2007年3月30-31日 行動計量学会　第10回春の合宿セミナー八王子セミナーハウス 2007-04-02 修正版

数量化と非線形多変量解析の方法

大学入試センター 研究開発部　大津起夫[1]

多変量データの次元縮約法、特に尺度推定を含む手法を概説.

- 主成分分析と特異値分解
- 対応分析（パターン分類の数量化）
- 尺度最適化を伴う主成分分析 (OSMOD)
  数量化３類（多重対応分析）の問題点を（部分的に）克服
- 非線形因子分析

[1] http://www.rd.dnc.ac.jp/~otsu

1 主成分分析法と特異値分解

主成分分析法: (principal cmponent analysis,PCA) は複数の連続変数間の相互関係の簡潔な表現を求めるために広く用いられている方法.
(**対応分析**: 2重分割表の構造を求める．共通点＝特異値分解)

分析対象となる変数: $X_j, (j = 1, .., p)$

観測されたデータの値: $x_{ij}, (i = 1, ..., n; j = 1, ..., p)$
(i: サンプル(被験者)の番号, j: 観測変数の番号)

いいたいこと

- 対応分析（パターン分析の数量化）＋Bertin グラフィックス
- MCA(数量化３類) より尺度最適化つき主成分分析 (OSMOD, PRINQUAL, ほか)

合成変量：

$$Z_m = \mu_m + b_{m1}X_1 + b_{m2}X_2 + \cdots + b_{mp}X_p,\ (m = 1, ..., q)$$

残差２乗和

$$SSQ_q = \sum_{j=1}^{p} E(|X_j - \mu_j - \sum_{m=1}^{q} a_{jm}Z_m|^2) \qquad (1)$$

$$E(X_j) = \mu_j + \sum_m a_{jm}E(Z_m)$$

(もし$X_j, (j = 1, ..., p)$の平均がゼロなら $\mu_j = 0$)

各変数X_jの平均がゼロの場合，観測されたデータの要素で表すと

$$SSQ_q = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \sum_{m=1}^{q} a_{jm}z_{im})^2$$

$X_1, ..., X_p$を要素とする縦ベクトルを　$\boldsymbol{x}$
（資料では大文字の太字になっているが，行列と間違えやすいので小文字にした）

a_{jm}を要素とする大きさ$p \times q$の行列を$\boldsymbol{A}$

$Z_1, ..., Z_q$を要素とする縦ベクトルを$\boldsymbol{z}$（資料は大文字）

$\boldsymbol{Az}$で $\boldsymbol{x}(p$次元$)$を近似したい．$\boldsymbol{z}$は$\boldsymbol{x}$の1次式.

$$\boldsymbol{z} = \boldsymbol{Bx} \text{ なので } \boldsymbol{Az} = \boldsymbol{ABx} = \boldsymbol{Cx}$$

（$m > 1$のときは$\boldsymbol{B}$は一通りには決まらない）
ただし$\mathrm{rank}\boldsymbol{C} = q$

上のSSQの式は行列記法を用いると次のように表される.

$$\begin{aligned} SSQ_q &= \mathrm{trace}E\{(\boldsymbol{x}-\boldsymbol{Az})(\boldsymbol{x}-\boldsymbol{Az})^T\} \\ &= \mathrm{trace}E\{(\boldsymbol{x}-\boldsymbol{Cx})(\boldsymbol{x}-\boldsymbol{Cx})^T\} \\ &= \mathrm{trace}(\boldsymbol{I}_p-\boldsymbol{C})E(\boldsymbol{xx}^T)(\boldsymbol{I}_p-\boldsymbol{C})^T \end{aligned}$$

記号trace（トレース）は，対角要素の和を表す記号.

確率変数のベクトル$\boldsymbol{x}$を，$\boldsymbol{Az}=\boldsymbol{Cx}$で平均的に近似する.

$$x_{ij} \sim \sum_{m=1}^{q} a_{jm}z_{ij} = \sum_{m=1}^{q} c_{jm}x_{im}$$

(x_{ij})の低ランク近似になっている.（qがpよりずっと小さいところがミソ）

主成分分析のモデルは理論的には単純 (観測変数の最小２乗近似を観測変数の１次式で行う).

因子分析 (factor analysis) はもっと凝った理論的背景を持っている．主成分分析より大胆な（攻撃的？）モデル推定を行う.

最小の SSQ_q をもたらす $Z_m, m = 1, .., q$ は，主成分スコアまたは単に主成分 (principal component) と呼ばれる.

主成分は変数の尺度に依存する.（因子分析の解は尺度に依存しない）

主成分の求め方: 変数の共分散行列$(p \times p)$: $E(\boldsymbol{x}\boldsymbol{x}^T) = \Big(E(X_j X_{j'})\Big) = \boldsymbol{\Sigma}_{XX}$

（1）**実対称行列の固有値分解**: $\boldsymbol{\Sigma}_{XX} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$

ここで$\boldsymbol{U}$は直交行列である(異なる列の内積,異なる行の内積がゼロ). また$\boldsymbol{\Lambda}$は固有値$(\lambda_1, ..., \lambda_p)$を成分とする対角行列である.

$\boldsymbol{\Lambda}$の対角要素は降順に並んでいるものとする.

$\boldsymbol{U}$の第1列から第q列の$p \times q$の行列を$\boldsymbol{U}_1$とする. $(\boldsymbol{U} = (\boldsymbol{U}_1|\boldsymbol{U}_2))$

$\boldsymbol{A} = \boldsymbol{B}^T = \boldsymbol{U}_1$とおくことにより$SSQ_q$を最小化できる.

固有値分解の数値例:

共分散行列

```
     [,1] [,2] [,3] [,4]
[1,]    4    3    2    1
[2,]    3    4    3    2
[3,]    2    3    4    3
[4,]    1    2    3    4
```

固有値分解の結果

```
  lambda_1  lambda_2   lambda_3   lambda_4
  11.099    3.414      0.901      0.586
```

```
       u_1    u_2    u_3    u_4
[1,] 0.448  0.653  0.547  0.271
[2,] 0.547  0.271 -0.448 -0.653
[3,] 0.547 -0.271 -0.448  0.653
[4,] 0.448 -0.653  0.547 -0.271
```

$$Z_1 = 0.448X_1 + 0.547X_2 + 0.547X_3 + 0.448X_4$$
$$Z_2 = 0.653X_1 + 0.271X_2 - 0.271X_3 - 0.653X_4$$

Z_1の分散は$\lambda_1 = 11.099$, Z_2の分散は$\lambda_2 = 3.414$.
これらの共分散はゼロ.

$\boldsymbol{C} = \boldsymbol{AB} = \boldsymbol{U}_1\boldsymbol{U}_1^T$($p \times p$の行列)は，上位$q$個の固有値に対応する固有ベクトルによって張られる変数の空間への直交射影子（対称で巾（べき）等な行列，つまり行列積で２乗しても元通りであること）になっている．

$Z_m = \boldsymbol{u}_m^T\boldsymbol{x}$を第$m$主成分と呼ぶ．($\boldsymbol{u}_m$は$\boldsymbol{U}$の第$m$列の縦ベクトルであり，$\boldsymbol{u}_m^T$はそれを転置した横ベクトルを表す．$\boldsymbol{u}_m^T\boldsymbol{x}$はこれら２つのベクトルの内積を表す.）($\boldsymbol{x}$は縦ベクトル)

$\Rightarrow \boldsymbol{U}_1^T$を$p$次元の縦ベクトル$\boldsymbol{x}$に左からかけることが情報を縮約する$(q$個の主成分$Z_1, ..., Z_q$にまとめる$)$ことになる．$(\boldsymbol{z} = \boldsymbol{U}_1^T\boldsymbol{x} \ \Rightarrow \ \boldsymbol{z}^T = \boldsymbol{x}^T\boldsymbol{U}_1\ )$

これに$\boldsymbol{U}_1(p \times q)$をさらに左からかけると，$q$個の主成分をもとの$p$次元のベクトルの世界（観測変数）に引き戻す．$(\hat{\boldsymbol{x}} = \boldsymbol{U}_1\boldsymbol{U}_1^T\boldsymbol{x} \ \Rightarrow \ \hat{\boldsymbol{x}}^T = \boldsymbol{x}^T\boldsymbol{U}_1\boldsymbol{U}_1^T\ )$

もうひとつの求め方　（データ行列を直接分解する）

(2)**特異値分解**(singular varlue decomposition)

一般的に$n \times p$の行列（正方行列とは限らない）$\boldsymbol{X}$は

$$\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T \tag{2}$$

と，直交行列（の一部）と対角行列の積に分解される．ここで$n \geq p$とする．

$\boldsymbol{V}$は$n \times p$の行列であり，その列ベクトルな各々長さ1であり互いに直交している．

$\boldsymbol{U}$は，$p \times p$の直交行列

$\boldsymbol{D}$は$p \times p$の非負要素の対角行列．

各変数の標本平均がゼロであると仮定すると，共分散行列は

$$\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$$

であり，上の特異値分解を用いると，$\boldsymbol{S} = \frac{1}{n}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T = \frac{1}{n}\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^T$ となる．

この$\boldsymbol{U}$の各列が$\boldsymbol{S}$の固有ベクトルであり，対角行列$\boldsymbol{D}$の要素の2乗が$\boldsymbol{S}$の固有値をn倍したものであることも分かる．

$\boldsymbol{U}$のうち，大きな特異値に対応する第1列から第q列を取り出すと，これが上の$\boldsymbol{U}_1$となっている．ここでq個の主成分は$n \times q$ の行列

$$\boldsymbol{X}\boldsymbol{U}_1 = (\boldsymbol{v}_1 \times d_1, ..., \boldsymbol{v}_q \times d_q)$$

で表され，また主成分を用いた各変数の近似は

$$\boldsymbol{X}\boldsymbol{U}_1\boldsymbol{U}_1^T$$

となる．

$\boldsymbol{V}$の列ベクトルを$\boldsymbol{v}_1, ..., \boldsymbol{v}_p$とし，また$\boldsymbol{U}$の列ベクトルを$\boldsymbol{u}_1, ..., \boldsymbol{u}_p$とすると，特異値分解の式は

$$\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T = \sum_m^p d_m \boldsymbol{v}_m \boldsymbol{u}_m^T \tag{3}$$

