〈研究ノート〉

Biplot と多変量データ解析について*

中　山　慶　一　郎**

1.　このテーマを取り上げた理由

いままで取り上げてきた多変量データ解析では、多次元の変数で示されたデータを低次元の変数に縮小したデータに変換し、その分析結果をグラフに表示して、視覚化するものである。このような手法は多変量データ分析に共通するものである。これらの分析に共通する手法は、データ行列の特異値分解 Singular Value Decomposition であり、その分析結果をグラフ化したものが Biplot である。

本稿では、このようなデータ解析の手法である SVD と Biplot を取り上げて多変量データ解析の理論とその意味を説明する。

Biplot は、n×p のデータ行列 data matrix の列と行をグラフ表示したものであり、列 rows は個体 individual とか、標本単位 sample units を示し、列 columns は変数 variables を示す。Biplot は、Gabriel（1971）によって、PCA の文脈のなかで定義されたものが始まりであり、データ行列の行と列を、点 point とベクトル vector で、2 次元のグラフ表示するものである。ここでは、まず、SVD の説明による縮約された計算結果を Biplot で表示する。

次に、Biplot の考え方を述べ、その活用の仕方を示し、R による program と、R の package の利用によるデータ解析を示すことにする。

2.　Biplot の理論

まず、データ行列 n×p が与えられる。このデータの性質によって、多変量解析の分析手法が決まる。数値をとる量的データの分析で代表的なものは主成分分析 Principal Component Analysis であり、質的データの代表的なものは、対応分析 Correspondence Analysis である。この場合、データは 2 項変数 binary variable であり、度数 frequency を示すことも多い。いずれにせよ、他にも、数多くの分析モデルが知られているが、それらに共通していることは、多変数のデータ行列を少数の次元（多くは 2 つの次元）に縮約し、それをグラフ化（Biplot）して解釈することを、目的としている。ここで、用いられるのは、特異値分解 Singular Value Decomposition を用いてモデル化した元のデータ行列を分解し、biplot display により分析結果を解釈することである。

a.　SVD について

いま、分解するデータ行列を X とすれば、SVD は

$$X = U\Gamma V^T,\ U^TU = V^TV = I,\ \Gamma = \mathrm{diag}(\sigma_1 \gneqq \sigma_2 \gneqq \cdots \sigma_K > 0)$$

と定式化される。この分解が如何なる意味をもつか、いま、行列 $X(n \times p)$ が 2 つの行列 $A(n \times R)$ と、

---

*キーワード：バイプロット、特異値分解、主成分分析、**R**、**FactoMineR**、**BiplotGUI**

**関西学院大学名誉教授

$B^T(k \times p)$ の積で表される。(この積は一意ではない)

$$X = AB^T$$

これが SVD を用いて表すと、

$$A = U\Gamma^{\alpha} \quad B = V\Gamma^{1-\alpha}$$

と表現できる。$(0 \leqq \alpha \leqq 1)$ また、$U \in R^{n\times n}$ で $V \in R^{p\times p}$、かつ直交行列 orthogonal matirix $(U^TU = VV^VT = I)$ である。Γ は正の singular value の対角行列で、rank(A) = k の対角行列である。また、singular value は A^TA と AA^T の非負の固有値の平方根 $(\lambda = \sigma^2)$ でもある。

また、$x_{ij} = \Sigma\sigma_i u_{ik} v^T_{kj}$ となる。

いま、rank = 2 の場合

$$X = [u_1 u_2] \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} [v_1 v_2]^T$$

となる。

更に、これを、$x_{ij} = \Sigma^2_{i=1}(u_{ik}\sigma^{\alpha}_k)\Sigma^2_{k=1}(\sigma^{1-\alpha}_k v_{kj})$　$(0 \leqq \alpha \leqq 1)$ とする。

b. Biplot について、

$(\Sigma u_{ik}\sigma^{\alpha}_k)$ は row effect (n 個の標本の座標の値) を表し、$\Sigma\sigma^{1-\alpha}_k v_{kj}$ は column effect (p 個の変数の座標の値) を表すものとなる。グラフ上では、row effect を点 points で示し、column effects を vectors で表示することが多い。グラフの表現では、α の値によって、異なる座標の値が得られる。

$\alpha = 1$ のとき、(row-metric preserving)

標本間の距離のユークリッド距離を保持し、変数間の角度は変数間の関連性を表す。

$\alpha = 0$ のとき、(column-metric preserving)

変数の軸 axes 間の $\cos\theta_{kl}$ は変動間の相関を表す。又、原点からの変数間の距離は変動の大きさを示す。

$\alpha = 1/2$ のとき、(symmetric biplots)

標本と変数の変動の相対的大きさを示している。

ここで、

$$F = U\Gamma^{\alpha}$$
$$G = V\Gamma^{1-\alpha}$$

とすると、$X = FG^T$ となる。

以上整理すると、SVD の分解した結果は、3 つの型の座標に分けられる。

Standard coordinates : Rows : $F_s = U$、Columns : $G_s = V$

Principal coordinate, Rows : $F_p = F_s\Gamma$、Columns : $G_p = G_s\Gamma$

Canonical coordinate, Rows : $F_p = F_s\Gamma^{1/2}$、Columns : $G_e = G_s\Gamma^{1/2}$

例えば、PCA では $F_p = F_s\Gamma$、$G_s = V$ の組み合わせが Biplot として実行される。

c. Biplot の精度

行列 X における全変動性 d^2 は、

全変動性 $= SS_x = \| X \|^2 == \Sigma^G_{g=1}\Sigma^E_{c=1}x^2_{gc}$　g = 行の数、c = 列の数

で定義され、X、$\hat{X}$ は直交性をもつので、

$$d^2(X, \hat{X}) = \Sigma_g\Sigma_c(x_{gc} - \hat{x}_{gc})^2 = \| X - \hat{X} \|^2$$

