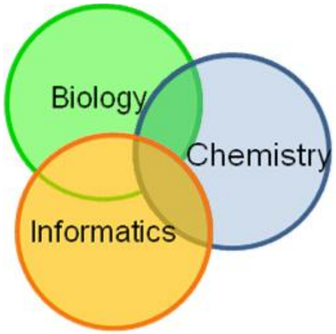


Principal Component Analysis (PCA) of metabolomic sample processing methods

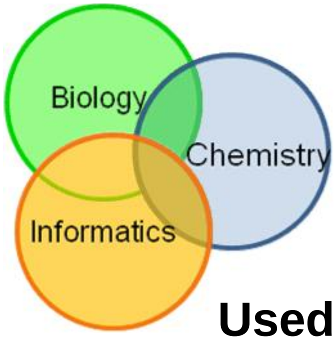
Principal Components Analysis

Goal:

Use PCA to identify the major modes of variance

Topics:

1. Principal component number selection
2. Data pretreatment
3. PCA results visualization



Principal Components Analysis

Used DATA: Pumpkin data 1.csv

Principal Components Analysis

Steps

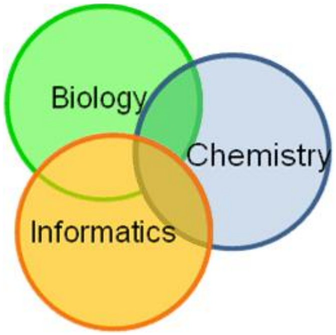
1.Calculate a PCA model
2.Select optimal model principal component (PC)
3.Overview PCA scores and loadings plots
4.Repeat steps 1-2 using data centering and scaling

Visualize:

1.Sample scores annotated by extraction and treatment
2.Leverage and DmodX (distance from model plane)
3.Variable loadings and biplots

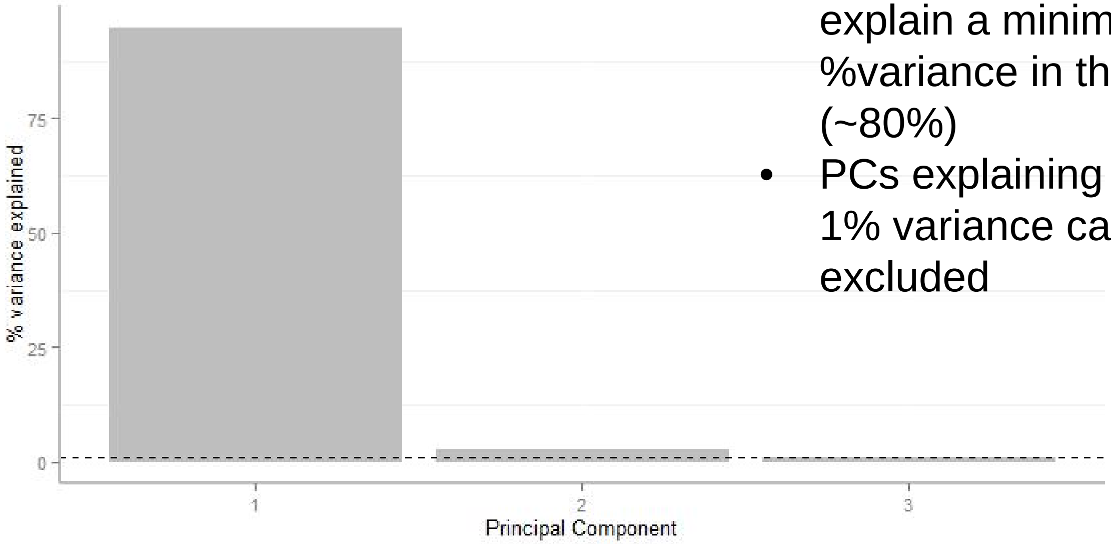
Exercise:

1.How many PCs are needed to capture 80% variance for raw data and scaled data?
2.Are their any moderate or extreme outliers?
3.What variables contribute most to the variance for raw and scaled data?

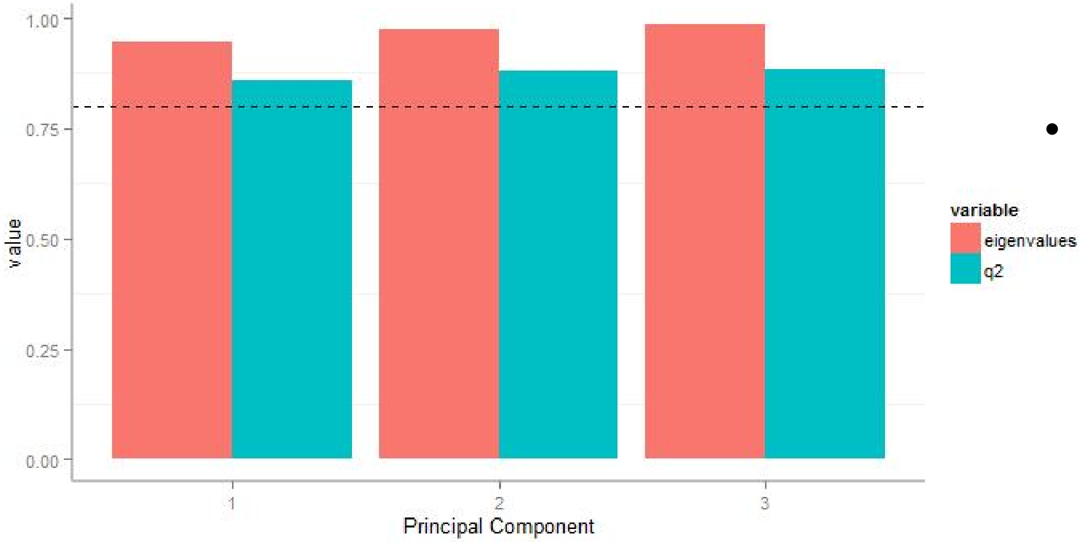


PCA Variance Explained (raw data)

