Multivariate Statistics in Ecology and Quantitative Genetics
Principal component analysis

Dirk Metzler & Noémie Becker

http://evol.bio.lmu.de/_statgen

Summer semester 2013

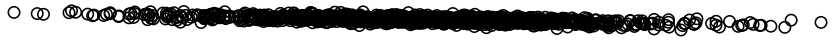
Contents

We wish to visualize multi-dimensional data
in order to identify patterns.

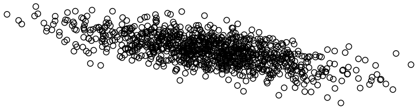
We wish to visualize multi-dimensional data
in order to identify patterns.

How do we visualize multi-dimensional data???

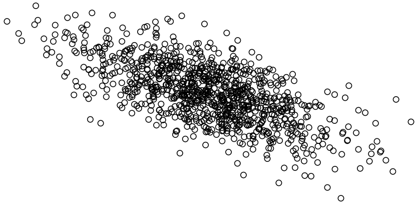
Example: 2-dim data in 3 dimensions
(Imagine the cloud to be rotated in 3 dimensions)



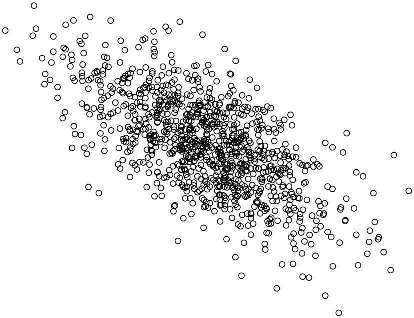
Example: 2-dim data in 3 dimensions
(Imagine the cloud to be rotated in 3 dimensions)



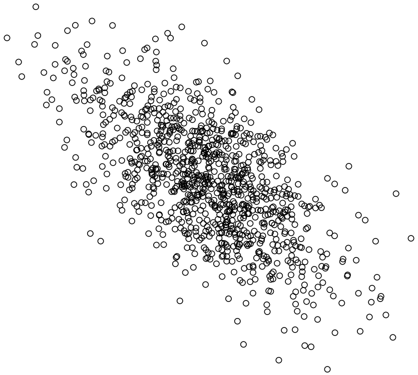
Example: 2-dim data in 3 dimensions
(Imagine the cloud to be rotated in 3 dimensions)



Example: 2-dim data in 3 dimensions
(Imagine the cloud to be rotated in 3 dimensions)



Example: 2-dim data in 3 dimensions
(Imagine the cloud to be rotated in 3 dimensions)



To have a good view on the data,
we wish to plot those components
which contribute most of the variation.

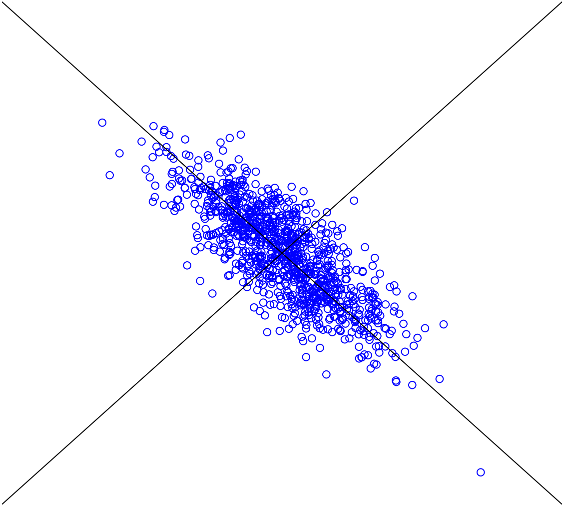
To have a good view on the data,
we wish to plot those components
which contribute most of the variation.