註　行列 X を低次元で近似した時の行列 X との距離は、

$$d(X, X)^2 = \sqrt{\Sigma_g\Sigma_c(x_{gc} - \hat{x}_{gc})^2}$$

従って、$\| X \|^2 = \| \hat{X} \|^2 + \| X - \hat{X} \|^2$
これを、singular value で表すと、

$$\Sigma_{s=1}^{K} \sigma_s^2 = \Sigma_{s=1}^{2} \sigma_s^2 + \Sigma_{s=3}^{K} \sigma_s^2$$

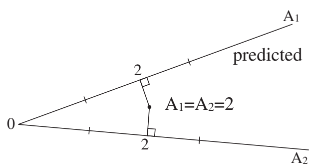
これを利用すると、biplot の精度は、plot された 2 つの軸の singular value によって示される。

$$q = \sigma_i^2 / \Sigma_j \sigma_j^2 \quad Q = q_1 + q_2$$
（ランク 2 の場合）

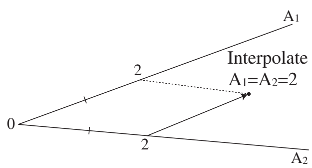
d. Biplot による分析結果の解釈について

多変量の分析結果をグラフに表示する時に用いられる Biplot では、標本 samples は、点 points で，変数 variables は、ベクトル vectors で表示される。Biplot による分析では、これらの点とベクトルによるグラフの説明が中心になる。各点の配置はデータ間の距離 distance を示し、その配置の形態はデータ間の興味ある情報をもたらす。ベクトル間の角度は変数間の相関を表すので、変数間の関連性を示すことになる。又、ベクトルの大きさは変数の分散を示し、変数の重要性を示す。グラフの上で、点とベクトルの線の関連を説明する幾何学的な用語として、interpolation と prediction が用いられる。Gower［1］によると、2 次元の biplot のグラフ上で両者を図示すると、

図（a）

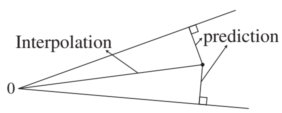


図（b）



図（a）では、点 P を指定すると、変数 $A_1 = 2$、$A_2 = 2$ が predict values となる。逆に点 P の prediction は、変数 $A_1 = 2$、$A_2 = 2$ の predict value を有し、その交わる角度は 90° である。（b）では、変数 A_2 のベクトル 2 と、変数 A_1 のベクトル 2 の vector–sum の点 P が得る。これを interpolation という。Interpolation と prediction との関連をグラフに表示すると、図（c）の様になる。

図（c）



データ分析の立場からは、prediction は、回帰分析において、回帰線と残差との関係に類似しているので、biplot において、変数（回帰線）と点（データ）との関連を示すものである。従って、適切な変数を回帰線として、適切な目盛付けして、各データ点からの predict value を推定することが出来る。Calibration による残差の表示をすることが可能である。Interpolation は、選択した点の関連がら、新しい変数を選択するので、点の選択はデータから推論する手法がのぞましい。この点から云うと。データから分析に応じて、グループにわけるこれまでの手法を利用するのがよいであろう。グループ分けについては、統計学的には、多変量分散分析、判別分析などのモデルが存在する。

註　Prediction については、次節で例を表示している。

3. Biplot のプログラミング

a. ここでは、Biplot の programing をデータが数値の場合である PCA（主成分分析）に即して、R をもちいて行うことにする。
X をデータ行列（37×7）Ocotea とする。

まず、データを読み込む
X<−Ocotea[,2 : 7] #pca の計算では、第 1 列は計算から除く。
データを変換する。
Xtr＜−scale(X,center＝TRUE,scale＝TRUE)　　#データの正規化をする。
Xtr＝(X−mean(X))/sd(X) と同じ。
X の SVD を行う。$X = U\Gamma V^T$
X<−Xtr
F_s<−svd(X)$u # Standard coordinate
G_s<−svd(X)$v #Standard coordinate
F_p<−F_s %*% diag(svd(X)$d)　#Pricipal coordinate
G_p<−diag(svd(X)$d) %*% Gs　　#Principal coordinate
F_c<−F_p %*% diag(1/sqrt(svd(X)$d))　#Canonical coordinate
G_c<−diag(1/sqrt(svd(X)$d)) %*% G_s　#Canonical coordinate
これらの座標を用いて biplot を作成するが、一般的に利用されているのは、
　Asymmetric map として、rows plot F_p と G_s（変数）が、PCA に用いられている。
columns plot F_s と G_p は、covariance biplot として、用いる。
一般に使われている関数 princomp を用いた program 例。
pca.results<−princomp(Xtr,cor＝TRUE)
F_s<−pca.results$scores
G_s<−pca.results$loadings
以下同様である。

b. Biplot を作成するには、soft により、また、分析者により同一ではない。一般に利用されている progu-ram には、より汎用性のある、加重 SVD を根拠にしているものと思われるので、weighted Singular Value Decomposition を述べることにする。それは、データ行列の行と列にウエイトになる対角行列 D_r、D_c を用い、

$D_r = (1/n)I$, $D_c = I$ とし、

weighted SVD は

$$D_r^{1/2} Z D_c^{1/2} = U\Lambda V^T, \ Z = X^T X \quad \Lambda = \Gamma^2$$

この分解定理を利用すると、

$$F_s = D_r^{-1/2} U, \qquad G_s = D_c^{-1/2} V$$
$$F_p = F_s\Gamma \qquad G_p = G_s\Gamma$$
$$F_c = F_s\Gamma^{1/2}, \qquad G_c = G_s\Gamma^{1/2}$$ とする。

