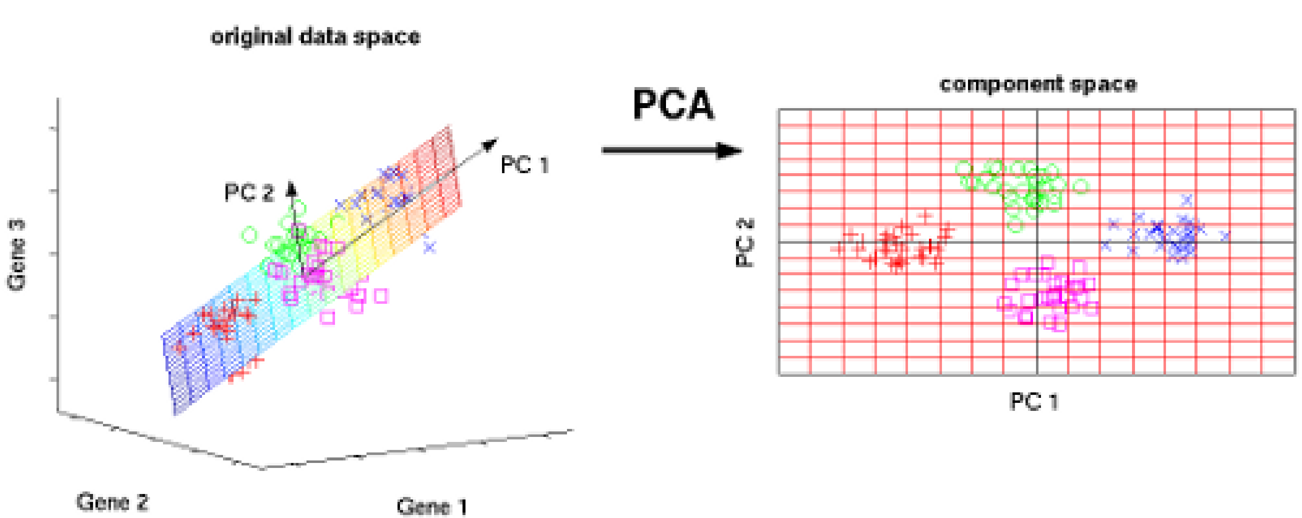
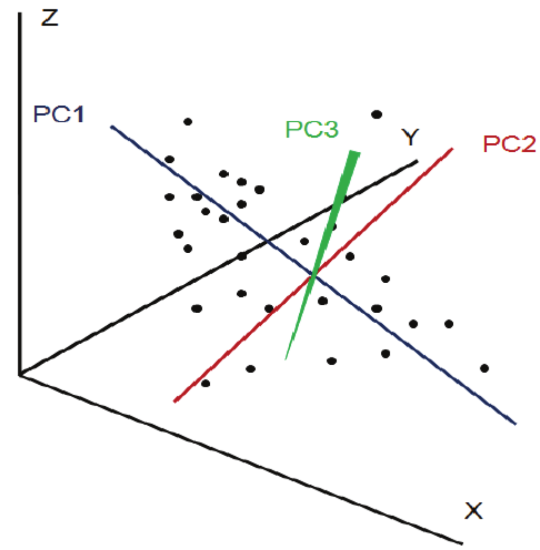
1. Principal Components Analysis





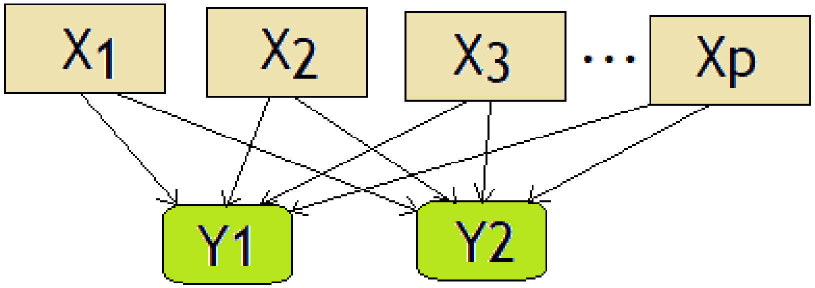
• **Unsupervised learning** is a machine learning technique in which the dataset has no target variable or no response value-Y.

→ Example unsupervised learning: **principal components analysis (PCA)**

→ PCA is a standard technique for visualizing high dimensional data and for data pre-processing. PCA reduces the dimensionality (the number of variables) of a data set by maintaining as much variance as possible.

(1) Formulation

PCA uses an **orthogonal transformation** to convert a set of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.



The first principal component (Y_1) is given by the linear combination of the variables $X_1, X_2, \ldots, X_p$

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.