となる．上記の主成分による近似は,

$$\sum_m^q d_m \boldsymbol{v}_m \boldsymbol{u}_m^T$$

であり，行列の階数（ランク）qであるもののうち，$\boldsymbol{X}$との残差２乗和が最小になることが確認できる．

特異値分解の数値例：

$X(5 \times 3)$　（各列の平均はゼロ）

```
     x1 x2 x3
[1,] -2  2 -6
[2,] -1 -1 -3
[3,]  0 -2  0
[4,]  1 -1  3
[5,]  2  2  6
```

特異値分解の結果 $X = VDU^T$

(注意：RやSPlus の特異値分解では　$X = UDV^T$ となっている．ここでは固有値分解の結果と一致させるためにUとVを逆にしてある)

特異値　d_1　　d_2　　d_3

10.000　3.742　0.000

v_1　　v_2　　v_3　(d_k × v_k が主成分Z_k)

```
[1,] -0.632 -0.535 -0.561 ( v_3の値は無意味)
[2,] -0.316  0.267  0.102
[3,]  0.000  0.535 -0.510
[4,]  0.316  0.267 -0.612
[5,]  0.632 -0.535 -0.204
```

```
          u_1      u_2      u_3
[1,]   0.316        0    0.949
[2,]   0.000       -1    0.000
[3,]   0.949        0   -0.316
```

（u_3も無意味）

（u_m は共分散行列の固有ベクトルになる）

主成分分析の（因子分析と比べての）特徴

因子分析と比べて，理論的には洗練されていないが，単純で計算が速い．（確率モデルを想定したあてはめではない）．

変数の尺度が影響を及ぼす（かならずしも欠点ではないが）．

最近の速いPCなら2000次元の実対称固有値計算が30秒以内で可能．(Intel MKLやACML (AMD Core Math Library)を使うと速い．)

変数の平均がゼロでなくても,特異値分解による行列の近似は可能．

2　対応分析（ パターン分類の数量化）

2重分割表の構造を把握したい.

χ^2検定で，独立性の仮説が棄却されても，それだけではあまり意味がない.

行と列に，データの構造を反映した尺度を与えたい.

表 1: NLSY79: 1990年における職業と年齢補正済AFQT
（クロスセクショナルサンプル） (大津, 2003)

	年齢補正済AFQT					
職業	1	2	3	4	5	計
事務 (Clerical)	115	194	219	187	132	847
製造 (Craft)	132	129	128	105	56	550
農業 (Farm)	23	10	16	8	7	64
労働 (Labor)	89	65	40	22	13	229
管理職 (Mgr)	56	90	135	164	152	597
熟練工 (Operatv)	149	97	89	59	27	421
専門技術職 (Prof/Tech)	31	80	115	224	428	878
営業 (Sales)	14	37	37	78	74	240
サービス (Service)	197	166	126	99	54	642
運輸 (Transpt)	44	41	27	28	7	147
計	850	909	932	974	950	4615

2.1 対応分析のモデル

表1のような2重分割表のi行j列の数値をn_{ij} とする．行数をIとし，列数をJとする．また，対象（被験者）の総数を

$$N = \sum_i \sum_j n_{ij}$$

とする．

$$P_{ij} = n_{ij}/N,\ i = 1, ..., I;\ j = 1, ..., J$$

は相対頻度を表す．また，周辺相対頻度を

$$P_{i.} = \sum_{j=1}^{J} P_{ij},\ P_{.j} = \sum_{i=1}^{I} P_{ij}$$

と表す．

行スコアを$x_i, i = 1, ..., I,$列スコアを$y_j, j = 1, ..., J$

i行j列のセルの座標(x_i, y_j)

この2次元上の位置にP_{ij}の確率が与えられているとすると，これによって定まる分布の平均は次のように表される．

$$\mu_x = \sum_i P_{i.} x_i,\ \mu_y = \sum_j P_{.j} y_j \tag{4}$$

以下では，μ_xとμ_yとが，ともにゼロベクトルに制約する．

この仮定のもとで，行スコアxと列スコアyの分散は，それぞれ

$$\sigma_x^2 = \sum_i P_{i.} x_i^2,\ \sigma_y^2 = \sum_j P_{.j} y_j^2 \tag{5}$$

となる．以下ではさらに$\sigma_x^2 = \sigma_y^2 = 1$の制約も，スコアが満たしているとする．

このときxとyとの相関は

$$r_{xy} = \sum_i \sum_j P_{ij} x_i y_j \tag{6}$$

となる．この値は，行と列との関係を表す一つの指標と考えられる．相関rが大きければ，行と列との間に強い関係があるとみなせる．

r_{xy}を最大にする行スコア，列スコアを求めたい.

これらのスコアは，データに内在する構造をよく表しているだろう（たぶん）.

周辺相対頻度を要素とする対角行列（対角線上の要素のみが非ゼロの行列）を定める.

$$\boldsymbol{F} = \mathrm{diag}(P_{1.}, ..., P_{I.}),\ \boldsymbol{G} = \mathrm{diag}(P_{.1}, ..., P_{.J})$$

これらの要素の平方根を要素とする対角行列を，それぞれ

$$\boldsymbol{F}^{1/2} = \mathrm{diag}(\sqrt{P_{1.}}, ..., \sqrt{P_{I.}}),\ \boldsymbol{G}^{1/2} = \mathrm{diag}(\sqrt{P_{.1}}, ..., \sqrt{P_{.J}})$$

とする.

相関(6)を制約のもとで最大化するスコアは，次の特異値分解を利用して得ることができる．

$$\boldsymbol{F}^{-1/2}\boldsymbol{P}\boldsymbol{G}^{-1/2} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T} \tag{7}$$

$\boldsymbol{D}$の対角要素はすべて非負であり，降順にならんでいるものとする．

変則的だが$\boldsymbol{U}$と$\boldsymbol{V}$の列ベクトルを，それぞれ添え字ゼロから始まるものとし，

$$\boldsymbol{u}_k,\ \boldsymbol{v}_k,\ (k = 0, 1, ..., K)$$

とする．また$K = \min(I - 1, J - 1)$である．

対角行列$\boldsymbol{D}$の要素を$d_0, d_1, ..., d_K$とする．

$$\boldsymbol{F}^{-1/2}\boldsymbol{u}_k = \boldsymbol{x}_k,\ \boldsymbol{G}^{-1/2}\boldsymbol{v}_k = \boldsymbol{y}_k$$

とおくと，$\boldsymbol{x}_1$と$\boldsymbol{y}_1$とが(6)を最大にし，制約を満たすスコアになり，$d_1 = r_{max}$となる．

また，$\boldsymbol{x}_0$と$\boldsymbol{y}_0$とは，要素が全て1のベクトルとなり，$d_0 = 1$となる．添え字ゼロに対応するスコアは，重みつき平均がゼロの制約を満たさないため，対応分析の解とはならない．

特異値分解の式 (7) の左から $\boldsymbol{F}^{1/2}$ を乗じ，また右から $\boldsymbol{G}^{1/2}$ を乗ずると

$$\begin{aligned}\boldsymbol{P} &= \boldsymbol{F}^{1/2}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{G}^{1/2} \\ &= \boldsymbol{F}\boldsymbol{F}^{-1/2}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{G}^{-1/2}\boldsymbol{G} \\ &= \boldsymbol{F}\left(\sum_{k=0}^{K} d_k\boldsymbol{x}_k\boldsymbol{y}_k^T\right)\boldsymbol{G}\end{aligned}$$

となる.

対応分析の解を考えることは，上の式で$k=0$と$k=1$の部分によって2重分割表$\boldsymbol{P}$の重みつきの近似を行っているともみなせる．

$$\boldsymbol{P} \sim \boldsymbol{F}\left(d_0\boldsymbol{x}_0\boldsymbol{y}_0^T + d_1\boldsymbol{x}_1\boldsymbol{y}_1^T\right)\boldsymbol{G}$$

また，対応分析の利用においては1次元より多くの解，つまり$k=2,3,...$に対応する$\boldsymbol{x}_k$および$\boldsymbol{y}_k$を解釈の対象とすることがある．

これらの$k=2,..$に対応する解は，１次元ほど多くの部分を説明しないが，データのより詳細な部分についての情報を与える場合がある（そうでない場合もある）．

$\boldsymbol{U}$が直交行列であることから,

$$\boldsymbol{x}_k^T \boldsymbol{F} \boldsymbol{x}_l = 0, \ (k \neq l)$$

であり，$\boldsymbol{V}$が直交行列であることから,

$$\boldsymbol{y}_k^T \boldsymbol{G} \boldsymbol{y}_l = 0, \ (k \neq l)$$

が成立する.

スコア$\boldsymbol{x}_2$と$\boldsymbol{y}_2$とは，平均と分散の制約に加え，上の直交条件を満たすもののうち，(6)を最大にするものである.

2.2 対応分析の利用例

表1のデータについて，具体的な数値例を示す.
各行の和と平方根は，表に示されているように

行和	847	550	64	229	597	421	878	240	642	147
平方根	29.10	23.45	8.00	15.13	24.43	20.52	29.63	15.49	25.34	12.12

となる.

一方，各列の和と平方根は

列和	850	909	932	974	950
平方根	29.15	30.15	30.53	31.21	30.82

となる.

$\boldsymbol{Q} = \boldsymbol{F}^{-1/2}\boldsymbol{P}\boldsymbol{G}^{-1/2}$ (各要素は$q_{ij} = p_{ij}/\sqrt{f_i g_j}$である.)

