Data analysis in metabolomics

Jasper Engel (jasper.engel@wur.nl)

WAGENINGEN
UNIVERSITY & RESEARCH

Learning objectives

- The understand the **importance of** (multivariate) **statistics** for data driven research

- To obtain an overview of the most commonly used methods for **univariate and multivariate data analysis** in (untargeted) metabolomics

- To understand the **limitations** of applying univariate and multivariate techniques to metabolomics data sets.

Outline

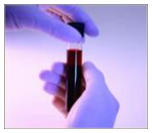
- Introduction
  - What is hypothesis generating research
  - Structure of metabolomics data
  - Univariate vs multivariate analysis
- Exploratory data analysis
  - Principal component analysis
  - Examples
- Discriminant analysis
  - The curse of dimensionality
  - Partial least squares – discriminant analysis
  - Model selection and validation
  - Examples

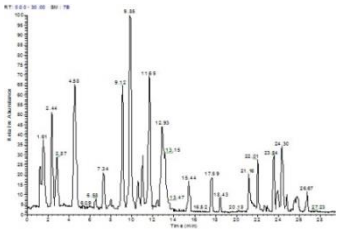
The metabolomics pipeline

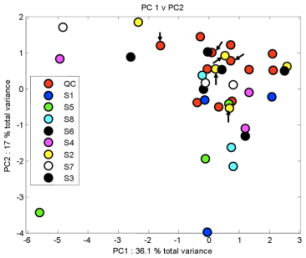
Raw data processing

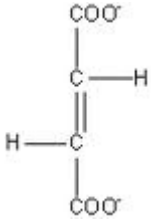
Univariate/ Multivariate statistics

SAMPLE → **METABOLIC PROFILE** → **DATA ANALYSIS** → **METABOLITE IDENTIFICATION** → **BIOLOGICAL INTERPRETATION**









WAGENINGEN
UNIVERSITY & RESEARCH

Data driven research

Data driven research

Hypothesis driven

Hypothesis

1

Data driven research

Hypothesis driven

Hypothesis

1

measurements

36	144	72	108
216	864	252	288
576	2304	432	648
1286	3024	1152	1458
3456	5184	2592	3168

Data Analysis

WAGENINGEN
UNIVERSITY & RESEARCH

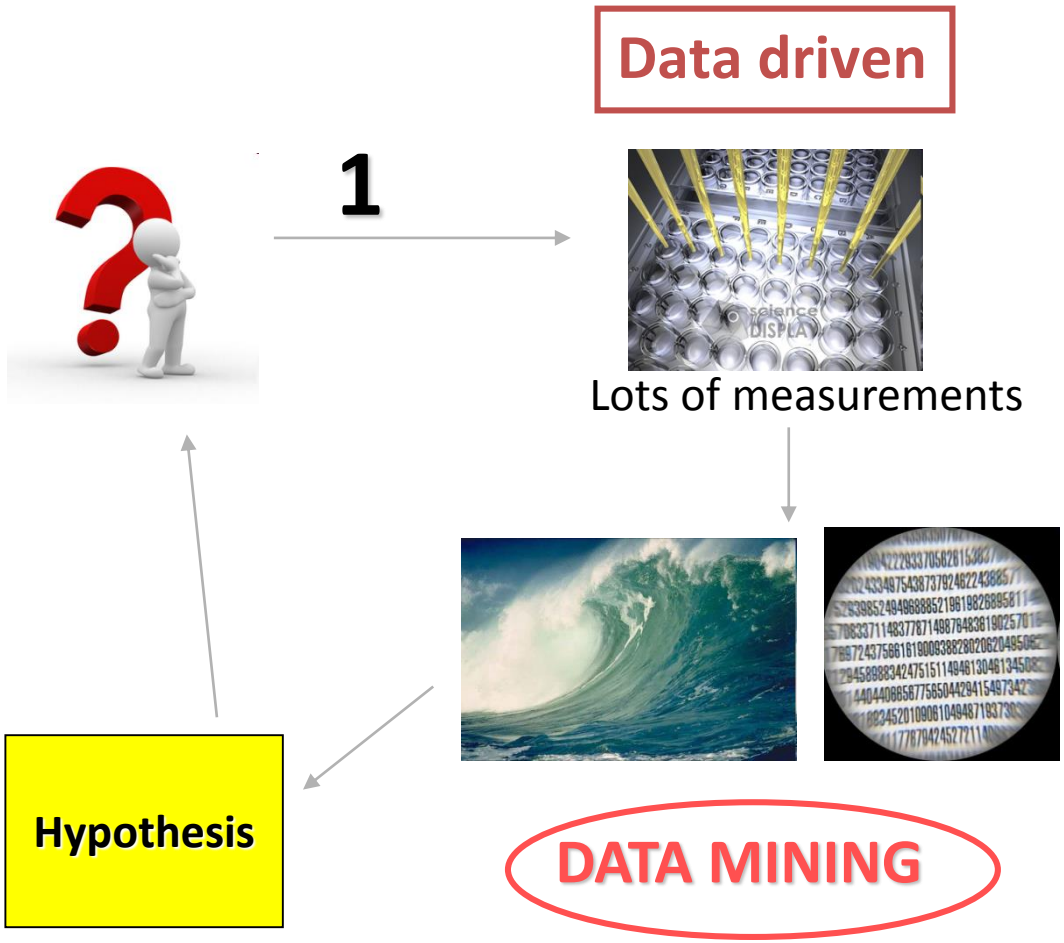
Data driven research

Data driven

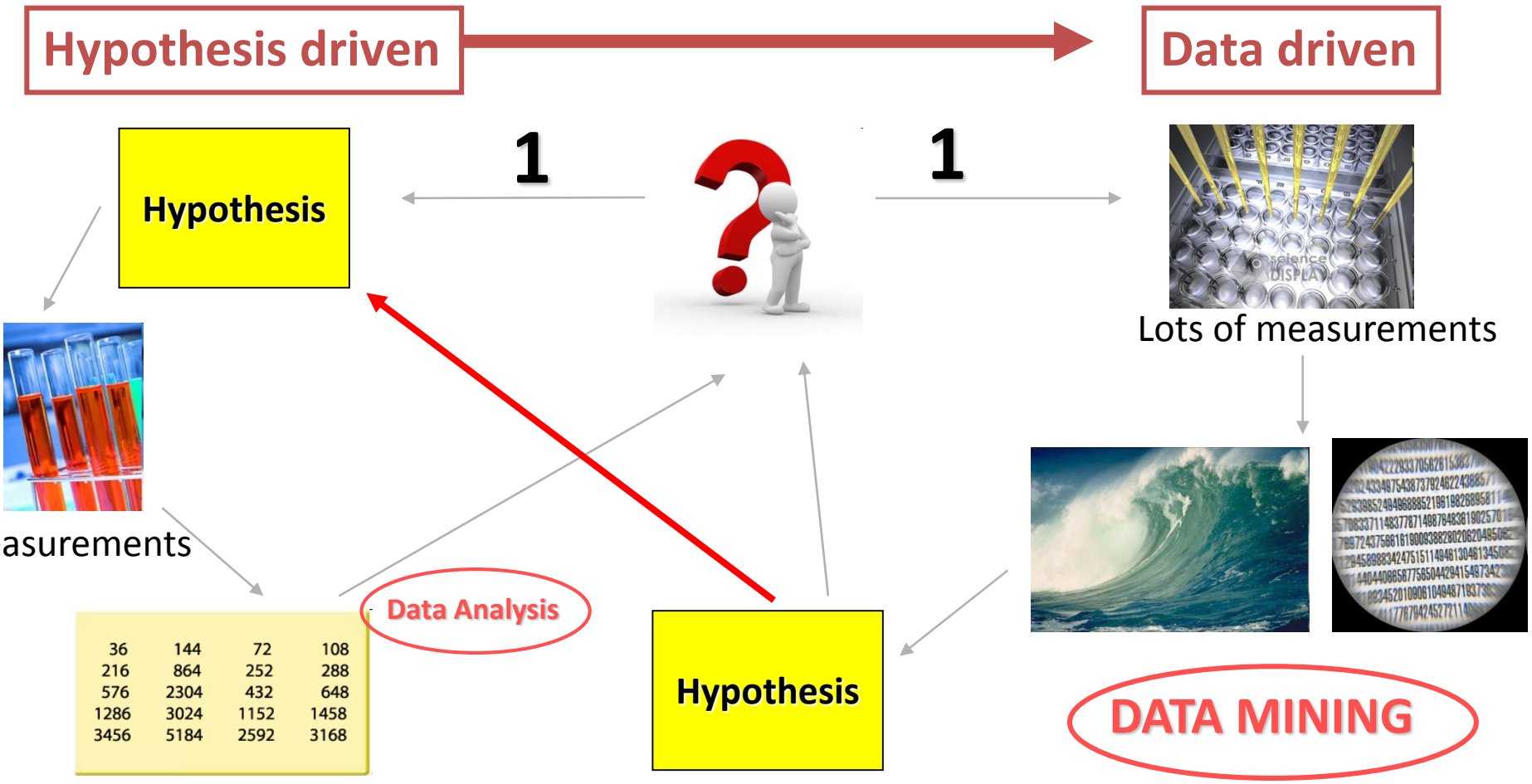
1

Lots of measurements

Data driven research

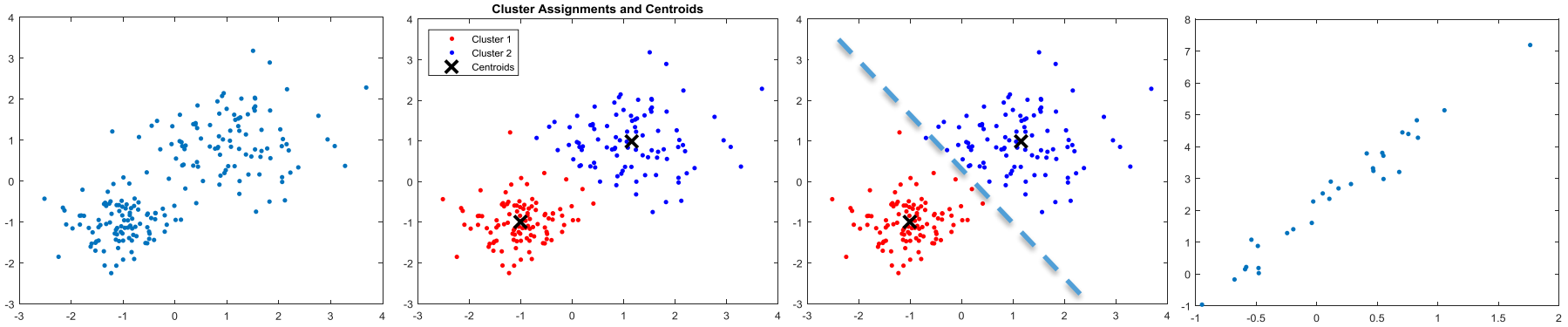


Data driven research



Paradigm change

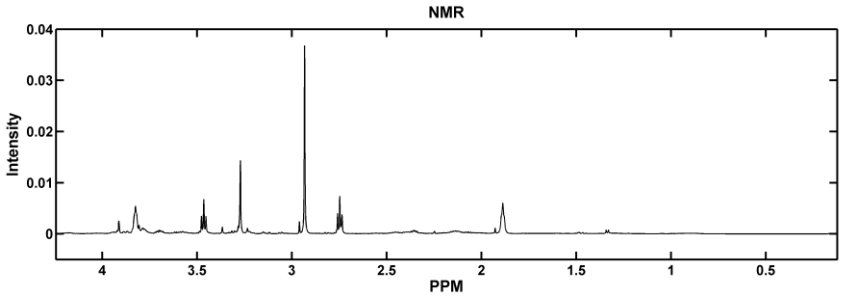
Data analysis objectives

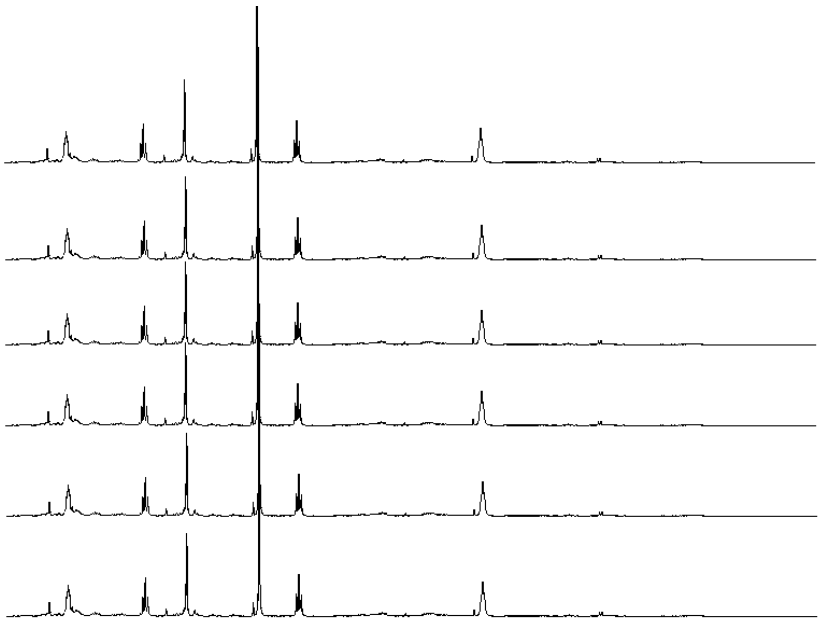


Overview	Clustering	Discriminant analysis / Classification	Regression
• Trends • Patterns • Clusters • Outliers • Quality assurance • Biological diversity	• Grouping of samples and / or variables • Determining group structure • Identification of subgroups • Biological diversity	• Pattern recognition • Discriminating between groups • Assigning samples to groups • Biomarker candidates	• Predicting continuous response • Comparing blocks of omics data
PCA	HCA, k-means	MANOVA, LDA, PLS-DA, O-PLS-DA	PCR, PLS2, O2-PLS

WAGENINGEN
UNIVERSITY & RESEARCH

Organizing your data: the data matrix





NMR spectrum = vector

	Peak 1	Peak 2	Peak3	Peak 4…	Peak N
Sample 1	19812	432	2309	4501882	5876

NMR spectra = matrix

	Peak 1	Peak 2	Peak3	Peak 4…	Peak N
Sample 1	19812	432	2309	4501882	5876
Sample 2	8994	654	5409	357890	312
Sample 3	15012	1098	3102	1342098	10879
Sample 4	9999	302	4231	1809282	890
Sample N	17531	789	4500	2200192	3456

Data = matrix

sample label		peak 1	peak 2	peak 3	...		X matrix of signal intensities
c1	1	0.0031722	-0.0381444	-0.0090	-0.01395	-0.0062389	0.0340444	5.0
c2	1	0.0038722	0.0460556	0.0065	0.00355	-0.0032389	-0.0037556	8.4
c3	1	-0.0036278	-0.0157444	0.0018	0.00365	-0.0072389	-0.0127556	9.2
c4	1	0.0200722	-0.0175444	0.0134	0.01555	0.0066611	-0.0163556	1.3
c5	1	-0.0004278	0.0175556	0.0052	0.00985	0.0027611	0.0056444	4.9
c6	1	0.0010722	0.0053556	0.0016	0.00125	0.0105611	0.0099444	7.9
indo1	2	-0.0075278	0.0100556	0.0005	-0.00415	-0.0045389	0.0225444	16.6
indo2	2	-0.0000278	-0.0231444	-0.0018	-0.01515	0.0046611	0.0047444	18.2
indo3	2	-0.0017278	0.0094556	0.0024	-0.00175	0.0021611	-0.0293556	14.4
indo4	2	-0.0003278	-0.0255444	-0.0054	-0.00165	-0.0015389	0.0089444	10.0
indo5	2	-0.0017278	-0.0212444	-0.0083	-0.00115	-0.0065389	-0.0296556	25.9
indo6	2	-0.0002278	-0.0327444	-0.0039	-0.00625	-0.0103389	-0.0192556	21.3
mpa1	3	0.0022722	0.0287556	-0.0012	0.02465	-0.0004389	-0.0073556	103.1
mpa2	3	-0.0002278	0.0363556	0.0010	0.00815	-0.0025389	0.0179444	69.9
mpa3	3	-0.0027278	-0.0068444	0.0013	0.00205	0.0045611	0.0061444	91.3
mpa4	3	-0.0055278	-0.0216444	-0.0011	-0.00815	0.0010611	0.0030444	98.5
mpa5	3	-0.0022278	0.0230556	-0.0025	-0.01055	0.0047611	0.0269444	48.1
mpa6	3	-0.0041278	0.0259556	-0.0005	-0.00595	0.0054611	-0.0214556	59.7

EITHER

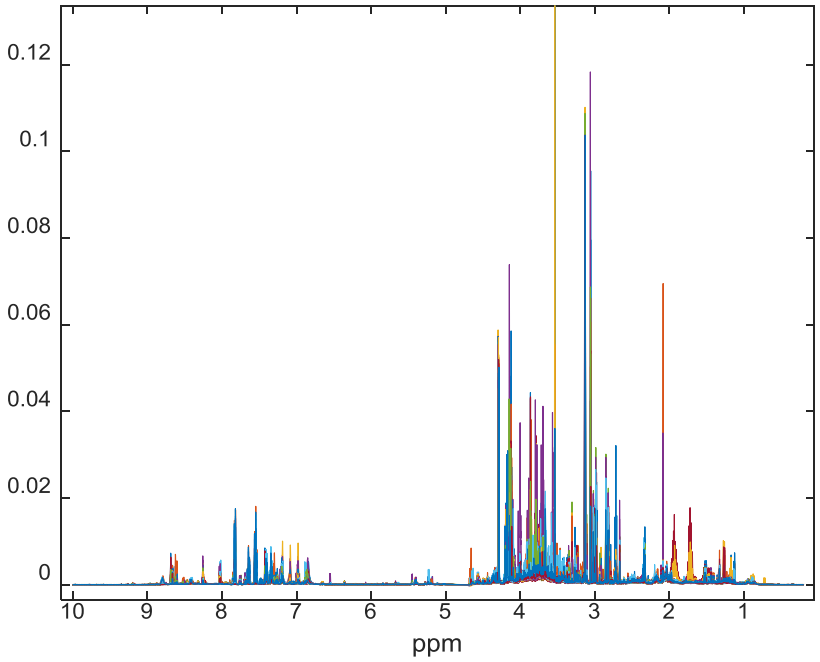
Y matrix = treatment group labels = discrete variable

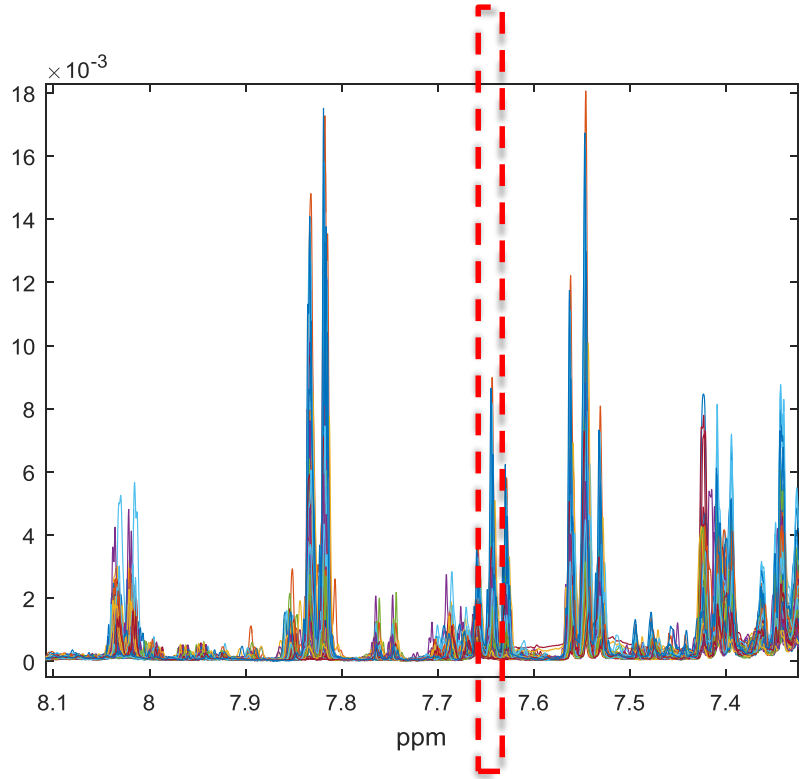
OR

Y matrix = separate non-metabolic measurement for each sample = continuous variable

