〈研究ノート〉

対応分析によるデータ解析*

中　山　慶一郎**

1．はじめに

幾つかの選択肢をもつ質問から構成される多くの調査データの解析に利用される手法の１つとして対応分析（Correspondence Analysis）は、かなり利用されつつある。本稿では、対応分析の理論の解説と、それの調査データの適応の仕方、および、R を用いた簡単な分析例を提示しようとするものである。

この論文でとりあげる分析手法は多くの異なる名前を持っており、そのうち主要なものを列挙すると、主成分尺度分析（Principal components of scale analysis）(Guttman; Lond)、質的データの要因分析 (factorial analysis of qualitative data) (Burt)，双対尺度法（dual scaling）(Nishisato)，数量化３類 (third method of quantification) (Hayashi)，多重対応分析（multiple correspondence analyisis）(BenZécri, Cazes, Lebart, Greeacre)，等質性分析 (homogeneity analysis) (Gifi) などが挙げられる。これらの手法は、多次元尺度法 (multidimensional scaling)、主成分分析 (principal component analysis)、尺度分析（scale analysis）などの考えを出発点として、様々なデータ解析から導出されてきた。

対応分析は単純な２次元表や多重表の行と列間の対応する測定値を分析する探索的データ解析の手法であり、また記述的データ解析の技法でもある。複雑なデータを単純化して、２次元または３次元での行と列のグラフィカルな表示は、変数間、対象物間の構造的関連性の発見に役立つ。対応分析はかなり柔軟にデータに適応する性質を持っている。対応分析は、１）データ行列が充分に大きく簡単な統計分析ではデータの構造がわからないとき、２）変数が同質で、行または列間の統計的距離を計算する意味があるとき、データ行列の行と列の幾何的図形による解釈ができ、分析を容易にし、関連性の探索に役立つ利点がある。

2．対応分析について

ここでは、Greenacre（2006, 2007）に従って、一般に用いられている多重対応分析（MCA）の説明を行うことにする。

Primitive matrix, N

元のデータ行列、N（*I, J*）は I × J のクロス表（contingency table）とし、この行列の要素は n_{ij} する。（i = 1, 2, ….., I),(j = 1, 2, ….., J）

プロファイル（Profiles）

クロス表の内容を理解するには各セルの実際の度数を比較するのはあまり意味がない。各行および各列は異なる反応数をもつので、データ全体数 n に対する比率で比較する。n_{ij} の周辺度数（marginal frequencies）を、n_{i+} と n_{+j} で表す。

$$n_{i+} = \sum_j n_{ij} \qquad n_{+j} = \sum_i n_{ij}$$

度数の総計は、$n = \sum_j \sum_i n_{ij}$ であるので、row profiles は $r_i = \frac{n_{i+}}{n}$、column profiles は $c_j = \frac{n_{+j}}{n}$ となる。

行プロファイルの行列（Matrix of Row Profiles）は

*キーワード：対応分析、多重対応分析、R

**関西学院大学名誉教授

Rows		Columns			Total
	1	2		J	
1	n_{11}/n_{1+}	n_{12}/n_{1+}	……	n_{1J}/n_{1+}	1
2	n_{21}/n_{2+}	n_{22}/n_{2+}	……	n_{2J}/n_{2+}	1
.	.	.	.	….	.
I	n_{I1}/n_{I+}	n_{I2}/n_{I+}	….	n_{IJ}/n_{I+}	1
Expected row profile c_j	n_{+1}/n	n_{+2}/n	….	n_{+J}/n	1

であり、列プロファイルの行列（Matrix of Column Profiles）は、

Rows		Columns			Expected column profile r_i
	1	2		J	r_i
1	n_{11}/n_{+1}	n_{12}/n_{+1}	……	n_{1J}/n_{+1}	n_{1+}/n
2	n_{21}/n_{+2}	n_{22}/n_{+2}	……	n_{2J}/n_{+2}	n_{2+}/n
.	.	.	.	….	.
I	n_{I1}/n_{+I}	n_{I2}/n_{+I}	….	n_{IJ}/n_{+I}	n_{I+}/n
total	1	1	….	1	1

χ^2 距離と χ^2 統計量

対応分析では変数間、個体間の距離を定義するために、χ^2 距離（Chi-square distance）を用いる。

いま、i 行の observed profile a_i から、i 行の expected profile c_j 間の χ^2 distance を

$$\left\| a_i - c_j \right\|_c = \sqrt{\frac{\sum_i (a_{ij} - c_j)^2}{c_j}}$$

と定義し、同様に j 列の observed profile b_j と、j 列の expected profile r_i 間の χ^2 distance は

$$\left\| b_j - r_i \right\|_r = \sqrt{\frac{\sum_i (b_{ij} - r_i)^2}{r_i}}$$

となる。

ここで、i 番目の行プロファイルに、inertia（標準化された分散）を、

$$\text{Inertia} = m\sum_j \frac{(r_{ij} - \bar{r}_j)^2}{\bar{r}_j}$$

$$r_{ij} = \frac{n_{ij}}{n_{i+}} \qquad \bar{r}_j = \frac{n_{+j}}{n}$$

と定義する。m は行、列のある量 n_{i+}、n_{+j} である。列 profile と、平均 profile（centroid）との距離の加重平均を inertia といい、χ^2 統計量との関連は次式で示される。

Total inertia を Φ^2 とすると、

$$\Phi^2 = \frac{\chi^2}{n} = \sum_i r_i \left\| a_i - c_j \right\|_c^2$$

$$= \sum_i r_i \sum_j (\frac{p_{ij}}{r_i} - c_j)^2 / c_j,$$

$$a_{ij} = \frac{n_{ij}}{n_{i+}} = \frac{n_{ij}/n}{n_{i+}/n} = \frac{p_{ij}}{p_{i+}} = \frac{p_{ij}}{r_i}$$

