MVDA-Exercise POPLAR

OPLS in multi-class metabolomics

Background

This exercise is the study of genetically modified poplar plants by GC/MS metabolomics. Two modifications are investigated, i.e., up-regulation and down-regulation within the *PttPME1* gene. This gene is involved in the production of pectin methyl esterase, PME, which is an enzyme that de-esterifies methylated groups within pectin. Pectin is a big complex polymer and will not be analyzed by this type of technique. Nevertheless, the metabolic profile was of interest as both lines indicated several symptoms of oxidative stress response.

Objective

The objective of this exercise is to shed some light on how OPLS may be used in state-of-the-art multi-class Metabolomics. In particular, the objectives are to:

- demonstrate how to extract putative biomarkers from the S-plot
- demonstrate how to provide statistical evidence to extracted biomarkers
- demonstrate how to extract information that is unrelated (orthogonal) to the modelled response, **y**
- demonstrate how to compare multiple classes by the use of a SUS-plot

Data

In total, the data set contains N = 26 observations (plants) and K = 78 variables (resolved and integrated GC/MS profiles by the use of H-MCR [1]). The observations (plants) are divided in three groups ("classes"):

- Control Wild type plant, 10 plants **"WT"**
- *PttPME1* down-regulated poplar, L5, 7 plants **"L5"**
- *PttPME1* up-regulated poplar, 9 plants **"2B"**

The GC/MS data are three-way by nature with time, absorbance and mass dimensions. In this example the three way data have been pre-processed by Hierarchical Multivariate Curve Resolution, H-MCR [1]. H-MCR resolves the chromatographic profiles, calculates the area of the resolved profiles and generates the corresponding mass spectrum. The resolved mass spectrum can be subjected to a library search and the compound can thus be identified. A more elaborate account of the experimental conditions is found in reference [2].

*1) Jonsson et. al Journal of Proteome Research **2006,** 5,1407-1414, Predictive metabolite profiling applying hierarchical multivariate curve resolution to GC-MS data-A potential tool for multi-parametric diagnosis*

*2) Wiklund et. al Analytical Chemistry **2008,** 80, 115-122, Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models*

Copyright Sartorius Stedim Data Analytics AB, 21-09-24

Tasks

Task 1 (Overview of data)

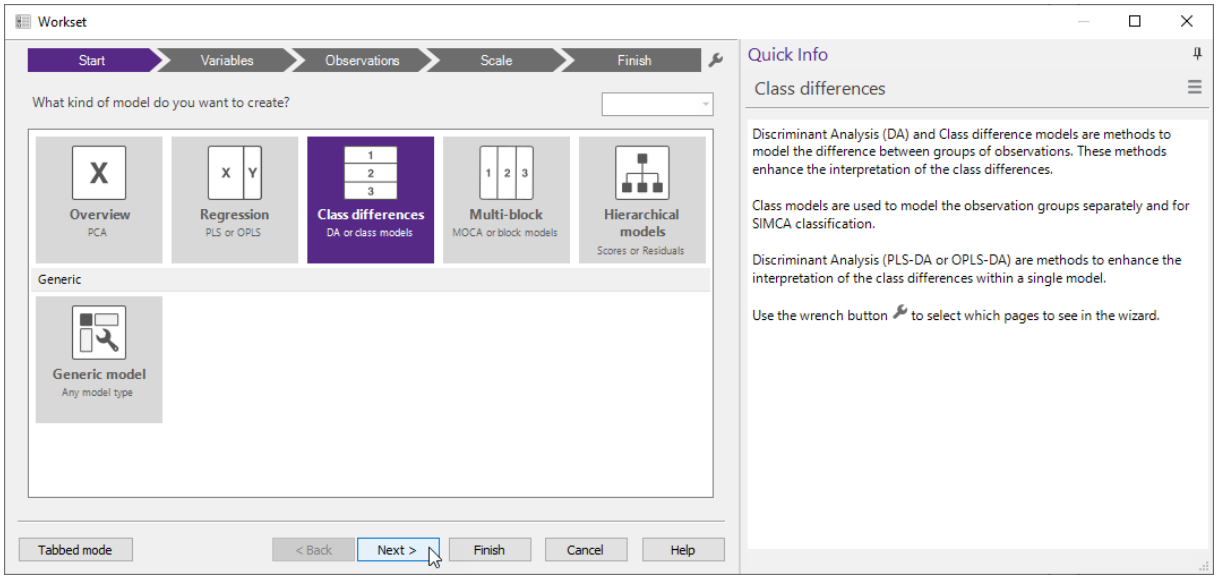
Create a new regular project in SIMCA using Poplar.xlsb. The dataset is pre-configured so that the import should run un-interrupted. Finish the import and give the project a suitable name.

