Pattern recognition on spatial data

Exploring structures : Principal Components Analysis

Pr Yves Brostaux, ULiège

Openspat ULisboa

June 2019

Contents

Introduction

- Introduction
- Example data

Objectives

- summarize data
- display data

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

Data

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

Introduction

- Introduction
- Example data

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

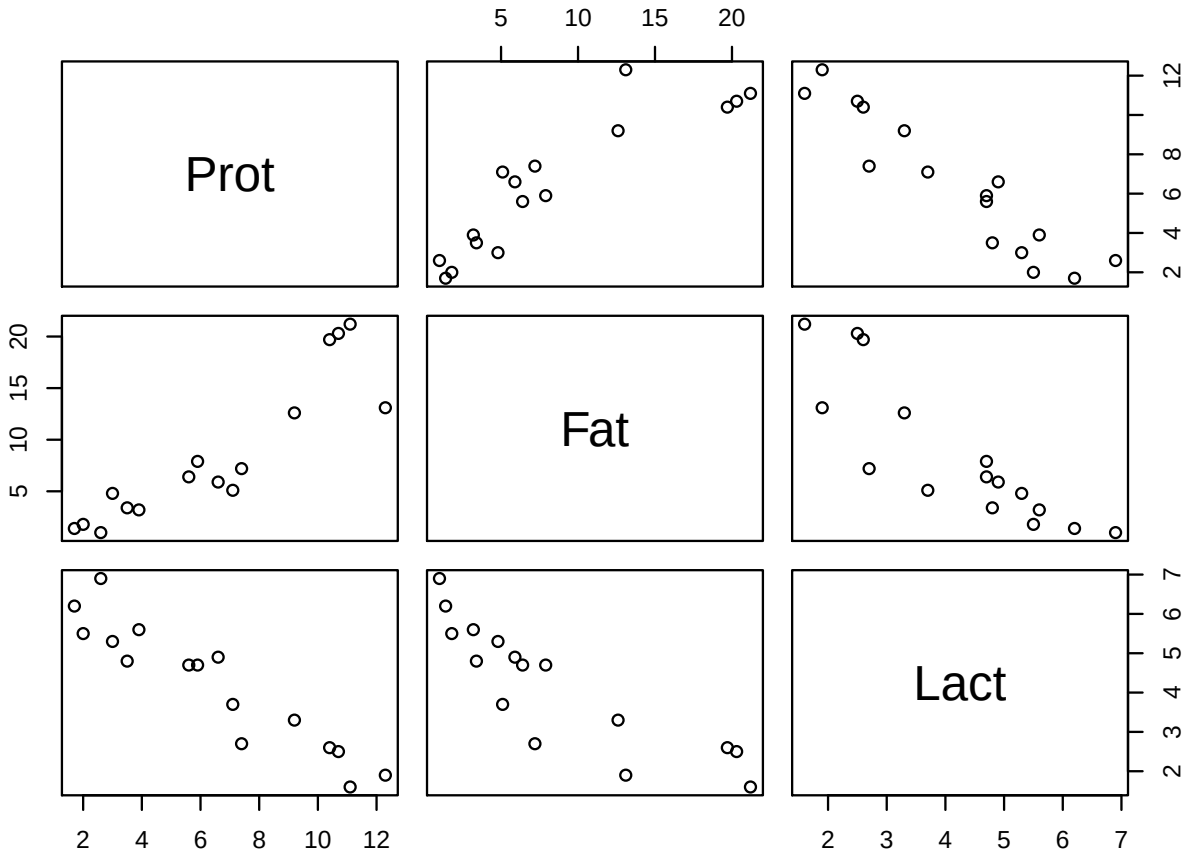
Univariate statistics

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



```r
# read raw data
mammi <- read.table("mammals.txt", row.names=1)
colnames(mammi) <- c("Name", "Prot", "Fat", "Lact")

# compute means and standard deviations
# ([-1] to exclude the first column)
lapply(mammi[-1], mean)
lapply(mammi[-1], sd)

# correlation matrix
cor(mammi[-1])

# matrix plot
plot(mammi[-1])
```

