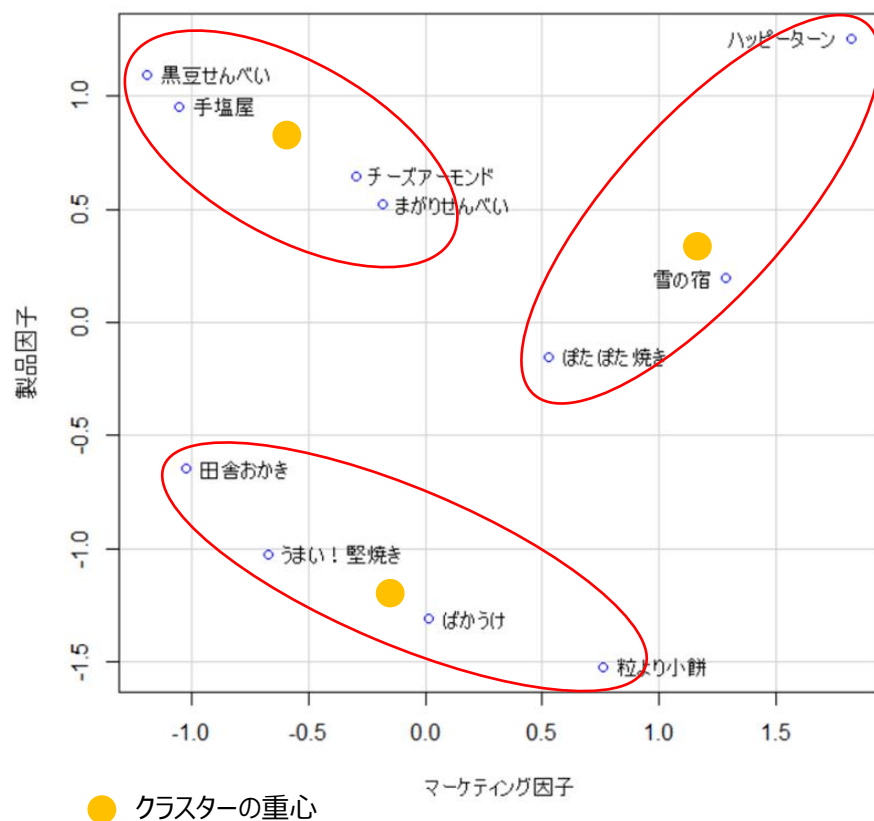
4つの場合

	rowname	味	パッケージデザイン	広告宣伝	素材栄養素	キャンペーンイベント	F1	F2	KMeans
1	ハッピーターン	89	63	32	51	37	1.82050452	1.2510722	1
2	雪の宿	73	46	25	48	32	1.28176903	0.1942184	1
3	ぼたぼた焼き	65	45	17	32	21	0.52926245	-0.1552535	1
4	黒豆せんべい	72	33	2	50	9	-1.18662160	1.0960546	3
5	まがりせんべい	70	29	11	35	21	-0.17977160	0.5197866	3
6	チーズアーモンド	71	20	10	38	24	-0.29757433	0.6461495	3
7	手塩屋	71	38	3	38	13	-1.05010910	0.9528750	3
8	ばかうけ	48	39	10	16	16	0.01599978	-1.3103923	2
9	粒より小餅	49	26	17	27	30	0.76018030	-1.5198705	2
10	田舎おかき	51	16	1	38	5	-1.02021352	-0.6475204	2
11	うまい！堅焼き	48	15	4	36	1	-0.67342593	-1.0271194	2



KMeans
4
4
1
2
2
2
2
1
1
3
3

3 セグメントの場合：K平均法



```
Rcmdr> .cluster <- KMeans(model.matrix(~-1 + F1 + F2, Dataset),  
  centers = 3, iter.max = 10, num.seeds = 10)
```

```
Rcmdr> .cluster$size # cluster sizes  
[1] 3 4 4
```

```
Rcmdr> .cluster$centers # cluster centroids
```

	new.x.F1	new.x.F2
1	1.2105120	0.4300123
2	-0.6785192	0.8037164
3	-0.2293648	-1.1262257