Poplar

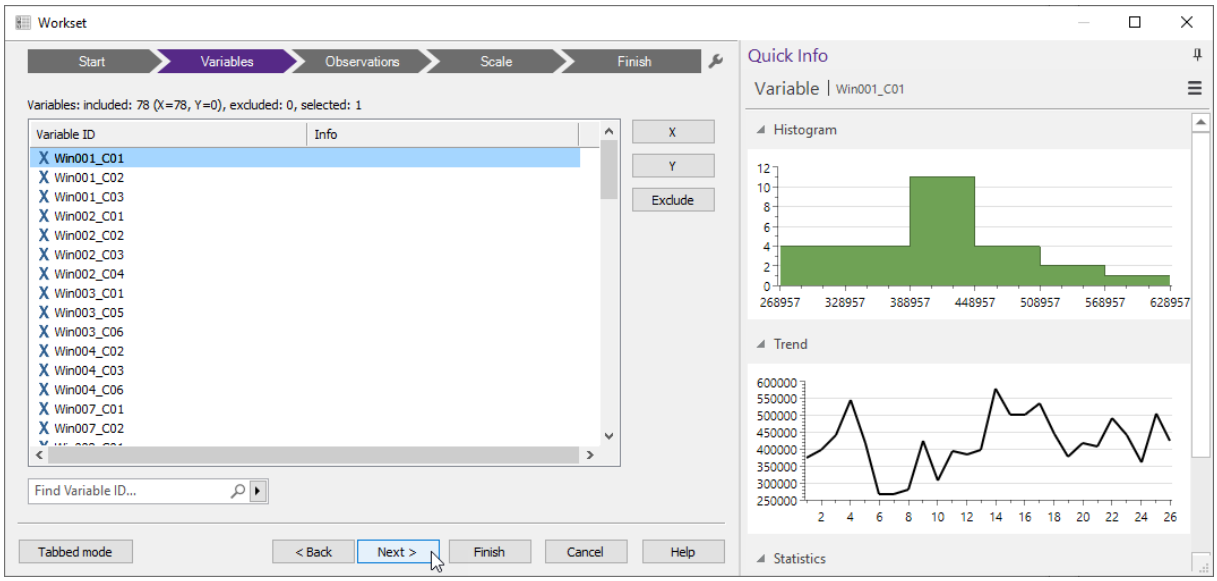
	Primary ID	Secondary ID	Class ID	4	5	6	7	8
Primary ID	Primary ID	Secondary ID	Class ID	Win001_C01	Win001_C02	Win001_C03	Win002_C01	Win002_C02
Secondary ID	Primary ID	Secondary ID	Class ID	Dodecamethylpentasiloxane	HEPTANOIC ACID-TMS	Dodecane	L-Valine (2TMS)	ETHANOLAMINE,N,N,O-TMS
3	1	WT_10	WT	375833	543556	38276,9	69192,9	184317
4	2	WT_1	WT	398184	537285	34609,7	217668	193095
5	3	WT_2	WT	443126	695953	38394,2	314980	148169
6	4	WT_3	WT	547311	627214	51519,6	205261	167343
7	5	WT_4	WT	417952	505763	30364	109736	166606

Before we make any OPLS modelling it is recommended that class-specific PCA models are calculated. This is done to ensure that no outlier exists in the data.

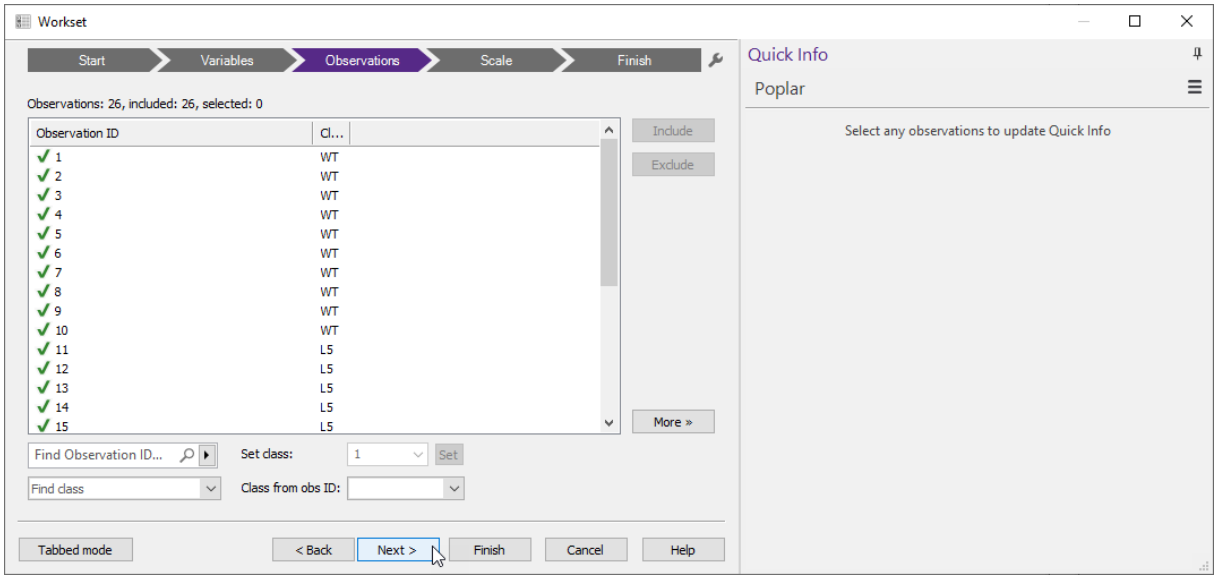
Click on Workset/New and select Class differences (DA or class models). Click on the Wrench button and add the scale page. Click Next.



