〈研究ノート〉

対応分析によるデータ解析*
── その3 ──

中　山　慶一郎**

1．はじめに

前稿で説明した正準対応分析Canonical Correspondence analysisによる分析を、本稿でも取り上げ、その理論と、データ解析にどのように応用するかについて説明する。終りに、対応分析について、データ構造、理論と応用について考察し、その解析の展望を述べることにする。

2．Canonical Correspondence Analysis（CCA）について

正準対応分析とは、対応分析と回帰分析を結びつけたものである。$\mathbf{Y}(n\times p)$はn個の個体とp変数のデータ行列とし、$\mathbf{X}(n\times m)$をn個の個体とm個の変数をもつデータ行列とする。$\mathbf{Y}$の変数ベクトル$\mathbf{y}$を、$\mathbf{y}=(y_1, y_2, \cdots, y_p)$とし、$\mathbf{X}$の変数ベクトル$\mathbf{x}$を、$\mathbf{x}=(x_1, x_2, \cdots, x_m)$とする。CCAでは、元の変数$\mathbf{y}$と回帰による推定量$\hat{\mathbf{y}}$を回帰モデルによって推定する。

$$\hat{\mathbf{y}}_i=b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_px_{im}$$

ここで、重回帰モデルの決定係数の平方

$$\mathbf{R}^2=[r(\mathbf{y},\hat{\mathbf{y}})]^2$$

を最大にする。この解は固有値問題となるが、CCAにおいては、correspondence analysisと同様$\mathbf{Y}$はχ^2距離を用いて計算したSを用いる。即ち、CCAでは説明変数$\mathbf{X}$について行列

$$\mathbf{S}=\frac{p_{ij}-p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}$$

の加重線形回帰によって得られる推定行列$\hat{\mathbf{Y}}$が得られる。行列$\mathbf{X}$にweightとして、対角行列$\mathbf{D}(p_{i+})^{1/2}$を用いて、多重回帰を実行すると、係数行列$\mathbf{B}$、$\mathbf{Y}$の推定値$\hat{\mathbf{Y}}$は次式で得られる。

$$\mathbf{B}=[\mathbf{X}^T\mathbf{D}(p_{i+})\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{D}(p_{i+})^{1/2}\mathbf{S}$$
$$\hat{\mathbf{Y}}=\mathbf{D}(p_{i+})^{1/2}\mathbf{XB}$$

$\hat{\mathbf{Y}}$の共分散行列$\mathbf{S}_{\hat{Y}^T\hat{Y}}$は

$$\mathbf{S}_{\hat{Y}^T\hat{Y}}=\mathbf{S}_{YX}\mathbf{S}^{-1}{}_{XX}\mathbf{S}^T{}_{YX}$$

となり、固有方程式は

$$(\mathbf{S}_{\hat{Y}^T\hat{Y}}-\lambda_k\mathbf{I})\mathbf{u}_k=0$$

となる。上式から、CCAにおける固有値の対角行列$\boldsymbol{\Lambda}$と固有ベクトル$\mathbf{U}$が計算される。これらの値を用いて、CCAにおけるn個のデータの座標とp個の変数の座標が求まる。

$$\hat{\mathbf{U}}=\mathbf{SU}\Lambda^{-1/2}$$
$$\mathbf{V}=\mathbf{D}(p_{+j})^{-1/2}\mathbf{U}$$
$$\hat{\mathbf{V}}=\mathbf{D}(p_{i+})^{-1/2}\hat{\mathbf{U}}$$
$$\hat{\mathbf{V}}=\mathbf{D}(p_{i+})^{-1/2}\mathbf{SU}\Lambda^{-1/2}$$
$$\mathbf{F}=\hat{\mathbf{V}}\Lambda^{1/2}\qquad\hat{\mathbf{F}}=\mathbf{V}\Lambda^{1/2}$$

標準座標（Standard coordinates）は、個体（row）については$\mathbf{V}$、変数（column）について$\hat{\mathbf{V}}$が得られ、主座標（Principal coordinate）は、$\mathbf{F}$（row），$\hat{\mathbf{F}}$（column）が求まる。

3．Canonical Correspondence Analysisの応用について

CCAの目的は2つのデータセットの関連性に

*キーワード：対応分析、正準対応分析、R
**関西学院大学名誉教授

あるが、前稿で試みに取り上げた２つの質問データでは分析結果の解釈が難しく、パラメータの推定にも質問の形式からランク落ちが生じるという問題点がある。回帰では独立変数から従属変数を説明するという意味をいかし **Y** を質問群とし、**X** を **Y** を説明する Demographic variable に設定するほうが分析目的に適合するものと考えて、前稿の分析に継続して、**Y** を日本の Q５(a, b, c)、ドイツの Q４（A, B, C）および **X** を各々 age, sex, school, revenue, size の Demographic variables のデータ行列とする２つのモデルを設定して計算する。

