多変量解析

目的

- データの情報集約 (主成分分析)
- データの背後にある要因の解析 (因子分析)
- データの分類，パターンわけ (クラスター分析)
- 市場予測 (重回帰分析)
- 新製品コンセプトの開発 (コンジョイント分析)

手法

- すべては正準相関分析
- データの相関を解析する

キーワード

- 平均，分散，標準偏差
- 共分散，相関
- 固有値，固有ベクトル

基本統計量

- データ: $x(i),\ i = 1, \ldots, N$
- 平均: $\bar{x} = \dfrac{1}{N}\sum_{i=1}^{N} x(i)$
- 平均偏差 (データの平均からのずれ): $x(i) - \bar{x}$
- 分散 (データの散らばり具合い．データセットの平均からのずれの度合い，標準偏差の 2 乗):

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x(i) - \bar{x})^2$$

- 共分散 (データのずれの関係): $s_{xy} = \dfrac{1}{N}\sum_{i=1}^{N}(x(i)-\bar{x})(y(i)-\bar{y})$
- 規準化 (正規化): $x_r(i) = \dfrac{x(i) - \bar{x}}{s_x}$
- 相関 (規準化されたデータの散らばり具合い): $r_{xy} = \dfrac{s_{xy}}{s_x s_y}$

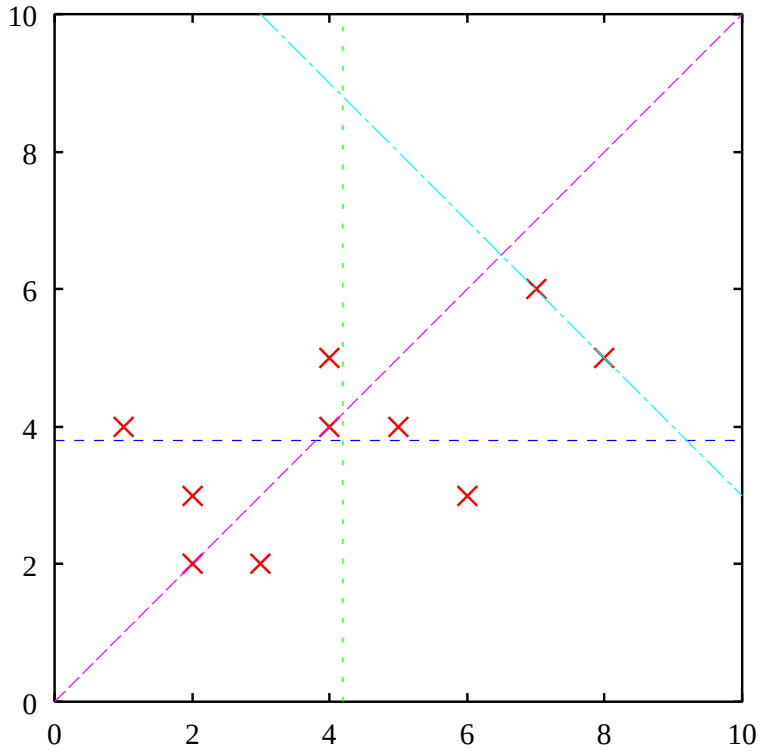
- ベクトル表現

$$\boldsymbol{x} = \begin{bmatrix} x(1) \\ \vdots \\ x(N) \end{bmatrix}, \tilde{\boldsymbol{x}} = \begin{bmatrix} x(1) - \bar{x} \\ \vdots \\ x(N) - \bar{x} \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix}, \tilde{\boldsymbol{y}} = \begin{bmatrix} y(1) - \bar{y} \\ \vdots \\ y(N) - \bar{y} \end{bmatrix}$$

- 共分散行列

$$S = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N}[\tilde{\boldsymbol{x}} \ \ \tilde{\boldsymbol{y}}]^T[\tilde{\boldsymbol{x}} \ \ \tilde{\boldsymbol{y}}]$$

基本統計量の例



- 10 人の学生 (A から J) の英語と数学の点数が以下のとおりであった．総合成績と文理度の順位をつけよ．また，その根拠を説明せよ．

科目	A	B	C	D	E	F	G	H	I	J
英語	2	1	2	3	5	4	8	6	7	4
数学	3	4	2	2	4	4	5	3	6	5

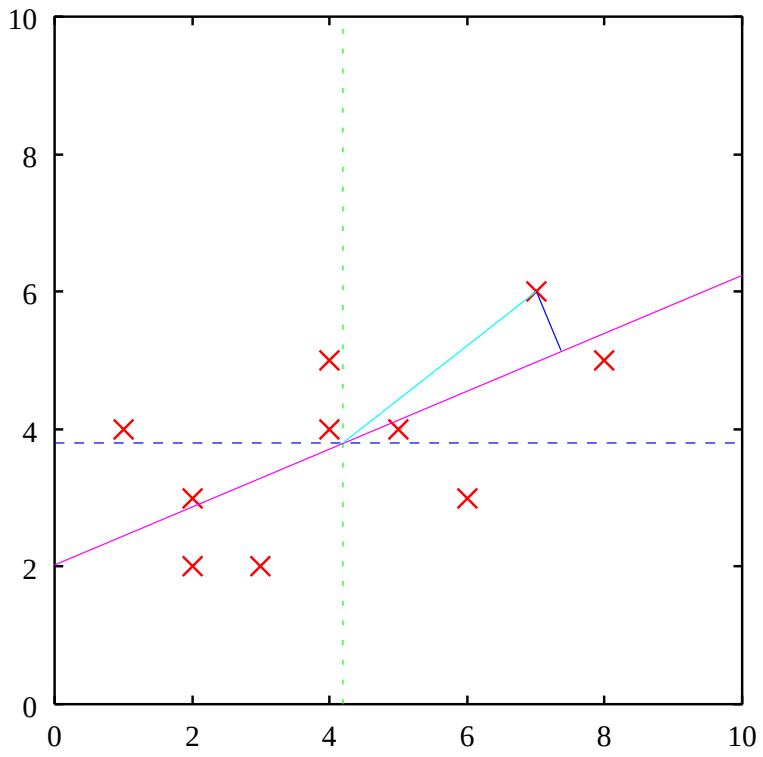
- データ: $E = [2\ 1\ 2\ 3\ 5\ 4\ 8\ 6\ 7\ 4]^T, M = [3\ 4\ 2\ 2\ 4\ 4\ 5\ 3\ 6\ 5]^T$
- 平均: $\bar{E} = 4.2, \bar{M} = 3.8$
- 分散: $s_E^2 = 4.76, s_M^2 = 1.56$
- 総合成績: $E + M$ による評価はデータを $x = y$ 上に射影して評価 (平均，分散が考慮されていない)