が、行に対して成り立ち、列に対しても同様に、

$$\Phi^2 = \sum_j c_j \sum_i (\frac{p_{ij}}{c_j} - r_j)^2 / r_i,$$

$$b_{ij} = \frac{n_{ij}}{n_{+j}} = \frac{p_{ij}}{c_j}$$

となる。

行プロファイルに対する χ^2 統計量は、

$$\chi^2 = \sum n_{i+} \times \frac{(n_{ij}/n_{i+} - n_{+j}/n)^2}{n_{+j}/n}$$

$$= \sum n_{i+} \times \frac{(P_{ij}/r_{i+} - c_j)^2}{c_j}$$

列プロファイルに対する χ^2 統計量は、

$$\chi^2 = \sum n_{+j} \times \frac{(n_{ij}/n_{+j} - n_{i+}/n)^2}{n_{i+}/n}$$

$$= \sum n_{+j} \times \frac{(p_{ij}/c_j - r_i)^2}{r_i}$$

ただし、$p_{ij} = \frac{n_{ij}}{n}$　$r_i = \frac{n_{i+}}{n}$　$c_j = \frac{n_{+j}}{n}$である。

これらの χ^2 統計量は、独立性の検定に使用されるものと同一である。

χ^2 統計量は、

$$\chi^2 = \sum \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}$$

$$= \sum \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

で相対度数 p_{ij} の標準化残差の平方和である。

行と列の双対性を考えて、整理すると、

$$\Phi^2 = \frac{\chi^2}{n} = \sum_{i,j} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

となる。$S = \frac{(p_{ij} - r_i c_j)}{\sqrt{r_i c_j}}$ を、標準残差 Standrard Residuals といい、対応分析の基礎となる。S は Correspondence Matrix ともいい、I × J の行列で、行変数から見ると、その profile は I 次元空間の I 個の点を表し、各点は行 profile から

centroid の距離を標準化したものである。

S を行列表示とし、ここで、$D_r = \mathrm{diag}(r_i)$、$D_c = \mathrm{diag}(c_j)$ とする。

S は連続変量における分散共分散行列に該当するもので、多変量解析のデータ行列の分解理論によると、$S = UD_\alpha V^T$ となる。

これを S の特異値分解 SVD（Singular Value Decomposition）といい、ここでは、（I × J）の行列とする。U は SS^T の固有ベクトルであり、V は S^TS の固有ベクトルである。D_α は S^TS の固有値 λ_k 平方根を要素とする対角行列 $D_\alpha = \mathrm{diag}(\lambda_k^{1/2})$ である。ただし、k = 1, 2,….., K, K = min{I−1, J−1} であり、$\lambda_k = \alpha_k^2$ 即ち、固有値（principal inertia）は特異値（singular value）の平方に等しい。

$$S^TS = VD_\alpha U^T UD_\alpha V^T = VD_\alpha^2 V^T = V\Lambda V^T$$
$$SS^T = UD_\alpha V^T VD_\alpha U^T = UD_\alpha^2 U^T = U\Lambda U^T$$
$$UU^T = VV^T = I$$

S の要素別の表現では、

$$S_{ij} = \sum_{k=1}^{K} \lambda_k^{1/2} u_{ik} v_{jk}$$

行と列との互いの対応関係を分析するのに固有ベクトル u_k と v_k に注目する。例えば、最初の２つの固有値が支配的であるとすると、$s_{ij} \approx \lambda_1^{1/2} u_{i1} v_{j1} + \lambda_2^{1/2} u_{i2} v_{j2}$ で近似される。

座標 u_{i1} と v_{j1} が他の座標に比較して同符号で大きければ、s_{ij} も大きく、i 番目の行と j 番目の列のカテゴリー間に正の連関が大である。又、異符号で大きければ、負の連関が大きくなる。

対応分析の応用では、最初の２つの固有値、λ_1、λ_2 が固有ベクトルで説明される χ^2 全体の比率の多くを占めるときグラフ表示されるのが普通である。対応分析では S の加重された行と列の射影 projection の値 f_k、g_k、によってグラフ表示される。

ここで、行と列の双対関係 dual relation

$$v_k = \frac{1}{\sqrt{\lambda_k}} S^T u_k \qquad u_k = \frac{1}{\sqrt{\lambda_k}} S v_k$$

を利用して、

$$f_k = D_r^{-1/2} S v_k = \sqrt{\lambda_k}\, D_r^{-1/2} u_k$$
$$g_k = D_c^{-1/2} S^T u_k = \sqrt{\lambda_k}\, D_c^{-1/2} v_k$$

成分座標値 Principal coordinate が得られる。Greenacre［1, 2］は f_k、g_k 以外に ϕ_k、γ_k を次のように定義し、標準座標値 Standard coordinate と呼んでいる。

$$\phi_k = D_r^{-1/2} u_k \qquad \gamma_k = D_c^{-1/2} v_k$$

行と列の座標を行列表示すると、

行の主成分座標（Principal coordinates of rows）：
$$F = D_r^{-1/2} UD_\alpha = \Phi D_\alpha$$
列の主成分座標（Principal coordinates of columns）：
$$G = D_c^{-1/2} VD_\alpha = \Gamma D_\alpha$$
行の標準座標（Standard coordinates of rows）：
$$\Phi = D_r^{-1/2} U$$
列の標準座標（Standard coordinates of columns）：
$$\Gamma = D_c^{-1/2} V$$

となる。各座標の加重平方和を計算すると、Principal coordinate では、

$$FD_rF^T = GD_cG^T = D_{\dot{\alpha}}$$

Standard coordinate では、

$$\Phi D_r \Phi^T = \Gamma D_c \Gamma^T = I$$

となるので、この両者のスケールの違いは、D_α（principal intertia α_k^2）だけである。

３．調査データへの適用

調査データに対応分析を用いることにする。一般に社会調査や意識調査に用いられる調査データは、幾つかの質問項目から構成されている。各質問は４つか５つの選択肢を持つものが多い。ここで、例として取り上げるのは、関西学院大学社会学部真鍋研究室によって2007年３月に実施された「価値観と生活意識に関する調査」[1] である。