Plotting the data

- Line plot of each spectrum

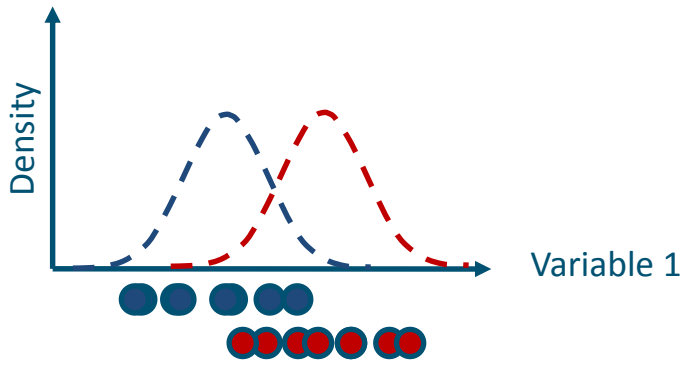




Plotting the data

- One variable → density plot

sample label		peak 1
c1	1	0.0031722
c2	1	0.0038722
c3	1	-0.0036278
c4	1	0.0200722
c5	1	-0.0004278
c6	1	0.0010722
indo1	2	-0.0075278
indo2	2	-0.0000278
indo3	2	-0.0017278
indo4	2	-0.0003278
indo5	2	-0.0017278
indo6	2	-0.0002278

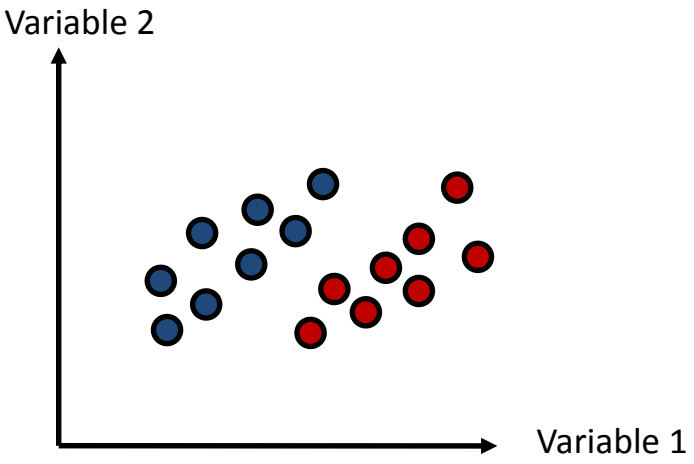


Plotting the data

- Two variables → scatter plot

  = multivariate analysis - variables are plotted against each other (instead of analysing them one at a time)

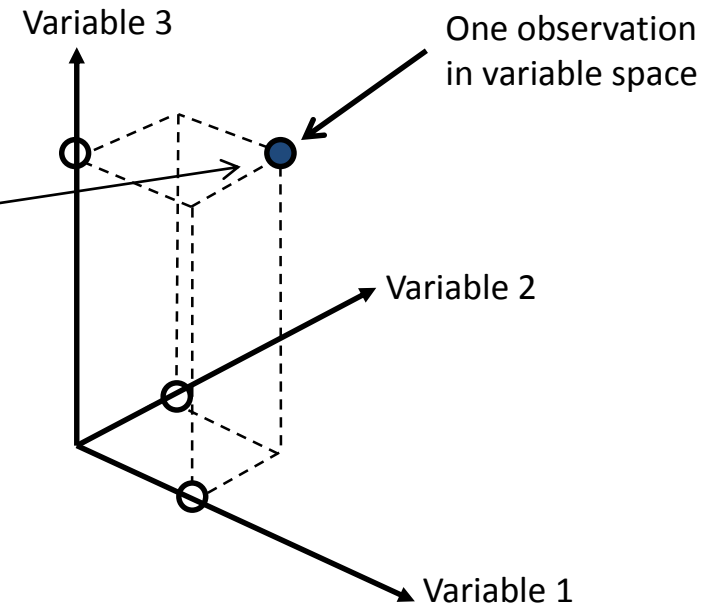
sample label		peak 1	peak 2
c1	1	0.0031722	-0.0381444
c2	1	0.0038722	0.0460556
c3	1	-0.0036278	-0.0157444
c4	1	0.0200722	-0.0175444
c5	1	-0.0004278	0.0175556
c6	1	0.0010722	0.0053556
indo1	2	-0.0075278	0.0100556
indo2	2	-0.0000278	-0.0231444
indo3	2	-0.0017278	0.0094556
indo4	2	-0.0003278	-0.0255444
indo5	2	-0.0017278	-0.0212444
indo6	2	-0.0002278	-0.0327444



Plotting the data

- From data to variable space

	var. 1	var.2	var.3			
1						
2						
3						
N						

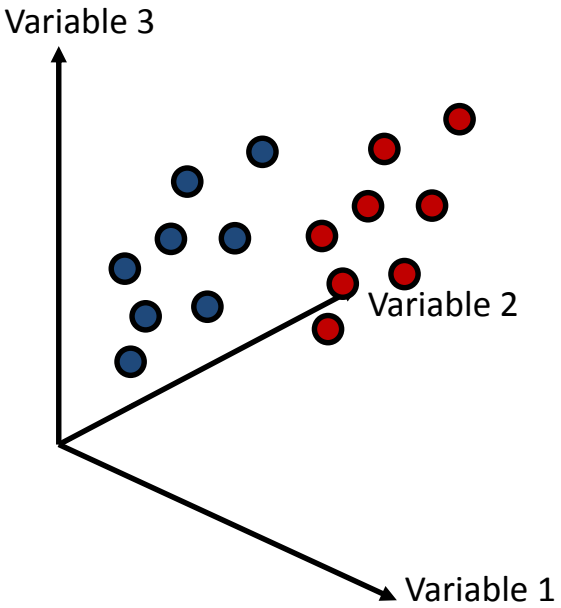


→ The whole data table produces points in variable space

Plotting the data

- From data to variable space

	var. 1	var.2	var.3			
1						
2						
3						
N						



→ The whole data table produces a cloud of points in variable space

Plotting the data

-0.2885	-0.9647	0.4738	1.1464	-2.3682	-0.0366	-1.1781	0.2178	-0.7105	0.7141	-0.4178	0.3614	-0.1499	-0.8445	-0.2342	-0.4725	1.6093	0.1271	-1.1256
-0.5610	-1.0793	-2.0080	2.2270	-0.2614	-0.9431	1.9671	-0.1192	0.1722	-0.5887	-0.7276	-0.8905	-0.2093	-0.7083	-0.2227	0.9231	-1.1733	-0.0802	-0.2159
-1.4942	-0.6843	0.2389	0.1963	-0.7781	-0.9703	-0.2140	0.6728	0.1532	0.4437	0.0850	0.5938	-0.0752	-0.1194	-0.8047	-1.7806	-1.7275	1.0613	-0.1515
1.1784	1.5411	-0.7294	-0.7056	0.5002	-0.3395	1.3686	0.8845	0.0523	-1.8751	-0.3918	1.0445	0.9909	0.1551	-0.6279	-1.4079	-1.2117	1.0801	0.2836
0.3946	0.1079	0.1165	-1.0427	1.2985	0.5630	0.5186	1.5534	0.6885	0.5043	0.9276	2.0917	1.6536	0.4336	-1.2396	0.1013	-0.2833	-0.3431	0.1347
-0.1976	0.9616	-0.4758	-0.0638	1.1760	-0.8341	-0.6553	0.4452	-0.7934	0.9440	-1.0293	0.1492	0.2699	-0.4304	0.7197	0.1376	0.9769	0.7718	-1.1217
0.6040	2.8095	0.1100	-0.1883	-1.2887	-0.6209	0.4519	0.9049	-0.4528	-1.1361	-1.6332	-0.2841	1.1272	0.3492	0.0246	-1.4865	0.6840	0.2739	-0.0961
2.5493	-0.4846	-1.6971	-0.6448	-0.7020	0.8858	0.6544	-2.7280	-0.5840	0.3451	1.4301	0.3992	1.3865	1.3772	-0.1542	0.6965	0.1148	-0.0424	0.6775
-0.6052	-0.9043	-0.1689	-0.4319	-1.0758	0.8917	0.1586	-0.1266	2.3342	-0.2773	2.1286	1.0292	1.8895	0.0528	0.0998	0.1698	0.1635	0.9269	0.9691
-0.0589	-0.9505	2.3193	-0.6232	1.0005	-0.3255	-1.8818	-1.7121	1.3795	1.5689	0.5213	-0.2757	1.3631	0.1084	0.2399	0.7997	-0.0605	0.3391	0.7002
-0.5811	-0.4474	1.0280	1.9958	-0.3134	-1.1572	-1.3709	1.1157	-0.3673	-0.7110	-0.5376	-1.2230	-0.6961	0.5216	-0.3346	-1.2490	0.0567	1.8877	0.2109
-1.7633	-1.8299	-0.5810	-0.2323	-1.3619	0.5400	-0.3119	0.2527	-0.9007	-1.1743	-0.5543	0.4616	-0.3534	-1.0967	0.2371	1.6737	0.6454	-0.5637	-0.5783
1.0087	-1.1588	-0.3893	0.8375	0.4459	-0.3473	-0.1427	-0.1417	-0.3130	-0.4885	-1.0336	0.0051	-1.5608	-0.4521	0.6088	-0.8646	-0.0995	0.1278	-0.9804
-0.0485	1.6557	-0.3071	-0.5564	-0.5211	-0.7456	-1.4932	-0.6925	-0.4768	0.5265	1.5623	-0.1347	-0.9694	0.1013	-0.6225	0.6130	-1.5608	1.5003	1.0153
-0.0056	-1.3776	3.0674	0.5932	-1.9910	1.8513	-2.0737	-1.8845	0.4226	-0.0090	0.2027	-1.2119	0.1443	-0.4798	0.8493	1.4771	-0.1481	-1.5274	-0.9529
0.1194	-1.6126	-0.4413	0.4836	0.1971	-0.8301	1.3888	0.1879	-0.7880	-0.1869	-2.1036	-1.7946	-0.6797	-0.5991	0.0528	-1.3158	-0.3805	0.5086	-0.8773
-0.9913	-0.0243	-1.6194	-0.2357	-0.0447	1.1622	-0.4479	0.8727	-1.0069	-2.1671	0.5657	0.6848	-0.8822	1.1720	0.0487	0.7252	-2.3069	-0.1330	-1.5332
0.2709	-1.7022	1.1874	0.2449	1.1109	1.2171	-0.3153	1.3442	-0.7777	0.8801	-1.4131	0.1835	1.3985	1.1424	0.7631	0.7409	2.4692	0.1878	-1.8953
-0.4269	0.0123	0.6787	0.1867	-0.6294	2.5526	0.0280	-0.0352	0.1835	-0.1280	0.6396	0.8420	-0.9023	0.4523	-0.1185	-0.6120	2.2298	0.6144	-0.3758
0.2492	0.4564	-0.1904	-1.0407	-1.2573	0.4603	0.8435	1.4324	0.5518	0.4677	-0.0141	-0.9001	1.4457	1.9747	1.1133	-0.1303	-0.3598	0.2244	0.5924
1.3100	1.3721	-1.6052	1.4344	-0.7130	-0.0899	-1.0514	0.4683	0.4715	0.3075	0.4735	-1.2072	-0.8856	1.5405	-0.0154	0.5488	1.8637	0.4059	-1.0521
0.5765	0.5671	0.9783	-0.5712	0.3337	-0.1446	0.0232	0.7426	0.1217	-1.8051	0.6570	0.6991	0.0994	0.0277	0.5778	0.5584	-1.1536	0.5259	-0.0541
0.4547	-0.1145	0.3483	1.3628	-0.1583	-1.5534	0.0719	0.0477	-0.7641	0.9581	-1.4712	-0.9815	0.1303	-0.1226	0.3092	-0.9176	-0.6105	-0.7388	0.7452
-1.2930	-0.2669	-0.0614	0.0943	-1.5473	1.1314	1.2724	1.2443	0.8829	0.4562	-1.0775	-0.8772	-0.4306	0.4269	-0.2727	0.2003	-1.6804	0.9697	0.1373
-1.3446	-1.8748	0.2917	-0.5912	0.1658	-0.5971	0.0596	1.0416	1.0436	0.5391	1.9350	0.6339	1.2907	0.5800	-0.0093	0.9417	-0.6400	0.4939	-0.2619
0.8749	-0.1700	-0.1880	0.1105	-0.1932	0.5403	0.4567	0.8731	0.4395	1.9887	-0.1237	0.5719	-1.5612	-2.3112	0.5763	-0.2512	-0.2756	1.4516	-2.0350
-0.9361	0.3463	-0.4862	-0.4542	0.7669	-0.3754	-0.6581	0.5948	0.8785	-0.5018	-0.4488	0.5410	-0.0987	-0.4573	-0.2998	-0.8753	1.3124	0.5133	0.7910
1.3917	1.0159	-0.3013	0.0379	-1.2993	1.0743	0.8889	1.3856	0.0533	1.3625	-0.6494	1.4279	-1.4840	0.6523	-0.8602	-1.7108	1.6111	2.0881	-1.3386
0.4477	0.6068	0.7775	0.0389	0.3289	1.5209	0.5583	1.2030	0.9088	-0.3559	0.7998	-0.3140	-0.5088	-0.5612	0.0327	1.0427	-1.5368	0.1487	0.2358
-0.9071	0.2054	-0.3637	-0.5292	0.5552	-0.7828	0.2623	-1.7432	0.8730	-1.5783	-0.9939	-1.6359	0.5656	0.4197	-0.4334	0.7721	0.7937	-0.3888	-0.4699
1.5304	0.8127	-0.2692	-0.8622	-0.3772	-0.8801	-1.8917	-1.3950	-0.3801	0.9426	-1.6171	-0.6060	-0.3529	-0.9126	-1.7918	1.6599	0.4376	0.7255	1.0104
0.3514	1.5774	1.0537	0.4426	0.3832	-0.3897	-2.0147	0.6647	-0.7921	-0.1282	-1.3085	0.1992	0.6171	-0.6923	1.6115	0.8673	-0.6239	1.0649	1.5248
0.1395	0.1824	-0.1771	0.3318	0.2199	-0.8975	-0.0268	-0.5821	0.8900	0.5126	0.7198	1.5835	2.1294	0.4826	0.9421	-0.4409	1.0468	0.4116	1.0894
2.2172	-1.9004	0.0410	-0.4062	-1.0665	-0.9580	-1.6771	0.0582	-0.2689	1.0826	-0.3029	0.3386	0.7284	0.8211	-0.8390	-1.5653	-1.3202	0.4908	0.7149
1.0612	-1.5357	-0.4389	-0.6367	0.4175	-1.2033	-0.3263	0.1524	-1.4511	0.9493	-0.1338	0.0017	0.9627	0.9439	1.0107	0.5749	-0.0070	-0.1533	0.7535

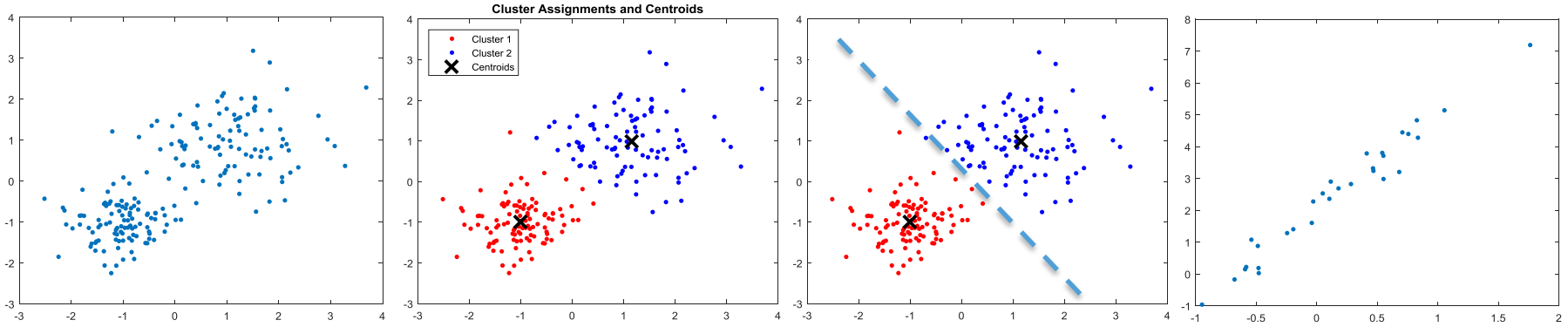
We usually have more than 3 variables

Data analysis in metabolomics

Explorative (unsupervised) analysis

WAGENINGEN
UNIVERSITY & RESEARCH

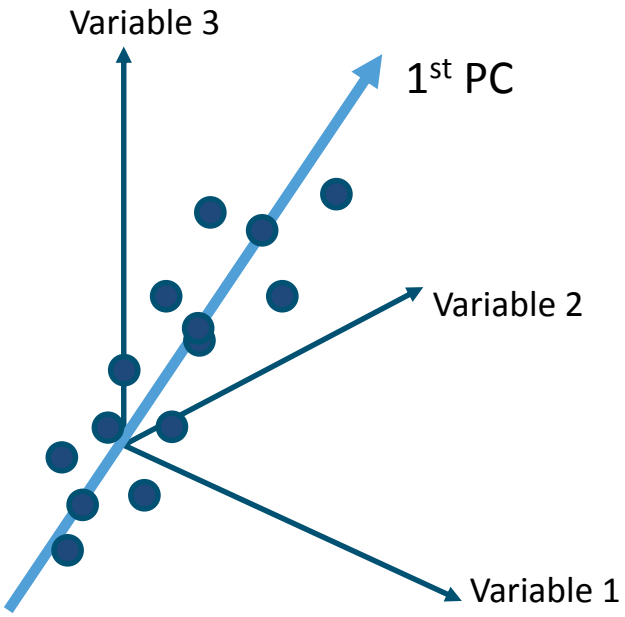
Principal component analysis



Overview	Clustering	Discriminant analysis / Classification	Regression
• Trends • Patterns • Clusters • Outliers • Quality assurance • Biological diversity	• Grouping of samples and / or variables • Determining group structure • Identification of subgroups • Biological diversity	• Pattern recognition • Discriminating between groups • Assigning samples to groups • Biomarker candidates	• Predicting continuous response • Comparing blocks of omics data
PCA	HCA, k-means	MANOVA, LDA, PLS-DA, O-PLS-DA	PCR, PLS2, O2-PLS

WAGENINGEN
UNIVERSITY & RESEARCH

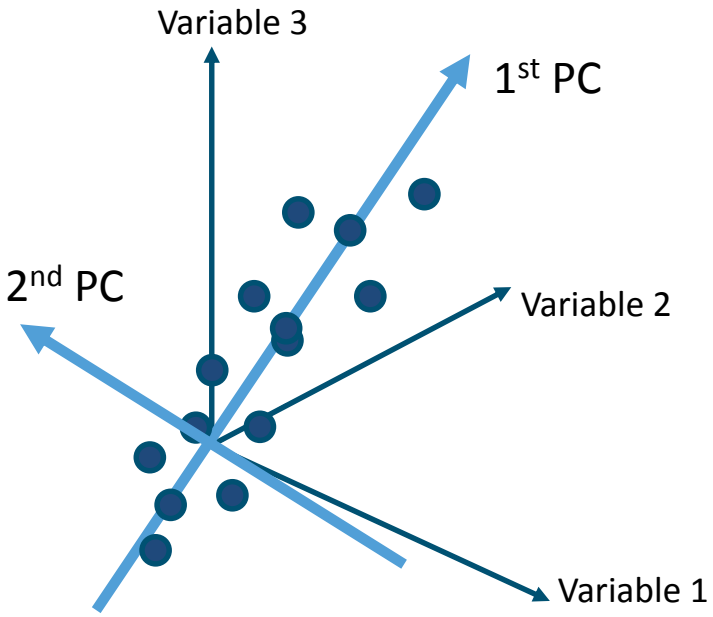
Principal component analysis



The first principal component (PC) describes the **largest amount of variance** in the data.

