



# 4 Wm

Workflow4metabolomics

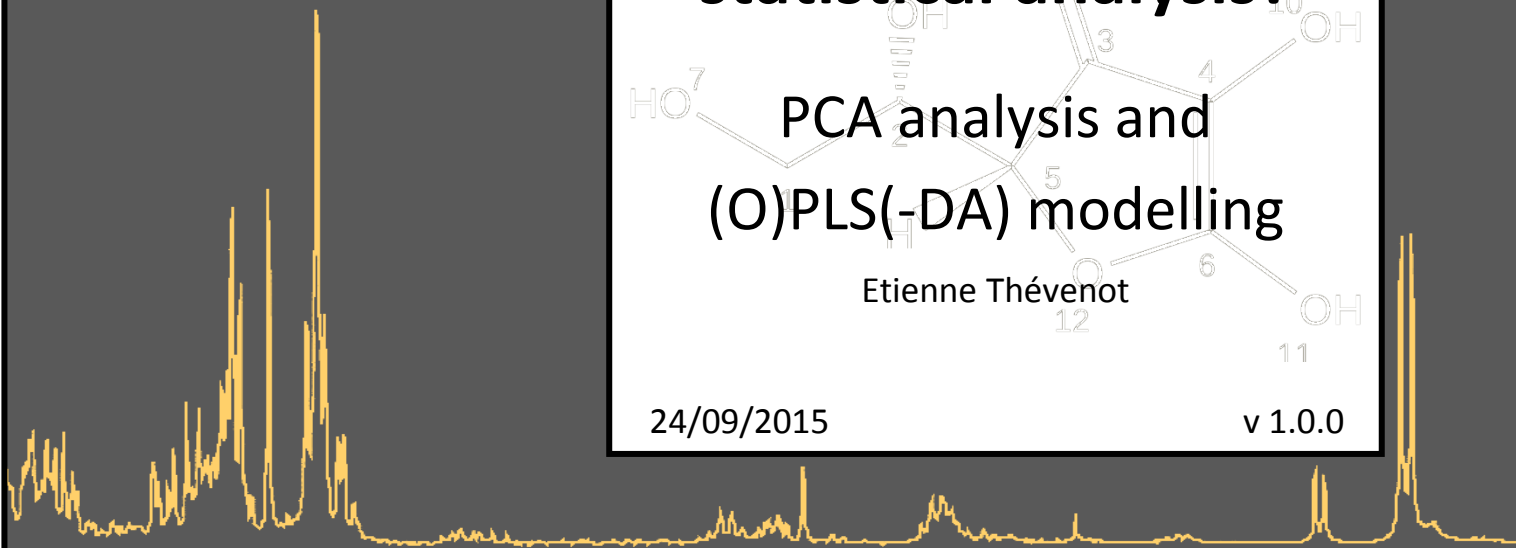
## How to perform statistical analysis?

PCA analysis and  
(O)PLS(-DA) modelling

Etienne Thévenot

24/09/2015

v 1.0.0



# The "Multivariate" module

Multivariate

- The "**Multivariate**" module allows you to perform:
  - Principal Component Analysis (**PCA**)
  - Partial Least-Squares regression (**PLS**) and discriminant analysis (**PLS-DA**)
  - Orthogonal Partial Least-Squares regression (**OPLS**) and discriminant analysis (**OPLS-DA**)
- It is available in the "Statistical Analysis" sections of LC-MS, GC-MS, and NMR

Galaxy / 4 / Meta

Tools

search tools

[Upload File from your computer](#)

[Export Data](#)

LC-MS

[Format Conversion](#)

[Preprocessing](#)

[Normalisation](#)

[Quality Control](#)

[Statistical Analysis](#)

Univariate Univariate statistics

**Multivariate PCA, PLS and OPLS**

[Anova](#) N-way anova. With ou Without interactions

[ACP](#) ellipsoid by factors

[Hierarchical Clustering](#) using ctc R package for java-treeview

- The Multivariate module uses internally the *ropls* R module from bioconductor

<http://bioconductor.org/packages/ropls>

- implements the original, NIPALS based, algorithms for PCA, PLS and OPLS
- diagnostics to detect outliers, overfitting
- graphics (scores, loadings, predictions)
- feature selection (VIP, regression coefficients)

Thévenot E.A., Roux A., Xu Y., Ezan E. and Junot C. (2015). Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *Journal of Proteome Research*, **14**:3322-3335.

<http://dx.doi.org/10.1021/acs.jproteome.5b00354>

# Objectives

---

- Multivariate analysis:
  1. PCA [unsupervised]: Visualize the structure of the **dataMatrix: X**
  2. (O)PLS(-DA) [supervised]: How can a factor of interest (response; column of **sampleMetadata**) be explained as a linear combination of **all** the variables (predictors) from **dataMatrix:  $y = f(X)$** 
    - a. when the response **y** is quantitative: (O)PLS regression
    - b. when **y** is qualitative: (O)PLS(-DA) classification

Complementary to univariate analysis (where variables are tested independently)

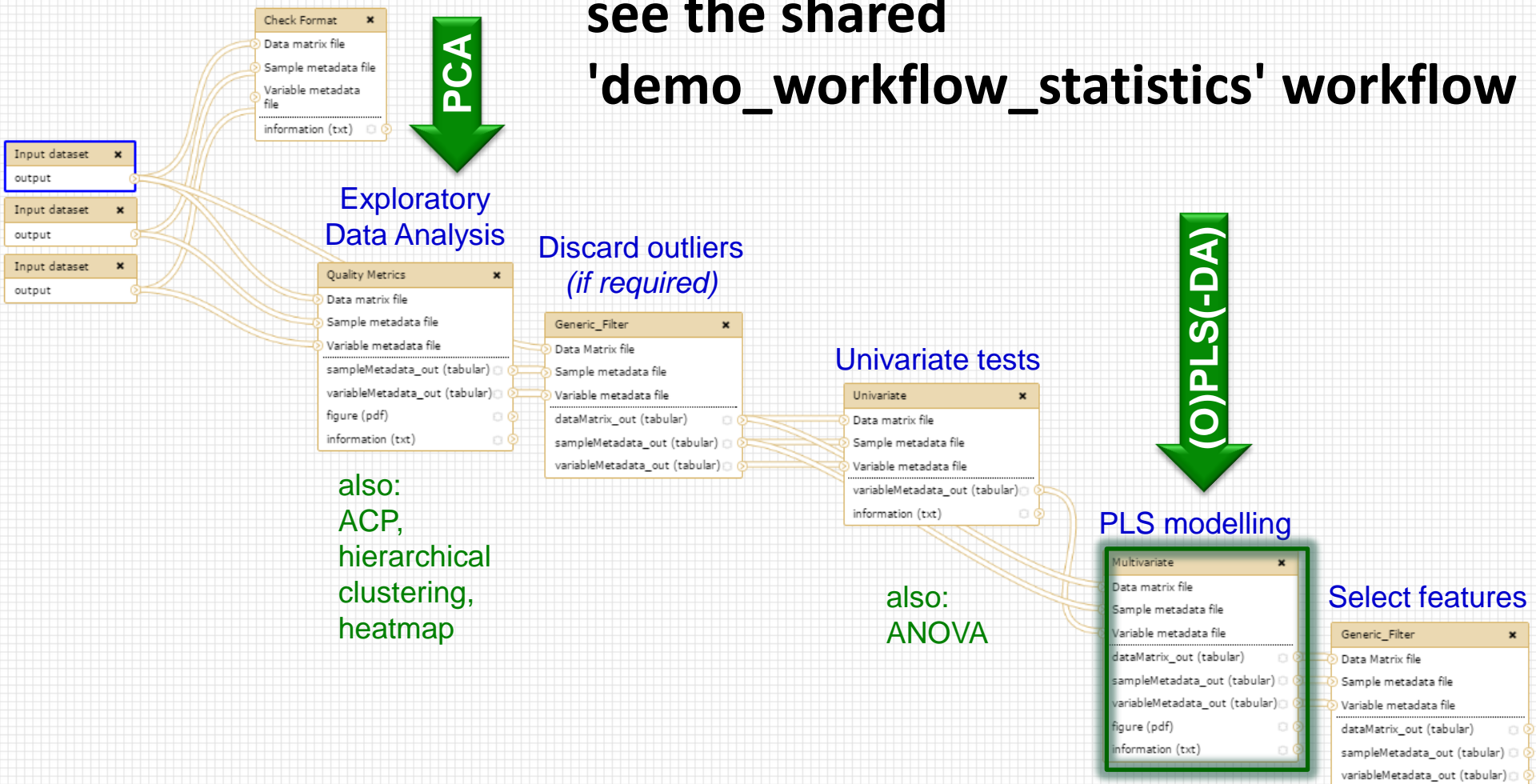
# Latent variable methods

---

- PCA and (O)PLS(-DA) are **latent variable** methods: new components are computed as linear combinations of the original variables
- The assumption is that a few components can efficiently represent the whole dataset (PCA) or model the factor of interest (O)PLS(-DA)
- **Other powerful multivariate methods** exist for regression and classification (Support Vector Machine, Random Forest, etc.)
  - soon available on W4M

# PCA and (O)PLS(-DA) steps in the analysis

see the shared  
'demo\_workflow\_statistics' workflow



# Open the "Multivariate" module

- and select your 3 files of interest:

The screenshot displays the Galaxy web interface for the 'Multivariate' tool. The interface is divided into three main sections: Tools, the tool configuration area, and History.

- Tools Panel (Left):** Shows a search bar and a list of tool categories. The 'Statistical Analysis' category is expanded, and 'Multivariate PCA, PLS and OPLS' is selected, indicated by a green box and callout '1'.
- Tool Configuration Area (Center):** Shows the 'Multivariate (version 2015-04-25)' tool configuration.
  - Callout 2:** Points to the 'Y Response (for PLS(-DA) and OPLS(-DA) only):' dropdown menu, which is set to 'none'.
  - Callout 3:** Points to the 'Data matrix file:' dropdown menu, which is set to '1: dataMatrix.tsv'.
  - Callout 4:** Points to the 'Sample metadata file:' dropdown menu, which is set to '2: sampleMetadata.tsv'.
  - Callout 5:** Points to the 'Variable metadata file:' dropdown menu, which is set to '3: variableMetadata.tsv'.
- History Panel (Right):** Shows the workflow history. The steps are:
  - 1: dataMatrix.tsv
  - 2: sampleMetadata.tsv
  - 3: variableMetadata.tsv
  - 4: Check Format information.txt

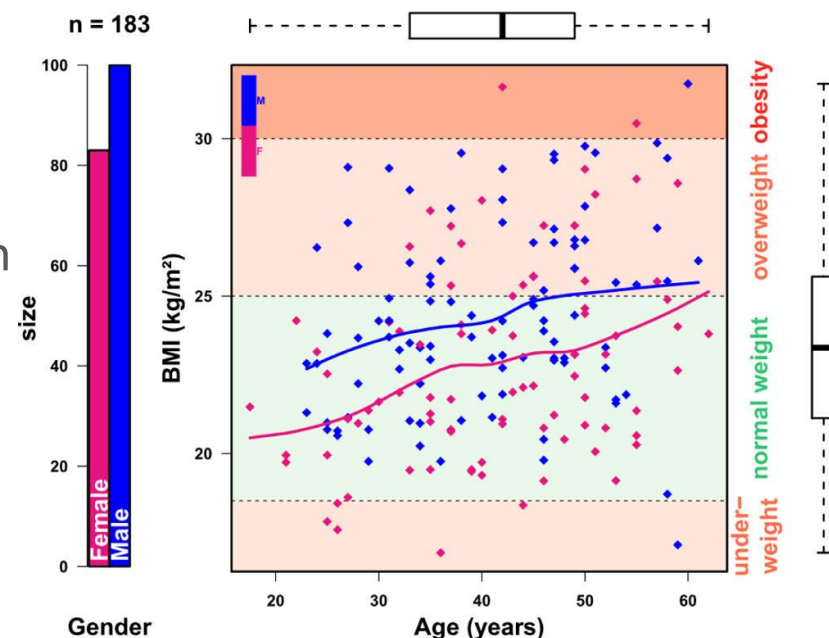
- you are now ready to start your multivariate analyzes!

# Sacurine dataset

- Objective: influence of age, body mass index and gender on metabolite concentrations in urine
- Cohort: 183 employees from the CEA institute
- Analytics: LTQ-Orbitrap (negative ionization mode)
- Annotation: 109 metabolites were identified or annotated at the MSI level 1 or 2.
- Pre-processing:
  - XCMS followed by Quan Browser
  - Signal drift and batch effect correction
  - Normalization to the osmolality
  - log<sub>10</sub> transformation

Thévenot E.A., Roux A., Xu Y., Ezan E. and Junot C. (2015). Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *Journal of Proteome Research*, **14**:3322-3335.

<http://dx.doi.org/10.1021/acs.jproteome.5b00354>





# PRINCIPAL COMPONENT ANALYSIS (PCA)



# Objectives

---

- Visualize the dataMatrix
  - by selecting a few components which capture most of the spread (variance) of the cloud of samples
- Detect outliers
  - which may bias the computation of the component
- Detect clusters of samples
  - which may suggest an internal structuration of the data



# Unsupervised analysis

**$p = 30$  (quantitative) variables**

**$n = 20$  samples**

	1,7-Dimethyluric acid	Dehydroepiandrosterone sulfate	Acetaminophen glucuronide
H011	2114	29025	44
H023	43274	639	2
H033	22386	325	1933
H042	8185	13938	933
H052	22385	357	5004
H062	6380	292	1
H073	10012	22781	1
H083	30414	105	1
H092	6637	35156	1
H103	12100	2	1
H114	33362	149041	46
H124	11197	84536	1
H134	18698	34053	254
H145	14435	212398	52
H157	31732	19317	2200
H168	10221	78	475
H180	22936	463	1
H189	14423	1039	220
H199	2888	12272	37
H209	12563	100236	2

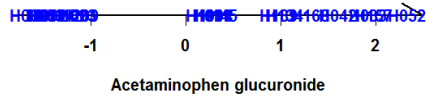
...