The component with the most variation
is rotated onto the x-axis,

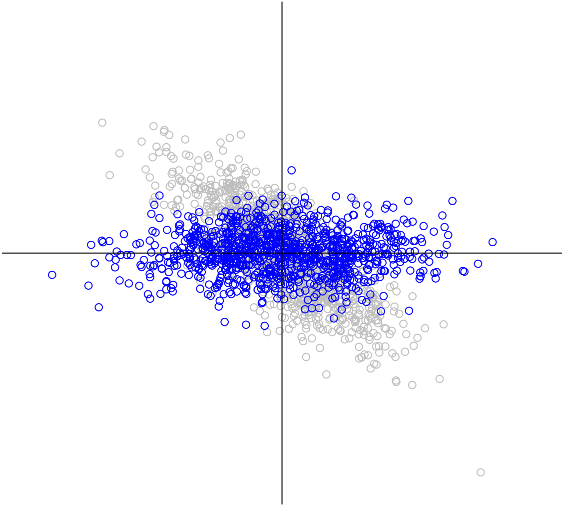
To have a good view on the data,
we wish to plot those components
which contribute most of the variation.

The component with the most variation
is rotated onto the x-axis,
the component with the second most variation
is rotated onto the y-axis.

Example: 2-dim data



Example: 2-dim data



The principal component analysis finds the components with the most contribution to the total variance.

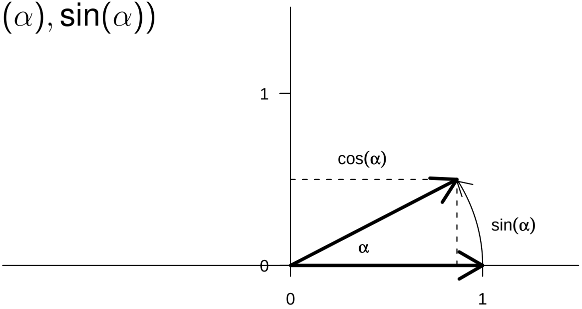
The principal component analysis finds
the components with the most contribution
to the total variance.

Before we investigate
how to obtain the optimal transformation,
we need to understand
how to rotate a data cloud.

Contents

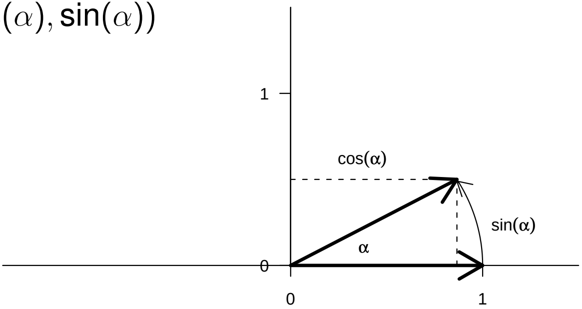
Rotation by angle α.

$(1, 0) \to (\cos(\alpha), \sin(\alpha))$

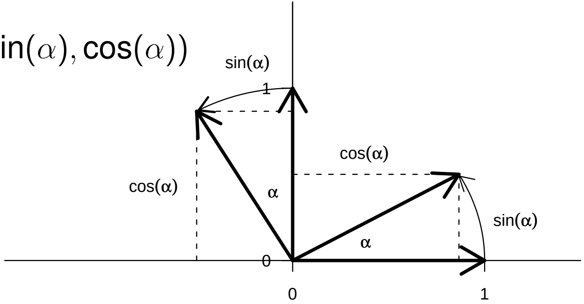


Rotation by angle α.

$(1, 0) \to (\cos(\alpha), \sin(\alpha))$



$(0, 1) \to (-\sin(\alpha), \cos(\alpha))$



Rotation by angle α of a vector (x, y):

$$(x, y) \to (x, y) \cdot \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

Every rotation matrix R has the property $R^T \cdot R = \mathbb{1}$.
Example

$$\begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \cdot \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

$$= \begin{pmatrix} \sin^2(\alpha) + \cos^2(\alpha) & 0 \\ 0 & \sin^2(\alpha) + \cos^2(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Every rotation matrix R has the property $R^T \cdot R = \mathbb{1}$.
Example

$$\begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \cdot \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

$$= \begin{pmatrix} \sin^2(\alpha) + \cos^2(\alpha) & 0 \\ 0 & \sin^2(\alpha) + \cos^2(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

From now on we consider matrices U with the property

$$U^T \cdot U = \mathbb{1}$$

These matrices are called orthogonal (also called orthonormal) and preserve distances. Such transformations are mixtures of rotations and reflections.