例として、問12の質問 a、墓参について（a1, a2, a3, a4），更に、性別（男（m）、女（f）），年齢別（（若年（young），中年（middle），老年（old））を取り上げる。データは回答者が設問に

1）この調査の概要については、関西学院大学社会学部紀要104，真鍋一史「日本的な「宗教意識」の構造」を参照されたい。

対して選択した項目の番号を示している。通常、調査データは表１のように質問に対して回答者が選択した項目の番号を示したものが、データとして得られる。以下の計算プロセスは R を用いる。

表 1　response pattern matrix

Q12	性	年齢
2	1	3
2	1	3
1	2	2
1	2	2
1	2	2
・	・	・
・	・	・
・	・	・
3	1	3

表 2　indicator (dummy) variable matrix

a1	a2	a3	a4	m	f	young	middle	old
0	1	0	0	1	0	0	0	1
0	1	0	0	1	0	0	0	1
1	0	0	0	0	1	0	1	0
1	0	0	0	0	1	0	1	0
1	0	0	0	0	1	0	1	0
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・	・
0	0	1	0	1	0	0	0	1

表１のデータを R に読み込むには、 ここでは、Excel 上のシートから直接読み込むことにする。Excel 上でデータ範囲を指定して、

```
>data<-read.table("clipboard",header=TRUE)
```

とするのが簡便である。元のデータから表２のダミー変数に変換するには、青木のプログラムを利用してから、変数名を書き込むことにする。

```
>data.dummy<-make.dummy[2] (data)
>colnames(data.dummy)<-c("a1","a2","a3","a4",
 "m","f","young","middle","old")
>data1<-data.dummy[,c(1:4)]
                        ＃質問 a のみのデータ
>data2<-data.dummy[,c(5:9)]
                        ＃demographic のみのデータ
```

１．Q12と性別、年齢別データ（Demographic Data）のクロス表を作成するために、質問の回答データと他と分割して計算を実行する。

```
>table.N<-t(data2)%*%data1
```

>table.N　　**表 3．クロス表 N**

	a1	a2	a3	a4
m	229	79	65	40
f	304	77	56	27
young	64	40	35	21
middle	153	59	46	32
old	316	57	40	14

２．クロスデータから、確率行列 P ＝ (1/n)N をつくり、行と列の周辺度数を求める。

Row and Column masses r, c は

$$r = P1 \qquad c = P^{T}1$$

$$r_i = \sum_{j=1}^{J} p_{ij} \qquad c_j = \sum_{i=1}^{J} p_{ij}$$

さらに、行と列の r と c の対角行列をもとめる。

$$D_r = \mathrm{diag}(r) \qquad D_c = \mathrm{diag}(c)$$

```
>table.P<-table.N/sum(table.N)
>r<-apply(table.P,1,sum)     ＃列の周辺度数の比率
>c<-apply(table.P,2,sum)     ＃行の周辺度数の比率
```

表 4　table P と r 及び c

	a1	a2	a3	a4	r
m	0.130559	0.04504	0.037058	0.022805	0.235462
f	0.173318	0.0439	0.031927	0.015393	0.264538
young	0.036488	0.022805	0.019954	0.011973	0.09122
middle	0.087229	0.033637	0.026226	0.018244	0.165336
old	0.18016	0.032497	0.022805	0.007982	0.243444
c	0.607754	0.177879	0.13797	0.076397	1

2）http://aoki2.si.gunma-u.ac.jp/R/index.html より、数量化３類の subprogram を利用した

3．S（対応行列、標準化残差行列）を計算する。

$$S = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$$

```
>Drmh<-diag(1/sqrt(r))          #D_r^{-1/2}を求める
>Dcmh<-diag(1/sqrt(c))          #D_c^{-1/2}を求める
>S<-Drmh %*% (table.P-r %o%c) %*% Dcmh
                                #S を求める
```

>S　　表 5

	a1	a2	a3	a4
m	−0.03316	0.015422	0.025363	0.035911
f	0.031285	−0.01455	−0.02393	−0.03388
young	−0.08049	0.051647	0.065683	0.059939
middle	−0.04181	0.024651	0.022606	0.049941
old	0.083728	−0.05193	−0.05884	−0.07785

4．S の特異値分解 (Singular value decomposition) を行う。

```
>S.svd<-svd(S)
>S.svd
```

表 6　Singular value (Eigen value), Eigen vector

$d				
singular value	2.17E−01	2.16E−02	7.78E−03	7.08E−17
eigen value	4.73E−02	4.67E−04	6.06E−05	5.01E−33
$u				
	[,1]	[,2]	[,3]	[,4]
m	−0.26084	0.245046	−0.63324	−0.35772
f	0.246089	−0.23119	0.597425	−0.37916
young	−0.59665	−0.67917	−0.01705	−0.36451
middle	−0.33029	0.642149	0.384548	−0.49073
old	0.637429	−0.11346	−0.30647	−0.59547
$v				
	[,1]	[,2]	[,3]	[,4]
a1	0.604901	0.136867	−0.08723	−0.77959
a2	−0.3663	−0.28759	0.777969	−0.42176
a3	−0.44448	−0.54029	−0.61038	−0.37144
a4	−0.54987	0.778876	−0.12081	−0.2764

5．固有根、固有ベクトルを用いて、行と列変数の、Principal coordinates F, G を計算する。

```
>F<-Drmh %*% S.svd$u %*% diag(S.svd$d)
>G<-Dcmh %*% S.svd$v %*% diag(S.svd$d)
>F
```

