1 ASCA Version 1.3

The program `asca(X, F, interactions, center)` does a principle component analysis on the effect matrices of a data matrix $\mathbf{X}$. The program calculates scores and loadings for each experimental factor and each interaction that is indicated in the function parameters (`interactions`). Only interactions between two factors can be calculated. *It is assumed the data are balanced, i.e. the number of subjects per level is the same.* The program does *not* check for unbalanced data. Output (scores, loadings singular values, projections and percentages explained variance) are given in the data structure ASCA. Fields of ASCA include:

`ASCA.data`	`Centered/scaled data matrix`
`ASCA.design`	`Experimental design vector`
`ASCA.factors.scores`	`ASCA experimental factor scores`
`ASCA.factors.loadings`	`ASCA experimental factor loadings`
`ASCA.factors.projected`	`Projected residuals`
`ASCA.factors.singular`	`Singular values`
`ASCA.factors.explained`	`Percentage variation explained`
`ASCA.interactions.scores`	`ASCA scores interactions`
`ASCA.interactions.loadings`	`ASCA loadings interactions`
`ASCA.interactions.singular`	`Singular values`
`ASCA.interactions.explained`	`Percentage variation explained`
`ASCA.effects`	`Percentage explained for each effect`

The structure elements are cells (except for `ASCA.data`, `ASCA.design` and `ASCA.effects` which are matrices) that can be accessed by e.g. `ASCA.factors.scores{1}` for the scores of the first factor. The program also provides a plotting routine to plot scores and loadings plots for each experimental factor. It also includes the projection of the residuals (see Zwanenburg et al. J Chemometrics, Volume 25, (2011), pages 561 - 567).

When we have an experimental design with two experimental factors, α and β the data matrix $\mathbf{X}$ is decomposed in

$$\mathbf{X} = \mathbf{X}_{\text{avg}} + \mathbf{X}_{\alpha} + \mathbf{X}_{\beta} + \mathbf{X}_{\alpha\beta} + \mathbf{E} \tag{1}$$

The data matrix $\mathbf{X}$ has N rows corresponding to N observations and J columns corresponding to J variables. The terms in the decomposition have the following meaning:

- $\mathbf{X}_{\text{avg}}$: matrix with column averages in each row
- $\mathbf{X}_{\alpha}$: matrix with level averages for first factor
- $\mathbf{X}_{\beta}$: matrix with level averages for second factor
- $\mathbf{X}_{\alpha\beta}$: matrix with level averages interaction between factor 1 and factor 2.
- $\mathbf{E}$: matrix with residuals

The input to the program are the data matrix $\mathbf{X}$, the matrix $\mathbf{F}$ that describes the experimental design, the interactions that are to be included and a parameter to indicate if centering and/or scaling is wanted. The scaling option includes centering of the data. In some fields this is known as standardization.

For each factor and interaction the percentages explained by the principal components are stored in `ASCA.factors.explained` and `ASCA.interactions.explained`.

2 Input example

As an example consider an experimental design with two factors, one with two and one with three levels. Each level has two subjects. In the experiment two variables are measured. The data matrix $\mathbf{X}$ has 12 rows and two columns:

$$\mathbf{X} = \begin{bmatrix} 1 & 0.6 \\ 3 & 0.4 \\ 2 & 0.7 \\ 1 & 0.8 \\ 2 & 0.01 \\ 2 & 0.8 \\ 4 & 1 \\ 6 & 2 \\ 5 & 0.9 \\ 5 & 1 \\ 6 & 2 \\ 5 & 0.7 \end{bmatrix} \qquad \mathbf{F} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \\ 2 & 1 \\ 2 & 1 \\ 2 & 2 \\ 2 & 2 \\ 2 & 3 \\ 2 & 3 \end{bmatrix}$$

the matrix $\mathbf{F}$ indicates the factor and level each row of $\mathbf{X}$ belongs to. The numbers in the first column of $\mathbf{F}$ are the levels for the first factor, the numbers in the second column of $\mathbf{F}$ indicate the levels of the second factor. In general, F has one column for each experimental factor.

For example, the first row of $\mathbf{F}$ is 11 indicating level 1 for the first factor and level 1 for the second factor. The first row in the data matrix are thus from a subject that was in the first level for each factor. The second row in $\mathbf{F}$ is also 11, the second row in the data matrix therefore contains the measurements of a subject that also was in the first level treatment of both factors. The last row of $\mathbf{F}$ is 23 indicating that the last row of the data matrix contains measurements for a subject that was treated according to level 2 for the first factor and level 3 for the second factor. After centering, the data matrix $\mathbf{X}$, the effect

matrices $\mathbf{X}_\alpha$ and $\mathbf{X}_\beta$ and the interaction matrix are found to be:

$$\mathbf{X} = \begin{bmatrix} -2.5000 & -0.3092 \\ -0.5000 & -0.5092 \\ -1.5000 & -0.2092 \\ -2.5000 & -0.1092 \\ -1.5000 & -0.8992 \\ -1.5000 & -0.1092 \\ 0.5000 & 0.0908 \\ 2.5000 & 1.0908 \\ 1.5000 & -0.0092 \\ 1.5000 & 0.0908 \\ 2.5000 & 1.0908 \\ 1.5000 & -0.2092 \end{bmatrix} \qquad \mathbf{X}_\alpha = \begin{bmatrix} -1.6667 & -0.3575 \\ -1.6667 & -0.3575 \\ -1.6667 & -0.3575 \\ -1.6667 & -0.3575 \\ -1.6667 & -0.3575 \\ -1.6667 & -0.3575 \\ 1.6667 & 0.3575 \\ 1.6667 & 0.3575 \\ 1.6667 & 0.3575 \\ 1.6667 & 0.3575 \\ 1.6667 & 0.3575 \\ 1.6667 & 0.3575 \end{bmatrix}$$

$$\mathbf{X}_\beta = \begin{bmatrix} 0 & 0.0908 \\ 0 & 0.0908 \\ -0.2500 & -0.0592 \\ -0.2500 & -0.0592 \\ 0.2500 & -0.0317 \\ 0.2500 & -0.0317 \\ 0 & 0.0908 \\ 0 & 0.0908 \\ -0.2500 & -0.0592 \\ -0.2500 & -0.0592 \\ 0.2500 & -0.0317 \\ 0.2500 & -0.0317 \end{bmatrix} \qquad \mathbf{X}_{\alpha\beta} = \begin{bmatrix} 0.1667 & -0.1425 \\ 0.1667 & -0.1425 \\ -0.0833 & 0.2575 \\ -0.0833 & 0.2575 \\ -0.0833 & -0.1150 \\ -0.0833 & -0.1150 \\ -0.1667 & 0.1425 \\ -0.1667 & 0.1425 \\ 0.0833 & -0.2575 \\ 0.0833 & -0.2575 \\ 0.0833 & 0.1150 \\ 0.0833 & 0.1150 \end{bmatrix}$$

3 Calculating the variation explained

The variation that is explained by the principal components can be calculated from the singular values. The matrix with singular values is diagonal with the square roots of the singular values on the diagonal. Let the elements on the diagonal of the singular matrix be s_i. The percentage explained for the first principal component, f_1 is then:

$$f_1 = \frac{s_1^2}{\sum_i s_i^2} \times 100\%$$

In the program `asca` the singular values are stored in the structure `ASCA` as:
`ASCA.factors.singular{1}` for the first factor and
`ASCA.factors.singular{2}` for the second factor and `ASCA.interactions.singular{1}` for the interactions. The first factor only has two levels, hence there is only one principal component that explains 100% of the variation. The second factor has three levels, and

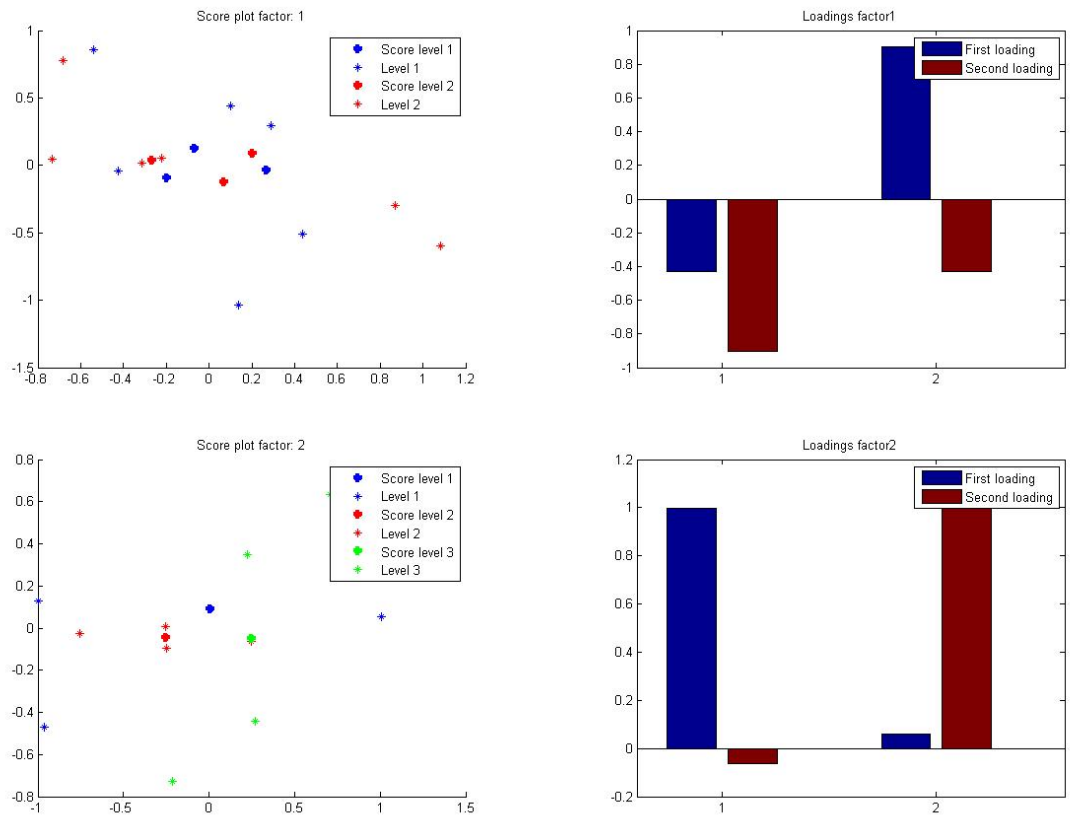

Figure 1: Scores and loadings for the two experimental factors for the data **X**. The scores plots also include the projected residuals.

