OPEN SPAT

Erasmus+

INSTITUTO SUPERIOR D AGRONOMIA
Hinc Patriam Sustinet
Universidade de Lisboa

Pattern recognition on spatial data

Exploring and visualizing : Principal Component Analysis

Pr Yves Brostaux, GxABT, University of Liege (Belgium)

OpenSpat, Gembloux

June 2018

Gembloux Agro-Bio Tech
Université de Liège

Contents

Gembloux Agro-Bio Tech
Université de Liège

Contents

Introduction

Objective

- summarize data
- display data

$\Rightarrow$ simultaneous study of the relashionships between p symetrical variables (no prior internal causality)

Introduction

Data

- Raw data : matrix of numerical data
  - n rows $\equiv$ individuals
  - p columns $\equiv$ variables
- Standardized data : for each variable j
  - $x_{ij} = (y_{ij} - \bar{y}_j)/\hat{\sigma}_j$
  - $\bar{x}_j = 0$
  - $\hat{\sigma}_{x_j} = 1$

Example I - simple classical data

Raw data : mammal's milk (16 mammals)

- y_{1i} : protein (percentage)
- y_{2i} : fat (percentage)
- y_{3i} : lactose (percentage)

Raw data

	Name	Prot	Fat	Lact
a	Donkey	1.7	1.4	6.2
b	Whale	11.1	21.2	1.6
c	Deer	10.4	19.7	2.6
d	Sheep	5.6	6.4	4.7
e	Buffalo	5.9	7.9	4.7
f	Camel	3.5	3.4	4.8
g	Guinea pig	7.4	7.2	2.7
h	Horse	2.6	1.0	6.9
i	Llama	3.9	3.2	5.6
j	Rabbit	12.3	13.1	1.9
k	Mule	2.0	1.8	5.5
l	Rat	9.2	12.6	3.3
m	Fox	6.6	5.9	4.9
n	Reindeer	10.7	20.3	2.5
o	Pig	7.1	5.1	3.7
p	Zebra	3.0	4.8	5.3

Descriptive statistics

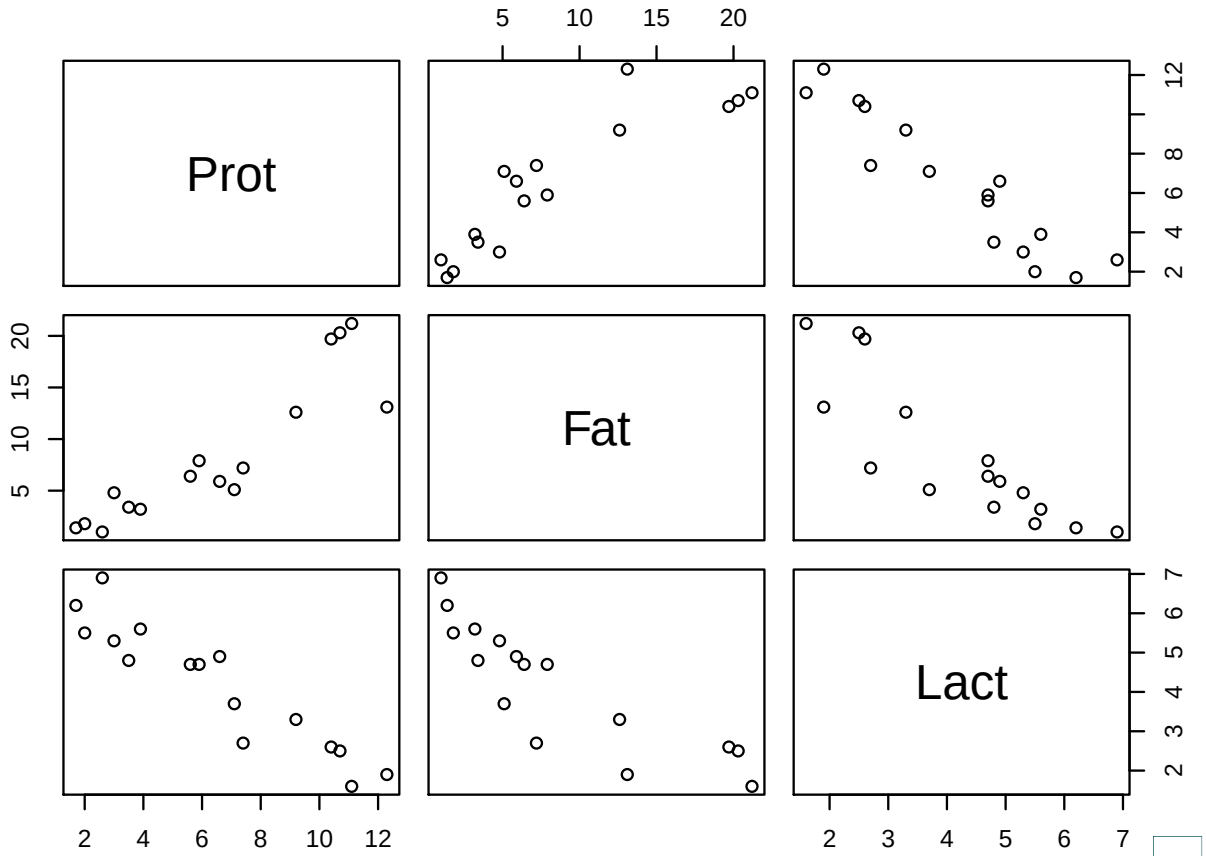
Univariate statistics

Variable	Mean	StDev
Prot	6.44	3.50
Fat	8.44	6.87
Lact	4.18	1.60

Correlation matrix

	Prot	Fat	Lact
Prot	1.00000	0.89731	-0.93845
Fat	0.89731	1.00000	-0.86543
Lact	-0.93845	-0.86543	1.00000

Matrix plot



Gembloux Agro-Bio Tech
Université de Liège

Standardized data

	Name	Prot	Fat	Lact
a	Donkey	-1.35362	-1.02435	1.26329
b	Whale	1.33219	1.85766	-1.61529
c	Deer	1.13218	1.63932	-0.98951
d	Sheep	-0.23929	-0.29657	0.32462
e	Buffalo	-0.15358	-0.07824	0.32462
f	Camel	-0.83931	-0.73324	0.3872
g	Guinea pig	0.27501	-0.18013	-0.92694
h	Horse	-1.09647	-1.08257	1.70134
i	Llama	-0.72502	-0.76235	0.88782
j	Rabbit	1.67506	0.67865	-1.42756
k	Mule	-1.2679	-0.96613	0.82525
l	Rat	0.78931	0.60588	-0.55147
m	Fox	0.04643	-0.36935	0.44978
n	Reindeer	1.2179	1.72666	-1.05209
o	Pig	0.18929	-0.48579	-0.30116
p	Zebra	-0.98218	-0.52946	0.70009