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

Principal components or z-scores

- Principles
- Mathematical aspects
- Geometrical meaning of principal components

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 $\boldsymbol{z}_2$ 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 $\boldsymbol{z}_3$ 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

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

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

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

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

Principal components or z-scores

- Principles
- Mathematical aspects
- Geometrical meaning of principal components

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

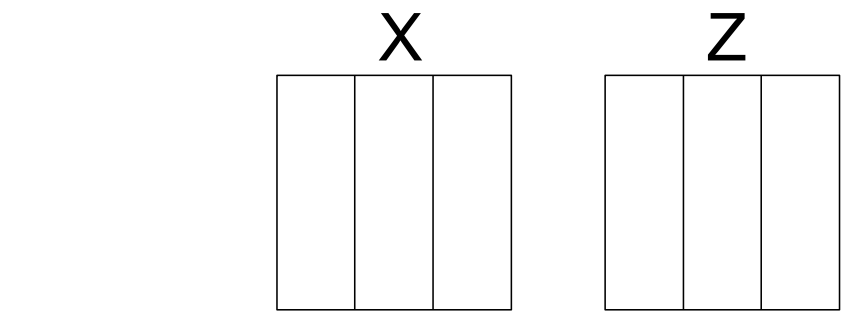
Computing scores

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

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

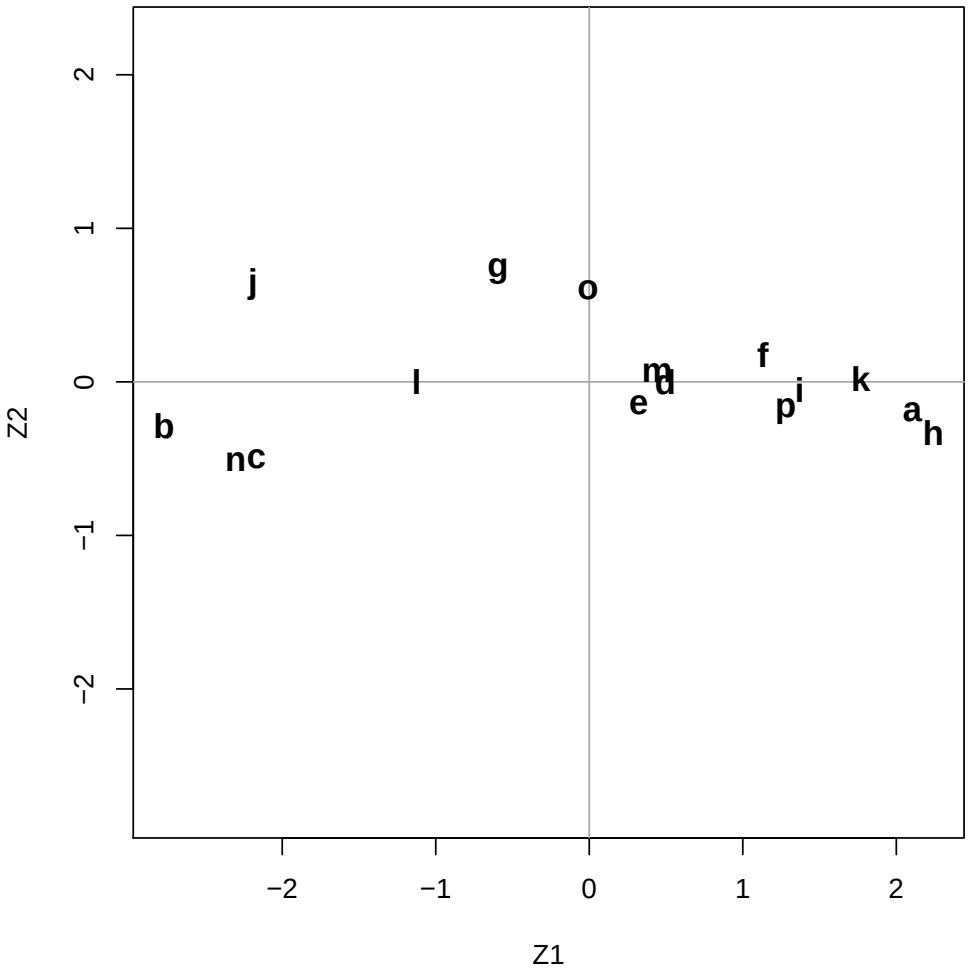