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.1355	0.2211	0.2465	0.2059	0.1472
[2,]	0.1931	0.1824	0.1788	0.1435	0.0775
[3,]	0.0986	0.0415	0.0655	0.0320	0.0284
[4,]	0.2017	0.1425	0.0866	0.0466	0.0279
[5,]	0.0786	0.1222	0.1810	0.2151	0.2018
[6,]	0.2491	0.1568	0.1421	0.0921	0.0427
[7,]	0.0359	0.0895	0.1271	0.2422	0.4686
[8,]	0.0310	0.0792	0.0782	0.1613	0.1550
[9,]	0.2667	0.2173	0.1629	0.1252	0.0691
[10,]	0.1245	0.1122	0.0729	0.0740	0.0187

行列の特異値分解$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$の，対角行列$\boldsymbol{D}$の成分は，

$$1.0,\ 0.4625,\ 0.1457,\ 0.0514,\ 0.0456$$

2番以降が対応分析によって得られる行と列の相関係数になる．第１次元の行と列の相関は$r_{xy} = 0.4625$

また, $\boldsymbol{U}$ の第１列(制約を満たさない部分) および第２, ３列は, 次のようになる.

```
          [,1]     [,2]     [,3]
 [1,] 0.4284   0.0421 -0.5933
 [2,] 0.3452   0.2109 -0.1372
 [3,] 0.1178   0.1059  0.1473
 [4,] 0.2228   0.2978  0.3540
 [5,] 0.3597 -0.1980 -0.3481
 [6,] 0.3020   0.3344  0.2738
 [7,] 0.4362 -0.7131  0.4678
 [8,] 0.2280 -0.2115 -0.1556
 [9,] 0.3730   0.3428  0.2068
[10,] 0.1785   0.1794  0.0119
```

これに左から$\boldsymbol{F}^{-1/2}$（相対頻度から求めたもの）をかけた$\boldsymbol{F}^{-1/2}\boldsymbol{U}$の第１～３列は次のようになる.

```
      [,1]    [,2]    [,3]
 [1,]    1  0.0982 -1.3850
 [2,]    1  0.6108 -0.3973
 [3,]    1  0.8994  1.2510
 [4,]    1  1.3369  1.5892
 [5,]    1 -0.5506 -0.9677
 [6,]    1  1.1070  0.9065
 [7,]    1 -1.6348  1.0725
 [8,]    1 -0.9272 -0.6825
 [9,]    1  0.9190  0.5545
[10,]    1  1.0050  0.0667
```

一方$\boldsymbol{V}$の第１列と第２列は，次になる．

```
        [,1]    [,2]    [,3]
[1,] 0.4292  0.5757  0.6072
[2,] 0.4438  0.2958 -0.1607
[3,] 0.4494  0.1172 -0.4990
[4,] 0.4594 -0.2298 -0.3793
[5,] 0.4537 -0.7173  0.4611
```

同様に，これに$\boldsymbol{G}^{-1/2}$を左からかけると

```
     [,1]     [,2]     [,3]
[1,]    1   1.3414   1.4149
[2,]    1   0.6666 -0.3621
[3,]    1   0.2607 -1.1103
[4,]    1 -0.5001 -0.8257
[5,]    1 -1.5810   1.0163
```

となる．この第２列と第３列とが，$\boldsymbol{y}_1$および$\boldsymbol{y}_2$である．

列についてのスコアは（当然予想されることだが），AFQTの層別の順を反映したものになっている．

表1の行の順を$\boldsymbol{x}_1$の大きさにしたがって昇順に置き換えると，次のような表が得られる．

表 2: 対応分析のスコアによって行を並べ替えた表

	AFQT1	AFQT2	AFQT3	AFQT4	AFQT5
Prof/Tech	31	80	115	224	428
Sales	14	37	37	78	74
Mgr	56	90	135	164	152
Clerical	115	194	219	187	132
Craft	132	129	128	105	56
Farm	23	10	16	8	7
Service	197	166	126	99	54
Transpt	44	41	27	28	7
Operatv	149	97	89	59	27
Labor	89	65	40	22	13

専門技術職(Prof/Tech)ではAFQTの高成績群が相対的に多く，一方，熟練工(Operatv)や労働(Labor) （職種の訳語は大津によるので，公的なものではない.）では，低成績群が多い．中ほどの事務職(Clerical)では，AFQT中成績群が多く，対応分析によるスコアは，すくなくとも表の見かけ上の構造を簡潔に把握する手助けとなることがわかる.

図1は，横軸に１次元目のスコア，縦軸に２次元目のスコアをとり，$(d_1 \times \boldsymbol{x}_1, d_2 \times \boldsymbol{x}_2)$および$(d_1 \times \boldsymbol{y}_1, d_2 \times \boldsymbol{y}_2)$を，同一のグラフ上にプロットしたものである．(文献やソフトウェアによってはd_kで重み付けないものもある．ここでのプロットはRのMASSライブラリ (Venables & Ripley,1999) を用いている．

対応分析（パターン分類の数量化）

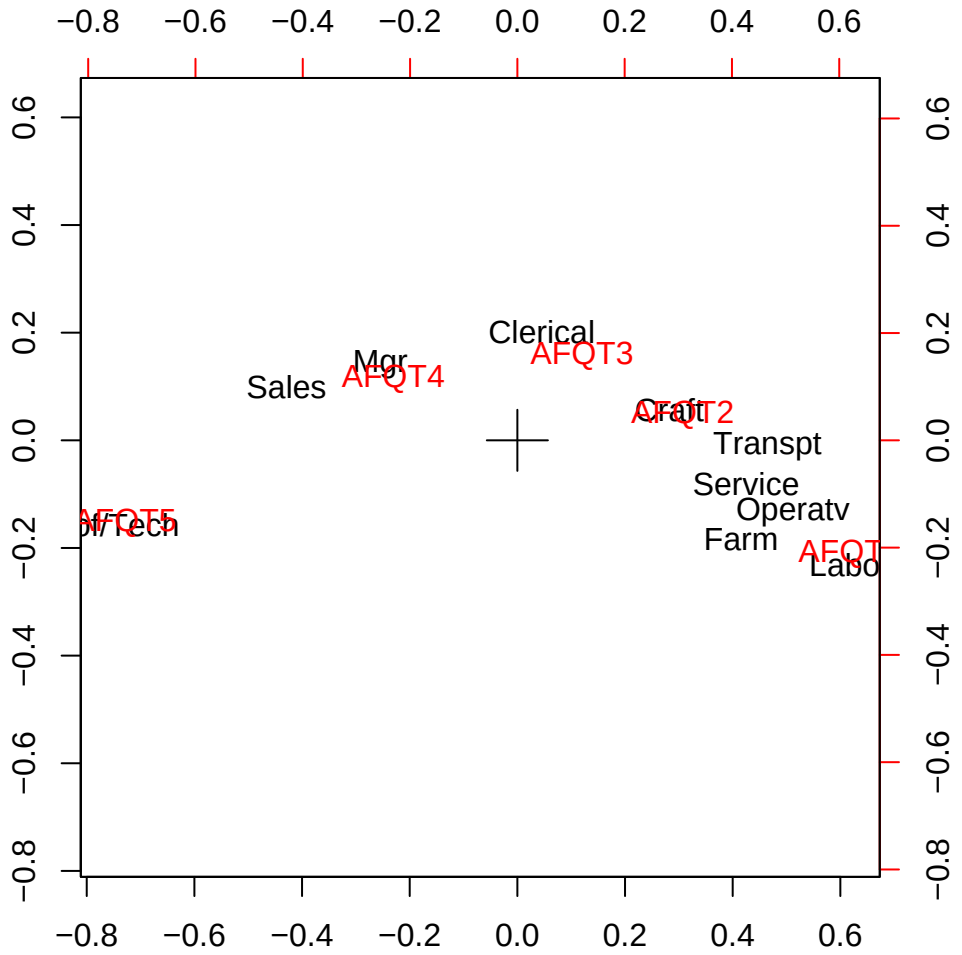

図 1: 対応分析の解，横軸：第１次元，縦軸：第２次元

Jaques Bertinの重み付き置換マトリックス．

図2は，表1の表示である．
各行の幅は，対応する周辺分布の大きさに比例．第i行の幅（ここでは縦方向の幅）は，p_{i+}の大きさに比例している．

各セルに対応する棒グラフの長さ（ここでは横方向の長さ）は，各セルの値を周辺度数で割った値p_{ij}/p_{i+}を表している．小さな長方形の面積が，セルの値を表す．

各列の破線は，p_{+j}に相当する大きさ（行と列が独立である場合の期待値）を表す．

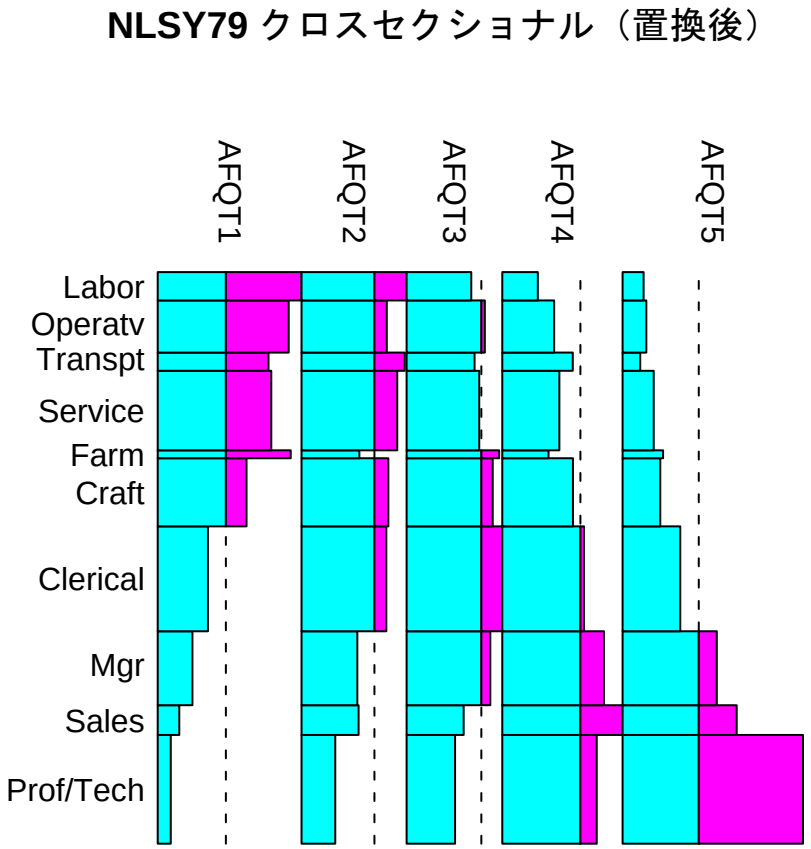

図 2: 重み付き置換マトリックス（第１次元スコアによる並べ替え）

2.3 馬蹄形問題

対応分析は，計算が単純で便利な方法であるが，実用上いくつか注意すべき点がある．ひとつは，カテゴリ間の関係を離散変数の間の相関によって捉えるため，周辺度数が偏っている場合（特にカテゴリーの水準の数が小さい場合）には，変数間の関係を適切に表現できない場合がありうることである．

さらに，多次元の解を利用する場合には，次のような特徴（馬蹄形問題）にも注意する必要がある．

表3に示す共分散を持つ4変量正規分布を想定する．σ_{13}, σ_{24} の値は，表4 に示す2つのケースを考える．次にこの分布に基づく乱数を生成し，次のような手順で2重分割表を作成する．

1. 表3に示す共分散行列に従う4変量正規乱数を5000個生成する.
2. 各変数の値を大きさの順に並べ替え，大きさの順に5つのカテゴリーに分類する．この際，カテゴリー区分は各々の頻度が等しくなるようにする.
3. 次にX_1とX_2を離散化して得られたカテゴリーのクロスカテゴリーを求める．同様にしてX_3とX_4からクロスカテゴリーを求める．各々25のカテゴリを持つ2つの変数XとYが得られる(図3).
4. 離散化されたXとYから25×25の2重分割表を求める.