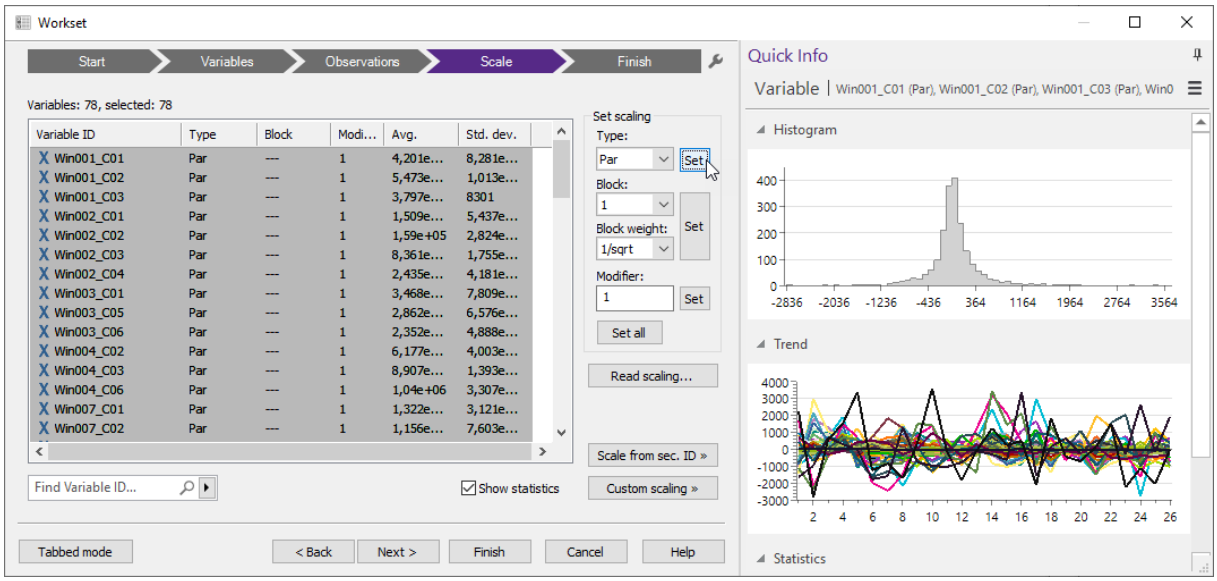
Include all 78 variables. Click Next.



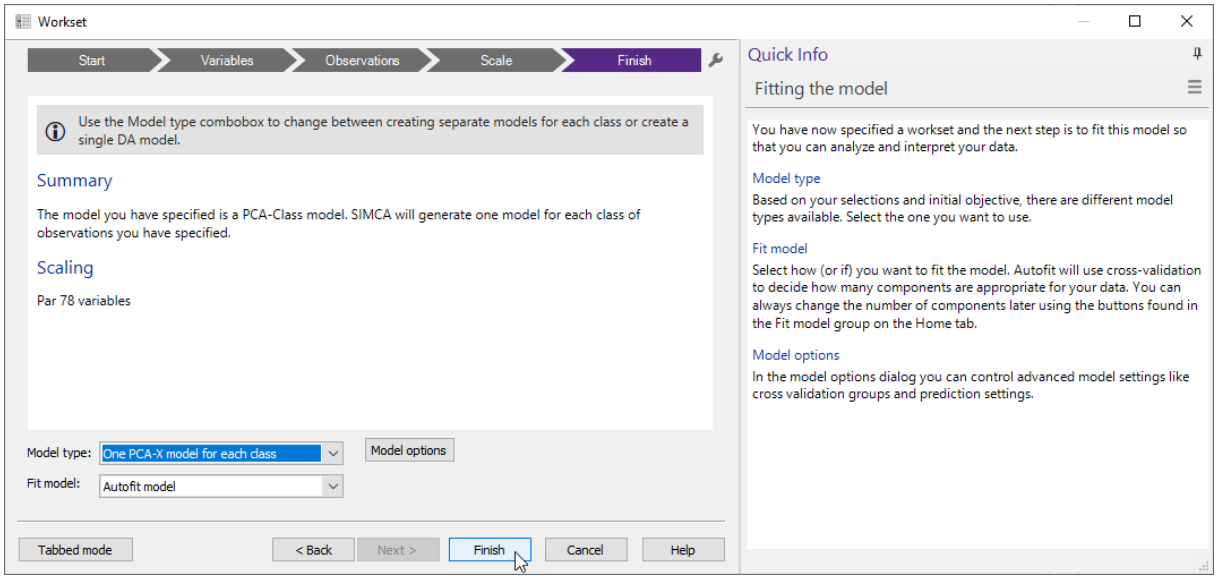
Include all 26 observations. Click Next.



Pareto scale the data. Click Next.



Review your selections and settings. Calculate individual PCA models for the three classes. Click Finish.

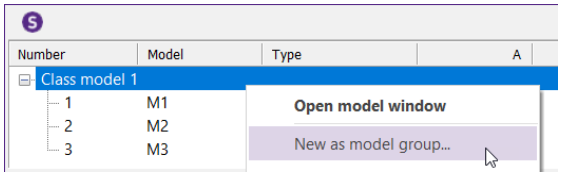


Interpret the PCA-class models. What do you see? Review the score- and DModX-plots and search for possible outliers.
Continue in the Solutions part, Task 1.

Task 2 (Contrasting two classes)

Contrasting WT and L5

Right click on Class model 1 in the project window and select “New as model group”.



Select Class differences and run through the workset wizard and exclude observations 2B. Set model type to OPLS-DA and Autofit the model.

Contrasting WT and 2B

Start by right clicking on the previous model and select "New as Model ??" (where ?? represents the previous model's number). Exclude all samples from L5 and include all samples from 2B. Compute the corresponding OPLS-DA model for WT vs 2B. Compare the results from the two models.

Specific questions

Can you separate WT from L5 and WT from 2B by OPLS-DA models?

How much of the variation in **X** is related to the separation between WT and L5, WT and 2B?

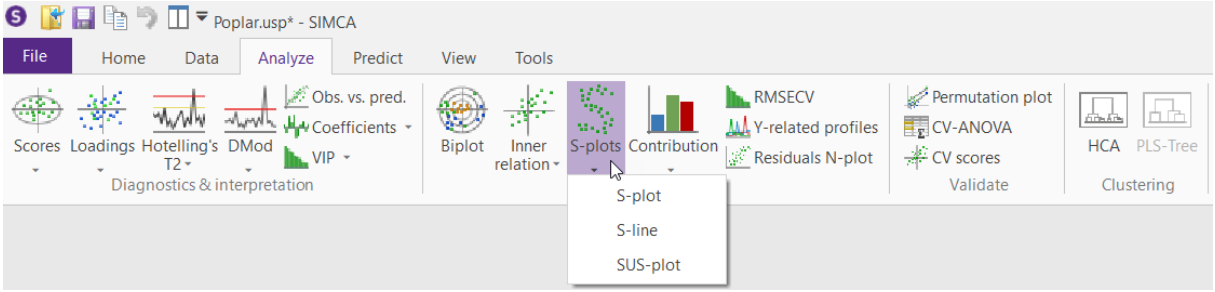
How much of the variation in **X** is systematic but uncorrelated to the separation between WT and L5, WT and 2B?