= direction of largest spread between the data points in variable space

Principal component analysis

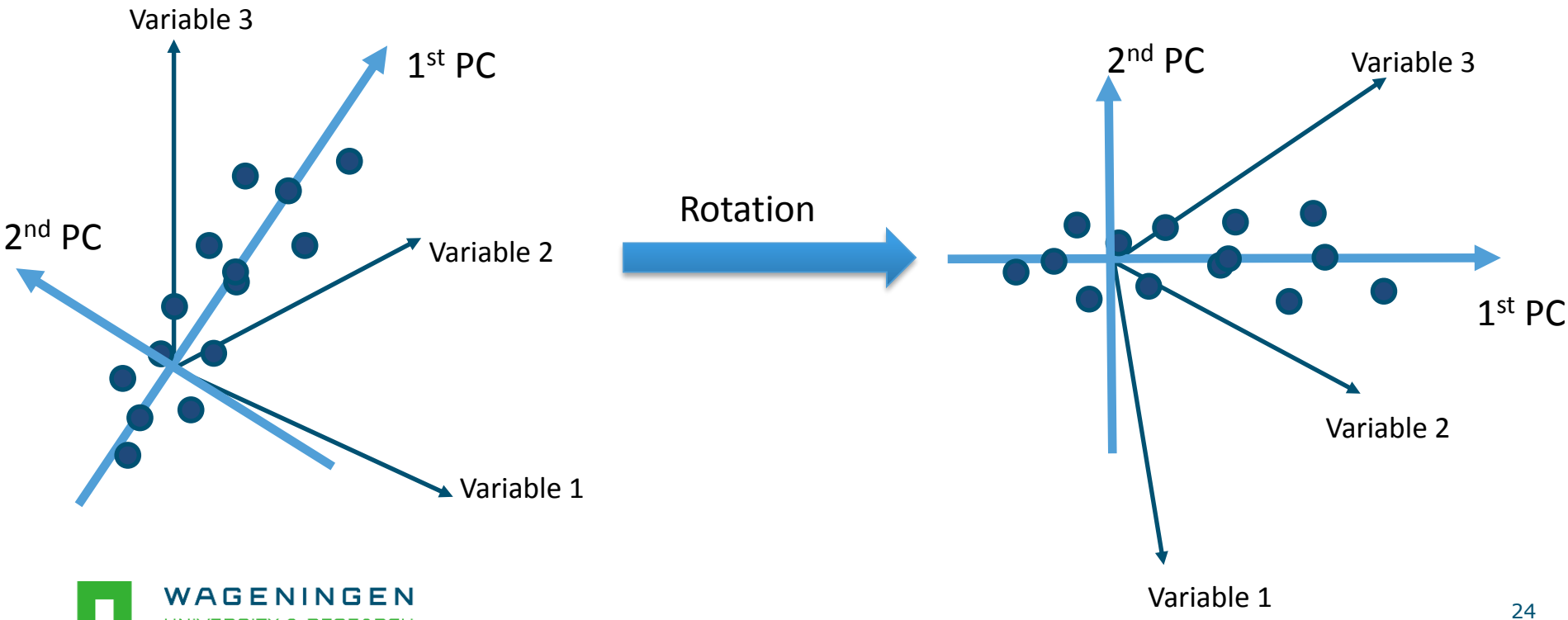


The second principal component (PC) describes the second **largest amount of variance** in the data.

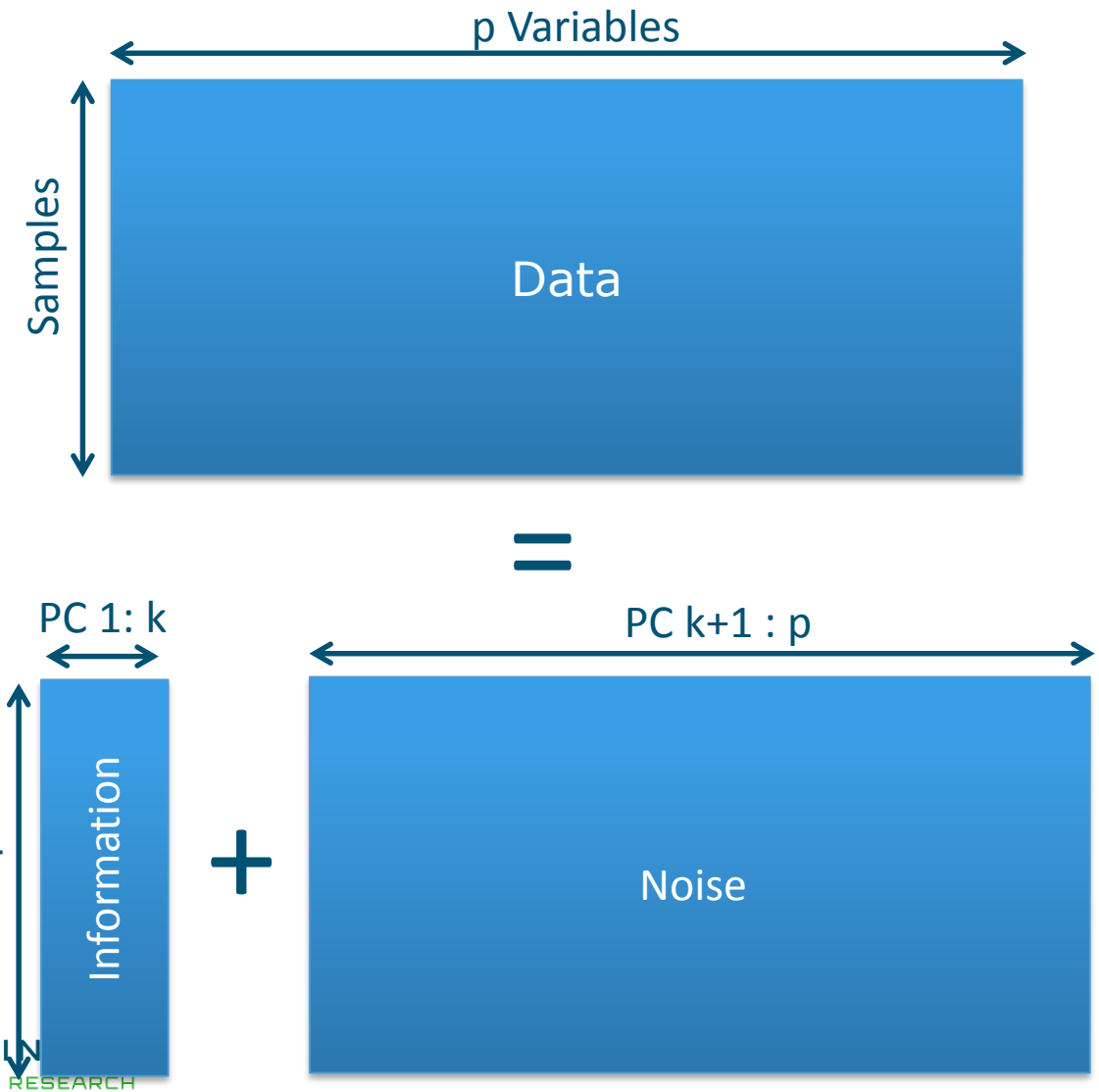
= direction of largest spread between the data points in variable space orthogonal to PC 1.

Principal component analysis

- PCA = rotation of the data such that the first variables (PCs) explain most of the variance



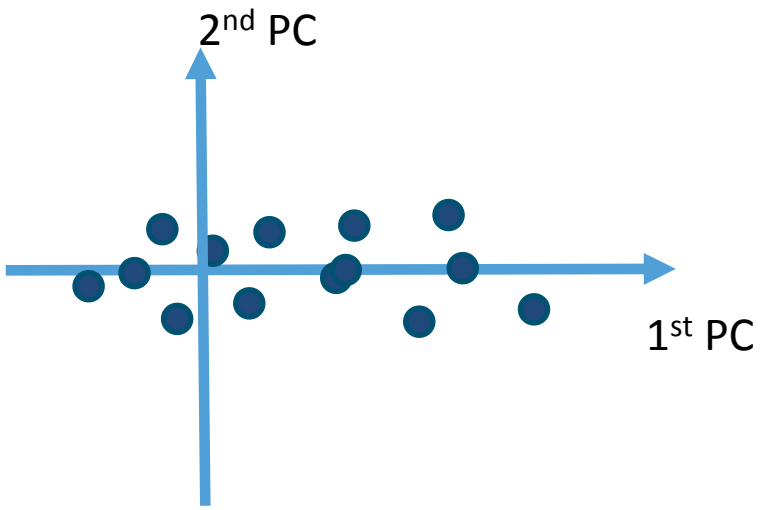
Dimension reduction with PCA



Data visualization with PCA

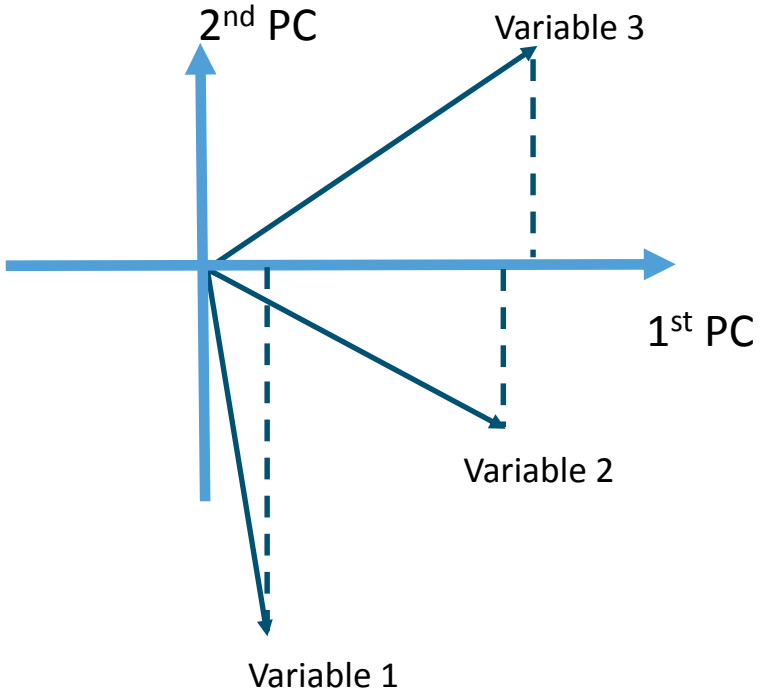
Scores

- Summarize the observations
- Separate signal from noise
- Observe patterns, clusters, etc.



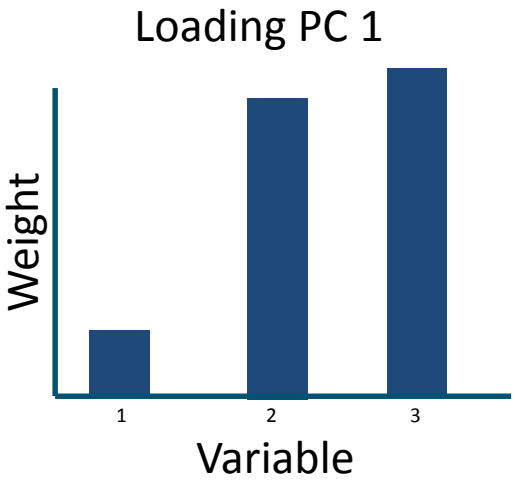
Loadings

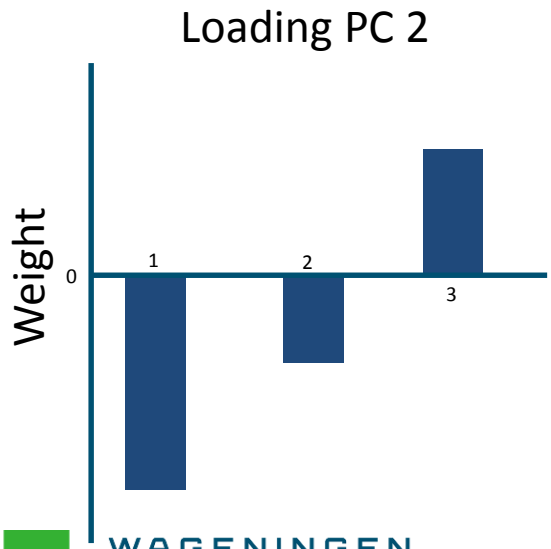
- Summarize the variables
- Explain the position of the observations in the score plot



WAGENINGEN
UNIVERSITY & RESEARCH

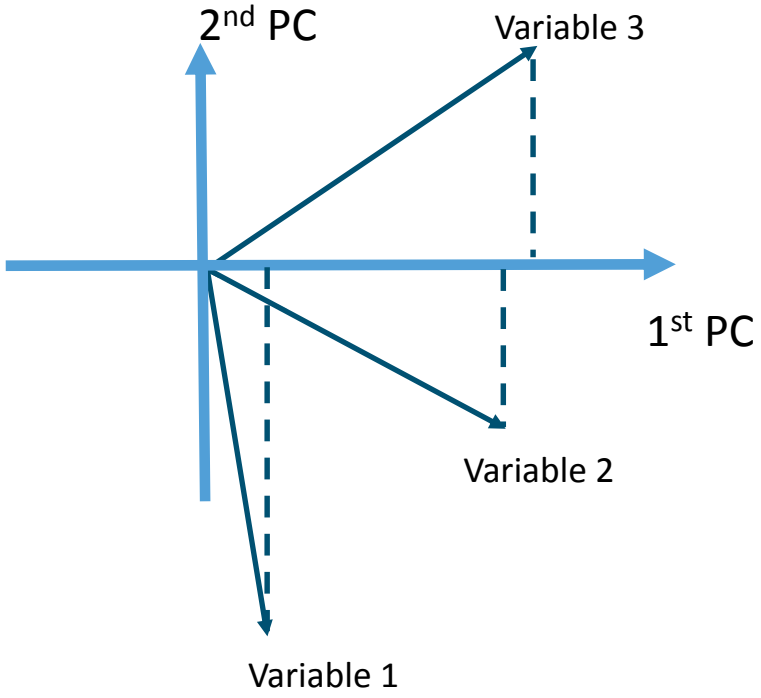
Data visualization with PCA





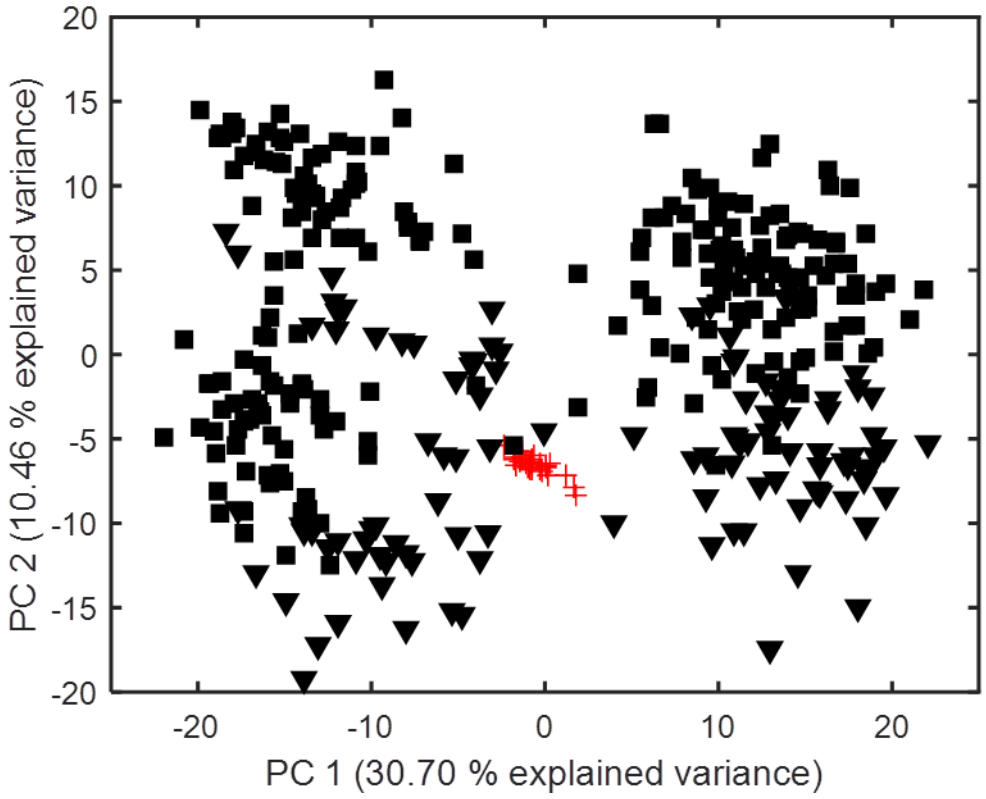
Loadings

- Summarize the variables
- Explain the position of the observations in the score plot



WAGENINGEN
UNIVERSITY & RESEARCH

Example: quality assurance

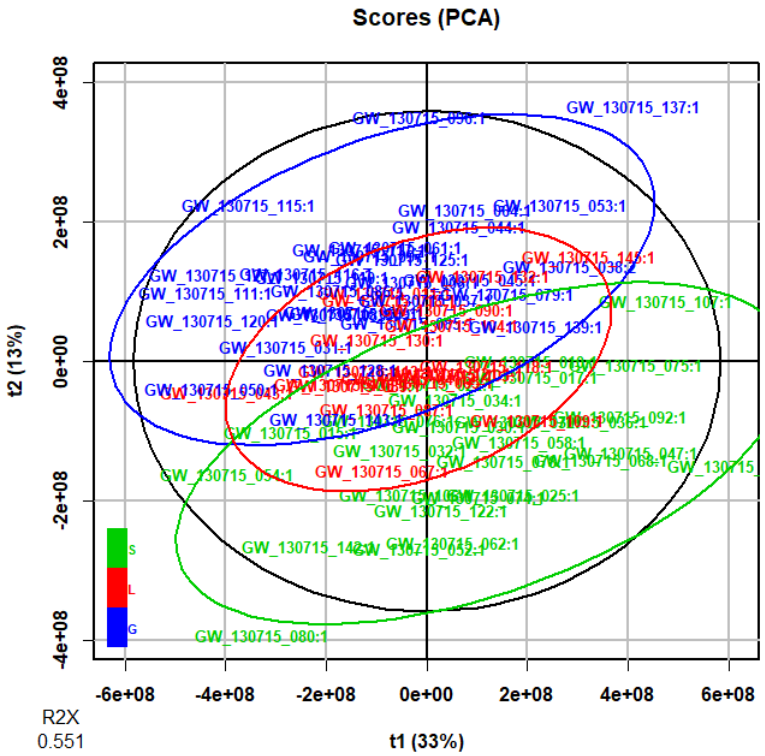


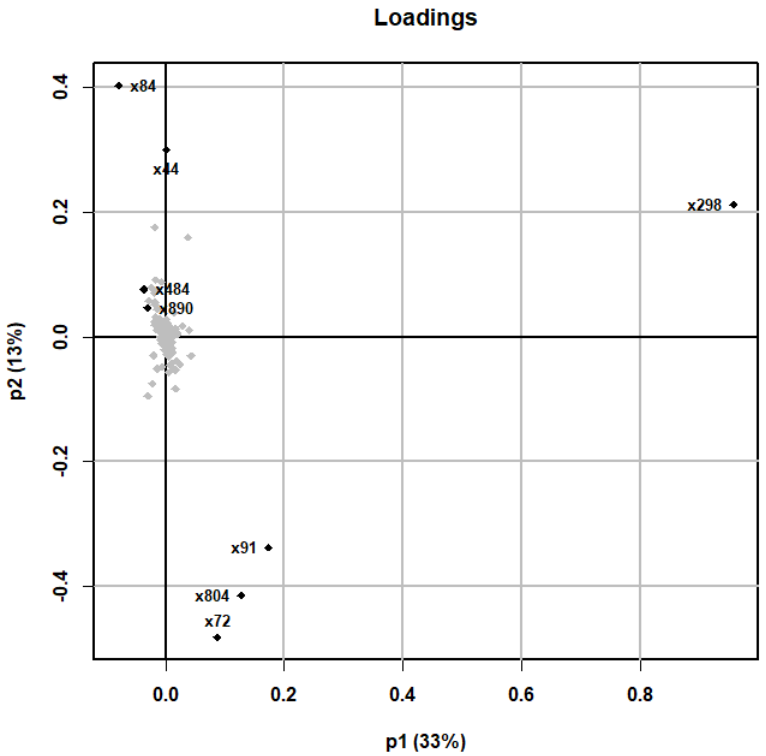
WAGENINGEN
UNIVERSITY & RESEARCH

Viant, Metabolites, 7, 2017

Example: assessment of regional differences in Lambrusco wines

- Wines from 3 regions
- Analysed by GCxGC-MS (76 x 1208)





Intermezzo: common artefacts in metabolomics data

- Baseline drift
- Peak misalignments
- Unwanted peak intensity differences
- Noise variables
- Batch effects
- Missing values
- Unequal peak weights
- Large RSD values
- ...