**X**

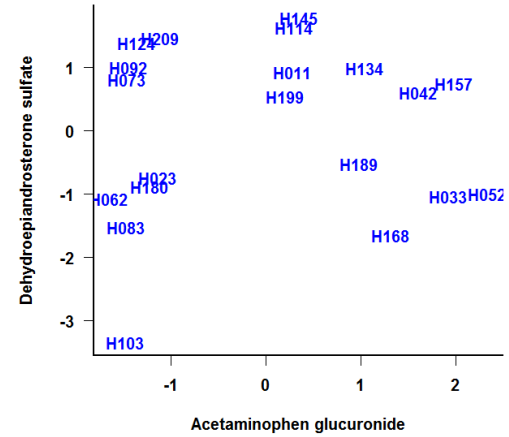


# How to visualize multivariate observations?

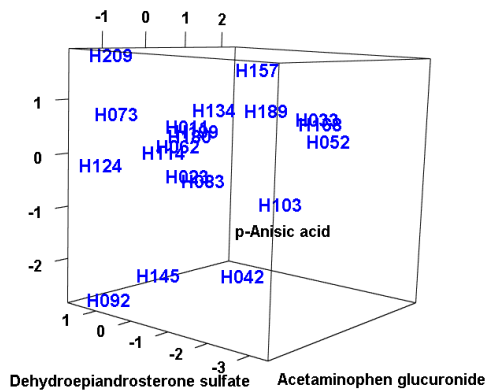
## 1 variable



## 2 variables



## 3 variables



## p variables



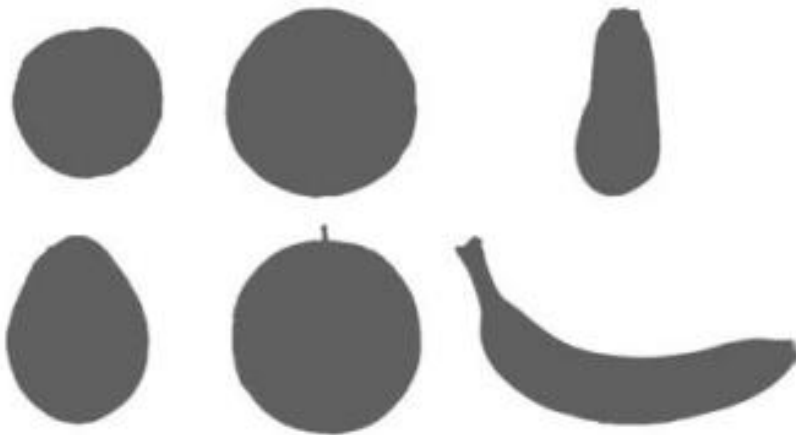
=> Dimension reduction



# Projection

---

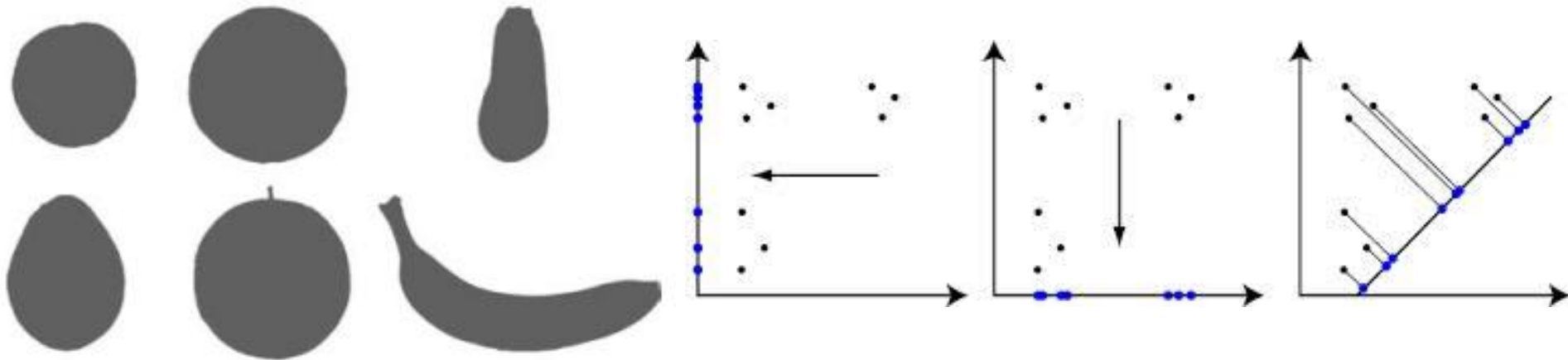
- Projected distances as high as possible



Husson and Pages (2011). Exploratory multivariate analysis by example using R. Chapman & Hall/CRC

# Projection on latent variables

- Projected distances as high as possible
- Define new variables as linear combination of original ones



Husson and Pages (2011). Exploratory multivariate analysis by example using R. Chapman & Hall/CRC

# Selection of PCA as the type of analysis

- Keep the "Y response" to 'none' for PCA (unsupervised analysis)

The screenshot displays the Galaxy 4 Metabolomics interface. The main window shows the configuration for the 'Multivariate' tool (version 2015-04-25). The configuration includes:

- Data matrix file:** 1: dataMatrix.tsv (variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular)
- Sample metadata file:** 2: sampleMetadata.tsv (sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular)
- Variable metadata file:** 3: variableMetadata.tsv (variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular)
- Y Response (for PLS(-DA) and OPLS(-DA) only):** none (highlighted with a green box)
- Number of predictive components:** NA
- Number of orthogonal components (for OPLS(-DA) only):** 0
- Advanced graphical parameters:** Use default

Notes for the Y Response field: 1) PCA: keep the default (none); 2) PLS(-DA) and OPLS(-DA): indicate the name of the column of the sample table to be modeled.

The interface also shows a 'Tools' sidebar on the left with categories like 'Upload File from your computer', 'Export Data', 'LC-MS', 'Format Conversion', 'Preprocessing', 'Normalisation', 'Quality Control', and 'Statistical Analysis'. The 'Statistical Analysis' section is expanded, showing options for 'Univariate statistics', 'Multivariate PCA, PLS and OPLS', 'Anova', 'ACP', 'Hierarchical Clustering', and 'Heatmap'. The 'History' panel on the right shows a list of datasets: '4: Check Format information.txt', '3: variableMetadata.tsv', '2: sampleMetadata.tsv', and '1: dataMatrix.tsv'.

# Automatic selection of the number of components

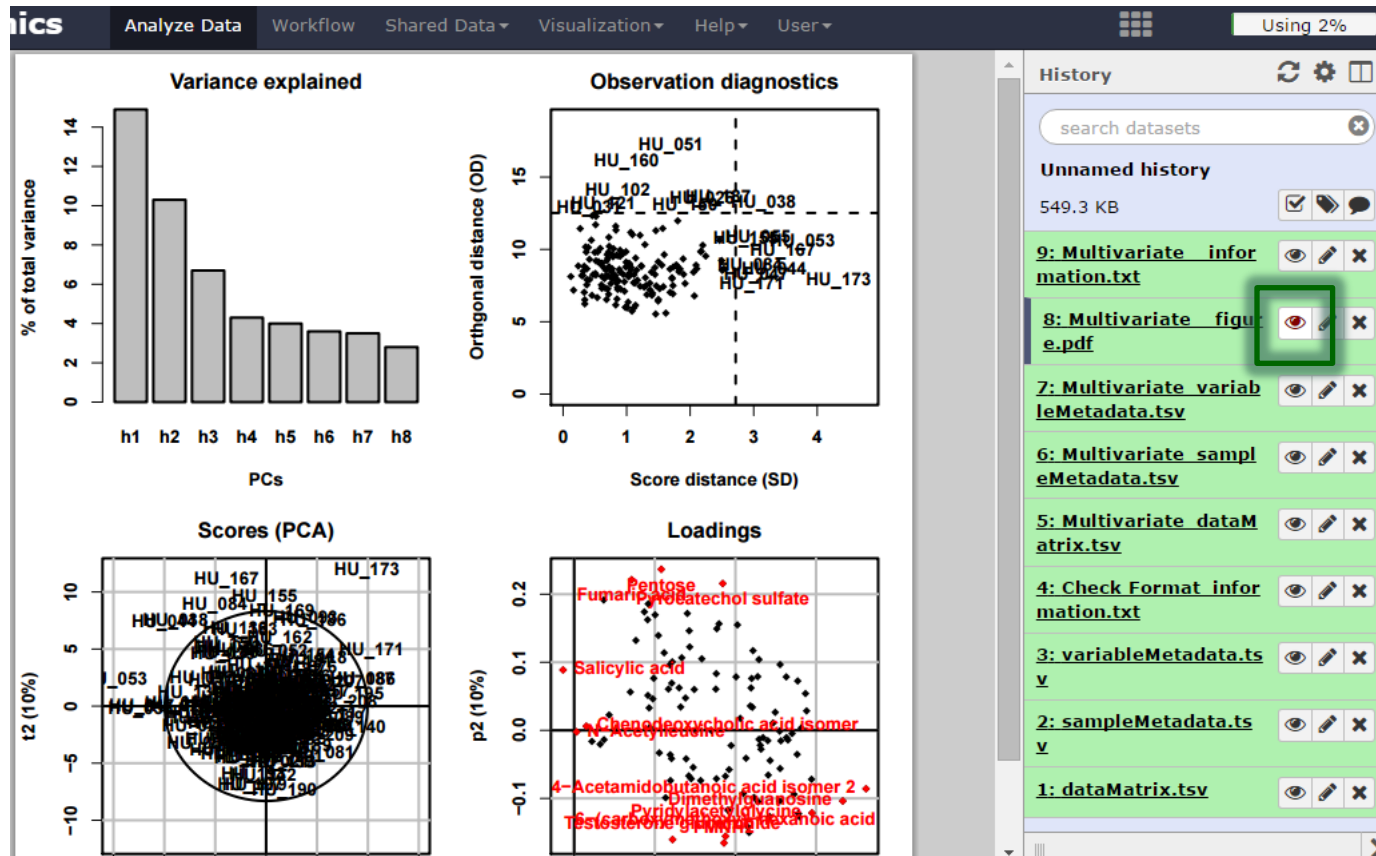
- Until the variance is less than the mean variance of all components

The screenshot displays the Galaxy 4 Metabolomics interface. The main window shows the configuration for the 'Multivariate (version 2015-04-25)' tool. The 'Number of predictive components' dropdown menu is highlighted with a green box and is set to 'NA'. Below it, the 'Number of orthogonal components (for OPLS(-DA) only):' dropdown is set to '0'. The 'Advanced graphical parameters' dropdown is set to 'Use default'. The 'Y Response (for PLS(-DA) and OPLS(-DA) only):' dropdown is set to 'none'. The 'Data matrix file' is '1: dataMatrix.tsv', the 'Sample metadata file' is '2: sampleMetadata.tsv', and the 'Variable metadata file' is '3: variableMetadata.tsv'. The 'History' panel on the right shows a list of datasets, including '4: Check Format information.txt', '3: variableMetadata.tsv', '2: sampleMetadata.tsv', and '1: dataMatrix.tsv'. The 'Tools' panel on the left lists various analysis tools such as 'Univariate statistics', 'Multivariate PCA, PLS and OPLS', 'Anova', 'ACP', 'Hierarchical Clustering', and 'Heatmap'.



# Graphical results

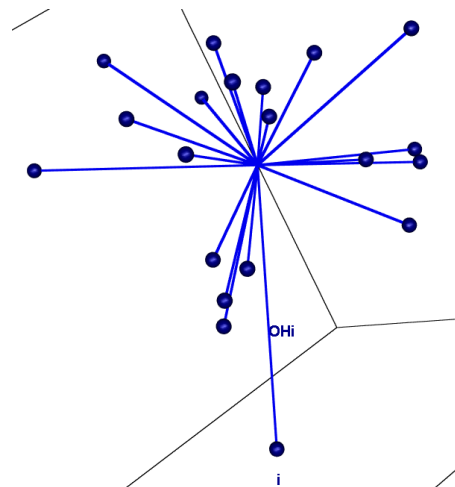
- scree plot, outliers, and the loading and score plots



# Diagnostics R2X: How much of the original inertia is still reflected by the model?

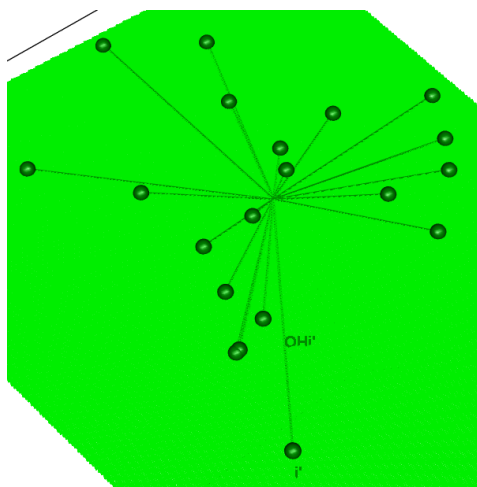
Total

$$TSS = \sum_{i=1}^n OH_i^2$$



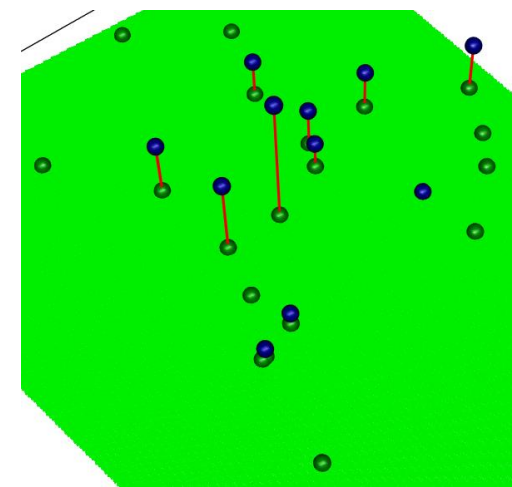
Explained

$$ESS = \sum_{i=1}^n OH'_i{}^2$$



Residual

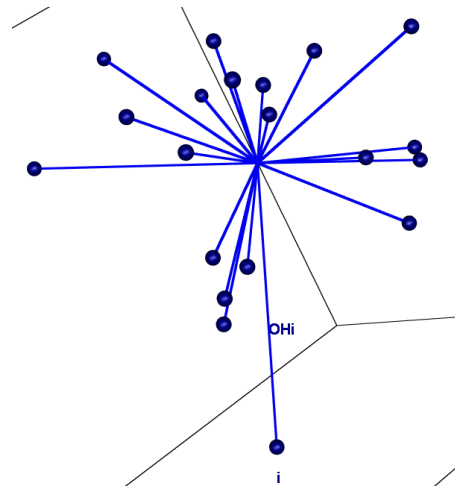
$$RSS = \sum_{i=1}^n HH'_i{}^2 = TSS - ESS$$



