Contents

1 Dimensionality Reduction

2 Stepwise Regression (using `R`

SPSS can be very opaque in determining how particularly statistical routines are carried out. Conversely the statistical programming language `R` is usually quite clear, once a familiarity with the language has been developed.

For variable selection procedures, `R` used the AIC criterion. When comparing multiple candidate models, the candidate model with the lowest AIC value is the best model. We will use `R` output to revise variable selection procedures. Recall that we used the ***mtcars*** data set. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (197374 models). For this data, we tried to determine the optimal set of independent variables to predict the dependent variables ***mpg*** (miles per gallon).

cyl Number of cylinders

disp Displacement (cu.in.)

hp Gross horsepower

drat Rear axle ratio

wt Weight (lb/1000)

qsec 1/4 mile time

vs V/S

am Transmission (0 = automatic, 1 = manual)

gear Number of forward gears

carb Number of carburetors

2.1 Backward Elimination

The initial model contains all of the independent variables. Candidate models, whereby each of the independent variables are individually removed from the model are fitted. The AIC value for each reduced model is computed. The unreduced model is also used for comparison. The AIC values are tabulated to determine which removal results in the lowest AIC value. In this first case, the removal of cyl would reduced the AIC value from 70.898 (see bottom row) to 68.915. Thus the independent variable cyl is removed from the set of IVs.

```
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280
```

In the second phase, the process is repeated. This time removing vs results in an AIC value of 66.973. It is then removed from the set of IVs. For this phase, the unreduced model is the model fitted by all independent variables except cyl, which was removed in the previous phase.

```
Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
```

```
       Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
- wt    1   27.0280 174.60 72.297
```

this process continues until the removal of an IV will not results in an improvement in AIC. This is indicated by having the $< none >$ (i.e unreduced model) having the lowest AIC value. At the end of the output is the optimal model, according to the backward elimination procedure, using the IVs : am , qsec and wt.

```
Step:  AIC=61.31
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC
<none>              169.29 61.307
- am    1    26.178 195.46 63.908
- qsec  1   109.034 278.32 75.217
- wt    1   183.347 352.63 82.790
```

```
Call:
lm(formula = mpg ~ wt + qsec + am)

Coefficients:
(Intercept)           wt         qsec           am
      9.618       -3.917        1.226        2.936
```

2.2 Stepwise Regression

Stepwise Regression differs from Backward Elimination, in that it allows IVs to be re-introduced. Hence the + signs from the second phase onwards.

```
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280

Step:  AIC=68.92
```

```
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
+ cyl   1    0.0799 147.49 70.898
- wt    1   27.0280 174.60 72.297
```

Again, the procedure finishes when it is found that the unchanged model has the lowest of all possible AIC values.

```
Step:  AIC=61.31
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC
<none>              169.29 61.307
+ hp    1     9.219 160.07 61.515
+ carb  1     8.036 161.25 61.751
+ disp  1     3.276 166.01 62.682
+ cyl   1     1.501 167.78 63.022
```

```
+ drat  1     1.400 167.89 63.042
+ gear  1     0.123 169.16 63.284
+ vs    1     0.000 169.29 63.307
- am    1    26.178 195.46 63.908
- qsec  1   109.034 278.32 75.217
- wt    1   183.347 352.63 82.790

Call:
lm(formula = mpg ~ wt + qsec + am)

Coefficients:
(Intercept)           wt         qsec           am
      9.618       -3.917        1.226        2.936
```

2.3 Scree plot

A scree plot displays the proportion of the total variation in a dataset that is explained by each of the components in a principle component analysis. It helps you to identify how many of the components are needed to summarise the data.

To create a scree plot of the components, use the screeplot function.

```
screeplot(modelname)
```

where modelname is the name of a previously saved principle component analysis, created with the princomp function as explained in the article Performing a principle component analysis in R.

Example: Scree plot for the iris dataset

To create a scree plot of the components, use the command:

```
screeplot(irispca)
```

The result is shown below.