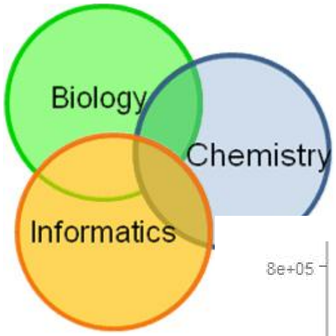
Principal Components Analysis



- PCs can be selected to explain a minimum %variance in the data (~80%)
- PCs explaining below 1% variance can be excluded

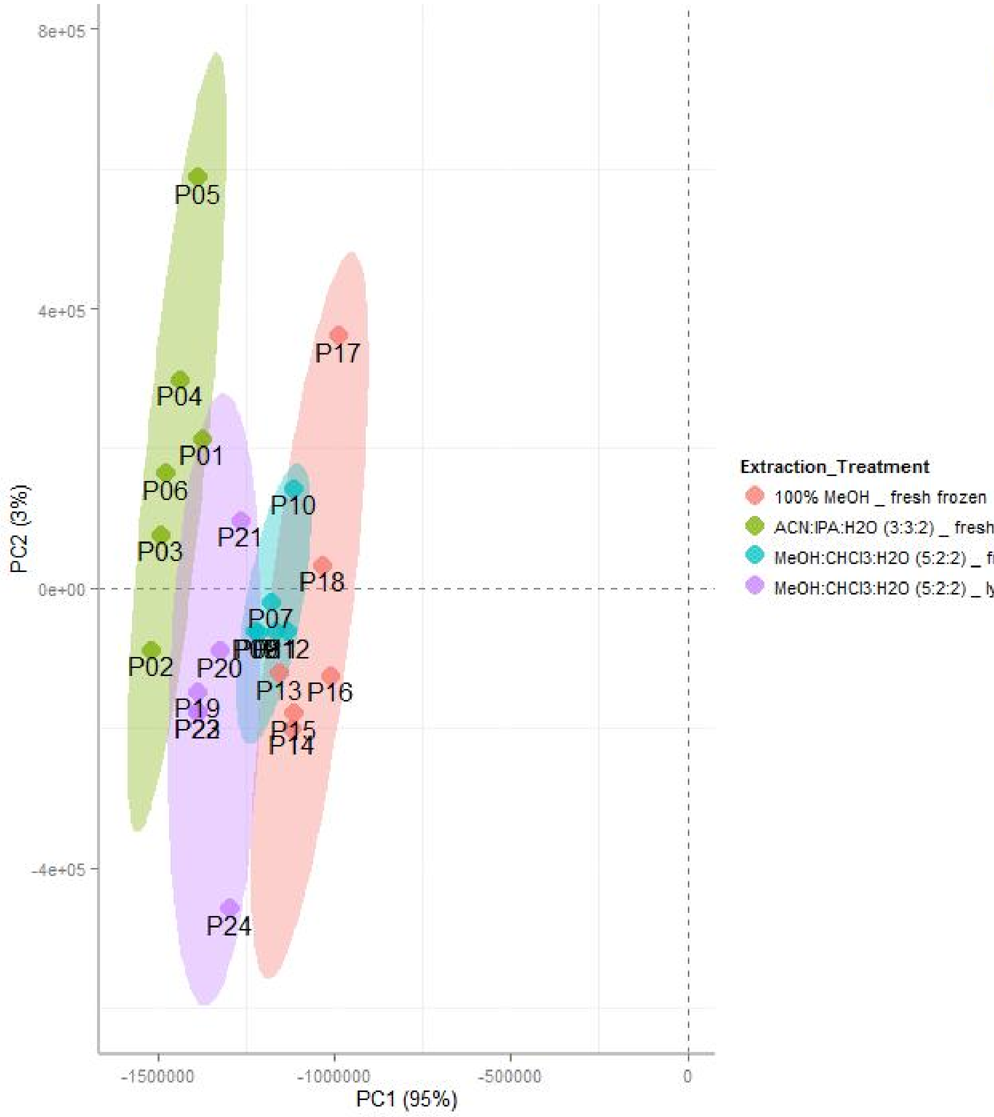


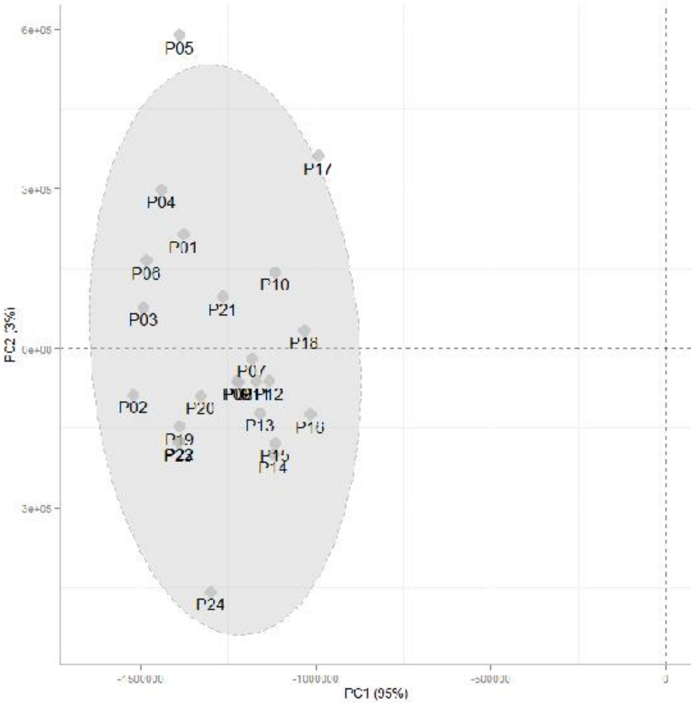
- q2 is the cross-validated PCA prediction of left out data