各クラスターの重心

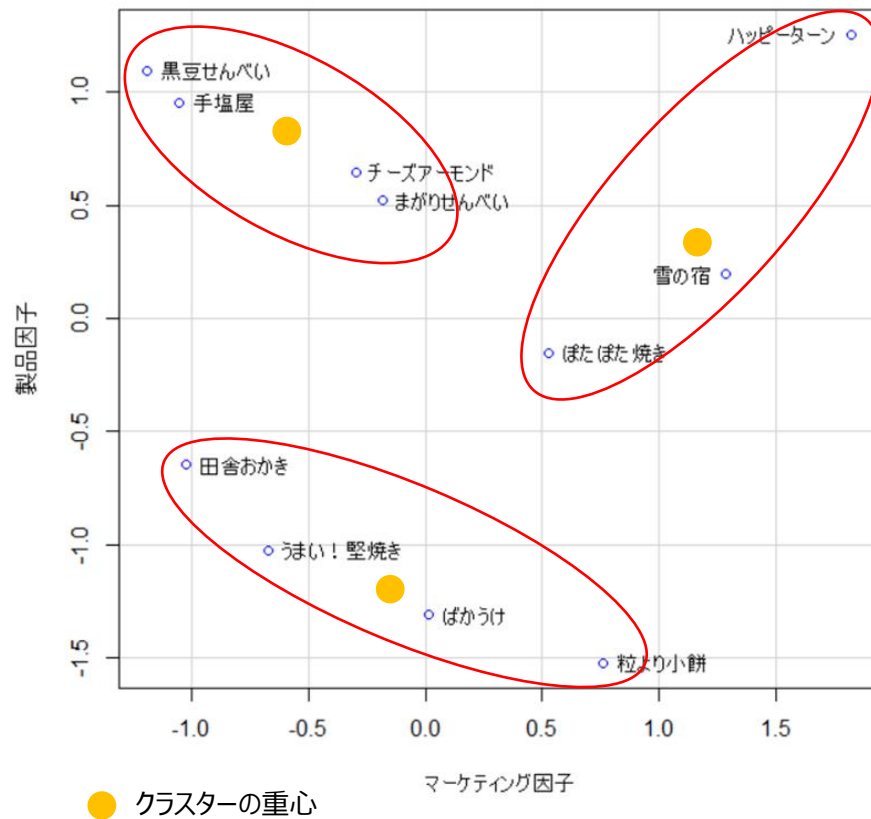
```
Rcmdr> .cluster$withinss # within cluster sum of squares  
[1] 1.913544 1.003269 2.289890
```

```
Rcmdr> .cluster$tot.withinss # total within sum of squares  
[1] 5.206702
```

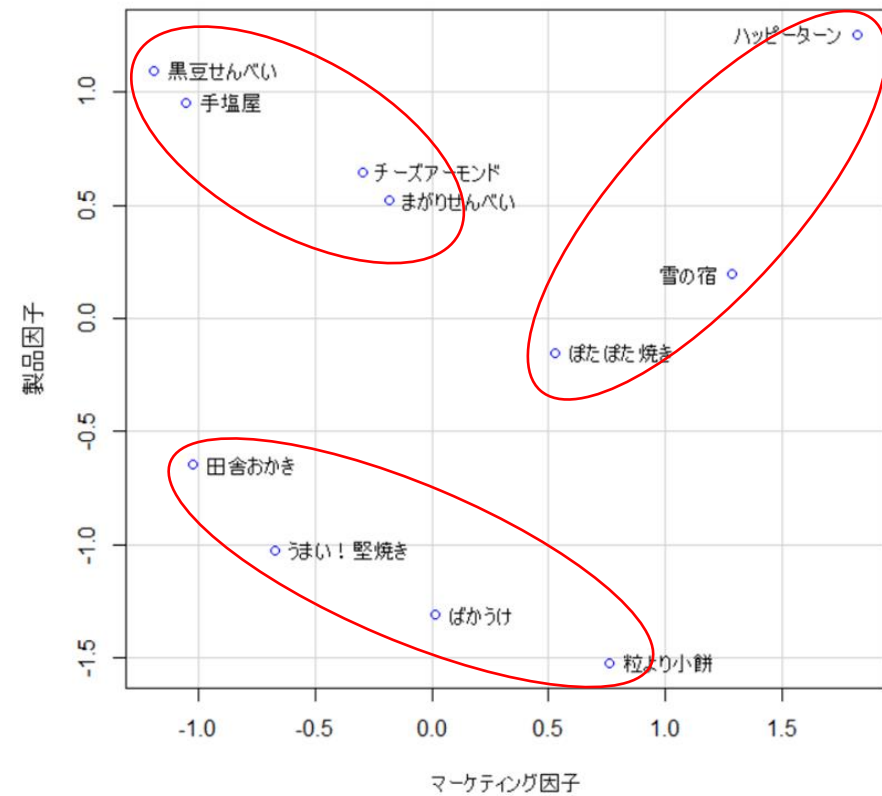
```
Rcmdr> .cluster$betweenss # between cluster sum of squares  
[1] 14.66011
```