表 3: 多変量正規分布の共分散行列

	X_1	X_2	X_3	X_4
X_1	1.0			
X_2	0.0	1.0		
X_3	σ_{31}	0.0	1.0	
X_4	0.0	σ_{42}	0.0	1.0

表 4: テストデータの共分散の値

	σ_{13}	σ_{24}
Case 1	0.8	0.5
Case 2	0.8	0.4

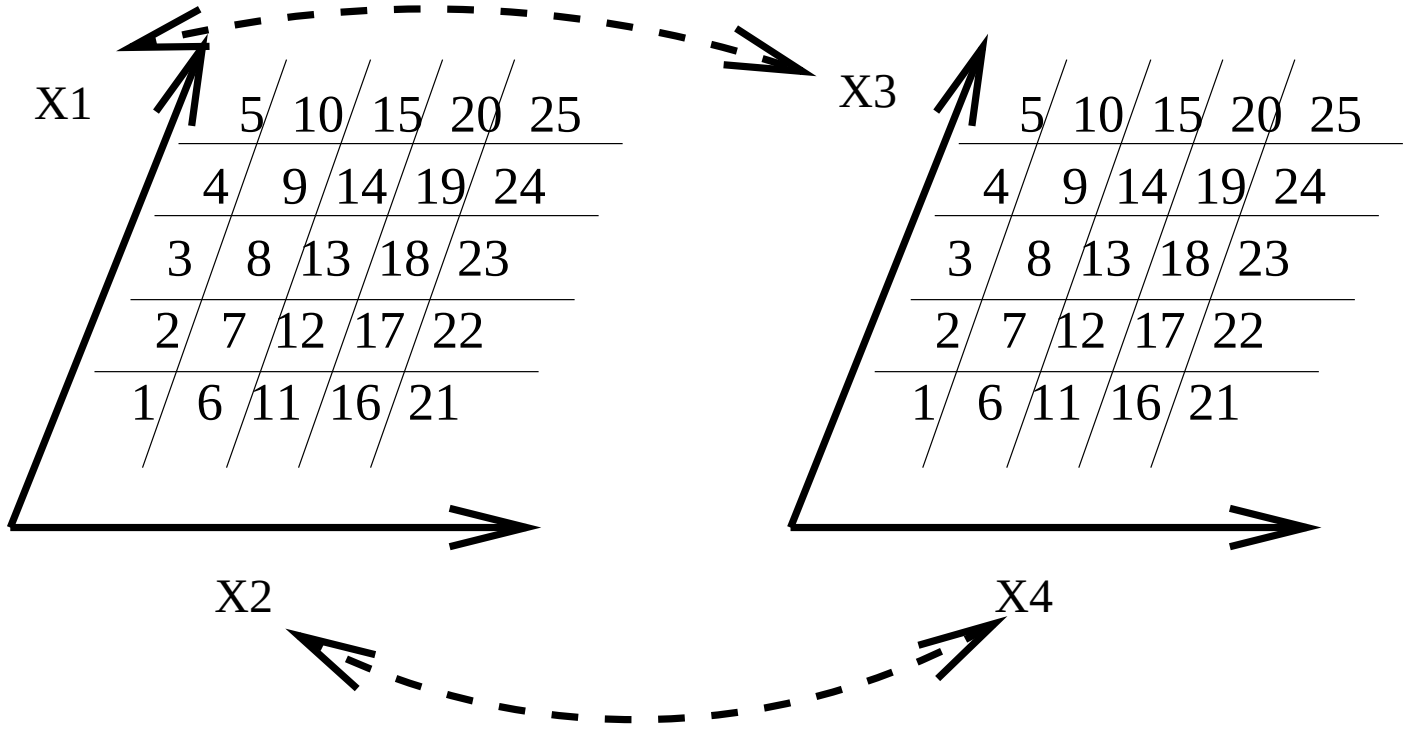
X1
5 10 15 20 25
4 9 14 19 24
3 8 13 18 23
2 7 12 17 22
1 6 11 16 21
X3
5 10 15 20 25
4 9 14 19 24
3 8 13 18 23
2 7 12 17 22
1 6 11 16 21
X2
X4

図 3: 2重分割表の作成

表 5: 対応分析によって推定された相関係数(2重分割表)

	r_1	r_2	r_3	r_4
Case 1	0.75	0.46	0.42	0.36
Case 2	0.74	0.41	0.36	0.25

この手続きによって生成された2重分割表に対応分析を適用する．表5は，対応分析を適用して得られた相関係数(特異値) を大きさの順に上位4個まで示したものである．ここでは自明な解$(r_0 = 1)$は除いてある．

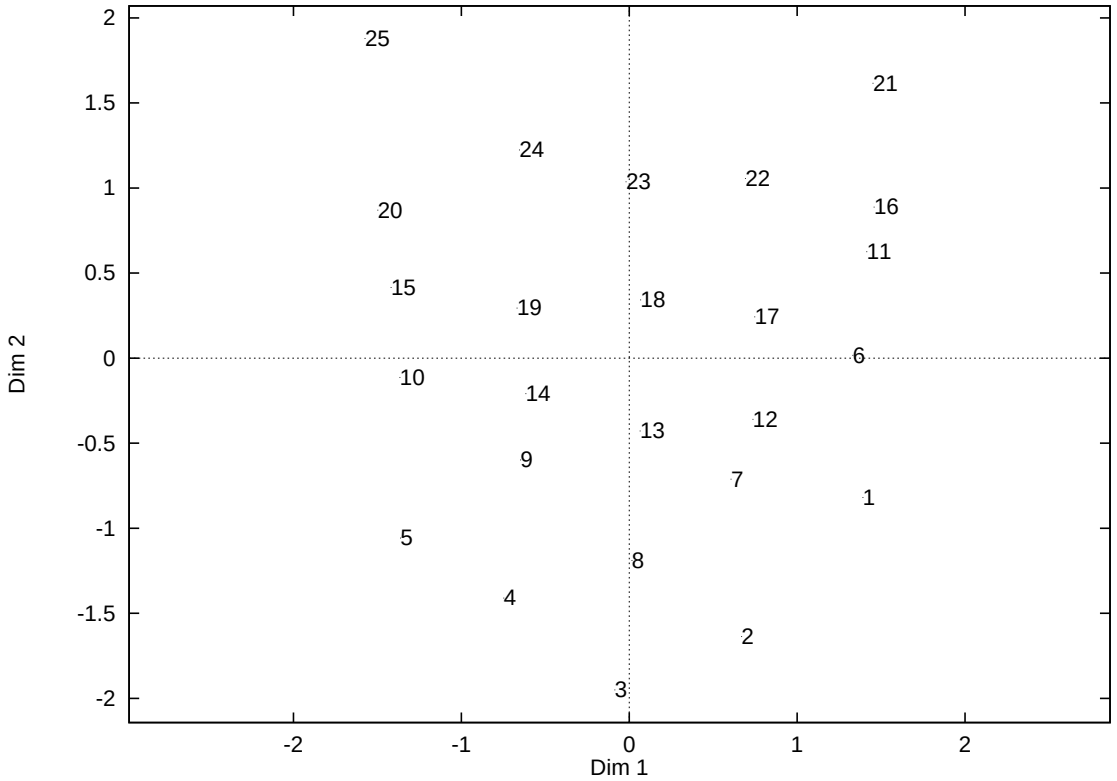

図 4: 対応分析によるスコア (Case1, 第 1 次元と第 2 次元)

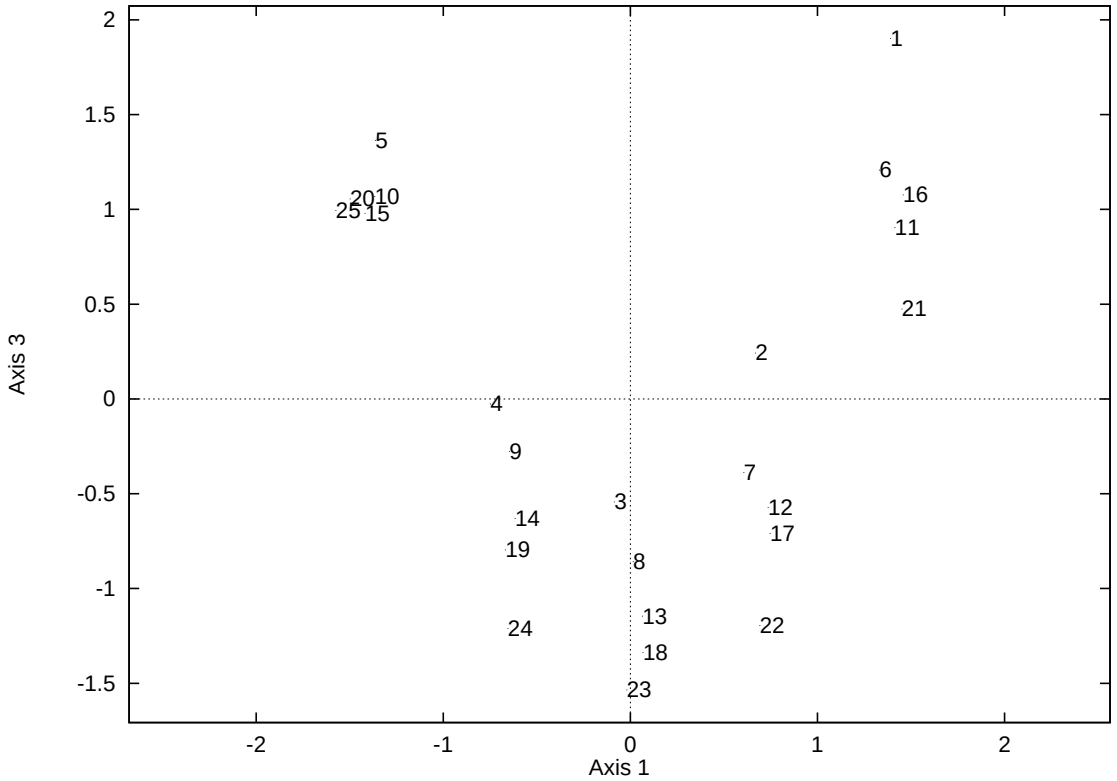

図 5: 対応分析によるスコア (Case1, 第 1 次元と第 3 次元)

