ANOVA/ASCA Exercises

Age K. Smilde[1], Federico Marini[2], Gooitzen Zwanenburg[1]

1 mei 2018

[1]Biosystems Data Analysis, Swammerdam Institute for Life Sciences,
University of Amsterdam, The Netherlands.
[2]University of Rome La Sapienza, Rome, Italy.

1 Introduction

The Caldana data set measures transcriptome and metabolome levels of Arabidopsis under different experimental conditions. The data analysed here are part of the complete data set and comprise metabolites measured at seven different times (0, 5, 10, 20, 40, 80, 160 minutes) under four different light conditions (D: Dark, L: light, LL: Low light, HL: High Light). All data in the present data set are measured at a temperature of 21°C. There are 5 repeats per treatment combination and the data have been normalized by dividing each original value by the median of all values of that metabolite. The data for first time point (t=0) is identical for all light conditions.

2 ANOVA/ASCA-GUI

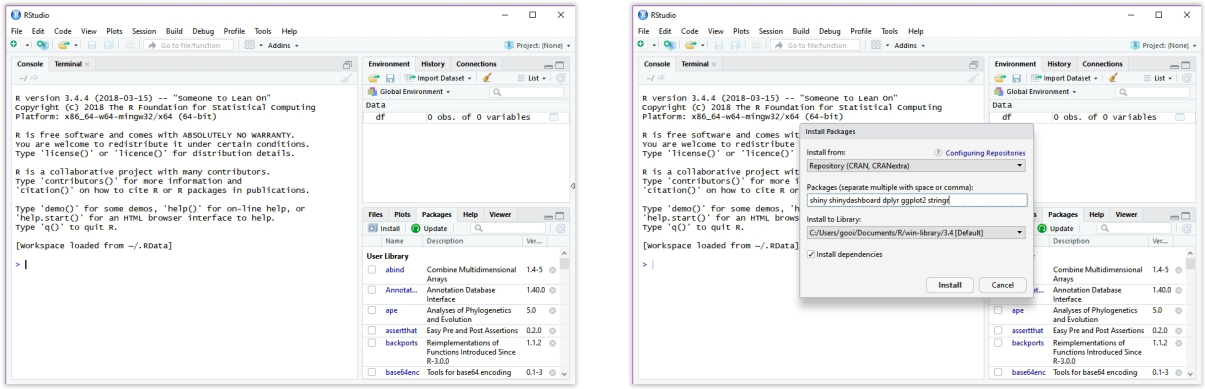
2.1 Some preliminaries

2.1.1 Obtaining and installing the software

The data analysis tool you will use, is written in the computer language `R`. For the tool to work you need to have `R` and the `R`-development environment `RStudio` installed on your computer. You can download `R` from `https://cran.r-project.org/index.html` and `RStudio` from `https://www.rstudio.com/products/rstudio/download/#download`. Installing the software on the different platforms is fairly straight forward: accept all the defaults and click on `OK` or `Finish` when appropriate.

Once `R` and `RStudio` are installed, you're almost good to go. `R` comprises a (large) number of base-packages that supply most of the functionality of `R`. However, additional functionality is supplied by extra packages that need to be installed separately. `RStudio` is a very handy tool to do this. First, open `RStudio`. You'll see something like Figure 1. The `R`-console is in the left hand panel. When you open a file to edit, it will open in an extra panel on the left, above the `R`-console. The bottom right panel is of interest to us now. It has a `Files`-tab with a file-explorer and a `Packages` tab. The latter we'll use to install missing packages. When you select the `Packages` tab, and click the `Install` tab, a window for easy package installation comes up (see right hand panel of Figure 1). You type the names of the packages you want to install in the window and make sure the `Install dependencies` box is checked. The data analysis tool uses the following packages that you may need to install:

shiny
shinydashboard
plyr
stringr
ggplot2



(a) (b)

Figuur 1: (a) The `RStudio` environment. `R`-files open in the panel on the left. When no files are open, as in this screen shot, you see the `R`-console with the `R`-prompt, `>`. The top right pane displays variables and other bits that are interesting for programmers. The pane on the bottom right has, among other things, a file explorer and package manager. (b) The `RStudio` package manager can be used to install missing packages. Fill in the names of the packages that you want to install, separated by a space or a comma, and click `Install`.

Before you can use a package, you need to load it into the workspace with the `library` function. The data analysis tool takes care of this, but when you run the tool for the first time, you may see some red printed text on the console. These are usually just warnings or info on the packages that are loaded and nothing to worry about.