A didactic Example

Before we go into applications,
we wish to see what the PCA does.

A didactic Example

Before we go into applications,
we wish to see what the PCA does.

We simulate a data cloud from a
multi-variate normal distribution with covariance matrix

$$\begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$$

that is, the two components are independent
and normally distributed with variances 5 and 1, respectively.

A didactic Example

Before we go into applications,
we wish to see what the PCA does.

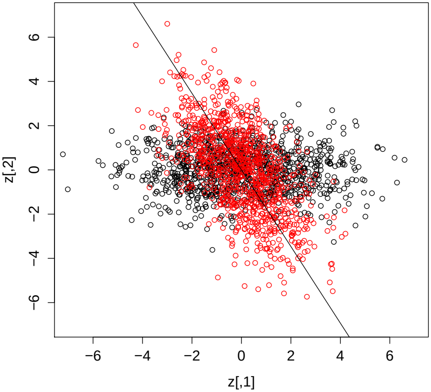
We simulate a data cloud from a
multi-variate normal distribution with covariance matrix

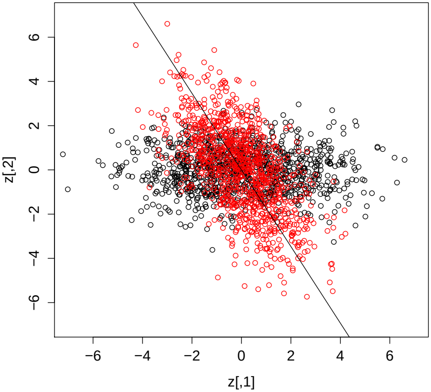
$$\begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$$

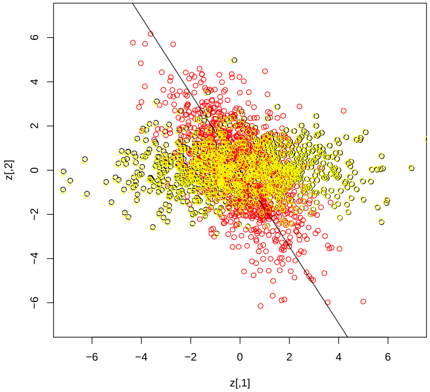
that is, the two components are independent
and normally distributed with variances 5 and 1, respectively.

We rotate the cloud by -60°
and apply the R-command `prcomp()`.

```
> library("mvtnorm")
> z <- rmvnorm(1000,sigma=matrix(c(5,0,0,1),nrow=2))
> RotMat <- matrix(c(cos(pi/3),sin(pi/3),
+                   -sin(pi/3),cos(pi/3)),nrow=2)
> x <- z %*% RotMat
> plot(z,xlim=c(-7,7),ylim=c(-7,7))
> points(x,col="red")
> abline(b=tan(-pi/3),a=0)
> pca <- prcomp(x)
> points(pca$x,col="yellow")
```


z[,2]
z[,1]
−6
−4
−2
0
2
4
6
−6
−4
−2
0
2
4
6


z[,2]
z[,1]
−6
−4
−2
0
2
4
6


z[,2]
z[,1]
−6
−4
−2
0
2
4
6

Further observations:

```
> names(pca)
[1] "sdev"     "rotation" "center"   "scale"    "x"
> pca
Standard deviations:
[1] 2.232067 1.008979

Rotation:
PC1        PC2
[1,]  0.5027292 0.8644439
[2,] -0.8644439 0.5027292

> ( pca$sdev )^2
[1] 4.982122 1.018038
```

```
> RotMat %*% pca$rotation
             PC1          PC2
[1,] 0.999995025 -0.003154303
[2,] 0.003154303  0.999995025
> t( pca$rotation ) %*% pca$rotation
    PC1 PC2
PC1   1   0
PC2   0   1
> cov(z)
[,1]       [,2]
[1,] 4.98180617 0.01204928
[2,] 0.01204928 1.01732926
> t( pca$rotation ) %*% cov(x) %*% pca$rotation
              PC1           PC2
PC1  4.9818427419 -0.0004560566
PC2 -0.0004560566  1.0172926950
```