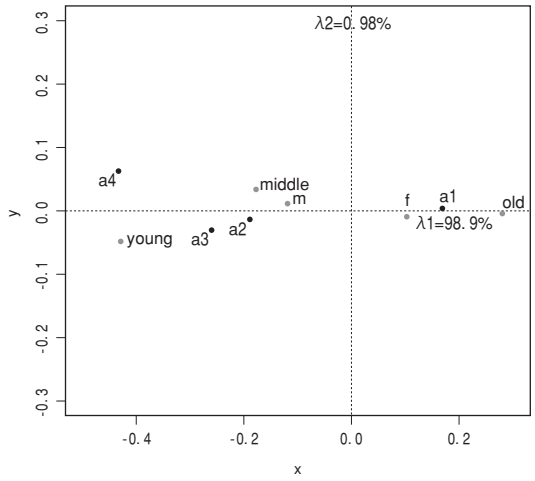
	[,1]	[,2]	[,3]	[,4]
m	−0.1169	0.010909	−0.01016	−5.22E−17
f	0.104055	−0.00971	0.009042	−5.22E−17
young	−0.42963	−0.04858	−0.00044	−8.54E−17
middle	−0.17666	0.034116	0.007362	−8.54E−17
old	0.280963	−0.00497	−0.00484	−8.54E−17

>G

	[,1]	[,2]	[,3]	[,4]
a1	0.168748	0.003793	−0.00087	−7.08E−17
a2	−0.18888	−0.01473	0.014359	−7.08E−17
a3	−0.26024	−0.03142	−0.01279	−7.08E−17
a4	−0.43265	0.060875	−0.0034	−7.08E−17

6．ここで、性別、年齢別データ及び回答データの計算結果である、F と G について、第1主成分軸と第2主成分軸の得点の散布図 biplot を描いてみる。

```
>x<-F[,1]
>y<-F[,2]
>plot(x,y,col="grey",pch=16,xlim=c(-0.5,0.3),ylim=c
 (-0.3,0.3))
>text(x,y,c("m","f","young","middle","old"),adj=c
 (0,0))
>x<-G[,1]
>y<-G[,2]
>par(new=T)
>plot(x,y,col="red",pch=16,xlim=c(-0.5,0.3),ylim=c
 (-0.3,0.3))
>text(x,y,c("a1","a2","a3","a4"),adj=c(1,1))
>abline(h=0,v=0,lty="13")
>a<-"λ1=98.9%";b<-"λ2=0.98%"
>text(0.19,0.01,a,cex=1.0)
>text(-0.02,0.3,b,cex=1.0)
>
```



7．実際の分析では、R のパッケージを利用するのがよい。R のパッケージとしては、いくつかのサイトがあるが、ここでは、ca を用いた分析を述べる。

>library(ca)
>ca(table.P)
>plot(ca(table.P))
を実行すれば、以下の結果が出力される[3)]。

	Principal	inertias	(eigenvalues):
	1	2	3
Value	0.047297	0.000467	6.10E-05
Percentage	98.90%	0.98%	0.13%

Rows:

	m	f	young	middle	old
Mass	0.235462	0.264538	0.09122	0.165336	0.243444
ChiDist	0.117851	0.104898	0.432365	0.180073	0.281049
Inertia	0.00327	0.002911	0.017053	0.005361	0.019229
Standard 1	−0.53755	0.478462	−1.97549	−0.8123	1.291911
Standard 2	0.504996	−0.44949	−2.24872	1.579253	−0.22995
Principal 1	−0.1169	0.104055	−0.42963	−0.17666	0.280963
Principal 2	0.010913	−0.00971	−0.0486	0.034128	−0.00497
C (1, r)	0.068038	0.06056	0.355992	0.109094	0.406315
C (2, r)	0.060048	0.053448	0.461277	0.412356	0.012872

Columns:

	a1	a2	a3	a4
Mass	0.607754	0.177879	0.13797	0.076397
ChiDist	0.168792	0.189998	0.262443	0.436928
Inertia	0.017315	0.006421	0.009503	0.014585
Standard 1	0.775927	−0.86851	−1.19663	−1.98941
Standard 2	0.175563	−0.6819	−1.45456	2.817935
Principal 1	0.168748	−0.18888	−0.26024	−0.43265
Principal 2	0.003794	−0.01474	−0.03143	0.060896
C(1,c)	0.365905	0.134175	0.197562	0.302358
C(2,c)	0.018733	0.08271	0.291909	0.606649

グラフの出力は上と同じである。

さらに、$FD_rF^T = D_\lambda$、$GD_cG^T = D_\lambda$ なる関係があるので、

$$r_i f_{ik}^2 = \lambda_k \qquad \begin{array}{l} \sum_{i,k} r_i f_{ik}^2 = \sum_{j,k} c_i g_{jk}^2 \\ \qquad\qquad (total\ ineritia) \\ \sum_i r_i f_{ik}^2 = \sum_j c_j g_{jk}^2 = \lambda_k \\ \qquad\qquad (principal\ ineritia) \end{array}$$

さらに standard coordinate と principal coordinate の間に $f_{ik} = \sqrt{\lambda_k}\,\phi_{ik}$ が存在するので、$\sum_k \frac{r_i f_{ik}^2}{\lambda_k} = \sum_k r_i \phi_{ik}^2 = 1$ となる。

同様に、$g_{jk} = \sqrt{\lambda_k}\,\delta_{jk}$ から、

$\sum_k \frac{c_i g_{jk}^2}{\lambda_k} = \sum_k c_j \delta_{jk}^2 = 1$ である。

これらは、f_{ik} の分散に対する行 i の absolute contribution であり、g_{jk} の分散に対する列 j の absolute contribution という。

C(i,r),C(i,c)i=1,2 は対応分析によるグラフを解釈するに役立つ。

Absolute contribution を R を用いて計算するには、次のプログラムを実行すればよい。
>F[,1]^2*r/S.svd$d[1]^2
>F[,2]^2*r/S.svd$d[2]^2
>G[,1]^2*c/S.svd$d[1]^2
>G[,2]^2*c/S.svd$d[2]^2

principal coordinate variable は平均 0 で分散は

3）ca()の出力は表の Standard までである。また、グラフの出力はすべてカラーで表わされているが、印刷ではカラーでは表示されていないので、正確なグラフの出力は R のプログラムを実行し、確かめることが出来る。