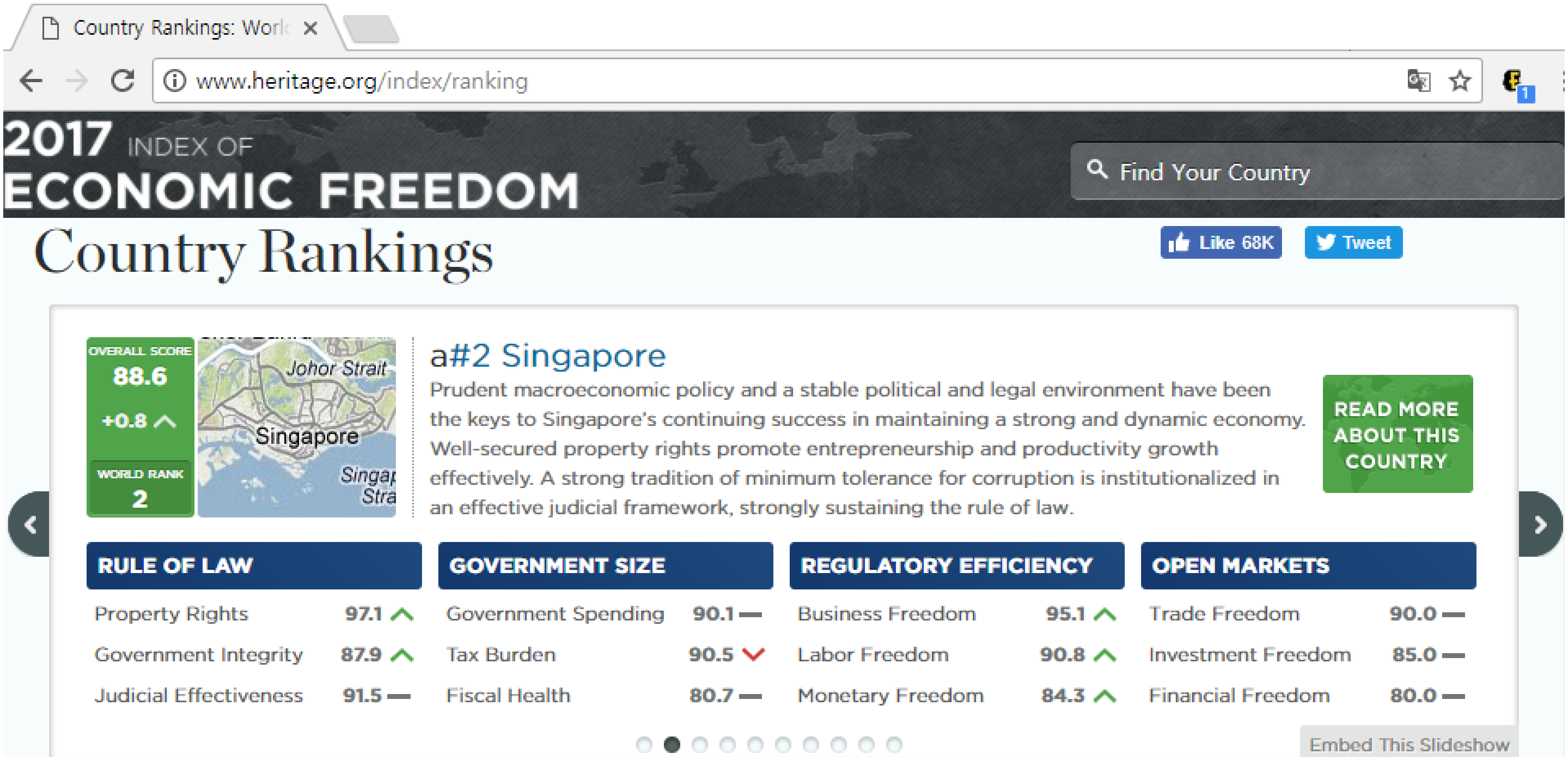
$$Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

Collectively, all of these transformations of the original variables to the principal components is

$$Y = AX$$

[1] https://onlinecourses.science.psu.edu/stat505/node/51/
[2] https://www.neuraldesigner.com/blog/principal-components-analysis

(2) Sample Data : Economic Freedom Dataset




```
> edat17 <- read.csv('index2017_data.csv', header=TRUE)
> names(edat17)
 [1] "CountryName"          "WorldRank"                "Score"
 [4] "PropertyRights"       "JudicalEffectiveness"     "FiscalHealth"
 [7] "BusinessFreedom"      "LaborFreedom"             "MonetaryFreedom"
[10] "TradeFreedom"         "InvestmentFreedom"        "FinancialFreedom"
```

(2) Sample Data : Economic Freedom Dataset

```
> edat <- edat17[,-c(1:2)] #exclude 1:2 columns
> str(edat)
'data.frame':   186 obs. of  10 variables:
 $ Score                 : Factor w/ 142 levels "27","33.9","4.9",..: 18 91 9
 $ PropertyRights        : Factor w/ 119 levels "12.3","12.6",..: 2 56 26 21 1
 $ JudicalEffectiveness: Factor w/ 138 levels "10","10.3","12.9",..: 27 28 3
 $ FiscalHealth          : Factor w/ 153 levels "0","1.2","10.6",..: 140 40 15
 $ BusinessFreedom       : num  54.2 79.3 62.1 58.5 57.3 78.5 89.3 76.9 71.5 6
 $ LaborFreedom          : num  59.9 50.7 49.5 40.4 46.1 72.4 84.1 67.6 75 71.
 $ MonetaryFreedom       : Factor w/ 137 levels "0","16.8","47.4",..: 26 93 19
 $ TradeFreedom          : Factor w/ 122 levels "0","47.8","50.5",..: 31 114 2
 $ InvestmentFreedom     : Factor w/ 21 levels "0","10","15",..: 1 15 7 6 11 1
 $ FinancialFreedom      : Factor w/ 11 levels "0","10","20",..: 1 8 4 5 6 8 1
```

edat[, c(1:4,7:10)] ~ factor

edat[, c(5:6)] ~ numeric

```
> # factor -> numeric
> for(i in c(1:4,7:10))
+   edat[,i] <- as.numeric(edat[,i])
```

(3) Assumptions

The three main assumptions a researcher should meet to conduct a PCA is related to ① sampling accuracy, ② sphericity, and ③ positive determinant of a correlation.

① Kaiser-Meyer-Olkin (KMO) Test (Sampling Adequacy)

This is a measure of how suited the data for analysis and quantifies the proportion of variance among variables that might be common variance. KMO value ranges: 0 ~ 1, where above 0.7 means an adequate sample.

paf(object, eigcrit=1, convcrit=.001) {rela} Principal Axis Factoring

```
> library(rela)
> paf_dat <- paf(as.matrix(edat))
> #KMO test
> paf_dat$KMO
[1] 0.86756
```

KMO = 0.868 is close to 1. We would conclude that n=186 with 10 variables is an adequate sample size.

② Bartlett's Test (Sphericity)

Bartlett's Test is concerned with determining if data samples are from populations with equal variances. The sphericity assumption is tested using the Bartlett's chi-square test, which checks whether an identity matrix (diagonal terms=1, off-diagonal terms=0) is present.

```
> #Bartlett test
> paf_dat$Bartlett
[1] 1295.3
```