Case1の計算結果を示す図4 においては，データを生成するために設定した格子状の構造が第1次元と第2次元によって再現されている．

Case 1で得られた相関係数は第1次元については0.75であり，また第2次元については0.46となっている．表4に示した正規分布の相関係数よりは幾分小さめではあるが，カテゴリーの背後に設定した分布の構造を推定することに成功している．

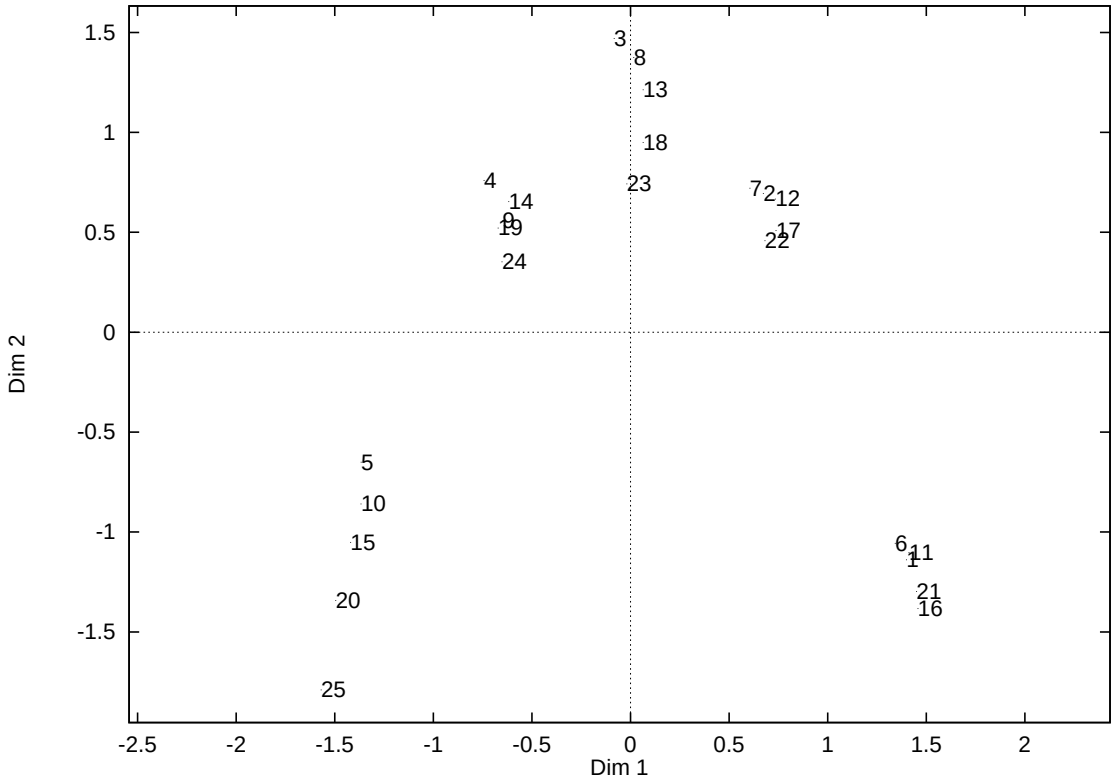

図 6: 対応分析によるスコア (Case2, 第 1 次元と第 2 次元)

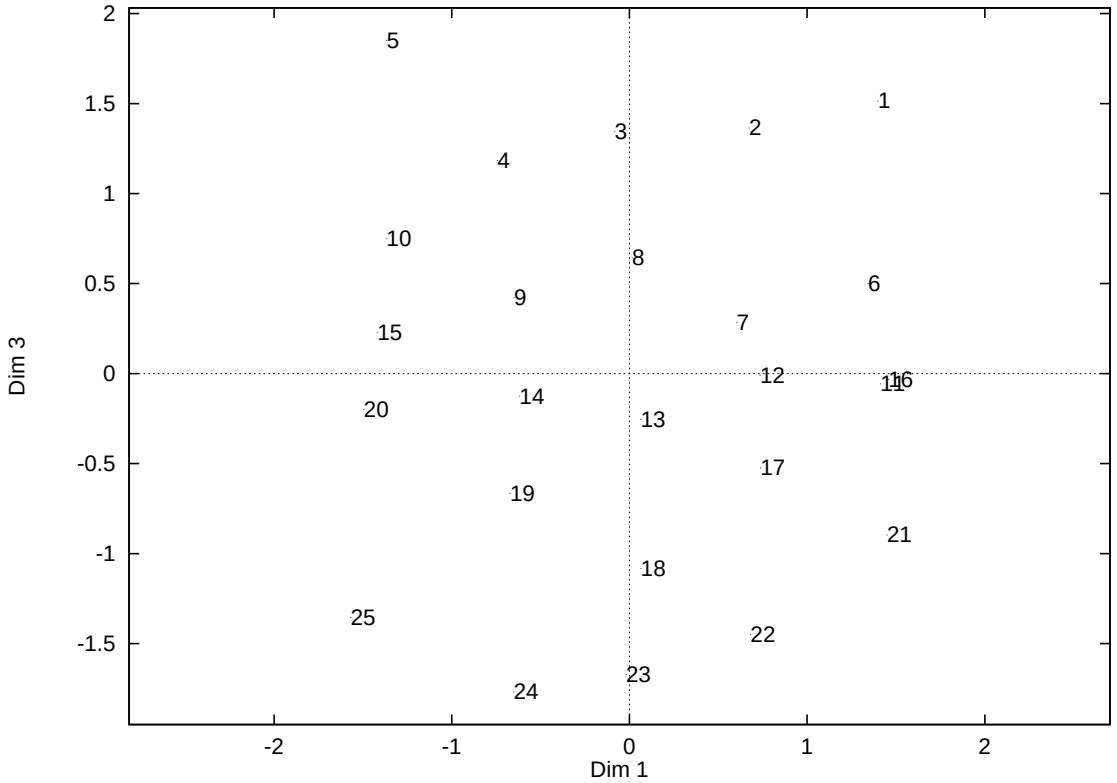

図 7: 対応分析によるスコア (Case2, 第 1 次元と第 3 次元)

一方，図6はCase 2におけるYのカテゴリースコアである．第2次元のスコアは第1次元のスコアの2次関数になっている．ここには示さなかったが，格子状の構造は第1次元と第3次元のスコアを合わせ見ることによって初めて確認できる．

Case1とCase2の違いはX_2とX_4の相関が各々0.5と0.4であることのみであり，わずかな相関構造の違いが大きな結果の違いをもたらしている．

理論的な検討によれば，潜在的な構造が多変量正規分布であり，データ十分多くあり，カテゴリーの分割が小さければ，最大の相関（特異値）がr_1であるとき，データの実質的な内容に関わらず，ほぼr_1^2の相関を持つ解が現れる．

より小さな相関を持つ微妙な成分である場合には，１次元目の２次関数成分が解として得られ，データの内容を表す成分を発見することができない．

ただし，実際の計算においては，カテゴリー数がそれほど多くはないため，２次関数やより高次の関数成分を表す解の相関は，r_1^2やr_1^3よりも，かなり小さくなる．

3 尺度最適化を伴う主成分分析

対応分析：２つの離散変数の関係から最適尺度を求める

$\Rightarrow r_{XY}$ の最大化

$\Rightarrow$ $\mathrm{var}(X+Y)$（主成分の分散）の最大化

多重対応分析(MCA,数量化３類): 多肢選択の多変量データを対象とする.

X_jは$1, ..., k_j$の離散値をとる.

$\Rightarrow k_j$個の$0-1$値をとるダミー変量を対応させる.

$n \times (\Sigma_j\, k_j)$の$0-1$値のテーブルに対応分析を適用.

$\Rightarrow$　サンプル(被験者）のスコアとアイテムの選択肢のスコアが得られる.

$\Rightarrow$ 第１次元は$X_1, X_2, ..., X_p$の尺度最適化による主成分

著名な利用例：統計数理研究所「日本人の国民性調査」
ピエール・ブルデュー「ディスタンクシオン」

対応分析と類似の馬蹄形問題が生じる．選択肢のスコアはしばしば不安定．特に第２次元目以降の解に問題がありがち．

- 多重対応分析（数量化３類）の不安定性を回避したい
- 離散値と連続値の混在するデータを分析したい
- 離散値の順序関係も考慮したい

尺度最適化を伴う主成分分析
(SAS:PRINQUAL, Gifi(Heiser et al.) HOMALS, OSMOD)
離散変数については最適スコアを与えつつ主成分分析モデルを当てはめる．

3.1 データとモデル

分析対象となるデータのサンプル件数（被験者数）をNとし，変数（調査項目）の個数をJとする．第j変数が間隔尺度を持つ数値データである場合には，第i番目のサンプル（被験者）の第j変数の値を$v_i(j)$と表す．また，第j変数が順序尺度または名義尺度のカテゴリー値をとる場合には，それらのカテゴリーを

$$\{C_{j1}, C_{j2}, \ldots, C_{jk_j}\}$$

とする．特に，第j変数が順序尺度である場合には

$$\{C_{j1} \prec C_{j2} \prec \cdots \prec C_{jk_j}\} \tag{8}$$

なる順序関係が定められているものとする．

第$1 \sim J_1$変数は間隔尺度の数値変数

第$J_1+1 \sim J_1+J_2$変数は順序尺度のカテゴリー値を持つ変数

第$J_1+J_2+1 \sim J_1+J_2+J_3$変数は名義尺度のカテゴリー値を持つ変数

ここで，$J_1+J_2+J_3=J$ とする．

また，カテゴリー値をとる第j変数における第iサンプルの反応を次のように表記する．

$$\delta_i(jk) = \begin{cases} 1 & (\text{第}i\text{サンプルが第}k\text{カテゴリーに反応}) \\ 0 & (\text{それ以外}) \end{cases}$$