From the scree plot we can see that the amount of variation explained drops dramatically after the first component. This suggests that just one component may be sufficient to summarise the data.

3 Sampling adequacy (KMO Statistic)

- Measured by the Kaiser-Meyer-Olkin (KMO) statistics, sampling adequacy predicts if the analyses are likely to perform well, based on correlation and partial correlation. KMO can also be used to assess which variables to drop from the model because they are too multi-collinear.
- There is a KMO statistic for each individual variable, and their sum is the KMO overall statistic. KMO varies from 0 to 1.0 and KMO overall should be 0.60 or higher to proceed with factor analysis. Values below 0.5 imply that factor analysis or PCA may not be appropriate. (Approach to overcoming this: If it is not, drop the **indicator variables** with the lowest individual KMO statistic values, until KMO overall rises above 0.60.)
- Kaiser-Meyer-Olkin To compute KMO overall, the numerator is the sum of squared correlations of all variables in the analysis (except the 1.0 self-correlations of variables with themselves, of course).
- The denominator is this same sum plus the sum of squared partial correlations of each variable i with each variable j, controlling for others in the analysis. The concept is that the partial correlations should not be very large if one is to expect distinct factors to emerge from factor analysis.

3.1 Bartlett's Test for Sphericity

Bartlett's measure tests the null hypothesis that the original correlation matrix is an identity matrix. For PCA and factor analysis to work we need some relationships between variables and if the R- matrix were an identity matrix then all correlation coefficients would be zero. Therefore, we want this test to be signifcant (i.e. have a significance value less than 0.05). A significant test tells us that the correlation matrix is not an identity matrix; therefore, there are some relationships between the variables we hope to include in the analysis.

For these data, Bartlett's test is highly significant (p ¡ 0.001), and therefore factor analysis is appropriate.

4 FactoMineR

The ***FactoMineR*** package offers a large number of additional functions for exploratory factor analysis. This includes the use of both quantitative and qualitative variables, as well as the inclusion of supplementary variables and observations. Here is an example of the types of graphs that you can create with this package.

```
# PCA Variable Factor Map
library(FactoMineR)
result <- PCA(mydata)
# graphs generated automatically
```

The GPARotation package offers a wealth of rotation options beyond varimax and promax.

5 Principal Component Analysis

5.1 Introduction

- Principal component analysis is appropriate when you have obtained measures on a number of (possibly) correlated observed variables and wish to develop a smaller number of **artificial uncorrelated variables** called **principal components** that will account for most of the variance in the observed variables.

- The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

- The principal components may then be used as predictor or criterion variables in subsequent analyses.

- Traditionally, principal component analysis is performed can be performed on raw data, on the symmetric **Covariance matrix** or on the symmetric **Correlation matrix**.
  (The covariance matrix contains scaled sums of squares and cross products. A correlation matrix is like a covariance matrix but first the variables, i.e. the columns, have been standardized.)

- If raw data is used, the procedure will create the original correlation matrix or covariance matrix, as specified by the user.

- If the correlation matrix is used, the variables are standardized and the total variance will equal the number of variables used in the analysis (because each standardized variable has a variance equal to 1).

- If the covariance matrix is used, the variables will remain in their original metric. However, one must take care to use variables whose variances and scales are similar.

- Unlike **factor analysis**, which analyzes the common variance, the original matrix in a principal components analysis analyzes the total variance.
- Also, principal components analysis assumes that each original measure is collected without measurement error.