以下は、Biplot の作成 program
F<−sqrt(nrow(X))*Fp　#F<−diag(1/sqrt(nrow(X))) %*% Fp
G<−sqrt(ncol(X))*Gs　#G<−diag(1/sqrt(ncol(X))) %*% Gs

```
plot(rbind(F[,1 : 2],G[,1 : 2]),type = ”n”,asp = 1,
+  xlim = c(−3.6,2.3),xlab = ”dim 1”,ylab = ”dim 2”,cex.axis = 0.7)
+  text(F[,1 : 2],labels = rownames(X),col = ”forestgreen”,cex = 0.7)
arrows(0,0,G[,1],G[,2],col = ”chocolate”,length = 0.1,angle = 1.0)
+ text(c(1.07,1.3,1.07,1.35,1.2,1.4)*G[,1],
+ c(1.07,1.07,1.05,1,1.16,1.1)*G[,2],
+ labels = colnames(X),col = ”chocolate”,font = 4,cex = 0.8）
```

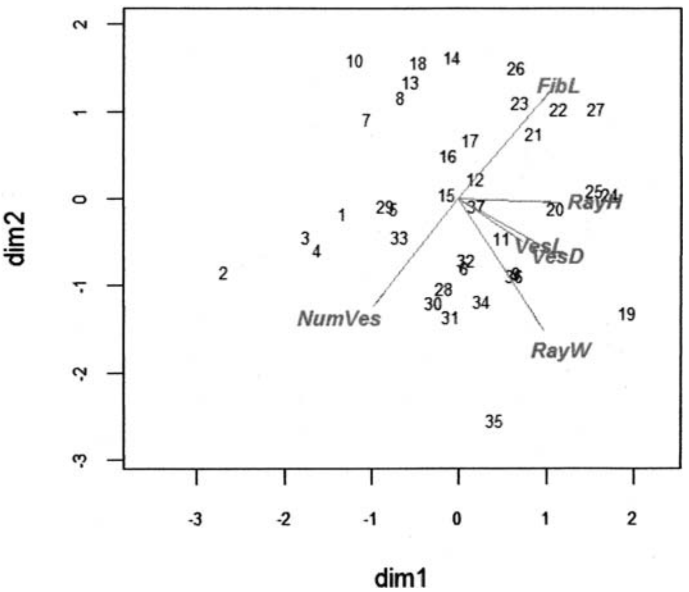
SVD 分解の精度は、次の program を実行する。

```
#scree plot
X_percents<−100*svd(X)$d^2/sum(svd(X)$d^2)
X_percents<−X_percents[seq(6,1)]
barplot(X_percents,horiz = TRUE,cex.axis = 0.7)
```

以下も同じ。

```
screeplot(pca.results)
summary(pca.results)
```

program による biplot の図は、以下の graph である。



一番簡単に biplot を作成するには、

```
pca,results<–princonmp(Xtr,cor = TRUE)
biplot(pca.results,cex = 0.7)
```

すればよい。

ここで利用したデータは次表 Ocotea である。Ocotea 南アメリカ産の植物で、ここでは 3 つの種からの 6 変数の測定値である。文献［5］，p 96.‘

Ocotea

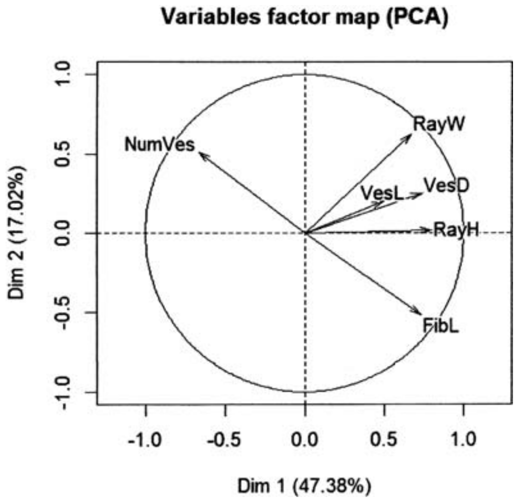
	Species	VesD	VesL	FibL	RayH	RayW	NumVes
1	Obul	79	383	941	333	30	17
2	Obul	78	346	961	223	24	31
3	Obul	82	361	1039	316	27	25
4	Obul	79	324	1048	369	29	26
5	Obul	85	418	1051	347	34	14
6	Obul	114	448	1096	379	40	13
7	Obul	76	320	1130	347	29	13
8	Obul	103	371	1165	326	26	10
9	Obul	129	406	1165	428	44	11
10	Obul	74	281	1175	324	26	11
11	Obul	102	567	1221	395	40	11
12	Obul	95	415	1225	416	38	10
13	Obul	91	372	1234	375	26	11
14	Obul	113	314	1253	466	23	10
15	Obul	93	541	1267	347	34	14
16	Obul	94	437	1271	336	36	10
17	Obul	119	359	1280	412	32	11
18	Obul	104	387	1290	381	22	12
19	Obul	114	569	1369	568	52	11
20	Obul	141	621	1527	419	34	15
21	Oken	147	402	1391	440	32	9
22	Oken	142	393	1468	443	35	6
23	Oken	125	322	1530	459	34	11
24	Oken	156	401	1588	512	42	11
25	Oken	162	502	1591	369	42	8
26	Oken	103	378	1655	441	34	11
27	Oken	126	414	1759	459	42	8
28	Opor	122	346	981	393	40	14
29	Opor	139	133	993	342	33	14
30	Opor	130	368	1005	356	39	16
31	Opor	127	331	1027	473	38	20
32	Opor	112	309	1044	358	47	8
33	Opor	115	352	1048	300	36	14
34	Opor	130	471	1072	409	39	15
35	Opor	153	419	1077	392	48	20
36	Opor	135	370	1104	531	38	15
37	Opor	130	325	1166	428	36	12