以降の表記を簡単にするため，反応頻度について次のような記号を導入する．式中，$\bar{x}$などのように横棒がついているものは，N件のサンプルについての平均を表す．

$$n_{jk} = \sum_{i=1}^{N} \delta_i(jk) = N \times \overline{\delta(jk)}$$

$$n_{jk,lm} = \sum_{i=1}^{N} \delta_i(jk)\delta_i(lm) = N \times \overline{\delta(jk)\delta(lm)}$$

ここで考慮の対象としているカテゴリカルデータは択一式のものである．従って次の等式が成立する．

$$\sum_{k=1}^{k_j} \delta_i(jk) = 1, \quad (i = 1, ..., N; j = J_1 + 1, ..., J)$$

ある変数のカテゴリーが順序尺度に基づくものである場合には，カテゴリースコアがその順序制約に従うことにする．

カテゴリースコアのベクトルを

$$\boldsymbol{w} = (\boldsymbol{w}'_{J_1+1}, \boldsymbol{w}'_{J_1+2}, \ldots, \boldsymbol{w}'_{J})'$$

とする．ここで，$\boldsymbol{w}_j = (w_{j1}, \ldots, w_{jk_j})'$であり，第$j$変数が順序尺度である場合には$(1)$の順序に基づいて，

$$w_{j1} \leq w_{j2} \leq \cdots \leq w_{jk_j}, \quad (j = J_1 + 1, \ldots J_1 + J_2) \quad (9)$$

という制約をスコアに課す.

何らかの基準に基づいて(J_2+J_3)個のカテゴリー変数に対応するスコアベクトルが得られたと仮定する．

次の式によって(J_2+J_3)個の数値変数が定義される．

$$s_{ij}=\sum_{k=1}^{k_j} w_{jk}\delta_i(jk)\,,\quad (j=J_1+1,\ldots,J;i=1,\ldots,N) \tag{10}$$

これにより，第$1\sim J_1$番目の数値変数と併せてJ個の数値変数が得られる．

カテゴリースコアの設定が適切なものであれば，これらJ個の変数の相関行列に基づく主成分分析によってデータの内的構造の簡潔な表現を得ることが期待できる．

どのような基準によってスコアベクトルを定めたら良いか.

ここでは，Saito & Otsu (1983,1988) に採用されている基準（分散和最大化基準）と一般化分散最小化基準の 2 つについて検討する.

Maximizing Total Variance (MTV)，と Minimizing Generalized Variance (MGV)

カテゴリースコアを一意に決定するために，(10)のJ個の変数は，全て平均0．分散1に正規化することにする．数値変数についてはこれらの制約は

$$\frac{1}{N}\sum_{i=1}^{N} v_i(j) = 0 \text{ および } \frac{1}{N}\sum_{i=1}^{N} v_i(j)^2 = 0\,,\quad (j = 1, \ldots, J_1) \tag{11}$$

となる．また，カテゴリー変数については

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{k_j} w_{jk}\delta_i(jk) = 0 \text{ および}$$

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{k_j} {w_{jk}}^2\delta_i(jk) = 1\,,\ (j = J_1 + 1, \ldots, J) \tag{12}$$

と表現される．

数値変数については

$$s_{ij} = v_i(j)\,,\ (j = 1, \ldots, J_1)$$

とし．またカテゴリー変数については (10) によって正規化条件 (11), (12) を満たす $J_2 + J_3$ 個の変数を定める．

これによって得られる $N \times J$ の行列を $S = (s_{ij})$ とする．S はカテゴリースコア $\boldsymbol{w}$ の関数であるので，$S(\boldsymbol{w})$ とかける．また，このとき J 個の変数の相関行列は

$$R(\boldsymbol{w}) = \frac{1}{N} S'(\boldsymbol{w}) S(\boldsymbol{w}) \tag{13}$$

となる．この相関行列の J 個の固有値を

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_J \geq 0$$

とする．

MTV基準は，あらかじめデータを縮約表現する分析次元数mを定めておき，次の値をモデル適合度の指標とする．

$$\theta = \lambda_1 + \lambda_2 + \cdots + \lambda_m \tag{14}$$

これを変数の平均と分散についての制約の下で最大化する．

実対称行列の性質より，m個のJ次元ベクトル$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$が正規直交系であるとの制約のもとで

$$\theta = \max_{X,w} \sum_{i=1}^{m} \boldsymbol{x}_i{}' R(\boldsymbol{w}) \boldsymbol{x}_i \tag{15}$$

がいえる．

MTV基準の最大化は$\boldsymbol{w}$とXとについての最大化によって実現される．

MGV基準の場合には

$$\det R = \prod_{j=1}^{J} \lambda_j$$

の最小化を行なえばよい．これは

$$\eta = -\sum_{j=1}^{J} \log \lambda_j \tag{16}$$

の最大化に等しい．後者のほうが数値的な取扱いが容易であるので，これをMGV基準の目的関数とすることにする．

それぞれの変数X_jがX_jを除く他の変数によって回帰したときの重相関係数の2乗R_j^2が大きければ，MGVの適合度が高まる．

相関行列の行列式$\det R$（一般化分散）が最も大きくなるのは，各変数が完全に無相関であるとき，すなわちR

が単位行列である場合である．また，$\det R$はRが半正定値であることから常に非負であり，0となるのはSの列ベクトルが一次従属となる場合である．

推定されたカテゴリースコアベクトルを$\hat{\boldsymbol{w}}$とし，それに基づく変数間の相関行列を$\hat{R} = R(\hat{\boldsymbol{w}})$とする．また，$\hat{R}$の$m$個の固有ベクトルを列ベクトルとして持つ$J \times m$の行列を$\hat{X} = (\hat{x}_{jl})$とする．これ以降，アイテム(変数)$j$の布置ベクトルとは，$\hat{X}$の第$j$行目(長さ$m$のベクトル)を意味することとする．また，第$j$アイテムの第$k$カテゴリーの布置とは，アイテム$j$の布置ベクトルに，スカラー$\hat{w}_{jk}$を掛けたものとする．名義尺度アイテム(変数)について，その布置とカテゴリースコアは符号不定性を持つが，カテゴリー布置は符号不定性を持たない．さらに，$S(\hat{\boldsymbol{w}})$の第i行目のベクトルに$\hat{X}$を右から掛けて得られるm次元ベクトルによって，サンプルiの布置を定義する．因子分析の用語を用いると，アイテム布置は因子負荷量に相当し，サンプル布置は因子得点に相当する．

3.2 OSMOD分析例: NLSY79の概要

NLSY79 (BLSによる調査)
被験者は1957年から1964年のあいだに生まれた男女であり全部で12,686名.

クロスセクショナルサンプル(6,111名)
サプリメンタルサンプル(5,295名)
軍隊サンプル(1,280名)

このうち分析対象 7,025名（検討対象の変数の欠測を原則として除く）

NLSY79, 19変数：

Age 年齢, WithMan 14歳時の同居男性, WithWo 14歳時の同居女性,
MEdu 母親の学歴, FEdu 父親の学歴, Sib 同胞の数(1979), Reli 被験者の宗教, RelAtt 宗教活動への参加, Race 人種, Sex 性別, REdu 被験者の学歴(1992), EmpStat 雇用状態(1996),
LkJob 仕事への好感度(1996), Delinq 非行, Jail 矯正施設への入所,
Resi 居住地域, Pov 貧困基準に該当, AFQT AFQT得点, FmInc 世帯収入(1979,1980)

3.3 NLSY79 分析結果

MTV基準($q = 3$)

表 6: OSMODによる固有値

	1	2	3	4	5	6	7
固有値(MTV $q = 3$)	3.72	1.72	1.67	1.49	1.29	1.14	1.03
累積寄与率	0.20	0.29	0.37	0.45	0.52	0.58	0.63
固有値(MGV)	3.63	1.74	1.68	1.48	1.40	1.16	1.06
累積寄与率	0.19	0.28	0.37	0.45	0.52	0.58	0.64

表 7: 回転後の主成分の分散共分散/相関行列

	V_1	V_2	V_3	V_4	V_5
V_1	2.91	-0.32	0.19	-0.07	-0.18
V_2	-0.78	2.00	-0.16	0.02	0.26
V_3	0.45	-0.31	1.85	-0.03	-0.10
V_4	-0.15	0.04	-0.05	1.50	0.03
V_5	-0.39	0.46	-0.17	0.05	1.62

(上三角部分は相関係数)