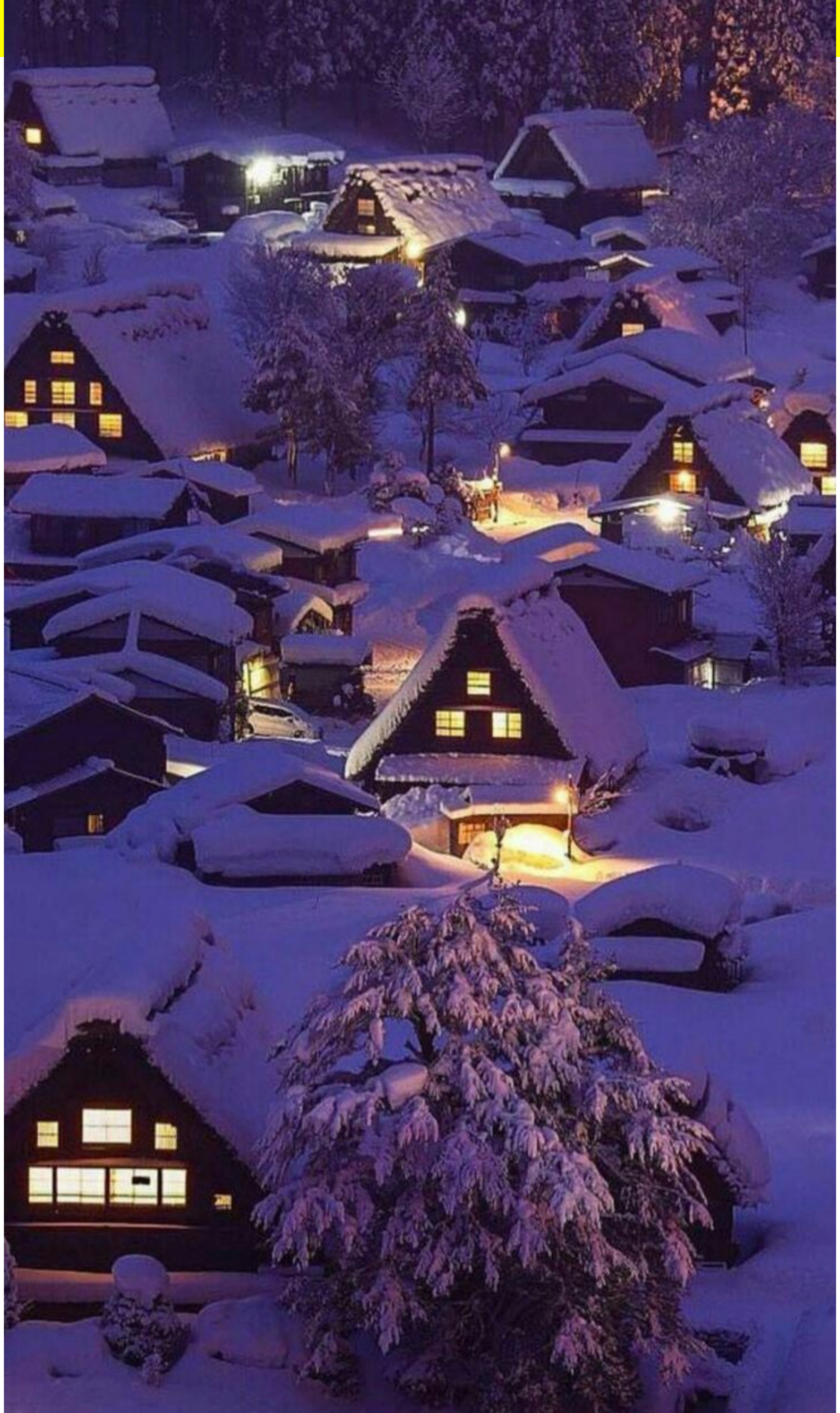
cortest.bartlett(R, n = NULL, diag=TRUE) {psych}

More useful for pedagogical purposes than actual applications.

The Bartlett test is asymptotically chi square distributed.

```
> library(psych)
> pcacor <- cor(edat)
> cortest.bartlett(pcacor, n=186)
$chisq
[1] 1295.3

$p.value
[1] 6.4916e-242

$df
[1] 45
```

Since the reported chi-square value is high and *p*-value<0.05, we reject the null hypothesis. The dataset is considered suitable for PCA.



③ Determinant of Correlation Matrix

If the determinant is zero, then a factor analytic solution cannot be obtained.

det(x, ...) {base}
Calculate the Determinant of a Matrix

Matrix

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

Determinant

$$\det A = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$
$$= a(ei - fh) - b(di - fg) + c(dh - eg)$$
$$= aei + bfg + cdh - ceg - bdi - afh.$$

```
> det(pcacor)
[1] 0.0007748
```

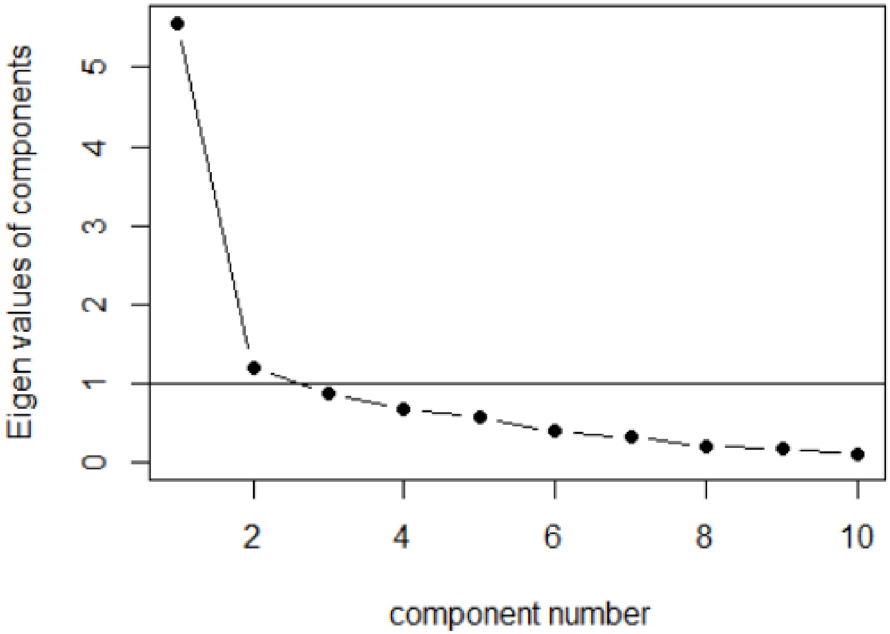
Dataset satisfies the assumptions for conducting PCA.