PCA Scores (raw data)

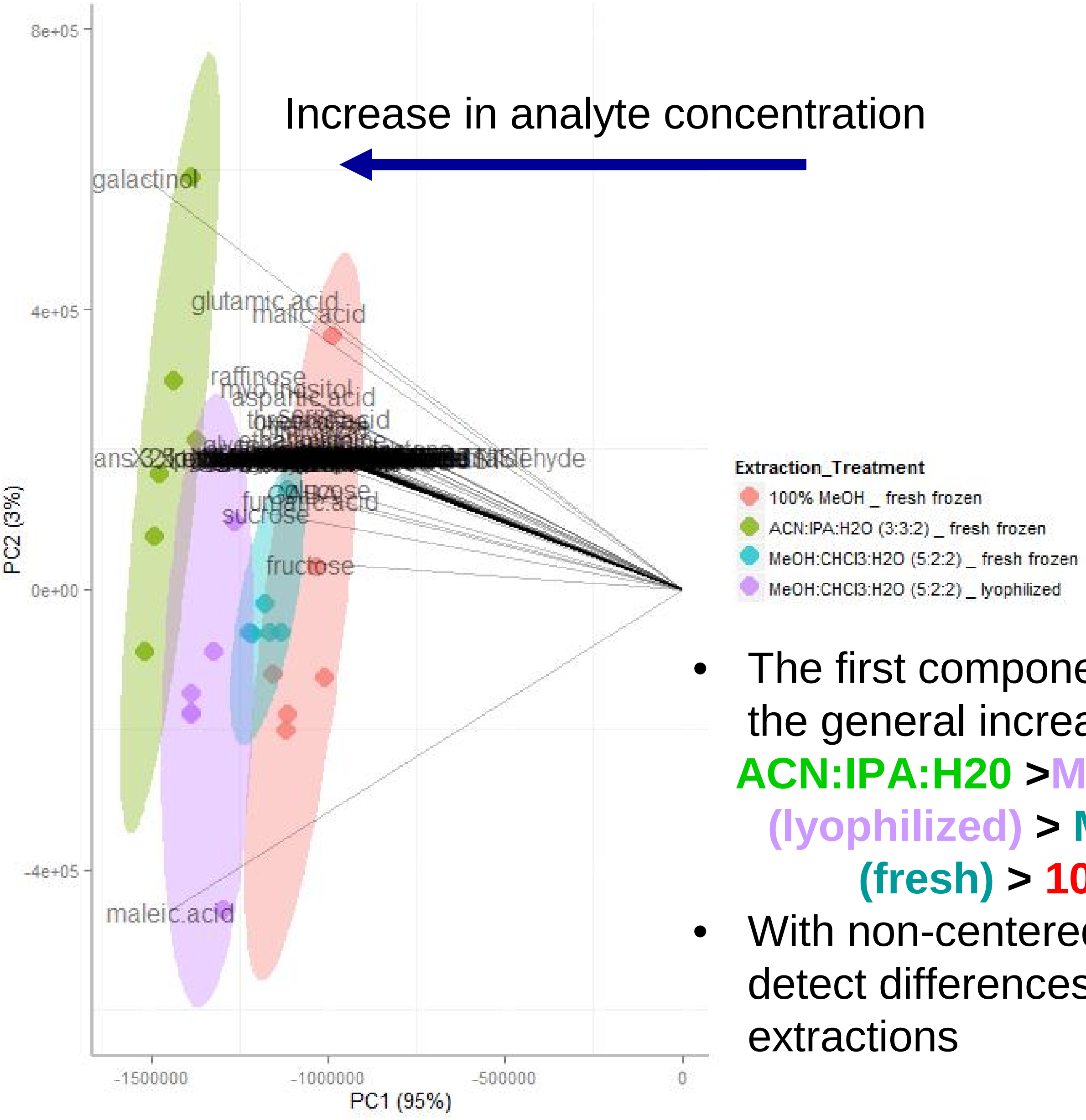
Principal Components Analysis





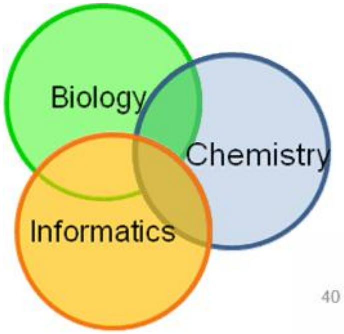
- Hotelling’s T^2 ellipse shows 95% CI for bivariate normal distribution
- Samples lying outside of the ellipse could be outliers

PCA Biplot (raw data)