Raw variables and scores



X : correlated variables
same importance
(same variance)

Z : non correlated variables
decreasing importance

Graphical result - Z_1 & Z_2



PCA using FactoMineR package

```r
# package loading
library(FactoMineR)

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

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

PCA object

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 16 individuals, described by 3 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$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"
```

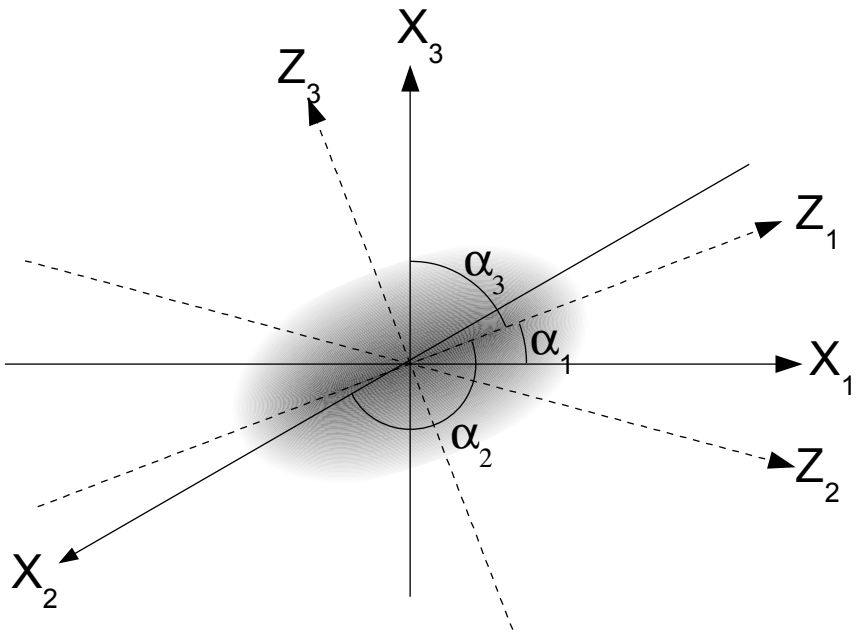
Individuals plot - Z_1 & Z_2

```
# with FactoMineR (first and second component by default)
plot(mammi.pca, choix="ind")
```

Principal components or z-scores

- Principles
- Mathematical aspects
- Geometrical meaning of principal components

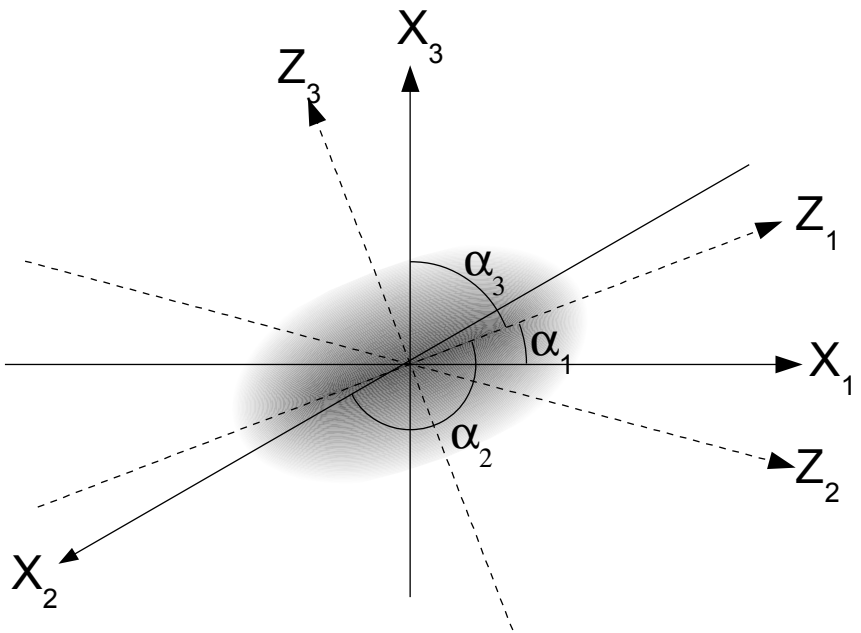
Geometrical meaning of principal components