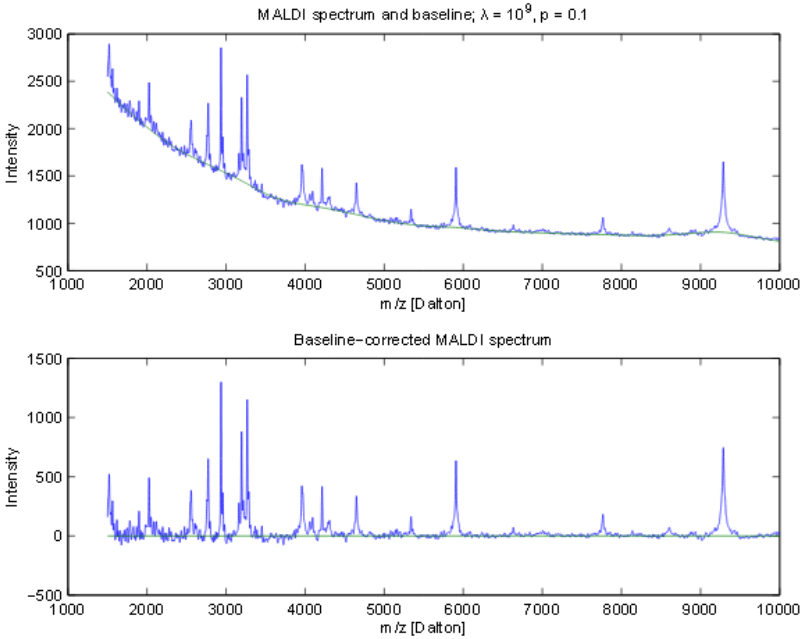
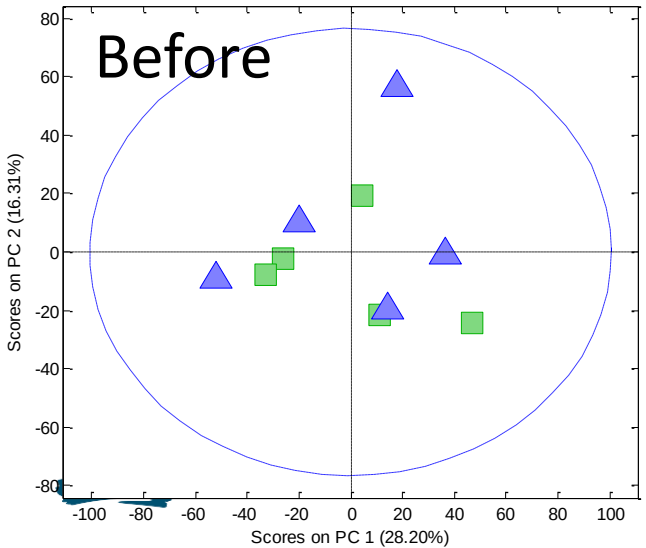


Figure 4: Baseline correction of a MALDI-TOF mass spectrum. Top: spectrum and estimated baseline; bottom: baseline corrected spectrum.

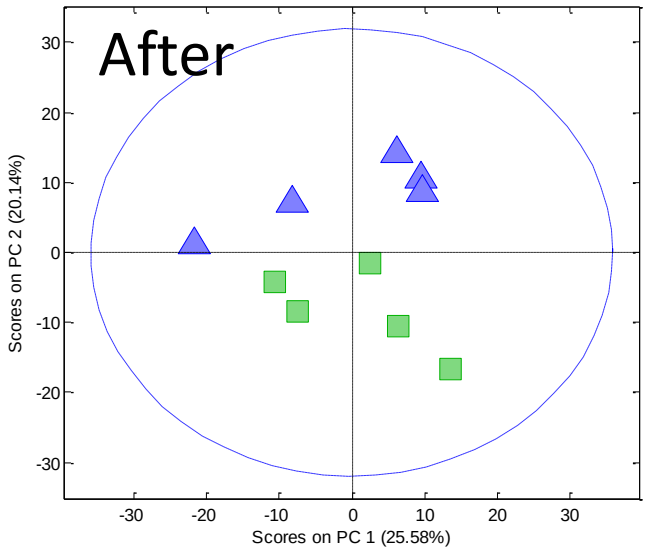
Taken from Eilers et al, Baseline Correction with Asymmetric Least Squares Smoothing (2005)

Intermezzo: why data preprocessing?

- Our analytical tools (NMR/MS) have produced multiple spectra
- Spectra must be cleaned up and processed e.g. to make:
  - Spectra comparable
  - Remove unwanted variation due to data artefacts
  - Make variables within spectra comparable
  - …

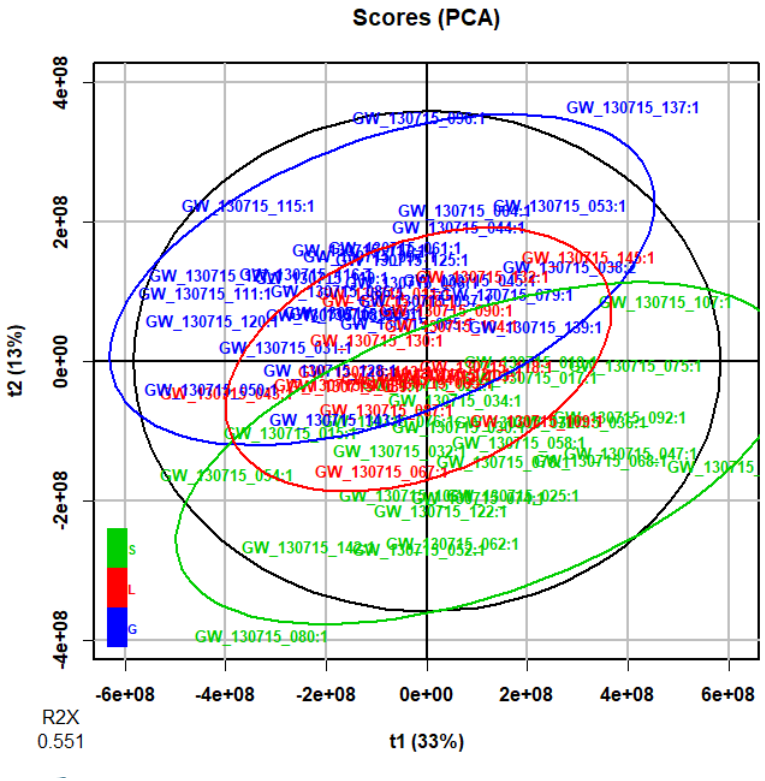


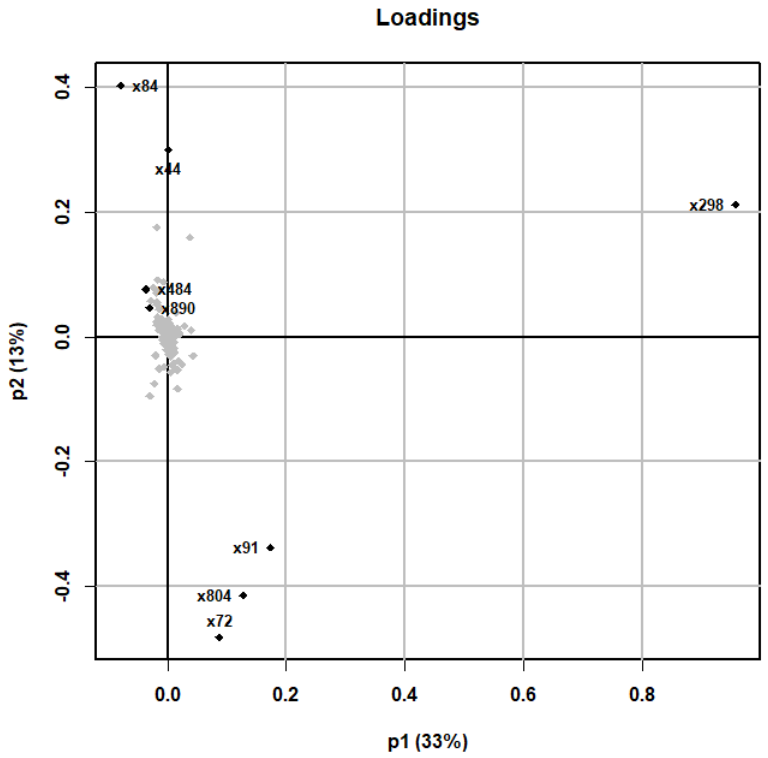
Same data!



Example: assessment of regional differences in Lambrusco wines

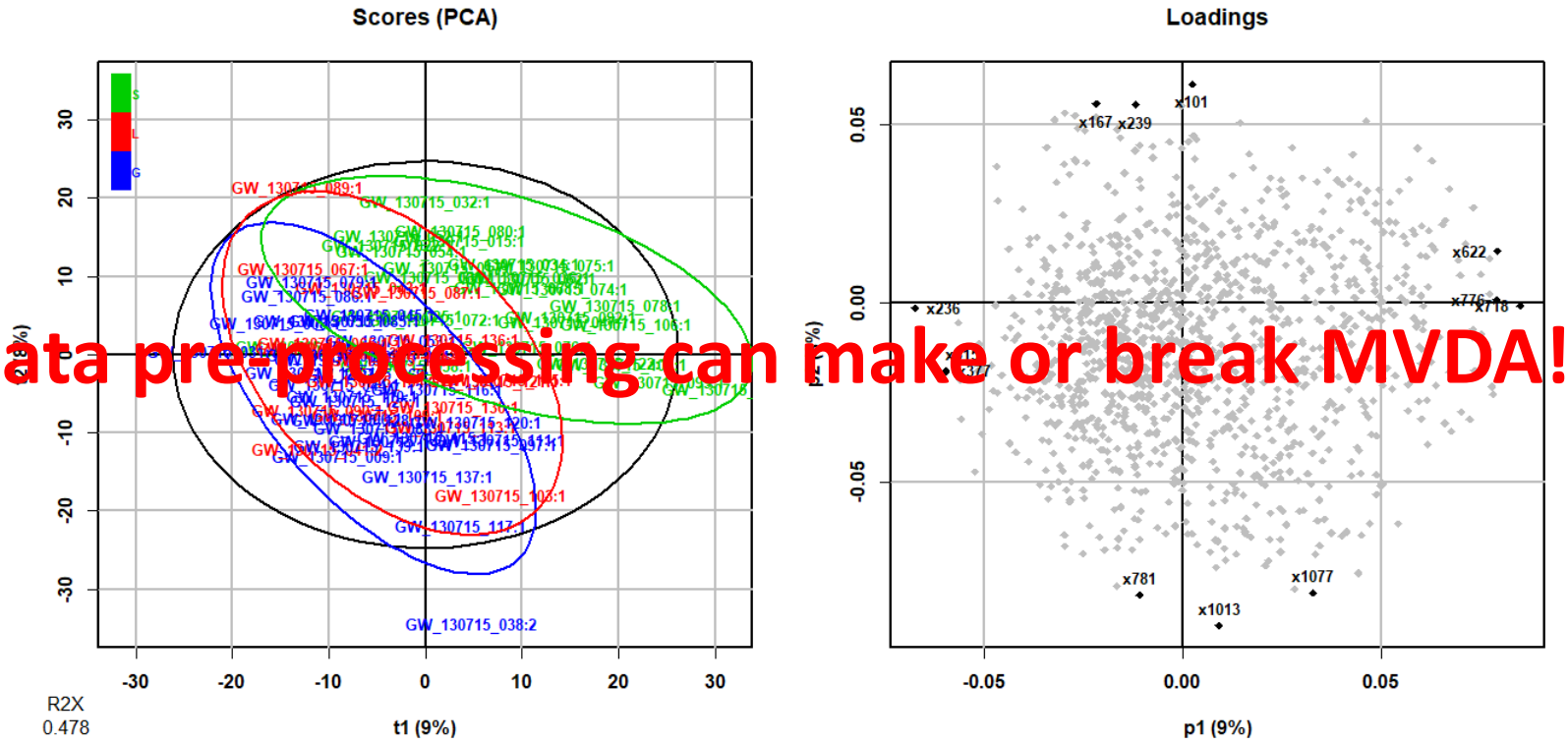
- Wines from 3 regions
- Analysed by GCxGC-MS (76 x 1208)





Example: assessment of regional differences in Lambrusco wines

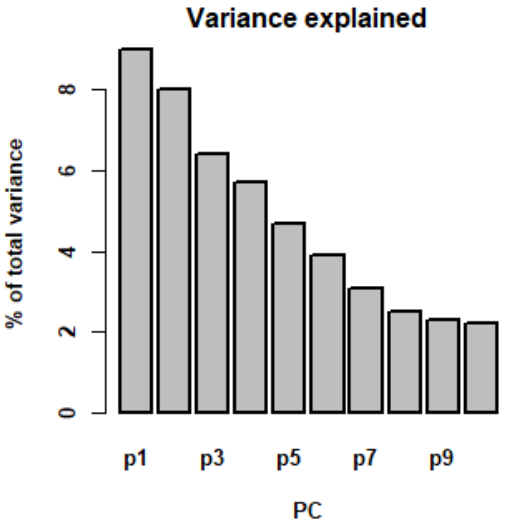
- Autoscaled data

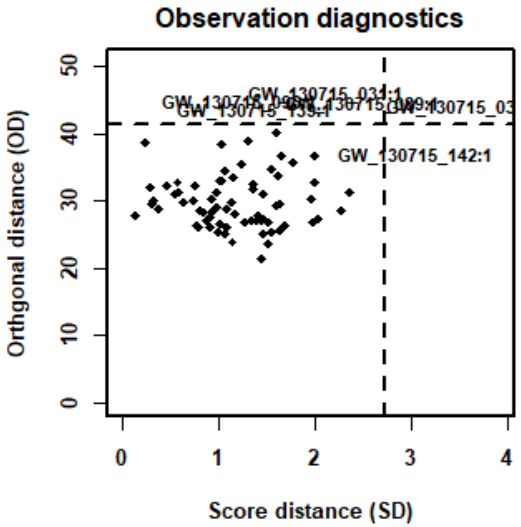


Data preprocessing can make or break MVDA!

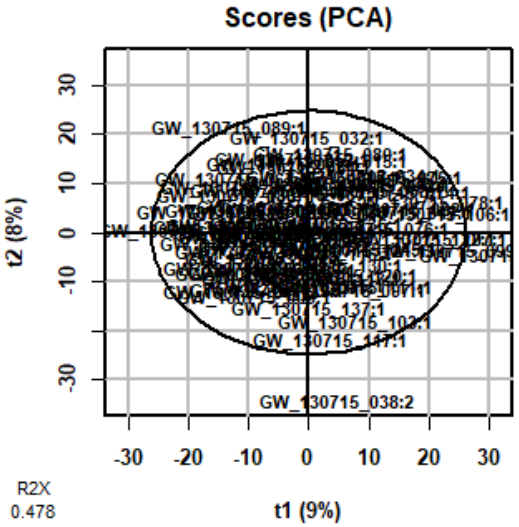
Example: assessment of regional differences in Lambrusco wines

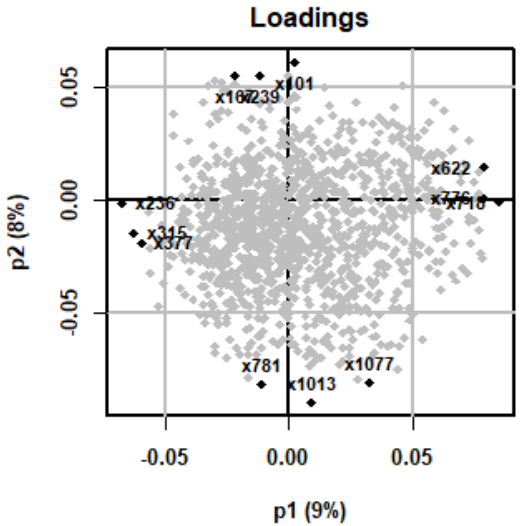
Select # PCs





Outlier detection





Example: untargeted disease diagnosis

Current approaches in disease diagnosis

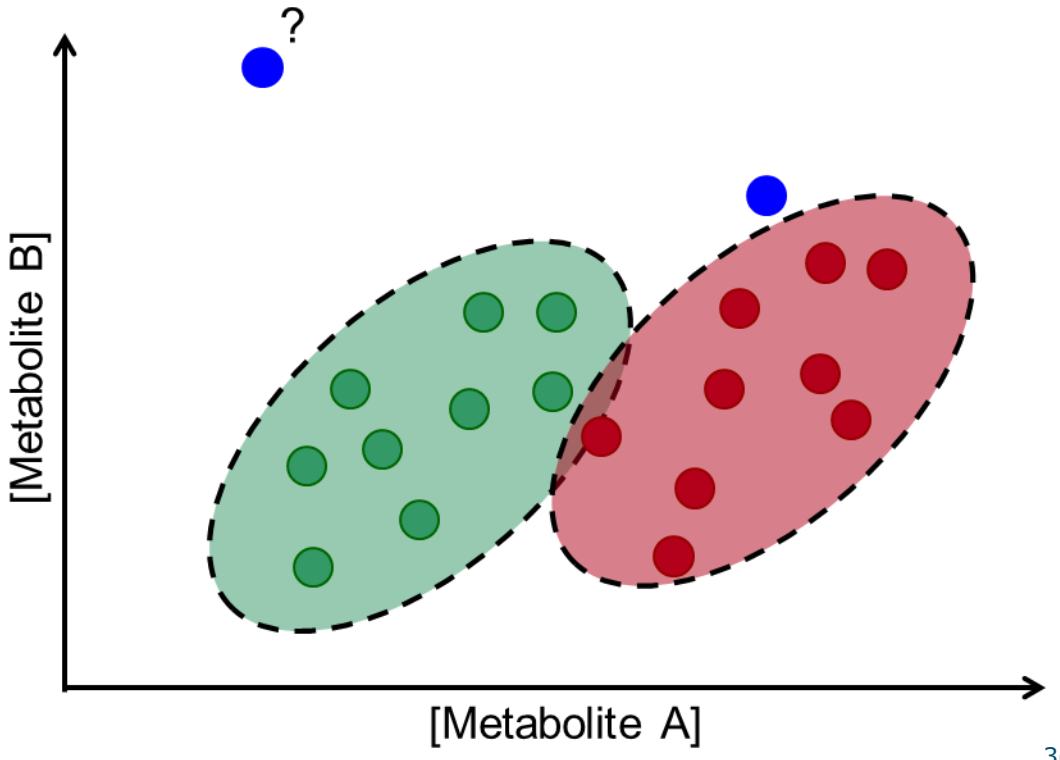
Focus on **one specific (known) disease** status