The vector `pca$sdev` is approx. $(\sqrt{5}, \sqrt{1})$
The matrix `pca$rotation` is the transformation matrix
The matrix `pca$x` is the transformed data

Contents

Obviously the height (in cm) and the shoe size of human beings are correlated variables. We also consider the weight (in kg). The following data is from a test questionnaire from a statistics course in 1999/2000 in Göttingen.

Obviously the height (in cm) and the shoe size of human beings are correlated variables. We also consider the weight (in kg). The following data is from a test questionnaire from a statistics course in 1999/2000 in Göttingen.

Problem:

How can we compare variation in height (cm) with variation in weight (kg)?

Obviously the height (in cm) and the shoe size of human beings are correlated variables. We also consider the weight (in kg). The following data is from a test questionnaire from a statistics course in 1999/2000 in Göttingen.

Problem:

How can we compare variation in height (cm) with variation in weight (kg)?

Answer: (Co-)variances should be measured in units of the standard deviation.

Obviously the height (in cm) and the shoe size of human beings are correlated variables. We also consider the weight (in kg). The following data is from a test questionnaire from a statistics course in 1999/2000 in Göttingen.

Problem:
How can we compare variation in height (cm)
with variation in weight (kg)?

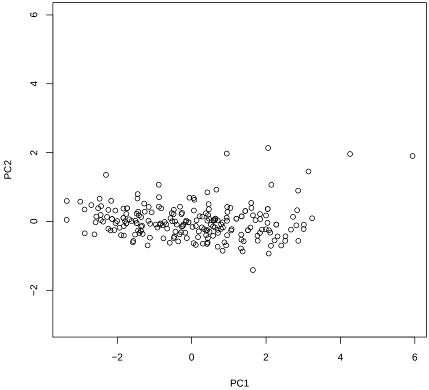
Answer: (Co-)variances should be measured
in units of the standard deviation.

This leads to considering
correlation matrices instead of covariance matrices.
In `R` simply use the option `scale=TRUE`.

```
shsw <-read.table("HeightShoeWeight.txt",header=TRUE)
attach(shsw)
head(shsw)
hsw <- shsw[,2:4]
head(hsw)
hsw.pca <- prcomp(hsw,scale=TRUE)
hsw.pca
fm.col <- character()
fm.col[sex==0] <- "blue"
fm.col[sex==1] <- "red"
sqrt( length(sex)-1 )      # = 15
```

Let us plot the transformed data.

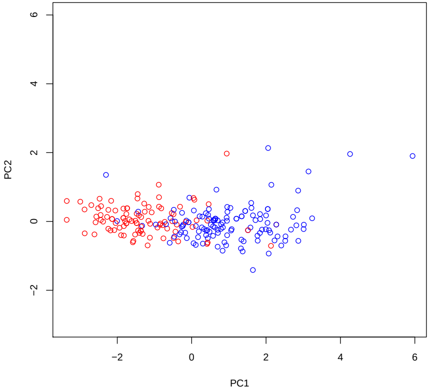
```
plot(hsw.pca$x,ylim=c(-3,6))
```



There is nothing special to see.

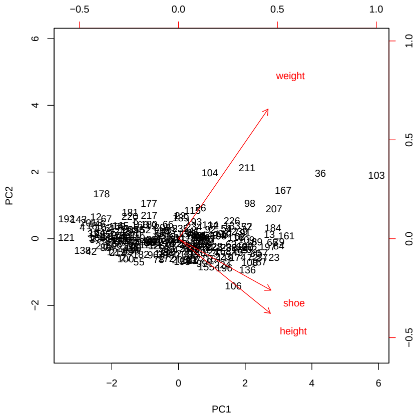
Which observation is from which sex:

```
plot(hsw.pca$x,ylim=c(-3,6),col=fm.col)
```

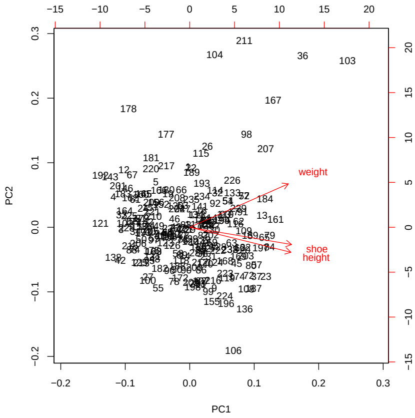


Why are guys on the right and girls on the left?