2.1.2 Cleaning of the Caldana data

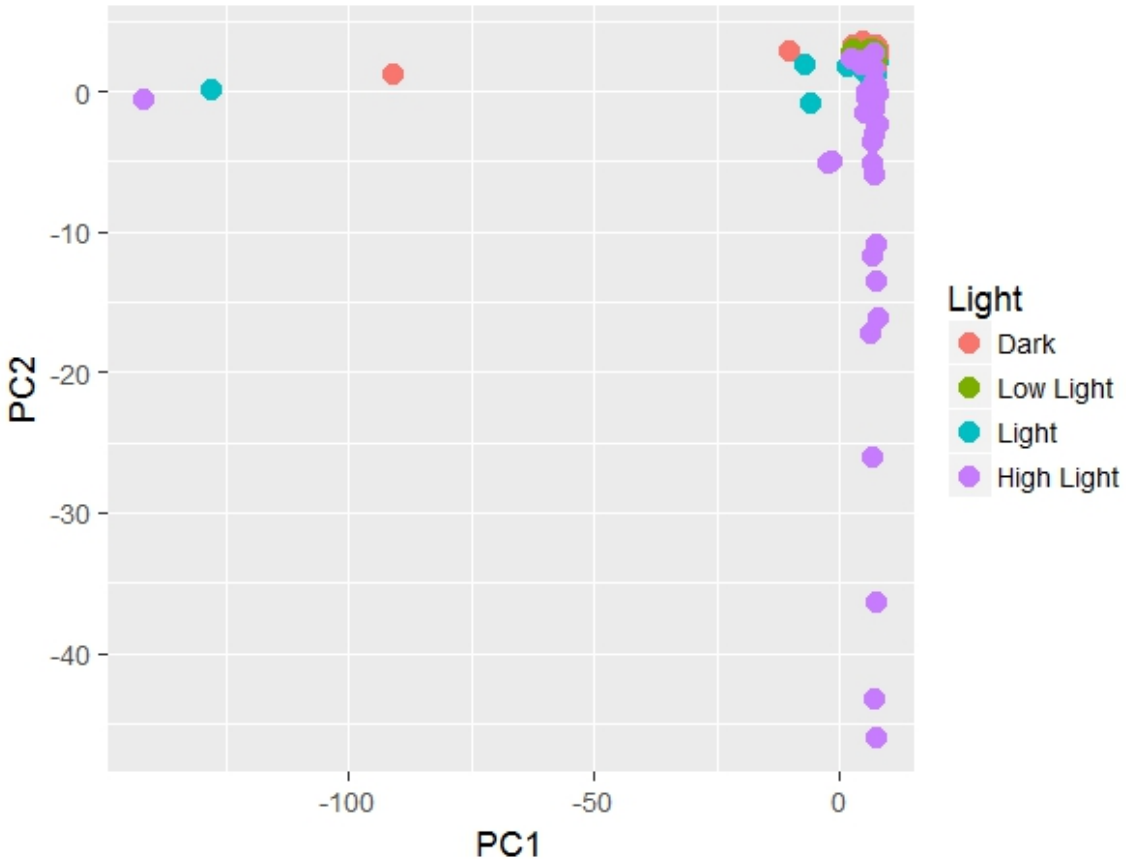
The Caldana data set measures transcriptome and metabolome levels of the plant Arabidopsis under different experimental conditions. The data we use here is part of a larger data set and comprises 69 variables (substances). There are two experimental factors. The first is time: the data are measured at 7 different times (0, 5, 10, 20, 40, 80, 160 minutes); the second is the light level. The measurements were done under four different light conditions (D: Dark, LL: Low light, L: light, HL: High Light). All data in the present data set is measured at a temperature of 21°C.

The original data had a fairly large number (117) of missing values and some of the measurements were clearly outliers. Just throwing away observations with a missing value would reduce the number of observations severely. Therefore, the missing values are replaced by the cell (a cell is combination of experimental factors, for example time = 20 minutes, light condition is `Dark`) average plus a random number taken from a normal distribution with the standard deviation of the measurements in the cell (the cells each have 5 measurements).