(4) Number of Components

In order to determine the number of principal components that need to be retained, we can draw a scree plot.

scree(rx, factors=TRUE, pc=TRUE) {psych}
Cattell's scree test is one of most simple ways of testing the number of components or factors in a correlation matrix. Here we plot the eigen values of a correlation matrix as well as the eigen values of a factor analysis.

```
library(psych)
scree(edat, factors=FALSE, pc=TRUE)
```



General rule is to select

$$\text{eigenvalue} \geq 1$$

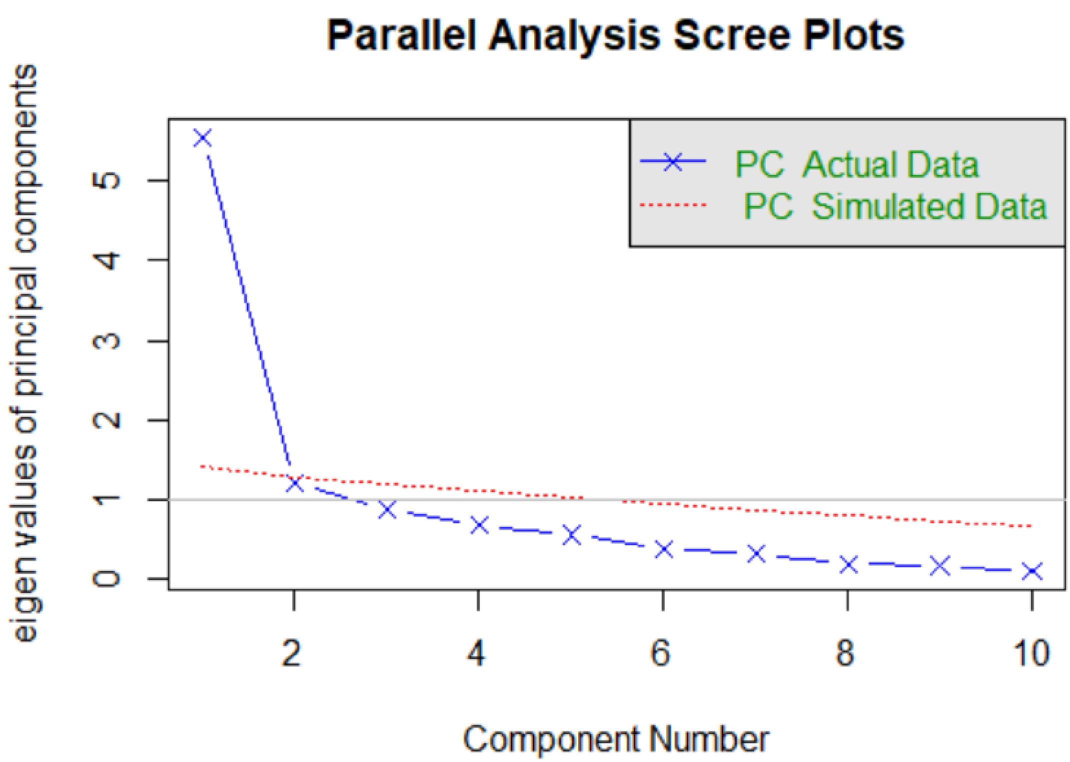
This scree plot indicates that **2 components** are present.

(4) Number of Components

fa.parallel(x, n.obs, fm, fa, ...) {psych}

Sharp breaks in the scree plot suggest the appropriate number of components or factors to extract. “Parallel" analyis is an alternative technique that compares the scree of factors of the observed data with that of a random data matrix of the same size as the original.

```
nc <- dim(edat)[1]
#fm="pa": principal factor solution
#fa="pc": principal components
fa.parallel(edat,n.obs=nc, fm="pa", fa="pc")
abline(h=1, col="grey")
```

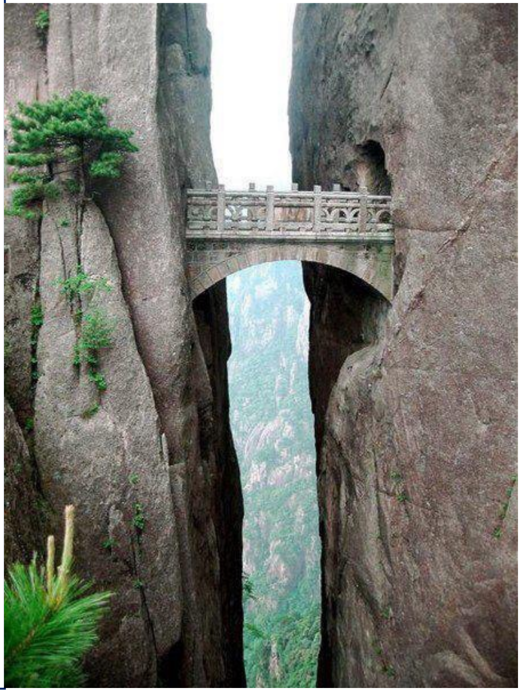


(5) Principal Components Loading

principal(r, nfactors=1, rotate="varimax", n.obs=NA, scores=TRUE, ...) {psych}
Does an eigen value decomposition and returns eigen values, loadings, and degree of fit for a specified number of components.

```
> library(psych)
> pca <- principal(edat, nfactors=2, rotate='none')
> pca
                        PC1   PC2   h2   u2 com
Score                  0.93  0.04 0.87 0.13 1.0
PropertyRights         0.88  0.14 0.79 0.21 1.0
JudicalEffectiveness   0.77  0.21 0.63 0.37 1.2
FiscalHealth           0.16  0.80 0.67 0.33 1.1
BusinessFreedom        0.80  0.19 0.68 0.32 1.1
LaborFreedom           0.54  0.28 0.37 0.63 1.5
MonetaryFreedom        0.70 -0.12 0.51 0.49 1.1
TradeFreedom           0.79  0.01 0.63 0.37 1.0
InvestmentFreedom      0.76 -0.46 0.80 0.20 1.6
FinancialFreedom       0.82 -0.39 0.81 0.19 1.4
                        PC1  PC2
SS loadings            5.56 1.20
Proportion Var         0.56 0.12
Cumulative Var         0.56 0.68
Proportion Explained   0.82 0.18
Cumulative Proportion  0.82 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 components are sufficient.
```

h2: communality
u2: residual
com: complexity



The eigenvalues for the 2 principal components are given in descending order: PC1=5.56(82%) and PC2=1.20(18%). The sum of eigenvalues explained variance equals 100%.