6 Mathematical background of PCA

- Principal Component Analysis is a linear **dimensionality reduction** technique, which identifies orthogonal directions of maximum variance in the original data, and projects the data into a lower-dimensionality space formed of a sub-set of the highest-variance components (Bishop, 1995).
- The mathematical technique used in PCA is called **eigen-analysis**. Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables.
- Software packages compute solutions for these weights by using a special type of equation called an ***eigenequation***. The weights produced by these eigenequations are optimal weights in the sense that, for a given set of data, no other set of weights could produce a set of components that are more successful in accounting for variance in the observed variables.
- The weights are created so as to satisfy a principle of least squares that is similar (but not identical) to the principle of least squares used in multiple regression.
- ***Remarks***
  - The words ***linear combination*** refer to the fact that scores on a component are created by adding together scores on the observed variables being analyzed.
  - ***Optimally weighted*** refers to the fact that the observed variables are weighted in such a way that the resulting components account for a maximal amount of variance in the data set.

6.1 Number of Extracted Components

- The preceding section may have created the impression that, if a principal component analysis were performed on data from the 7-item job satisfaction questionnaire, only two components would be created. However, such an impression would not be entirely correct.
- In reality, the number of components extracted in a principal component analysis ***is equal*** to the number of observed variables being analyzed. This means that an analysis of your 7-item questionnaire would actually result in seven components, not two.
- However, in most analyses, only the first few components account for meaningful amounts of variance, so only these first few components are retained, interpreted, and used in subsequent analyses (such as in multiple regression analyses).
- For example, in your analysis of the 7-item job satisfaction questionnaire, it is likely that only the first two components would account for a meaningful amount of variance; therefore only these would be retained for interpretation.
- You would assume that the remaining five components accounted for only trivial amounts of variance. These latter components would therefore not be retained, interpreted, or further analyzed.

6.2 Characteristics of Principal Components

- The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables. Under typical conditions, this means that the first component will be correlated with at least some of the observed variables. In fact it may be correlated with many.
- The second component extracted will have two important characteristics. First, this component will account for a maximal amount of variance in the data set that was not accounted for by the first component. Again under typical conditions, this means that the second component will be correlated with some of the observed variables that did not display strong correlations with component 1.
- The second characteristic of the second component is that it will be uncorrelated with the first component. If you were to compute the correlation between components 1 and 2, that correlation should be close to zero.
- The remaining components that are extracted in the analysis display the same two characteristics: each component accounts for a maximal amount of variance in the observed variables that was not accounted for by the preceding components, and is uncorrelated with all of the preceding components.
- A principal component analysis proceeds in this fashion, with each new component accounting for progressively smaller and smaller amounts of variance (this is why only the first few components are usually retained and interpreted).
- When the analysis is complete, the resulting components will display varying degrees of correlation with the observed variables, but are completely uncorrelated with one another.

6.3 Total Variance in the context of PCA

- To understand the meaning of **total variance** as it is used in a principal component analysis, remember that the observed variables are standardized in the course of the analysis. This means that each variable is transformed so that it has a mean of zero and a variance of one.
- The total variance in the data set is simply the sum of the variances of these observed variables.
- Because they have been standardized to have a variance of one, each observed variable contributes one unit of variance to the total variance in the data set.
- Because of this, the total variance in a principal component analysis ***will always be equal*** to the number of observed variables being analyzed.
- For example, if seven variables are being analyzed, the total variance will equal seven. The components that are extracted in the analysis will partition this variance: perhaps the first component will account for 3.2 units of total variance; perhaps the second component will account for 2.1 units.
- The analysis continues in this way until all of the variance in the data set (i.e. the remaining 1.7 units). has been accounted for.

7 Orthogonal versus Oblique Solutions

- This course will discuss only principal component analysis that result in **orthogonal solutions**. An orthogonal solution is one in which the components remain uncorrelated (orthogonal means uncorrelated).
- It is possible to perform a principal component analysis that results in correlated components.
- Such a solution is called an **oblique solution**. In some situations, oblique solutions are superior to orthogonal solutions because they produce cleaner, more easily-interpreted results.
- However, oblique solutions are also somewhat more complicated to interpret, compared to orthogonal solutions. For this reason, we will focus only on the interpretation of orthogonal solutions

8 Determining the Number of Meaningful Components to Retain

Previously it was stated that the number of components extracted is equal to the number of variables being analyzed, necessitating that you decide just how many of these components are truly meaningful and worthy of being retained for further analysis.