After the missing values were filled in, the data were visualized with PCA. The scores of the first two principal components are shown in Figure 2. From this figure we see that there are possibly three outliers with a very large (in absolute value) first principal component and four outliers (possibly even nine) with a very large second principal component. To find the associated variables, we plot the loadings for the first and second principal components, as shown

in Figure 3. A plot of the loadings of the first principal component shows that variable 56 (raffinose) is very dominant. Inspection of the data revealed that this is caused by a value of 149.20 for $t = 0$, while the other measurements in this cell are all of the order of 1. Hence, we conclude that 149 is an outlier. On closer inspection, more seemed wrong with the measurements of raffinose: also values of 136, 98.9 and some smaller values of around 15 were found where other measurements are at least an order of magnitude smaller. Because multiple entries in the raffinose data appeared to be outliers, we concluded something went wrong with the raffinose measurements and removed the variable `raffinose` from the data.

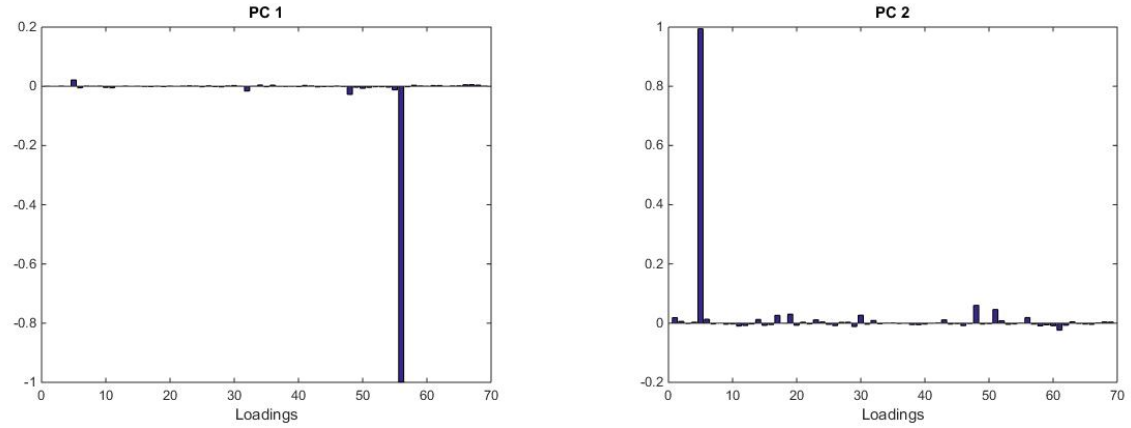
The loadings of the second principal component suggest something is going on with variable 5, `glycine`. When we inspected the glycine data we noticed that for the low light (LL) condition, the values for glycine increase with time. From the data alone it is hard to say wether this is real or an artifact. Because the values for glycine dominated the other data in a PCA plot, wether right or wrong, `glycine` was removed from the data set. It could be analysed further, separately.



Figuur 2: PCA score plot of the first two principal components of the Caldana data set. The plot shows a number of possible outliers.