Interpretation of the Principal Components

You need to determine at what level the correlation value will be of phenomenological importance.
In general, a correlation value above **0.5** will be deemed important.

https://online.stat.psu.edu/~ajw13/stat505/fa06/16_princomp/06_princomp_interpret.html

The principal component equation to generate the scores is computed using the first set of weights.

$$Y1 = 0.93\,X1 + 0.88\,X2 + 0.77\,X3 + 0.16\,X4 + 0.80\,X5 + 0.54\,X6 + 0.70\,X7 + 0.79\,X8 + 0.76\,X9 + 0.82\,X10$$

where X1=Score, X2=PropertyRights, X3=JudicalEffectiveness,
X4=FiscalHealth, X5=BusinessFreedom, X6=LaborFreedom,
X7=MonetaryFreedom, X8=TradeFreedom, X9=InvestFreedom,
X10=FinancialFreedom

(6) Cronbach's Alpha Reliability Coefficient

alpha(x) {psych}
Find two estimates of reliability: Cronbach's alpha and Guttman's Lambda 6.
Internal consistency measures of reliability.
x : A data.frame or matrix of data, or a covariance or correlation matrix

Cronbach's alpha	Internal consistency
α ≥ 0.9	Excellent
0.9 > α ≥ 0.8	Good
0.8 > α ≥ 0.7	Acceptable
0.7 > α ≥ 0.6	Questionable
0.6 > α ≥ 0.5	Poor
0.5 > α	Unacceptable

https://en.wikipedia.org/wiki/Cronbach%27s_alpha

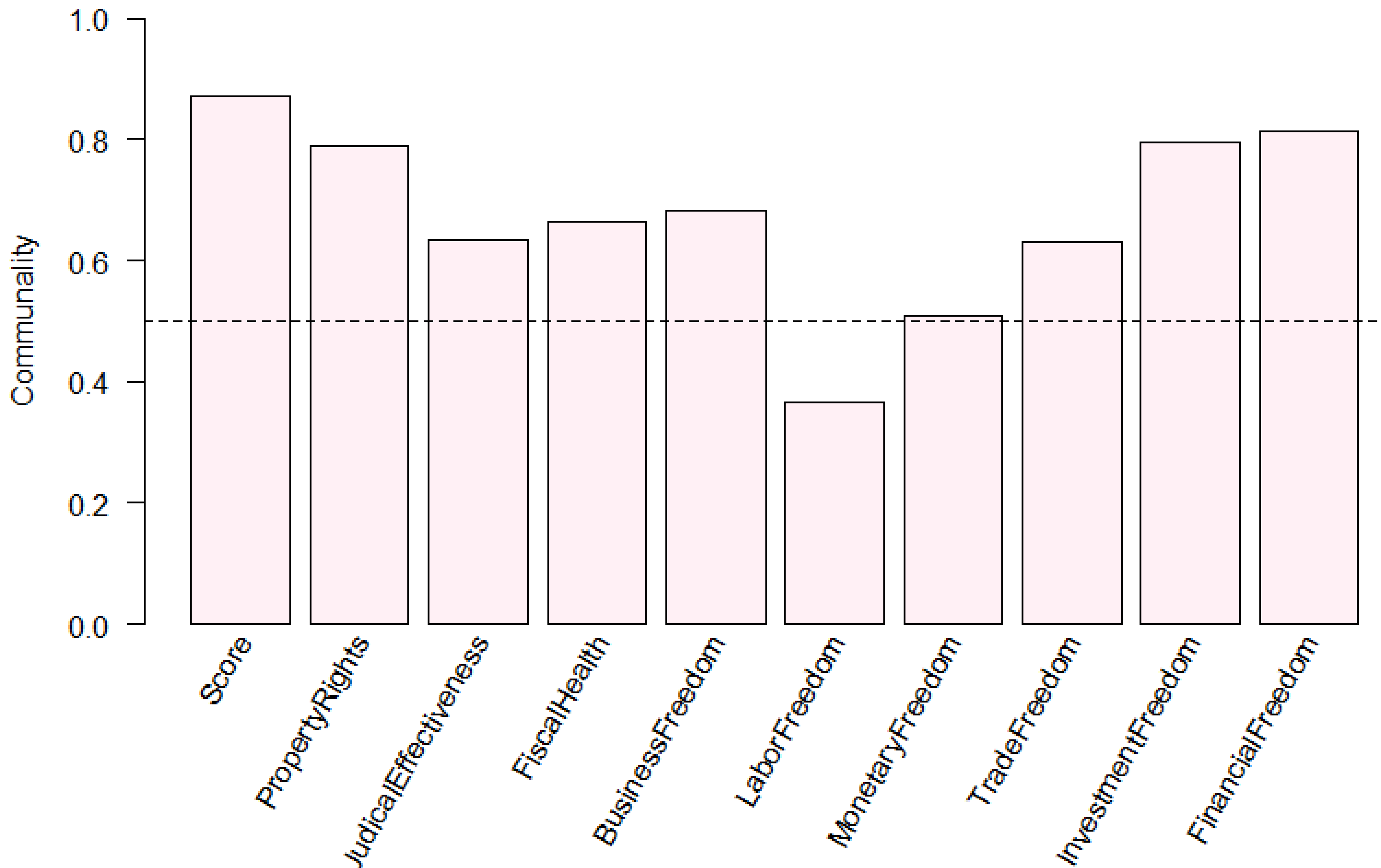
```
> alpha(pcacor)

Reliability analysis
Call: alpha(x = pcacor)

  raw_alpha std.alpha G6(smc) average_r S/N
       0.9       0.9    0.93      0.46 8.7
```

A check of the Cronbach's alpha reliability coefficient indicates high internal consistency of response (α=0.9); thus it does not affect the PCA results.

(7) Communality

A communality is the extent to which an item correlates with **all** other items.
Higher communalities are better.

```
> pca$communality #h2
              Score      PropertyRights JudicalEffectiveness          FiscalHealth
          0.8718668           0.7881048            0.6339253             0.6652405
    BusinessFreedom         LaborFreedom      MonetaryFreedom          TradeFreedom
          0.6816035           0.3656683            0.5078353             0.6309383
  InvestmentFreedom    FinancialFreedom
          0.7952352           0.8136502
```