情報量最大化の概念

- 総合成績 $\rightarrow$ 2変量のデータを1変量で表す $\rightarrow$ 軸上への投影
- 軸上への投影: 点の位置ベクトル (x, y) と軸の方向 (a, b) の内積 $z = ax + by$. ただし $a^2 + b^2 = 1$
- 情報量を最大: 軸上に投影されたデータの分散が最大
- 軸 (方向ベクトル (a, b)) に射影するとして，分散を計算

$$
\begin{aligned}
s_z^2 &= \frac{1}{N}\sum_{i=1}^{N}(z(i)-\bar{z})^2 = \frac{1}{N}\sum_{i=1}^{N}(ax(i)+by(i)-a\bar{x}-b\bar{y})^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(a(x(i)-\bar{x})+b(y(i)-\bar{y}))^2 \\
&= a^2 s_x^2 + 2ab s_{xy} + b^2 s_y^2
\end{aligned}
$$

[注意:] $ax(i)+by(i)-a\bar{x}-b\bar{y}$ は点 $(x(i), y(i))$ と平均点 $(\bar{x}, \bar{y})$ を通り方向 (a, b) を持つ直線との距離.



情報量最大化の手法

- Lagrange の未定乗数法

$$\begin{aligned} F(a,b,\lambda) &= a^2 s_x^2 + 2abs_{xy} + b^2 s_y^2 - \lambda(a^2+b^2-1) \\ 0 = F_a &= 2as_x^2 + 2bs_{xy} - 2\lambda a \\ 0 = F_b &= 2as_{xy} + 2bs_y^2 - 2\lambda b \\ 0 = F_\lambda &= -a^2 - b^2 + 1 \end{aligned}$$

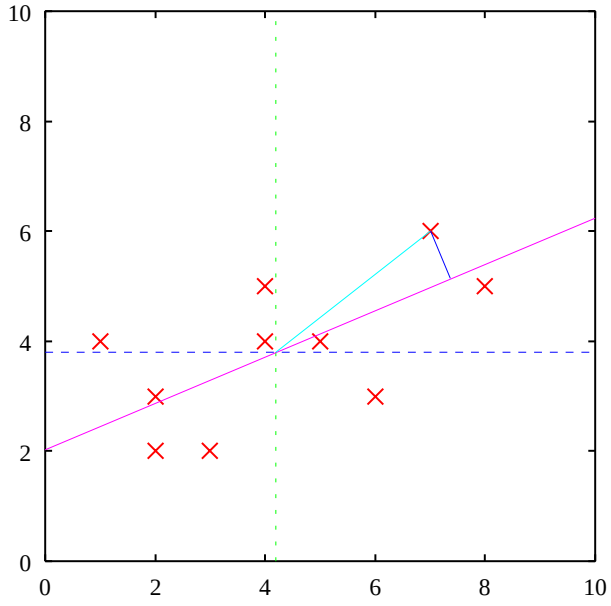
- 固有値問題

$$\begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \end{bmatrix}$$

- 対称行列ゆえ固有値 (λ) は実数で正.

$$s_z^2 = -abs_{xy} + \lambda a^2 - abs_{xy} + \lambda b^2 + 2abs_{xy} = \lambda$$

- **固有値を $\lambda_1,\ \lambda_2\ (\lambda_1 > \lambda_2 > 0)$ とするとき，分散の最大値は λ_1 であり，投影すべき軸の方向は固有ベクトルに等しい.**
- 軸への投影だけで評価する場合，軸からの距離の情報は無視される.



座標変換

- データ $(x(i), y(i)),\ i = 1, \ldots N$ を情報量最大化の観点から変量 $(z(i), w(i))$ に変換する.

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} x(1) - \bar{x} \\ \vdots \\ x(N) - \bar{x} \end{bmatrix}, \tilde{\boldsymbol{y}} = \begin{bmatrix} y(1) - \bar{y} \\ \vdots \\ y(N) - \bar{y} \end{bmatrix}, \boldsymbol{z} = \begin{bmatrix} z(1) \\ \vdots \\ z(N) \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} w(1) \\ \vdots \\ w(N) \end{bmatrix}$$

- $\boldsymbol{z} = a\tilde{\boldsymbol{x}} + b\tilde{\boldsymbol{y}}, \boldsymbol{w} = c\tilde{\boldsymbol{x}} + d\tilde{\boldsymbol{y}}$ とすると，係数 (a, b) はデータ $(\boldsymbol{x}, \boldsymbol{y})$ の共分散行列

$$S = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}$$

の最大固有値 λ_1 に対応する固有ベクトル $\boldsymbol{v}_1$ $(\|\boldsymbol{v}_1\| = 1)$ にとるのがよい.

$$\boldsymbol{z} = [\bar{\boldsymbol{x}}\ \ \bar{\boldsymbol{y}}] \begin{bmatrix} a \\ b \end{bmatrix} \boldsymbol{z} = X\boldsymbol{v}_1$$

- 係数 c, d: $\boldsymbol{v}_1$ に直交するようにとる．つまり $\boldsymbol{v}_2$ とする
  - 第 2 固有値 λ_2 に対応する固有ベクトル $\boldsymbol{v}_2$
- 座標変換
  - 共分散行列 S の対角化: $S = V\Lambda V^T,\ V = [\boldsymbol{v}_1\ \boldsymbol{v}_2]$