# Diagnostics R2X: How much of the original inertia is still reflected by the model?

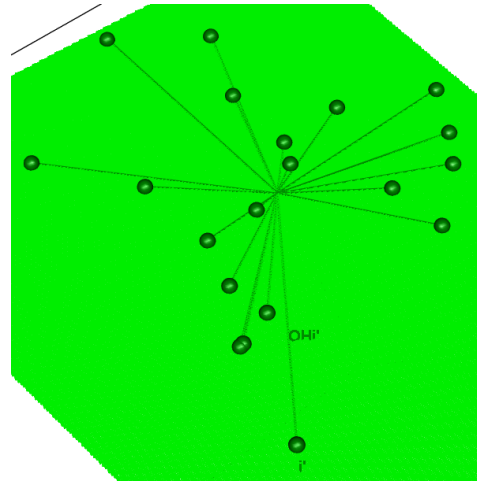
Total

$$TSS = \sum_{i=1}^n OH_i^2$$



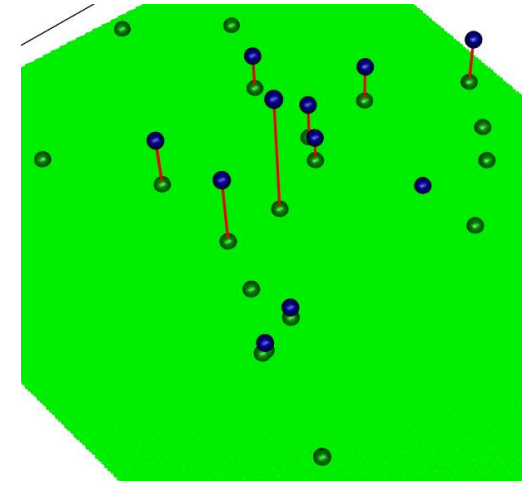
Explained

$$ESS = \sum_{i=1}^n OH'_i{}^2$$



Residual

$$RSS = \sum_{i=1}^n HH'_i{}^2 = TSS - ESS$$

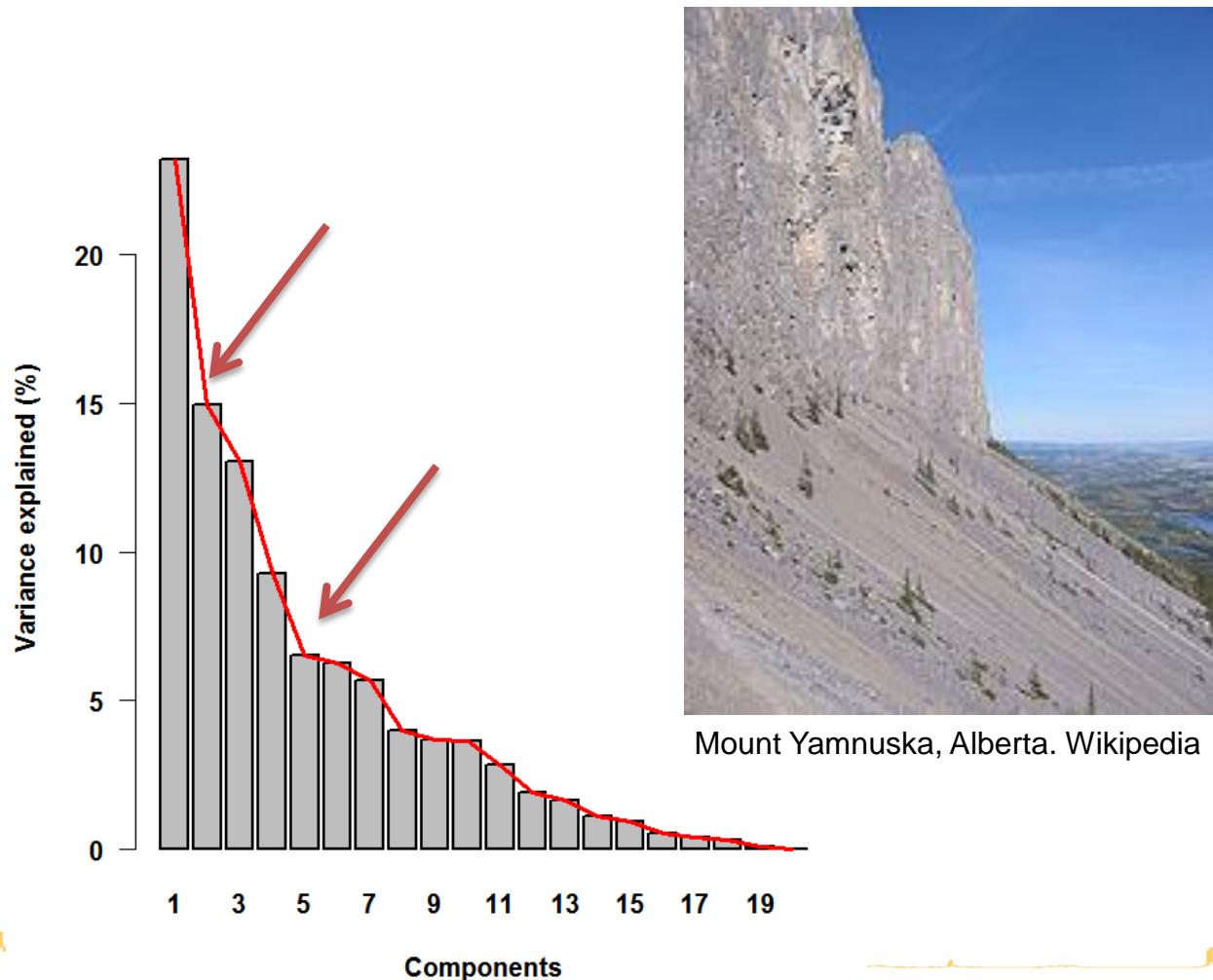


$$R2X = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad 0 \leq R2X \leq 1$$

- R2X increases with the number of components in the model
- For a given number of components, the higher the R2X, the more inertia is captured by the model (projection)

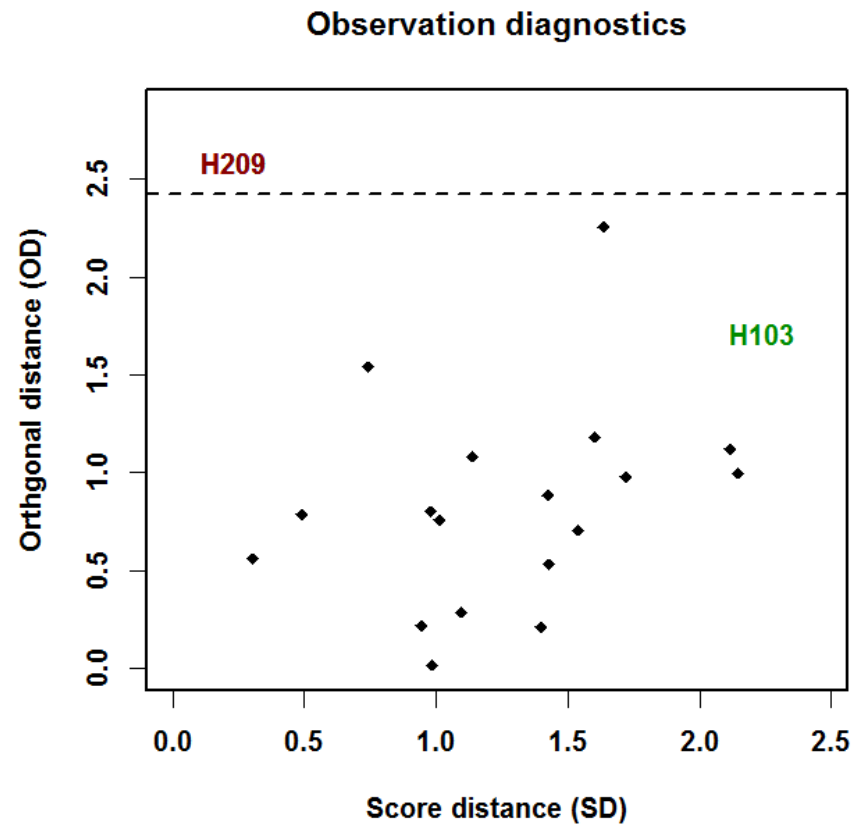
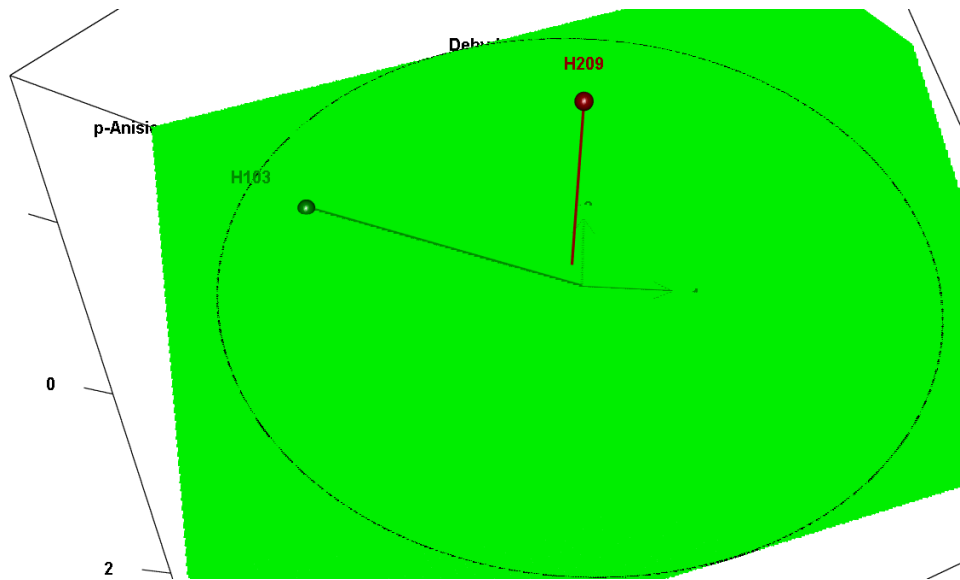
# Scree plot

- Check that the first components capture most of the variance

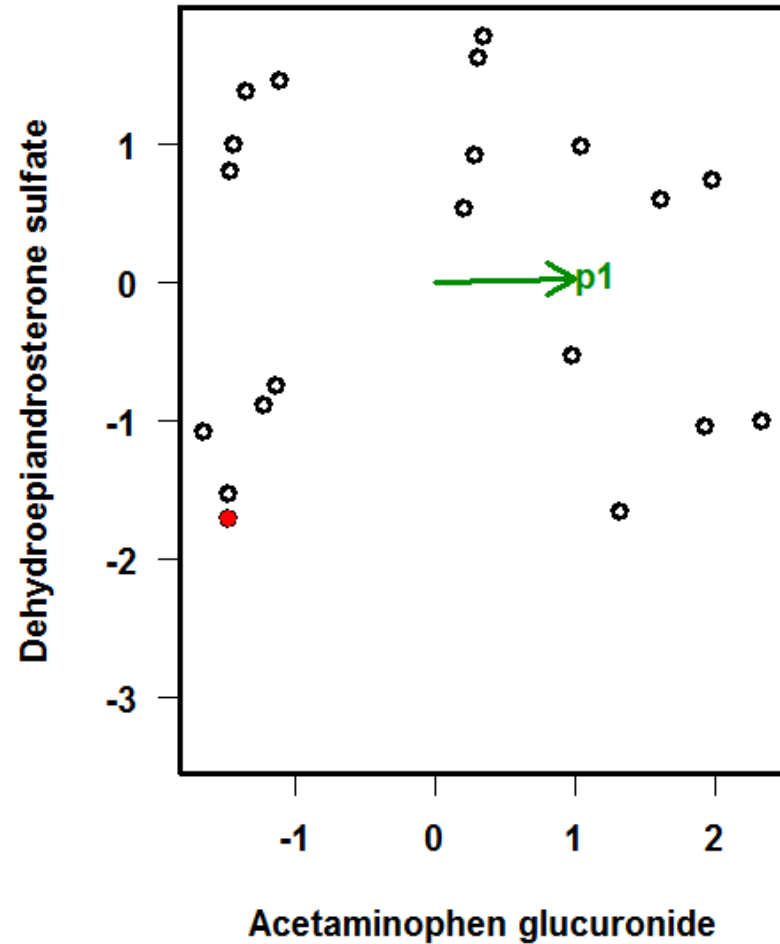
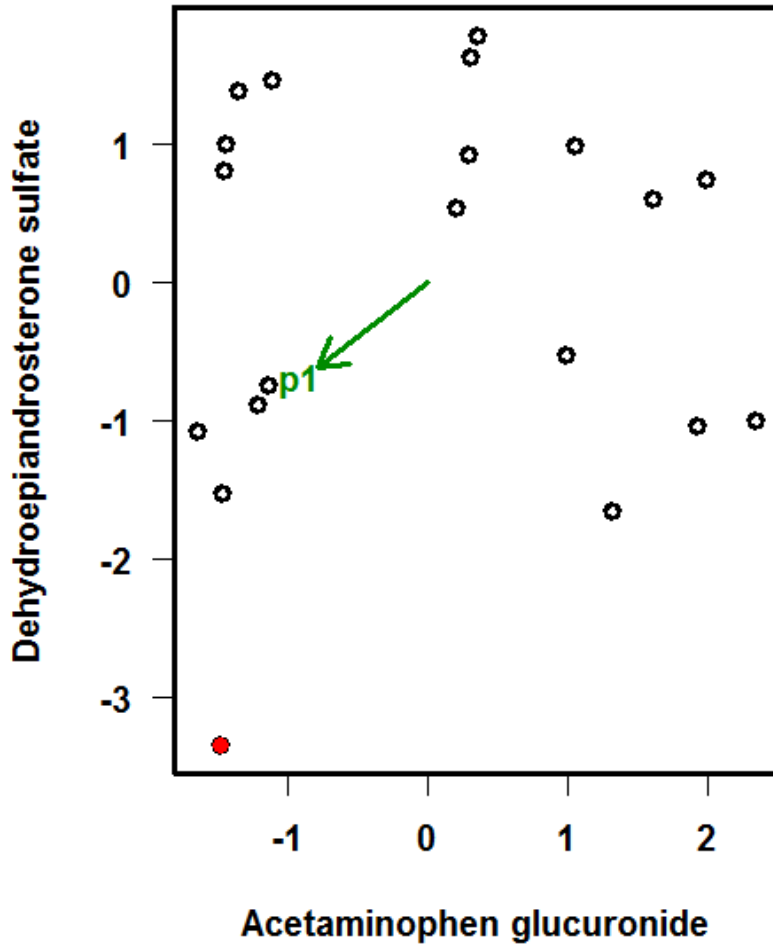


# Observation diagnostics

- Samples which may bias the PCA computation and/or may not be faithfully visualized by the score plot



# Sensitivity to outliers



# Numerical results

- Numerical results (including the percentage of explained inertia) can be viewed in the "information.txt" file

The screenshot displays the Workflow4Metabolomics software interface. The main window shows the results of a PCA analysis performed using the 'svd' algorithm. The number of components is 8, and there are 183 reference observations (100%). A table of correlations between variables and components is shown, with variables like Salicylic acid and N-Acetylleucine. A 'Model overview' table is highlighted with a green box, showing R2X and R2X(cum) values for components h1 through h8. The history panel on the right lists several datasets, with '9: Multivariate information.txt' highlighted and its eye icon circled in green.