a．Q４を **X** について、回帰した計算結果[1]は、summary(Q 4.cca)で大略表示されるが、回帰式は

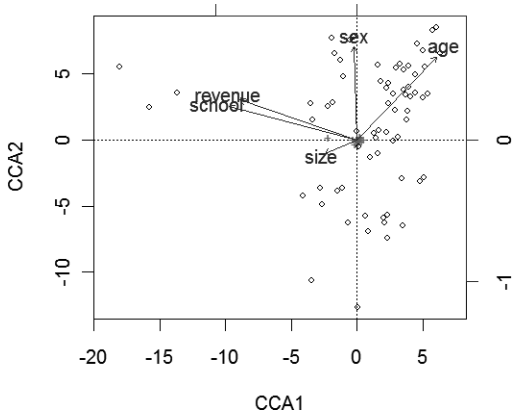
cca(formula=Q 4～age+sex+school+revenue+size, data=Dv)

で示される。

Partitioning of mean squared contingency coefficient:

	Inertia	Proportion
Total	4	1
Constrained	0.0798	0.01995
Unconstrained	3.9202	0.98005

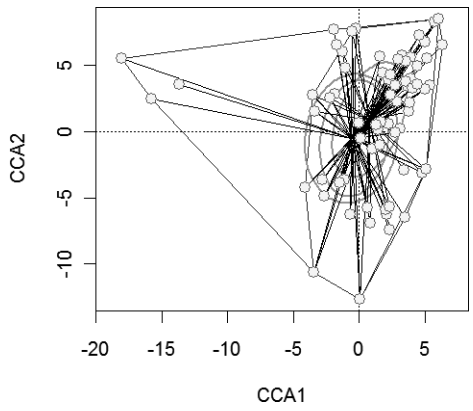
これから、制約式による Inertia（分散）の説明力は２％程度にすぎない。また、Canonical axes の固有値の合計は、0.0798で、Non-canonical axes の固有値の合計は、3.902である。他に **Y** の変数の座標点（Species scores）が $\hat{\mathbf{F}}$、データの点（Site scores）が $\hat{\mathbf{V}}$が得られる。グラフは、>plot(Q4.cca)で図では、**Y** の変数の点が座標の原点の近くに集まり、データを表す点が散らばっており、**X** の変数が Biplot(dashed arrows)で示されている。



Biplot は **X** の変数の大きさと方向を示している。

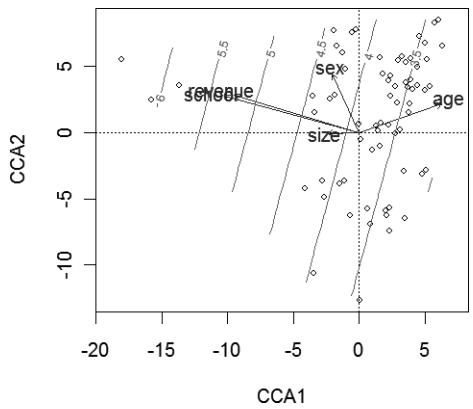
b．確率実験を伴う分散分析を、vegan package を用いて算出[2]すると、モデルは有意であり、変数では、age, school が、軸では CCA１が有意である。

c．CCA の分析におけるデータ点のグラフを表示し分布の状態を観察する[3]。



d．ここで、グラフ表示で、有用なものとして、vegan の package から、**X** の変数のグラフに直交する等高線を引くことにより、データの変動の様子が観察することができる[4]。

School, revenue の level とデータの分布



この図は、**X** の変数の水準によるデータの分布を観察するのに便利である。学歴が高く、所得の低いデータが集団から離れていることが観察される。

e．モデル選択について、vegan には回帰分析に通常含まれている最適な **X** の変数を選択するプログラムが含まれているので、それを利用する

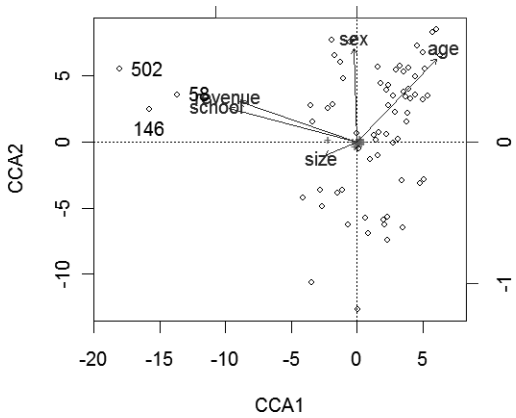
と、**Y** を説明する最適な **X** の変数を含むモデルを定めることができる[5]。結果は、Q 4 を説明する変数として、age, school をえらぶ model を選択する。

f. 異常値（Outlier）の検出

また、vegan のグラフ機能を使って outlier を見つけ、データの性質を調べることも可能である。

```
> fig<-ordiplot(Q 4.cca)
> identify(fig,"sites")
[1] 58 146 498 （498 は誤りで 502 とする）
```