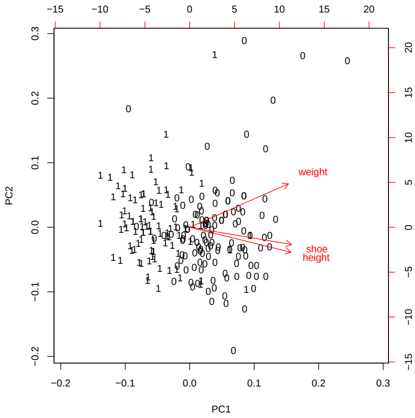
```
biplot(hsw.pca,scale=0)
```



```
biplot(hsw.pca,scale=1)
```



```
biplot(hsw.pca,scale=1,xlabs=sex)
```



The first component can be interpreted as size.
As guys are on average taller than girls, this explains why guys are on the right and girls on the left.

The first component can be interpreted as size.
As guys are on average taller than girls, this explains why guys are on the right and girls on the left.

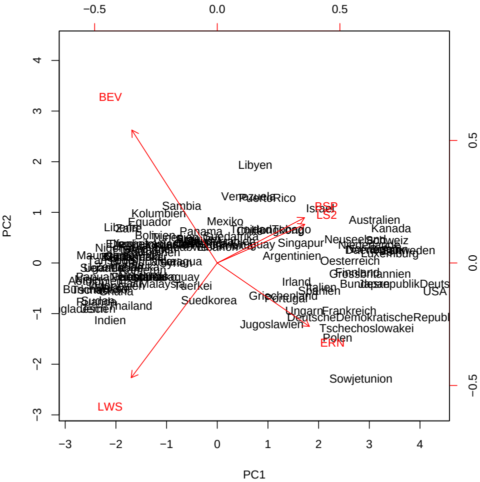
The second component is
„weight which is not explained by the first component 'size' “.
Thus students with overweight are on top of the last figure whereas students with underweight are on the bottom of the last figure.

Contents

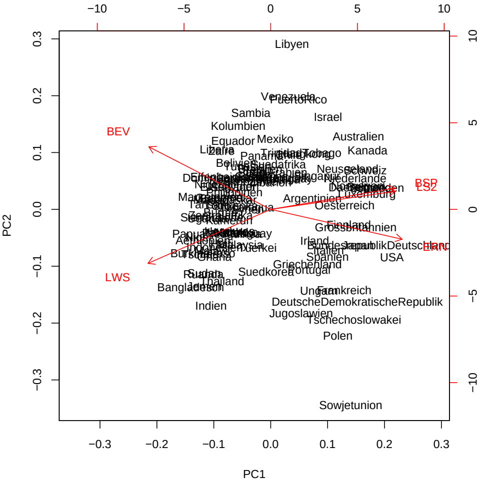
The file Countries.txt contains data from
Kockluner: Angewandte Regessionsanalyse mit SPSS, Vieweg 1988, S. 7:

Variables:
ERN: nutrition index (Ernährungsindex)
BSP: gross national product per person
(Bruttosozialprodukt pro Kopf)
LWS: agriculture index (Landwirtschaftsindex)
LS2: cost of living index (Lebenshaltungsindex 2)
BEV: index of inhabitants (Bevölkerungsindex)

```
countries <- read.table("Countries.txt",header=TRUE)
cntr.pca <- prcomp(countries,scale=TRUE); cntr.pca
plot(cntr.pca$x)
biplot(cntr.pca,scale=0)
```



```
biplot(cntr.pca,scale=1)
```



Contents

The mathematical background is explained on the board.