Can you see any patterns in the score plot except for the separation between the classes?

Continue in the Solutions part, Task 2.

Task 3 (Identifying putative biomarkers between Control and Treated)

Identify putative biomarkers by using the S-plot (p[1] vs. p(corr)[1]) and the loading column plot (p[1]) with confidence intervals. Use the two plots interactively. An S-plot can be plotted from the predictive components under Analyse.



Specific questions:

Why is the S-plot important? What does the confidence interval seen in the loading column plot (p[1]) mean? What are the default settings in SIMCA for the column loadings confidence intervals (what is α by default)? In what regions should you be careful when extracting putative biomarkers from the S-plot? Why is it recommended not to use only the column plot with confidence intervals for putative biomarker identification?

Identify the specific pattern seen in the orthogonal components by plotting an S-plot with the orthogonal components (no plots given in the solutions section).
Continue in the Solutions part, Task 3.

Task 4 (Investigate if the same biomarkers appear in two models)

In order to compare the outcome from two models the *shared and unique structure* plot, SUS-plot, is useful. This plot can be found at the same location as the S-plot.

Specific questions:

In what regions will you find shared information?

In what regions will you find up regulated and unique information for L5?

In what regions will you find up regulated and unique information for 2B?

Continue in the Solutions part, Task 4.

SOLUTIONS to Poplar

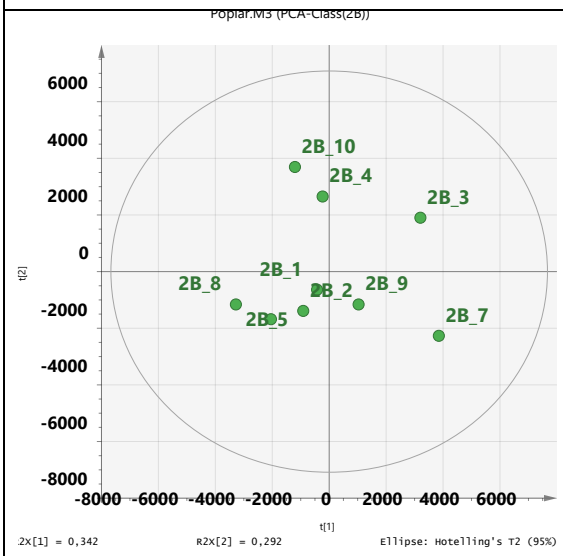
Task 1

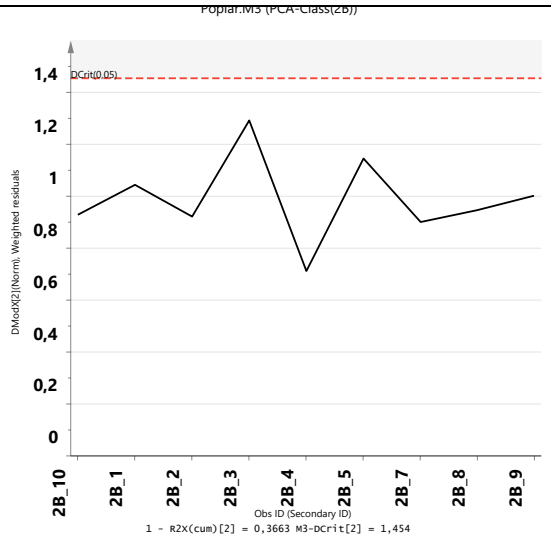
No obvious outliers were found in the three classes.

Poplar [M1]

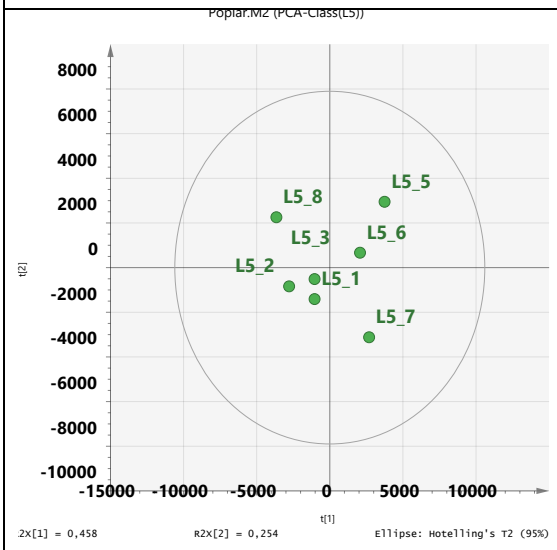
Number	Model	Type	A	N	R2X(cum)	R2Y(cum)	Q2(cum)	Title	Hierarchical
Class model 1									
1	M1	PCA-Class(WT)	2	10	0,701		0,419		
2	M2	PCA-Class(L5)	2	7	0,712		0,109		
3	M3	PCA-Class(2B)	2	9	0,634		-0,0348		