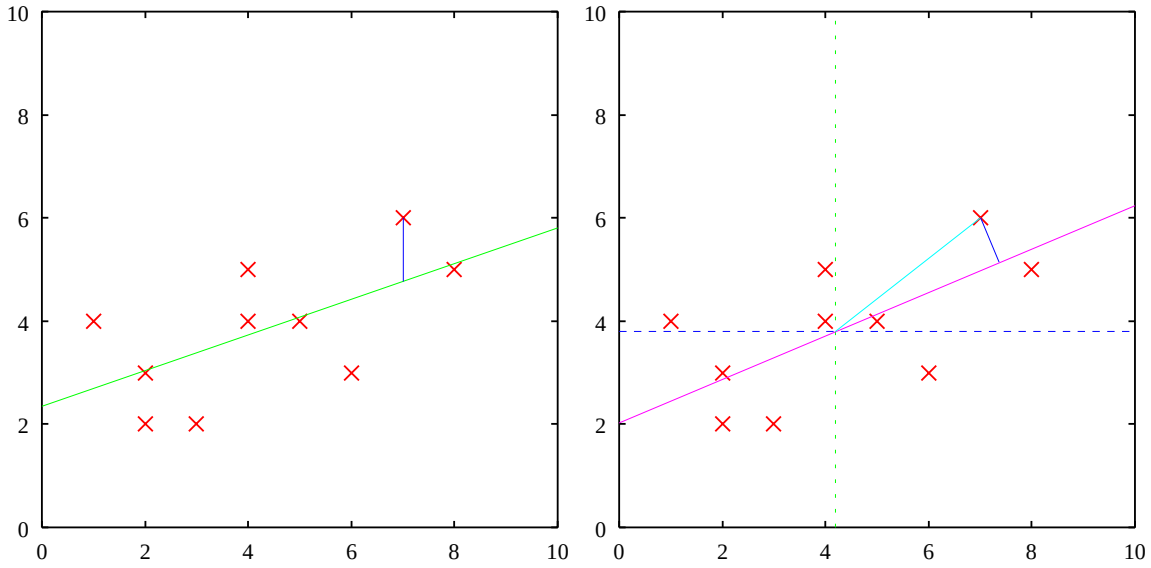
$$[\boldsymbol{z}\ \ \boldsymbol{w}] = [\bar{\boldsymbol{x}}\ \ \bar{\boldsymbol{y}}][\boldsymbol{v}_1\ \ \boldsymbol{v}_2] = XV$$

  - 共分散行列 S':

$$S' = \frac{1}{N}[\boldsymbol{z}\ \boldsymbol{w}]^T[\boldsymbol{z}\ \boldsymbol{w}] = V^T\frac{1}{N}[\tilde{\boldsymbol{x}}\ \tilde{\boldsymbol{y}}]^T[\tilde{\boldsymbol{x}}\ \tilde{\boldsymbol{y}}]V = V^TSV = \Lambda$$

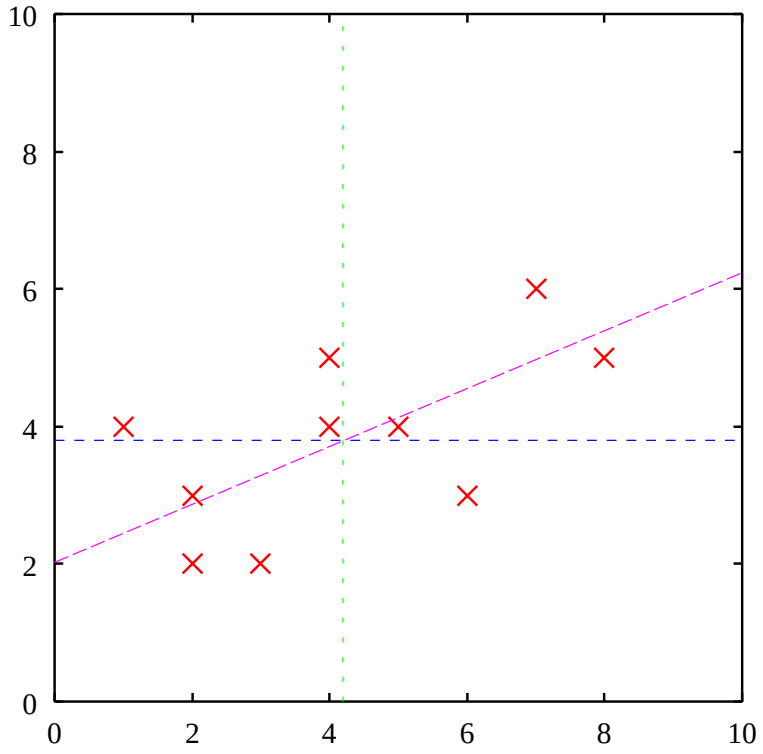
- Λ は対角行列: $\boldsymbol{z}$ と $\boldsymbol{w}$ は無相関

最小2乗法との違い



- 最小2乗法: 点から縦方向に計った直線との距離を最小化
- 主成分分析:
  - 第一主成分: 軸への投影データ
    第一主成分の分散が最大. 情報量が最大.
  - 第二主成分: 軸への垂線の足データ ... 無視される
    第二主成分の分散が最小. 無視されるデータの情報が最小.