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

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

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

Contents

Interpretation

- Principal components
- Individuals
- Number of components

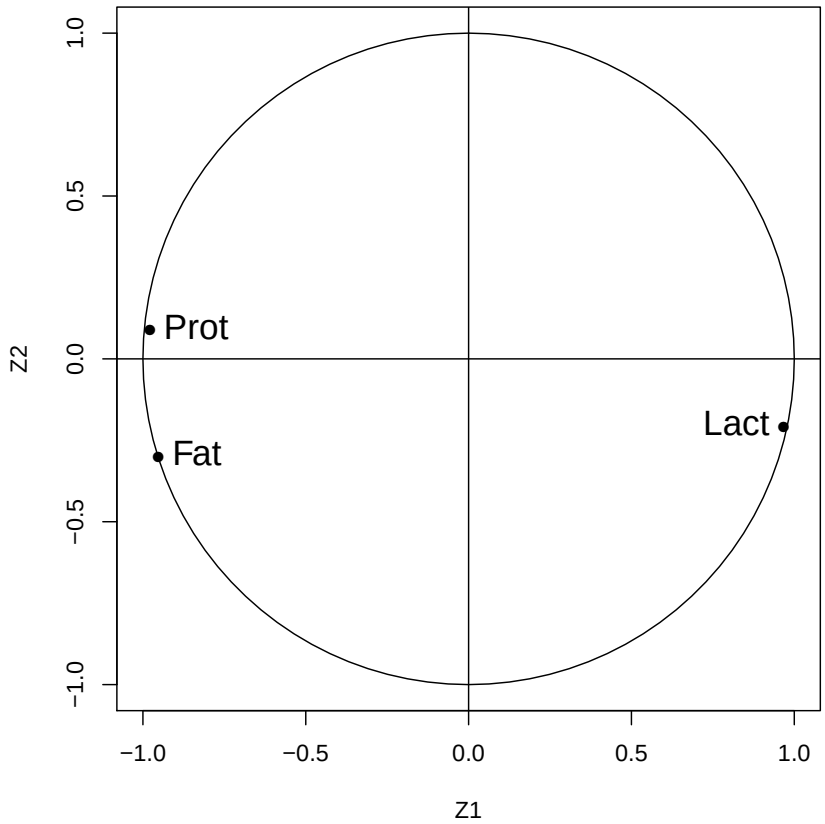
Correlation between z and x

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

	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

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