In general, you expect that only the first few components will account for meaningful amounts of variance, and that the later components will tend to account for only trivial variance.

The next step of the analysis,therefore, is to determine how many meaningful components should be retained for interpretation. The followings section will describe four criteria that may be used in making this decision:

- the eigenvalue-one criterion,
- the scree test,
- the proportion of variance accounted for,
- the interpretability criterion.

8.1 The Eigenvalue-One Criterion

- In principal component analysis, one of the most commonly used criteria for solving the number-of-components problem is the eigenvalue-one criterion, also known as the **Kaiser criterion** (Kaiser, 1960). With this approach, you retain and interpret any component with an eigenvalue greater than 1.00.
- The rationale for this criterion is straightforward. Each observed variable contributes one unit of variance to the total variance in the data set. Any component that displays an eigenvalue greater than 1.00 is accounting for a greater amount of variance than had been contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance, and is worthy of being retained.

- On the other hand, a component with an eigenvalue less than 1.00 is accounting for less variance than had been contributed by one variable. The purpose of principal component analysis is to reduce a number of observed variables into a relatively smaller number of components; this cannot be effectively achieved if you retain components that account for less variance than had been contributed by individual variables. For this reason, components with eigenvalues less than 1.00 are viewed as trivial, and are not retained.

8.2 Advantages and Disadvantages

- The eigenvalue-one criterion has a number of positive features that have contributed to its popularity. Perhaps the most important reason for its widespread use is its simplicity: You do not make any subjective decisions, but merely retain components with eigenvalues greater than one.

- On the positive side, it has been shown that this criterion very often results in retaining the correct number of components, particularly when a small to moderate number of variables are being analyzed and the variable communalities are high. Stevens (1986) reviews studies that have investigated the accuracy of the eigenvalue-one criterion, and recommends its use when less than 30 variables are being analyzed and communalities are greater than .70, or when the analysis is based on over 250 observations and the mean communality is greater than or equal to 0.60.

- There are a number of problems associated with the eigenvalue-one criterion, however. As was suggested in the preceding paragraph, it can lead to retaining the wrong number of components under circumstances that are often encountered in research (e.g., when many variables are analyzed, when communalities are small).

- Also, the mindless application of this criterion can lead to retaining a certain number of components when the actual difference in the eigenvalues of successive components is only trivial. For example, if component 2 displays an eigenvalue of 1.001 and component 3 displays an eigenvalue of 0.999, then component 2 will be retained but component 3 will not; this may mislead you into believing that the third component was meaningless when, in fact, it accounted for almost exactly the same amount of variance as the second component.

- In short, the eigenvalue-one criterion can be helpful when used judiciously, but the thoughtless application of this approach can lead to serious errors of interpretation.

9 The Scree test

With the scree test (*Cattell, 1966*), you plot the eigenvalues associated with each component and look for a break between the components with relatively large eigenvalues and those with small eigenvalues.

The components that appear before the break are assumed to be meaningful and are retained for rotation; those apppearing after the break are assumed to be unimportant and are not retained.

Remark: The word scree refers to the loose rubble that lies at the base of a cliff. When performing a scree test, you normally hope that the scree plot will take the form of a cliff: At the top will be the eigenvalues for the few meaningful components, followed by a break (the edge of the cliff). At the bottom of the cliff will lie the scree: eigenvalues for the trivial components.