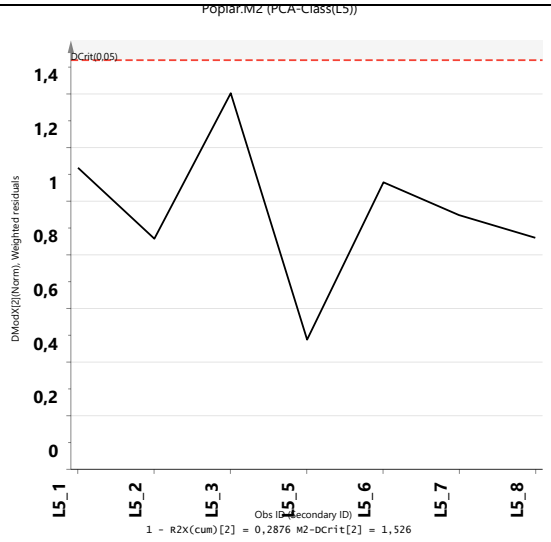
2B



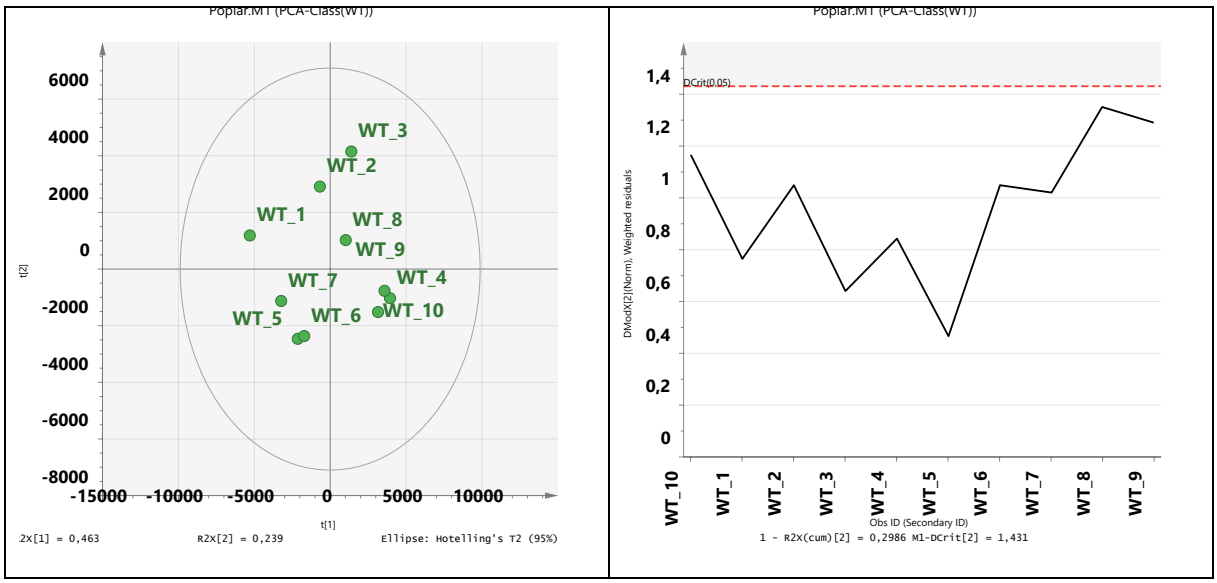


L5





WT



Return to the Tasks section, Task 2.

Task 2 (Contrasting two classes)

A basic requirement to answer all the questions in this task is that we get a reasonably good OPLS-DA model with a good Q2. To answer the questions look into the model overview window.

Model WT vs. L5. Note we are only using the 1+1 model to avoid over-interpreting the data.

Model M4

Model type: OPLS-DA, observations (N): 17, variables (K): 80 (X=78, Y=2)

Title: WT vs L5 Workset Options

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0.532			0.778			0.412		1		
Predictive		0.171			0.778			0.412		1		
P1	0.171	0.171	2.9	0.778	0.778	0.412	0.01	0.412	1	1	2	R1
Orthogon...		0.362			0							
O1	0.362	0.362	6.15	0	0							R1

Model WT vs. 2B.

Model M5

Model type: OPLS-DA, observations (N): 19, variables (K): 80 (X=78, Y=2)

Title: WT vs 2B Workset Options

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0.411			0.827			0.645		1		
Predictive		0.14			0.827			0.645		1		
P1	0.14	0.14	2.66	0.827	0.827	0.645	0.01	0.645	1	1	2	R1
Orthogon...		0.271			0							
O1	0.271	0.271	5.15	0	0							R1

Copyright Sartorius Stedim Data Analytics AB, 21-09-24

Answers:

Can you separate WT from L5 and WT from 2B by OPLS-DA models?

WT vs. L5: R2Y=0,778, Q2Y=0,412, good class separation and high predictive ability.

WT vs. 2B: R2Y=0,827, Q2Y=0,645, good class separation and high predictive ability.

Better separation between WT vs. 2B than WT vs. L5

How much of the variation in **X** is related to the separation between WT and L5, WT and 2B?