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

Correlation

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

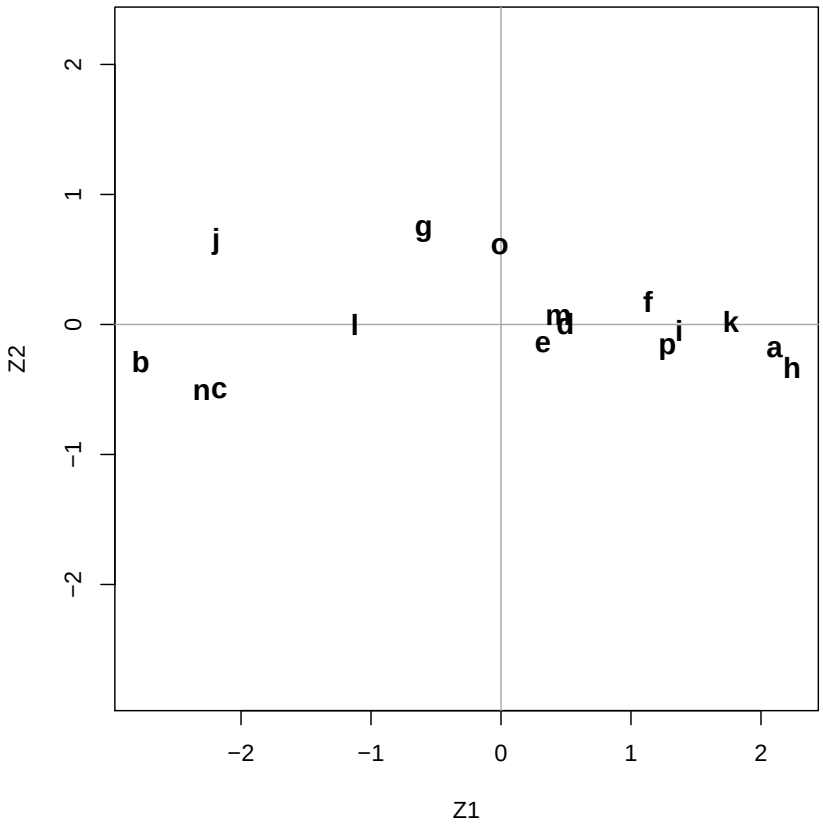
```r
## With FactoMineR
# Correlations
mammi.pca$var$cor
dimdesc(mammi.pca, proba=1) # proba = alpha risk of signif

# Correlations circle ((first and second component by defa
plot(mammi.pca, choix="var")
```

Interpretation

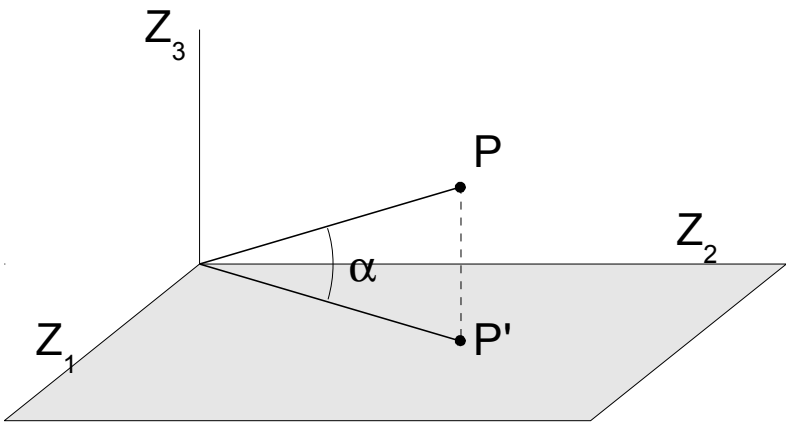
- Principal components
- Individuals
- Number of components