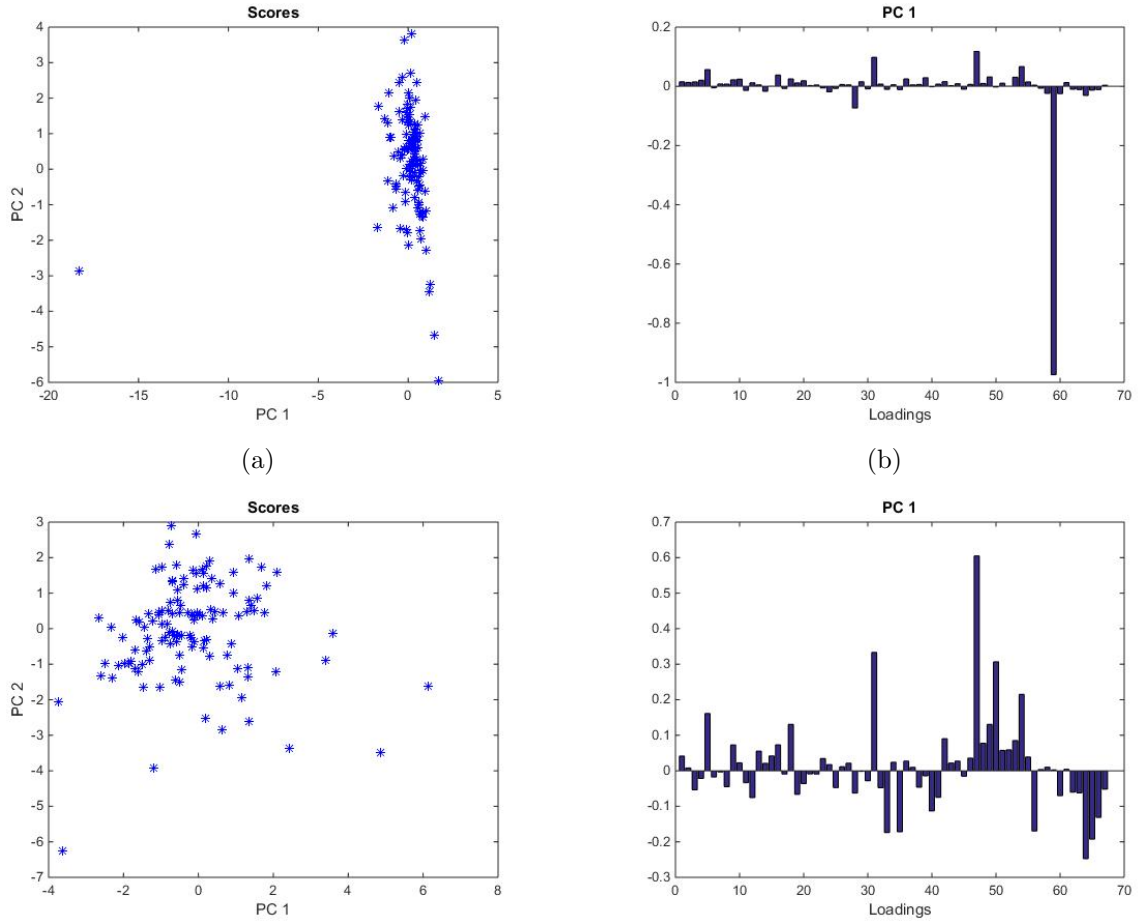
A new PCA with the glycine and raffinose data removed shows (see Figure 4) that there seems to be one more outlier for variable 59 (glycerol). We replaced the value 20.01 for the light (L) condition at time $t = 80$ with the cell average. The PCA on the data, after replacing this outlier, is shown in the bottom graphs of Figure 4. This figure shows no obvious outliers anymore and this is the data set used in the ASCA data analysis tool. In the appendix we included a list with the measured variables.

For use with the ASCA data analysis tool the centered data are stored in two comma separated



(a) (b)

Figuur 3: Loading plots for the Caldana data. Left hand panel: the loadings for the first principal component. The right hand panel shows the loadings for the second principal component. From the plots we see that variable 56 (raffinose) has a dominant contribution to the first principal component whereas variable 5 (Glycine) gives a very large contribution to the second principal component.



(a) (b)

(c) (d)

Figuur 4: Score (a) and loadings (b) plot after PCA on the data set after glycine and raffinose have been removed. There seems to be one outlier left on variable 59 (glycerol). After cleaning, the Caldana data set shows no more obvious outliers in the scores (c) and loadings (d) plots.

values (csv) files: `Caldana_data.csv` and `Caldana_F.csv`. The tool itself is in the file `asca.R`. The file `Caldana_data.csv` contains the measured data, the file `Caldana_F.csv` represents the

design of the experiment. Each row in this file corresponds to a measurement and has two values, the first value (first factor) represents the light level of the measurement, the second value (second factor) the time value of the measurement. The tool and the two data files are packed together in a `.zip` file, `Caldana.zip`. To run the tool, unpack the zip-file at a place you can find back. Make sure the two data files, `Caldana_F.csv` and `Caldana_data.csv`, and the program `asca.R` are in the same folder. Open `asca.R` in `RStudio` either by dubble clicking `asca.R` or by opening the file using the `RStudio` file menu. Once the program is open in `RStudio`, you start the program with the `Run App` button on the top right corner of the top left pane. The `Run App` button has three options, hidden under the small downward pointing black triangle: `Run in Window`, `Run in Viewer Pane` and `Run External`. The first option starts a separate `RStudio` window with the data analysis tool, the third starts the tool in a webbrowser. The `Run in Viewer Pane` option is not recommended because it runs the program in the, way too small, bottom right panel of `RStudio`.

If for some reason the data analysis tool crashes, you can restart the tool with the `Run App` button. In some cases, however, nothing seems to happen. This is because in the background the tool is still running. You can stop the process by clicking the little red `Stop` sign at the top right of the console pane. After the process is stopped, you can restart the tool with the `Run App` button.