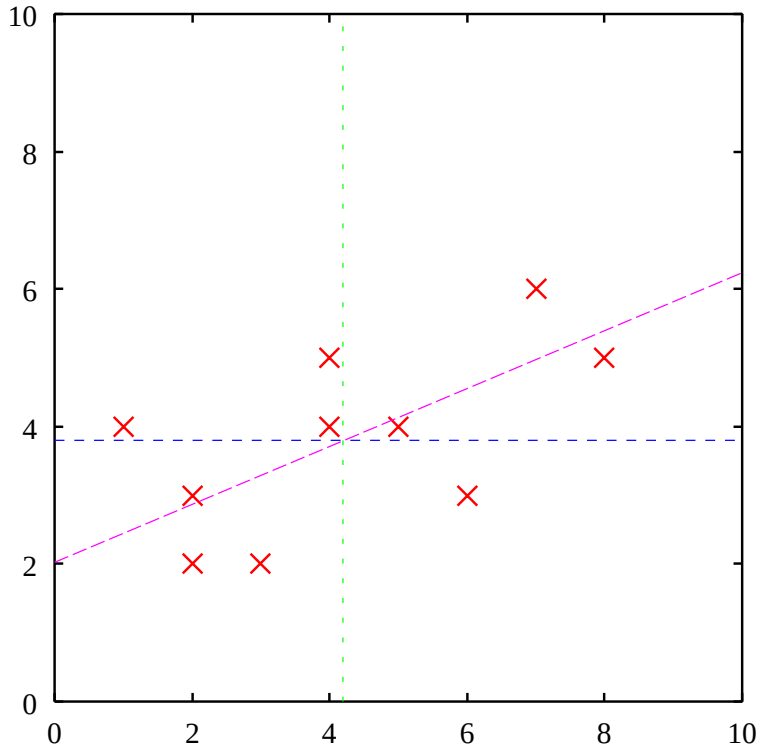
情報量最大化の例



- データ: $E = [2\ 1\ 2\ 3\ 5\ 4\ 8\ 6\ 7\ 4]^T, M = [3\ 4\ 2\ 2\ 4\ 4\ 5\ 3\ 6\ 5]^T$
- 平均: $\bar{E} = 4.2, \bar{M} = 3.8$
- 分散: $s_E^2 = 4.76, s_M^2 = 1.56$　共分散: $s_{EM} = 1.64$
- 対角化

$$
\begin{aligned}
S &= \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix} = \begin{bmatrix} 4.76 & 1.64 \\ 1.64 & 1.56 \end{bmatrix} \\
&= \begin{bmatrix} 0.9215 & -0.38838 \\ 0.38838 & 0.9215 \end{bmatrix} \begin{bmatrix} 5.4512 & 0 \\ 0 & 0.8688 \end{bmatrix} \begin{bmatrix} 0.9215 & 0.38838 \\ -0.38838 & 0.9215 \end{bmatrix} \\
&= V \Lambda V^T
\end{aligned}
$$

座標変換の例



- データ: $E = [2\ 1\ 2\ 3\ 5\ 4\ 8\ 6\ 7\ 4]^T, M = [3\ 4\ 2\ 2\ 4\ 4\ 5\ 3\ 6\ 5]^T$
- 平均からの差: $\tilde{E} = E - \bar{E}, \tilde{M} = M - \bar{M}$
- 座標変換: $[E'\ M'] = [\tilde{E}\ \tilde{M}]V$

科目	A	B	C	D	E	F	G	H	I	J
英語	2	1	2	3	5	4	8	6	7	4
数学	3	4	2	2	4	4	5	3	6	5
合計	5	5	4	5	9	8	13	9	13	9
単純順位	7	7	10	7	3	6	1	3	1	3
第 1 成分	-2.34	-2.87	-2.73	-1.81	0.82	-0.11	3.97	1.35	3.44	0.28
情報最大	8	10	9	7	4	6	1	3	2	5
第 2 成分	0.12	1.43	-0.80	-1.19	-0.13	0.26	-0.37	-1.44	0.94	1.18
文理度?	6	10	3	2	5	7	4	1	8	9