この点は、グラフから、元のデータセットから、確かめる事ができる[6]。

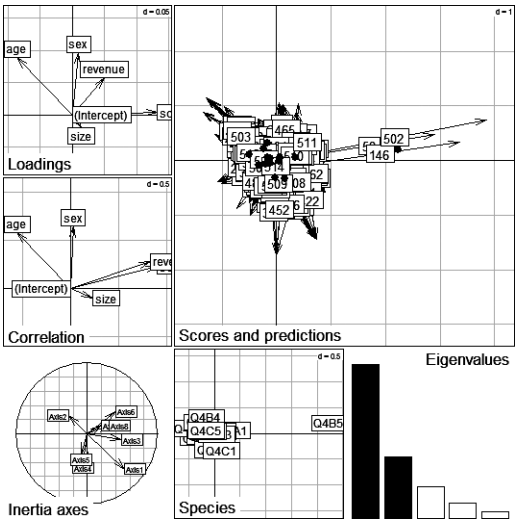


グラフからデータ番号 58, 146, 502 を outlier として指定する。また、これらのデータを教育水準が高く、収入が低く、若年層とみなすと、19人の層が指定される。この中で、質問 QB 4 で 5 を選択した人が outlier であることが分かる。

4．補　論

1．ここでは、参考までに package ade 4 を用いた、CCA を実行した結果を示す。vegan と同じ結果をもたらすが、グラフ表示はかなり異なるが内容的には同じである。

```
>library(ade 4)
>Q 4<-Ddata 1[,c(1:3)]
>Dv<-Ddata 1[,-c(1:3)]
>Q 4<-make.dummy(Q 4)
>Q 4<-data.frame(Q 4);Dv<-data.frame(Dv)
>colnames(Q 4)<-c("Q 4 A 1","Q 4 A 2","Q 4 A 3","Q 4 A 4","Q 4 A 5","Q 4 B 1","Q 4 B 2","Q 4 B 3","Q 4 B 4","Q 4 B 5","Q 4 C 1","Q 4 C 2","Q 4 C 3","Q 4 C 4","Q 4 C 5")
>ivl<-cca(Q 4,Dv,scan=FALSE)
>plot(ivl)
```



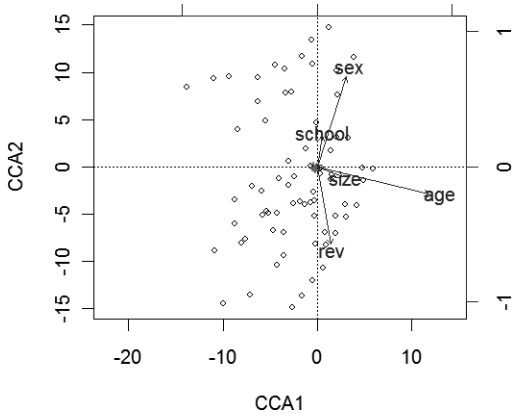
```
Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' ' 1
P values based on 999 permutations.
```

2．日本のデータについて、Q 5 a, Q 5 b, Q 5 c と Demographic variables との、CCA 分析は、以下の program を実行すればよい。

```
>Jdata 1[1:2,]
```

	Q5a	Q 5 b	Q5c	age	sex	school	rev	size
1	2	2	2	3	1	2	2	2
2	2	2	3	3	1	4	2	2

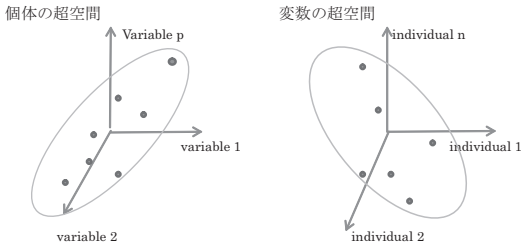
```
>Q 5<-Jdata 1[,c(1:3)]
>Dv<-Jdata 1[,-c(1:3)]
>Q 5<-make.dummy(Q 5)
>library(vegan)
>Q 5.cca<-cca(Q 5～,data=Dv)
>plot(Q 5.cca)
```