Contents

Gembloux Agro-Bio Tech
Université de Liège

First component

$$z_{1i} = u_{11}x_{i1} + u_{21}x_{i2} + u_{31}x_{i3}$$

with $u_{11}^2 + u_{21}^2 + u_{31}^2 = 1$
variance of $\mathbf{z}_1$ maximum

Solution

$u_{11} = -0.585 \quad u_{21} = -0.569 \quad u_{31} = 0.578$
Variance $= 2.80$

First component

Solution

$u_{11} = -0.585 \quad u_{21} = -0.569 \quad u_{31} = 0.578$

Variance $= 2.80$

Value of z_{11} (donkey)

$$x_{11} = -1.354 \quad x_{12} = -1.024 \quad x_{13} = 1.263$$

$$\begin{aligned} z_{11} &= (-0.585)(-1.354) + (-0.569)(-1.024) \\ &\quad + (0.578)(1.263) \\ &= 2.105 \end{aligned}$$

Second component

$$z_{2i} = u_{12}x_{i1} + u_{22}x_{i2} + u_{32}x_{i3}$$

with $u_{12}^2 + u_{22}^2 + u_{32}^2 = 1$

$u_{11}u_{12} + u_{21}u_{22} + u_{31}u_{32} = 0$

variance of $\mathbf{z}_2$ maximum

Solution

$u_{12} = 0.233 \quad u_{22} = -0.801 \quad u_{32} = -0.552$

Variance $= 0.14$

Third component

$$z_{3i} = u_{13}x_{i1} + u_{23}x_{i2} + u_{33}x_{i3}$$

with
$$u_{13}^2 + u_{23}^2 + u_{33}^2 = 1$$
$$u_{11}u_{13} + u_{21}u_{23} + u_{31}u_{33} = 0$$
$$u_{12}u_{13} + u_{22}u_{23} + u_{32}u_{33} = 0$$
variance of $\mathbf{z}_3$ maximum

Solution

$u_{13} = -0.777 \quad u_{23} = -0.188 \quad u_{33} = 0.601$

Variance $= 0.06$

Principal components or z-scores

Synthetic indexes as

$$z_{ji} = u_{1j}x_{i1} + u_{2j}x_{i2} + \ldots + u_{pj}x_{ip} \quad (1)$$

$$u_{1j}^2 + u_{2j}^2 + \ldots + u_{pj}^2 = 1 \quad (2)$$

$$u_{1j}u_{1k} + u_{2j}u_{2k} + \ldots + u_{pj}u_{pk} = 0, \forall j \neq k \quad (3)$$

$$\text{variance of } z_j \text{ maximum} \quad (4)$$

Computing scores

R : correlation matrix (rank $r \leq p$)

- $l_1 \geq l_2 \geq \ldots \geq l_r$: eigenvalues of R
  - $\rightarrow$ solution of $|\mathbf{R} - l_i\mathbf{I}| = 0$
  - $\rightarrow$ equal to the **variances** of the corresponding components
- $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$: eigenvectors of R
  - $\rightarrow$ solution of $(\mathbf{R} - l_i\mathbf{I})\mathbf{u}_i = 0$

Computing scores - example

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.897 & -0.938 \\ 0.897 & 1.000 & -0.865 \\ -0.938 & -0.865 & 1.000 \end{bmatrix}$$

$$l_1 = 2.801 \quad l_2 = 0.142 \quad l_3 = 0.057$$

$$\mathbf{U} = \begin{bmatrix} -0.585 & 0.233 & 0.777 \\ -0.569 & -0.801 & -0.188 \\ 0.578 & -0.552 & 0.601 \end{bmatrix}$$

Note : signs of u_i are arbitrary

Gembloux Agro-Bio Tech
Université de Liège

Computing scores

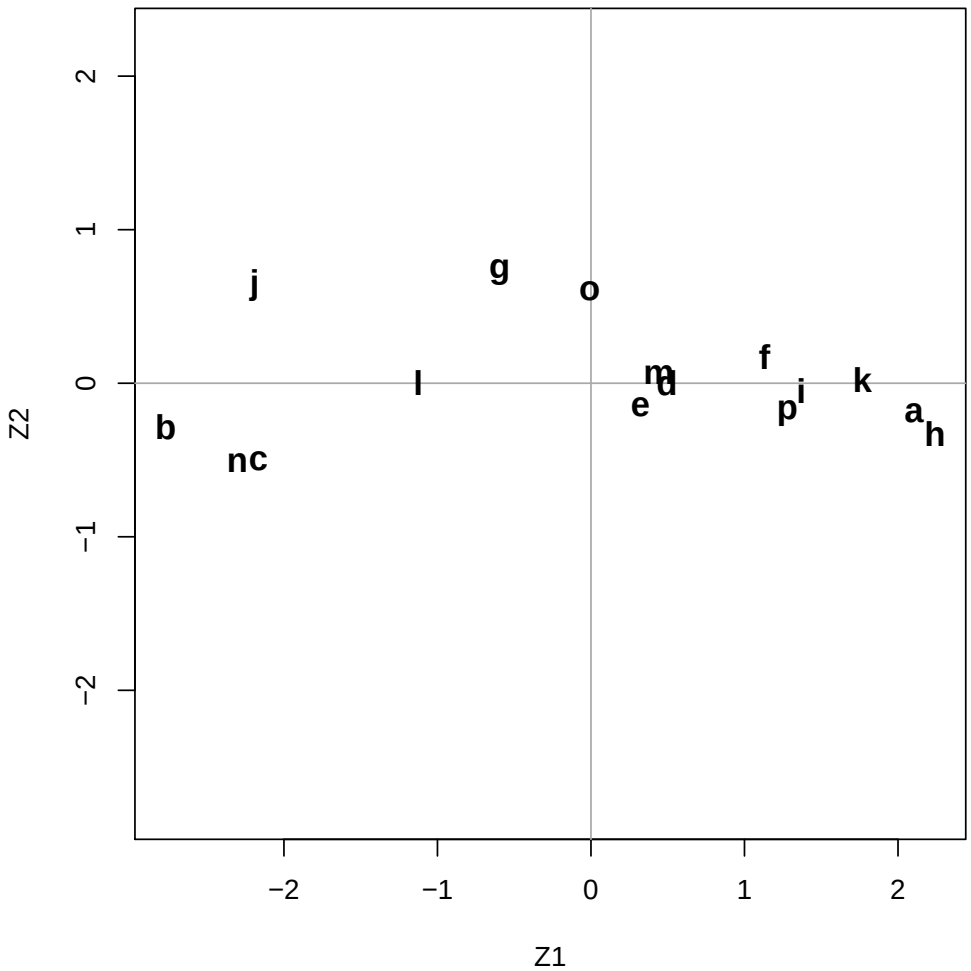
$$\left.\begin{aligned} \mathbf{z}_1 &= \mathbf{X}\mathbf{u}_1 \\ \mathbf{z}_2 &= \mathbf{X}\mathbf{u}_2 \\ &\vdots \\ \mathbf{z}_r &= \mathbf{X}\mathbf{u}_r \end{aligned}\right\} \mathbf{Z} = \mathbf{X}\mathbf{U}$$