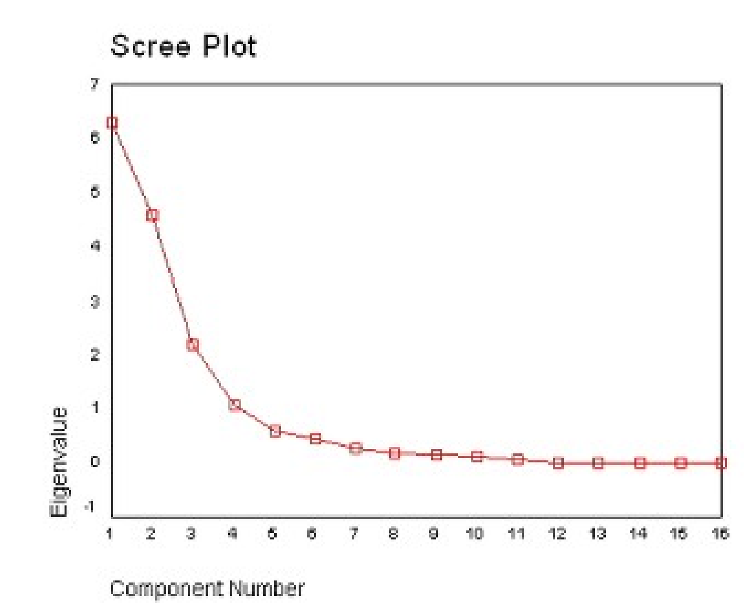

Figure 1: Example of a Scree Plot

Sometimes a scree plot will display several large breaks. When this is the case, you should look for the last big break before the eigenvalues begin to level off. Only the components that appear before this last large break should be retained.

9.0.1 Advantages and Disadvantages

The scree test can be expected to provide reasonably accurate results, provided the sample is large (over 200) and most of the variable communalities are large (Stevens, 1986). However, this criterion has its own weaknesses as well, most notably the ambiguity that is often displayed by scree plots under typical research conditions: Very often, it is difficult to determine exactly where in the scree plot a break exists, or even if a break exists at all.

9.1 Proportion of Variance Accounted For

A third criterion in solving the number of factors problem involves retaining a component if it accounts for a specified proportion (or percentage) of variance in the data set. For example, you may decide to retain any component that accounts for at least 5% or 10% of the total variance. This proportion can be calculated with a simple formula:

$$\text{Proportion} = \frac{\text{Eigenvalue for the component of interest}}{\text{Total eigenvalues of the correlation matrix}}$$

In principal component analysis, the total eigenvalues of the correlation matrix is equal to the total number of variables being analyzed (because each variable contributes one unit of variance to the analysis).

An alternative criterion is to retain enough components so that the cumulative percent of variance accounted for is equal to some minimal value. Suppose that,in a PCA procedure, that components 1, 2, 3, and 4 accounted for approximately 37%, 33%, 13%, and 10% of the total variance, respectively.

Suppose that it was required to account for 90% of the variance. Adding these percentages together results in a sum of 93%. This means that the cumulative percent of variance accounted for by components 1, 2, 3, and 4 is 93%.

	Eigenvalue	Difference	Proportion	Cumulative
1	3.88155846	2.24699126	0.4313	0.4313
2	1.63456720	0.57436630	0.1816	0.6129
3	1.06020090	0.10480537	0.1178	0.7307
4	0.95539554	0.42415968	0.1062	0.8369
5	0.53123586	0.10477119	0.0590	0.8959
6	0.42646467	0.13882496	0.0474	0.9433
7	0.28763971	0.16930381	0.0320	0.9752
8	0.11833590	0.01373414	0.0131	0.9884
9	0.10460176		0.0116	1.0000

Eigenvalues of the Correlation Matrix

Figure 2: Eigenvalue Table

9.1.1 Advantages and Disadvantages

The proportion of variance criterion has a number of positive features. For example, in most cases, you would not want to retain a group of components that, combined, account for only a minority of the variance in the data set (say, 30%). Nonetheless, many critical values discussed earlier are obviously arbitrary. Because of these and related problems, this approach has sometimes been criticized for its subjectivity (Kim and Mueller, 1978).

9.2 The Interpretability Criteria

Perhaps the most important criterion for solving the ***number of-components*** problem is the interpretability criterion: interpreting the substantive meaning of the retained components and verifying that this interpretation makes sense in terms of what is known about the constructs under investigation.