5．Correspondence Analysis のデータ分析について

a．ここで、対応分析が適用されるデータ構造について、簡単にふれておく。多変量解析の1つである correspondence analysis で分析されるデータは、幾つかの質的変数 categorical variables に対する測定単位の n×p のデータになっている。これまで分析したデータは1つの質問が幾つかの選択肢 k_i を持つものであった。従ってデータ行列は、$n\times\Sigma p_i$ の、0, 1 の値をもつ binary matrix で表現される。2つ以上の質問を分析する場合は multiple correspondence analysis という。測定単位が個体でなく集合体であるときは、データは n×p の行列で、その値は頻度 frequency を表している。従って、対応分析は多次元の分割表で表わされる一般的なクロスデータで説明されることが多い。

b．分析の対象となるデータ **X** を n 列（個体）、p 行（変数）の表とする。この表で n 次元空間の p 個の点を示すものとすると、各点は変数を表し、各次元は個体に対応する。点の座標は変数に対応する個体のもつ値である。また、各点を p 次元空間の n 個の点とすれば、点は個体を表し、各次元は変数に対応し、点の座標は個体に対応する変数のもつ値である。この関係はグラフ[7]で表示すると理解しやすい。



c．次に問題となるのは、変数間の関連性、個体間の類似性、またはその差を測定することになる。これらは、距離という概念で測定されているが、距離の定義は分析手法によって極めて多くの試みが存在する。CA では、カイ2二乗（χ^2）距離で表される。

CA における距離の定義は、カイ二乗距離で定義される。行と列について、中心からの距離と正規化を考慮して、次の関係がある。

行 a_{ij} と列 b_{ij} のプロファイル（observed row and column profile）を $a_i=n_{i+}/n$, $b_j=n_{+j}/n$ とする。ただし、分割表（n×p）の各セルの値を n_{ij} とし、表全体の合計を n とする。表の周辺度数は、行については n_{i+} とし、列については n_{+j} とする。また、行と列の平均プロファイル（average row and column profile），$r_i=\Sigma_j n_{ij}/n$, $c_j=\Sigma_i n_{ij}/n$ とすると、各行と列の χ^2−distance を $\{(a_{ij}-c_j)^2/c_j\times r_i\}^{1/2}$,$\{(b_{ij}-r_i)^2/r_i\times c_j\}^{1/2}$ と定義し、これらの式を n 倍すると、元の表の χ^2 になる。これから実際に計算する場合には、次式を用いる。

$$\mathbf{S}=\frac{p_{ij}-p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}$$

S を行列表示すれば　$\mathbf{S}=\mathbf{D}_r^{-1/2}(\mathbf{P}-\mathbf{rc}^T)\mathbf{D}_c^{-1/2}$ と書いて、**S** の特異値分解を用いて、

$$\mathbf{S}=\mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$$
$$\mathbf{U}^T\mathbf{U}=\mathbf{V}^T\mathbf{V}=\mathbf{I}$$

となる。更に、

$$\mathbf{S}\mathbf{S}^T=\mathbf{U}\Lambda\mathbf{U}^T$$
$$\mathbf{S}^T\mathbf{S}=\mathbf{V}\Lambda\mathbf{V}^T$$
$$\mathbf{U}^T\mathbf{U}=\mathbf{I}$$
$$\mathbf{V}^T\mathbf{V}=\mathbf{I}$$
$$\mathbf{D}^2_\alpha=\Lambda$$

なる関係がある。

CAは多変量のデータの元の位置関係を出来るだけ保存するような2,3次元の位置を求める手法である。その手段として特異値分解を利用する。

$\mathbf{S}=\mathbf{U\Lambda V}^{T}$ において、$\mathbf{U}$ と $\mathbf{V}$ は左、右の特異ベクトルで r 個の直交列をもつ行列である。$\mathbf{\Lambda}$ は、$\delta_1 \geq \cdots \geq \delta_r > 0$ の大きさをもつ特異値の対角行列である。また、$\mathbf{S}=\mathbf{U\Lambda V}^{T}=\Sigma^{r}_{k=1}\delta_k u_k v^{T}_k$ と展開すると、k=2 に対応する特異ベクトルと最初から2つの特異値を用いて、

$$\hat{\mathbf{S}}=[(\mathbf{u}_1\ \ \mathbf{u}_2)]\begin{bmatrix}\delta_1 & 0\\ 0 & \delta_2\end{bmatrix}[(\mathbf{v}_1\ \ \mathbf{v}_2)]^{T}$$

とすると、$\hat{S}$ は S のランク2の最小2乗近似である。この適合度は、$(\delta_1+\delta_2)/\Sigma\ \delta_k$ である。上式を $\hat{\mathbf{S}}=\mathbf{FG}^{T}=[\delta_1{}^{\alpha}u_1,\delta_2{}^{\alpha}u_2][\delta_1{}^{1-\alpha}v_1,\delta_2{}^{1-\alpha}v_2]^{T}$ と分解する。但し、$\delta_i{}^2=\lambda_i(\lambda_i=\delta_i{}^{1/2})$ なる関係があり、$\Lambda=\mathrm{diag}(\lambda_i)$

ここで、α=0,1 として、

α=0 のとき、

$\mathbf{F}=[\mathbf{u}_1,\mathbf{u}_2],\mathbf{G}=[\delta_1\mathbf{v}_1,\delta_2\mathbf{v}_2]\Rightarrow$