2.2 The data analysis tool

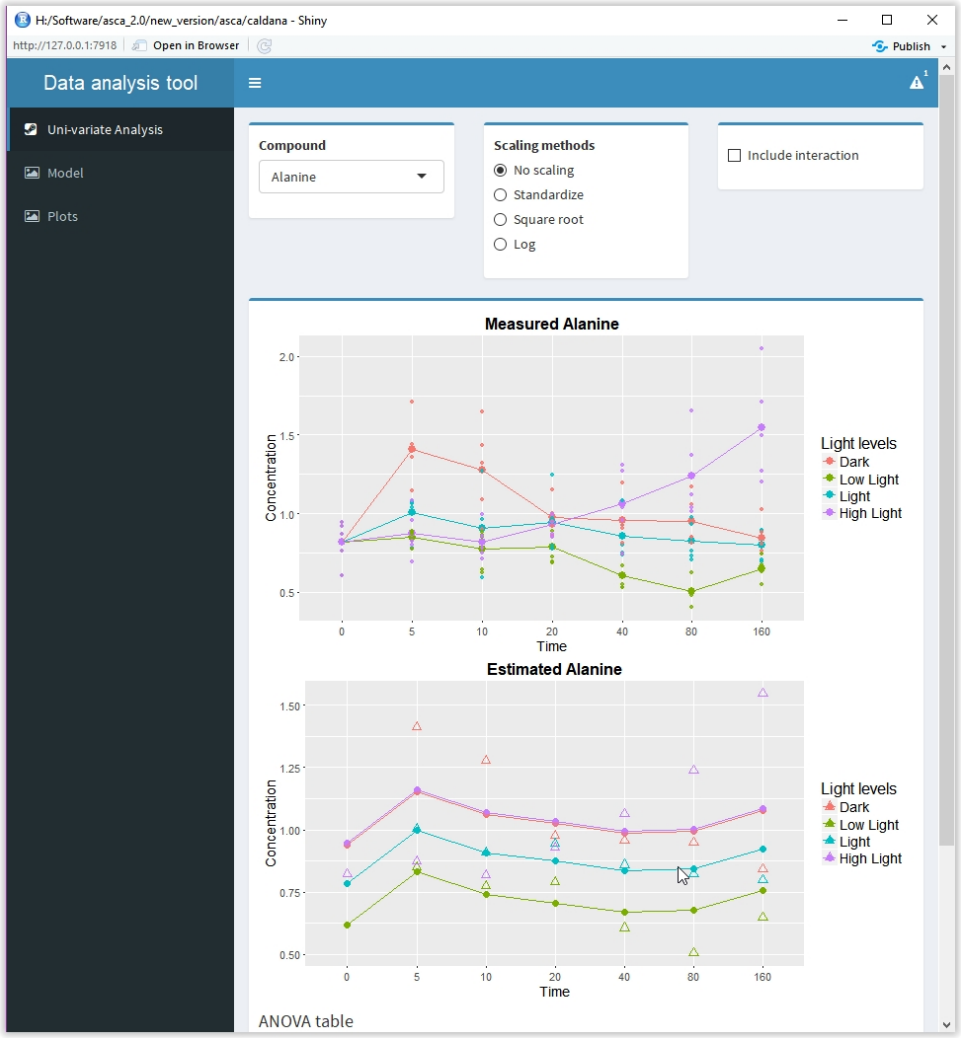
2.2.1 Univariate model

The ASCA data analysis tool is written in `R` using the `shinydashboard` package. In short, the tool comprises a univariate and a multivariate part. When you start the tool it opens with the uni-variate window, as is shown in Figure 5. The univariate part of the tool displays the measured data for a selected variable (small dots) together with the average of the measured data (larger dots) in the top graph. The second graph gives the model estimate of the measured averages for a simple linear model with `Time` and `Light` as independent variables. The model can be extended to include the interaction between `Time` and `Light` by checking the `Include interaction` checkbox in the top right. The estimates are represented in the second graph by filled circles, the averages of the measurements by open triangles. Below the graphs is the ANOVA-table for the model. Different scalings can be applied to the data with the `Scaling methods` buttons. Scaling of the last data that is selected here carries over to the multivariate analysis.

2.2.2 Multivariate model

On the left hand side-bar of the tool, you can select the `Model` option to display the model selection boxes. With the `Include factors`, `Include interaction` and `Combine terms` selection boxes the desired model can be selected. The first box,`Include factors`, allows to include two, one or no factors in the model. With the `Include interaction` box the interaction between `Light` and `Time` can be included into the model. The model that is selected is displayed in the `Model` box below the selection boxes, while the explained variances for each of the models are given in the bottom box.