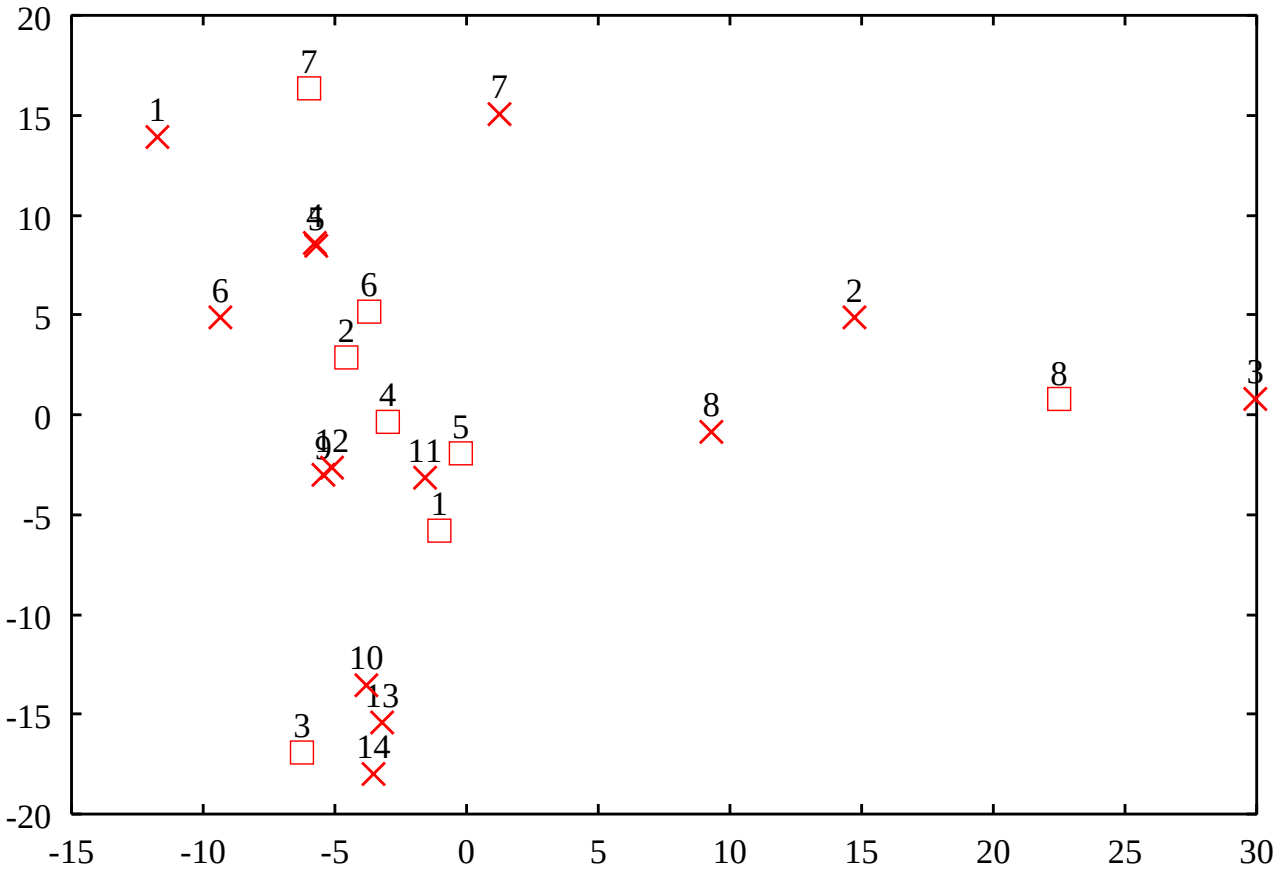
多次元情報の縮約

- 情報の要約と次元の縮小
  - 情報の要約: 相関関係の整理
  - 次元の縮小: 分散の小さなデータ (情報量が少ない) の棄却
- 共分散行列を用いて主成分分析 (PCA)

- 例:

メーカー \ イメージ	A	B	C	D	E	F	G	H	計
IBM	1	11	3	4	4	3	23	1	50
HP	3	3	1	2	6	1	9	24	49
ルーセント	1	3	0	1	3	1	0	38	47
東芝	3	6	3	8	0	12	13	6	51
NEC	2	7	4	3	4	11	14	6	51
日立	1	13	7	6	1	8	11	3	50
三菱	0	3	0	5	2	6	21	13	50
三洋	6	2	6	3	1	5	5	19	47
シャープ	12	12	5	6	6	0	4	4	49
エプソン	2	7	21	5	4	1	1	8	49
キャノン	1	4	10	10	6	4	5	9	49
ドコモ	1	1	12	4	7	15	5	6	51
J-Phone	14	3	18	3	4	1	0	7	50
AU	6	2	24	6	5	0	0	8	51
計	53	77	114	66	53	68	111	152	

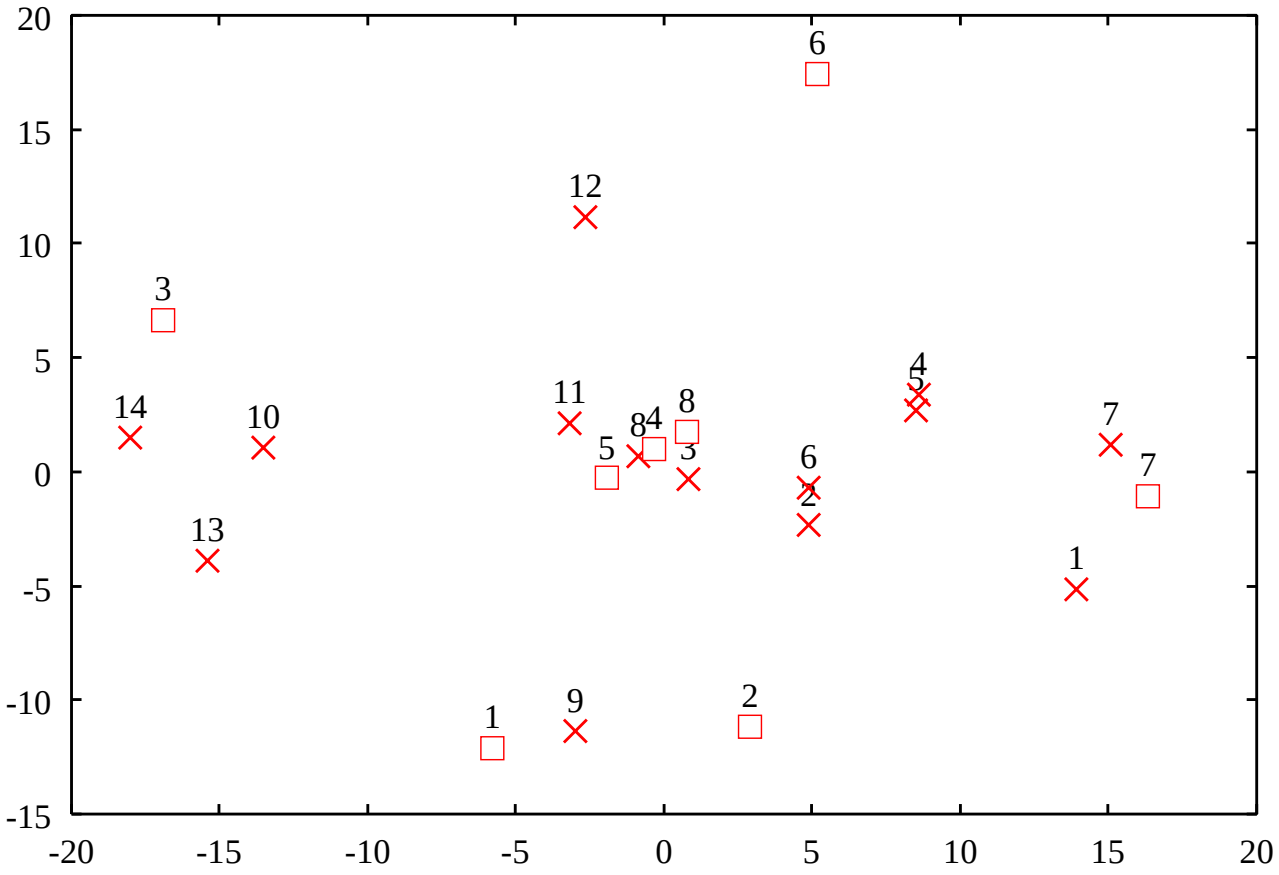
主成分分析1(知名度 vs 伝統)