**PCA ('svd' algorithm)**  
Number of components: 8

Number of reference observations: 183 (100%)

Correlations between variables and components:

	h1	h2	cor_h1	cor_h2
Salicylic acid	-0.0069	NA	-0.028	NA
N-Acetylleucine	0.0015	NA	0.006	NA
Chenodeoxycholic acid isomer	0.0075	NA	0.030	NA
Pyridylacetylglycine	0.1500	NA	0.590	NA
Dimethylguanosine	0.1700	NA	0.670	NA
4-Acetamidobutanoic acid isomer 2	0.1800	NA	0.730	NA
FMNH2	NA	-0.17	NA	-0.56
Testosterone glucuronide	NA	-0.16	NA	-0.54
6-(carboxymethoxy)-hexanoic acid	NA	-0.16	NA	-0.52
Pyrocatechol sulfate	NA	0.22	NA	0.72
Fumaric acid	NA	0.22	NA	0.74
Pentose	NA	0.24	NA	0.79

**Model overview:**

	R2X	R2X(cum)	Iter.
h1	0.149	0.149	0
h2	0.103	0.252	0
h3	0.067	0.319	0
h4	0.043	0.362	0
h5	0.040	0.402	0
h6	0.036	0.438	0
h7	0.035	0.473	0
h8	0.028	0.501	0

**Model summary:**

	R2X(cum)	ncp	nco
h8	0.501	8	0

**History Panel:**

- 9: Multivariate information.txt (highlighted)
- 8: Multivariate figure.pdf
- 7: Multivariate variableMetadata.tsv
- 6: Multivariate sampleMetadata.tsv
- 5: Multivariate dataMatrix.tsv
- 4: Check Format information.txt
- 3: variableMetadata.tsv
- 2: sampleMetadata.tsv
- 1: dataMatrix.tsv

# Score and loading values

- The score (resp. loading) values of the selected components have been added as columns in the **sampleMetadata** (resp. **variableMetadata**) files

The screenshot displays the olomics software interface. The main window shows a table with columns for sampleMetadata, age, bmi, gender, PCA\_XSCOR-h1, and PCA\_XSCOR-h2. The PCA\_XSCOR-h1 and PCA\_XSCOR-h2 columns are highlighted with a green box. On the right side, there is a History panel showing a list of datasets. The dataset '6: Multivariate sampleMetadata.tsv' is highlighted with a green box and a red eye icon, with a '1' in a green box next to it. The dataset '7: Multivariate variableMetadata.tsv' is also highlighted with a green box and a red eye icon, with a '2' in a green box next to it.

sampleMetadata	age	bmi	gender	PCA_XSCOR-h1	PCA_XSCOR-h2
HU_011	29	19.75	M	-8.74400891504494	0.29249883857013
HU_014	59	22.64	F	-1.86532133217634	0.285366844636407
HU_015	42	22.72	M	-6.74648640072742	-0.561605063374045
HU_017	41	23.03	M	-4.23534187957954	-0.641487554413452
HU_018	34	20.96	M	1.59252091681441	-2.89331923169429
HU_019	35	23.41	M	-1.2535250688467	0.200242710800258
HU_020	59	17.1	M	-5.47756634951485	-0.378911997626029
HU_021	34	23.36	M	1.08538964511728	-4.94025884576605
HU_022	51	28.23	F	-3.66836013881533	5.14176542596851
HU_023	51	29.55	M	-4.66609702458129	-1.17204283780617
HU_024	57	29.86	M	-0.794642666784698	-1.22728974524632
HU_025	53	21.6	M	-2.2313493995232	-2.91021037882818
HU_026	34	23.46	F	-8.79694543308979	-0.000101601980933629
HU_027	37	24.82	M	-7.0432093146523	-1.70548152914905
HU_028	41	23.92	F	-0.443606341382212	-3.16113671135982
HU_029	37	27.78	M	-4.50849252383876	-1.54412237704366
HU_030	49	25.88	M	0.60173477063632	-2.47896644698659
HU_031	25	20.76	M	0.209079981357257	-1.36514244700848
HU_032	38	24.09	F	2.3788535799504	2.08848500995035
HU_033	44	18.36	F	1.87769511456898	2.57155836373107
HU_034	52	23.37	M	-3.22008044172578	2.86622150577896
HU_035	37	20.7	F	3.2801149214796	-1.24975766384474
HU_036	47	29.51	M	-2.47266540217536	4.88240826458344
HU_037	35	25.62	M	-4.74331054976355	-2.89213123664626
HU_038	52	22.72	M	-8.90649077106328	7.54124509052761
HU_039	45	24.9	M	-4.23718132839903	4.62422497226667



# Tuning the parameters

- You can recall the page with your parameters, modify them, and restart the analysis

The screenshot displays the Workflow4Metabolomics software interface. The main window shows the 'Analyze Data' workflow with the following parameters:

- Sample metadata file:** 2: sampleMetadata.tsv
- Variable metadata file:** 3: variableMetadata.tsv
- Y Response (for PLS(-DA) and OPLS(-DA) only):** none
- Number of predictive components:** 3
- Number of orthogonal components (for OPLS(-DA) only):** 0
- Advanced graphical parameters:** Use default
- Advanced computational parameters:** Use default

An 'Execute' button is located at the bottom left. A green box with the number '4' highlights this button. A green box with the number '3' highlights the 'Number of predictive components' dropdown menu.

On the right side, a 'History' panel is visible, showing a list of saved analysis states. A green box with the number '1' highlights the '9: Multivariate information.txt' entry. A green box with the number '2' highlights the '8: Multivariate figure.pdf' entry. A green box with the number '1' also highlights the '8: Multivariate figure.pdf' entry.

# Advanced parameters: Scaling

- Variables are mean-centered for PCA
- By default, they are also unit-variance scaled
  - absence of variance scaling or changing to Pareto scaling can be selected in the advanced computational parameters

The screenshot shows the 'omics' software interface. The main panel displays the 'Advanced computational parameters' section, which is highlighted with a green box and a red '1'. Below it, the 'Scaling:' dropdown menu is set to 'pareto', also highlighted with a green box and a red '2'. The interface includes a navigation bar with 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User' menus. A 'History' panel on the right shows a list of datasets, including '9: Multivariate information.txt', '8: Multivariate figure.pdf', '7: Multivariate variableMetadata.tsv', '6: Multivariate sampleMetadata.tsv', '5: Multivariate dataMatrix.tsv', and '4: Check Format information.txt'. The 'Using 2%' indicator is visible in the top right corner.

**omics** Analyze Data Workflow Shared Data Visualization Help User Using 2%

none

Notes: 1) PCA: keep the default (none); 2) PLS(-DA) and OPLS(-DA): indicate the name of the column of the sample table to be modeled

**Number of predictive components:**

3

Notes: 1) PCA and PLS(-DA): NA can be selected to get a suggestion of the optimal number of predictive components; 2) OPLS(-DA) modeling: select 1 predictive component

**Number of orthogonal components (for OPLS(-DA) only):**

0

Notes: 1) PCA and PLS(-DA): keep the default value (0); 2) OPLS(-DA): NA can be selected to get a suggestion of the optimal number of orthogonal components

**Advanced graphical parameters:**

Use default

**Advanced computational parameters:**

Full parameter list

**Scaling:**

pareto

Select Standard mean-centering and unit-variance scaling

**Permutation testing: Number of permutations:**

0

'0' means that no permutation testing will be performed

**Log10 transformation:**

no

**History**

search datasets

Unnamed history

549.3 KB

9: Multivariate information.txt

8: Multivariate figure.pdf

15.1 KB

format: pdf, database: ?

Image in pdf format

7: Multivariate variableMetadata.tsv

6: Multivariate sampleMetadata.tsv

5: Multivariate dataMatrix.tsv

4: Check Format information.txt

3: variableMetadata

# Advanced parameters: Ellipses

- Indicate the column name of **sampleMetadata** to be used

**Workflow4metabolomics** Analyze Data Workflow Shared Data Visualization Help User

Variable metadata file:

variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

**Y Response (for PLS(-DA) and OPLS(-DA) only):**

Notes: 1) PCA: keep the default (none); 2) PLS(-DA) and OPLS(-DA): indicate the name of the column of the sample table to be modeled

**Number of predictive components:**

Notes: 1) PCA and PLS(-DA): NA can be selected to get a suggestion of the optimal number of predictive components; 2) OPLS(-DA) modeling: select 1 predictive component

**Number of orthogonal components (for OPLS(-DA) only):**

Notes: 1) PCA and PLS(-DA): keep the default value (0); 2) OPLS(-DA): NA can be selected to get a suggestion of the optimal number of orthogonal components

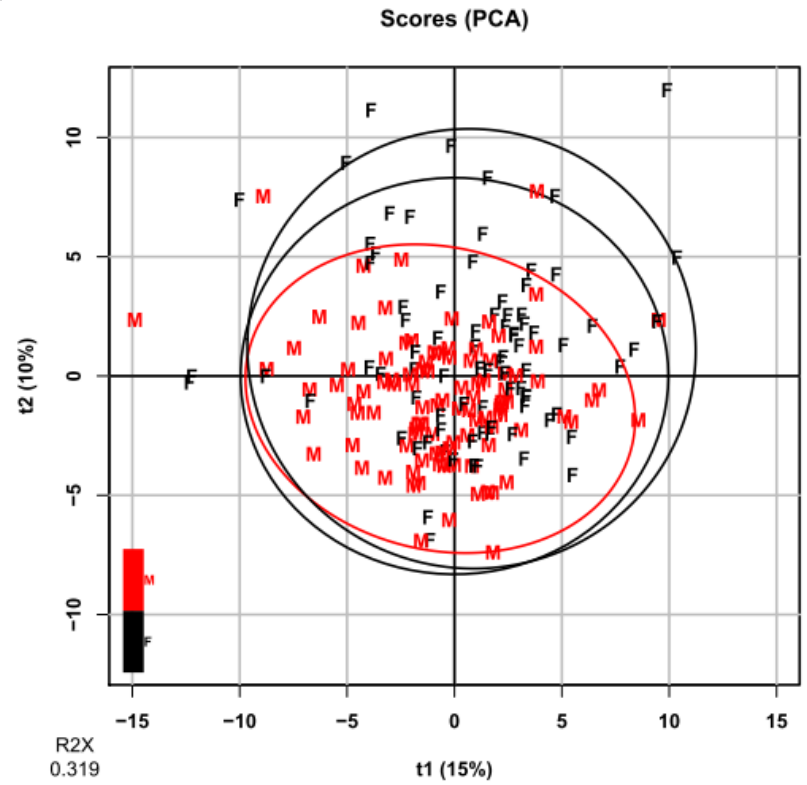
**Advanced graphical parameters:** 1

**Graphic:**

**Mahalanobis ellipses:** 2

Notes: 1) Name of the sample metadata column with the classes to be used for drawing ellipses (for (O)PLS-DA: indicate the same name as the 'Response' argument above); If you do not want ellipses, keep the default, none

**Sample colors:**



# References

---

- Husson F., Le S. and Pages J. (2011). Exploratory multivariate analysis by example using R. *Chapman & Hall/CRC*
- Ringner M. (2008). What is principal component analysis? *Nature Biotechnology*, **26**:303-304.  
<http://dx.doi.org/10.1038/nbt0308-303>
- Baccini A. (2010). Statistique descriptive multidimensionnelle (pour les nuls). [www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf](http://www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf)

**PARTIAL LEAST SQUARES  
REGRESSION (PLS) AND  
DISCRIMINANT ANALYSIS (PLS-DA)**

# PLS(-DA) modelling

---

- Powerful regression method when

$$n_{samples} < p_{variables}$$

- **Complementary to univariate hypothesis testing** (where variables are tested independantly)
- **Risk of overfitting:** i.e., building a model whose (apparently) good performances result from chance only



# Supervised analysis (i.e. with labels)

1 response

$p = 30$  (quantitative) variables

$n = 20$  samples

	bmi
H011	19.8
H023	29.6
H033	18.4
H042	19.8
H052	20.1
H062	22.2
H073	25.4
H083	29.8
H092	21.8
H103	26.8
H114	29.4
H124	22.2
H134	22.9
H145	29.1
H157	22.0
H168	20.8
H180	23.7
H189	19.4
H199	21.0
H209	21.5

	1,7-Dimethyluric acid	Dehydroepiandrosterone sulfate
H011	3.33	4.46
H023	4.64	2.81
H033	4.35	2.51
H042	3.91	4.14
H052	4.35	2.55
H062	3.80	2.47
H073	4.00	4.36
H083	4.48	2.02
H092	3.82	4.55
H103	4.08	0.21
H114	4.52	5.17
H124	4.05	4.93
H134	4.27	4.53
H145	4.16	5.33
H157	4.50	4.29
H168	4.01	1.89
H180	4.36	2.67
H189	4.16	3.02
H199	3.46	4.09
H209	4.10	5.00

...