WT vs. L5: predictive component R2X=0,171→17.1%

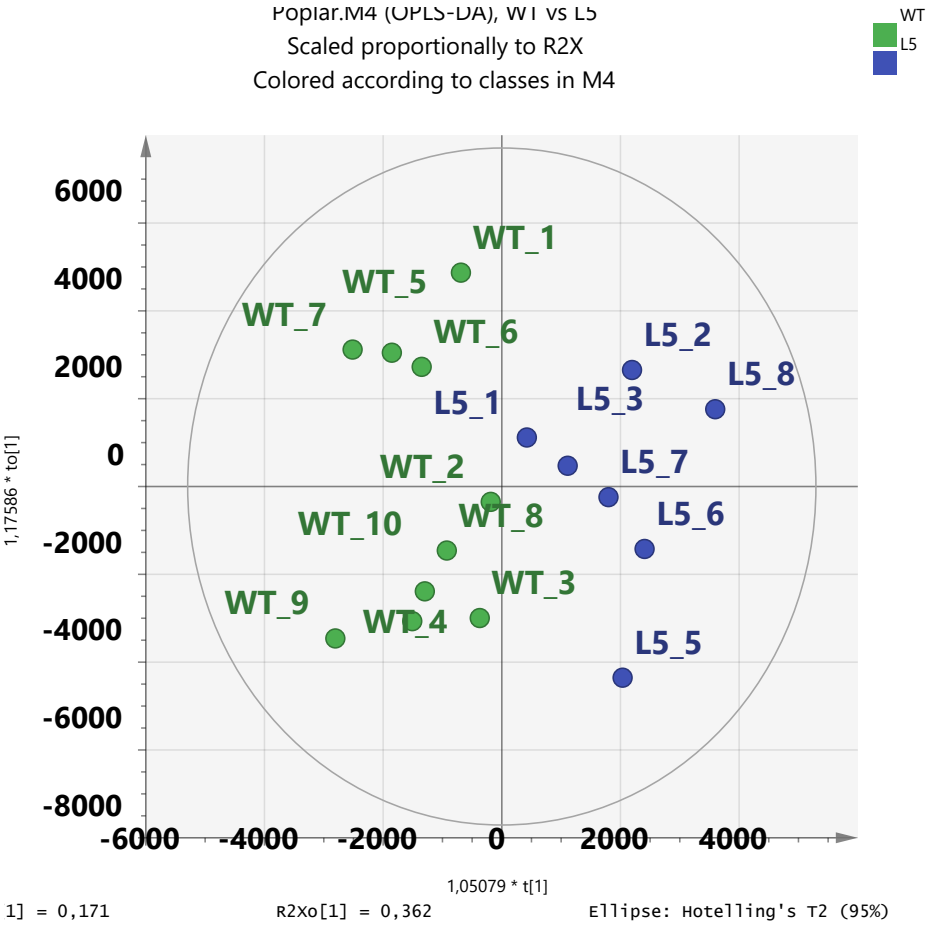
WT vs. 2B: predictive component R2X=0,14→14%

How much of the variation in **X** is systematic but uncorrelated to the separation between WT and L5, WT and 2B?

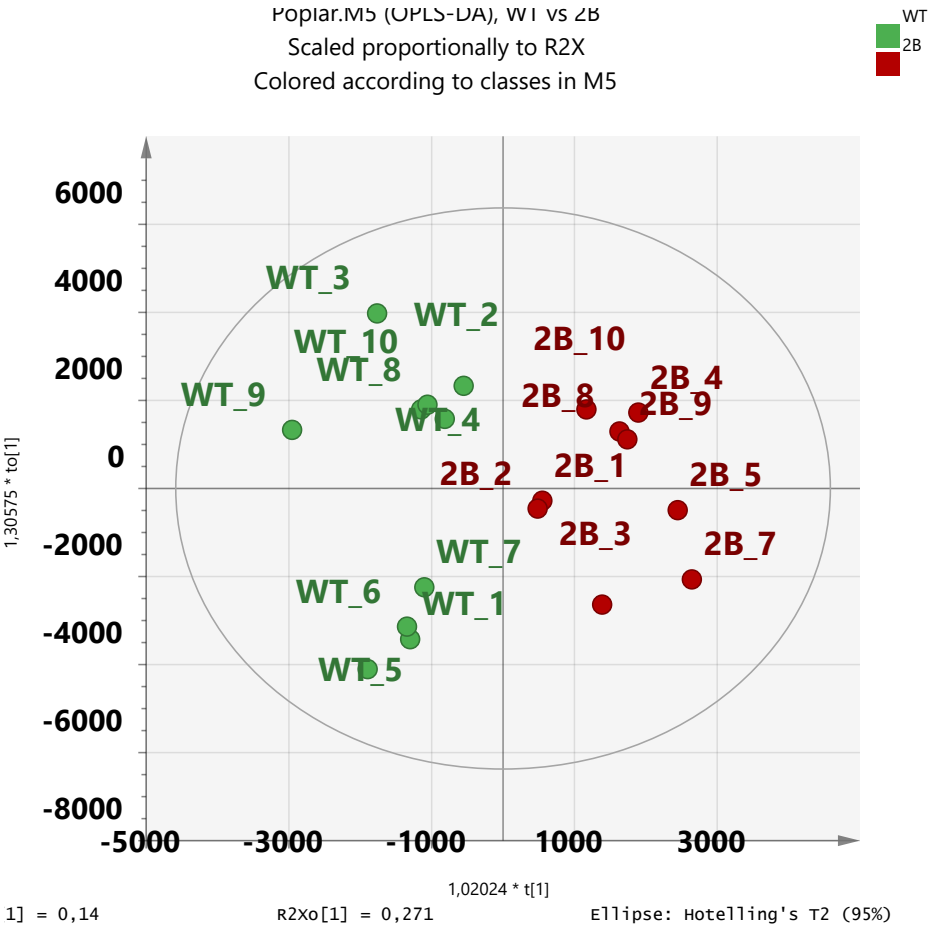
WT vs. L5: R2X=0,361→36,2%

WT vs. 2B: R2X=0,271→27,1%

Can you see any patterns in the score plot except for the separation between the classes?



A clear separation between the WT and L5 can be seen in the first OPLS component, t[1]. This visualized separation must be combined with a high Q2 for a good class separation. Only a clear separation in the score plot is NOT a valid class separation. It is also seen in the first orthogonal score vector, to[1], that the WT class separates into two subsets.