Gembloux Agro-Bio Tech
Université de Liège

Principal components or z-scores

	Name	Z1	Z2	Z3
a	Donkey	2.10479	-0.19321	-0.10011
b	Whale	-2.76998	-0.28459	-0.28459
c	Deer	-2.16701	-0.5019	-0.02292
d	Sheep	0.49637	0.00238	0.06487
e	Buffalo	0.32199	-0.1524	0.09045
f	Camel	1.13191	0.17736	-0.28168
g	Guinea pig	-0.59417	0.72003	-0.30952
h	Horse	2.2408	-0.32837	0.37385
i	Llama	1.37106	-0.04899	0.11342
j	Rabbit	-2.19098	0.63564	0.31606
k	Mule	1.76829	0.02198	-0.30768
l	Rat	-1.12515	0.00359	0.16804
m	Fox	0.44307	0.05825	0.37573
n	Reindeer	-2.30301	-0.51727	-0.01033
o	Pig	-0.00833	0.59930	0.05734
p	Zebra	1.28036	-0.19182	-0.24293

Graphical result - Z_1 & Z_2



Gembloux Agro-Bio Tech
Université de Liège

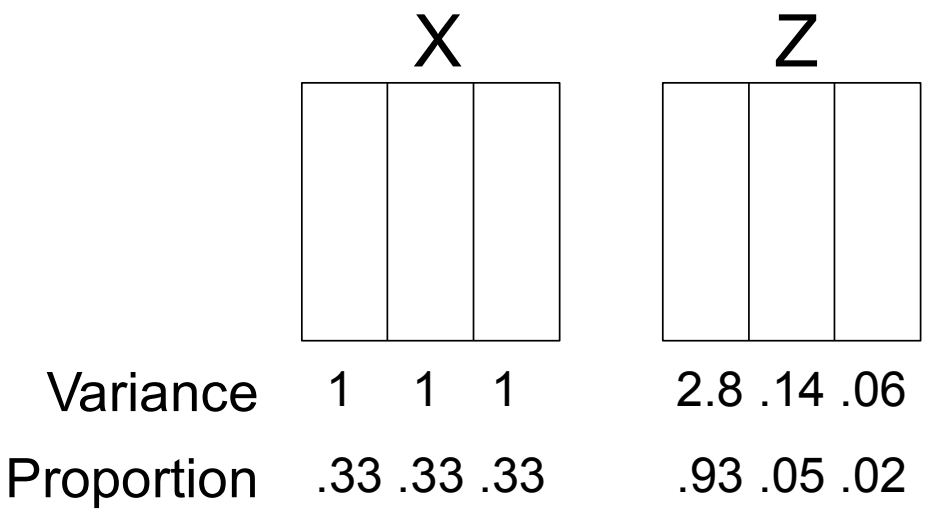
PCA

```
# PCA using FactoMineR package
##########################################

# package loading
library(FactoMineR)

# PCA on scaled data (default)
mammi.pca <- PCA(mammi[-1])

# stored informations in PCA object
mammi.pca
```

PCA

	name	description
1	"$eig"	"eigenvalues"
2	"$var"	"results for the variables"
3	"varcoord"	"coord. for the variables"
4	"varcor"	"correlations variables - dimensions"
5	"varcos2"	"cos2 for the variables"
6	"varcontrib"	"contributions of the variables"
7	"$ind"	"results for the individuals"
8	"indcoord"	"coord. for the individuals"
9	"indcos2"	"cos2 for the individuals"
10	"indcontrib"	"contributions of the individuals"
11	"$call"	"summary statistics"
12	"$call$centre"	"mean of the variables"
13	"$call$ecart.type"	"standard error of the variables"
14	"$call$row.w"	"weights for the individuals"
15	"$call$col.w"	"weights for the variables"

Raw variables and scores

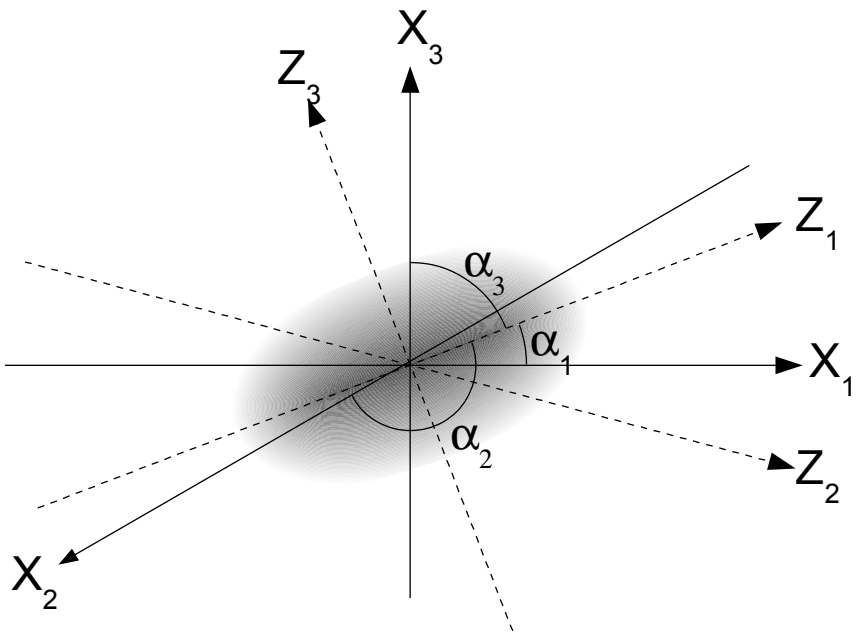


X : correlated variables
same importance
(same variance)