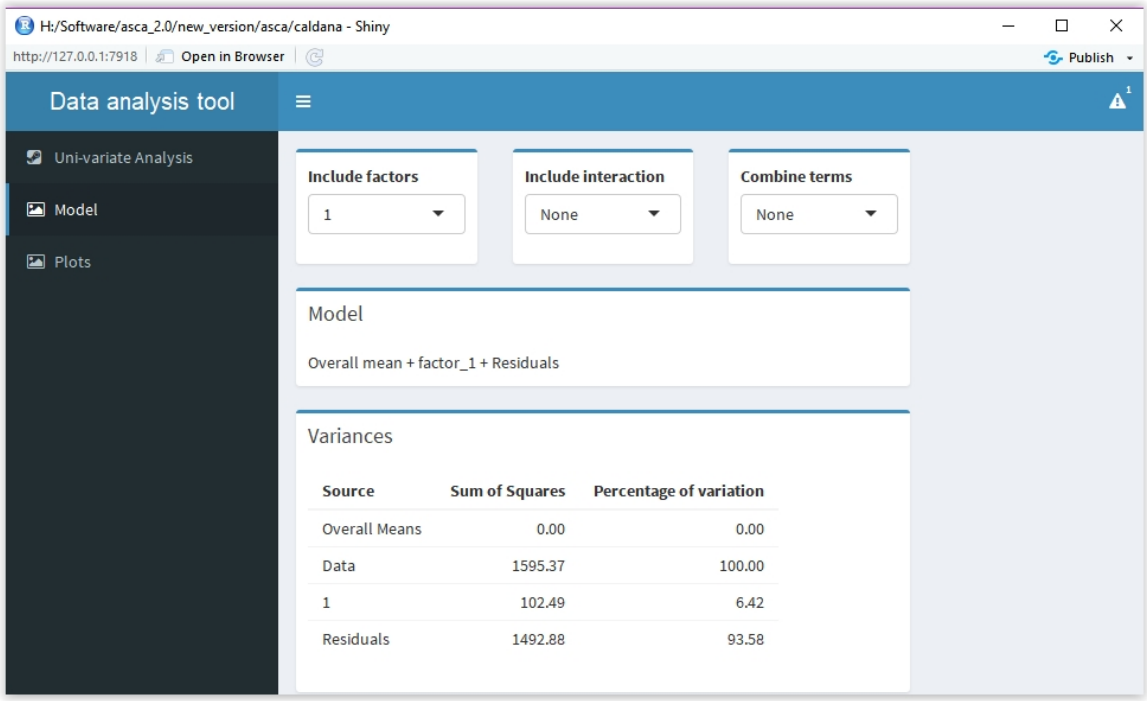
As an example, say we want a model in which we only include the effect of light levels, i.e. the first factor in the experiment. Select from the `Include factors` box `1` for factor 1, and `None` from the other two selection boxes. In the `Model` display the model is shown: $\mathbf{X} = \mathbf{M} + \mathbf{L} + \mathbf{R}$ where $\mathbf{X}$ is the data matrix, $\mathbf{L}$ the matrix with average light levels and $\mathbf{R}$ the matrix with



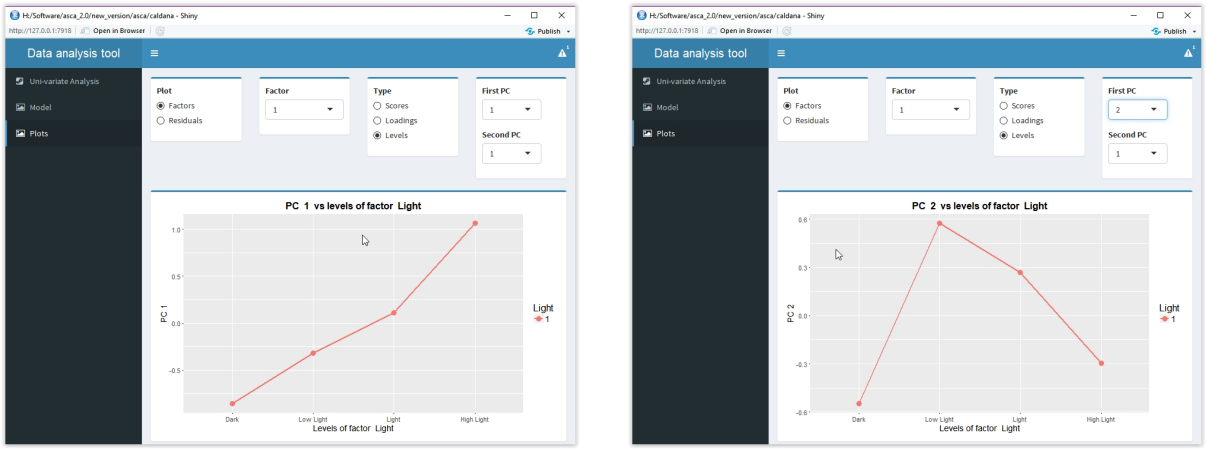
Figuur 5: The opening window of the ASCA data analysis tool. The tool opens in the univariate window. In this window the measured data of a single variable (small dots) and the average of the measurements (larger dots) are shown in the top graph. The variable that is displayed can be chosen from the drop-down box. The second graph gives the estimated values for a simple linear model without or with interactions (checkbox top right). The triangles represent the measured values, the dots the model estimates. An ANOVA-table (not seen in this figure) for the model is included below the graphs. Different scalings can be applied to the data with the `Scaling methods` buttons. It is important to know that the scaling that is chosen last is also used in the multivariate data analysis.