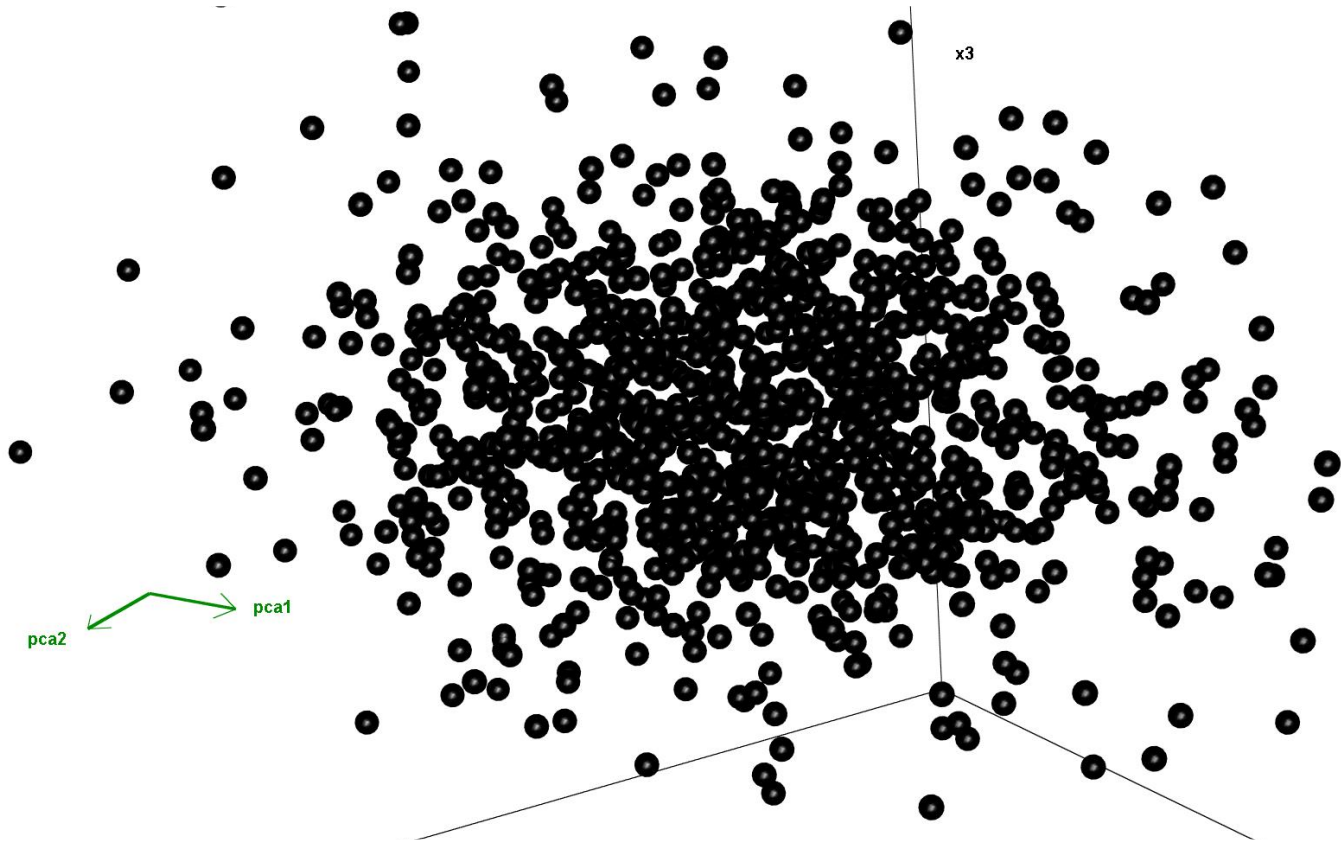
$y$

$X$



# PLS vs PCA

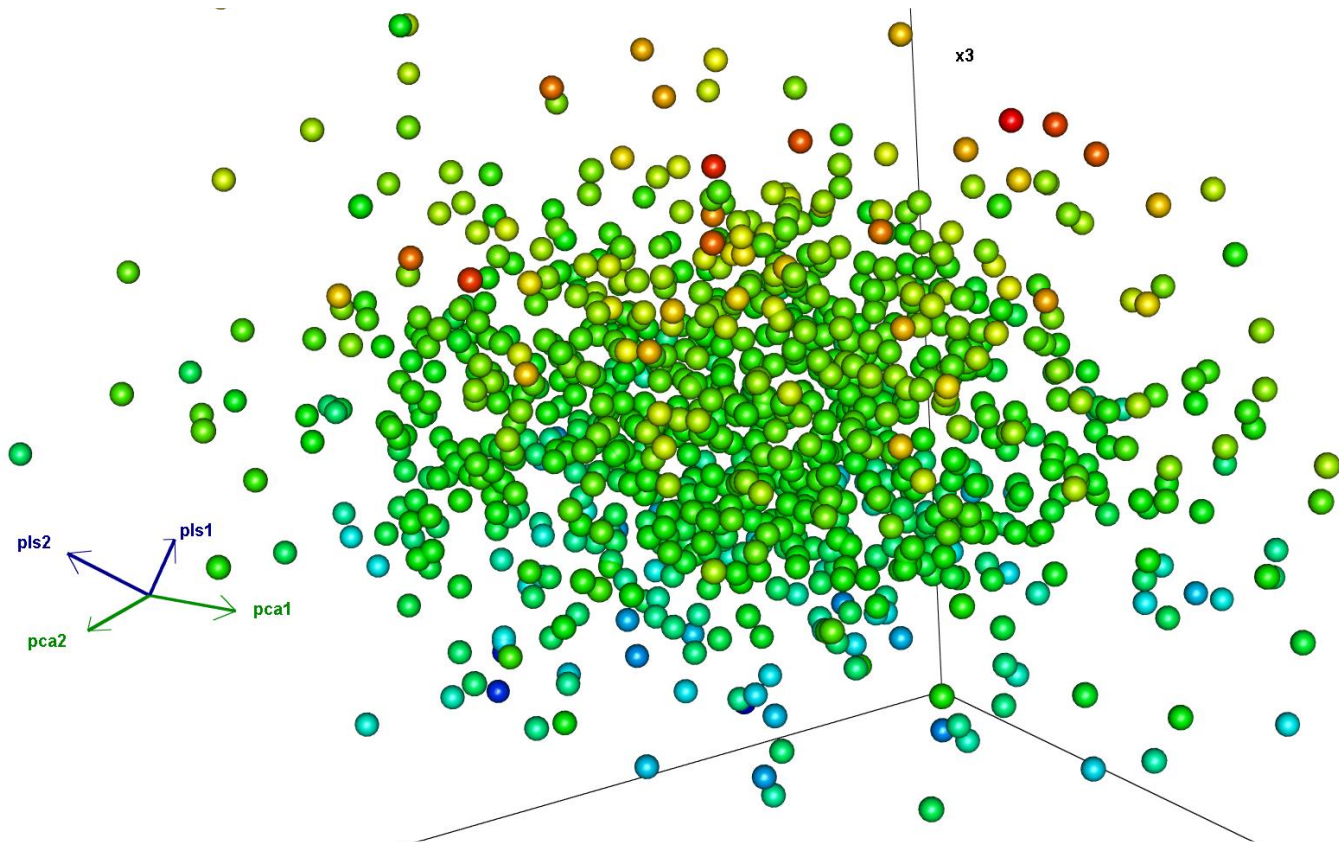
- PCA finds the directions of maximum variance





# PLS vs PCA

- PLS includes the labels into the model



# Selection of PLS(-DA) as the type of analysis

- Select the "Y response" to be modelled (column of **sampleMetadata**):
  - column of numbers (age, bmi): **PLS** regression
  - column of characters ('M'/'F', 'patient'/'control'): **PLS-DA** classification

**Galaxy / 4 / Metabolomics** Analyze Data Workflow Shared Data Visualization Help User Using 2%

**Tools** search tools

**Upload File from your computer**  
**Export Data**

**LC-MS**  
**Format Conversion**  
**Preprocessing**  
**Normalisation**  
**Quality Control**  
**Statistical Analysis**

Univariate Univariate statistics  
Multivariate PCA, PLS and OPLS  
Anova N-way anova. With ou Without interactions  
ACP ellipsoid by factors  
Hierarchical Clustering using ctc R package for java-treeview  
Heatmap Heatmap of the dataMatrix

**Annotation**  
GC-MS  
Preprocessing  
Normalisation  
Quality Control  
Statistical Analysis  
Annotation

**Multivariate (version 2015-04-25)**

**Data matrix file:** 1: dataMatrix.tsv  
variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular

**Sample metadata file:** 2: sampleMetadata.tsv  
sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

**Variable metadata file:** 3: variableMetadata.tsv  
variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

**Y Response (for PLS(-DA) and OPLS(-DA) only):**  
bmi  
Notes: 1) PCA: keep the default (none); 2) PLS(-DA) and OPLS(-DA): indicate the name of the column of the sample table to be modeled

**Number of predictive components:**  
NA  
Notes: 1) PCA and PLS(-DA): NA can be selected to get a suggestion of the optimal number of predictive components; 2) OPLS(-DA) modeling: select 1 predictive component

**Number of orthogonal components (for OPLS(-DA) only):**  
0  
Notes: 1) PCA and PLS(-DA): keep the default value (0); 2) OPLS(-DA): NA can be selected to get a suggestion of the optimal number of orthogonal components

**Advanced graphical parameters:**  
Use default

**Advanced computational parameters:**  
Use default

**Execute**

**History** search datasets

**multivariate\_example**  
14 shown, 10 deleted  
1.1 MB

14: Multivariate information.txt  
13: Multivariate figure.pdf  
12: Multivariate variableMetadata.tsv  
11: Multivariate sampleMetadata.tsv  
10: Multivariate dataMatrix.tsv  
9: Multivariate information.txt  
8: Multivariate figure.pdf  
15.1 KB  
format: pdf, database: ?  
Image in pdf format

7: Multivariate variableMetadata.tsv  
6: Multivariate sampleMetadata.tsv  
5: Multivariate\_data

**Author Etienne Thevenot (etienne.thevenot@cea.fr)**

34

# Automatic selection of the number of components

- A new component  $h$  is added to the model if:
  - $R^2Y_h \geq 1\%$
  - $Q^2Y_h \geq 0$  (or 5% if  $n_{samples} \leq 100$ )

Note:  $Q^2Y_h = 1 - \frac{PRESS_h}{RSS_{h-1}}$  where  $PRESS_h$  is estimated by cross-validation

The screenshot shows the Galaxy 4 / Metabolomics interface. The main panel displays the configuration for the 'Multivariate (version 2015-04-25)' tool. The configuration includes:

- Data matrix file:** 1: dataMatrix.tsv (variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular)
- Sample metadata file:** 2: sampleMetadata.tsv (sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular)
- Variable metadata file:** 3: variableMetadata.tsv (variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular)
- Y Response (for PLS(-DA) and OPLS(-DA) only):** bmi
- Number of predictive components:** NA (highlighted with a green box)
- Number of orthogonal components (for OPLS(-DA) only):** 0

Notes: 1) PCA: keep the default (none); 2) PLS(-DA) and OPLS(-DA): indicate the name of the column of the sample table to be modeled.

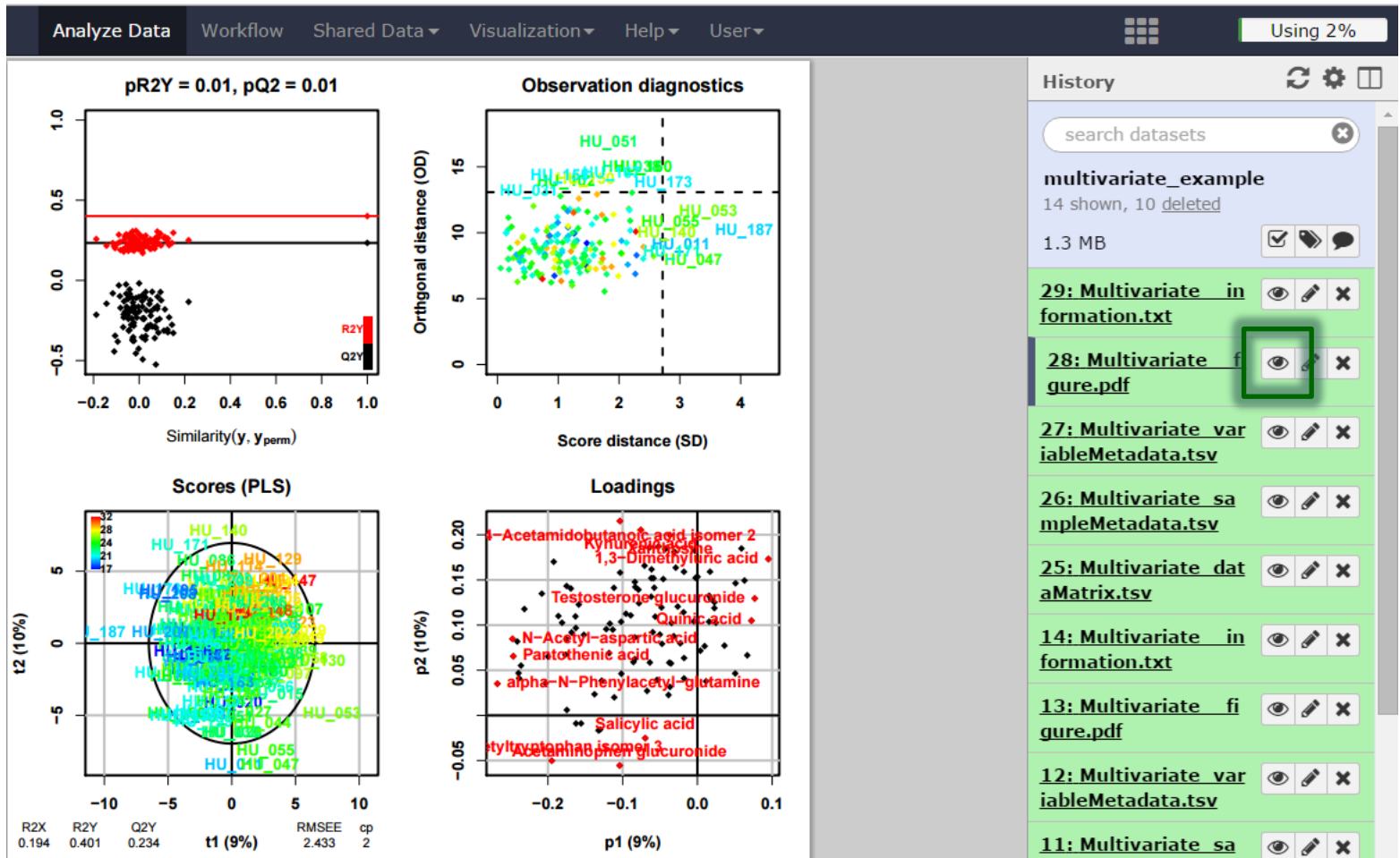
Notes: 1) PCA and PLS(-DA): NA can be selected to get a suggestion of the optimal number of predictive components; 2) OPLS(-DA) modeling: select 1 predictive component

Notes: 1) PCA and PLS(-DA): keep the default value (0); 2) OPLS(-DA): NA can be selected to get a suggestion of the optimal

The right sidebar shows the History panel with a list of datasets, including '14: Multivariate\_in\_formation.txt', '13: Multivariate\_fi\_gure.pdf', '12: Multivariate\_var\_iableMetadata.tsv', '11: Multivariate\_sa\_mpleMetadata.tsv', '10: Multivariate\_dat\_aMatrix.tsv', '9: Multivariate\_inf\_ormation.txt', and '8: Multivariate\_fig\_ure.pdf'.