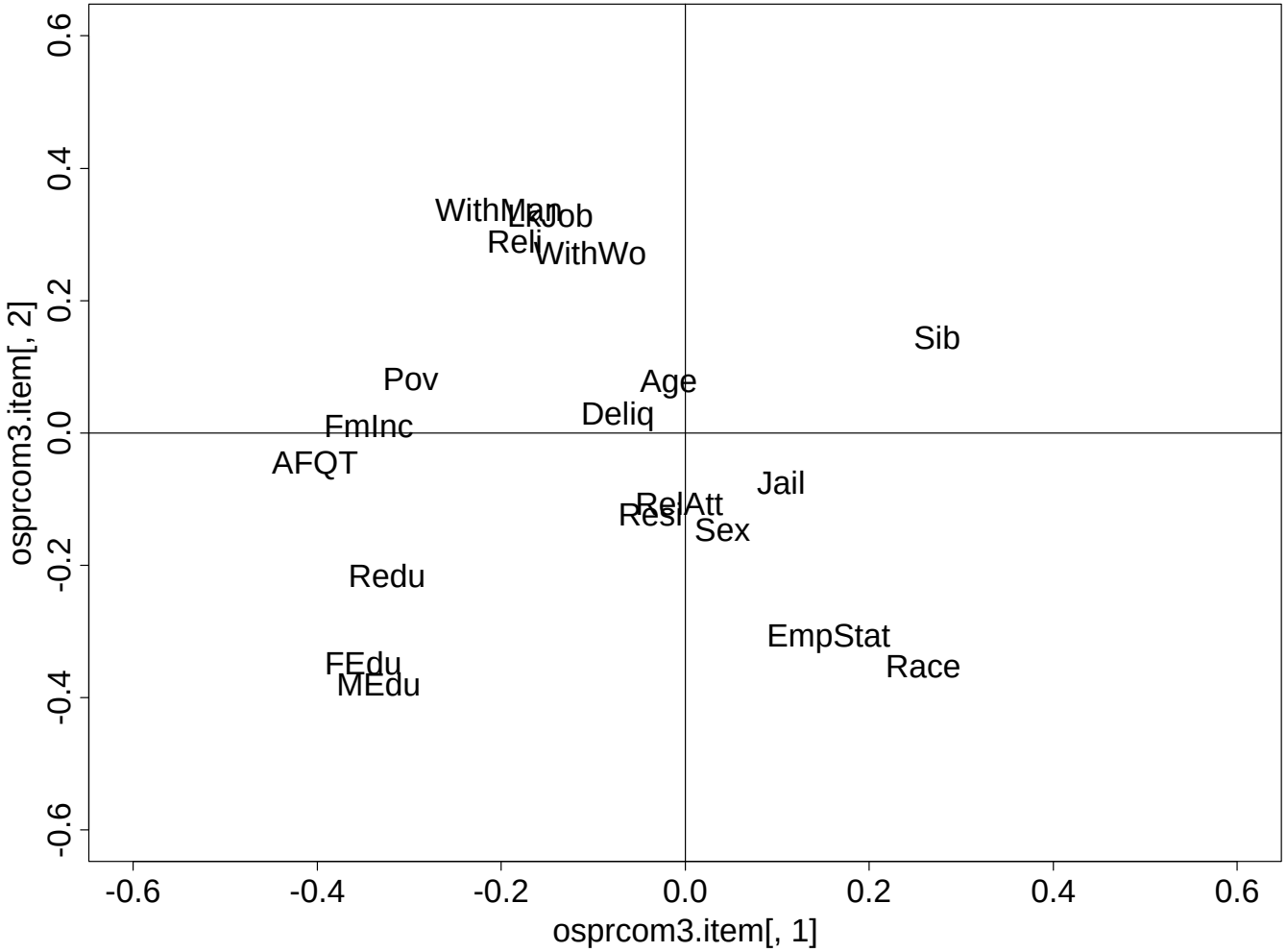

図 8: 回転前固有ベクトル（因子負荷量）のプロット

横軸 1 次元目，縦軸 2 次元目, MTV, $m = 3$

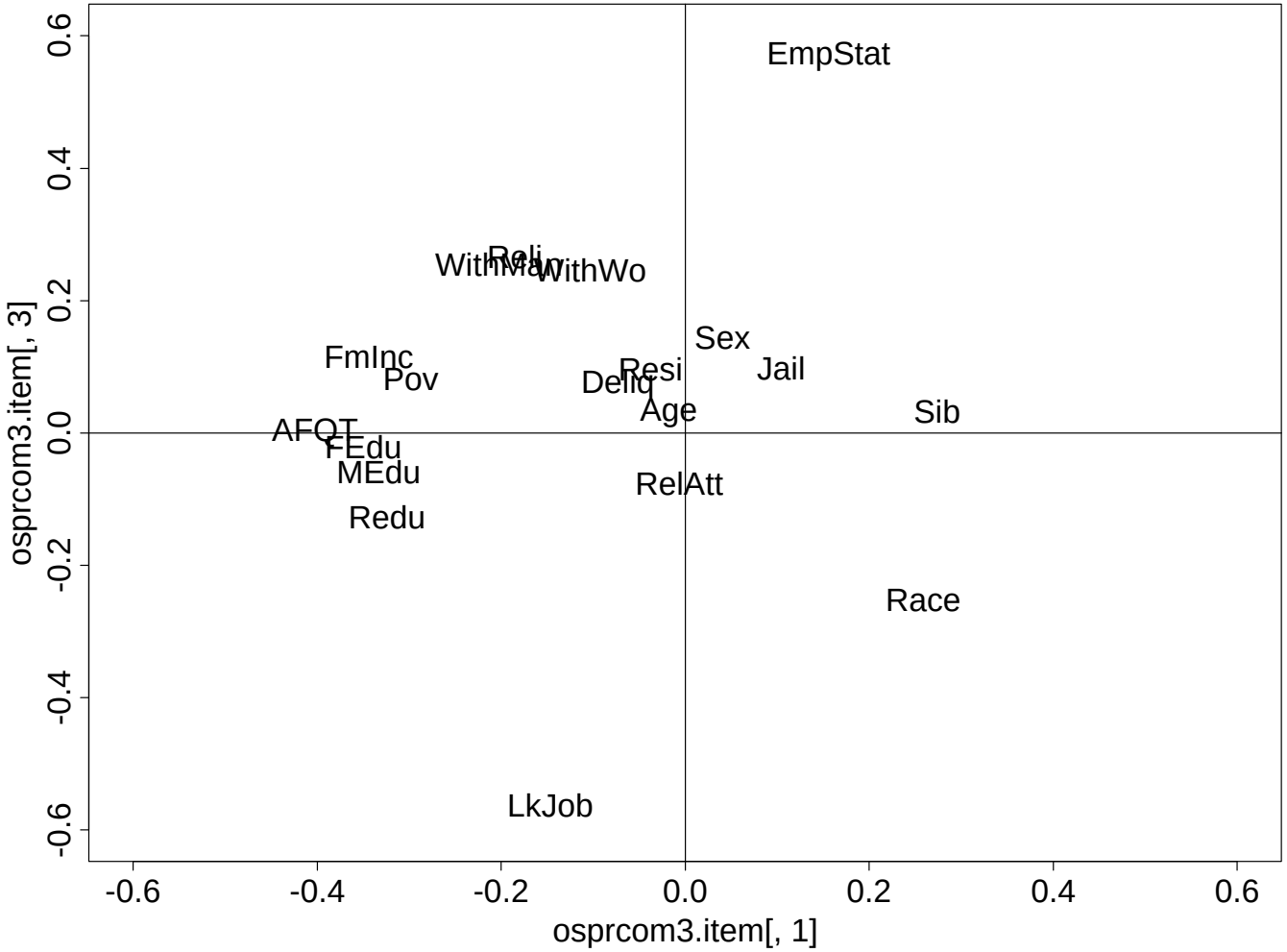

図 9: 回転前固有ベクトル（因子負荷量）のプロット

横軸 1 次元目，縦軸 3 次元目, MTV, $m = 3$

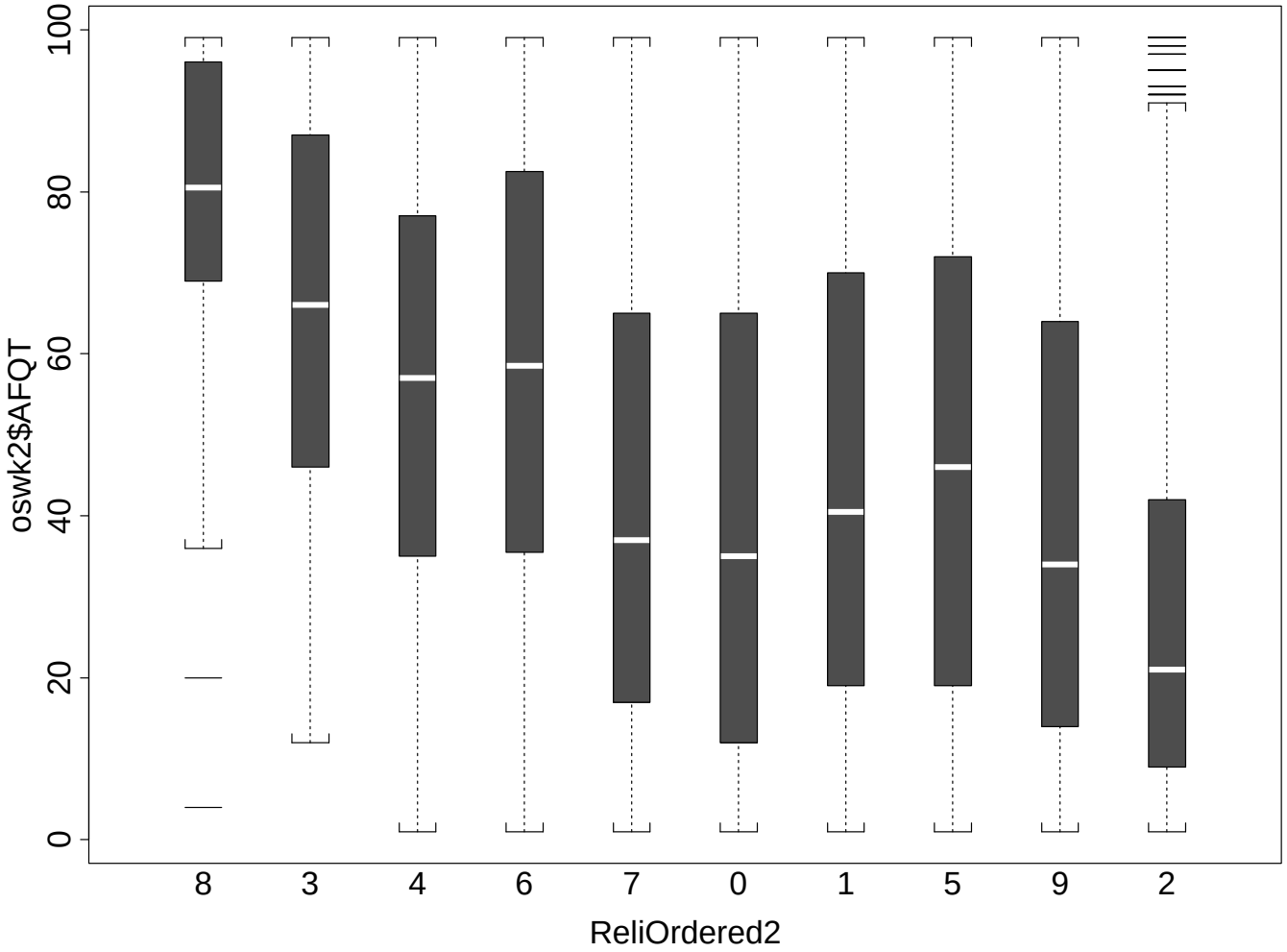

図 10: 宗教(スコア順)とAFQT

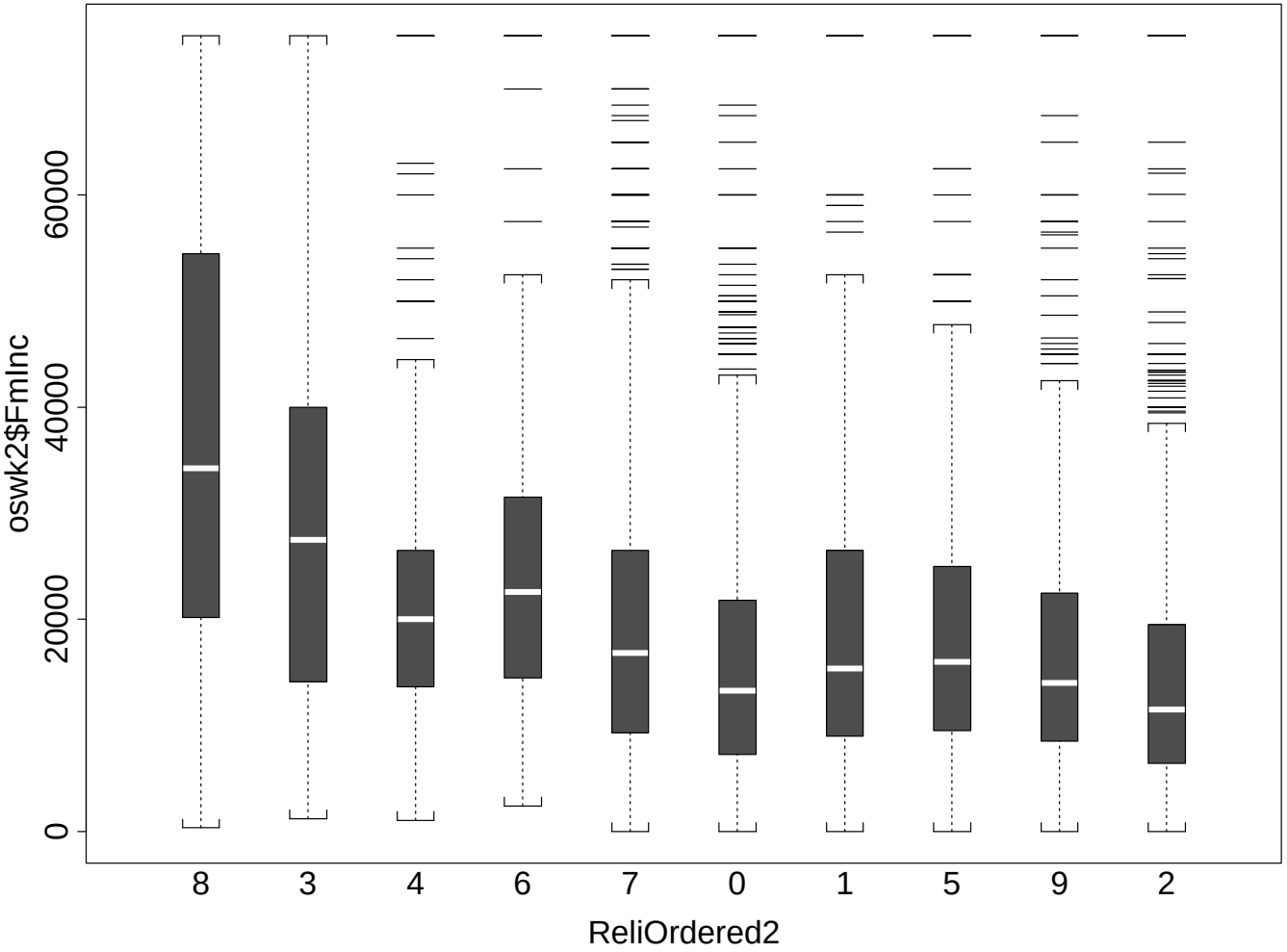

図 11: 宗教(スコア順)と世帯収入

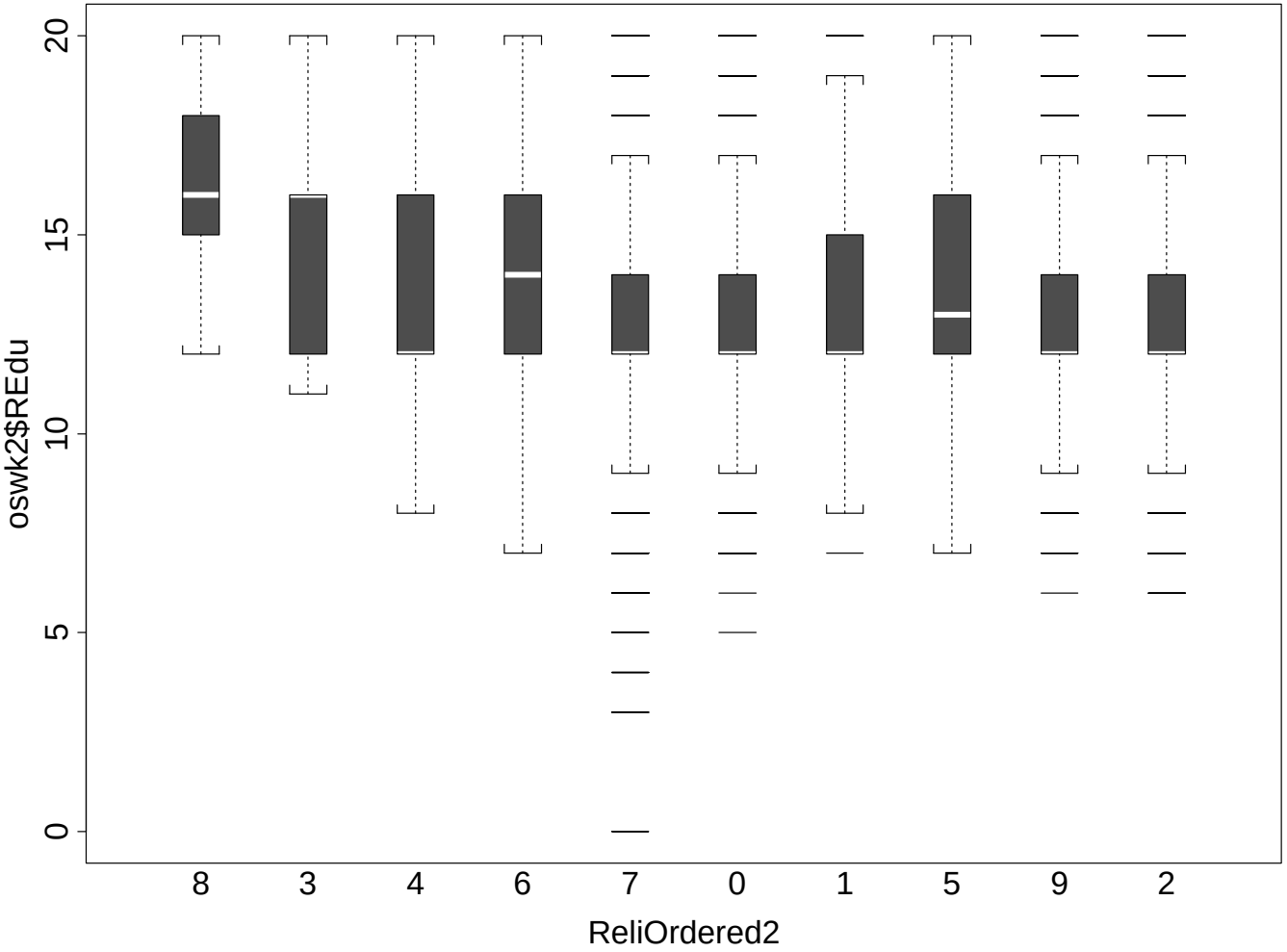

図 12: 宗教(スコア順)と被験者の教育年数

プログラムについて

- 対応分析
  SAS Corresp, SPSS Categries ,
  R & Splus (`corresp` in Modern Applyed Statistics with S (MASS). Venables & Ripley, Springer)
- 尺度最適化をともなう主成分分析
  SAS PRINQUAL, SPSS Category?, OSMOD (free)
- 大津の OSMOD サイト
  `http://www.rd.dnc.ac.jp/~otsu/osmod`
  ソースコード＋Winバイナリ, MinGW のg77 でコンパイル.

OSMODプログラム

1. Fortran 77 で記述（スタイルは古い）, 無償、無保証
2. 連続変数の非線形変換の機能はなし
3. 変数の重み付けは一律に固定(相関行列を分析)
4. Fortranの実数データとしてデータを読み込み（文字型データの入力不可)
5. コマンドウィンドウで実行。パラメータを標準入力から読み込み, パラメータで指定されたファイルにバイナリの計算結果を保持。標準出力にログ出力。