K平均法 VS. ウォード法

k平均法による

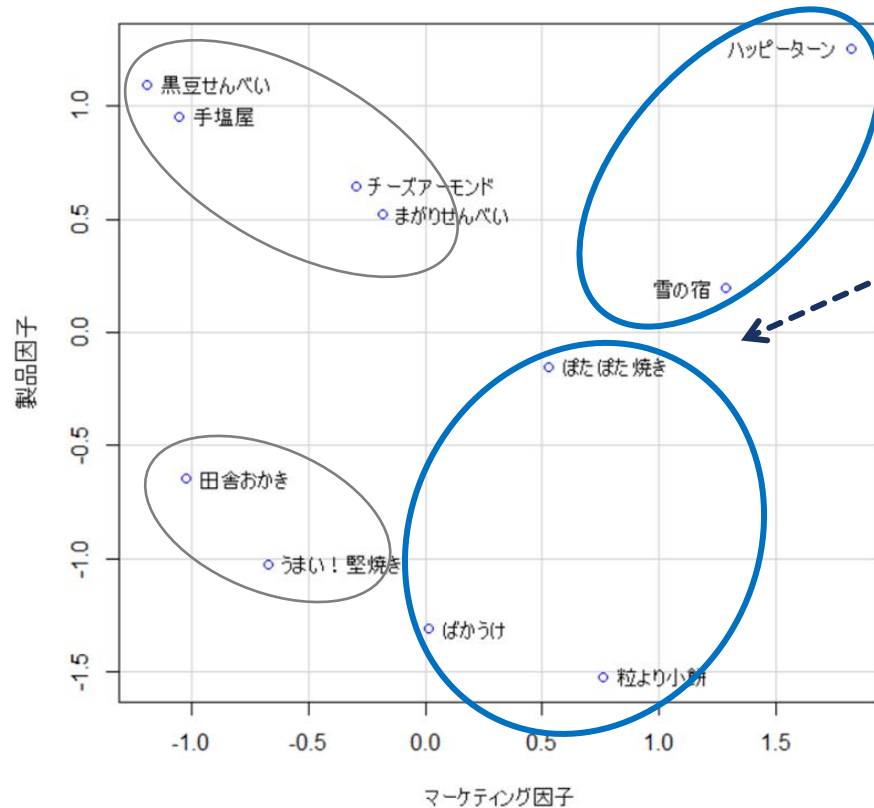


ウォード法による



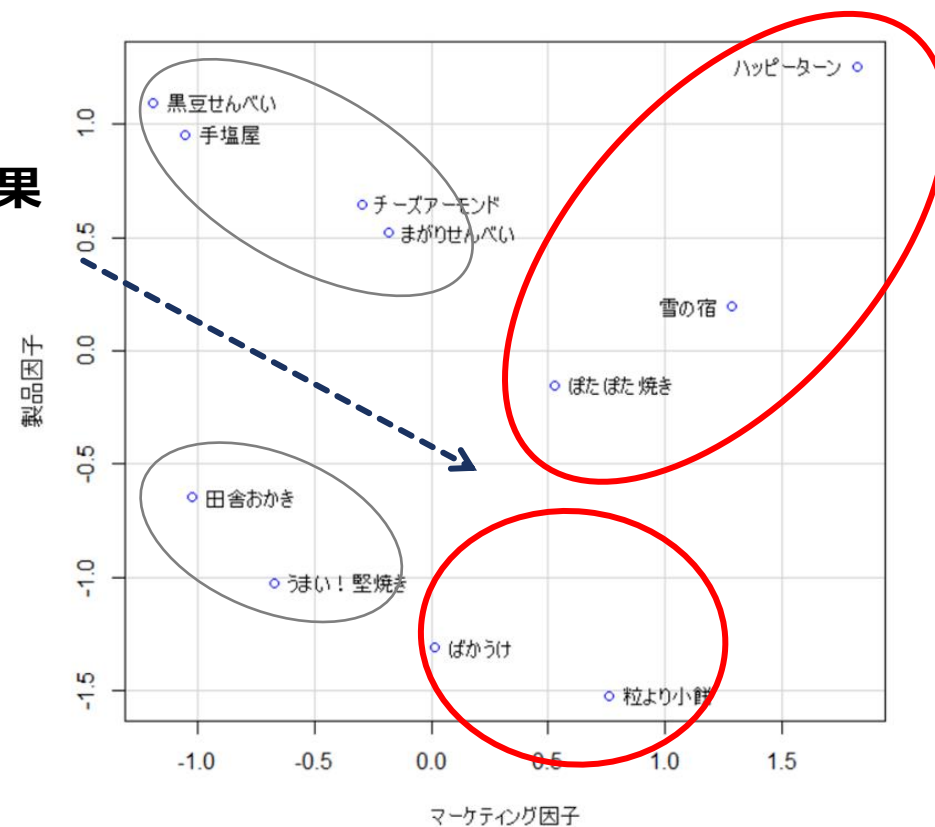