Plot of the observations



- coordinates : z-scores
- quality : squared cosines

Quality of individuals' representation



= **squared cosines**

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

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

Example - Donkey in first factorial plane

	Z1	Z2	Z3
Donkey	2.105	-0.193	-0.100

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

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

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

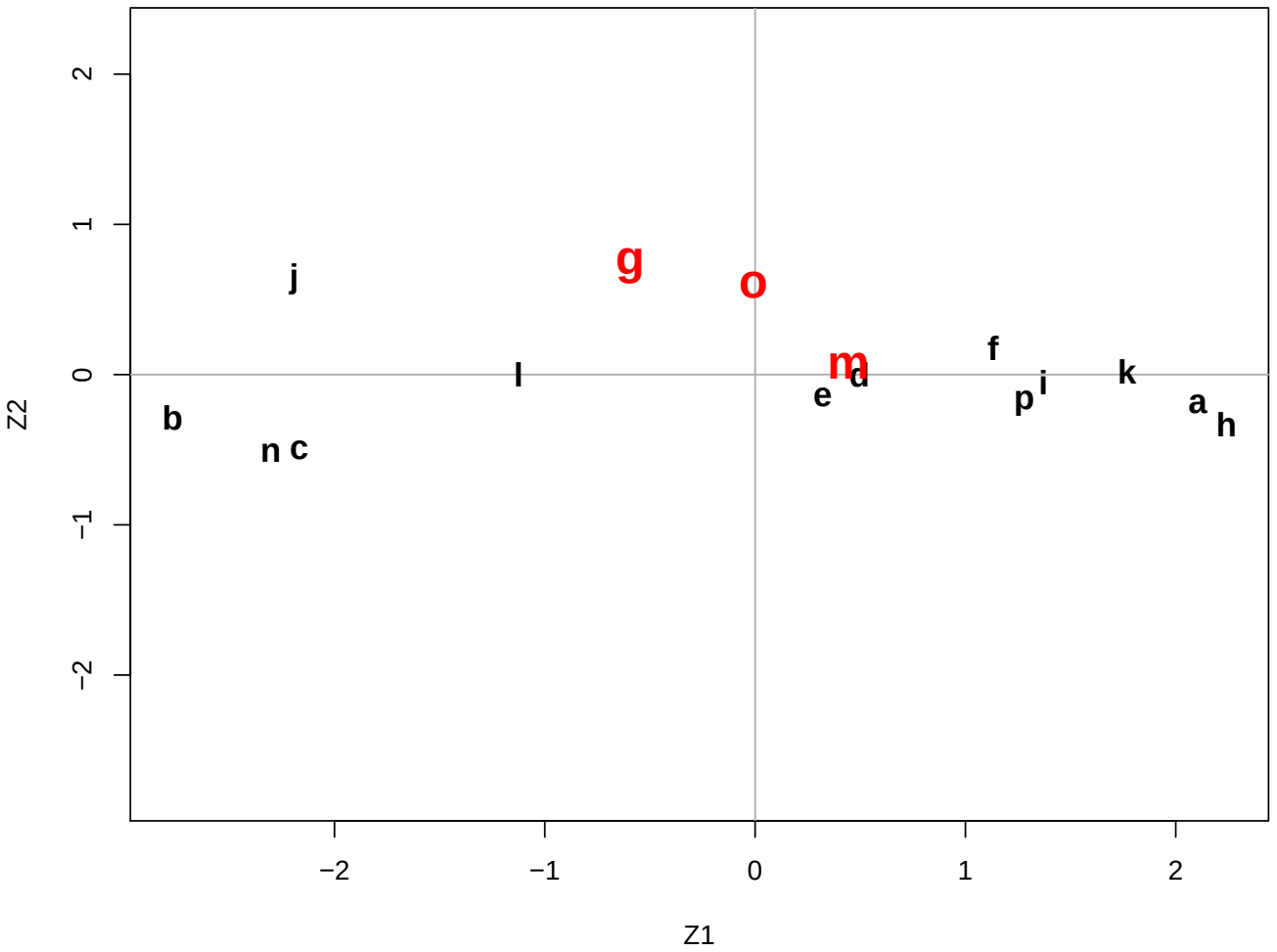
Example - Squared cosines

	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



Quality of individual's representation

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

What to look for ?