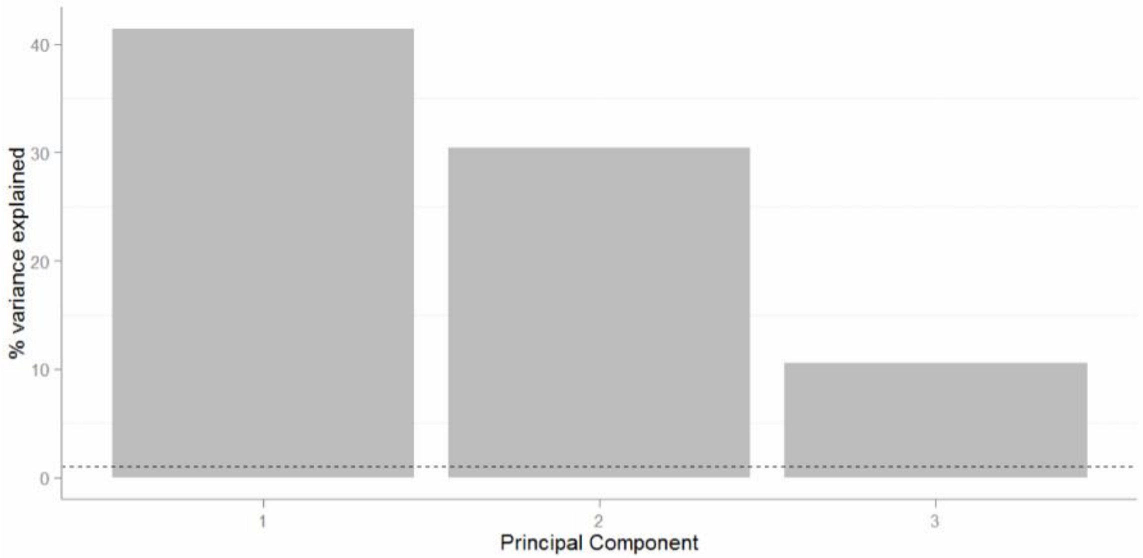
Increase in analyte concentration

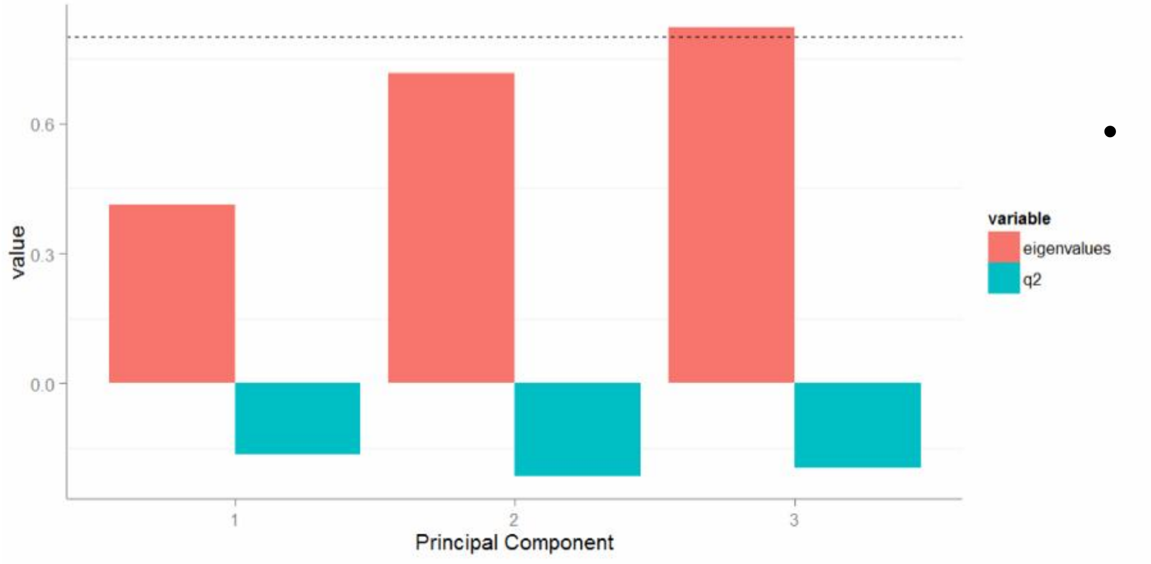
- The first component (PC1) captures the general increase in all analytes **ACN:IPA:H20 >MeOH:CHCL3:H20 (lyophilized) > MeOH:CHCL3:H20 (fresh) > 100% methanol**
- With non-centered data it is hard to detect differences between extractions



PCA Variance Explained (mean centered)