The following list provides four rules to follow in doing this.
(A later section shows how to actually interpret the results of a principal component analysis; the following rules will be more meaningful after you have

completed that section).

1. Are there at least three variables (items) with significant loadings on each retained component? A solution is less satisfactory if a given component is measured by less than three variables.

2. Do the variables that load on a given component share the same conceptual meaning? For example, if three questions on a survey all load on component 1, do all three of these questions seem to be measuring the same construct?

3. Do the variables that load on different components seem to be measuring different constructs? For example, if three questions load on component 1, and three other questions load on component 2, do the first three questions seem to be measuring a construct that is conceptually different from the construct measured by the last three questions?

4. Does the rotated factor pattern demonstrate simple structure? Simple structure means that the pattern possesses two characteristic:

    (a) Most of the variables have relatively high factor loadings on only one component, and near zero loadings on the other components, and

    (b) most components have relatively high factor loadings for some variables, and near-zero loadings for the remaining variables.

10 Review of Important Definitions

- An observed variable can be measured directly, is sometimes called a measured variable or an indicator or a manifest variable.
- A principal component is a linear combination of weighted observed variables. Principal components are uncorrelated and orthogonal.
- A latent construct can be measured indirectly by determining its influence to responses on measured variables. A latent construct could is also referred to as a factor, underlying construct, or unobserved variable.
- Factor scores are estimates of underlying latent constructs.
- Unique factors refer to unreliability due to measurement error and variation in the data.
- Principal component analysis minimizes the sum of the squared perpendicular distances to the axis of the principal component while least squares regression minimizes the sum of the squared distances perpendicular to the x axis (not perpendicular to the fitted line).
- Principal component scores are actual scores.
- Eigenvectors are the weights in a linear transformation when computing principal component scores. Eigenvalues indicate the amount of variance explained by each principal component or each factor.
- Orthogonal means at a 90 degree angle, perpendicular. Obilque means other than a 90 degree angle.
- An observed variable ***loads*** on a factors if it is highly correlated with the factor, has an eigenvector of greater magnitude on that factor.
- Communality is the variance in observed variables accounted for by a common factors. Communality is more relevant to EFA than PCA.

- `princomp`
- `prcomp`
-

```
Proportion of Variance 0.6200604 0.2474413 0.0891408 0.04335752
Cumulative Proportion  0.6200604 0.8675017 0.9566425 1.00000000
> loadings(pc.cr)

# note that blank entries are small but not zero

Loadings:
Comp.1 Comp.2 Comp.3 Comp.4
Murder   -0.536  0.418 -0.341  0.649
Assault  -0.583  0.188 -0.268 -0.743
UrbanPop -0.278 -0.873 -0.378  0.134
Rape     -0.543 -0.167  0.818

Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
```

```
## The signs of the columns are arbitrary
> plot(pc.cr) # shows a screeplot.
> biplot(pc.cr)
>
> ## Formula interface
> princomp(~ ., data = USArrests, cor = TRUE)
Call:
princomp(formula = ~., data = USArrests, cor = TRUE)

Standard deviations:
Comp.1    Comp.2    Comp.3    Comp.4
1.5748783 0.9948694 0.5971291 0.4164494

4  variables and  50 observations.
```

```
> ## NA-handling
> USArrests[1, 2] <- NA
> pc.cr <- princomp(~ Murder + Assault + UrbanPop,
+                    data = USArrests, na.action = na.exclude,
cor = TRUE)
>
```

`prcomp`

Performs a principal components analysis on the given data matrix and returns the results as an object of class `prcomp`.

```
> prcomp(~ Murder + Assault + Rape,
data = USArrests, scale = TRUE)
Standard deviations:
[1] 1.5380939 0.6691898 0.4318010

Rotation:
PC1        PC2        PC3
Murder  -0.5822314  0.5349197 -0.6122643
Assault -0.6063988  0.2159081  0.7652870
Rape    -0.5415599 -0.8168504 -0.1986662
```