(8) Biplots

biplot.psych {psych}

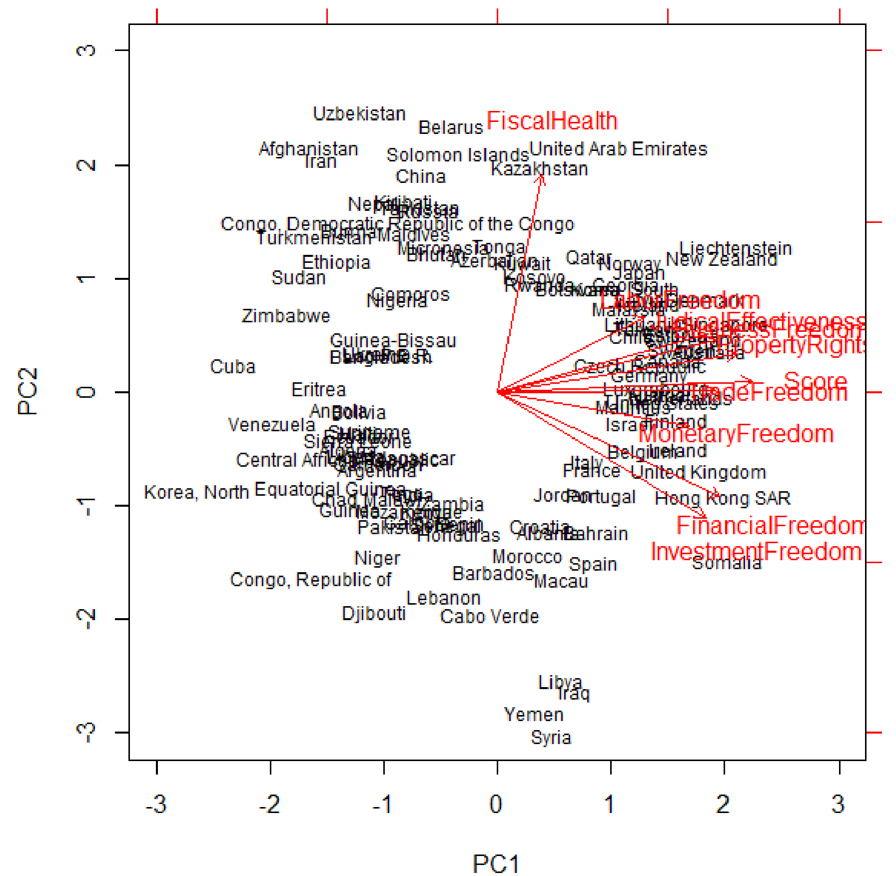
Draw biplots of factor or component scores by factor or component loadings

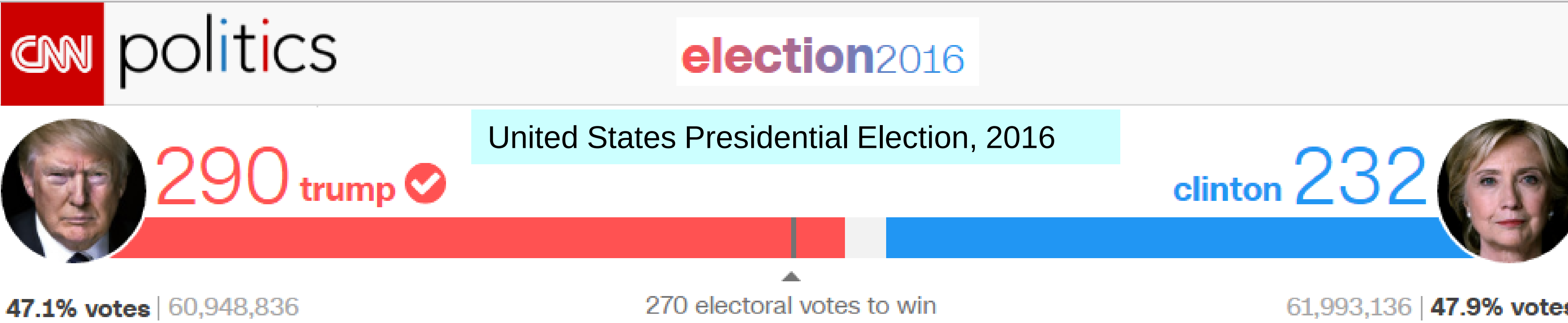
S3 method for class 'psych'

biplot(x, labels, cex=c(.75,1), arrow.len, ...)

```
library(psych)
pca <- principal(edat, nfactors=2, rotate='none')
biplot.psych(pca, col=c("black","red"), cex=c(0.5,1), arrow.len=0.08,
             main=NULL, labels=edat17[,1])
```

These vectors are pinned at the origin of PCs (PC1 = 0 and PC2 = 0). Their project values on each PC show how much weight they have on that PC. For example, Score and X-Freedom strongly influence PC1, while FiscalHealth have more influence on PC2. When two vectors are close, the two variables they represent are positively correlated. If two vectors form 90° angle, they are not likely to be correlated.