Z : non correlated variables
decreasing importance

Gembloux Agro-Bio Tech
Université de Liège

Geometrical meaning of principal components

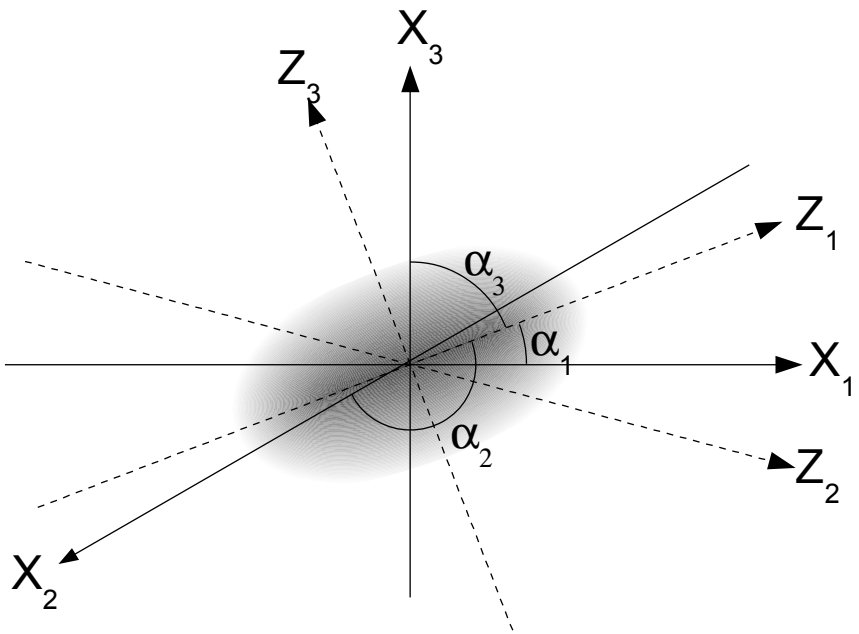


Projection of the original data in a new coordinates system with axis

- of decreasing importance (variance)
- non correlated (orthogonal)

Gembloux Agro-Bio Tech
Université de Liège

Geometrical meaning of principal components



$$u_{11} = \cos \alpha_1$$

$$u_{21} = \cos \alpha_2$$

$$u_{31} = \cos \alpha_3$$

$$u_{11}^2 + u_{21}^2 + u_{31}^2 = 1$$

l_1 is the variance of the scores along the new axis Z_1

Gembloux Agro-Bio Tech
Université de Liège

Correlation between z and x

$$r(x_i, z_j) = u_{ij}\sqrt{l_j}$$

	z_1	z_2	z_3
x_1	-0.979	0.088	0.186
x_2	-0.952	-0.301	-0.045
x_3	0.968	-0.208	0.144

$$\begin{aligned}
r(x_1, z_1) &= u_{11}\sqrt{l_1} &&= -0.585\sqrt{2.8} &&= -0.979\\
r(x_2, z_1) &= u_{21}\sqrt{l_1} &&= -0.569\sqrt{2.8} &&= -0.952\\
r(x_3, z_1) &= u_{31}\sqrt{l_1} &&= 0.578\sqrt{2.8} &&= 0.968\\
r(x_1, z_2) &= u_{12}\sqrt{l_2} &&= 0.233\sqrt{0.14} &&= 0.088\\
\text{etc.}
\end{aligned}$$

Gembloux Agro-Bio Tech
Université de Liège

Correlation

```
# Correlation between scaled variables and Z-scores
cor(X, Z)

# With FactoMineR
mammi.pca$var$cor
dimdesc(mammi.pca, proba=1)
```

Z-scores as a decomposition of X

U : orthonormal matrix : $U'U = UU' = I$

$$\begin{aligned} X &= XUU' = ZU' \quad \text{or} \quad X = ZU' \\ X &= z_1 u_1' + z_2 u_2' + \cdots + z_r u_r' \\ X &= X_1 + X_2 + \cdots + X_r \end{aligned}$$

Z-scores as a decomposition of X - example

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3$$

$$\begin{bmatrix} -1.3536 & -1.0243 & 1.2633 \\ \vdots & \vdots & \vdots \\ -0.9822 & -0.5295 & 0.7001 \end{bmatrix} = \begin{bmatrix} -1.2307 & -1.1978 & 1.2168 \\ \vdots & \vdots & \vdots \\ -0.7487 & -0.7286 & 0.7402 \end{bmatrix}$$

$$+ \begin{bmatrix} -0.0451 & 0.1547 & 0.1066 \\ \vdots & \vdots & \vdots \\ -0.0448 & 0.1536 & 0.1059 \end{bmatrix}$$

$$+ \begin{bmatrix} -0.0778 & 0.0188 & -0.0602 \\ \vdots & \vdots & \vdots \\ -0.1887 & 0.0456 & -0.146 \end{bmatrix}$$

Gembloux Agro-Bio Tech
Université de Liège

First component ($\mathbf{X}_1 = \mathbf{z}_1\mathbf{u}_1'$)

	Prot1	Fat1	Lact1
	-1.2307	-1.1978	1.2168
	1.6197	1.5764	-1.6014
	1.2671	1.2332	-1.2528
	-0.2902	-0.2825	0.287
	-0.1883	-0.1832	0.1861
	-0.6619	-0.6442	0.6544
	0.3474	0.3381	-0.3435
	-1.3103	-1.2752	1.2954
	-0.8017	-0.7803	0.7926
	1.2811	1.2469	-1.2666
	-1.034	-1.0063	1.0223
	0.6579	0.6403	-0.6505
	-0.2591	-0.2521	0.2561
	1.3467	1.3106	-1.3314
	0.0049	0.0047	-0.0048
	-0.7487	-0.7286	0.7402
SSQ	14.366	13.608	14.043

Information recovery

$$14.366/15 = .96 = r^2(x_1, z_1) = (-.979)^2$$
$$13.608/15 = .91 = r^2(x_2, z_1) = (-.952)^2$$
$$14.043/15 = .94 = r^2(x_3, z_1) = (.968)^2$$

$$(14.366 + 13.608 + 14.043)/(15 + 15 + 15) = .93 = \frac{l_1}{3}$$

Gembloux Agro-Bio Tech
Université de Liège

Second component ($\mathbf{X}_2 = \mathbf{z}_2\mathbf{u}_2'$)

	Prot2	Fat2	Lact2
	-0.0451	0.1547	0.1066
	-0.0664	0.2278	0.1571
	-0.1171	0.4018	0.2770
	0.0006	-0.0019	-0.0013
	-0.0356	0.1220	0.0841
	0.0414	-0.1420	-0.0979
	0.1680	-0.5764	-0.3974
	-0.0766	0.2629	0.1813
	-0.0114	0.0392	0.0270
	0.1484	-0.5088	-0.3509
	0.0051	-0.0176	-0.0121
	0.0008	-0.0029	-0.002
	0.0136	-0.0466	-0.0321
	-0.1207	0.4141	0.2855
	0.1399	-0.4798	-0.3308
	-0.0448	0.1536	0.1059
SSQ	0.1158	1.3618	0.6474

Information recovery

$$0.1158/15 = .01 = r^2(x_1, z_2) = (.088)^2$$
$$1.3618/15 = .09 = r^2(x_2, z_2) = (-.301)^2$$
$$0.6474/15 = .04 = r^2(x_3, z_2) = (-.208)^2$$

$$(0.1158 + 1.3618 + 0.6474)/(15 + 15 + 15) = .05 = \frac{l_2}{3}$$

Gembloux Agro-Bio Tech
Université de Liège

First two components ($\mathbf{X}_1 + \mathbf{X}_2$)