K平均法 VS. ウォード法（4セグメントの場合）

k平均法による



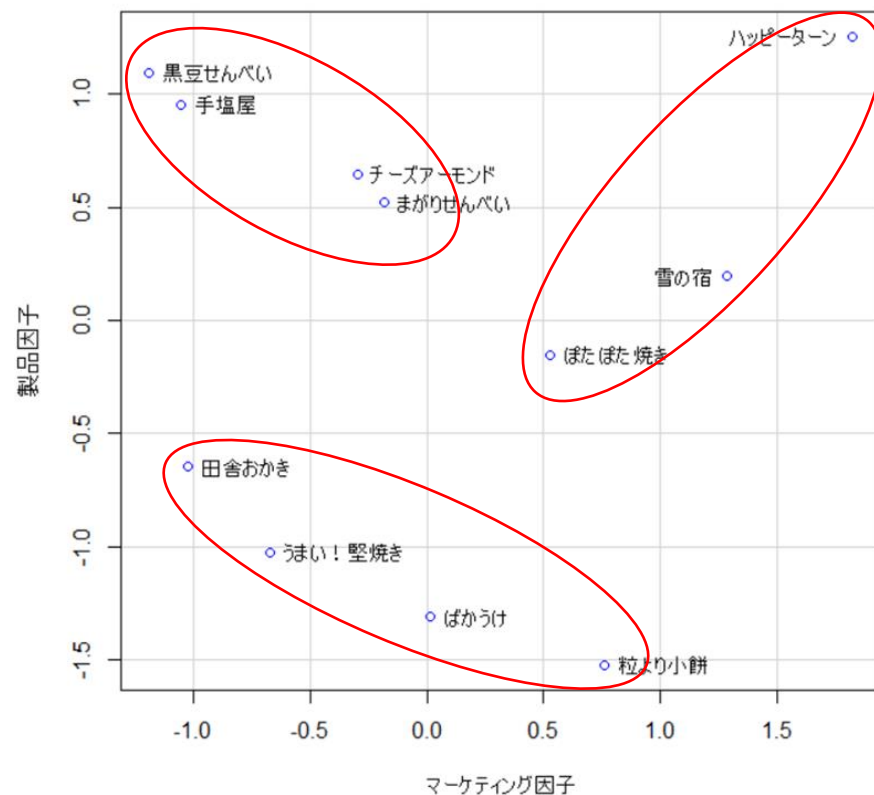
セグメント結果
が異なる!