λ_k である。

4．R を用いた多重対応分析（MCA）の計算例

実際の調査データを R を用いる例で示すことにする。前節で用いたデータを拡大し、宗教的行動に関する質問群 Q12[4]から、 a, c, e, f, i を選び、Demographic variable である sex, age との関連について分析する。以下は単に分析の手順を示すことを主な目的とする。

```
<-data<-read.table(“clipboard”,header=TRUE)
                                    ＃データの読み込み
```

	a	c	e	f	i	sex	age
1	2	3	4	4	1	1	3
2	2	3	4	4	2	1	3
3	1	2	3	4	1	2	2
4	1	1	1	1	1	2	3
5	1	3	4	4	2	2	2
882	4	3	4	4	4	2	1

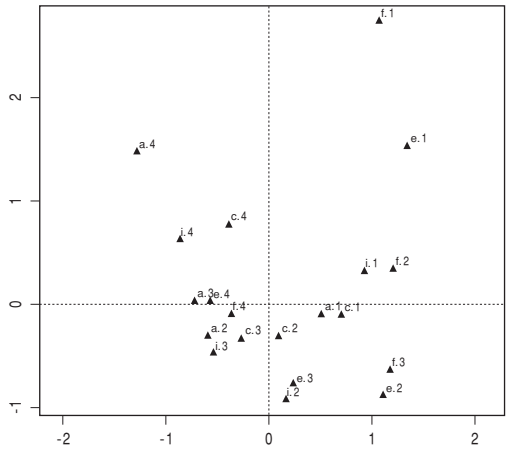
1．データの精査

集計したデータには欠測値が含まれるのが普通であるので、データを精査してプログラムに入力する。R では、簡単に処理できる。今の場合，欠測値[5]をデータから除くことにする。

```
>attach(data)    ＃変数名を data.frame に登録する
>missing<-a==9|c==9|e==9|f==9|i==9|sex==99|
 age==99                    ＃変数名の欠測値を指定
>data<-data[!missing,]
                    ＃欠測値を除いたデータを data にする
>dim(data)
[1] 853  7
>data1<-data[,c(1:5)]
                    ＃変数名だけのデータを data1 とする
```

2．data1 の多重対応分析（MCA）を行う。５つの質問を一括して処理する。

```
>library(ca)[6]          ＃package ca()を呼び出す
>z1<-mjca(data1,lambda=“indicator”)
                    ＃MCA を計算、ca の package から
                      mjca()を使用
>plot(z1,what=c(“none”,“all”))
                    ＃変数のみのグラフを書く
```



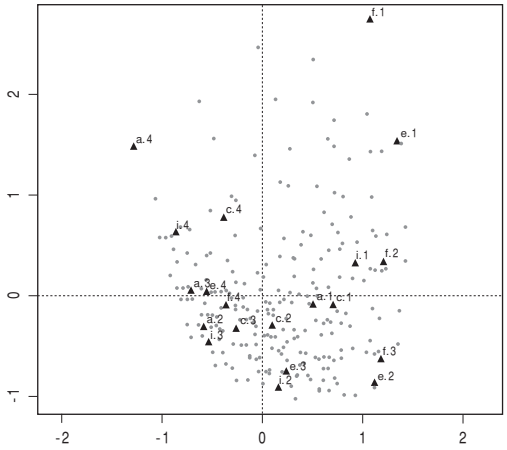
4）宗教的な行動についての質問

	記号	内容
Q12a	a1−a4	お盆やお彼岸などに墓参りをする
Q12c	c1−c4	お守りやおふだを買う
Q12e	e1−e4	ふだんから礼拝やお勤めなど宗教的な行いをする
Q12f	f1−f4	聖書や経典など宗教関係の本を読む
Q12i	i1−14	仏壇を拝む

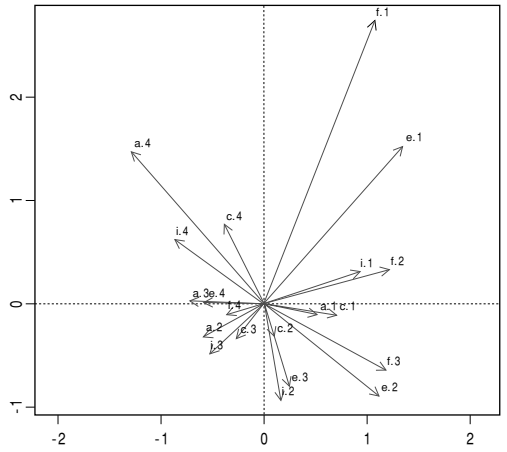
5）欠測値は、9, 99で指定されている場合である。

6）MCA のパッケージとして、代表的なものである。対応分析について R でのパッケージでは、MASS の corres, mca, Facto MineR での CA. MCA などがある。詳しくは、R のサイト、CRAN Task View:multivariate Statistics を参照されたい。

>plot(z1,labels=c(0,2))　　＃変数と個人のグラフ



>plot(z1,what=c(“none”,“all”),arrows=c(FALSE, TRUE))　　＃変数のベクトルを矢で示す



これらのグラフは変数の2次元における配置（response pattern）を表す。変数がどのようにまとまっているか、原点の左右にどの変数があるか、原点からの距離などから、軸の意味づけなど考慮する。

2．Response pattern の分析

a．2次元の response pattern の分析より、さらに高次の3次元空間の点の分布を見ることが出来る。

>library(scatterplot3d)[7]

＃package scatterolot3d（）を呼び出す

>x<-z1$colcoord[,1]*z 1$sv[1]

＃変数の第2主座標の値

>y<-z1$colcoord[,2]*z1$sv[2]