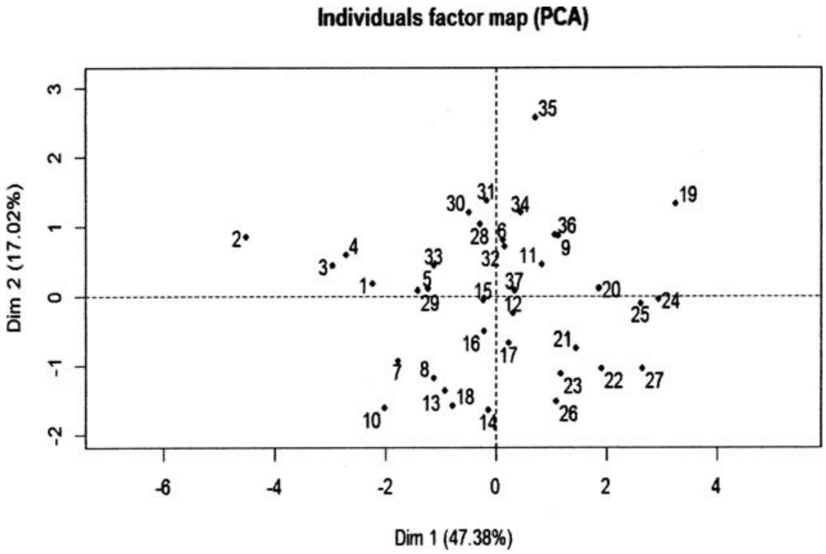
c. R の package を用いて同じデータで計算を実行する。ここで、利用した package は FactoMineR と BiplotGUI である。

```
>library(FactomineR)
>X<−Ocotea[,2 : 7]
>res.pca<−PCA(X)
```

この package では、変数と標本は別々にグラフ化されるが分かり易い、更に次節で述べる多くの機能を持つ。

註　FactoMineR については、文献［14］, BiplotGUI に「ついては、文献［3］を参照されたい。



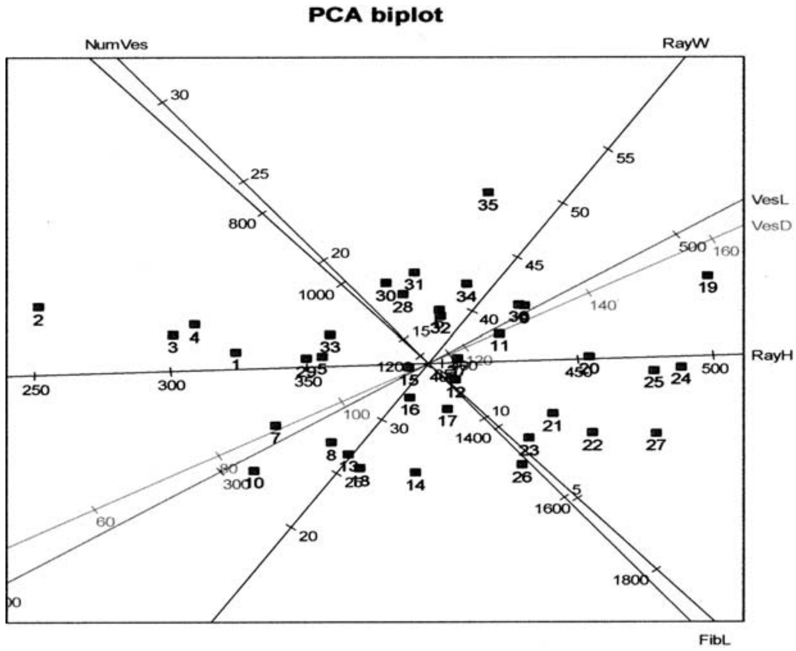


>Library(BiplotGUI)

>X<−Ocotea[,2 : 7]

>Biplots(Data＝X)

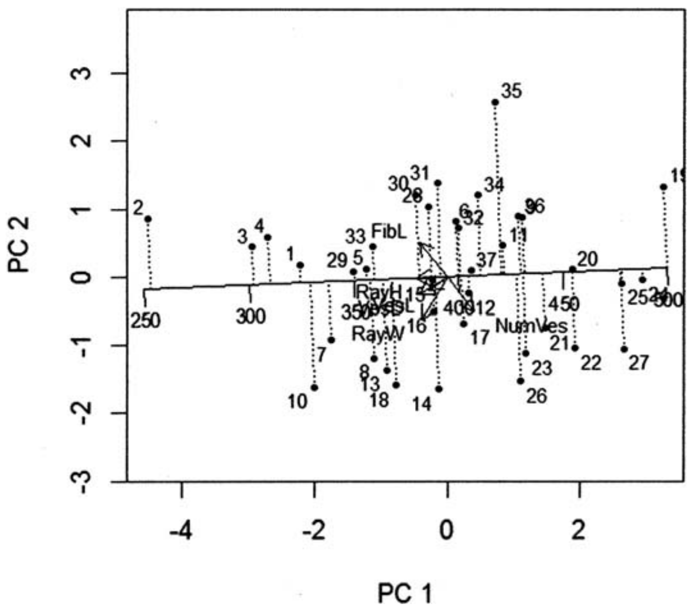
この package は比較的新しく多くの機能をもち、biplot 図は、目盛りが調整される calibration 機能をもち、更に多くの分析が可能である。