6. R & SPlus インタフェースも用意（データはファイル渡し)。

OSMODの実行:

コマンドウィンドウで以下を実行

```
osmod   < param.a1   > a1output.txt
osmodpr < param.b1   > b1output.txt
```

param.a1の中身 LARGE

```
*-------------------------------------
infile    'sample.data'
format    '*'
outfile   sample.out
nitem     5
* kfunc 1:MTV  2:MGV
kfunc     2
* level  0: nominal, 1:ordinal, 2:interval(numeric)
level      1   1
level      2   1
level      3   1
level      4   0
level      5   2
```

```
title   'DATA ANDREW HERTZBERG TB63.2'
* nr:分析次元 MGV のときは推定には関係しない
nr        2
* nvec: 固有値計算の反復のための固有ベクトル数
nvec      3
* limcat カテゴリ区分総数の上限
limcat    200
*-------------------------------------
```

4 まとめなど

- 名義尺度変数を含むデータの分析が面白い.
- 数値・順序尺度の場合は，多少の単調変換で結果が大きく影響されるのはまれ。ただし2値はかなり情報が落ちる。
- 数量化３類で分析していたデータをosmodなどで再分析すると，新たな発見があるかも。
- 精密なモデル評価には耐えないが，それなりには有用。欠測の扱いは実用上課題。
- Bertinのグラフィックスは商用システムでも欲しい. Mondrian システム (Java based GUI data analysis)