	Prot12	Fat12	Lact12
	-1.2758	-1.0432	1.3235
	1.5533	1.8042	-1.4443
	1.1500	1.6350	-0.9757
	-0.2897	-0.2844	0.2856
	-0.2239	-0.0612	0.2703
	-0.6205	-0.7861	0.5565
	0.5155	-0.2383	-0.7409
	-1.3869	-1.0124	1.4767
	-0.8131	-0.7410	0.8197
	1.4295	0.7380	-1.6175
	-1.0289	-1.0239	1.0101
	0.6588	0.6374	-0.6524
	-0.2455	-0.2988	0.2240
	1.2259	1.7247	-1.0459
	0.1447	-0.4750	-0.3356
	-0.7934	-0.5751	0.8461
SSQ	14.482	14.970	14.690

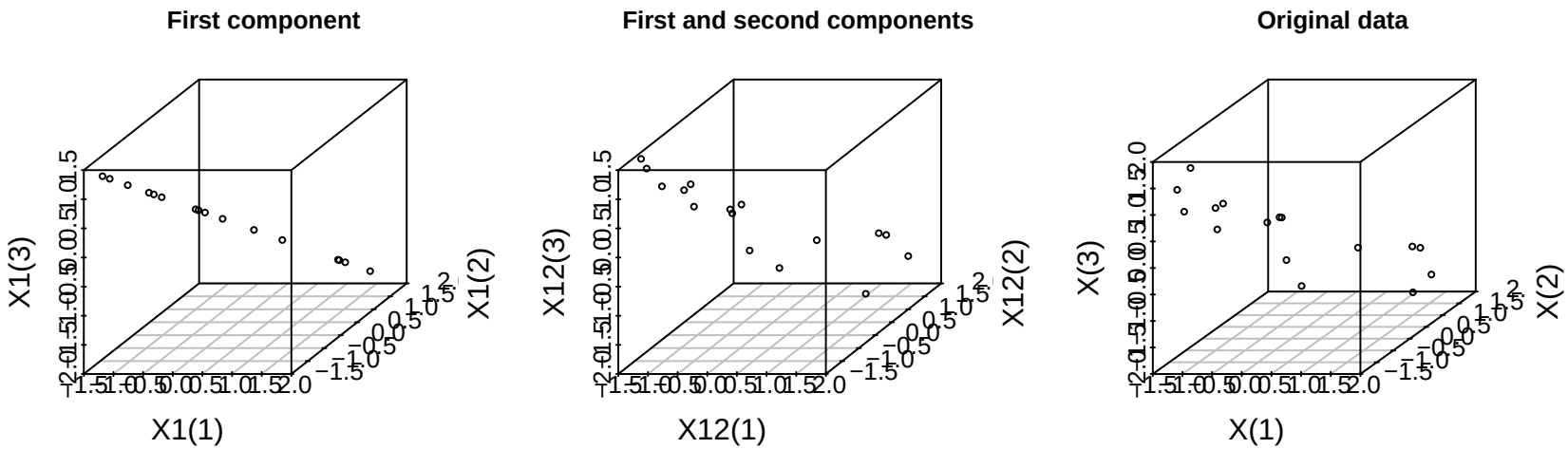
Information recovery

$$
\begin{aligned}
14.482/15 = \quad 0.97 &= r^2(x_1, z_1) + r^2(x_1, z_2) \\
&= \quad 0.96 \quad + \quad 0.01 \\
14.970/15 = \quad 1.00 &= r^2(x_2, z_1) + r^2(x_2, z_2) \\
&= \quad 0.91 \quad + \quad 0.09 \\
14.690/15 = \quad 0.98 &= r^2(x_3, z_1) + r^2(x_3, z_2) \\
&= \quad 0.94 \quad + \quad 0.04
\end{aligned}
$$

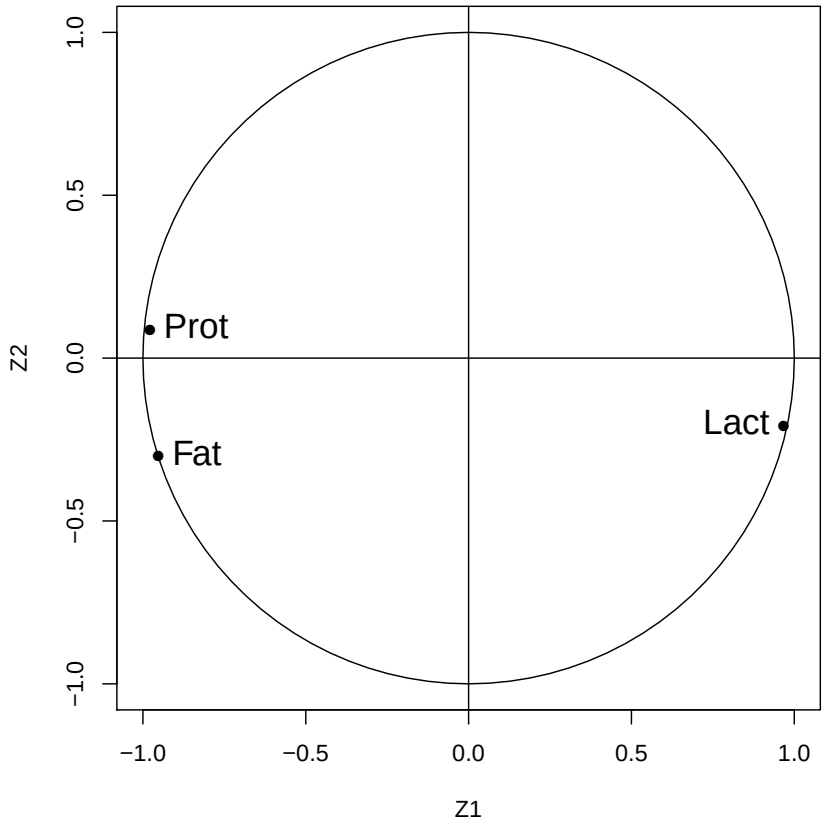
$$
\begin{aligned}
(14.482 + 14.970 + 14.690)/45 &= .98 \\
&= \frac{l_1 + l_2}{3}
\end{aligned}
$$

Gembloux Agro-Bio Tech
Université de Liège

Scatterplots



Plot of the variables (correlation circles)