Contents

Reading biplots

Distance biplot (scale=0 in R)

- Angles between lines are meaningless.
- The lines are projections of length 1 vectors into the plane of the first two principal components. So the length indicates how well the corresponding variable is represented by the first two components.
- Distances between points/labels approximate distances of the observations for different objects.
- The projection of a point onto a vector at right angle approximates the position of the corresponding object along the corresponding variable.

Correlation biplot (scale=1 in R)

- The cosine of the angle between two lines is approximately equal to the correlation between the corresponding variables.
- If the PCA used `scale=FALSE`, then the length of a line is approximately $\sqrt{N-1}$ times the estimated standard deviation of the corresponding variable. If the PCA used `scale=TRUE`, then the lines are projections of length $\sqrt{N-1}$ vectors into the plane of the first two principal components. So the length indicates how well the corresponding variable is represented by the first two components.
- Distances between points/labels are meaningless.
- The projection of a point onto a vector at right angle approximates the position of the corresponding object along the corresponding variable.

Due to the projection, the approximation of quantities such as distance between points or correlation between variables can be poor.

Due to the projection, the approximation of quantities such as distance between points or correlation between variables can be poor.

The approximations are reasonably good of the first two principal components explain $70\% - 80\%$ of the total variation (or even more).

Due to the projection, the approximation of quantities such as distance between points or correlation between variables can be poor.

The approximations are reasonably good of the first two principal components explain $70\% - 80\%$ of the total variation (or even more).

In applications the first two components typically explain far less then 70% of the total variation. PCA is still used as there is not better method. But be careful and think twice.

Contents

One problem with PCA is to decide how many components to present, and there are various rules of thumb.

- 80%-rule: Present the first k axes that explain 80% of the total variation.

One problem with PCA is to decide how many components to present, and there are various rules of thumb.

- 80%-rule: Present the first k axes that explain 80% of the total variation.
- ellbow-rule: Plot the eigenvalues as vertical lines or bars next to each other. Use k axes if the 'elbow' is at $k+1$.

One problem with PCA is to decide how many components to present, and there are various rules of thumb.

- 80%-rule: Present the first k axes that explain 80% of the total variation.
- ellbow-rule: Plot the eigenvalues as vertical lines or bars next to each other. Use k axes if the 'elbow' is at $k+1$.
- broken-stick-rule: If a stick of unit length is broken at random in p pieces, then the expected length of the j-th largest piece is given by

$$L_j = \frac{1}{p} \sum_{i=j}^{p} \frac{1}{i} \tag{1}$$

  If the eigenvalue of the j-th axis is larger than L_j, then it can be considered as important.

One problem with PCA is to decide how many components to present, and there are various rules of thumb.

- 80%-rule: Present the first k axes that explain 80% of the total variation.
- ellbow-rule: Plot the eigenvalues as vertical lines or bars next to each other. Use k axes if the 'elbow' is at $k+1$.
- broken-stick-rule: If a stick of unit length is broken at random in p pieces, then the expected length of the j-th largest piece is given by

$$L_j = \frac{1}{p} \sum_{i=j}^{p} \frac{1}{i} \tag{1}$$

  If the eigenvalue of the j-th axis is larger than L_j, then it can be considered as important.

The broken-stick-model is the most reliable rule of thumb.

Example: Height and weight data.

```
> gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 )
[1] 0.86984879 0.08035589 0.04979531
> cumsum( gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 ) )
[1] 0.8698488 0.9502047 1.0000000
> screeplot( gsg.pca, type="lines")
```

Example: Height and weight data.