residuals. Note that the matrix with overall means, $\mathbf{M}$, is $\mathbf{0}$ in all models because the data is centered. From the table of variances we see that the factor `Light` contributes for about 6.5% to the variance of the data (see Figure 6).

To visually inspect the model, the `Plots` option in the left hand side-bar is selected. With the fist selection box, we select what we want to plot, in the example, `Factors` or `Residuals`. With the second selection box we choose which factor we want to plot. In the example there is no choice, since the model only contains the first factor. With the `Type` selection box we select `Scores`, `Loadings` or `Levels`. The first two are straight forward, the last requires some explanation. The `Levels` option plots the selected principal component (most right selection box) against the levels of the selected factor. In Figure 7 the first principal component is shown as a function of the light level in the left panel; the right panel shows the second principal component as a function of the factor levels. Note that changing the `Second PC` has no effect on the 'level' plots.



Figuur 6: Example in which we look at a model with just the factor Time. The model display gives the terms in the model, the variances display shows the explained variation for each of the terms in the model.



(a) (b)

Figuur 7: First (left) and second (right) principal component as a function of light level.

The `Combine` selection box also needs some clarification. With the choices `1 + 1:2` and `2 + 1:2` the interaction matrix and the matrix of the first, respectively, second factor are combined (added). When a `Combine` option other than `None` is selected, the interaction and the selected factor can no longer be selected independently. For example, if `1 + 1:2` is selected, the model is $\mathbf{X} = \mathbf{M} + \mathbf{T} + (\mathbf{L} + \mathbf{I}) + \mathbf{R}$, with $\mathbf{I}$ the interaction between light levels and time. The matrices for the interaction and the first factor are added in the model to form a single effect matrix and, therefore, the interaction and the factor `Light` cannot be selected to be part of the model anymore: they have been 'used', so to speak. This is reflected in the `Model` and `Variances` displays. These restrictions may not be immediately visible in the `Include factors` and `Include interaction` selection boxes, but they do restrict the factors and interactions interactions that can be plotted under the `Plots` option in the side bar.

3 ANOVA Exercises

All exercises have the same structure which is detailed below and should be followed for the metabolites proline, valine, serine, alanine, leucine and fructose subsequently.

3.1 Plot the data

We start with plotting the data. If there is a time factor involved in data, it is always informative to use the time as the x-axis. The different light treatments are indicated with colors, and also the cell-means and individual measurements are shown. The questions to be answered for proline are the following:

1. Do the ANOVA assumptions hold for this metabolite?
2. Do you see main effects? If so, describe those.
3. Do you see an interaction? If so, describe this.

Next, we may want to try whether a variance stabilizing transformation helps. The option `Standardize` means that all values are standardized to standard deviation one. Try out all three types of transformations and look at the plots again.

4. Do the transformations make a difference?.

Now answer questions 1-3 for the other metabolites (valine, serine, alanine, leucine and fructose) working with untransformed data only.

3.2 Two-way ANOVA without interaction

We are going to make our first ANOVA model for proline in its most simple form using untransformed data: only with the two factors as main effects. The plot you get back from the GUI is the estimate of the model for the different factor levels. Answer the following questions:

1. Interpret the ANOVA table. Are the main effects significant?
2. Is the model adequately describing the data? Is there lack-of-fit?
3. Use the same three transformations as before. Does the ANOVA table change? Discuss the results.

Now answer questions 1,2 for the other metabolites (valine, serine, alanine, leucine and fructose) working with untransformed data only.