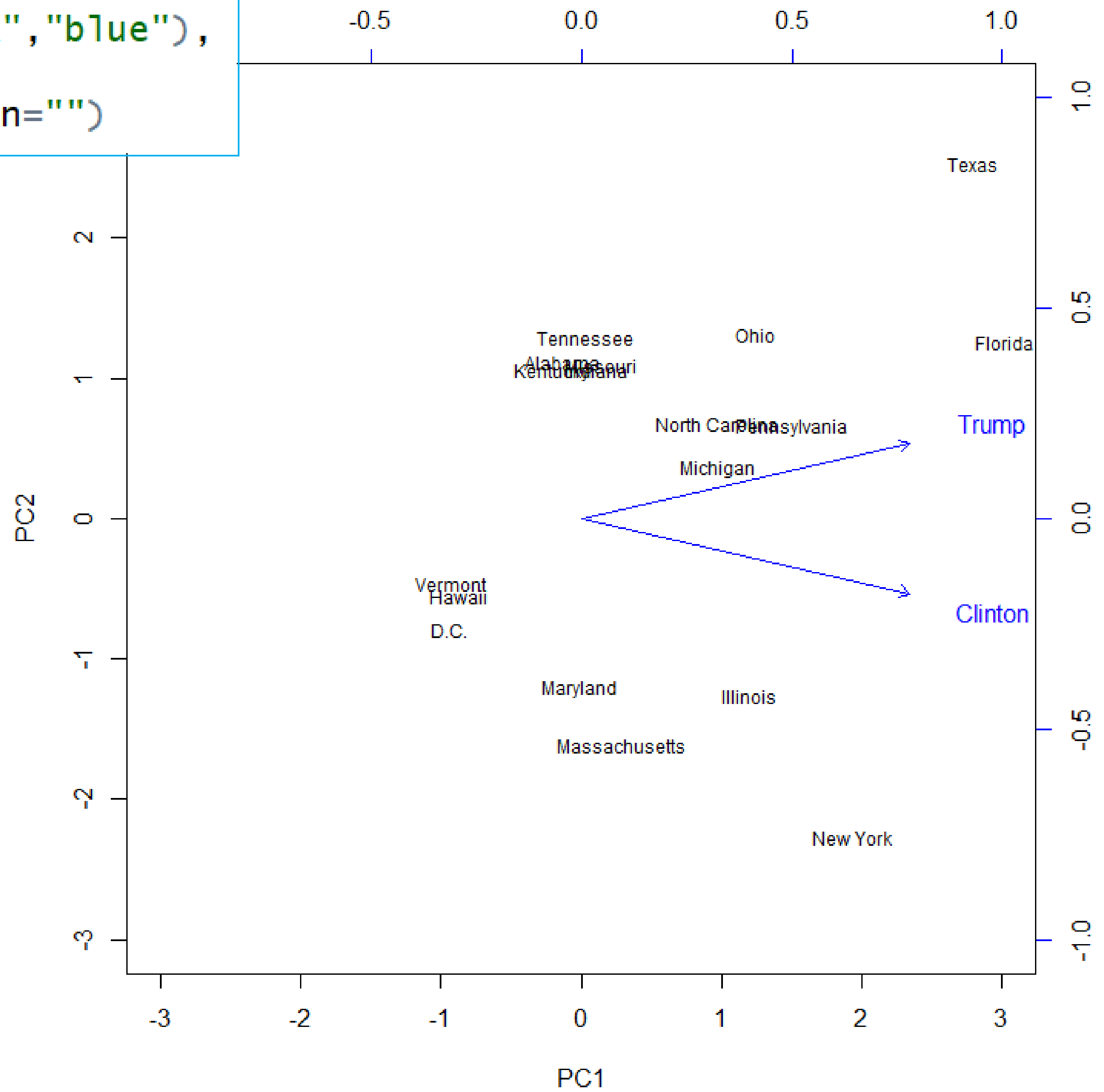





```
> # Data Source for Final vote count
> # https://en.wikipedia.org/wiki/United_States_presidential_election,_2016
> dat <- read.csv("USA2016PresidentElection.csv",header=T)
> head(dat,4)
     State Clinton   Trump Johnson Stein
1  Alabama  718084 1306925   43869  9287
2   Alaska   93007  130415   14593  4445
3  Arizona  888374  972900   75082 23697
4 Arkansas  378632  681765   29593  9406
> xd <- dat[,-c(1,4,5)] #exclude state names and two candidates
> library(psych)
> pcusa <- principal(xd, nfactors=2, rotate="none")
> pcusa$loadings

Loadings:
        PC1    PC2
Clinton  0.975 -0.222
Trump    0.975  0.222

                 PC1   PC2
SS loadings    1.901 0.099
Proportion Var 0.951 0.049
Cumulative Var 0.951 1.000
```

Let's draw a biplot of principal component scores for the United States presidential election in 2016.

```
biplot(pcusa,col=c("black","blue"),
       arrow.len=0.08,
       labels=dat[,1],main="")
```



2. Correspondence Analysis

Correspondence analysis (**CA**) is conceptually similar to principal component analysis, but applies to **categorical** rather than continuous data.

ca(obj, ...) {ca}
Computation of simple correspondence analysis.

data(**smoke**) {ca}
Artificial smoke dataset in Greenacre (1984)
Table containing 5 rows (staff group) and 4 columns (smoking categories), giving the frequencies of smoking categories in each staff group in a fictional organization.
[Reference] M. J. Greenacre, Theory and Applications of Correspondence Analysis(Academic Press, London, 1984).

```
> library(ca)
> data(smoke)
> smoke
   none light medium heavy
SM    4     2      3     2
JM    4     3      7     4
SE   25    10     12     4
JE   18    24     33    13
SC   10     6      7     2
```