ウォード法による

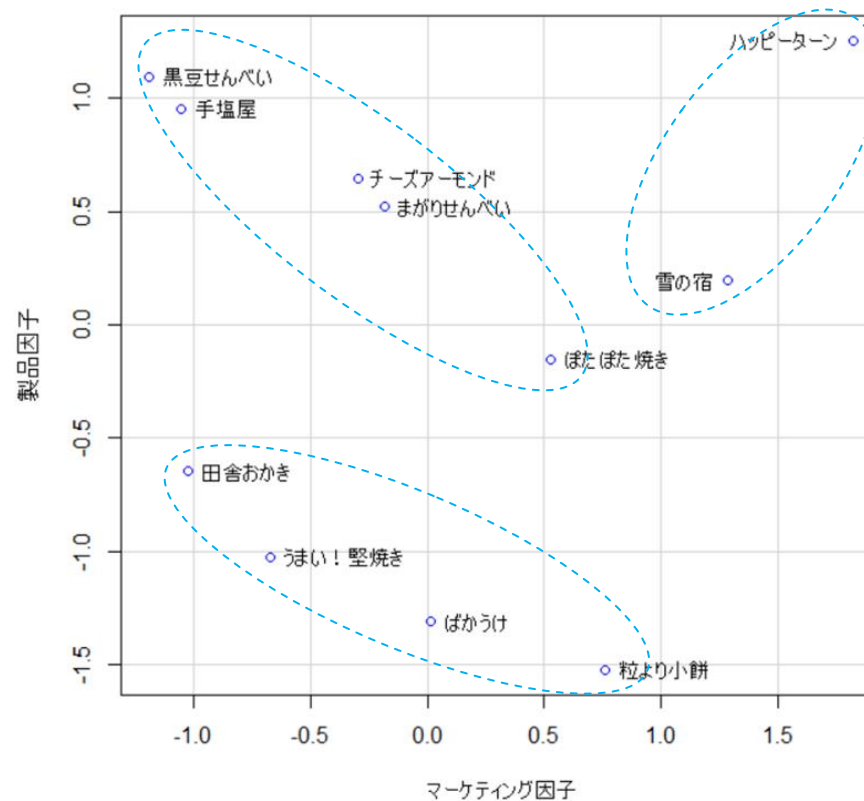


K平均法（入力情報が異なる場合）

因子得点(F1、F2)による



アンケート5項目による



課題：ノートパソコン購入に関するクラスター分析

100人の消費者に対して、ノートパソコンを購入する際に、以下の6項目をどれだけ重視するか5尺度でアンケート調査した。その際に得られた評価の平均点が下表に示される。この結果に対して、因子分析を実施して得られる因子得点を用いて、3セグメントのクラスター分析（ウォード法およびk平均法）を行い、結果を考察せよ。

製品	デザイン	ディスプレイ	バッテリー	カラー	アプリ	メモリー
AA	4.78	4.72	4.67	4.72	4.84	4.84
BU	4.82	4.41	4.01	4.83	4.91	4.42
CG	4.72	4.38	4.29	4.77	4.54	4.39
DS	4.61	4.24	3.83	4.63	4.72	4.79
EP	4.54	3.69	4.22	4.51	4.28	3.96
FA	4.51	4.13	4.18	4.45	4.38	3.85
GT	4.73	4.64	4.04	4.64	4.73	5.02
HK	4.64	4.11	4.36	4.72	4.91	3.48



データマイニングを楽しもう！