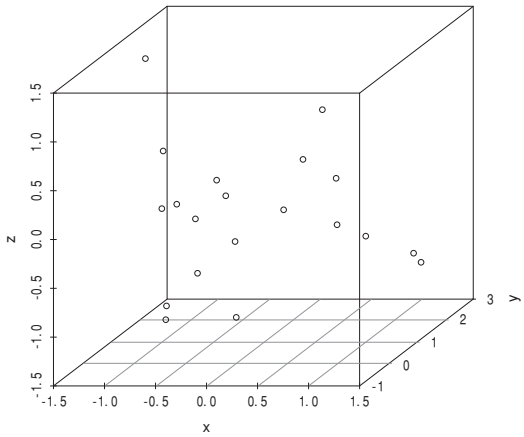
＃変数の第2主座標の値

>z<-z1$colcoord[,3]*z1$sv[3]

＃変数の第3主座標の値

>scatterplot3d(x,y,z,pch=1)

＃変数の3次元の配置を書く



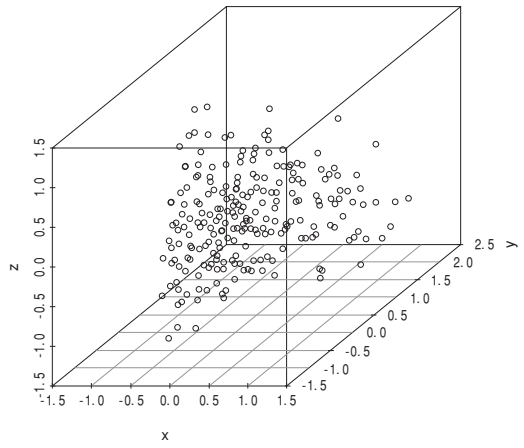
>x<-z1$rowcoord[,1]*z 1$sv[1]

>y<-z1$rowcoord[,2]*z 1$sv[2]

>z<-z1$rowcoord[,3]*z 1$sv[3]

>scatterplot3d(x,y,z,pch=1)

＃個人データの3次元配置を書く



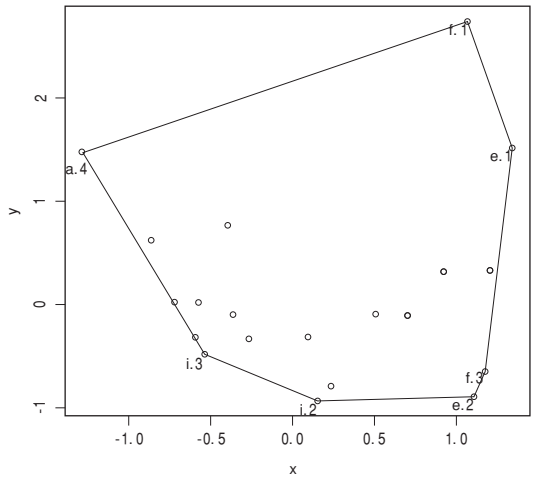
b．2変量データの凸包（convex hull）

2変量散布図（biplot）は縮約したデータのパターンを表すが、convex hull[8]を書くことによりデータの外れ値（outlier）を見つけることが出来

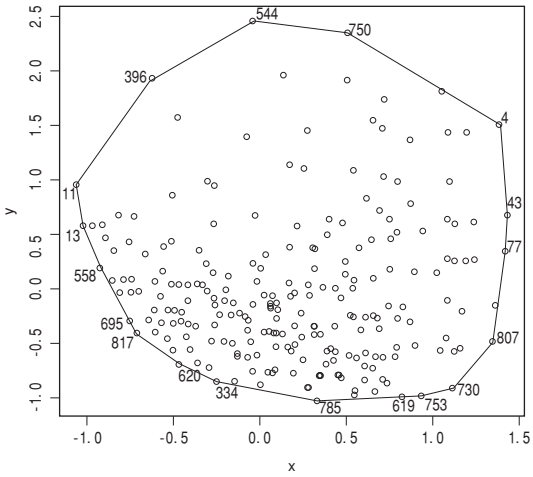
7）多変量データの Graphical Procedure のパッケージ

8）参考文献（6）p. 24

る。
列変数について、
>x<-z1$colcoord[,1]*z1$sv[1]
＃principal coordinate の第１座標
>y<-z1$colcoord[,2]*z1$sv[2] ＃主軸の第２座標
>plot(x,y,pch=1) ＃図をかく
>hull<-chull(x,y)
＃x,y の convex hull の値をもとめる
>polygon(x[hull],y[hull]) ＃convex hull を描く
>text(x[hull],y[hull],z 1.mca$levelnames[hull],adj=c(1,1)) ＃変数名をかく



行変数（個人のデータ）について
>x<-z1$rowcoord[,1]*z1$sv[1]
＃principal coordinate の第１座標
>y<-z1$rowcoord[,2]*z1.mca$sv[2]
＃主軸の第２座標
>plot(x,y,pch=1)
>hull<-chull(x,y)
>polygon(x[hull],y[hull])
>text(x[hull],y[hull],z1$rownames[hull],adj=c(0,0))



３．Demographic variable とのクロス分析

次に、性別、年齢別変数を導入して、質問変数と個人データとのクロス表との構造解析を行うことにする。

ａ．元のデータから Demograpic variable と Question のクロス表を作成し、対応分析を実行する。

>dim(data)
[1] 853 7
>library(ca) ＃package ca を呼び出す
>z2<-mjca(data,lambda=Burt[9]) ＃MCA の計算
>z3<-z2$Burt[21:25,1:20]
＃バート表から求めるクロス表を取り出す
>z3 クロス表

質問×性別・年齢別のクロス表

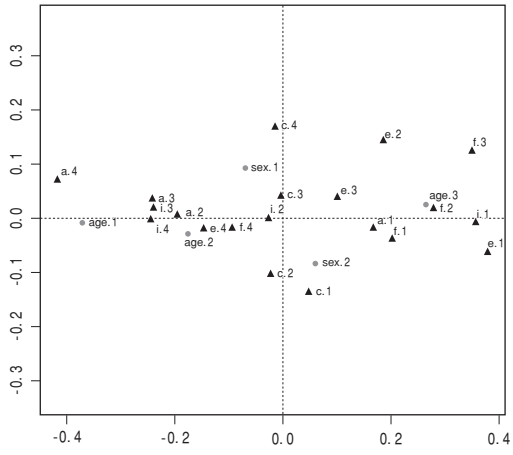
	a.1	a.2	a.3	a.4	c.1	c.2	c.3	c.4
sex.1	223	78	64	40	71	78	127	129
sex.2	290	75	56	27	118	115	126	89
age.1	63	40	35	21	31	41	47	40
age.2	149	59	45	32	67	65	84	69
age.3	301	54	40	14	91	87	122	109