# Graphical results

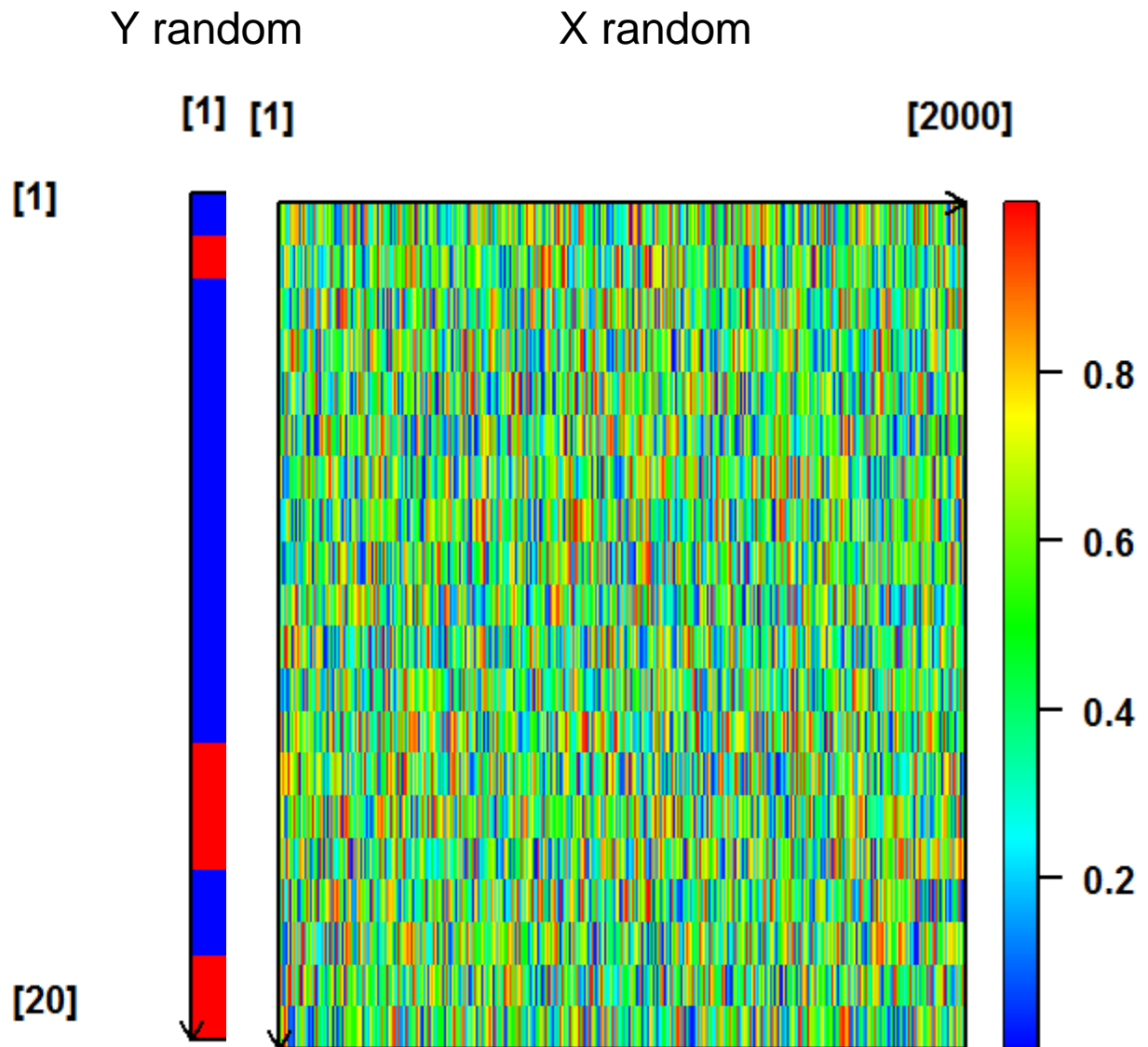
- permutation, overview, outlier, and score plots displayed as the default ('summary')



# Overfitting

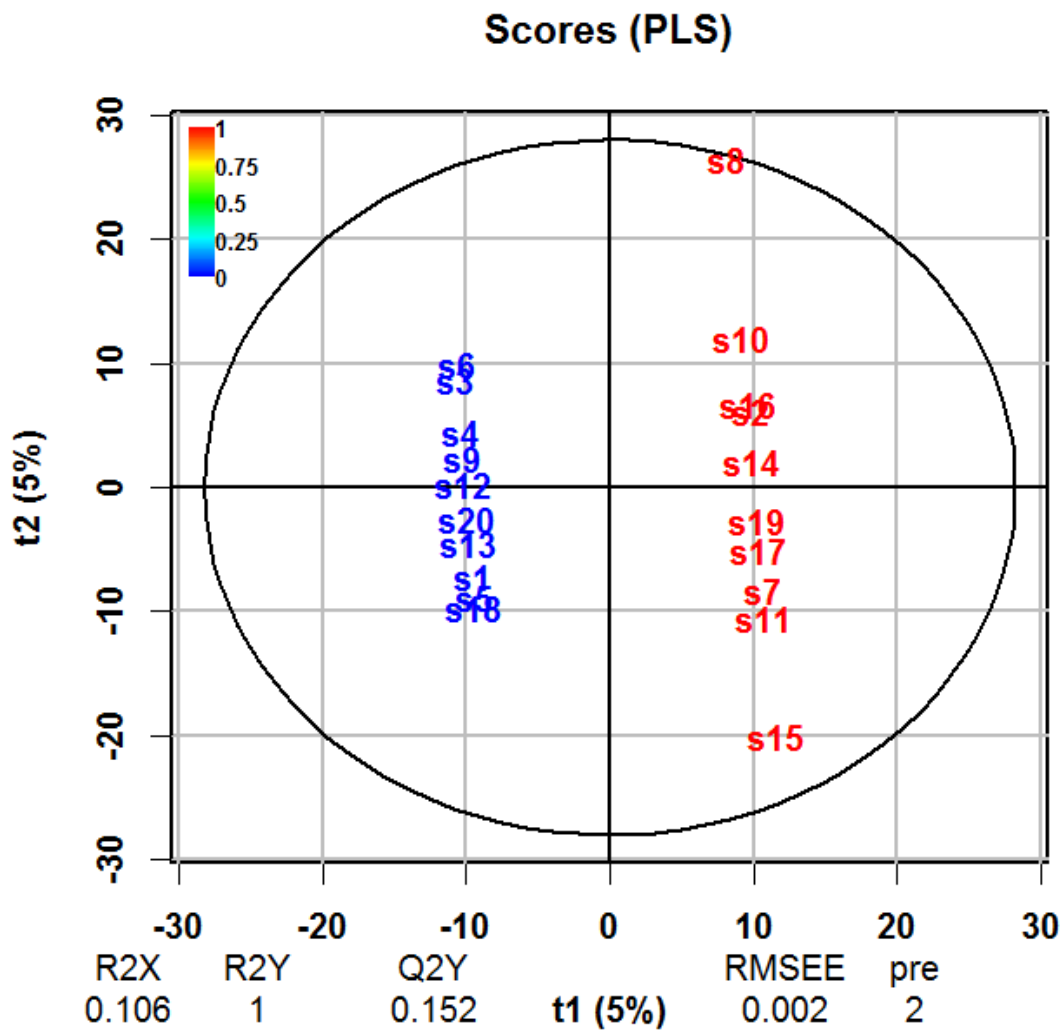


- X:  $20 \times 2,000$  matrix of **random** numbers
  - Uniform distribution between 0 and 1
- Y:  $20 \times 1$  matrix of **random** labels
  - 0 or 1 values



adapted from Wehrens (2011).  
Chemometrics with R. Springer.

# Score plot!





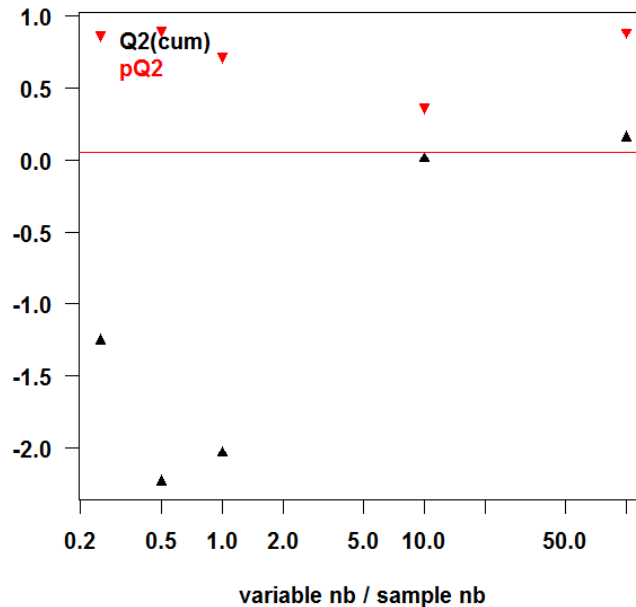
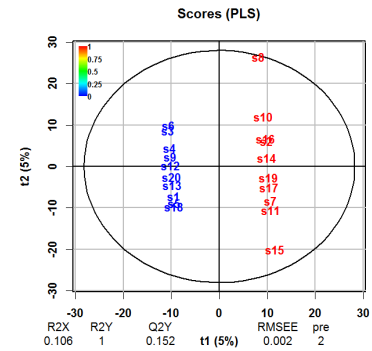
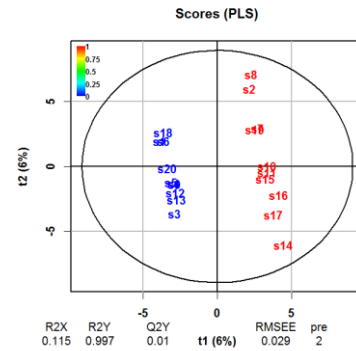
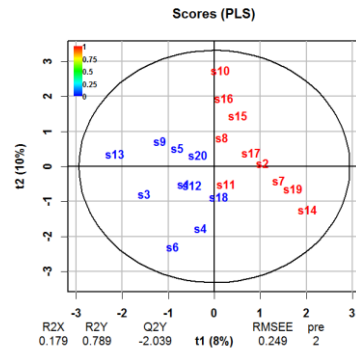
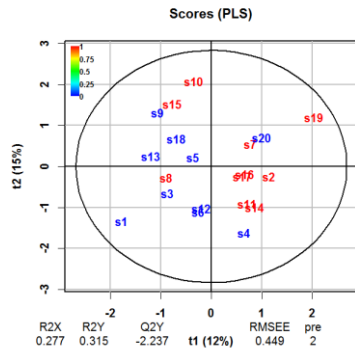
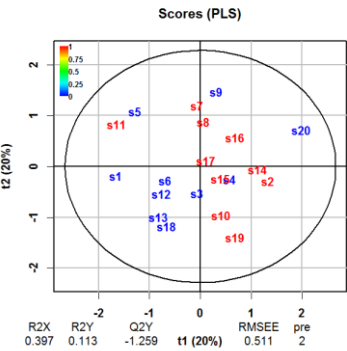


# Risk of overfitting when $n < p$



variables  
= samples

0.2                      0.5                      1                      10                      100





# Significance of the model

- The algorithm randomly permutes the **y** labels, builds the models and computes the  $R^2X$ ,  $R^2Y$ ,  $Q^2Y$

1 response

$p = 30$  (quantitative) variables

$n = 20$  samples

	bmi
H011	19.8
H023	29.6
H033	18.4
H042	19.8
H052	20.1
H062	22.2
H073	25.4
H083	29.8
H092	21.8
H103	26.8
H114	29.4
H124	22.2
H134	22.9
H145	29.1
H157	22.0
H168	20.8
H180	23.7
H189	19.4
H199	21.0
H209	21.5



	1,7-Dimethyluric acid	Dehydroepiandrosterone sulfate
H011	3.33	4.46
H023	4.64	2.81
H033	4.35	2.51
H042	3.91	4.14
H052	4.35	2.55
H062	3.80	2.47
H073	4.00	4.36
H083	4.48	2.02
H092	3.82	4.55
H103	4.08	0.21
H114	4.52	5.17
H124	4.05	4.93
H134	4.27	4.53
H145	4.16	5.33
H157	4.50	4.29
H168	4.01	1.89
H180	4.36	2.67
H189	4.16	3.02
H199	3.46	4.09
H209	4.10	5.00

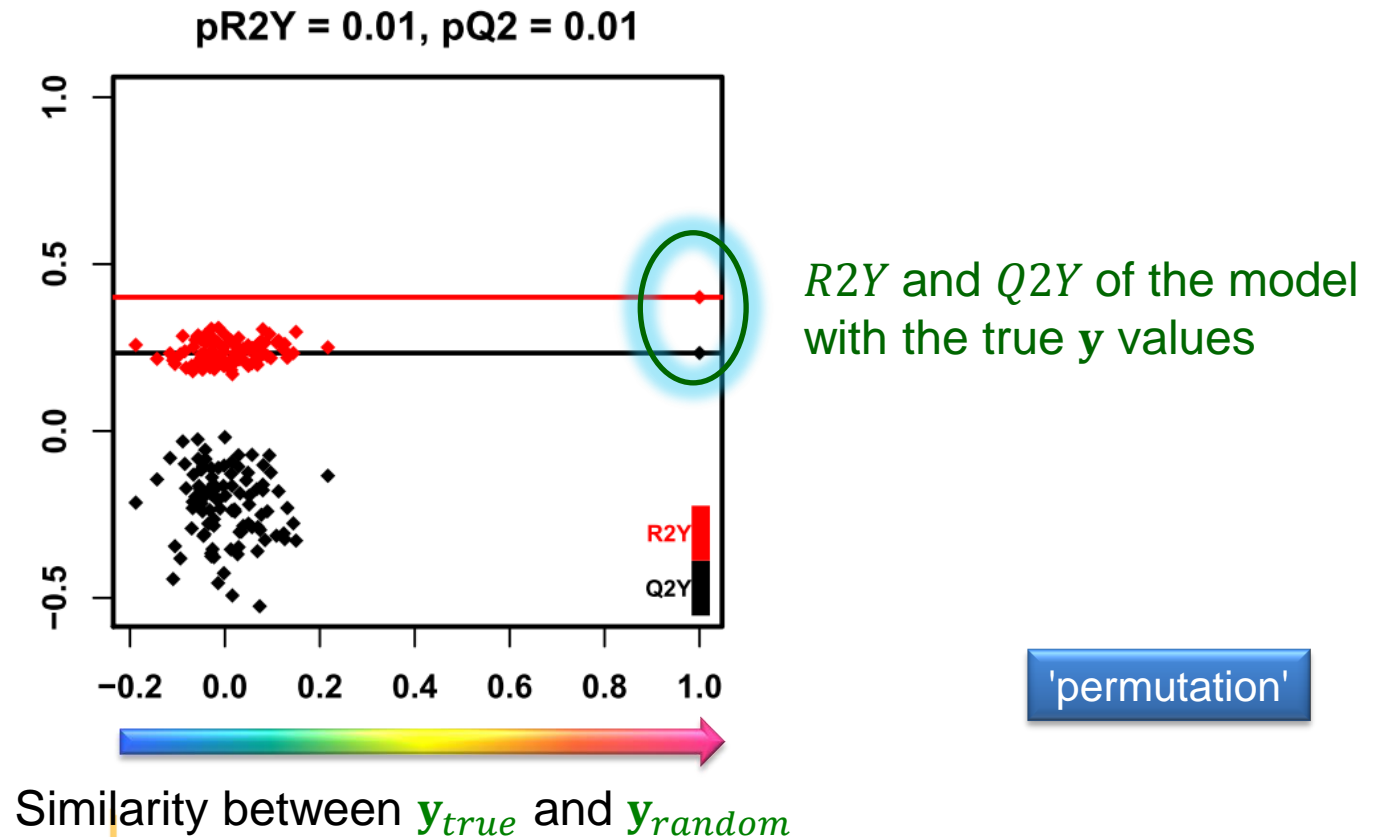
...