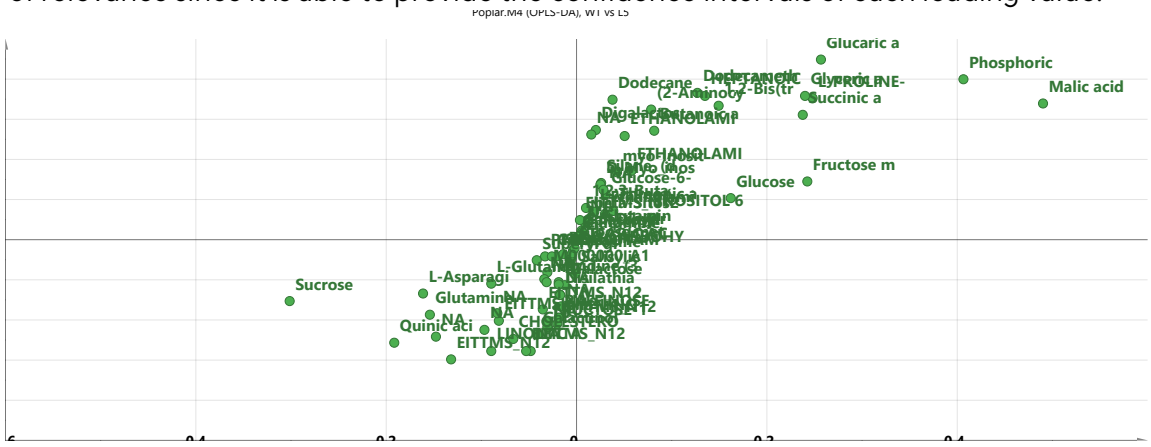


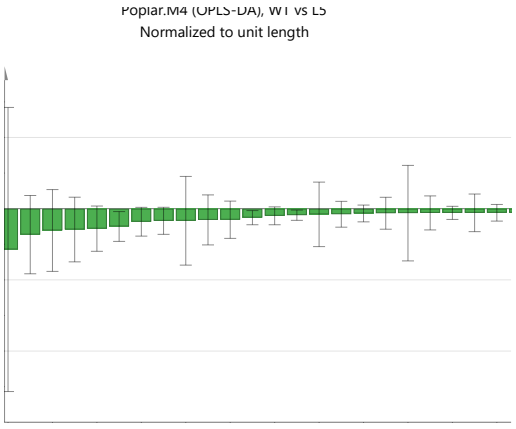
A separation between the WT and 2B can be seen in the first OPLS component, t[1]. The class separation is similar to the model WT vs. L5.

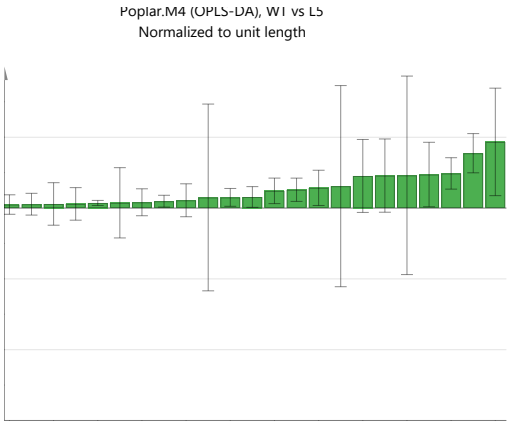
Return to the Tasks section, Task 3.

Task 3 (Identifying putative biomarkers between Control and Treated)

The predictive S-plot is a good way to identify putative biomarkers. The column plot of p[1] is also of relevance since it is able to provide the confidence intervals of each loading value.







Hint: sort the loading plot to ascending values.

It is clearly seen that e.g. sucrose is highly uncertain as a putative biomarker. In the S-plot sucrose has a high magnitude but a low reliability. This is confirmed from the loading plot p[1] where the confidence limit crosses zero. An interesting putative biomarker is linoleic acid which both has high magnitude and high reliability and the confidence interval is also small. Other putative biomarkers are glucaric acid, phosphoric acid and malic acid. All of these are of interest for further investigations. Remember that these putative biomarkers are only statistically significant.

Hint: Make a list of all interesting biomarkers by marking those of interest in the S-plot and then right click on the plot and *create/list*.

<u>Why is the S-plot important?</u>

The S-plot is an easy way to understand metabolomics data and is used to filtering out putative biomarkers.

In this plot both magnitude (intensity) and reliability is visualised. In spectroscopic data the peak magnitude is important as peaks with low magnitude are close to the noise level and are thus at higher risk for spurious correlation. The meaning of high reliability means high effect and lower uncertainty for putative biomarker.

What does the confidence interval seen in the loading column plot (p[1]) mean?

The confidence interval reflects the variable uncertainty and is directly correlated to the reliability. The confidence interval is very useful in reporting the significance of a putative biomarker.

What are the default settings in SIMCA for the column loadings confidence intervals (what is α by default)?

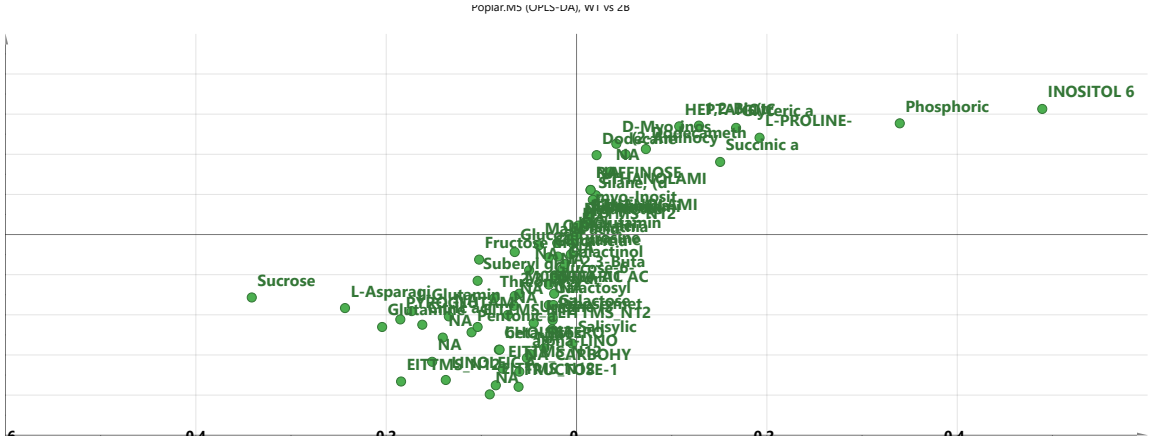
SIMCA will by default set the α to 0,05. The meaning of this is that a metabolite with a confidence interval which does not cross 0 is by 95% statistically safe. An alternative interpretation is: the probability of making the wrong decision is 5%. The default setting can be changed.