- Two-class models

Healthy vs disease

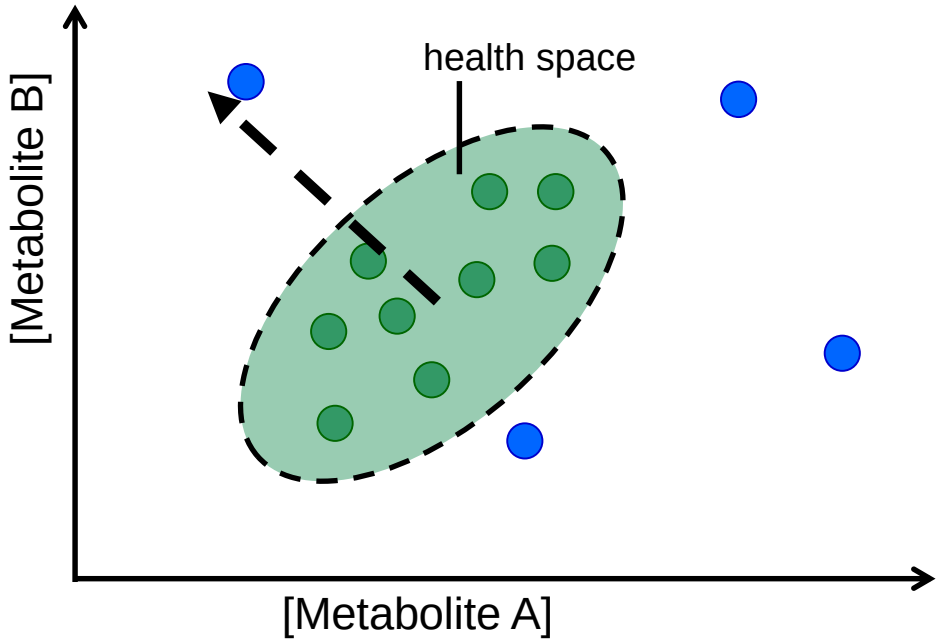
- One-class models

Disease model



Example: untargeted disease diagnosis

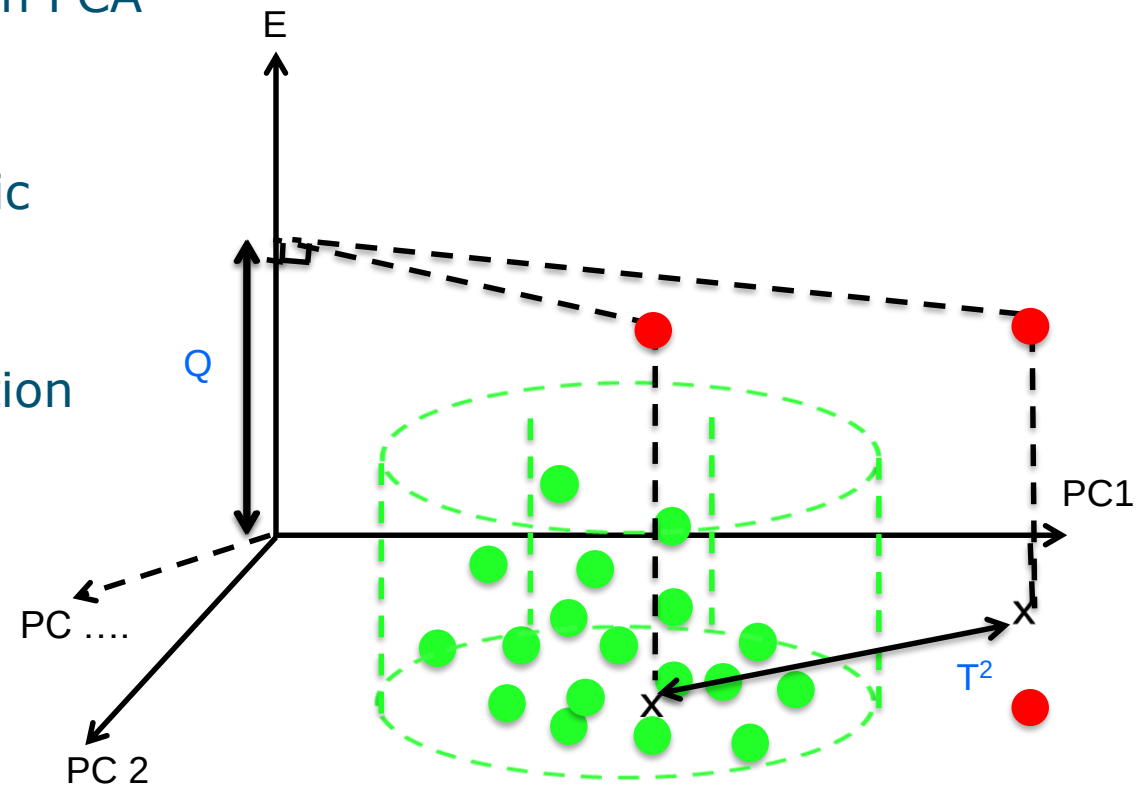
Approach: class model of **healthy** controls



{Engel et al, PLoS One 2014}

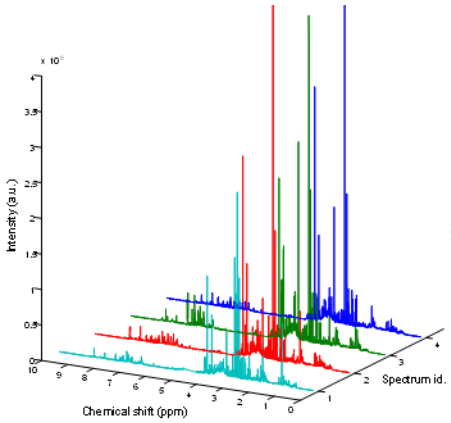
Example: untargeted disease diagnosis

- Model health space with PCA
- Detection via Q-statistic
- Diagnosis via contribution plots

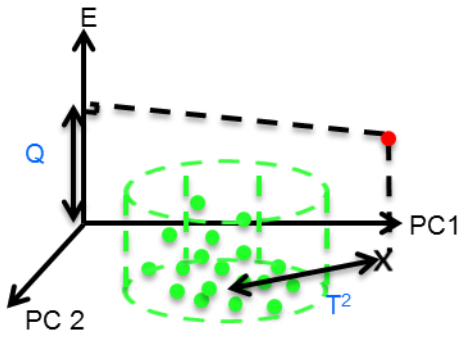


{Engel et al, PLoS One 2014}

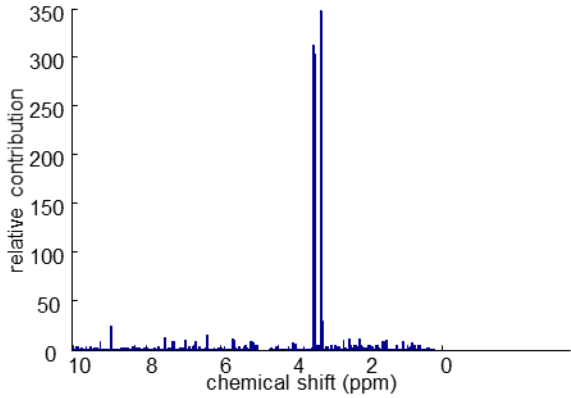
Example: untargeted disease diagnosis



Detection



Diagnosis



Validation data:

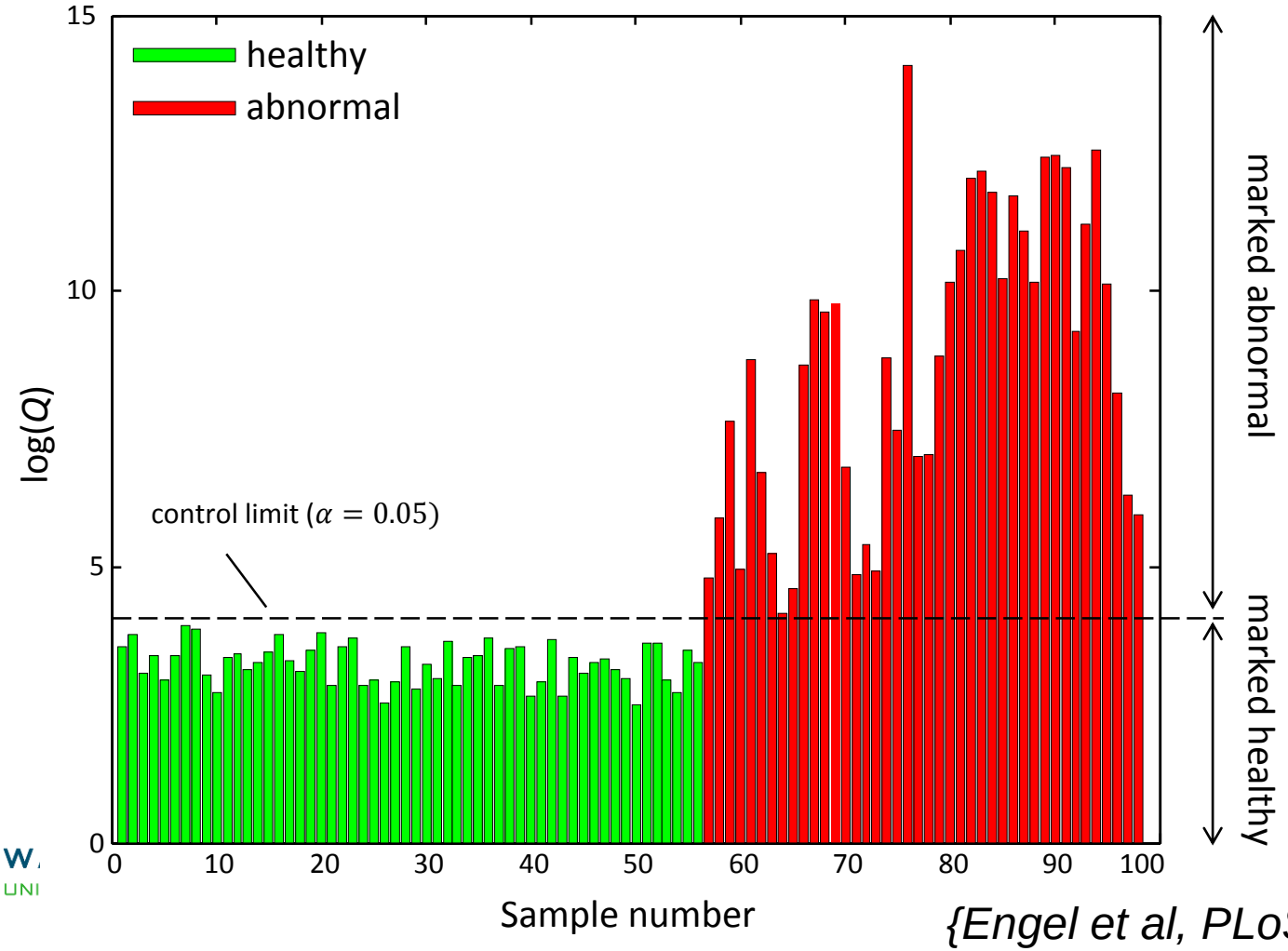
- 120 healthy children to construct model
- 98 validation samples (other children including 42 abnormal)
  - 8 different inborn errors of metabolism
  - 10 abnormalities related to common medication and diet

WAGENINGEN
UNIVERSITY & RESEARCH

{Engel et al, PLoS One 2014}

Example: untargeted disease diagnosis

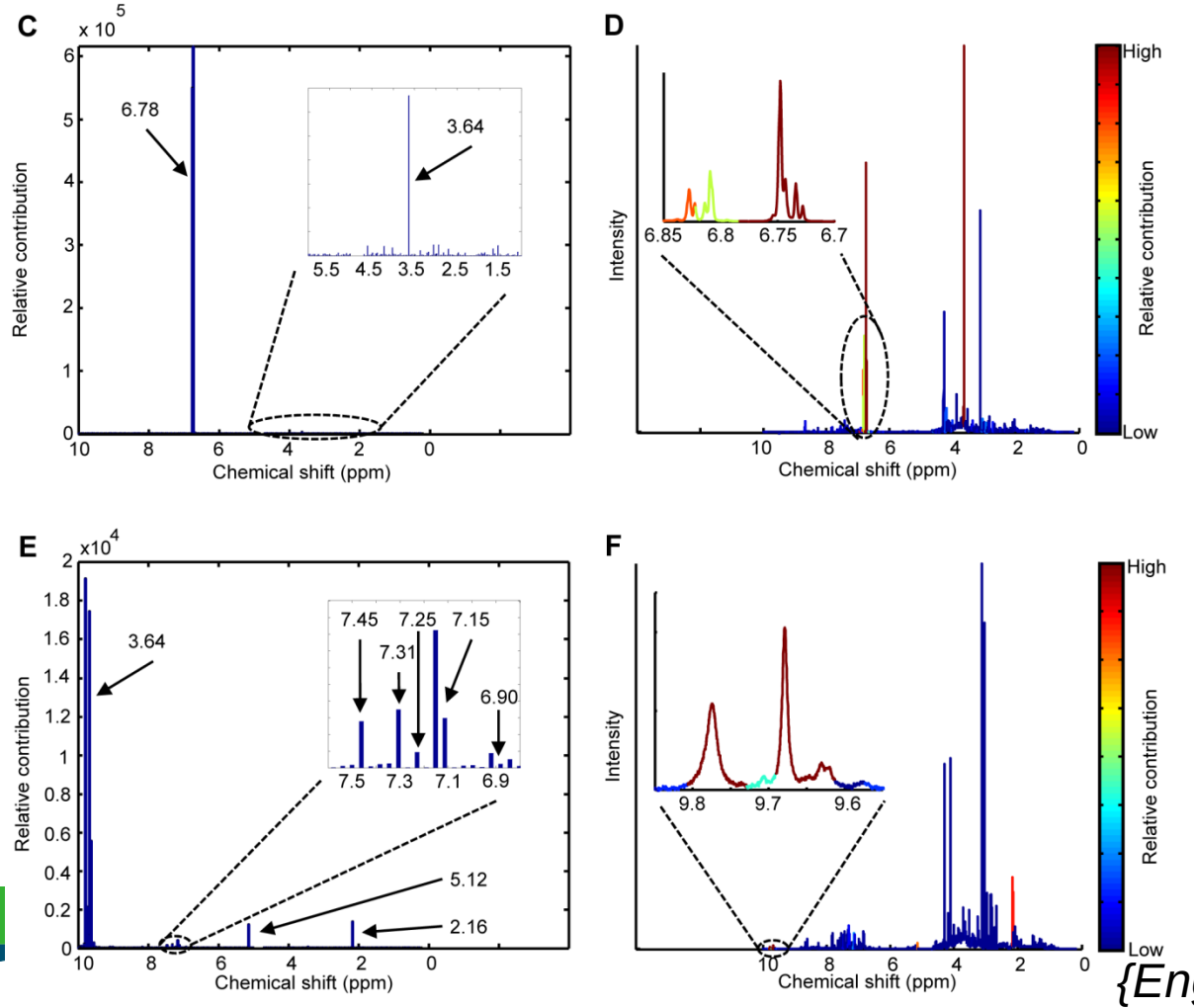
Step 1: detection of abnormal metabotype



{Engel et al, PLoS One 2014}

Example: untargeted disease diagnosis

Step 2: diagnosis of metabolic abnormality



Alkaptonuria disease

Paracetamol consumption

{Engel et al, PLoS One 2014}

Summary – principal component analysis

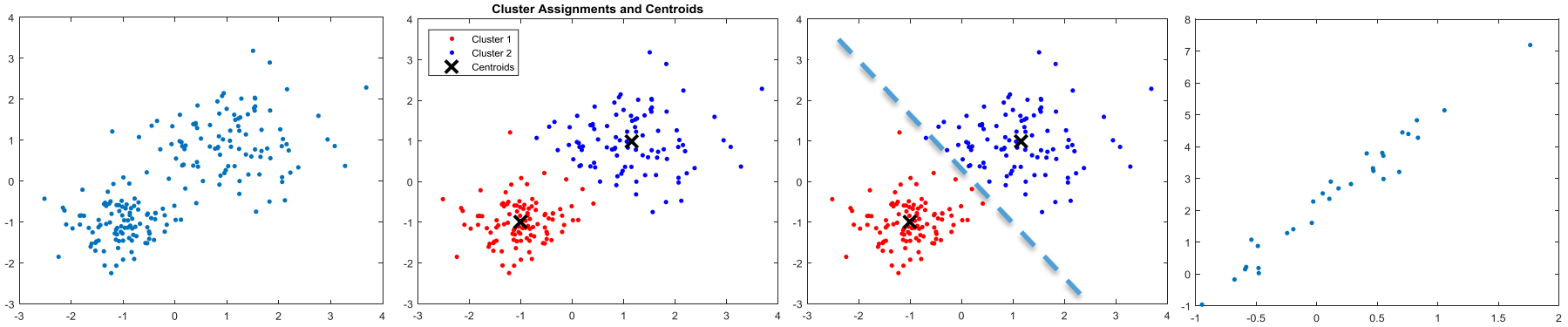
- PCA provides a graphical overview of the data
- Natural starting point for data analysis to visualize trends, groupings, outliers, etc.
- PCA gives
  - Scores: summary of observations
  - Loadings: summary of variable space
- PCA is often used as a dimension reduction step before other statistical methods are applied

Data analysis in metabolomics

Supervised data analysis

WAGENINGEN
UNIVERSITY & RESEARCH

Discriminant analysis



Overview	Clustering	Discriminant analysis / Classification	Regression
• Trends • Patterns • Clusters • Outliers • Quality assurance • Biological diversity	• Grouping of samples and / or variables • Determining group structure • Identification of subgroups • Biological diversity	• Pattern recognition • Discriminating between groups • Assigning samples to groups • Biomarker candidates	• Predicting continuous response • Comparing blocks of omics data
PCA	HCA, k-means	MANOVA, LDA, PLS-DA, O-PLS-DA	PCR, PLS2, O2-PLS

WAGENINGEN
UNIVERSITY & RESEARCH

Supervised statistical analysis

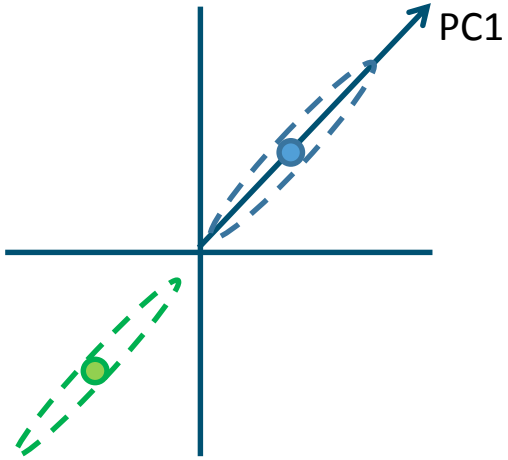
sample label		peak 1	peak 2	peak 3	...		X matrix of signal intensities	
c1	1	0.0031722	-0.0381444	-0.0090	-0.01395	-0.0062389	0.0340444	5.0
c2	1	0.0038722	0.0460556	0.0065	0.00355	-0.0032389	-0.0037556	8.4
c3	1	-0.0036278	-0.0157444	0.0018	0.00365	-0.0072389	-0.0127556	9.2
c4	1	0.0200722	-0.0175444	0.0134	0.01555	0.0066611	-0.0163556	1.3
c5	1	-0.0004278	0.0175556	0.0052	0.00985	0.0027611	0.0056444	4.9
c6	1	0.0010722	0.0053556	0.0016	0.00125	0.0105611	0.0099444	7.9
indo1	2	-0.0075278	0.0100556	0.0005	-0.00415	-0.0045389	0.0225444	16.6
indo2	2	-0.0000278	-0.0231444	-0.0018	-0.01515	0.0046611	0.0047444	18.2
indo3	2	-0.0017278	0.0094556	0.0024	-0.00175	0.0021611	-0.0293556	14.4
indo4	2	-0.0003278	-0.0255444	-0.0054	-0.00165	-0.0015389	0.0089444	10.0
indo5	2	-0.0017278	-0.0212444	-0.0083	-0.00115	-0.0065389	-0.0296556	25.9
indo6	2	-0.0002278	-0.0327444	-0.0039	-0.00625	-0.0103389	-0.0192556	21.3
mpa1	3	0.0022722	0.0287556	-0.0012	0.02465	-0.0004389	-0.0073556	103.1
mpa2	3	-0.0002278	0.0363556	0.0010	0.00815	-0.0025389	0.0179444	69.9
mpa3	3	-0.0027278	-0.0068444	0.0013	0.00205	0.0045611	0.0061444	91.3
mpa4	3	-0.0055278	-0.0216444	-0.0011	-0.00815	0.0010611	0.0030444	98.5
mpa5	3	-0.0022278	0.0230556	-0.0025	-0.01055	0.0047611	0.0269444	48.1
mpa6	3	-0.0041278	0.0259556	-0.0005	-0.00595	0.0054611	-0.0214556	59.7