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

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

Interpretation

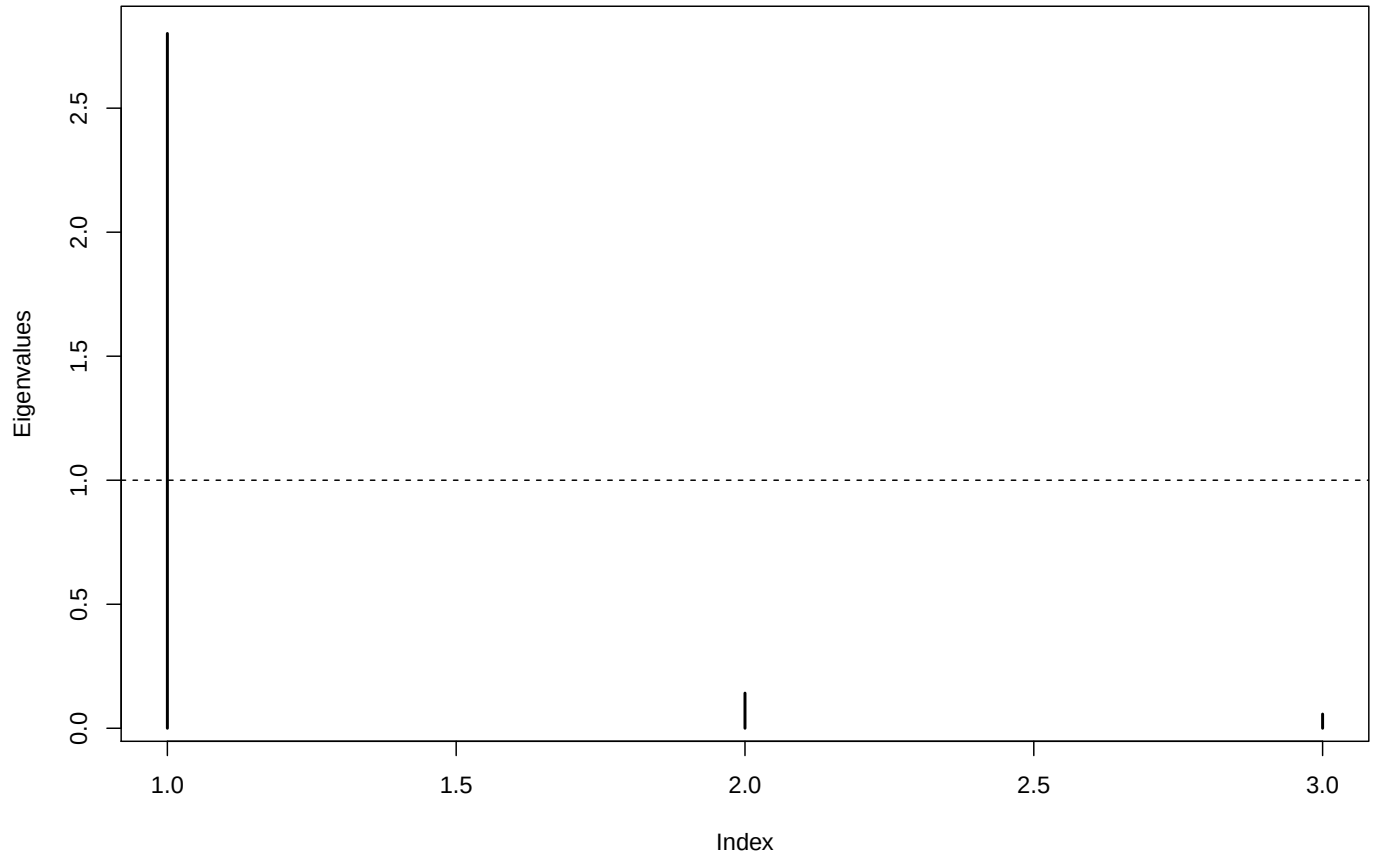
- Principal components
- Individuals
- Number of components

Number of components

How to choose the number of principal components to retain/interpret ?

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

Screeplot



Screeplot

```
# screeplot
plot(mammi.pca$eig[,1], type="b")
# horizontal line at 1 (limit for standardized data)
abline(h=1, lty="dashed")
```

Contents

Transformation of variables

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

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