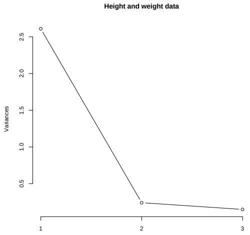
```
> gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 )
[1] 0.86984879 0.08035589 0.04979531
> cumsum( gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 ) )
[1] 0.8698488 0.9502047 1.0000000
> screeplot( gsg.pca, type="lines")
```

80%-rule: one component is enough

Example: Height and weight data.

```
> gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 )
[1] 0.86984879 0.08035589 0.04979531
> cumsum( gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 ) )
[1] 0.8698488 0.9502047 1.0000000
> screeplot( gsg.pca, type="lines")
```

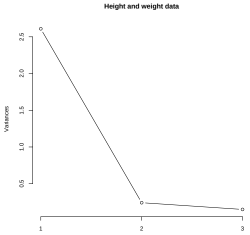
80%-rule: one component is enough



Example: Height and weight data.

```
> gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 )
[1] 0.86984879 0.08035589 0.04979531
> cumsum( gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 ) )
[1] 0.8698488 0.9502047 1.0000000
> screeplot( gsg.pca, type="lines")
```

80%-rule: one component is enough



ellbow-rule: one component is enough

```
> gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 )
[1] 0.86984879 0.08035589 0.04979531
> p<-length(gsg.pca$sdev)
> L<-matrix(ncol=p)
> for (i in 1:p) {
+ L[i]<-round(1/p*sum(1/seq(from=i, to=p)),2)
+ }
> L
     [,1] [,2] [,3]
[1,] 0.61 0.28 0.11
```

```
> gsg.pca$sdev^2/sum( (gsg.pca$sdev)^2 )
[1] 0.86984879 0.08035589 0.04979531
> p<-length(gsg.pca$sdev)
> L<-matrix(ncol=p)
> for (i in 1:p) {
+ L[i]<-round(1/p*sum(1/seq(from=i, to=p)),2)
+ }
> L
     [,1] [,2] [,3]
[1,] 0.61 0.28 0.11
```

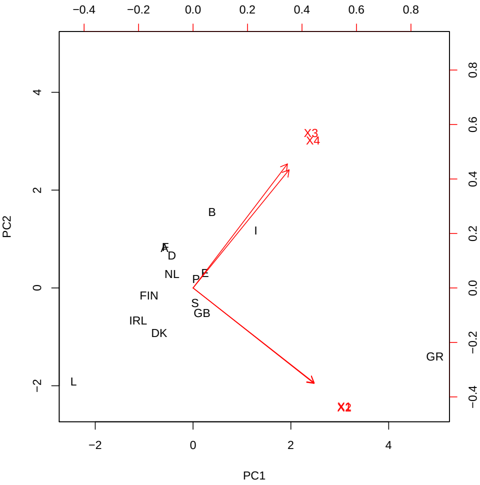
broken-stick-rule: one component is enough
($0.87 >= 0.61$, $0.08 < 0.28$, $0.05 < 0.11$)

Contents

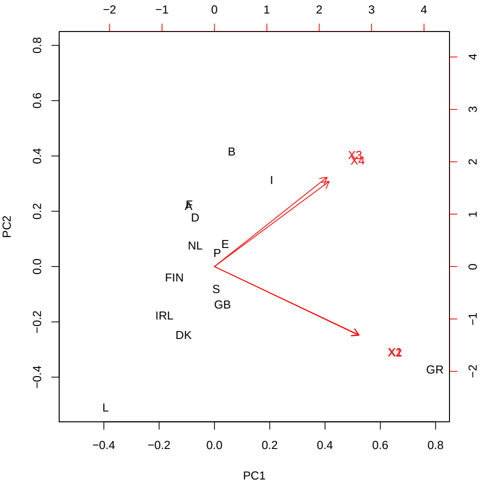
The file 'EWU.txt' contains data of European countries. (From Rinne (2000,p21.)). Let's find out.

```
ewu <- read.table("EWU.txt",header=TRUE)
ewu1 <- ewu[,2:5]
ewu.pca <- prcomp(ewu1, scale=TRUE)
biplot(ewu.pca,scale=0,xlabs=ewu$Staat)
biplot(ewu.pca,scale=1,xlabs=ewu$Staat)
```

Distance biplot (scale=0):



Correlation biplot (scale=1):



The variables X1 and X2 are highly positively correlated.
The variables X3 and X4 seem to be highly positively correlated.