d.　ここで、package（calibrate）を用いて、prediction を求めてみる。ここでは、RayH 変数を軸として計算した。

```
library(calibrate)
> #data(Ocotea)
> X<−Ocotea[,2 : 7]
> pca.results<−princomp(X,cor = TRUE)
> Fp<−−(pca.results$scores)
> Gs<−pca.results$loadings
> plot(Fp[,1],Fp[,2],pch = 16,asp = 1,xlab = ”PC 1”,ylab = ”PC 2”,cex = 0.5)
> textxy(Fp[,1],Fp[,2],rownames(X),cex = 0.75) >
> arrows(0,0,Gs[,1],Gs[,2],length = 0.1)
> textxy(Gs[,1],Gs[,2],colnames(X),cex = 0.75)
> #Calibration of RayH
> ticklab<−seq(250,500,by = 50)
> ticklabc<−ticklab−mean(X[,4])
> yc<−(X[,4]−mean(X[,4]))
> g<−Gs[4,1 : 2]
> #g<−Fp[4,1 : 2]
> Calibrate.RayH<−calibrate(g,yc,ticklabc,Fp[,1 : 2],ticklab,lm = TRUE,
+                         tl = 0.25,dp = TRUE,labpos = 4,cex.axislab = 0.75,axislab = ”RayH”)
−−−−−−−−−−− Calibration Results for RayH −−−−−−−−−−−
Length of 1 unit of the original variable     = −0.0314
Angle                                         = 2.18 degrees
Optimal calibration factor                    = −0.0667
Used calibration factor                       = −0.0667
```

Goodness−of−fit　　　　　　＝0.6315
Goodness−of−scale　　　　　＝0.6315
−−−
>

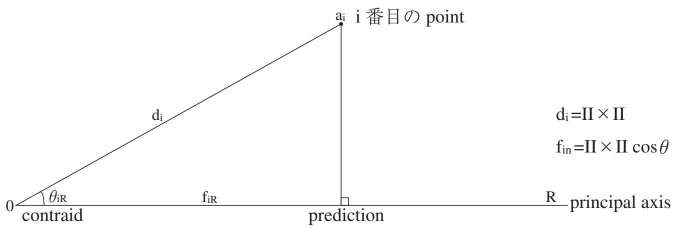


e. Biplot の解釈について

主成分分析の Biplot では、標本 individuals と、変数 variables の分析値を 2 次元の主成分軸上 principal axes に点 points と、ベクトル vectors で表示される。即ち、Fp（Scores）と Gs（Loadings）の値は principal axes の座標の値を示している。

Biplot から、我々は様々な情報が得られ分析結果の参考になる。個々の point はデータの位置を示し点間の距離と点の集合はデータのパターンを示している（Individual Study）．また、変数間の線形性が変数間の関連性を示し、変数の重要度がベクトルの大きさによって示される。又、付加した変数によるデータのグループ化が、点と変数の関係を理解するのに役立つ。

これらの諸点について中心的な役割を示すのに、inertia（分散）という概念である。Inertia とは、分散と同意であり、元のデータ（n×p）を縮約した新しい n×k のグラフについて計算される。この関係をグラフに表すと、



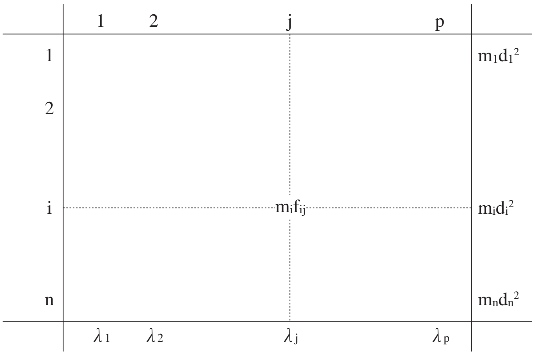
となる。図から各点の total inertia,Σλ は、
total inertia＝multidimensional variance＝データとその中心との距離を示している。

point の total inertia は，$\Sigma_i m_i d_i^2 = \Sigma_i m_i \Sigma_k f_{ik}^2 = \Sigma_k \lambda_k$

i 番目の点の inertia は、$m_i d_i^2 = m_i \Sigma_k f_{ik}^2$

i 番目の点の k 主成分に対する inertia contribution は $m_i f_{ik}^2$ である。

即ち、



これらの結果からデータ observations、変数 variables の計算値である scores と loadings の各成分 conponent に対する貢献度 contribution と cos squared が計算される。（ここでは、ウエイト m＝1）

この計算は、$ct_{r_{i,l}} = \dfrac{f_{il}^2}{\lambda_l}$　$ctr_{jl} = \dfrac{g_{jl}^2}{\lambda_l}$

$$\cos_{il}^2 = \frac{f_{il}^2}{d_{rj}^2} \quad \cos_{jl}^2 = \frac{g_{jl}^2}{d_{ij}^2}$$

これを実際に求めるには、R の package である FactoMineR を用いるのが簡単である。

```
>library(FactoMoneR)
>X<−Ocotea[,2 : 7]
>res.pca<−PCA(X)
>res.pca$ind$coord  #scores,Fp を求める。
>res.pca$ind$cos 2  #cos2 を表示
>res.pca$ind$contrib  #data の貢献度
>res.pca$var#coord  #loadings(Fp)
>res.pca$var$cos 2  #cos2 の表示
>res.pca$var$contrib  #variable の貢献度
>summary(res.pca)  #計算結果が要約されて出力される。
```

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Variance	2.843	1.021	0.932	0.536	0.436	0.232
%	47.384	17.015	15.538	8.93	7.262	3.871
Cumulative	47.384	64.399	79.938	88.867	96.129	100

Individuals（the 10 first）

	Dist	Dim.1	ctr	cos 2	Dim.2	ctr	cos 2
1	2.393	−2.231	4.73	0.869	0.18	0.086	0.006
2	4.96	−4.507	19.312	0.826	0.855	1.936	0.03
3	3.217	−2.948	8.26	0.84	0.44	0.514	0.019
4	3.192	−2.7	6.931	0.716	0.592	0.928	0.034
5	1.652	−1.227	1.431	0.552	0.115	0.035	0.005
6	1.142	0.128	0.016	0.012	0.808	1.728	0.5
7	2.143	−1.76	2.946	0.675	−0.927	2.276	0.187
8	1.877	−1.118	1.188	0.355	−1.182	3.702	0.397
9	1.591	1.121	1.194	0.497	0.868	1.993	0.298
10	2.716	−2.012	3.846	0.549	−1.61	6.864	0.352