- coordinates : coefficients of correlation

	z_1	z_2
Prot	-0.979	0.088
Fat	-0.952	-0.301
Lact	0.968	-0.208

- quality : sum of squared coordinates

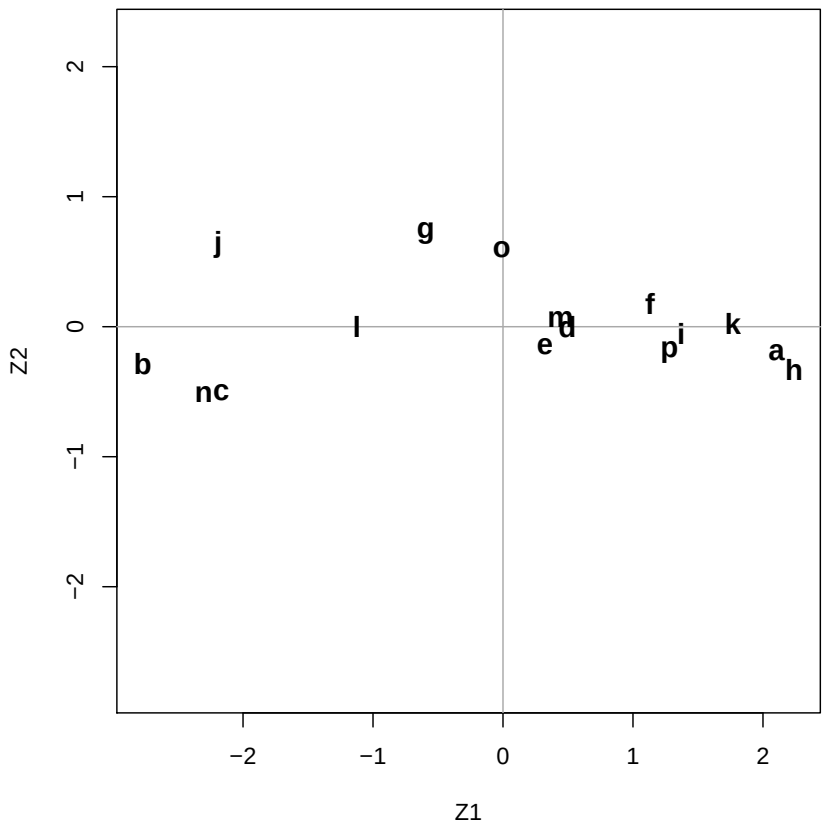
Prot	0.97 =	0.96	+	0.01
Fat	1.00 =	0.91	+	0.09
Lact	0.98 =	0.94	+	0.04

Gembloux Agro-Bio Tech
Université de Liège

Correlation circle

```
# correlation circle (FactoMineR)
plot(mammi.pca, choix="var")
```

Plot of the observations

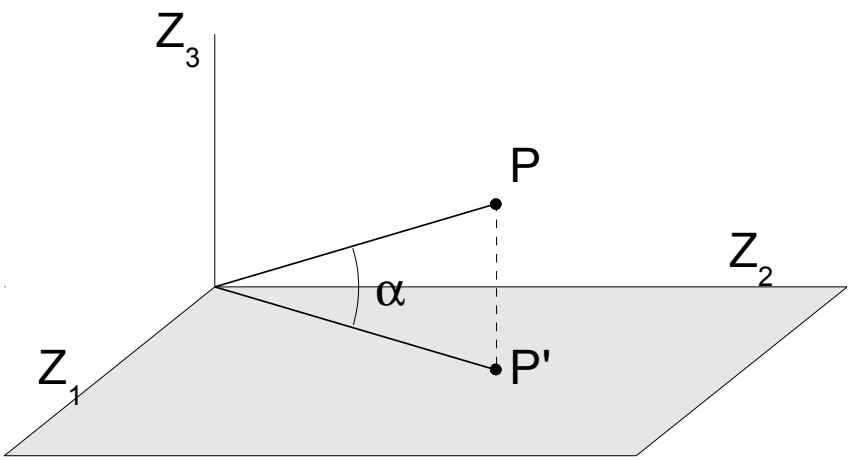


- coordinates : z-scores
- quality : squared cosines

Individuals

```
# with FactoMineR
plot(mammi.pca, choix="ind")
```

Quality of individuals' representation



= **squared cosines**

= ratio of squared distances in the chosen subspace and in the global space

$$= \frac{\sum_{i=1}^{r'} z_{ij}^2}{\sum_{i=1}^{r} z_{ij}^2}$$

Gembloux Agro-Bio Tech
Université de Liège

Example - Donkey in first factorial plane

	Z1	Z2	Z3
Donkey	2.105	-0.193	-0.100

$$d^2_{12} = z^2_{11} + z^2_{12} = 2.105^2 + (-0.193)^2 = 4.468$$

$$d^2_{123} = z^2_{11} + z^2_{12} + z^2_{13} = 2.105^2 + (-0.193)^2 + (-0.100)^2 = 4.478$$

$$\begin{aligned} \cos^2_{12} &= 4.468/4.478 = 0.998 \\ &= \cos^2_1 + \cos^2_2 \end{aligned}$$

Gembloux Agro-Bio Tech
Université de Liège

Quality of individual's representation

```
# individual's cos2 with FactoMineR
mammi.pca$ind$cos2
```