**Y** random

**X**

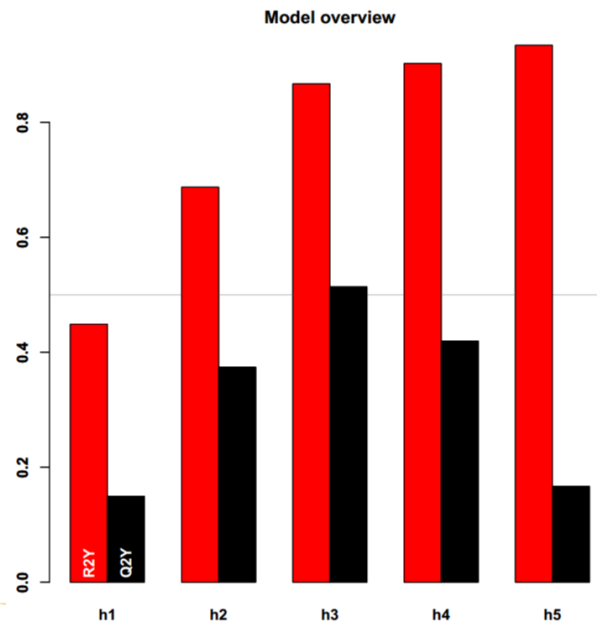
# Significance of the model

- Counting the number of  $R2Y$  (and  $Q2Y$ ) metrics from random models which are superior to the values of the true model gives an indication of the significance of the PLS modelling



# Diagnostic metrics

- $0 \leq R2X \leq 1$ : percentage of X inertia explained by the model
  - $0 \leq R2Y \leq 1$ : percentage of Y inertia explained by the model
  - $0 \leq Q2Y \leq 1$ : estimation of the predictive performance of the model by cross-validation
- 
- $R2X$  and  $R2Y$  increase with the number of components while  $Q2Y$  reaches a maximum (due to overfitting):



'overview'

# Numerical results

- The details of the  $R2X$ ,  $R2Y$ , and  $Q2Y$  values are stored in the "information.txt" file

The screenshot displays the olomics software interface. The main window shows the following text:

```
Y: mean-centering and unit-variance scaling
PLS ('nipals' algorithm)
Number of predictive components: 3
Number of reference observations: 183 (100%)
Correlations between variables and components:
```

	h1	h2	cor_h1	cor_h2
alpha-N-Phenylacetyl-glutamine	-0.220	NA	-0.64	NA
Phe-Tyr-Asp (and isomers)	-0.220	NA	-0.63	NA
Glucuronic acid and/or isomers	-0.220	NA	-0.62	NA
Asp-Leu/Ile isomer 1	0.080	NA	0.22	NA
6-(carboxymethoxy)-hexanoic acid	0.097	NA	0.27	NA
Testosterone glucuronide	0.180	NA	0.50	NA
Acetaminophen glucuronide	NA	-0.093	NA	-0.240
p-Anisic acid	NA	-0.066	NA	-0.180
Malic acid	NA	-0.030	NA	-0.078
p-Hydroxymandelic acid	NA	0.200	NA	0.520
1-Methyluric acid	NA	0.200	NA	0.530
Porphobilinogen	NA	0.200	NA	0.530

Model overview:

	R2X	R2X(cum)	R2Y	R2Y(cum)	Q2	Q2(cum)	Signif.	Iter.
h1	0.0984	0.0984	0.4791	0.479	0.401	0.401	R1	1
h2	0.0861	0.1846	0.1892	0.668	0.256	0.555	R1	1
h3	0.0907	0.2752	0.0615	0.730	0.065	0.584	R1	1

Model summary:

	R2X(cum)	R2Y(cum)	Q2(cum)	RMSEE	ncp	nco
h3	0.275	0.73	0.584	0.262	3	0

The right-hand side of the interface shows a 'History' panel with a list of datasets. A green box highlights the entry '39: Multivariate\_information.txt', with a green '1' in a box next to its eye icon, indicating it is the selected dataset.

# Scores, loadings and VIPs

- The score (resp. loading and VIPs) of the selected components have been added as columns in the **sampleMetadata** (resp. **variableMetadata**) files

The screenshot displays the olomics software interface. The main window shows a table with the following columns: msiLevel, hmdb, chemicalClass, gender\_PLSDA\_XLOAD-h1, gender\_PLSDA\_XLOAD-h2, and gender\_PLSDA\_VIP. The table contains 30 rows of data. A green box highlights the last three columns. On the right side, there is a History panel showing a list of datasets with their sizes and actions. Two green boxes with numbers 1 and 2 are overlaid on the History panel, pointing to specific entries.

msiLevel	hmdb	chemicalClass	gender_PLSDA_XLOAD-h1	gender_PLSDA_XLOAD-h2	gender_PLSDA_VIP
2		Organi	-0.0398502158539864	-0.0118906818365882	0.413402576648655
2		AA-pep	0.045506179215717	0.189853829891156	1.48654320826344
1	HMDB03099	AroHeP:Xenobi	-0.0892685224945862	0.200473082255006	0.994358885831879
1	HMDB10738	AroHeP	-0.0925960283984577	0.166237293630931	0.909198577023911
1	HMDB01857	AroHeP	-0.0533869298019096	0.166793890177945	0.703482789417141
1	HMDB11103	AroHeP	-0.105555888603966	0.129654344183481	0.68032554007513
2		AroHoM	-0.139031345364493	0.0256580978838288	0.930587981757499
1	HMDB00510	AA-pep	-0.123797451802098	0.122573314497015	0.901219803935142
1	HMDB59709	AroHoM	-0.0859289153376191	0.080533734055351	0.550144194269479
1	HMDB00402	Organi	-0.00500169475467362	0.164041655306413	1.1135503438424
1	HMDB11723	AA-pep:AcyGly	-0.146406017195434	0.00205394318915884	1.15042106154043
1		Lipids	-0.00866480699319381	0.117644113800042	0.543551532664842
1	HMDB59712	AroHoM	-0.0550063618628605	0.0437260467146582	0.65956729426584
1	HMDB00440	AroHoM	-0.0910480750747919	0.0263696450305611	0.594653447171177
2	HMDB13189	Carboh	-0.00243590621997017	0.0588028800259373	0.747217045999356
1	HMDB00491	Lipids	0.0464961899862177	0.112804940847864	0.820925594575721
1	HMDB00459	AA-pep:AcyGly	-0.128640803025914	0.0765010378278105	0.879948860811061
1	HMDB02441	Lipids	-0.0572183256960898	0.113224239823584	0.495244006648848
1	HMDB01336	AroHoM	-0.0760295060324308	0.0379713701648879	0.754733936526486
2		AroHoM	-0.137003034145239	0.0383124974603868	1.0070259405318
1	HMDB01982	AroHeP	-0.0287380299762852	0.179401841616721	0.797138454685613
2		Lipids	-0.043696294430725	0.18755264988441	0.737596864407318

History panel entries (from top to bottom):

- 39: Multivariate in formation.txt
- 38: Multivariate figure.pdf (highlighted with '2')
- 37: Multivariate variableMetadata.tsv (highlighted with '1')
- 36: Multivariate sampleMetadata.tsv
- 35: Multivariate dataMatrix.tsv
- 34: Multivariate in formation.txt
- 33: Multivariate figure.pdf



# PLS-DA

- The two response levels are encoded as numbers

## Qualitative

## Quantitative

## Quantitative

## Qualitative

*n* = 20 samples

	gender
H011	M
H023	M
H033	F
H042	M
H052	F
H062	M
H073	M
H083	M
H092	M
H103	M
H114	M
H124	M
H134	M
H145	M
H157	F
H168	F
H180	F
H189	F
H199	M
H209	F



	gender
HU_017	0.5
HU_028	0.5
HU_034	-0.5
HU_051	0.5
HU_060	-0.5
HU_078	0.5
HU_091	0.5
HU_093	0.5
HU_099	0.5
HU_110	0.5
HU_130	0.5
HU_134	0.5
HU_138	0.5
HU_149	0.5
HU_152	-0.5
HU_175	-0.5
HU_178	-0.5
HU_185	-0.5
HU_204	0.5
HU_208	-0.5

PLS



	gender
H011	0.40
H023	0.10
H033	-0.61
H042	0.39
H052	-0.47
H062	0.46
H073	0.36
H083	0.11
H092	0.47
H103	0.23
H114	0.25
H124	0.56
H134	0.12
H145	0.93
H157	-0.19
H168	-0.49
H180	-0.20
H189	0.00
H199	0.54
H209	0.05

	pred
H011	M
H023	M
H033	F
H042	M
H052	F
H062	M
H073	M
H083	M
H092	M
H103	M
H114	M
H124	M
H134	M
H145	M
H157	F
H168	F
H180	F
H189	M
H199	M
H209	M

*y*

*y*

*y*<sub>fitted</sub>

*y*<sub>fitted</sub>

# PLS-DA

- Automatically selected when the response is qualitative (i.e. the column of **sampleMetadata** only contains characters)

**Workflow4Metabolomics** Analyze Data Workflow Shared Data Visualization Help User Using 2%

Multivariate (version 2015-04-25)

**Data matrix file:**  
1: dataMatrix.tsv  
variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular

**Sample metadata file:**  
2: sampleMetadata.tsv  
sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

**Variable metadata file:**  
3: variableMetadata.tsv  
variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

**Y Response (for PLS(-DA) and OPLS(-DA) only):**  
gender

Notes: 1) PCA: keep the default (none); 2) PLS(-DA) and OPLS(-DA): indicate the name of the column of the sample table to be modeled

**Number of predictive components:**  
NA  
Notes: 1) PCA and PLS(-DA): NA can be selected to get a suggestion of the optimal number of predictive components; 2) OPLS(-DA) modeling: select 1 predictive component

**Number of orthogonal components (for OPLS(-DA) only):**  
0  
Notes: 1) PCA and PLS(-DA): keep the default value (0); 2) OPLS(-DA): NA can be selected to get a suggestion of the optimal

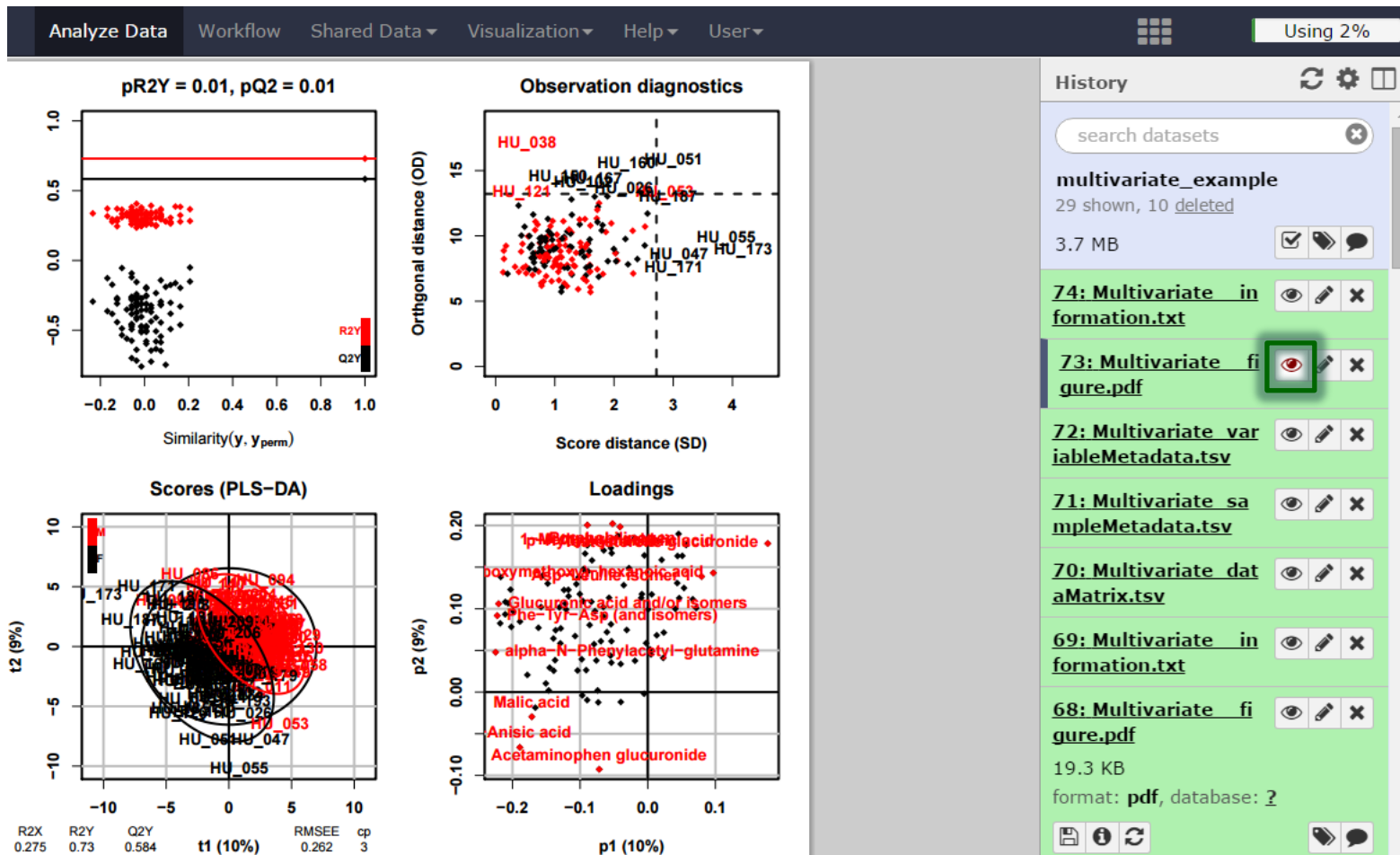
**History**  
search datasets  
multivariate\_example  
29 shown, 10 deleted  
3.7 MB  
74: Multivariate\_in\_formation.txt  
73: Multivariate\_fi\_gure.pdf  
72: Multivariate\_var\_iableMetadata.tsv  
71: Multivariate\_sa\_mpleMetadata.tsv  
70: Multivariate\_dat\_aMatrix.tsv  
69: Multivariate\_in\_formation.txt  
68: Multivariate\_fi\_gure.pdf  
19.3 KB