Principal Components Analysis



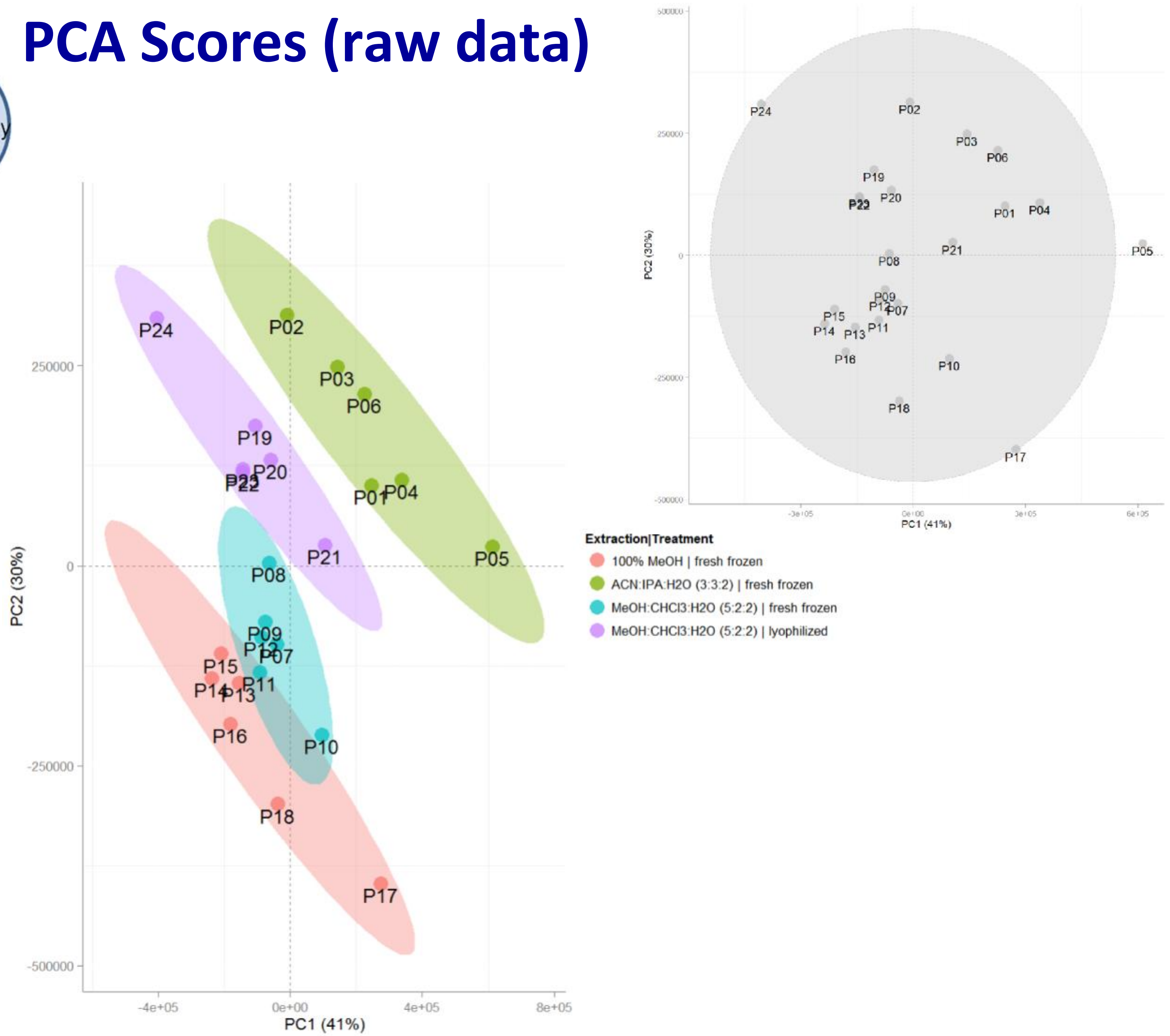


- q2 is low due to instability in the mean of each analyte

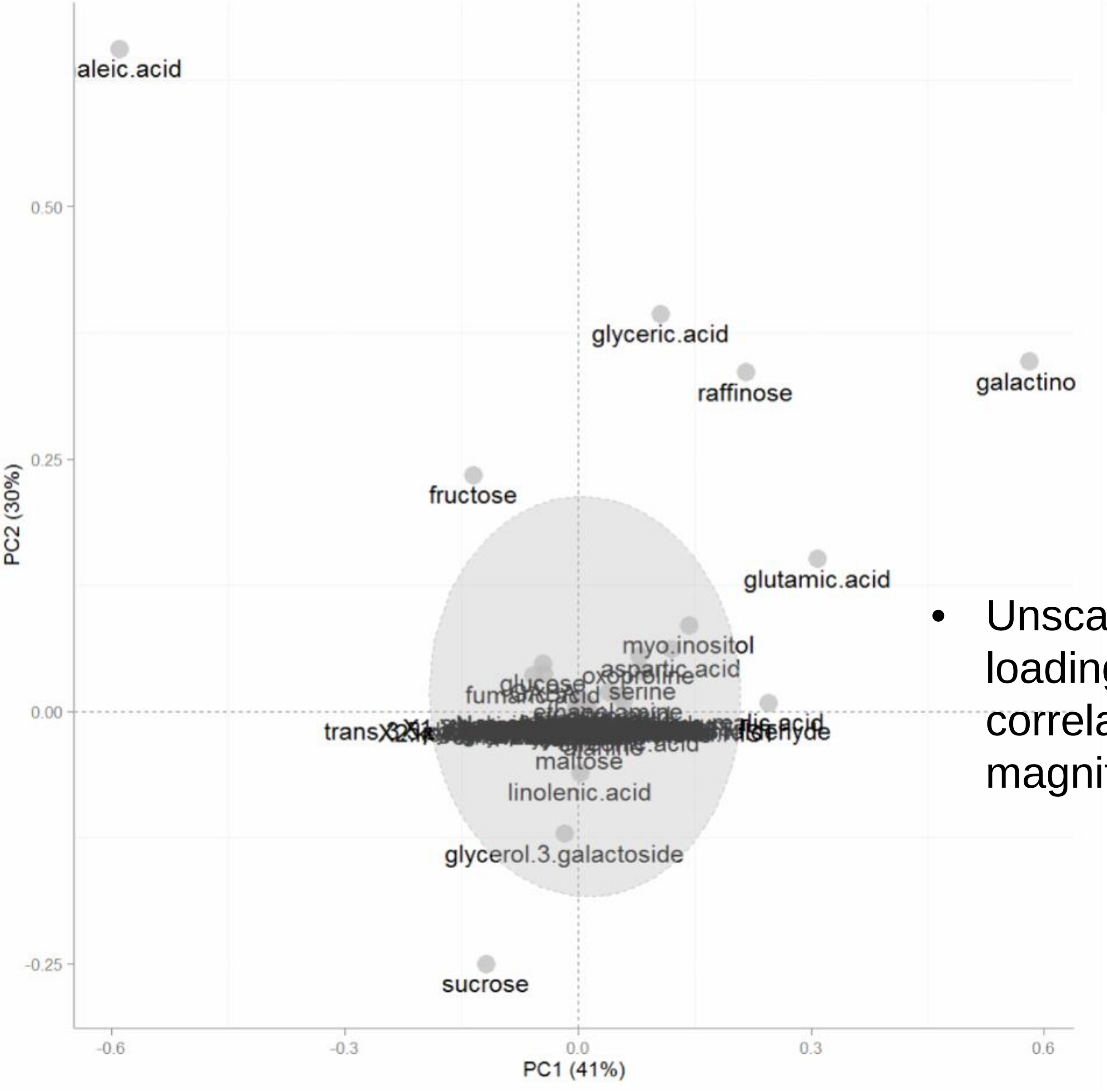
Biology Chemistry Informatics

PCA Scores (raw data)