9）MCA のクロス表、２次元のクロス表を一般化したもの

	e.1	e.2	e.3	e.4	f.1	f.2	f.3	f.4
sex.1	43	51	63	248	18	22	54	311
sex.2	73	44	65	266	26	27	55	340
age.1	5	12	14	128	6	2	9	142
age.2	28	23	45	189	11	15	19	240
age.3	83	60	69	197	27	32	81	269

	i.1	i.2	i.3	i.4
sex.1	118	74	96	117
sex.2	171	81	86	110
age.1	19	31	49	60
age.2	63	52	74	96
age.3	207	72	59	71

```
>ca(z3)                          ＃通常の計算
>plot(ca(z3)                     ＃Biplot を描く
```

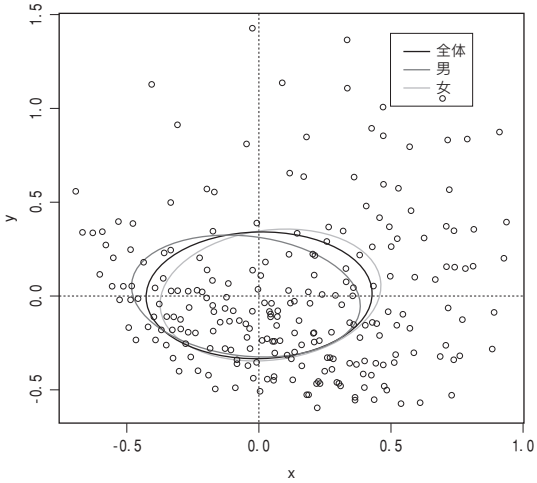


この計算例では、前節で述べたように、ダミー変数に変換しないで、直接関数から計算した。図からは、ほぼ1次元上の位置から変数の意図することが理解できよう。

b．個人と質問のクロス表から、Domographic factors の効果を分析する。

ここでは、MCA の計算結果を性別、年齢別に分割し、各要因ごとに、楕円体（ellipsoid）と convex hull を組み合わせてみることにする。

```
>z1<-mjca(data1,lambda="Burt")    ＃MCA の計算
>x<-z1$rowcoord[,1]*z1$sv[1]
                                 ＃各個人の第1主成分軸の値
>y<-z1$rowcoord[,2]*z1$sv[2]
                                 ＃各個人の第2主成分軸の値
>ds<-cbind(x,y,data[6:7])
  ＃各個人のデータと sex, age のデータを作る
>s1<-subset(ds,sex==1,select=c(x,y))
                    ＃男性のデータ（x,y）を取り出す
>s2<-subset(ds,sex==2,select=c(x,y))
                                 ＃女性のデータを取り出す
>s3<-subset(ds,age==1,select=c(x,y))
                                 ＃若年のデータを取り出す
>s4<-subset(ds,age==2,select=c(x,y))
                                 ＃中年のデータを取り出す
>s5<-subset(ds,age==3,select=c(x,y))
                                 ＃老年のデータを取り出す
性別についての楕円体[10]をかく
>plot(x,y,pch=1)
>abline(h=0,v=0,lty="13")        ＃軸を描く
>draw.ellipse(x,y,col="red")
                    ＃全体のデータの楕円を赤で描く
>draw.ellipse(s1,col="blue")     ＃男性を青で描く
>draw.ellipse(s2,col="green")    ＃女性を緑で描く
>legend(0.5,1.4,c("全体","男","女"),lty=1,
 col=c("red","blue","green"))    ＃凡例
```

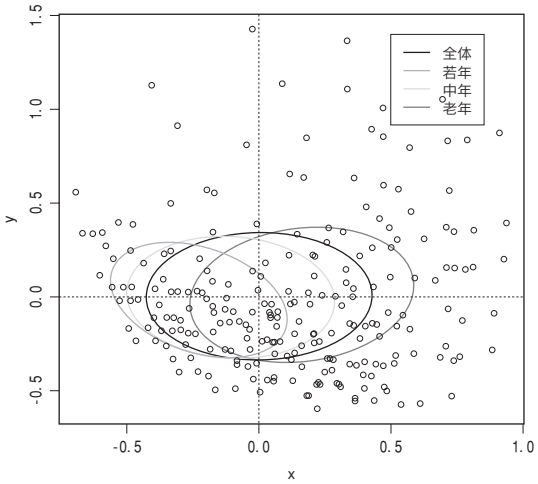


年齢別の楕円体

```
>plot(x,y,pch=1)
>abline(h=0,v=0,lty="13")
```

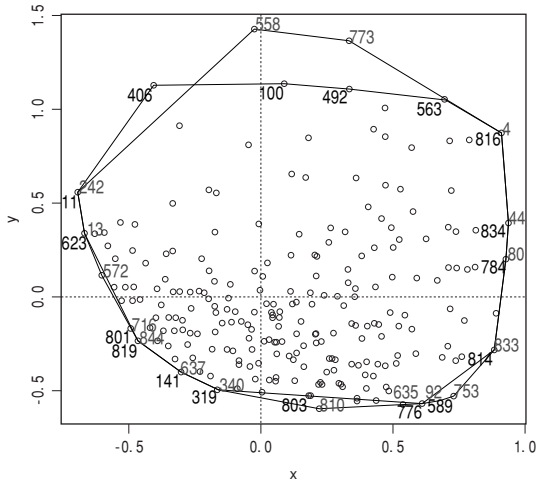
10）R には、ellipse パッケージがある。

```
>draw.ellipse[11](x,y,col="red")
>draw.ellipse(s3,col="blue")
>draw.ellipse(s4,col="green")
>draw.ellipse(s5,col="navy")
>legend(0.5,1.4,c("全体","若年","中年","老年"),lty=
 1,col=c("red","blue","green","navy"))
```