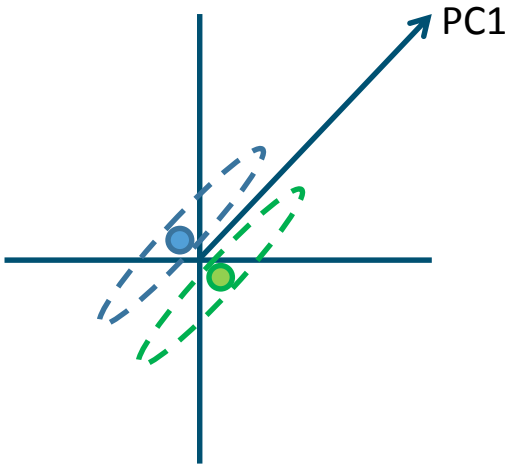
EITHER
Y matrix = treatment group labels = discrete variable

OR
Y matrix = separate non-metabolic measurement for each sample = continuous variable

“Disadvantage” PCA

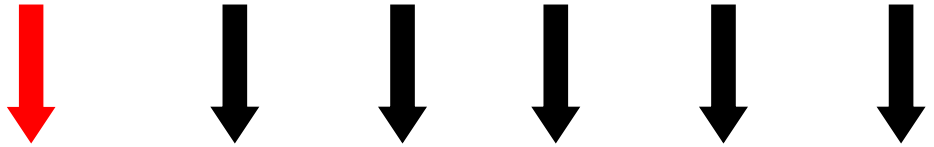
- Group separation is not always observed





Univariate data analysis

sample label		peak 1	peak 2	peak 3	…		
c1	1	0.0031722	-0.0381444	-0.0090	-0.01395	-0.0062389	0.0340444
c2	1	0.0038722	0.0460556	0.0065	0.00355	-0.0032389	-0.0037556
c3	1	-0.0036278	-0.0157444	0.0018	0.00365	-0.0072389	-0.0127556
c4	1	0.0200722	-0.0175444	0.0134	0.01555	0.0066611	-0.0163556
c5	1	-0.0004278	0.0175556	0.0052	0.00985	0.0027611	0.0056444
c6	1	0.0010722	0.0053556	0.0016	0.00125	0.0105611	0.0099444
indo1	2	-0.0075278	0.0100556	0.0005	-0.00415	-0.0045389	0.0225444
indo2	2	-0.0000278	-0.0231444	-0.0018	-0.01515	0.0046611	0.0047444
indo3	2	-0.0017278	0.0094556	0.0024	-0.00175	0.0021611	-0.0293556
indo4	2	-0.0003278	-0.0255444	-0.0054	-0.00165	-0.0015389	0.0089444
indo5	2	-0.0017278	-0.0212444	-0.0083	-0.00115	-0.0065389	-0.0296556
indo6	2	-0.0002278	-0.0327444	-0.0039	-0.00625	-0.0103389	-0.0192556
mpa1	3	0.0022722	0.0287556	-0.0012	0.02465	-0.0004389	-0.0073556
mpa2	3	-0.0002278	0.0363556	0.0010	0.00815	-0.0025389	0.0179444
mpa3	3	-0.0027278	-0.0068444	0.0013	0.00205	0.0045611	0.0061444
mpa4	3	-0.0055278	-0.0216444	-0.0011	-0.00815	0.0010611	0.0030444
mpa5	3	-0.0022278	0.0230556	-0.0025	-0.01055	0.0047611	0.0269444
mpa6	3	-0.0041278	0.0259556	-0.0005	-0.00595	0.0054611	-0.0214556



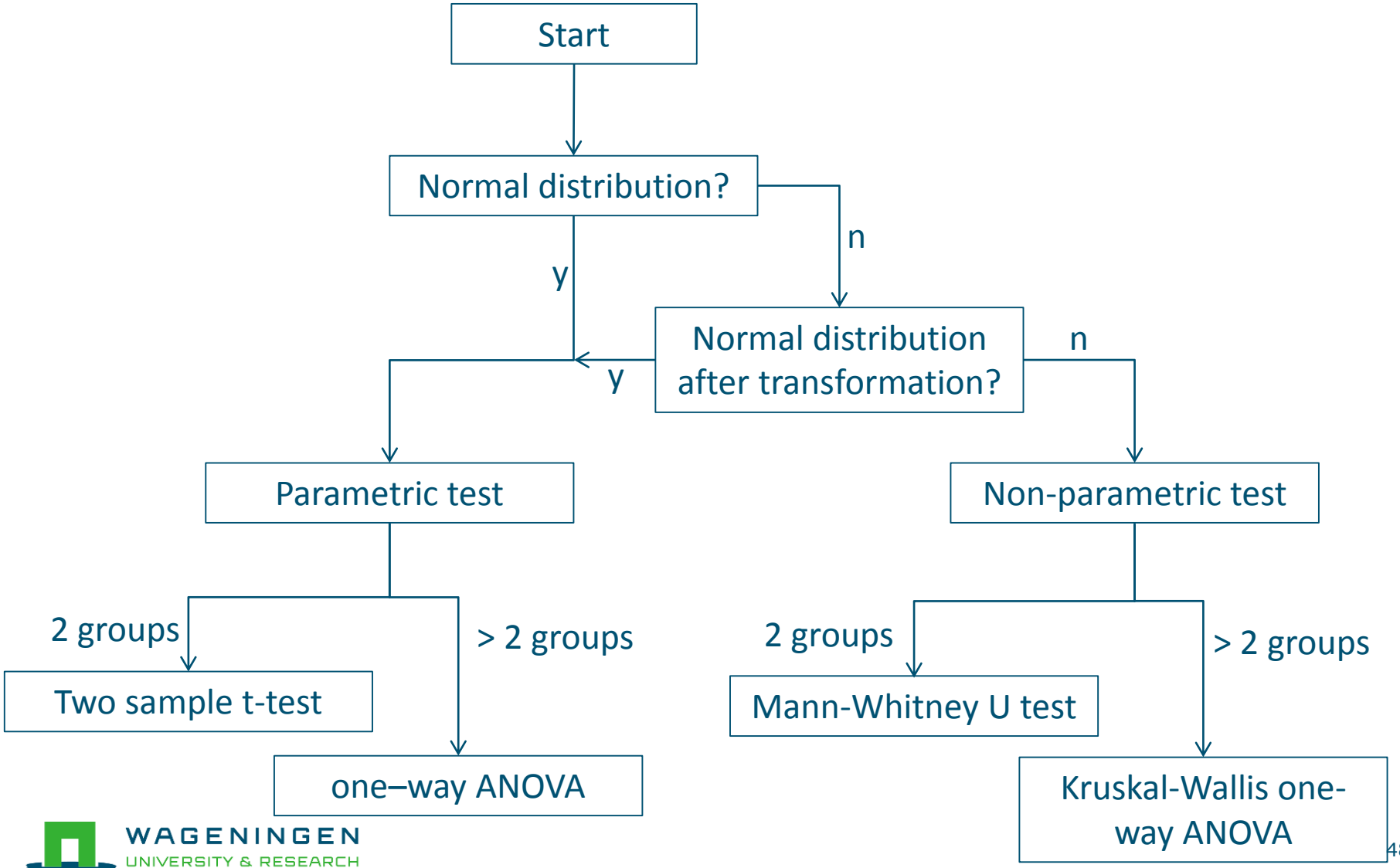
t-test or ANOVA

t-test or ANOVA…

with false discovery rate (FDR) correction

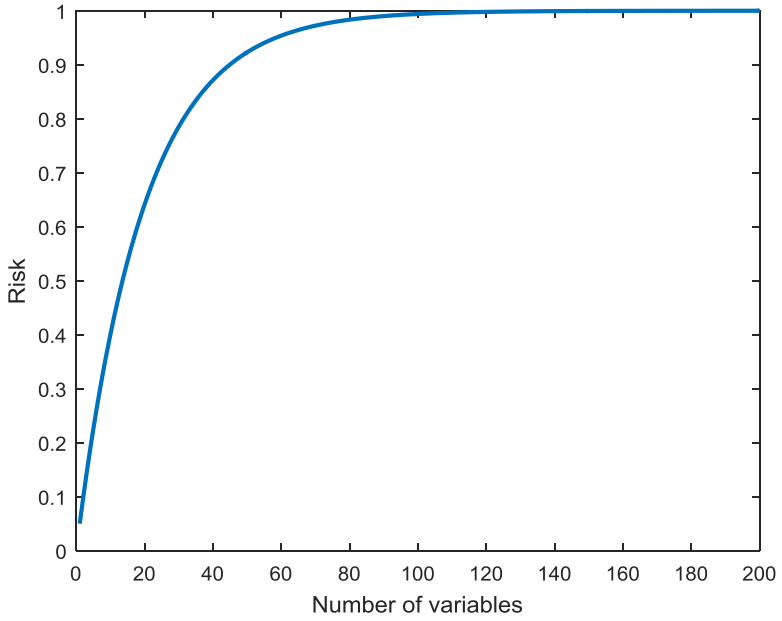
WAGENINGEN UNIVERSITY & RESEARCH

Commonly used methods



Multiple comparisons

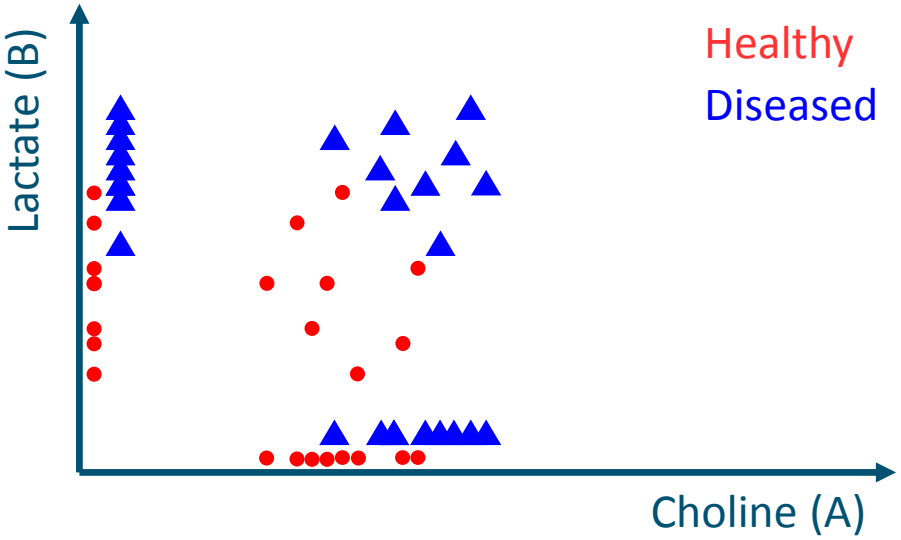
- Typically a 5% significance level is employed for each statistical test that is carried out
    - Type I errors: false positives, spurious results
    - Type II errors: false negatives, risk of not identifying relevant peaks
- Risk type I error = $1 - 0.95^K$

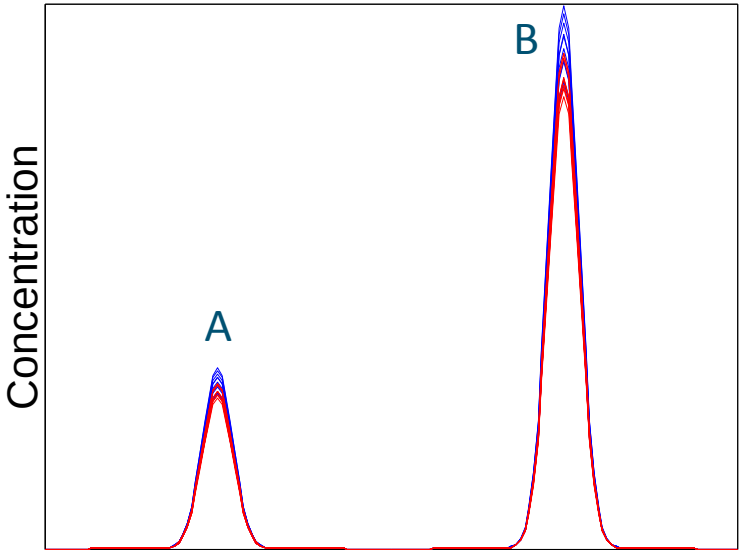


Multiple testing correction

- Bonferroni-Holm procedure controls the FWER:
  - Use adjusted significance level $\alpha_{adj} \approx \alpha \backslash \mathrm{p}$
  - Often overly strict
  - Many false negatives
- Benjamini-Hochberg procedure controls the FDR
  - Controls proportion of false positives, i.e. the number of false positives amongst the set of significant variables
  - Consequence: at the most 5% false positives amongst the set of significant variables

Univariate vs multivariate analysis





Variables (metabolites) should be studied together!

(Additionally, there is the issue of multiple-testing in univariate statistics)

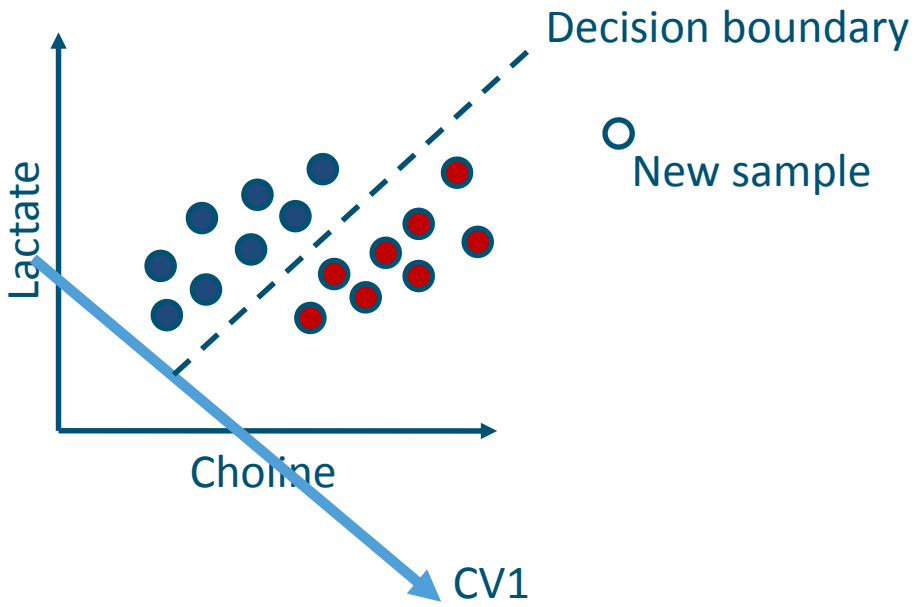
Multivariate data analysis

sample label		peak 1	peak 2	peak 3	...		
c1	1	0.0031722	-0.0381444	-0.0090	-0.01395	-0.0062389	0.0340444
c2	1	0.0038722	0.0460556	0.0065	0.00355	-0.0032389	-0.0037556
c3	1	-0.0036278	-0.0157444	0.0018	0.00365	-0.0072389	-0.0127556
c4	1	0.0200722	-0.0175444	0.0134	0.01555	0.0066611	-0.0163556
c5	1	-0.0004278	0.0175556	0.0052	0.00985	0.0027611	0.0056444
c6	1	0.0010722	0.0053556	0.0016	0.00125	0.0105611	0.0099444
indo1	2	-0.0075278	0.0100556	0.0005	-0.00415	-0.0045389	0.0225444
indo2	2	-0.0000278	-0.0231444	-0.0018	-0.01515	0.0046611	0.0047444
indo3	2	-0.0017278	0.0094556	0.0024	-0.00175	0.0021611	-0.0293556
indo4	2	-0.0003278	-0.0255444	-0.0054	-0.00165	-0.0015389	0.0089444
indo5	2	-0.0017278	-0.0212444	-0.0083	-0.00115	-0.0065389	-0.0296556
indo6	2	-0.0002278	-0.0327444	-0.0039	-0.00625	-0.0103389	-0.0192556
mpa1	3	0.0022722	0.0287556	-0.0012	0.02465	-0.0004389	-0.0073556
mpa2	3	-0.0002278	0.0363556	0.0010	0.00815	-0.0025389	0.0179444
mpa3	3	-0.0027278	-0.0068444	0.0013	0.00205	0.0045611	0.0061444
mpa4	3	-0.0055278	-0.0216444	-0.0011	-0.00815	0.0010611	0.0030444
mpa5	3	-0.0022278	0.0230556	-0.0025	-0.01055	0.0047611	0.0269444
mpa6	3	-0.0041278	0.0259556	-0.0005	-0.00595	0.0054611	-0.0214556

Analyse data in its entirety

Linear discriminant analysis

- Rotation of the data (similar to PCA)
- Rotation such that SSB/SSW is maximized
- Classification of (new) observations



WAGENINGEN
UNIVERSITY & RESEARCH

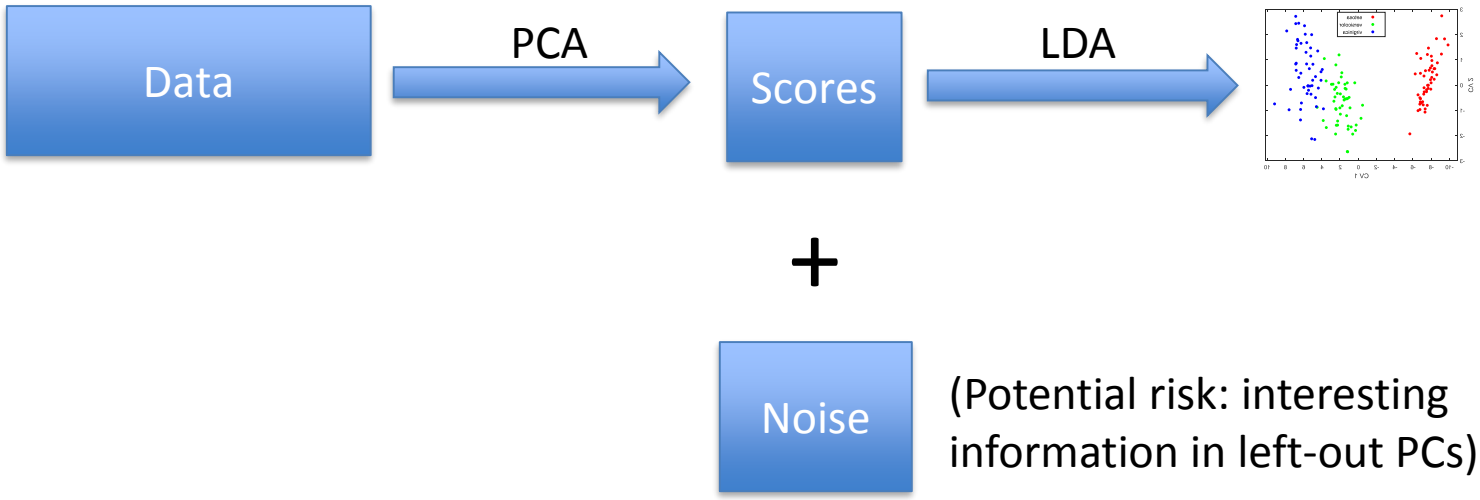
Disadvantage of LDA for metabolomics

- LDA is a traditional statistical method designed for long and slim data tables

- Metabolomics data is short and fat → traditional methods break down (curse of dimensionality)