The variables X1 and X2 are highly positively correlated.
The variables X3 and X4 seem to be highly positively correlated.

So what are X1, X2, X3 and X4?

- X1 is the inflation rate 1997 in %
- X2 is the long term interest rate 1997 in %
- X3 is the new indebtedness 1997 in % of the GDP
- X4 is the public debt level 1997 in % of the GDP

The fitness of candidates for the European currency union has been measured with these four variables.

Contents

In many cases the different variables are on different scales. Then you are recommended to scale the variables with their standard deviations, that is, to use the correlation matrix rather than the covariance matrix.
Otherwise the first principal component might be dominated by the variable with the largest scale.

In many cases the different variables are on different scales. Then you are recommended to scale the variables with their standard deviations, that is, to use the correlation matrix rather than the covariance matrix.
Otherwise the first principal component might be dominated by the variable with the largest scale.
For you this means to use the argument `scale=TRUE` in the `prcomp()` command.

In many cases the different variables are on different scales. Then you are recommended to scale the variables with their standard deviations, that is, to use the correlation matrix rather than the covariance matrix.
Otherwise the first principal component might be dominated by the variable with the largest scale.
For you this means to use the argument `scale=TRUE` in the `prcomp()` command.

If the values of the variables are of comparable order, then it is also fine to not scale the variables, that is, to apply PCA to the covariance matrix.
In R this means to use the argument `scale=FALSE`.

Contents

Summary

Principal component analysis is a transformation (rotation and reflection) of the data such that
most of the variation is on the first axis,
the second most variation is on the second axis...

Summary

Principal component analysis is a transformation (rotation and reflection) of the data such that most of the variation is on the first axis, the second most variation is on the second axis...

Used for:

- Visualizing multi-variate data (we have no better method)

Summary

Principal component analysis is a transformation (rotation and reflection) of the data such that most of the variation is on the first axis, the second most variation is on the second axis...

Used for:

- Visualizing multi-variate data (we have no better method)
- Get a feeling on the dependencies

Summary

Principal component analysis is a transformation (rotation and reflection) of the data such that most of the variation is on the first axis, the second most variation is on the second axis...

Used for:

- Visualizing multi-variate data (we have no better method)
- Get a feeling on the dependencies
- Find clusters in the variables (e.g. $\{X1, X2\}$ and $\{X3, X4\}$ in the EWU data set)

Summary

Principal component analysis is a transformation (rotation and reflection) of the data such that most of the variation is on the first axis, the second most variation is on the second axis...

Used for:

- Visualizing multi-variate data (we have no better method)
- Get a feeling on the dependencies
- Find clusters in the variables (e.g. $\{X1, X2\}$ and $\{X3, X4\}$ in the EWU data set)
- Find clusters in the set of objects/individuals (e.g. girls and guys in the height and weight data)

Be aware:

- Principal components can often not be interpreted
  $2 * \text{shoe} + 3 * \text{height}$ is a measure for size
  But how shall we interpret $2 * \text{shoe} - \text{height}$?

Be aware:

- Principal components can often not be interpreted
  $2 * \text{shoe} + 3 * \text{height}$ is a measure for size
  But how shall we interpret $2 * \text{shoe} - \text{height}$?
- If first two principal components explain less than 70%, then consider forgetting PCA

Be aware:

- Principal components can often not be interpreted
  $2 * \text{shoe} + 3 * \text{height}$ is a measure for size
  But how shall we interpret $2 * \text{shoe} - \text{height}$?
- If first two principal components explain less than 70%, then consider forgetting PCA
- Biplots are easily misread. Be careful!

Be aware:

- Principal components can often not be interpreted
  $2 * \text{shoe} + 3 * \text{height}$ is a measure for size
  But how shall we interpret $2 * \text{shoe} - \text{height}$?
- If first two principal components explain less than 70%, then consider forgetting PCA
- Biplots are easily misread. Be careful!
- It's spelled 'principal' (main, Haupt-),
  not 'principle' (Prinzip, Grundsatz)