In what regions should you be careful when extracting putative biomarkers from the S-plot?

Metabolites with: low reliability plus low magnitude, or high reliability plus low magnitude are uncertain. The importance of magnitude depends on how the spectral data was pre-processed. Always remember to go back and check the raw data.

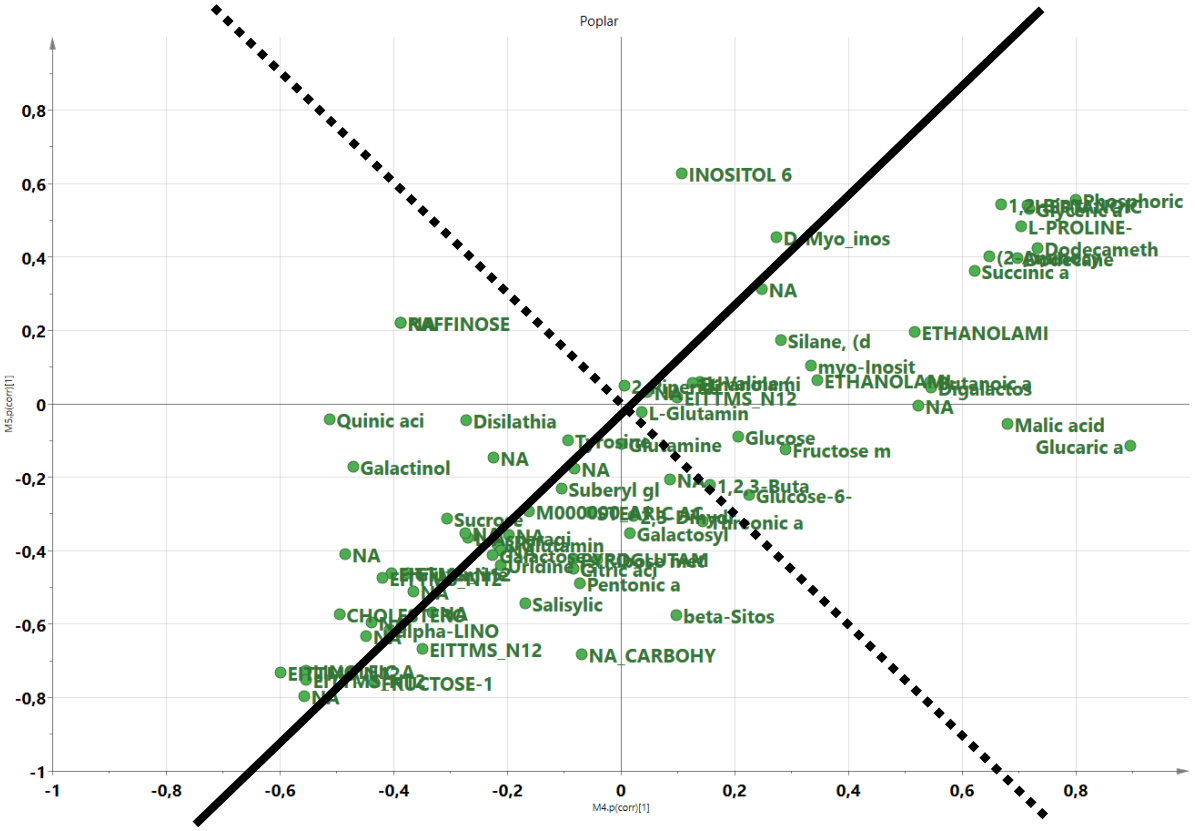
Why is it not recommended to only use the column plot with confidence intervals for putative biomarker identification?

The signal to noise overview is much easier to see in the S-plot. The combination of both plots also makes it easier to extract putative biomarkers.



Return to the Tasks section, Task 4.

Task 4 (Investigate if the same biomarkers appear in two models)



It is seen in the SUS-plot that many metabolites have the same effect on both L5 and 2B. All metabolites on the diagonal have same effect in both plants e.g. linoleic acid (down regulated) and phosphoric acid (up-regulated). The variables on the broken line have the same effect but in opposite directions in both transgenic plants. Some of the unique metabolites found in L5 was glucaric acid, malic acid, ethanolamine, butanoic acid. These were up regulated in L5. Quinic acid was also found as a unique down regulated metabolite in L5, but this metabolite was also highly uncertain. In 2B it was found that inositol was up regulated and several unassigned metabolites were found down regulated.

Copyright Sartorius Stedim Data Analytics AB, 21-09-24

Conclusions

- The S-plot can be done using PCA or PLS-DA ONLY if a clear class separation is seen in the first component in the score plot. If not the vectors p[1] and p(corr)[1] will be confounded by variation that is NOT related to class separation which will lead to an misleading interpretation.
- We obtain a list of potential biomarkers which are statistically significant and which separate one class from another.
- These biomarkers are **statistically significant**, but **not necessarily biochemically significant**.
- They may have biochemical significance and they may be the biomarkers we are interested in, however, this must be established through extensive testing.
- Metabonomics/Metabolomics allows the user to mine complex data and provides information which allows us to propose intelligent hypotheses.
- OPLS-DA is an excellent tool for this purpose.

Copyright Sartorius Stedim Data Analytics AB, 21-09-24