$\mathbf{F}=\mathbf{D}_r{}^{-1/2}\mathbf{U}\ (=\mathbf{V}),\ \mathbf{G}=\mathbf{D}_c{}^{-1/2}\mathbf{V\Lambda}(=\hat{\mathbf{F}})$

α=1 のとき、

$\mathbf{F}=[\delta_1\mathbf{u}_1,\delta_2\mathbf{u}_2],\mathbf{G}=[\mathbf{v}_1,\mathbf{v}_2]\Rightarrow$

$\mathbf{F}=\mathbf{D}_c{}^{-1/2}\mathbf{U\Lambda}(=\hat{\mathbf{V}}),\mathbf{G}=\mathbf{D}_r{}^{-1/2}\mathbf{V}(=\mathbf{F})$

どちらも、standard coordinates と principal coordinates を組み合わせた座標でグラフ化されて、表示されることが多い[8]。

6．まとめ

多変量解析は多くの変数を処理する統計的現象を取り扱っている。多変量回帰分析に代表される統計モデルは、母集団が抽出された標本からモデルのパラメータを推定する構造を持っているが、もう一方では、データから母集団の構造を発見するデータマイニング（data mining）の性格を持っている。データ解析の立場からみると、現代では、企画された調査資料を分析する以外に、大量の複雑なデータが、グローバル化した世界の多方面の機関から急速に集まる事態になっている。このような環境に適応する統計分析と、それを支える計算能力の整備が必要となる。多次元のデータを縮約する correspondence analysis の手法も、そのような局面に向いているものである。探索的多変量解析の諸手法は多次元から縮約された次元でのグラフ表示によるパターン認識を目的としている。このような観点からみると現在の統計理論の展開が理解できるように感じる。R の package である vegan, ade 4 など ecology，環境問題で、Ordination を分析する方法や、概念には統計分析に重要な視点を与えるものであろう。

参考文献

1）M. Greeanncre（2007）Correspondence Analysis in Practice, Chapman & Hall/CDC

2）Julian Izenman （2008） Modern Multivariate Statistical Technuques, Spriger

3）P. Legendre and L. Legendre （2000） Numerical Ecology, 2nd, ed, Elsevier

4）ヴェナブルス、リプリー（2009）S-PLUS による統計解析、第2版、Springer

5）F. Cox and A. Cox （2000） Multidimensional Scaling, 2nd, ed, Chapman & Hall/CDC

6）L. Rizzo （2008） Statistical Computing with R, Chapman & Hall/CDC

7）S. Dray & Dufour The ade 4 Package:Implementing the Duality Diagram for Ecologists, Journal of Statistical Software 2007/9, vol. 22

注

1）データ行列 Y(Q 4)　X(Demographic Variable) (Dv)

	Q4A	Q4B	Q4C	age	sex	school	revenue	size
1	2	4	4	1	2	4	1	3
2	1	1	4	3	1	3	1	3
3	4	2	2	3	1	3	1	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
513	3	2	3	3	2	4	1	3
514	2	2	3	2	1	4	1	3

Y は、各質問は 5 つの選択肢を持つので、binary data に変換してから分析する。X はそのまま用いることにする。R による Program は、CCA は、Package, vegan, ade 4 に装備されている。どちらも Ecological, Environmental Data の分析のために開発されているが、vegan の方が使いやすいようである。

```
>library(vegan)  #package を load する。
>Q 4<-make.dummy(Q 4) #Q 4 を binary data に変換する。
>colnames(Q 4)<-c("Q 4 A 1","Q 4 A 2","Q 4 A 3","Q 4 A 5","Q 4 B 1","Q 4 B 2",Q 4 B 3","Q 4 B 4",
"Q 4 B 5","Q 4 C 1","Q 4 C 2","Q 4 C 3","Q 4 C 4","Q 4 C 5")  #列名を入力する。
>Q 4.cca<-cca(Q 4～.,data=Dv)  #CCA を実行する。
> Q 4.cca
Call: cca(formula=Q 4～age+sex+school+revenue+size,data=Dv)
              Inertia Rank
Total         4.00000
Constrained   0.07979    5
Unconstrained 3.92021   12
Inertia is mean squared contingency coefficient
Eigenvalues for constrained axes:
  CCA 1   CCA 2   CCA 3   CCA 4   CCA 5
0.045532 0.018290 0.009373 0.004532 0.002066
Eigenvalues for unconstrained axes:
 CA 1  CA 2  CA 3  CA 4  CA 5  CA 6  CA 7  CA 8  CA 9  CA 10  CA 11  CA 12
0.5651 0.4881 0.4017 0.3951 0.3353 0.3141 0.2917 0.2691 0.2482 0.2257 0.2036 0.1825
>summary(Q 4.cca) #実行結果の要約
Call:
cca(formula=Q 4～age+sex+school+revenue+size,data=Dv)

Partitioning of mean squared contingency coefficient:

              Inertia  Proportion
Total         4        1
Constrained   0.0798   0.01995
Unconstrained 3.9202   0.98005

Eigenvalues, and their contribution to the mean squared contingency coefficient

Importance of components
                        CCA 1   CCA 2   CCA 3   CCA 4   CCA 5   CA 1   CA 2
Eigenvalue              0.0455  0.01829 0.00937 0.00453 0.00207 0.565  0.488
Proportion Explained    0.0114  0.00457 0.00234 0.00113 0.00052 0.141  0.122
Cumulative Proportion   0.0114  0.01596 0.0183  0.01943 0.01995 0.161  0.283

                        CA 3    CA 4    CA 5    CA 6    CA 7    CA 8   CA 9
Eigenvalue              0.402   0.3951  0.3353  0.3141  0.292   0.2691 0.248
Proportion Explained    0.1     0.0988  0.0838  0.0785  0.073   0.0673 0.062
Cumulative Proportion   0.384   0.4824  0.5663  0.6448  0.718   0.785  0.847
```