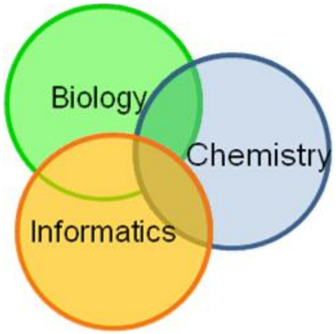
Principal Components Analysis



PCA Loadings (mean centered)

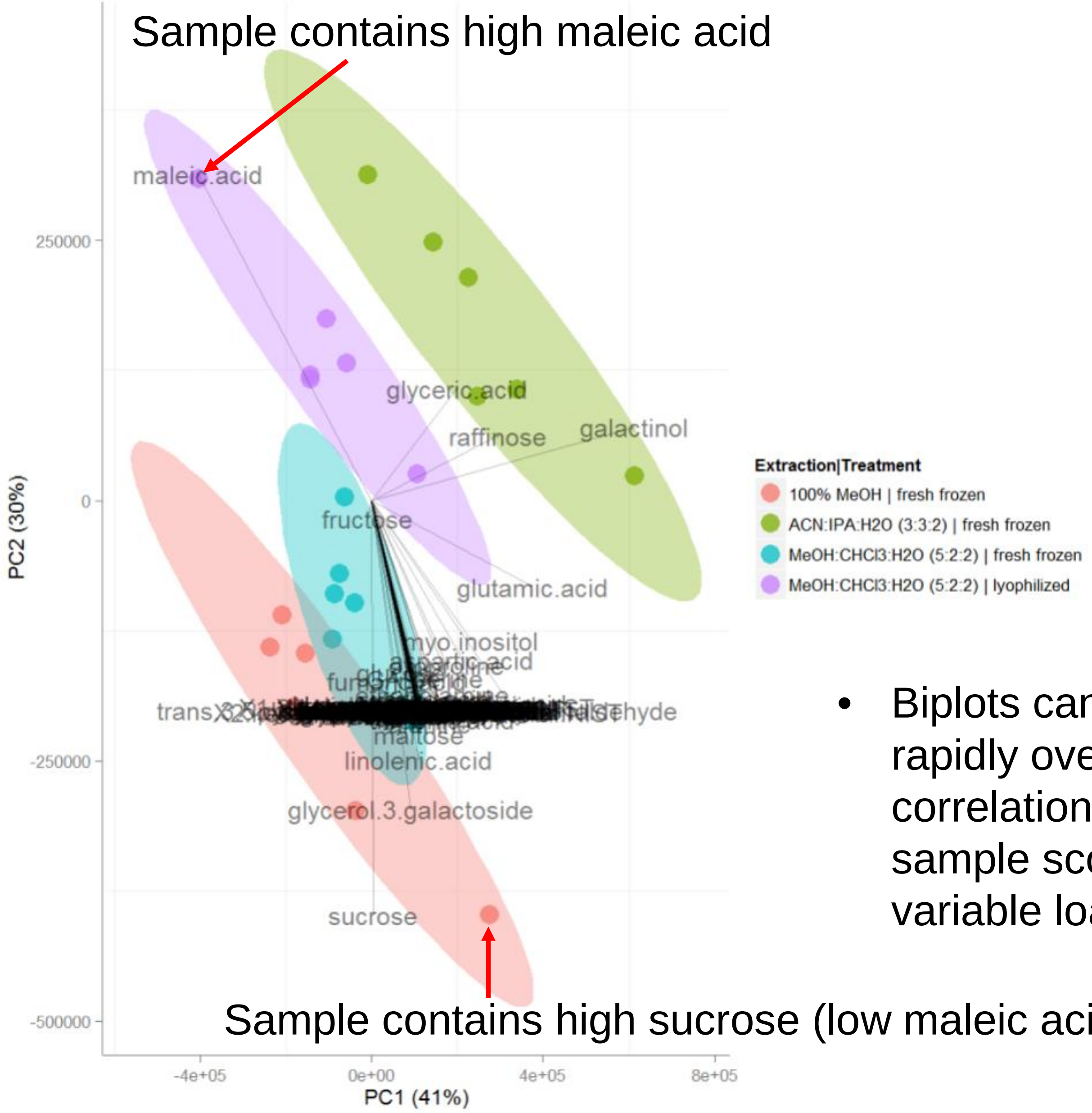


- Unscaled data PCA loadings are highly correlated with magnitude



PCA Biplot (raw data)