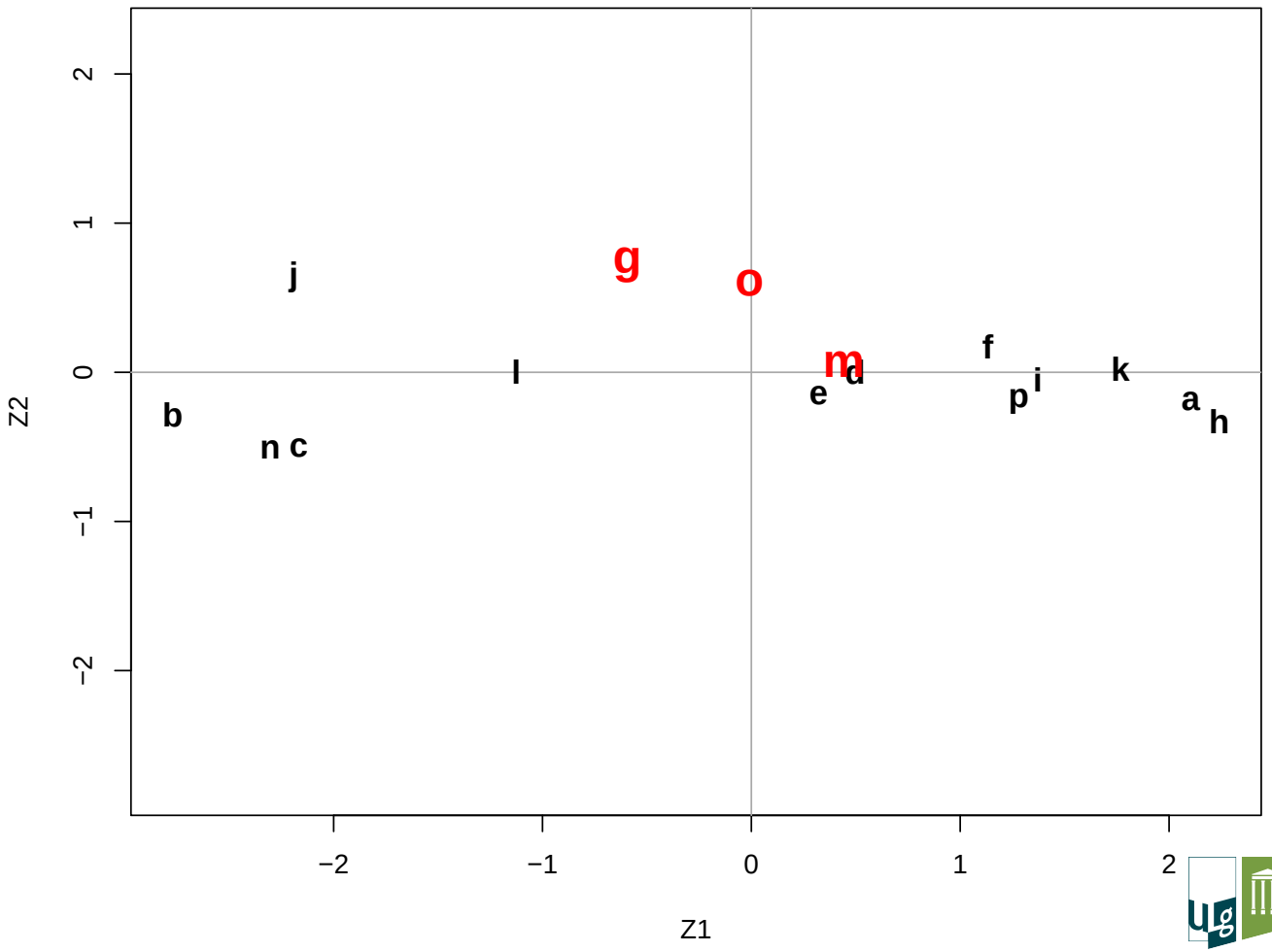
Example - Squared cosines

	Name	Axis 1	Axis 2	Plane(1, 2)
a	Donkey	0.9894	0.0083	0.9978
b	Whale	0.9793	0.0103	0.9897
c	Deer	0.9490	0.0509	0.9999
d	Sheep	0.9832	0.0000	0.9832
e	Buffalo	0.7675	0.1719	0.9394
f	Camel	0.9204	0.0226	0.9430
g	Guinea pig	0.3650	0.5360	0.9010
h	Horse	0.9530	0.0205	0.9735
i	Llama	0.9919	0.0013	0.9932
j	Rabbit	0.9050	0.0762	0.9812
k	Mule	0.9705	0.0001	0.9706
l	Rat	0.9782	0.0001	0.9782
m	Fox	0.5759	0.0100	0.5858
n	Reindeer	0.9520	0.0480	1.0000
o	Pig	0.0002	0.9907	0.9909
p	Zebra	0.9448	0.0212	0.9660

Example - Squared cosines

	Name	Axis 1	Axis 2	Plane(1, 2)
a	Donkey	0.9894	0.0083	0.9978
b	Whale	0.9793	0.0103	0.9897
c	Deer	0.9490	0.0509	0.9999
d	Sheep	0.9832	0.0000	0.9832
e	Buffalo	0.7675	0.1719	0.9394
f	Camel	0.9204	0.0226	0.9430
g	Guinea pig	0.3650	0.5360	0.9010
h	Horse	0.9530	0.0205	0.9735
i	Llama	0.9919	0.0013	0.9932
j	Rabbit	0.9050	0.0762	0.9812
k	Mule	0.9705	0.0001	0.9706
l	Rat	0.9782	0.0001	0.9782
m	Fox	0.5759	0.0100	0.5858
n	Reindeer	0.9520	0.0480	1.0000
o	Pig	0.0002	0.9907	0.9909
p	Zebra	0.9448	0.0212	0.9660

Example - Quality of individuals' representation



Gembloux Agro-Bio Tech
Université de Liège

Contents

Supplementary variables

Why ? - particular nature
- missing data

How ? calculate correlations of the new variables with existing components

```
# with FactoMineR
PCA(x, quanti.sup=...)
```

Supplementary observations

Why ? - centroids of existing groups
- particuler individuals (outliers, etc.)

How ? $\mathbf{Z}_s = \mathbf{X}_s\mathbf{U}$

```
# with FactoMineR
PCA(x, ind.sup=...)
```

Example - Cow

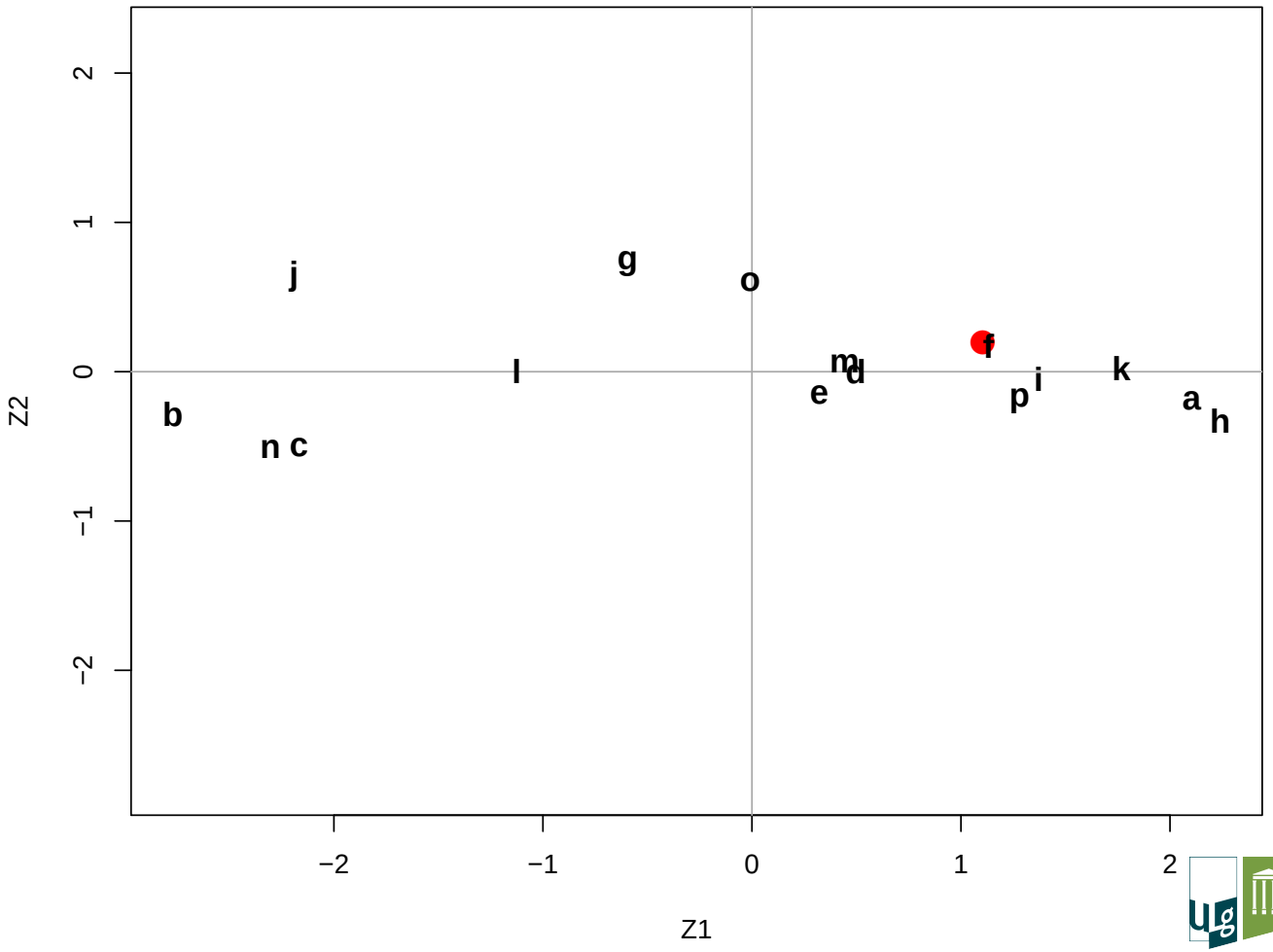
	Raw data			Standardized data		
	Prot	Fat	Lact	X1	X2	X3
Cow	3.4	3.5	4.7	-0.8679	-0.7187	0.3246