1: IBM
2: HP
3: ルーセント
4: 東芝
5: NEC
6: 日立
7: 三菱
8: 三洋
9: シャープ
10: エプソン
11: キャノン
12: ドコモ
13: J-Phone
14: au

1: ユニークな製品がある
2: 研究開発に熱心
3: 宣伝広告に熱心
4: 製品 (サービス) の質がよい
5: 近代的である
6: 安定性がある
7: 伝統がある
8: よい印象がない (知らない)

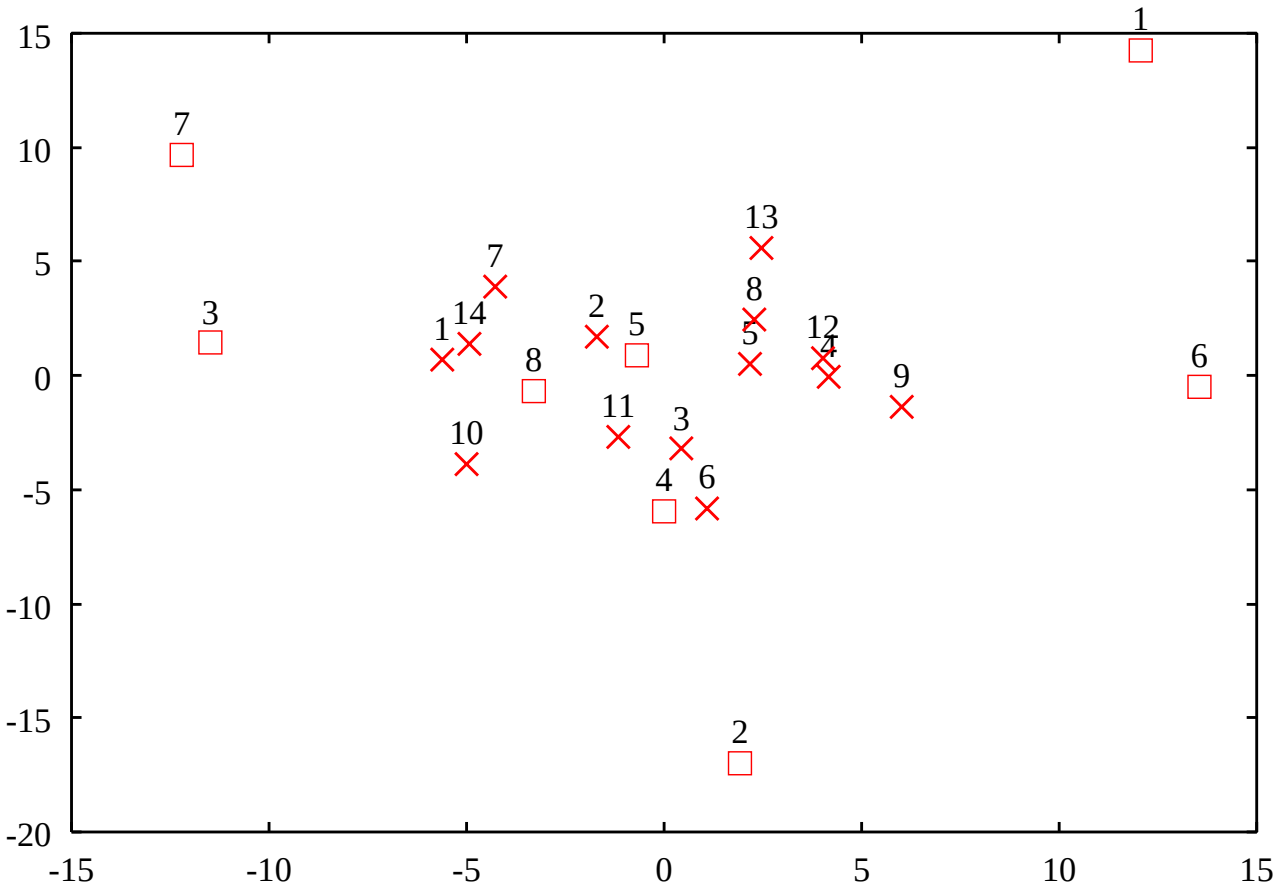
主成分分析2(伝統 vs 安定性)



1: IBM	6: 日立	11: キャノン
2: HP	7: 三菱	12: ドコモ
3: ルーセント	8: 三洋	13: J-Phone
4: 東芝	9: シャープ	14: au
5: NEC	10: エプソン	

1: ユニークな製品がある	5: 近代的である
2: 研究開発に熱心	6: 安定性がある
3: 宣伝広告に熱心	7: 伝統がある
4: 製品 (サービス) の質がよい	8: よい印象がない (知らない)