Variables

	Dim.1	ctr	cos 2	Dim.2	ctr	cos 2
VesD	0.736	19.075	0.542	0.251	6.175	0.063
VesL	0.497	8.676	0.247	0.201	3.97	0.041
FibL	0.731	18.778	0.534	−0.517	26.196	0.267
RayH	0.794	22.202	0.631	0.018	0.032	0
RayW	0.67	15.774	0.448	0.624	38.079	0.389
NumVes	−0.664	15.495	0.441	0.511	25.548	0.261

4. Grouping による Biplot Analysis

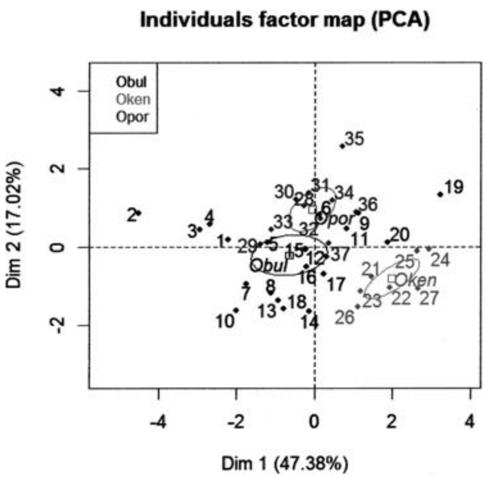
a. Biplot によるデータ分析で、補助的変数を用いて、個々のデータ点をグループ別に表しデータの性質を示す手法を述べることにする。主成分軸上に示されたデータ点 individual points の集合はデータの性質を示すものである。一般的なものは，点をグループに分けて表すことである。R の package である FactoMineR では、補助的変数 supplementary variable をデータに組み込んで、グループ別に点を表示することが出来る。ここでは、この機能を利用してグループ別のデータ解析を行い、更に、統計モデルを利用した分析モデルを紹介する。

```
<library(FactoMineR)
<−X<−Ocotea
<−res.pca<−PCA(X,quali.sup = 1)      #データの第 1 行は 3 つの種 species から構成される。
<−plot(res.pca,habillage = 1)         #グループ別にグラフ表示する。
<−aa<−cbind.data.frame(X[,1],respca$ind$coord）
<−bb<−coord.ellipse(aa,bary = TRUE)
<−plot.PCA(res.pca,habillage = 1,ellipse = bb)
```

ここでの補助的変数 supplementary variabls は Obul、Oken、Opur なる三つの Species 種である。この変数は、PCA の individuals の座標の値には影響を与えず、主成分と補助的変数との相関係数に相当する座標の値を与えるがけである。

$$G_s(k') = \frac{1}{\sqrt{\lambda_s}} \Sigma x_{ik} \qquad F_{s(i)} = r(k, F_s)$$

これは、グラフの点の座標の位置で表されている。このデータでは、Species による 3 つのグループによって全体のデータが分類されることは明白である。

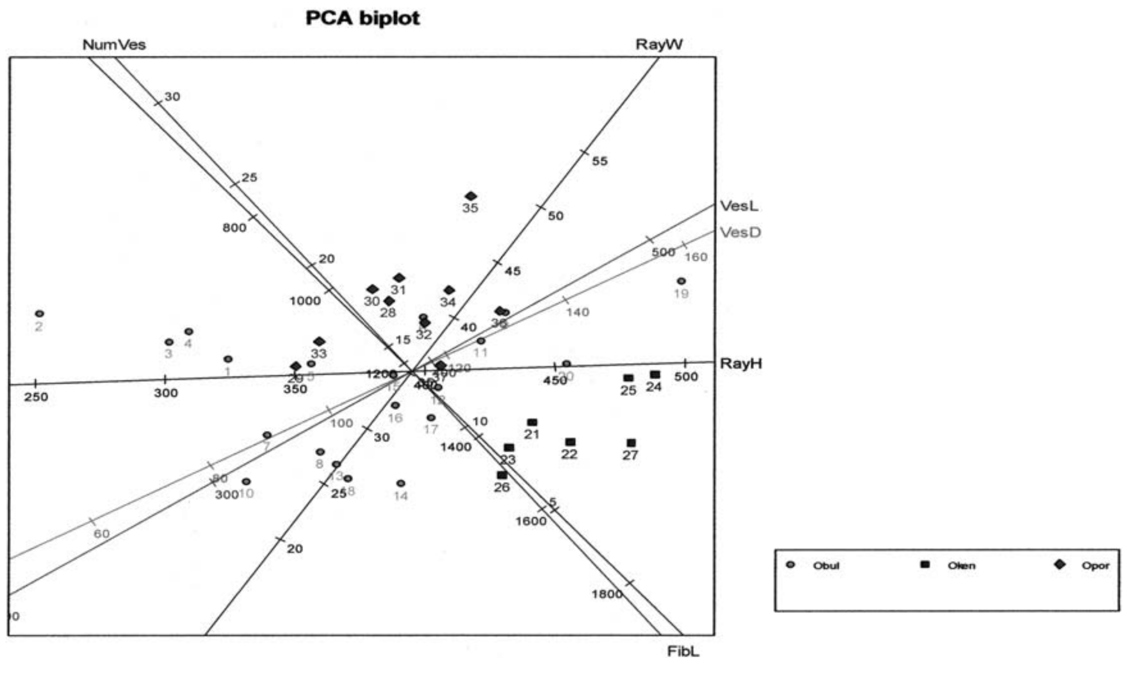


b. 最後に biplot のグラフにグループ別データによる分類を示す R の package、BiplotGUI による結果を示す。この package では、単なる biplot を表示すると云うより、統計モデルによる分類データの表示が示される。ここでは CVA（Canonical Variate Analysis）による計算結果を示している。これは、MANOVA 多変量分散分析を用いたもので、他にも判別分析 Discrimanal Analysis を用いても同様の結果が生じる。