Warning: Use balanced datasets (similar proportions of samples in each of the two classes)



# PLS-DA



**ORTHOGONAL PARTIAL LEAST SQUARES  
REGRESSION (OPLS)  
AND DISCRIMINANT ANALYSIS (OPLS-DA)**



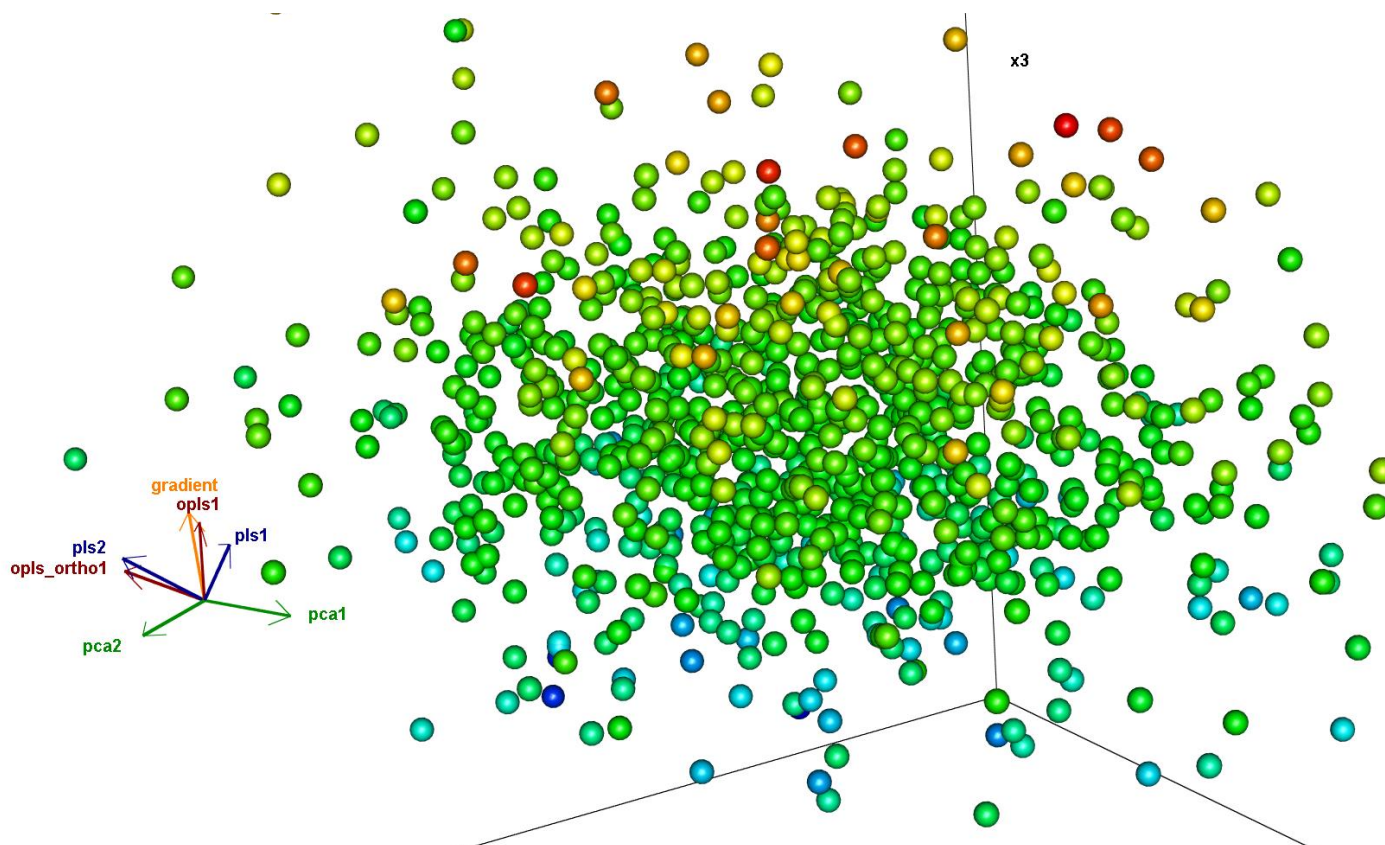
# Principles

---

- Separately models the variations of the predictors correlated and orthogonal to the response
- Improves the interpretation of the components but not the overall predictive performance of the model
- Only one predictive component required for single response models
- Note: As with PLS, care should be taken to avoid too many (orthogonal) components (which would result in overfitting)

# OPLS vs PLS

- Variation not correlated to the response (e.g., technical bias) is modelled separately by the orthogonal component(s)
- => The first predictive component is strongly correlated to the response



# Selection of OPLS(-DA) as the type of analysis

- Set the number of predictive component to 1
- Select the number of orthogonal components (e.g., NA)

The screenshot shows the Galaxy web interface for the 'Multivariate' tool (version 2015-04-25). The configuration is as follows:

- Data matrix file:** 1: dataMatrix.tsv (variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular)
- Sample metadata file:** 2: sampleMetadata.tsv (sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular)
- Variable metadata file:** 3: variableMetadata.tsv (variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular)
- Y Response (for PLS(-DA) and OPLS(-DA) only):** bmi
- Number of predictive components:** 1
- Number of orthogonal components (for OPLS(-DA) only):** NA
- Advanced graphical parameters:** Use default
- Advanced computational parameters:** Use default

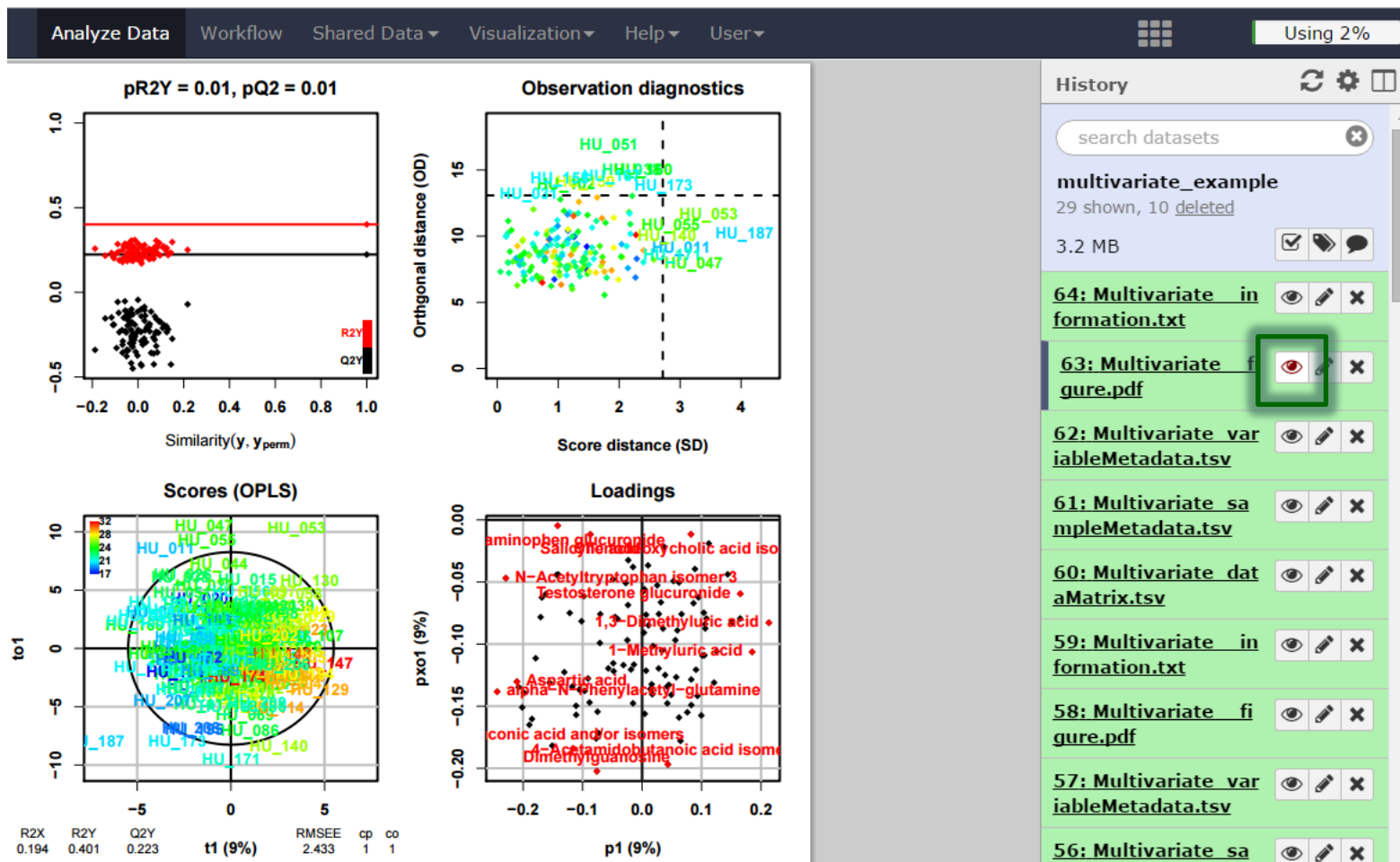
Notes for the component settings:

- Notes: 1) PCA and PLS(-DA): NA can be selected to get a suggestion of the optimal number of predictive components; 2) OPLS(-DA): NA can be selected to get a suggestion of the optimal number of orthogonal components
- Notes: 1) PCA and PLS(-DA): keep the default value (0); 2) OPLS(-DA): NA can be selected to get a suggestion of the optimal number of orthogonal components

The right sidebar shows a history of datasets, including '59: Multivariate\_in\_formation.txt', '58: Multivariate\_fi\_gure.pdf', '57: Multivariate\_var\_iableMetadata.tsv', '56: Multivariate\_sa\_mpleMetadata.tsv', '55: Multivariate\_dat\_aMatrix.tsv', '54: Multivariate\_in\_formation.txt', '53: Multivariate\_fi\_gure.pdf', '52: Multivariate\_var\_iableMetadata.tsv', and '51: Multivariate\_sa\_mpleMetadata.tsv'.

# Graphical results

- permutation, overview, outlier, and score plots displayed as the default ('summary')



# Numerical results

- The details of the  $R2X$ ,  $R2Y$ , and  $Q2Y$  values are stored in the "information.txt" file

The screenshot displays the Workflow4Metabolomics software interface. The main window shows analysis results for a multivariate model. The top bar includes the 'olomics' logo and navigation menus for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The status bar indicates 'Using 2%' of resources.

Analysis parameters and results:

- Max. : 4.804 Max. : 5.428 Max. : 5.56739
- X: mean-centering and unit-variance scaling
- Number of Y variables: 1
- Y: mean-centering and unit-variance scaling
- OPLS ('nipals' algorithm)
- Number of orthogonal components: 1
- Number of predictive components: 1
- Number of reference observations: 183 (100%)

Correlations between variables and components:

	h1	o1	cor_h1	cor_o1
alpha-N-Phenylacetyl-glutamine	-0.24	NA	-0.54	NA
N-Acetyltryptophan isomer 3	-0.23	NA	-0.51	NA
Aspartic acid	-0.21	NA	-0.47	NA
Testosterone glucuronide	0.17	NA	0.37	NA
1-Methyluric acid	0.19	NA	0.41	NA
1,3-Dimethyluric acid	0.21	NA	0.47	NA
Dimethylguanosine	NA	-0.2000	NA	-0.680
4-Acetamidobutanoic acid isomer 2	NA	-0.2000	NA	-0.660
Gluconic acid and/or isomers	NA	-0.1800	NA	-0.610
Salicylic acid	NA	-0.0120	NA	-0.039
Chenodeoxycholic acid isomer	NA	-0.0120	NA	-0.039
Acetaminophen glucuronide	NA	-0.0047	NA	-0.016

Model overview:

	R2X	R2X(cum)	R2Y	R2Y(cum)	Q2	Q2(cum)	Signif.
h1	0.0930	0.0930	0.285	0.285	0.1796	0.1796	R1
rot	-0.0396	0.0534	NA	0.401	NA	0.2232	<NA>
o1	0.1410	0.1410	0.116	0.116	0.0436	0.0436	R1
sum	NA	0.1944	NA	0.401	NA	0.2232	<NA>

Model summary:

	R2X(cum)	R2Y(cum)	Q2(cum)	RMSEE	ncp	nco
sum	0.194	0.401	0.223	2.43	1	1

The right-hand panel shows the 'History' of datasets. A green box highlights the entry '64: Multivariate information.txt', which is the file mentioned in the text. A 'View data' tooltip is visible over this entry. Other entries in the history include '63: Multivariate gure.pdf', '62: Multivariate variableMetadata.tsv', '61: Multivariate sampleMetadata.tsv', '60: Multivariate dataMatrix.tsv', '59: Multivariate information.txt', '58: Multivariate gure.pdf', '57: Multivariate variableMetadata.tsv', '56: Multivariate sampleMetadata.tsv', and '55: Multivariate dataMatrix.tsv'.



# References

---

- Wold S., Sjöström M. and Eriksson L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**:109-130.  
[http://dx.doi.org/10.1016/S0169-7439\(01\)00155-1](http://dx.doi.org/10.1016/S0169-7439(01)00155-1)
- Trygg J., Holmes E. and Lundstedt T. (2007). Chemometrics in Metabonomics. *Journal of Proteome Research*, **6**:469-479.  
<http://dx.doi.org/10.1021/pr060594q>
- Brereton R.G. and Lloyd G.R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, **28**:213-225.  
<http://dx.doi.org/10.1002/cem.2609>