	CA 10	CA 11	CA 12
Eigenvalue	0.2257	0.2036	0.1825
Proportion Explained	0.0564	0.0509	0.0456
Cumulative Proportion	0.9035	0.9544	1

Accumulated constrained eigenvalues
Importance of components:

	CCA 1	CCA 2	CCA 3	CCA 4	CCA 5
Eigenvalue	0.0455	0.0183	0.00937	0.00453	0.00207
Proportion Explained	0.5706	0.2292	0.11747	0.05679	0.02589
Cumulative Proportion	0.5706	0.7998	0.91732	0.97411	1

Scaling 2 for species and site scores
* Species are scaled proportion to eigenvalues
* Sites are unscaled weighted disperion equal on all dimensions

Species scores
$\hat{F} = V\Lambda$ v.eig

	CCA 1	CCA 2	CCA 3	CCA 4	CCA 5	CA 1
Q 4 A 1	-0.35346	0.05955	-0.0027	-0.08609	0.00915	0.4556
Q 4 A 2	0.12179	-0.05349	-0.07522	0.079332	0.011174	0.1946
Q 4 A 3	0.32119	-0.08247	0.036012	0.065144	-0.07885	-0.5396
Q 4 A 4	0.50033	0.10067	0.244628	-0.07088	-0.01341	-1.6023
Q 4 A 5	0.20528	-0.06162	0.540342	-0.09823	0.098612	-4.2173
Q 4 B 1	0.03148	-0.31201	-0.00691	-0.07459	-0.01353	0.5611
Q 4 B 2	-0.06576	0.04012	-0.02645	0.018921	-0.00136	0.2215
Q 4 B 3	0.24548	0.21943	0.167311	-0.02007	0.029959	-0.4344
Q 4 B 4	0.15883	0.3143	-0.05312	0.055736	0.010525	-1.8554
Q 4 B 5	-2.23261	0.2027	0.893445	0.748165	0.008633	-3.4883
Q 4 C 1	-0.14651	-0.33211	0.184358	0.120885	0.057144	0.8712
Q 4 C 2	-0.06478	0.05063	-0.0473	0.00645	-0.09318	0.6207
Q 4 C 3	-0.04016	-0.01531	0.046622	0.006756	0.027018	0.157
Q 4 C 4	0.11969	0.03986	-0.08741	-0.02782	0.078758	-0.6652
Q 4 C 5	0.16521	0.0542	0.204437	-0.09611	-0.07057	-2.0882

Site scores (weighted averages of species scores)
$\hat{V} = D(p_{i+})^{-1/2}\hat{U}$ wa

	CCA 1	CCA 2	CCA 3	CCA 4	CCA 5	CA 1
1	2.93059	5.47956	-7.6727	7.8885	16.2084	-1.52171
2	-1.48098	-3.87464	-3.4503	-13.8656	12.001	0.339322
3	2.7071	3.48861	6.0768	-3.3476	-17.4167	-0.3164
513	1.57596	-1.05094	1.9981	6.6804	-8.5818	0.008163
514	0.11614	-0.52277	-1.9576	7.724	5.9434	0.195311

Site constraints (linear combination of constraining variables)
U　u

	CCA 1	CCA 2	CCA 3	CCA 4	CCA 5	CA 1
1	-0.537821	-0.36253	-1.15306	1.47148	0.93249	-1.52171
2	1.122727	-0.11634	0.244144	-0.013	-1.32906	0.339322
3	1.122727	-0.11634	0.244144	-0.013	-1.32906	-0.3164
513	0.508765	1.349149	-1.32594	0.18715	-0.20937	0.008163
514	0.079623	-0.94701	-1.12211	0.21716	-0.90837	0.195311