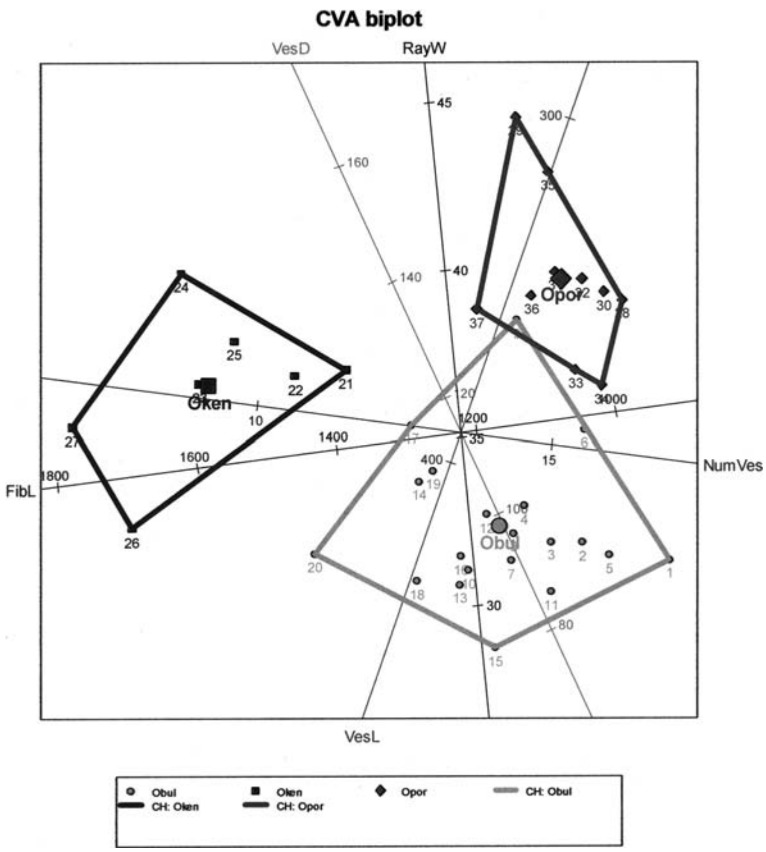
<−library(BiplotGUI)

<− attach(Ocotea)

<−Biplots(Data = Ocotea[,−1],group = Spesies)



この図では、PCA biplot については、FactoMineR と同じであるが CVA biplot になると、Species の差が明白に図示される。この図では、各グループを凸包 convex hull で示したのが、cva biplot である。



$Obul

	Dim 1	Dim 2
15	0.077389	−0.48468
20	−0.32999	−0.27175
17	−0.11309	0.017059
9	0.125329	0.254673
1	0.470639	−0.28684
15	0.077389	−0.48468

$Oken

	Dim 1	Dim 2
26	−0.74325	−0.21425
27	−0.87673	0.013615
24	−0.63247	0.3595
21	−0.25987	0.14128
26	−0.74325	−0.21425

$Opor

	Dim 1	Dim 2
34	0.318871	0.106227
33	0.259026	0.141141
37	0.036069	0.278893
29	0.12465	0.712025
35	0.197423	0.588725
28	0.365086	0.298923
34	0.318871	0.106227

補論　実際の分析に際しては、データの変換、分類されたデータの分布、その統計量が分析結果の解釈に役立つので、その program と、例のデータの結果を表示する。

データの整備

X<−Ocotea[,2 : 7]

Xtr<−scale(X,center = TRUE,scale = TRUE)

Ocotea 1<−cbind(Ocotea[,1],Xtr)

Group 別ヒストグラムの作成、

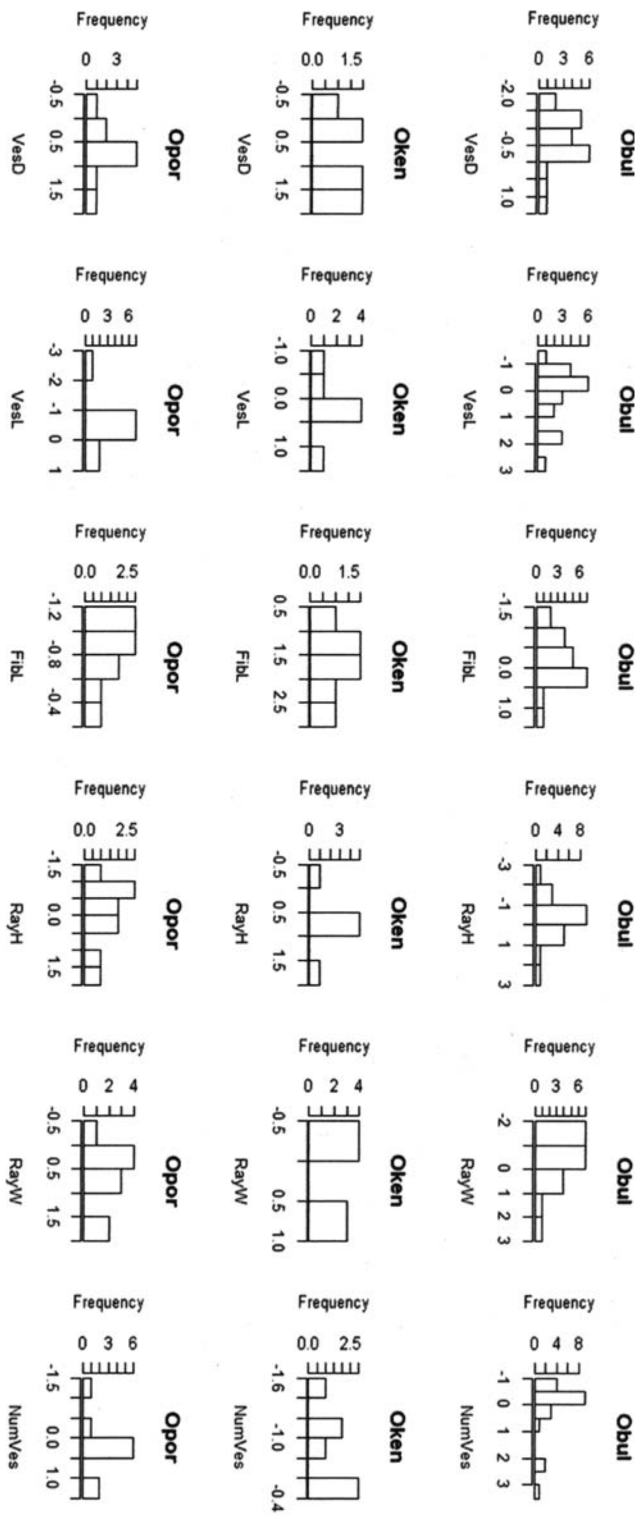
par(mfrow = c(3,6))