性別の convex hull

```
>plot(x,y,pch=1)
>abline(h=0,v=0,lty="13")
>hull<-chull(s1$x,s1$y)
>polygon(s1$x[hull],s1$y[hull])
                          ＃男性の convex hull
>text(s1$x[hull],s1$y[hull],rownames(s1)[hull],adj
 =c(1,1))                          ＃名前をつける
> hull<-chull(s2$x,s2$y)
> polygon(s2$x[hull],s2$y[hull])
                          ＃女性の convex hull
>text(s2$x[hull],s2$y[hull],rownames(s2)[hull],adj
 =c(0,0),col="red")
```



年齢別の convex hull

```
>plot(x,y,pch=1)
>abline(h=0,v=0,lty="13")
>hull<-chull(s3$x,s3$y)
>polygon(s3$x[hull],s3$y[hull])
>text(s3$x[hull],s3$y[hull],rownames(s3)[hull],adj
 =c(1,1))
```

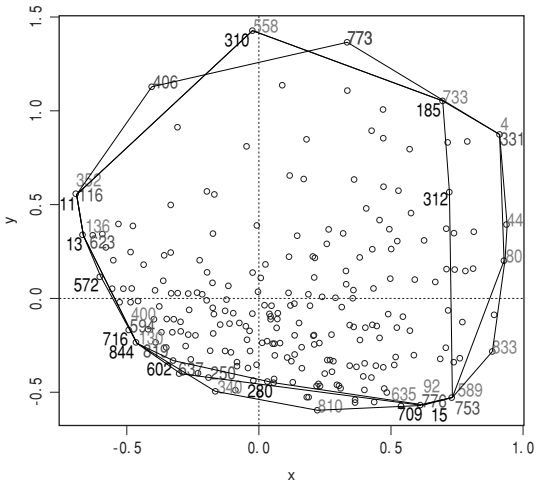
11）http://zoonek2.free.fr/UNIX/48.R/all.html より、

```
>draw.ellipse
function(
      x,y=NULL,N=100,method=lines,...)
  {
  if (is.null(y)){
        y<-x[,2]
        x<-x[,1]
   }
  center<-c(mean(x),mean(y))
  m<-matrix(c(var(x),cov(x,y),
                  cov(x,y),var(y)),
              nr=2,nc=2)
  e<-eigen(m)
  r<-sqrt(e$values)
  v<-e$vectors
  theta<-seq(0,2*pi,length=N)
  x<-center[1]+r[1]*v[1,1]*cos(theta)+r[2]*v[1,2]*sin(theta)
  y<-center[2]+r[1]*v[2,1]*cos(theta)+r[2]*v[2,2]*sin(theta)
  method(x,y,...)
}
```

```
>hull<-chull(s3$x,s3$y)
>polygon(s3$x[hull],s3$y[hull])
>text(s3$x[hull],s3$y[hull],rownames(s3)[hull],adj
 =c(0,0),col="red")
```

S3 を S4、S5 に変えて、プログラムを続ける。

a では、Demogrphic　variable による構造的変化が見て取れるし、b では、更に、外れ値の情報が得られる。また、各グループについて、通常の統計分析を行うのが便利である。



ここの分析では、性別については殆ど差がなく、年齢別でも、差がない。わずかに年齢別で宗教心が薄れていくのが見られる。

5．おわりに

ここまで分析した手法は、MCA と geometric analysis の一部であるが、実証分析が蓄積されるに従って、分析の有効性が明白になると思う。CA は多くの多方面にわたる理論的研究が進行中であり、分析結果の安定性についての議論も存在する。対応分析は、2 次元のクロス表から説明されることが多いし、類似の他の解析方法によっても計算上ほぼ同じ結果がもたらされる、これらを統一して説明する理論ができることが、期待される。また、分析結果はグラフ表示されるが、その解釈は常に明瞭であるとは云えない。

一方、データ構造をみると、データの個数×変数（n×m）という行列で表されるのが普通であるが、クロスデータのデータ構造から理論の展開を見ると、1．変数×変数、2．個数×変数との 2 種類あるが、ca では、1 は lambda=Burt、2 は lambda=indicator として区別している。一般に理論の解説では 1 の場合から説明することが多い。この場合、行変数と列変数は相互に入れ替えることが出来、Biplot の解釈も各変数の相対的位置の相違及び、互いの変数の位置関係で説明されることがある。

これに対して、2 の場合は変数の配置と個人の配置とは別々に説明され、個人の配置の分析は Geometric data analysis と呼ばれている。この点については、4 節で少しとりあつかっている。

参考文献

（1） M. Greenacre and J. Blasias ed.（2006）Multiple Correspondence Analysis and Related methods, Chapman & Hall/CDC
（2） M. Greeancre（2007）Correspondence Analysis in Plactice, Chapman & Hall/CRC
（3） B. Le Rowx and H. Ronanet（2004）Geometric Data Analysis, Kluwer Academic Publishers
（4） F. Murtagh（2005）Correspondence Analysis and Data Cording with Java and R, Chapman & Hall/CRC
（5） 大津起夫　社会調査データからの推論（2003）、言語と心理の統計　岩波書店
（6） B. エヴェリット、石田基弘訳（2005），R と S-Plus による多変量解析　Springer Japan
（7） 間瀬茂（2007）R プログラミングマニュアル　数理工学社
（8） P. Murrell（2006）R Graphics,Chapman & Hall/CRC

Statistical Data Analysis by the method of correspondence analyisis

ABSTRACT

Correspondence analysis is a statistical method to analyze and describe graphically and synthetically large amounts of data, which are the results of social investigation. I explain the essence of the theory of correspondence analysis and show how to apply it to the social investigation by implementing the software R program.

Key Words: correspondence analysis, multiple correspondence analysis, R