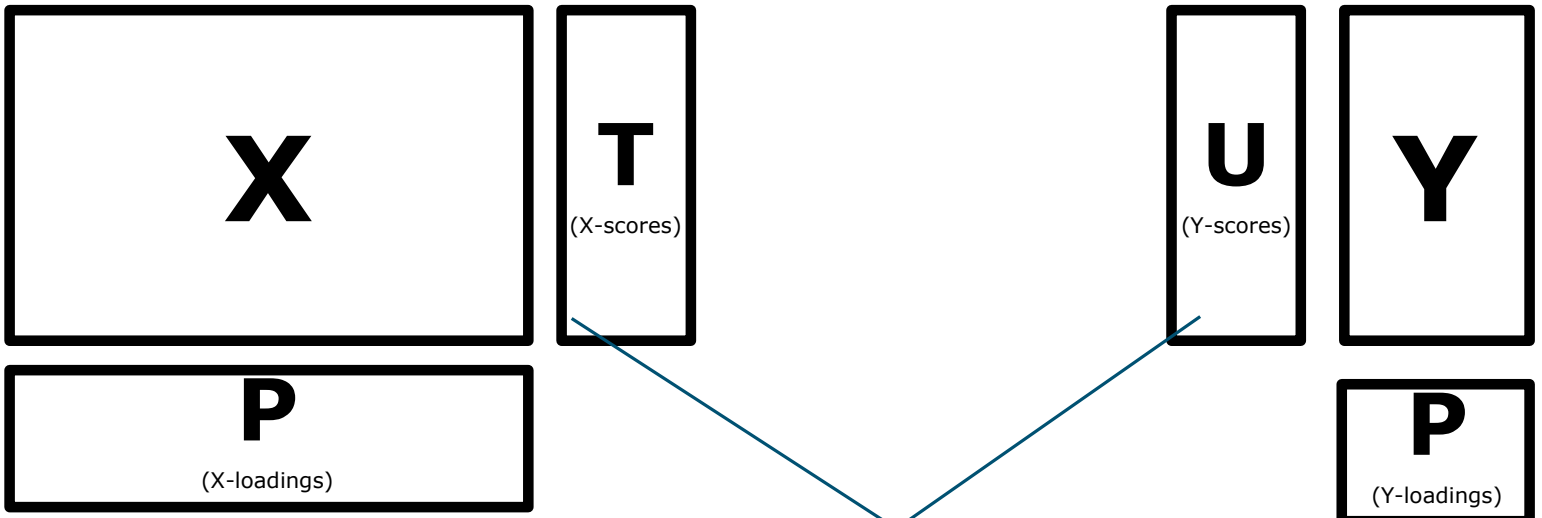
PCA + LDA

- Application of LDA to PCA scores



Partial least squares – discriminant analysis

- "Linked" dimension reduction of **X** and **Y** matrix

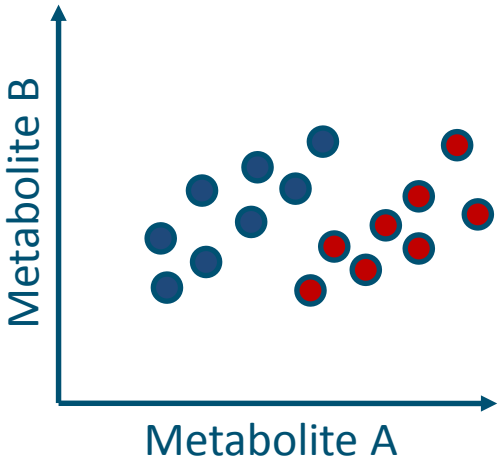


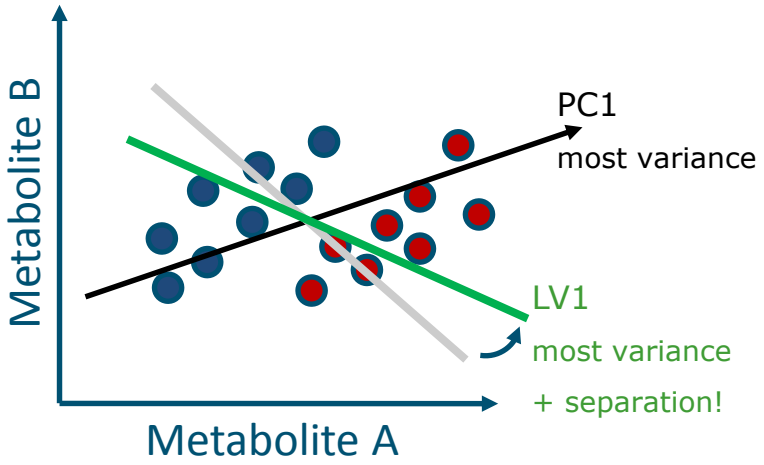
$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

$$\mathbf{u}_1 = r_1\mathbf{t}_1$$
$$\mathbf{u}_2 = r_2\mathbf{t}_2$$
etc.

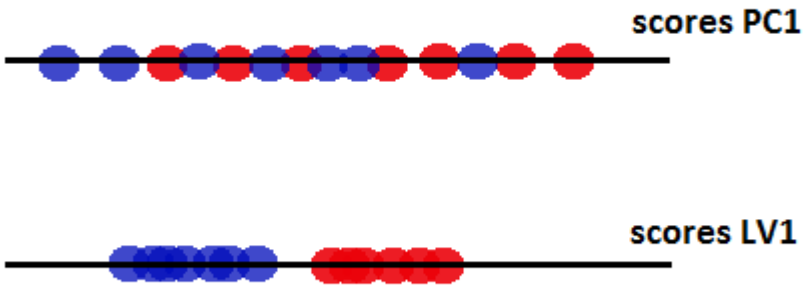
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}$$

Partial least squares – discriminant analysis (PLS-DA)



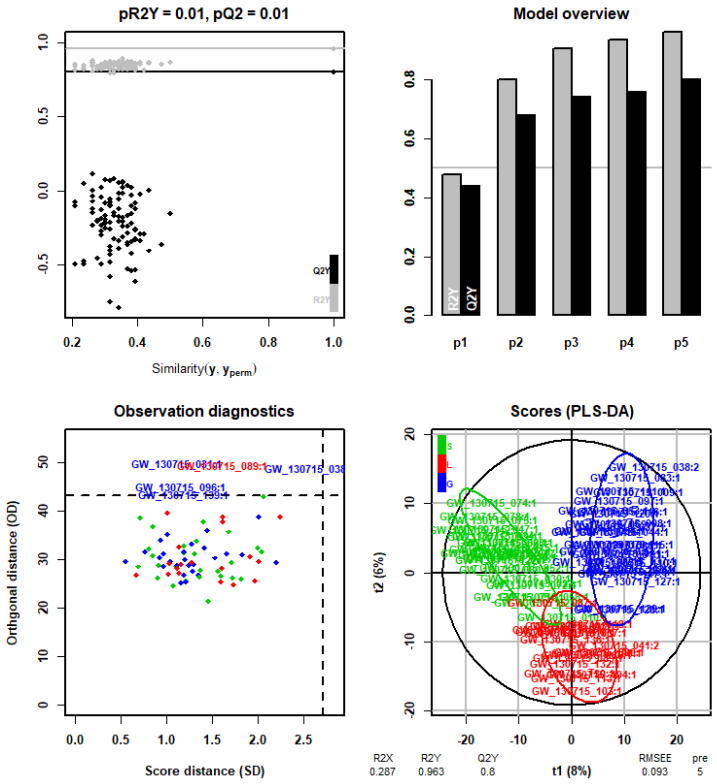


- PCA = best description
- LDA = best separation
- PLS-DA
  - Describes variance in data like PCA
  - **Separates classes**
- N.B.: "Latent Variables"



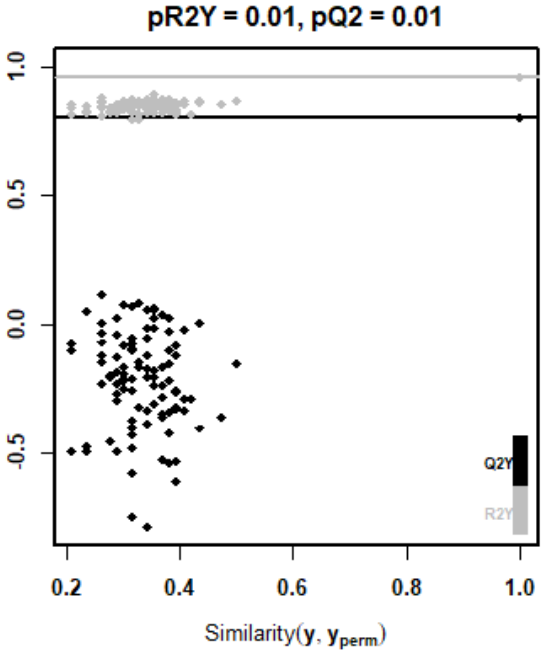
Example: assessment of regional differences in Lambrusco wines

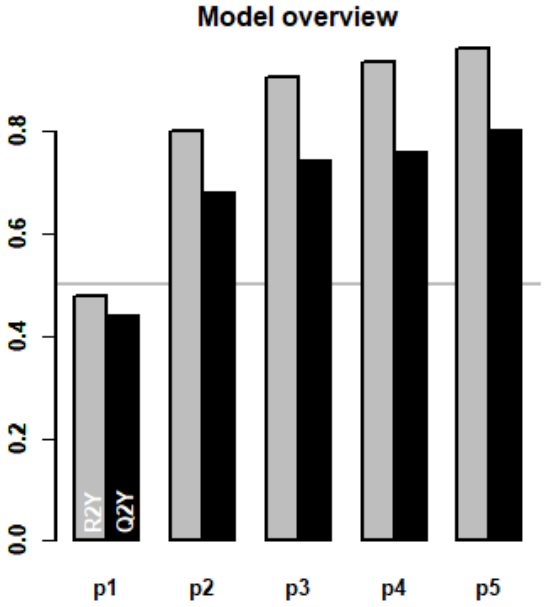
- Wines from 3 regions
- Analysed by GCxGC-MS (76 x 1208)

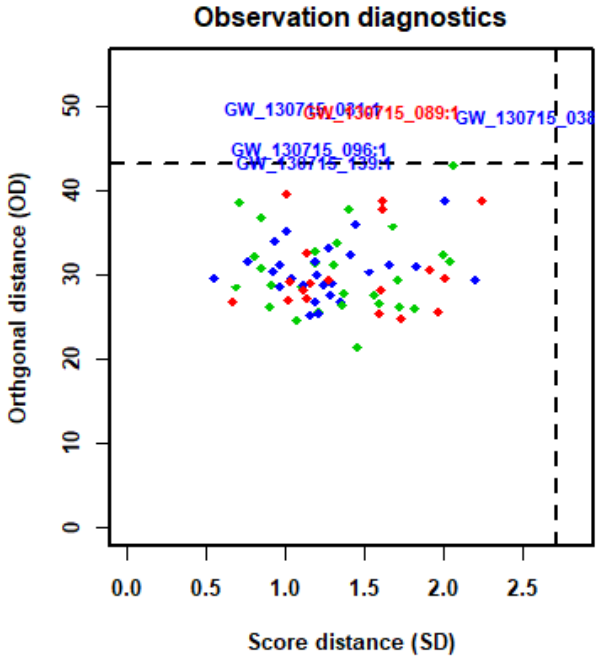


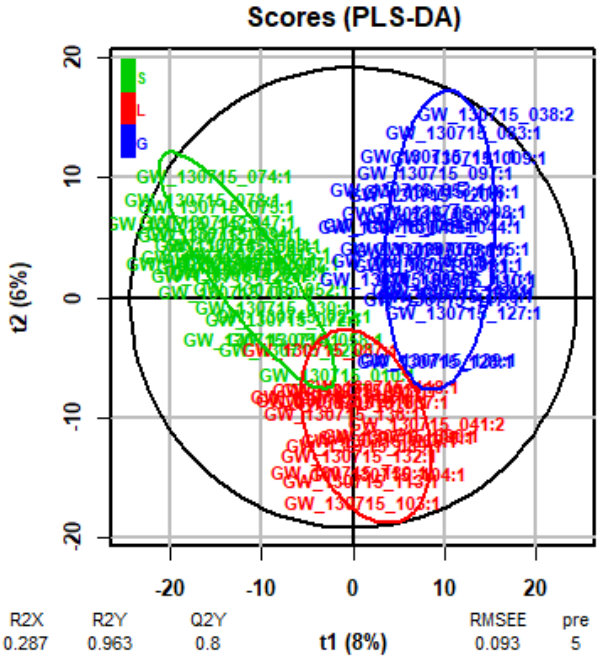
Examp in Lambr

- Wine
- Analy



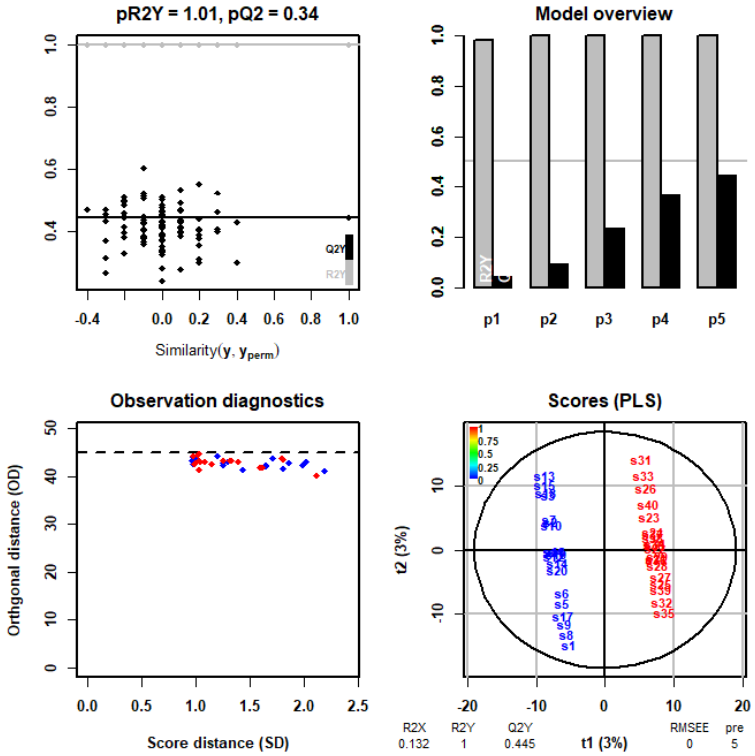




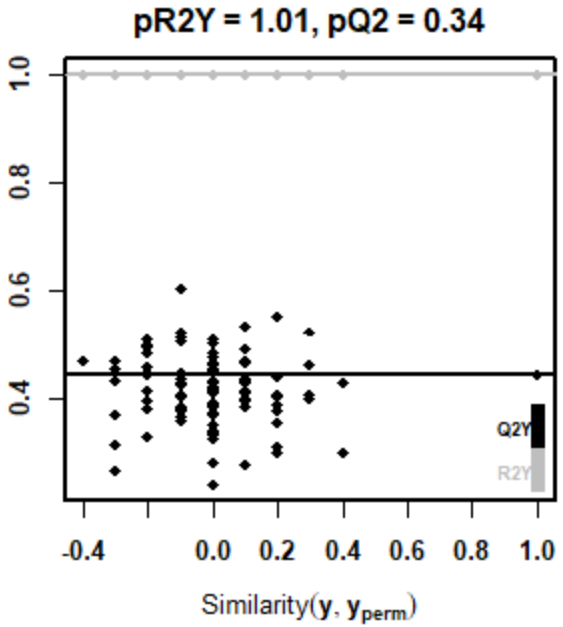


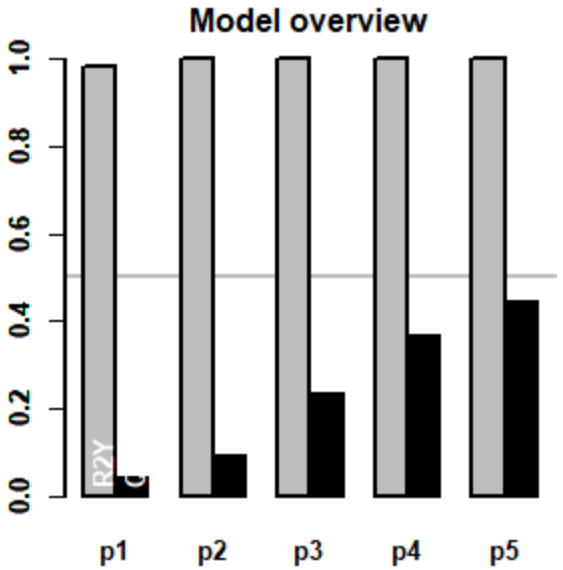
Application to nonsense data

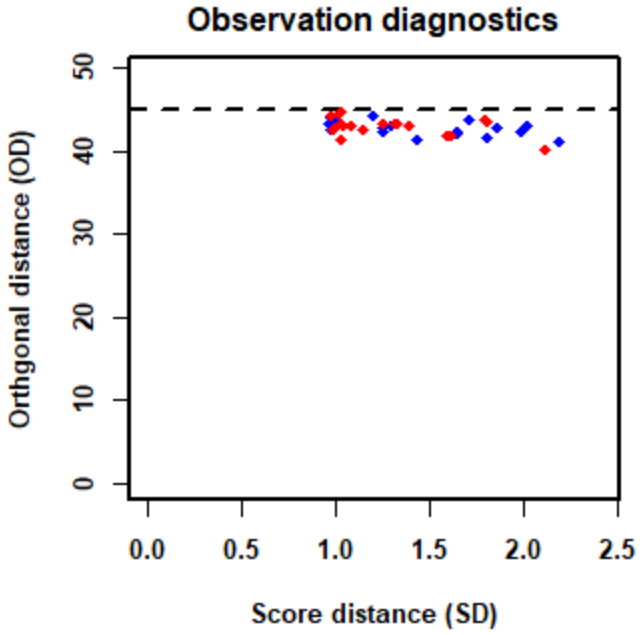
- Randomly generated data
- Two groups
- 20 observations per group
- 2000 variables

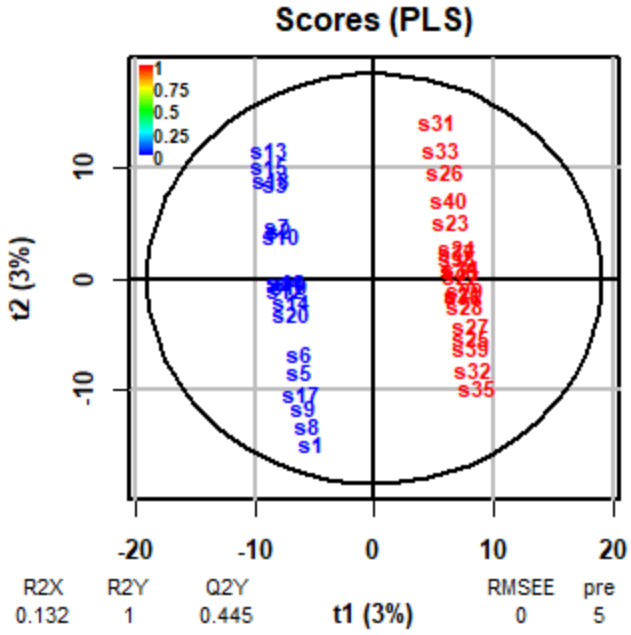


Appl

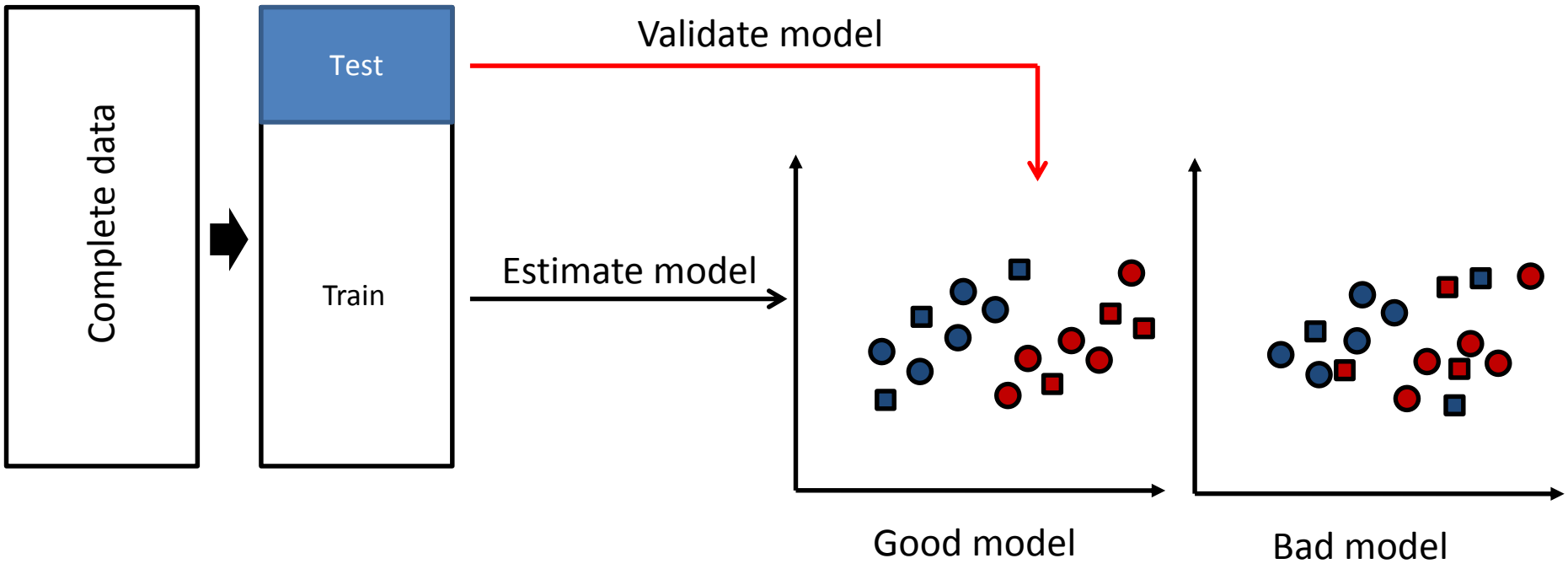

pR2Y = 1.01, pQ2 = 0.34
1.0
0.8
0.6
0.4
Q2Y
R2Y
-0.4
0.0
0.2
0.4
0.6
0.8
1.0
Similarity(y, y_perm)