Principal Components Analysis

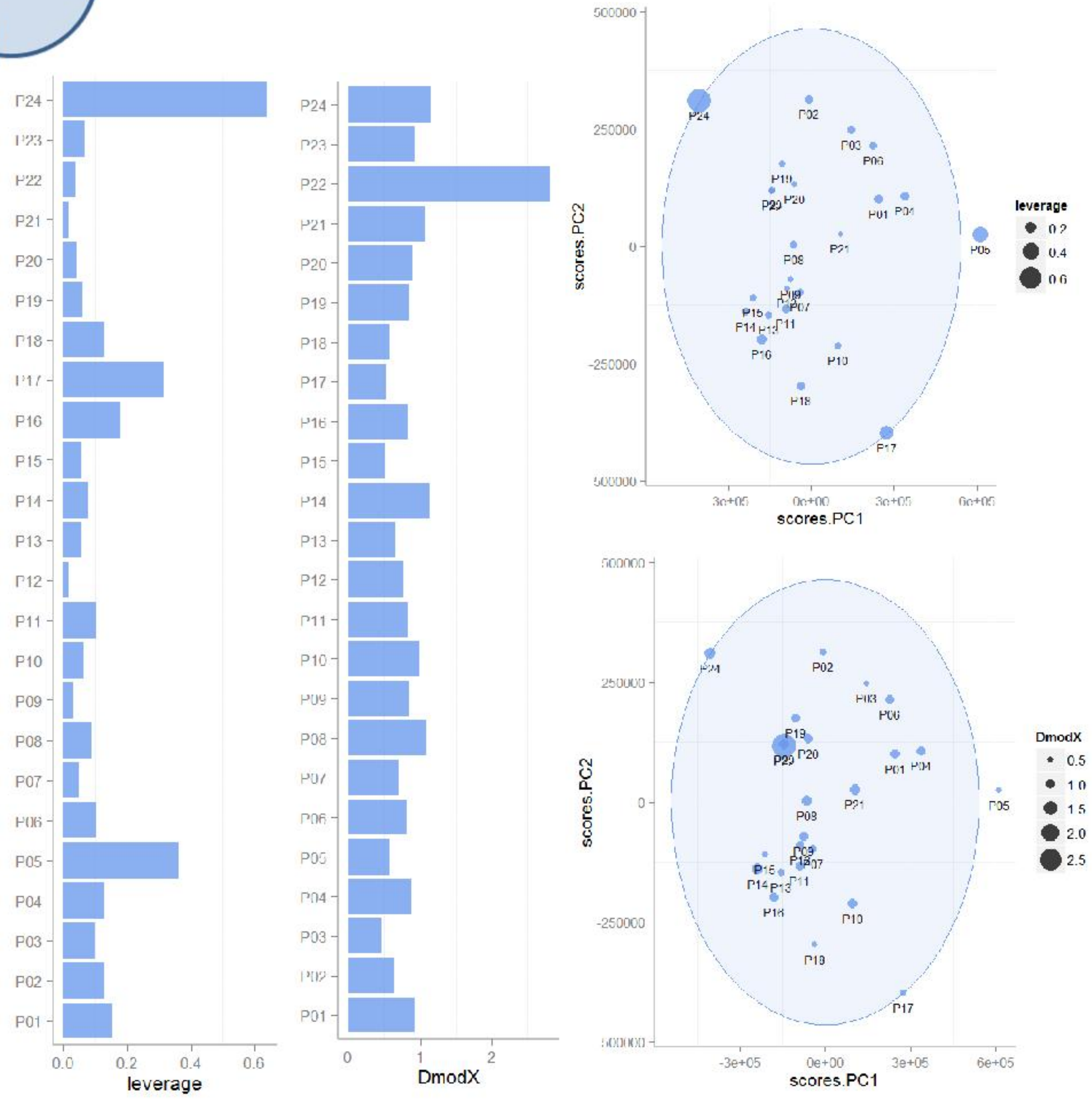


Sample contains high maleic acid

Sample contains high sucrose (low maleic acid)

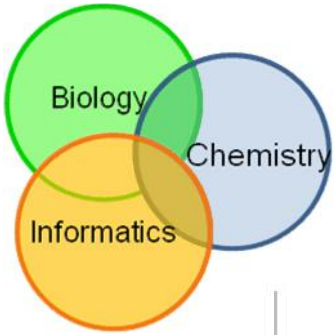
- Biplots can be used to rapidly overview the correlation between sample scores and variable loadings

PCA Leverage and DmodX (mean centered)



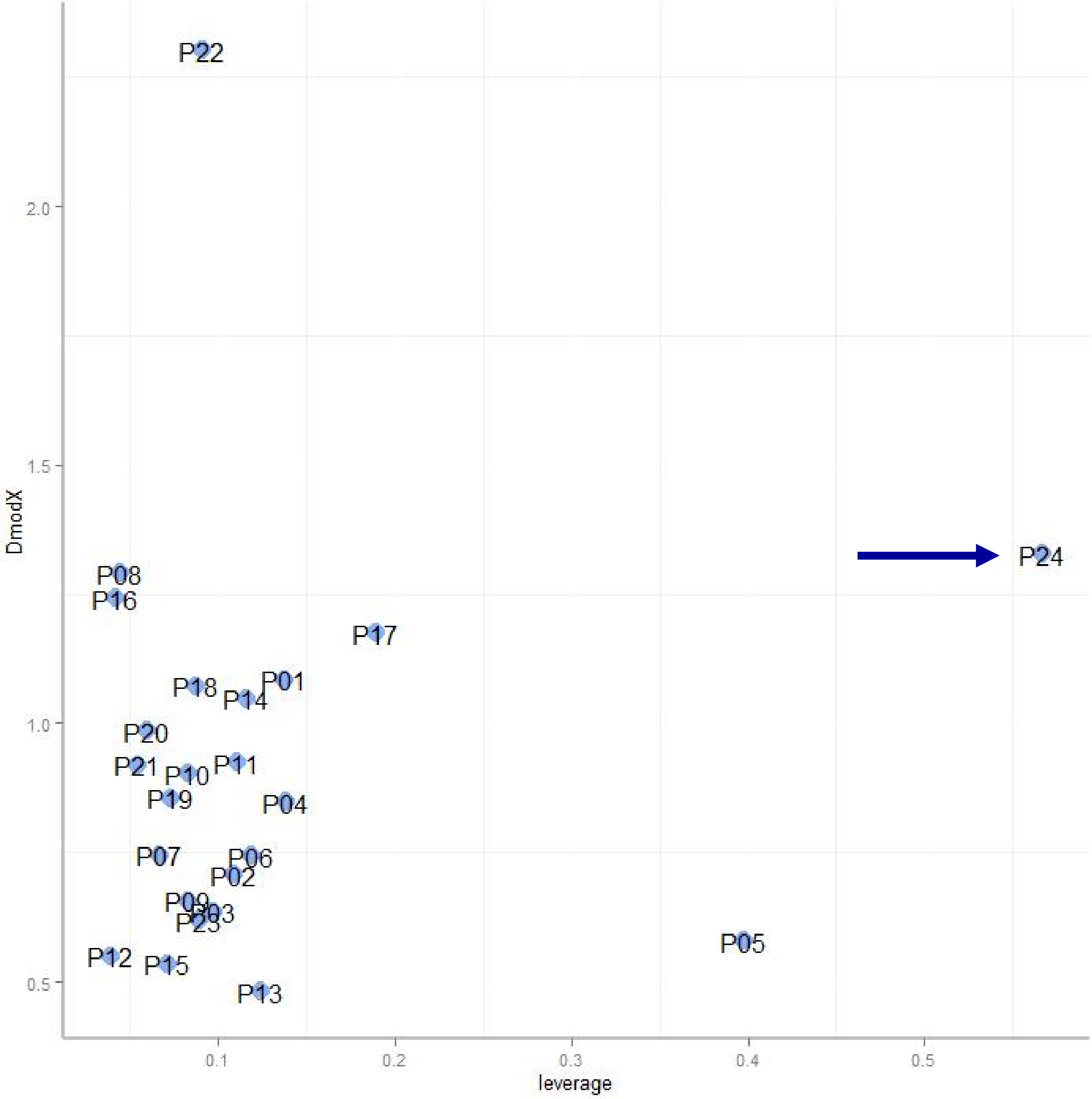
Leverage is the distance to samples center in the PCA plane (extreme outliers)