3.3 Two-way ANOVA with interaction

Our final model is an ANOVA model with interaction (untransformed data). The plot you get back from the GUI is the estimate of the model for the different factor levels and interaction. Answer the following questions:

1. Interpret the ANOVA table. Are the main effects and interaction significant?
2. What is the difference with the ANOVA model without interaction?
3. Does the model adequately describe the data? Which model do you recommend?
4. Does your final model differ from the one you guessed in Section 3.1? If so, why?

4 ASCA Exercises

4.1 Full model

For the ASCA analysis we are going to use untransformed data. We start with a full model (factor 1= Light; factor 2 = Time; and Interaction). Answer the following questions:

1. Interpret the table of variances. What is the largest contribution?
2. Plot the scores 1,2 of the factor Light and interpret those. Use also the Level plots for this. Plot also the loadings and identify the metabolites that contribute most to the first ASCA-Light component.
3. Plot the scores 1,2 of the factor Time and interpret those. Use also Level plots for this. Plot also the loadings and identify the metabolites that contribute most to the first ASCA-Time component.
4. Plot the scores of the interaction using level plots (Time on x-axis) and interpret those. Plot also the loadings and identify the metabolites that contribute most to the first ASCA-Interaction component. Does the second ASCA-Interaction component have a clear profile?
5. Make score plot 1,2 of the residuals. Investigate suspicious points and relate those to possible deviating metabolites.

4.2 Combined model

It is also possible to combine factors and interactions. The most sensible thing to do is to combine Light and Interaction (the factor Time has a clear interpretation on its own, but for Light this is more problematic). We repeat the analogous set of questions:

1. Interpret the table of variances. What is the largest contribution?
2. Plot the scores 1,2 of the factor Time. Are they different from the previous ones? Explain your results.
3. Plot the scores 1,2 of the Combined terms using level plots (Time on x-axis) and interpret those. Plot also the loadings and identify the metabolites that contribute most to the first ASCA-Interaction component. Does the second ASCA-Interaction component have a clear profile? Which metabolites contribute most?
4. Look again at the residuals. Are they different from the previous ones? Explain your results.

5 ASCA or ANOVA?

The big question is now, of course, what to use: ASCA or multiple ANOVA's? This is an easy question, but the answer is not easy. First, multiple ANOVA's and the associated testing suffers from the problem of multiple testing bias which need to be corrected for by using a more stringent cut-off value for deciding on significance. Secondly, ASCA uses the correlations between metabolites, whereas ANOVA's do not do that. You can test yourself whether the results of ASCA and ANOVA coincide by running again ANOVA's for the metabolites you found interesting in the ASCA results. Interesting metabolites are Fructose, Maleic-acid, O-acetyl-serine,....

6 Appendix: list of variables

variable name	variable number	variable name	variable number
Alanine	1	4-hydroxy-benzoic-acid	35
Valine	2	Dehydroascorbic-acid-dimer	36
Leucine	3	Gluconic-acid	37
Isoleucine	4	Dehydroascorbic-acid	38
Proline	5	Ascorbic-acid	39
Serine	6	4-Hydroxycinnamic-acid	40
Threonine	7	Similar-to-Adenine	41
beta-alanine	8	Shikimate	42
Hydroxyproline	9	Erythritol	43
GABA	10	Arabinose	44
Aspartate	11	Arabitol	45
Asparagine	12	Fucose	46
Methionine	13	Fructose	47
O-acetyl-serine	14	Mannitol	48
Glutamate	15	Galactose	49
Phenylalanine	16	Glucose	50
Ornithine	17	Sucrose	51
Glutamine	18	Maltose	52
Lysine	19	Trehalose	53
Tyrosine	20	Galactinol	54
Threonic-acid	21	myo-inositol	55
Citrulline-Arginine	22	Uracil	56
Pyruvic-acid	23	Putrescine	57
Citric-acid	24	Ethanolamine	58
Succinic-acid	25	Glycerol	59
Fumaric-acid	26	Indole-3-acetonitrile	60
Malic-acid	27	Sinapic-acid	61
Lactic-acid	28	Palmitic-acid	62
Glycolic-acid	29	Octadecanoic-acid	63
Benzoic-acid	30	Docosanoic-acid	64
Maleic-acid	31	Tetracosanoic-acid	65
Nicotinic-acid	32	Hexacosanoic-acid	66
Itaconic-acid	33	Octacosanoic-acid	67
Citramalate	34		

Tabel 1: Variables of the Caldana data. After removal of `raffinose` and `glycine`, there are 67 variables.