$$\mathbf{z}_s = \mathbf{x}_s\mathbf{U}$$

$$= \begin{bmatrix} -0.8679 & -0.7187 & 0.3246 \end{bmatrix} \begin{bmatrix} -0.585 & 0.233 & 0.777 \\ -0.569 & -0.801 & -0.188 \\ 0.578 & -0.552 & 0.601 \end{bmatrix}$$

$$= \begin{bmatrix} 1.1042 & 0.1936 & -0.3442 \end{bmatrix}$$

Example - Cow



Gembloux Agro-Bio Tech
Université de Liège

Transformation of variables

- Standardisation
  - raw variables
  - weighting
- Transformation
  - interpretation
  - normality of the initial variables

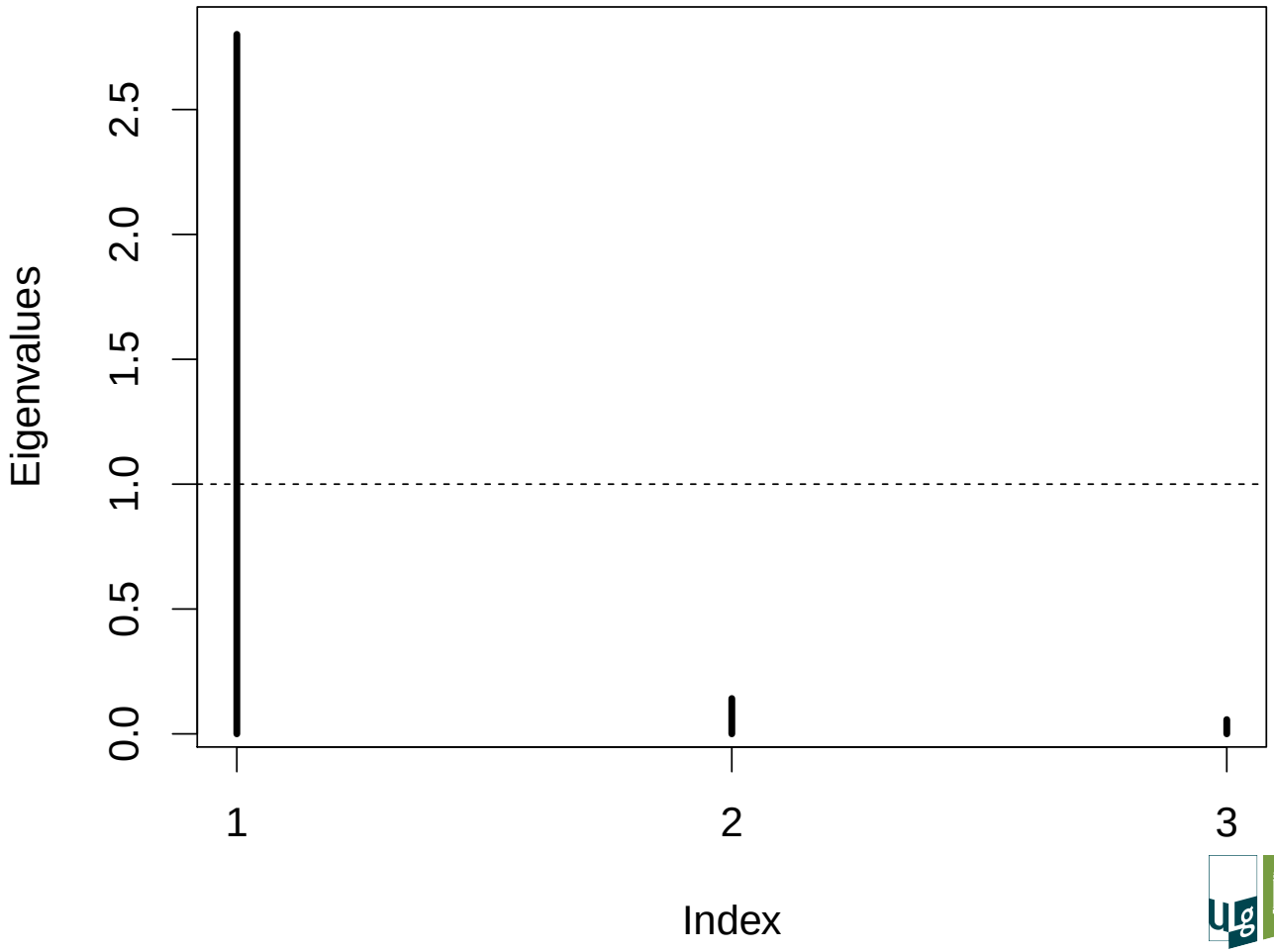
Number of components

- proportion of variance explained
- average eigenvalue
- scree plot
- interpretability

Screeplot

```
plot(mammi.pca$eig[,1], type="b")
abline(h=1, lty="dashed")
```

Number of components



Gembloux Agro-Bio Tech
Université de Liège

Interpretation of components

- **Correlation circle**
  Size effect, shape effect
- **Individuals**
  Detection of ouliers

Uses of PCA

- Description of the multivariate structure of the data (exploratory data analysis)
  - prior or posterior groups
  - outliers
- Reduction of dimensions
  - number of variables and orthogonality
  - clustering, ANOVA, regression

Example II - full spatial data

Raw data : WorldClim data

- 19 bioclimatic variables, averages for 1970-2000
- 10 minutes spatial resolution ($\simeq$ 340 km^2)

Available at `http://worldclim.org/version2` or directly from within R (package `raster`)

Example II - full spatial data

BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

Gembloux Agro-Bio Tech
Université de Liège