Model overview
1.0
0.8
0.6
0.4
0.2
0.0
R2Y
p1
p2
p3
p4
p5


Observation diagnostics
Orthgonal distance (OD)
50
40
30
20
10
0
0.0
0.5
1.0
1.5
2.0
2.5
Score distance (SD)


Scores (PLS)
1
0.75
0.5
0.25
0
t2 (3%)
10
0
-10
-20
-10
0
10
20
t1 (3%)
R2X
0.132
R2Y
1
Q2Y
0.445
RMSEE
0
pre
5

(Cross)-Validation



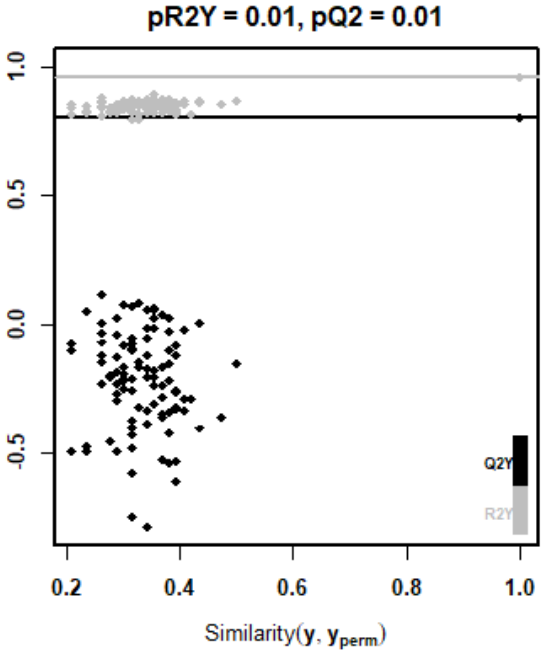
training observation

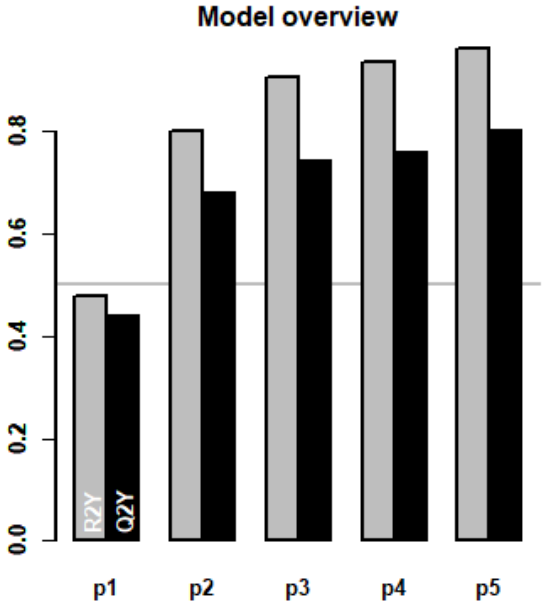
test observation

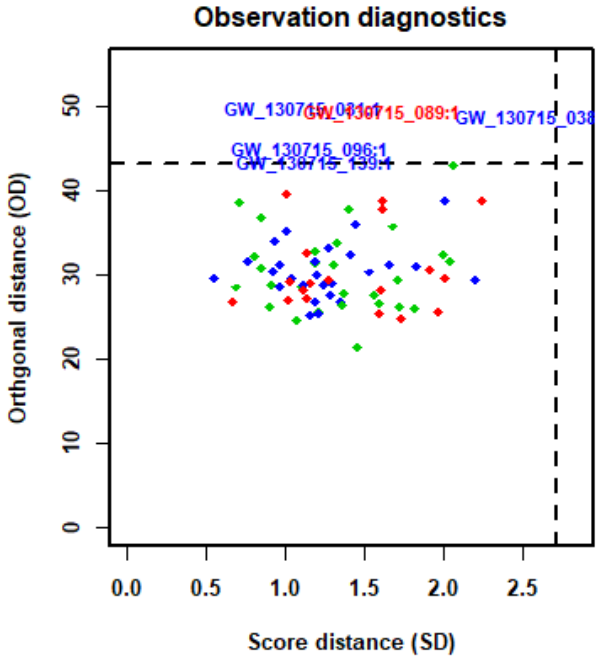
WAGENINGEN
UNIVERSITY & RESEARCH

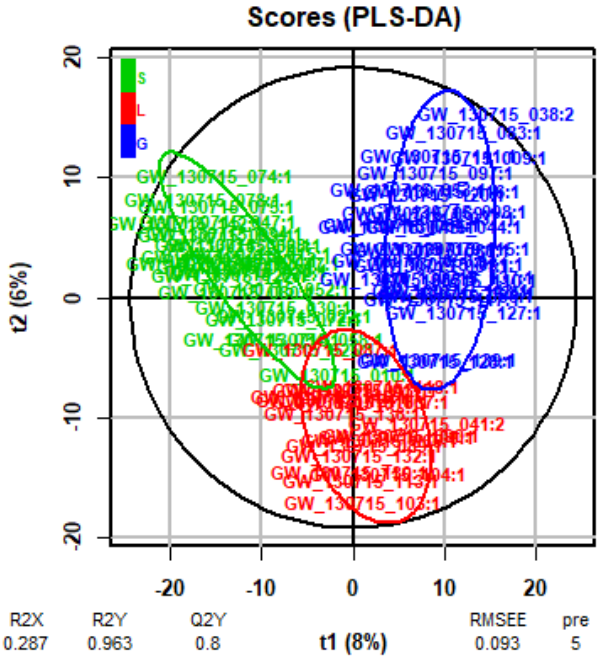
Examp in Lambr

- Wine
- Analy









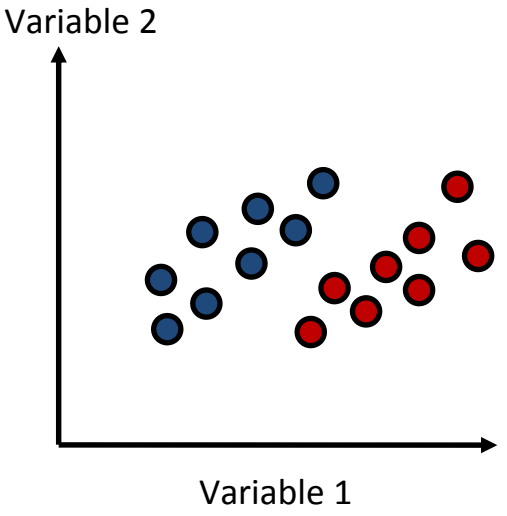
Permutation testing: how it works

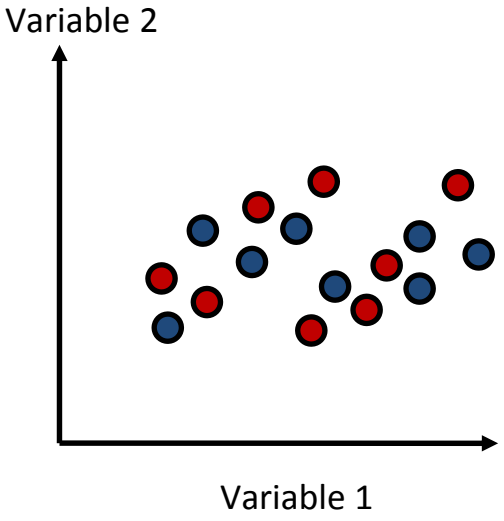
- Permutation testing is used to assess the statistical significance

1) Fit model using correct class labels

2) Fit model using permuted class labels

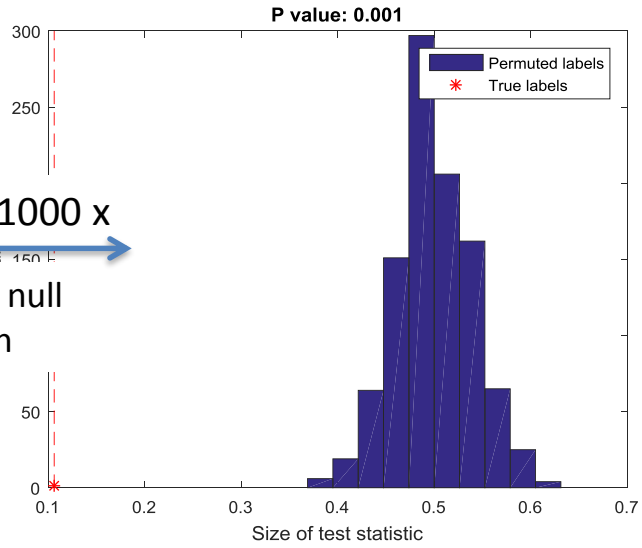
3) Estimate p-value





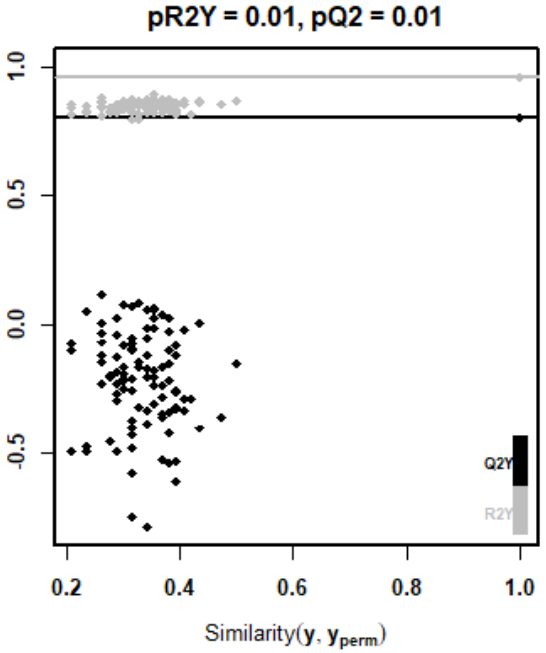
Repeat +- 1000 x

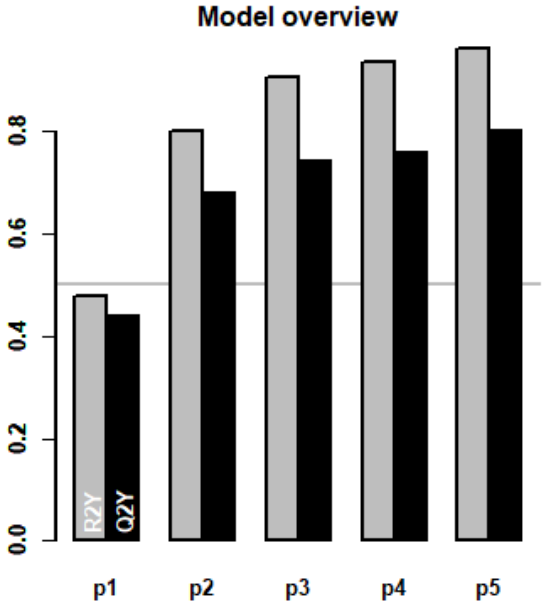
Determine null distribution

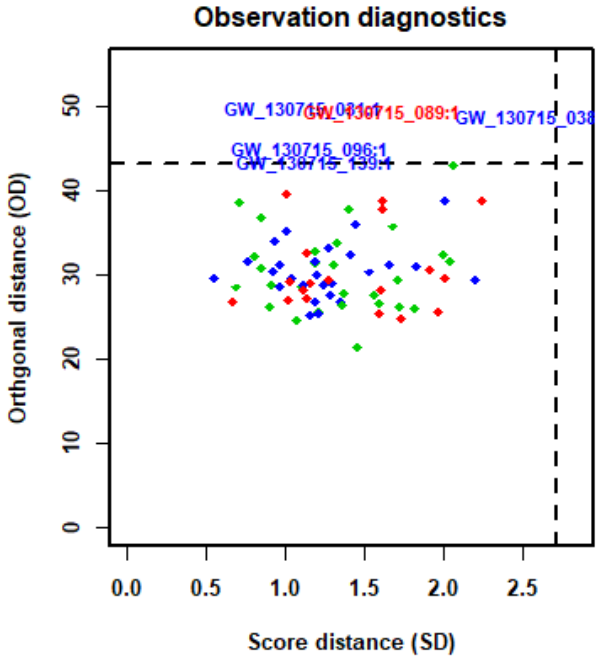


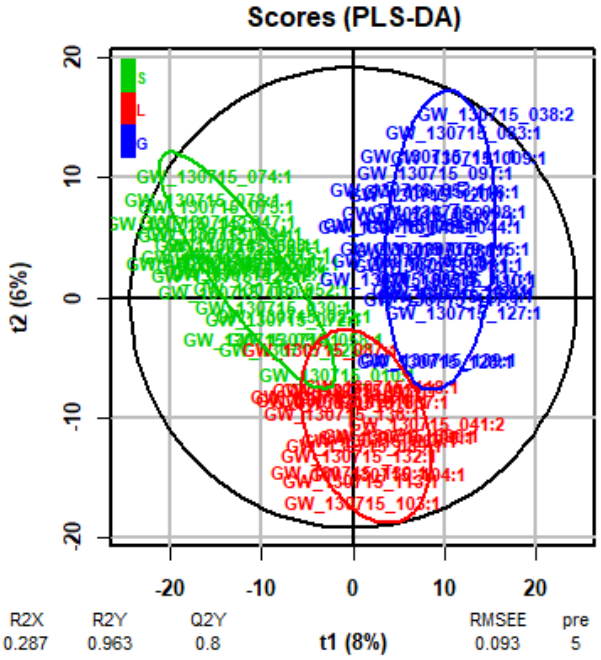
Examp in Lambr

- Wine
- Analy





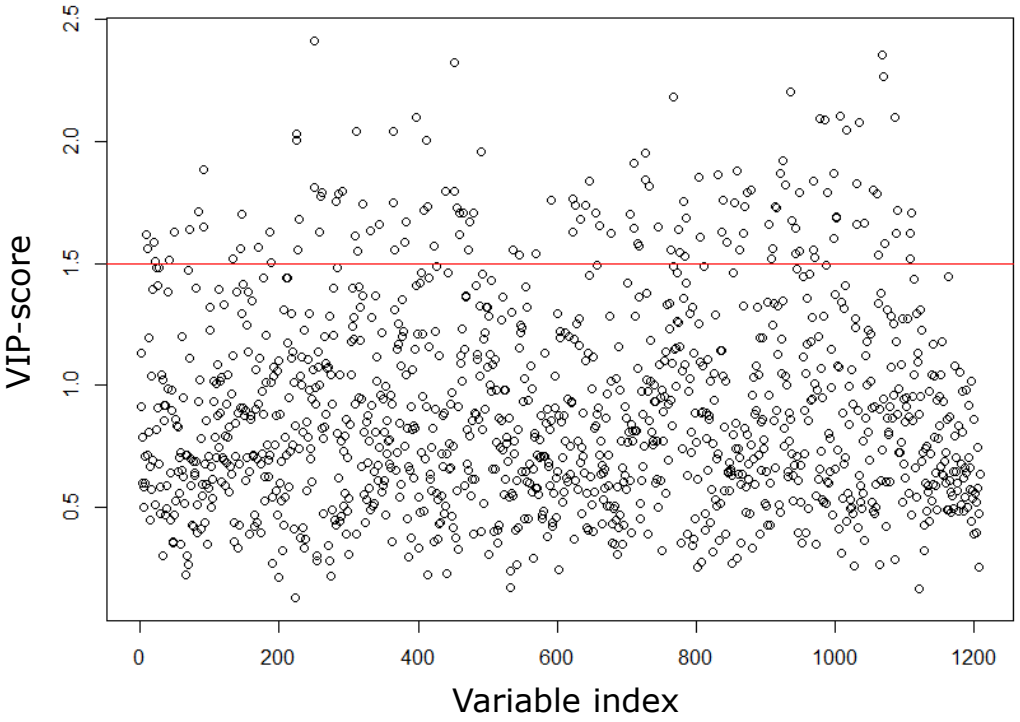


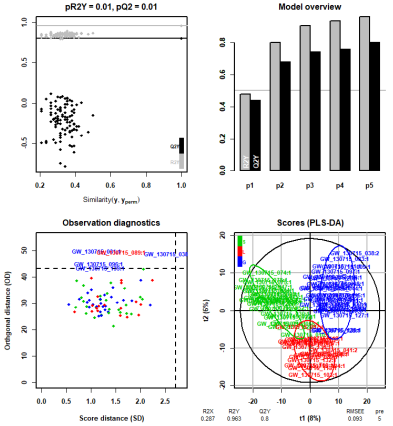


Example: assessment of regional differences in Lambrusco wines

- Interpretation: VIP-score
- Many other possibilities

(target projection, selectivity ratio, simple multivariate correlation)





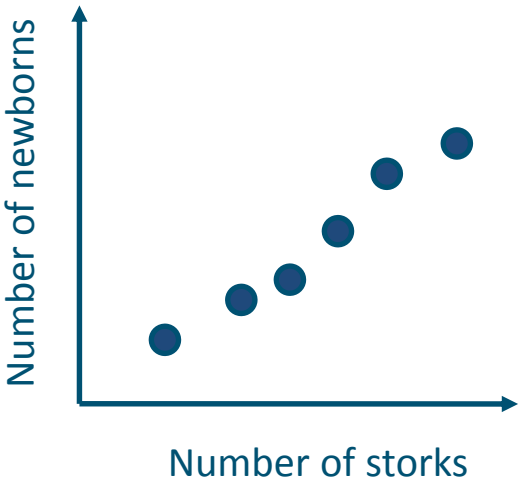
Summary – partial least squares

- For classification and regression problems
- "Directed" dimension reduction (response is taken into account)
- PLS gives
  - Predictions, scores, loadings, variable importance measures
- PLS is prone to overfitting → validation is crucial!

Some words of caution

- Correlation ≠ causation
- Real, but non-causal signal?
- Spurious correlation?

→ Further evidence is required



Typical workflow

1. Data visualization with PCA
    1. Study technical variability (spread QCs vs spread biological observations)
    2. Detect outliers
    3. Detect trends and clusters

2. Univariate analysis
    1. Identification of (statistically) significant peaks

3. Supervised multivariate analysis (e.g. PLS-DA)
    1. Identification of (statistically) significant patterns of peaks

Thank you for your attention!

Jasper Engel (jasper.engel@wur.nl)

WAGENINGEN
UNIVERSITY & RESEARCH