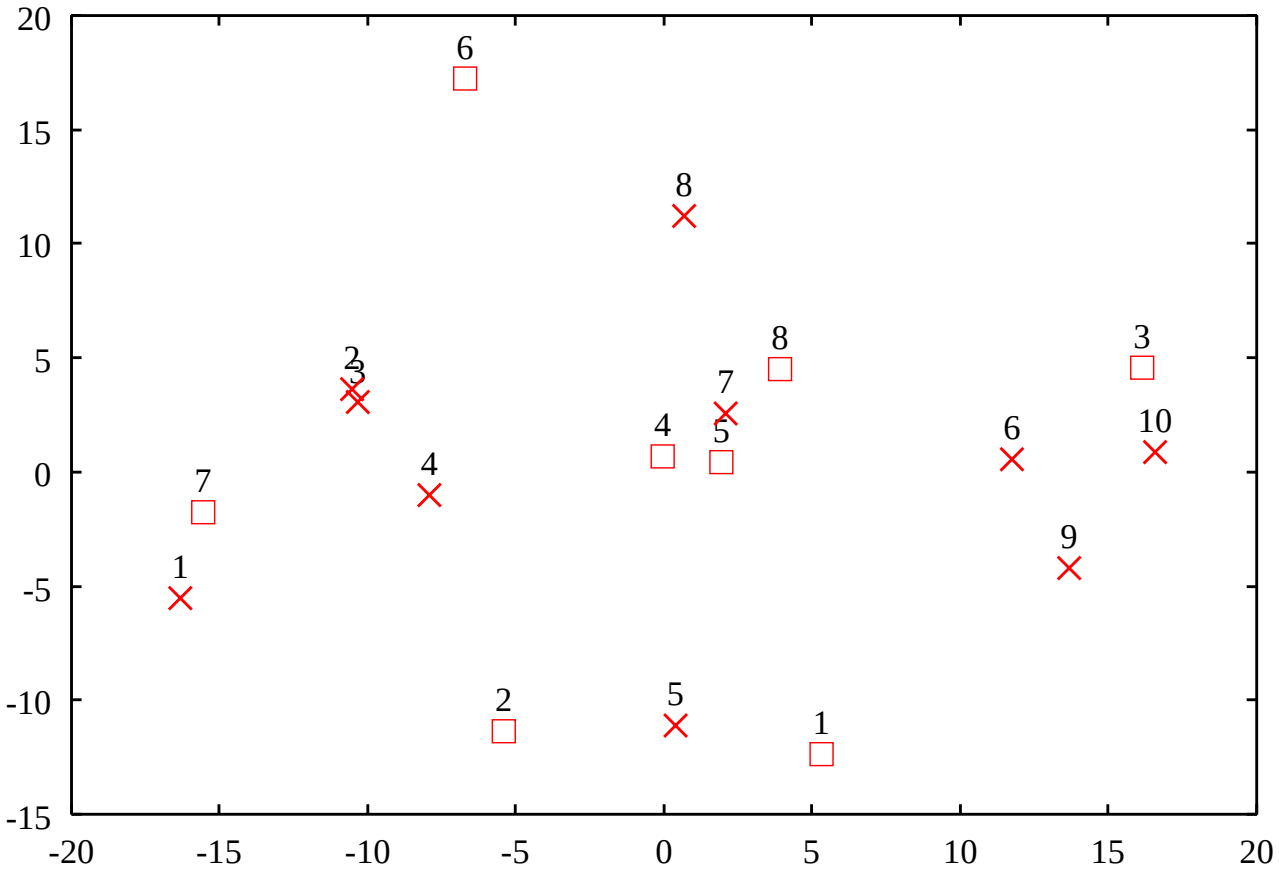
主成分分析3(安定性 vs 研究指向)



1: IBM
2: HP
3: ルーセント
4: 東芝
5: NEC
6: 日立
7: 三菱
8: 三洋
9: シャープ
10: エプソン
11: キャノン
12: ドコモ
13: J-Phone
14: au

1: ユニークな製品がある
2: 研究開発に熱心
3: 宣伝広告に熱心
4: 製品 (サービス) の質がよい
5: 近代的である
6: 安定性がある
7: 伝統がある
8: よい印象がない (知らない)

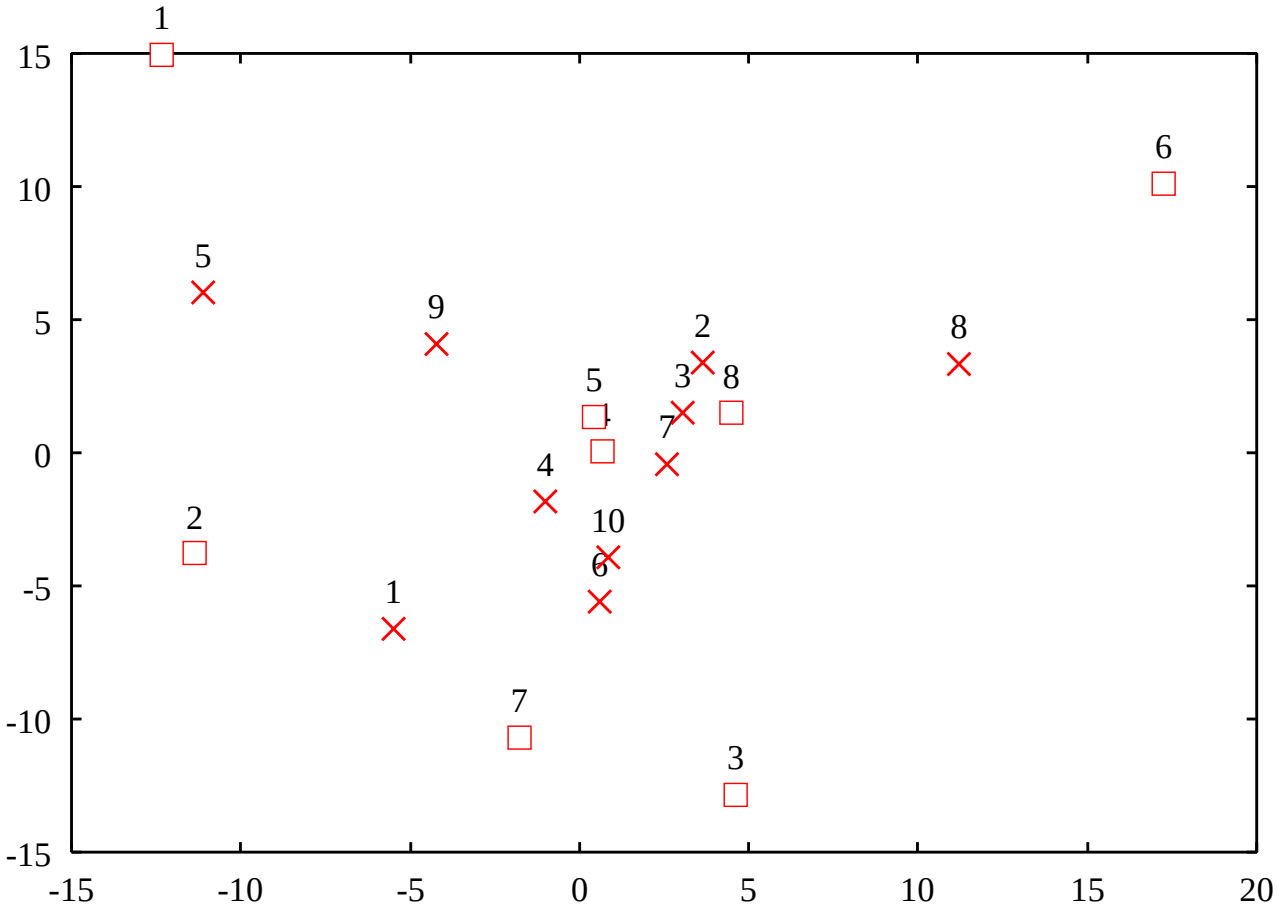
主成分分析 II-1 [知名度の低い 4 社を除く] (伝統 vs 安定性)