Biplot scores for constraining variables

	biplot					
	CCA 1	CCA 2	CCA 3	CCA 4	CCA 5	CA 1
age	0.55931	0.58501	-0.0338	-0.39163	-0.4364	0
sex	-0.02369	0.65603	-0.1229	0.24751	0.7019	0
school	-0.90336	0.23345	-0.1745	-0.16138	-0.2701	0
revenue	-0.83116	0.28181	0.3545	-0.06936	-0.315	0
size	-0.22772	-0.09697	0.2528	-0.79167	0.4981	0

2）> library(vegan)
This is vegan 1.17-4
> Dv<-data.frame(Dv)　#X(=Dv)を data.frame にする。
> anova(Q 4.cca)　　#Model の分散分析を行う。
Permutation test for cca under reduced model
Model: cca(formula=Q 4～age+sex+school+revenue+size,data=Dv)

	Df	Chisq	F	N.Perm	Pr(>F)
Model	5	0.0798	2.0517	199	0.005**
Residual	504	3.9202			

Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
回帰モデルは有意であることが分かる。
> anova(Q 4.cca,by=”term”,step=200) #各変数毎に分散分析を行う。
Permutation test for cca under reduced model
Terms added sequentially (first to last)

Model: cca(formula = Q 4～age+sex+school+revenue+size,data=Dv)

	Df	Chisq	F	N.Perm	Pr(>F)
age	1	0.0216	2.7773	199	0.005**
sex	1	0.0097	1.2429	199	0.230
school	1	0.0337	4.3276	199	0.005**
revenue	1	0.0102	1.3110	199	0.245
size	1	0.0047	0.5999	199	0.790
Residual	504	3.9202			

Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
変数 age と school が有意である。これから回帰モデルとしては、

Q 4～age+school が選択される。

> anova(Q 4.cca,by="margin",perm=500) #統計実験を行ってモデルを探す。

Permutation test for cca under reduced model(Permutation test については、参考文献[3]に詳細な例がある。)

Marginal effects of terms

Model: cca(formula = Q 4～age+sex+school+revenue+size,data=Dv)

	Df	Chisq	F	N.Perm	Pr(>F)
age	1	0.0163	2.0896	199	0.005**
sex	1	0.0107	1.3694	99	0.140
school	1	0.0128	1.6441	499	0.074.
revenue	1	0.0100	1.2863	99	0.190
size	1	0.0047	0.5999	99	0.830
Residual	504	3.9202			---

Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1

この結果は、変数として、age のみとする。

> anova(Q 4.cca,by="axis",perm=1000) #変数軸について調べる。

Permutation test for cca under reduced model

Model: cca(formula = Q 4～age+sex+school+revenue+size,data=Dv)

	Df	Chisq	F	N.Perm	Pr(>F)
CCA 1	1	0.0455	5.8538	199	0.005**
CCA 2	1	0.0183	2.3515	99	0.170
CCA 3	1	0.0094	1.2051	99	0.730
CCA 4	1	0.0045	0.5826	99	0.980
CCA 5	1	0.0021	0.2656	99	0.990
Residual	504	3.9202			

Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1

第 1 軸のみが有意である。

3）例として、次のようなプログラムを組み合わせる。ここでは、変数 age についてグラフを表示しているが、変数名を変更すれば、各変数について同様の分析ができる。

```
>attach(Dv)
>plot(Q 4.cca,disp="sites",type="n")
>ordihull(Q 4.cca,age,col="blue")  #convexhull のグラフを描く。
>ordiellipse(Q 4.cca,age,col=3,lwd=2) #楕円形を描く。
>ordispider(Q 4.cca,age,col="red") #中心より各点への直線を描く。
>points(Q 4.cca,disp="sites"pch=21,col="red",bg="yellow",cex=1.3)
```

4）> Q 4.cca.fit<-envfit(Q 4.cca～.,data=Dv,perm=1000)#変数の Biplot を計算
> Q 4.cca.fit
***VECTORS

	CCA 1	CCA 2	r 2	Pr(>r)
age	0.944038	0.329838	0.0420	0.000999***
sex	-0.418106	0.908398	0.0230	0.005994**
school	-0.965910	0.258879	0.1086	0.000999***
revenue	-0.958399	0.285433	0.0958	0.000999***
size	-0.999805	-0.019755	0.0058	0.226773

Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
P values based on 1000 permutations.
> plot(Q 4.cca,dis=”sites”)
> plot(Q 4.cca.fit) #Biplot のグラフを描く。
> attach(Dv)
> ordisurf(Q 4.cca,school,add=TRUE)
Loading required package: mgcv
vegan の関数 envfit は cca の biplot に類似した結果をもたらすようである。また、関数 ordisurf は変数の smooth surfaces を示す。

5）> mod 1<-cca(Q 4～.,Dv) #すべての変数を含むモデル。
> mod 0<-cca(Q 4～1,Dv) #定数項のみ含むモデル。
> mod<-step(mod 0,scope=formula(mod 1),test=”perm”)
Start: AIC=1269.3
Q 4～1