for (j in levels(Ocotea 1$Species)) {

　Ocotea 2<−Ocotea 1[Ocotea 1$Species = = j,2 : 7]

for (i in 1 : 6){

```
hist(Ocotea 2[,i],xlab = names(Ocotea 2[,i]),main = j) }
 }
```



Group 別統計量の作成。

Mean

> aggregate(・～Species,data＝Ocotea 1,mean)

	Species	VesD	VesL	FibL	RayH	RayW	NumVes
1	Obul	−0.6433358	0.2319666	−0.1701059	−0.28615257	−0.4013268	0.160839
2	Oken	0.9484135	0.1170189	1.6213491	0.74903788	0.2977603	−0.8259899
3	Opor	0.6227821	−0.5458464	−0.7947326	0.04797862	0.5942213	0.2565149

SD

> aggregate（・～Species,data＝Ocotea 1,sd）

	Species	VesD	VesL	FibL	RayH	RayW	NumVes
1	Obul	0.769701	1.0523701	0.6559068	1.0121786	1.0833577	1.1499561
2	Oken	0.8375674	0.5983739	0.564975	0.6173082	0.6307957	0.3734806
3	Opor	0.480962	0.9808166	0.2613274	0.980923	0.6616548	0.6738195

まとめ

今回とりあげた点は、多変量データ解析の分析の根拠となる特異値分解 SVD の視点と、その分析結果の表現 Biplot によって、データ解析を理解する点にある。この点は多くの多変量データ解析の理論的根拠となる。

このような分析が如何に役立つか、実際の分析者の立場に立つと取り上げたデータの分析目的に対する包括的な見方を与えるものである。この点から云えば探索的多変量解析という言葉がその本質を表している。

元のデータ行列の個々の標本 individual のグラフ上の位置はデータの相互関連を表し、変数の位置は変数間の重要性と主軸との関係を示すものとなる。これらの情報は問題となるデータの背景を探るための情報である。

文献

1　Gabriel, The biplot graphic display of matrices with application to principal component analysis, Biometrics（1971）, 58, 3

2　Grange, Le Roux and Gardner-Lube, BiplotGUI : Interative Biplots in R, Journal of Statistical Software（2009）60, 12

3　Udina, Interactive Biplot Construction, Journal of Statistical Software（2005）, 13, 5

4　Gower, JC, The geometry of biplot scaling（2004）, Biometrika, 91

5　Gower, Lubbe and le Roux, Understanding Biplots（2011）, Wiely

6　HBrand, PCA and CVA biplots : Astudy of their underlying theory and quality measures, Stellenbosch university（2013）

7　Gower, Unified Biplot Geometry, Ljubljiana ; FDV（2003）

8　Amenta, Interpolative and Predictive Biplots applied to Co-Inertia Analysis, Electron. Journal of Applied Statistical Analysis,（2013）vol.6, 2

9　M. Greenacre, Biplots in Practice, Fundation BBVA（20010）

10　S-Plus による統計解析（第 2 版）, W. N. Venables and B. D. Ripley, Springer（2002）

11　Jan Graffelman. A Guide to Scatterplot and Biplot Calibration. （2012）

12　W. Hardle, L. Simar, Applied Multivariate Statistical Analysis, Springer（2003）

13　Jan Graffelman. A Guide to catterplot and Biplot Calibration（2012）

14　F. Husson, S Le, and J Pages, Exploratory Multivariate Analysis by Example Using R. CRC Press（2011）

15　H. Abdi and J. Williams, Principal component analysis, WIREs Comp Stat（2010）, Jhon Wiley

16　U. Kohler and M. Luniak, Data inspection using biplots. Stata Journal（2005）5, Number 2
17　BiplotGUI : Interactive Biplots in R, Grange, Roux and Gardner-Lubbe, Journal of Statistical Software.（2009）, vol 30, 12

Biplot and Multivariate Statistical Data Analysis

ABSTRACT

Biplots are useful graphical representations of multidimensional data for displaying the rows and columns of a data matrix. Any element of the matrix is represented by the inner product of the vectors corresponding to its rows columns by the singular value decomposition (SVD) of a matrix.

This paper demonstrates principal component analysis (PCA) and SVD routines and biplots applied as graphical results using R programs.

Key Words: biplot, singular value decomposition (SVD), PCA, R, FactoMineR, Biplot GUI