1: IBM	5: シャープ	9: J-Phone
2: 東芝	6: エプソン	10: au
3: NEC	7: キャノン	
4: 日立	8: ドコモ	

1: ユニークな製品がある	5: 近代的である
2: 研究開発に熱心	6: 安定性がある
3: 宣伝広告に熱心	7: 伝統がある
4: 製品 (サービス) の質がよい	8: よい印象がない (知らない)

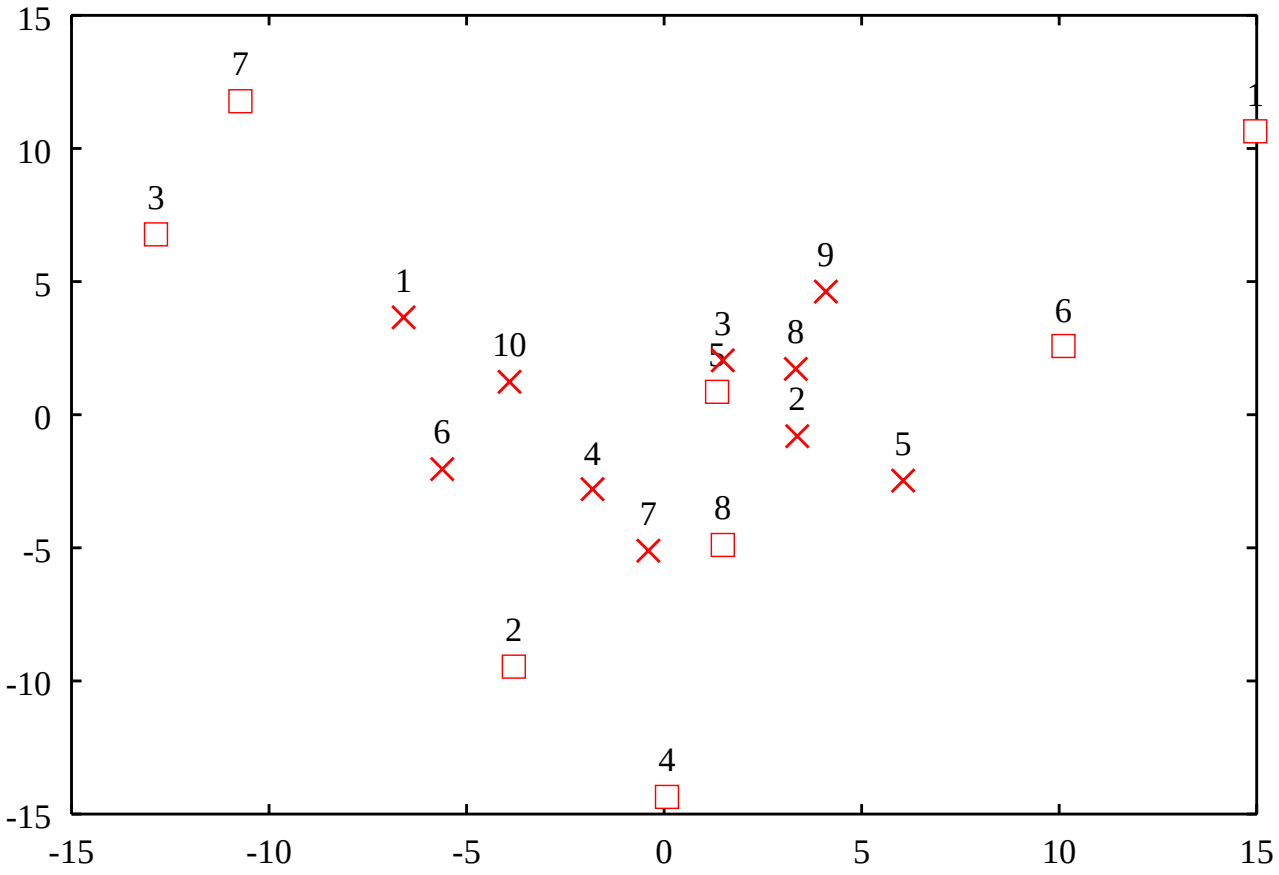
主成分分析II-2 [知名度の低い4社を除く](??)



1: IBM	5: シャープ	9: J-Phone
2: 東芝	6: エプソン	10: au
3: NEC	7: キャノン	
4: 日立	8: ドコモ	

1: ユニークな製品がある	5: 近代的である
2: 研究開発に熱心	6: 安定性がある
3: 宣伝広告に熱心	7: 伝統がある
4: 製品 (サービス) の質がよい	8: よい印象がない (知らない)

主成分分析 II-3 [知名度の低い 4 社を除く](??)



1: IBM	5: シャープ	9: J-Phone
2: 東芝	6: エプソン	10: au
3: NEC	7: キャノン	
4: 日立	8: ドコモ	

1: ユニークな製品がある	5: 近代的である
2: 研究開発に熱心	6: 安定性がある
3: 宣伝広告に熱心	7: 伝統がある
4: 製品 (サービス) の質がよい	8: よい印象がない (知らない)