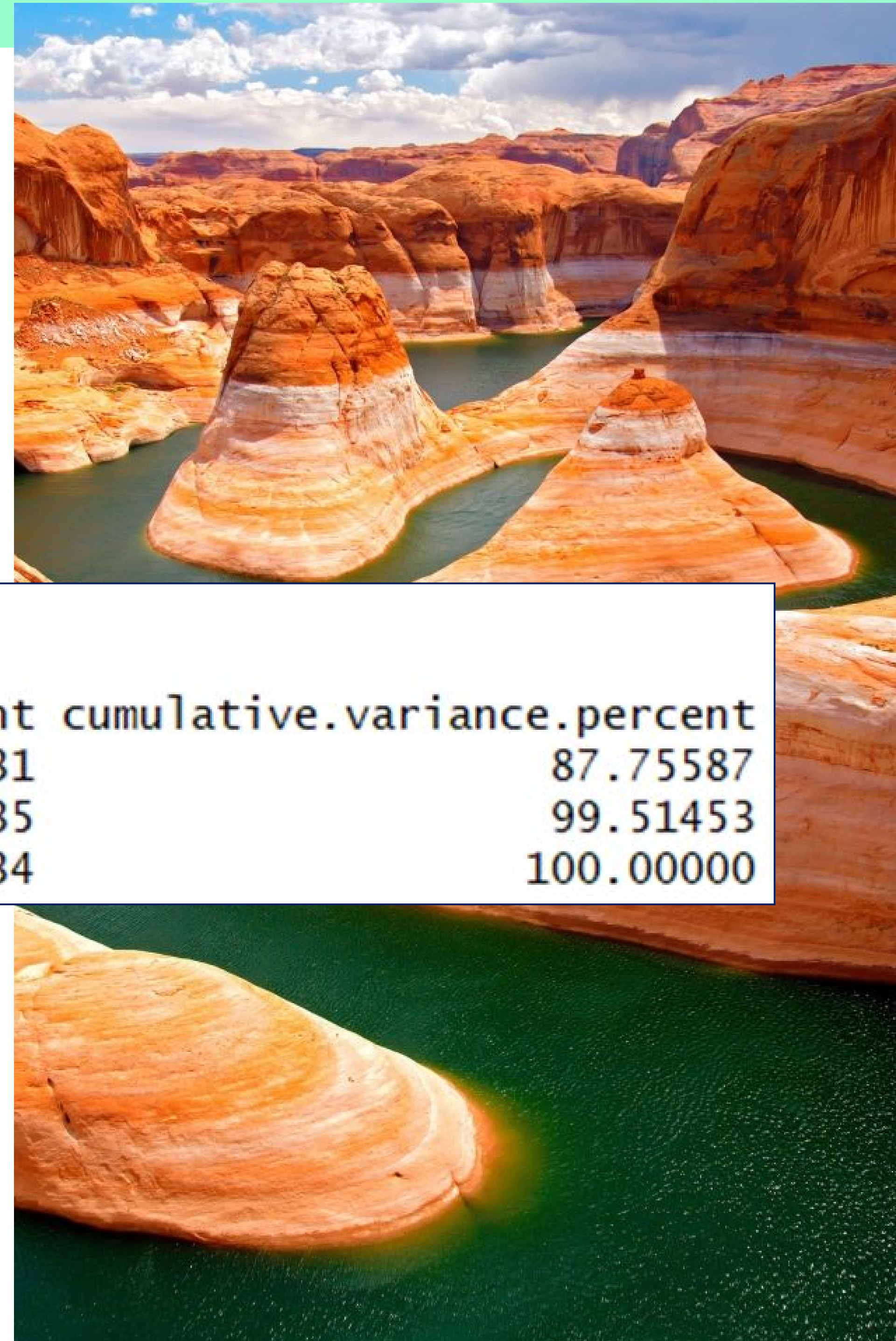
(1) Eigenvalues

How many dimensions are sufficient for the data interpretation? The number of dimensions to retain in the solution can be determined by examining the table of eigenvalues.

```
> smoke_ca <- ca(smoke)
> #content of result object
> names(smoke_ca)
 [1] "sv"         "nd"         "rownames"   "rowmass"    "rowdist"    "rowinertia"
 [7] "rowcoord"   "rowsup"     "colnames"   "colmass"    "coldist"    "colinertia"
[13] "colcoord"   "colsup"     "N"          "call"
> summary(smoke_ca)

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.074759  87.8  87.8  **********************
 2      0.010017  11.8  99.5  ***
 3      0.000414   0.5 100.0
        -------- -----
 Total: 0.085190 100.0
```

(1) Eigenvalues

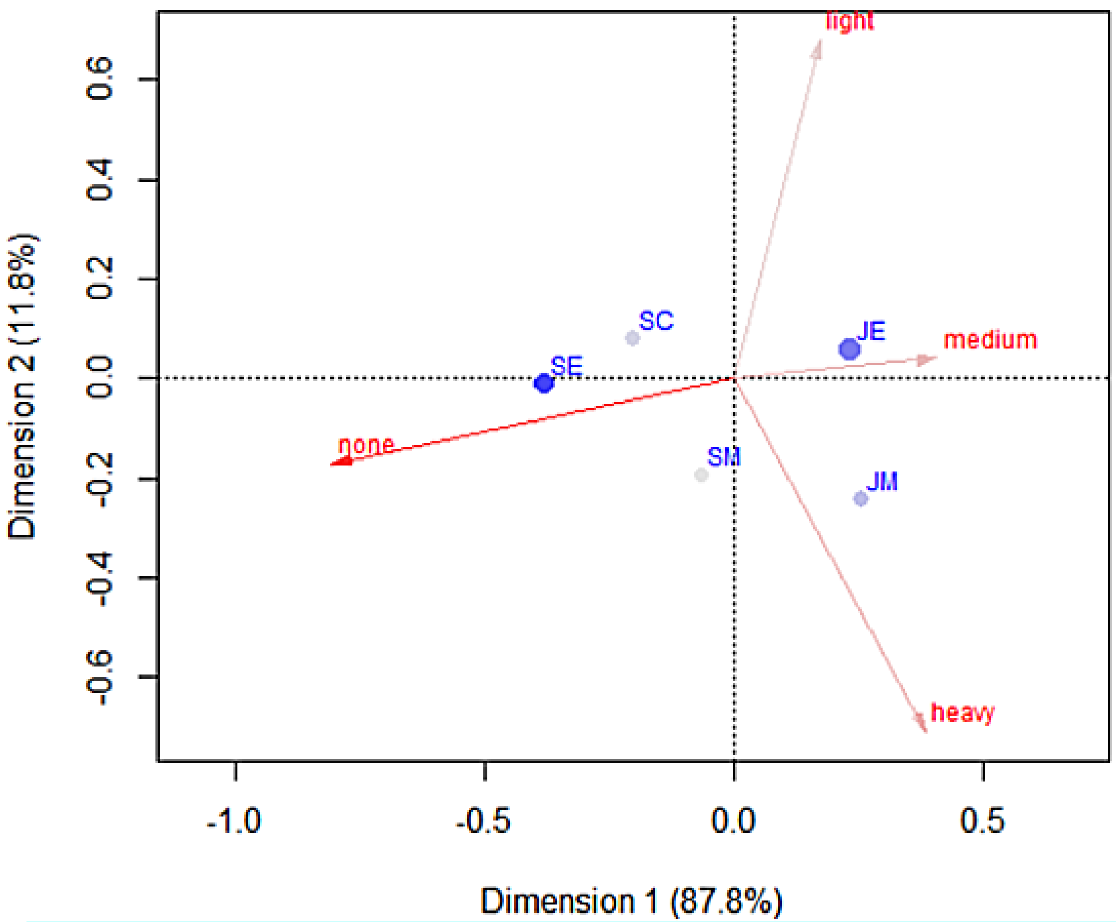
The proportion of variances retained by the different dimensions (axes) can be also extracted using the function **get_eigenvalue()** in *factoextra.*



```
> library(factoextra)
> get_eigenvalue(smoke_ca)
        eigenvalue variance.percent cumulative.variance.percent
Dim.1 0.0747591059       87.7558731                    87.75587
Dim.2 0.0100171805       11.7586535                    99.51453
Dim.3 0.0004135741        0.4854734                   100.00000
```

(2) CA Biplot


```
plot(smoke_ca, mass=TRUE, contrib="absolute",
     map="rowgreen", arrows=c(FALSE, TRUE))
```



The distance between any row points or column points gives a measure of their similarity (or dissimilarity).
The distance between any row and column items is not meaningful! You can only make a general statements about the observed pattern.

- Columns (smoking categories) are represented by arrows.
- Point intensity (shading) corresponds to the absolute contributions for the rows (staff group). **JM group corresponds to heavy smoke, JE to medium, and SE to none.**