	Df	AIC	F	N.Perm	Pr(>F)
+ school	1	1266.3	4.9639	199	0.005**
+ revenue	1	1266.9	4.3954	199	0.005**
+ age	1	1268.5	2.7584	199	0.005**
<none>		1269.3			
+ sex	1	1270.1	1.1882	99	0.310
+ size	1	1270.5	0.8249	99	0.630

Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
Step: AIC=1266.34
Q 4～school

	Df	AIC	F	N.Perm	Pr(>F)
+ age	1	1266.3	2.0786	199	0.015*
<none>		1266.3			
+ revenue	1	1267.1	1.2617	99	0.250
+ sex	1	1267.1	1.2338	99	0.250
+ size	1	1267.7	0.6292	99	0.760
-school	1	1269.3	4.9639	199	0.005**

```
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
Step: AIC=1266.26
Q 4～school+age
             Df   AIC      F      N.Perm  Pr(>F)
<none>           1266.3
-age         1   1266.3  2.0786   199    0.040*
+ sex        1   1267.0  1.2922    99    0.200
+ revenue    1   1267.0  1.2687    99    0.200
+ size       1   1267.7  0.5923    99    0.830
-school      1   1268.5  4.2768   199    0.005**
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
> mod #最適なモデル。
Call: cca(formula = Q 4～school+age,data=Dv)
              Inertia Rank
Total         4.00000
Constrained   0.05488   2
Unconstrained 3.94512  12
Inertia is mean squared contingency coefficient

> modb<-step(mod 1,scope=list(lower=formula(mod 0),
+                upper=formula(mod 1),trace=0))  #model 0 と model 1 の範囲で最適なモデル
(modb)を探すプログラムを実行する。
>modb
Call: cca(formula = Q 4～age+school,data=Dv)
              Inertia Rank
Total         4.00000
Constrained   0.05488   2
Unconstrained 3.94512  12
Inertia is mean squared contingency coefficient
> modb$anova
      Step  Df   Deviance  Resid. Df  Resid. Dev    AIC
1           NA     NA        -6       5997.916  1269.026
2  -size     1   7.138631    -5       6005.055  1267.633
3  -revenue  1  15.601322    -4       6020.656  1266.956
4  -sex      1  15.374695    -3       6036.031  1266.257
最適モデルは、いずれの計算でも同じである。
```

6）

```
> Ddata 1[1:2,] #参考までに元のデータセットの一部を表示する。
  Q 4 A  Q 4 B  Q 4 C  age  sex  school  revenue  size
1   2      4      4     1    2     4       1       3
2   1      1      4     3    1     3       1       3
> attach(Ddata 1)
>subset(Ddata 1,age==1&school==6&revenue==2)
```

	Q 4 A	Q 4 B	Q 4 C	age	sex	school	revenue	size
58	5	5	5	1	2	6	2	7
63	2	3	3	1	1	6	2	3
105	2	2	2	1	1	6	2	2
115	1	1	2	1	1	6	2	5
146	2	5	3	1	1	6	2	3
165	2	4	4	1	2	6	2	6
177	1	2	4	1	2	6	2	7
189	2	1	1	1	1	6	2	6
250	4	2	3	1	2	6	2	6
251	2	2	2	1	1	6	2	6
268	2	2	2	1	2	6	2	6
269	1	2	2	1	1	6	2	6
330	1	2	1	1	2	6	2	7
332	1	2	3	1	2	6	2	7
356	3	3	5	1	1	6	2	7
387	2	4	2	1	2	6	2	7
453	2	4	2	1	2	6	2	5
462	1	2	1	1	2	6	2	4
502	1	5	4	1	2	6	2	4

> subset(A,Q 4 B==5)

	Q 4 A	Q 4 B	Q 4 C	age	sex	school	revenue	size
58	5	5	5	1	2	6	2	7
146	2	5	3	1	1	6	2	3
502	1	5	4	1	2	6	2	4

7）このグラフは文献 7 ）を利用している。

8）個々の記号 F, G は CCA の解説の記号 V, F と異なるが（=V）のように注釈をつけている。 実際には、R の package の manual の plot の scale の説明を参照されたい。

Statistical data analysis using correspondence analysis

ABSTRACT

Based on previous research, this research examines the applications of canonical correspondece analysis (CCA) of data using R. CCA is combined with correspondence analysis for regression analysis. Question variable is connected to the demographic variables to develop geometric data analysis in correspondence analysis. From the point of view of statistical theory, such analysis has many interesting implications. CCA has wide application in the field of numerical ecology, such as in Ordination analysis. It is hoped that an exploratory multivariate statistical analysis will contribute to the field of social science.

Key Words: correspondence analysis, canonical correspondence analysis, R