therefore two principal components. To calculate the percentage of explained variation we first get the singular values and apply the above relation:

```
 s = ASCA.factors.singular{2}

s =

    0.7083
    0.2221

>> f_1 = (s(1)^2/(s(1)^2 + s(2)^2))*100

f_1 =

   91.0459

>> f_2 = (s(2)^2/(s(1)^2 + s(2)^2))*100
```

```
f_2 =

    8.9541
```

There are, as expected two singular values, the first explains 91% of the variation in the data, the second one 9%. Data sets with more variables will have more principal components, usually the first few will explain most of the variation. The program supplies the cells `ASCA.factors.explained` and `ASCA.interactions.explained` to return the percentage explained by the PC's of each factor and principal component:

```
>> s = ASCA.factors.explained{1}

s =

  100.0000
    0.0000

>> s = ASCA.factors.explained{2}

s =

   91.0459
    8.9541
```

4 Contribution of the different effects

The percentage each effect (overall mean, factors $\mathbf{X}_\alpha$, interactions $\mathbf{X}_{\alpha\beta}$ and residuals $\mathbf{E}$) contributes to the sum of squares of the data matrix $\mathbf{X}$ is given in `ASCA.effects` and also part of the standard output of the program. This is possible in ASCA because the subspaces spanned by the loadings vectors are all perpendicular to each other. Thus, the sum of squares of the elements of the data matrix can be decomposed as

$$\|\mathbf{X}\|^2 = \|\mathbf{X}_{\text{avg}}\|^2 + \|\mathbf{X}_\alpha\|^2 + \|\mathbf{X}_\beta\|^2 + \|\mathbf{X}_{\alpha\beta}\|^2 + \|\mathbf{E}\|^2 \tag{2}$$

For the example, the results are:

```
Percentage each effect contributes to the total sum of squares
Overall means
   76.3905

Factors
   15.0212    0.3154
```

```
Interactions
    1.3346

Residuals
    6.9383
```

or as given in `ASCA.effects`:

```
>> ASCA.effects

ans =

   76.3905   15.0212    0.3154    1.3346    6.9383
```

5 Score plot of interactions

Score and loadings plots of interactions can be plotted now (as of Version 1.1) with the function `plot_interactions` which is default included. The function plots the group averages for each interaction and labels each group average with the factors that comprise the group. Projections of the data are included, but these are unmarked to prevent the plot from cluttering up.

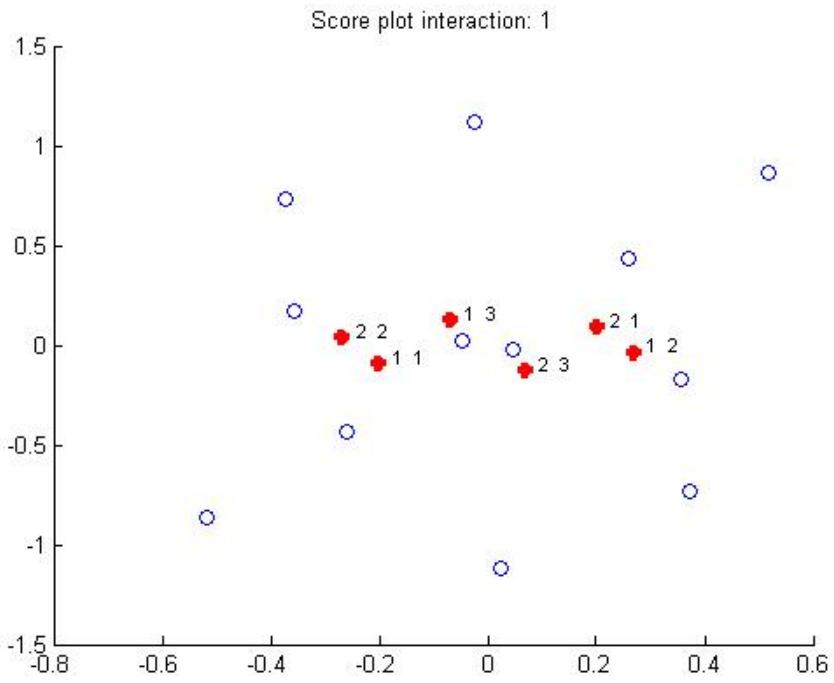

Figure 2: Scores for the interactions. The group averages (*) are labeled by their contributing factors, projected data points (o) are unlabeld.