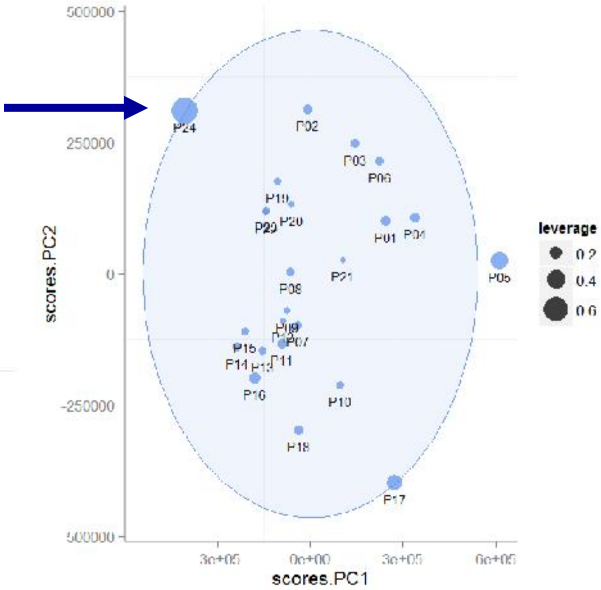
Distance to model X (**DmodX**) is the orthogonal distance to the PCA plane (moderate outliers)



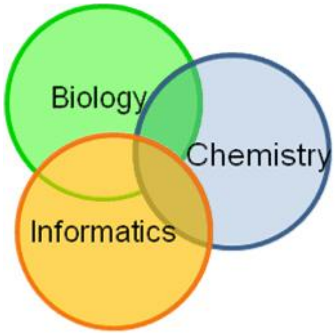
Detecting outliers

Principal Components Analysis



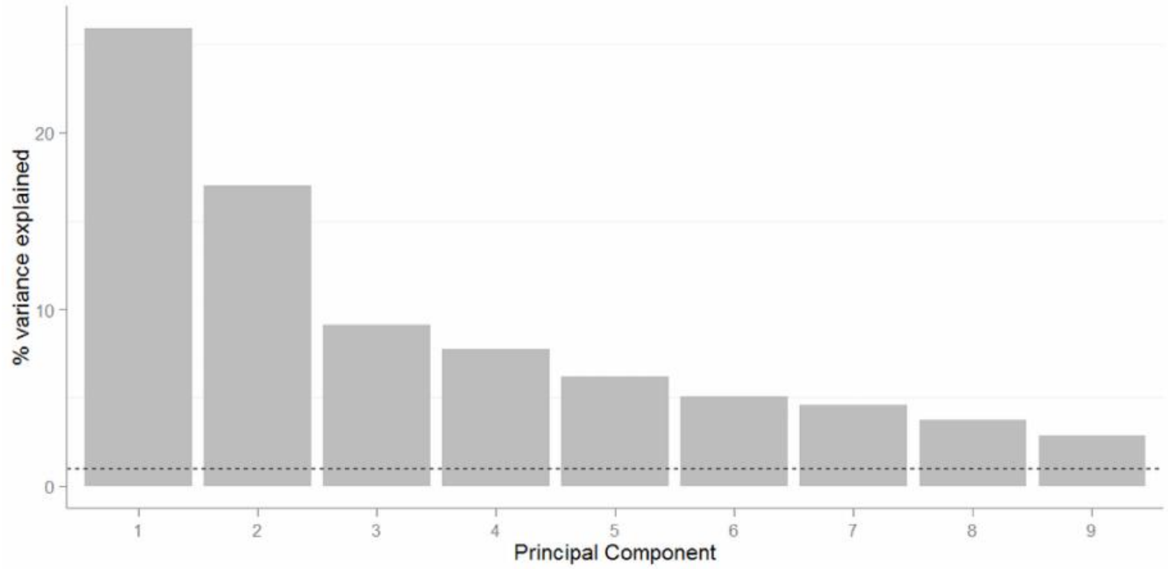


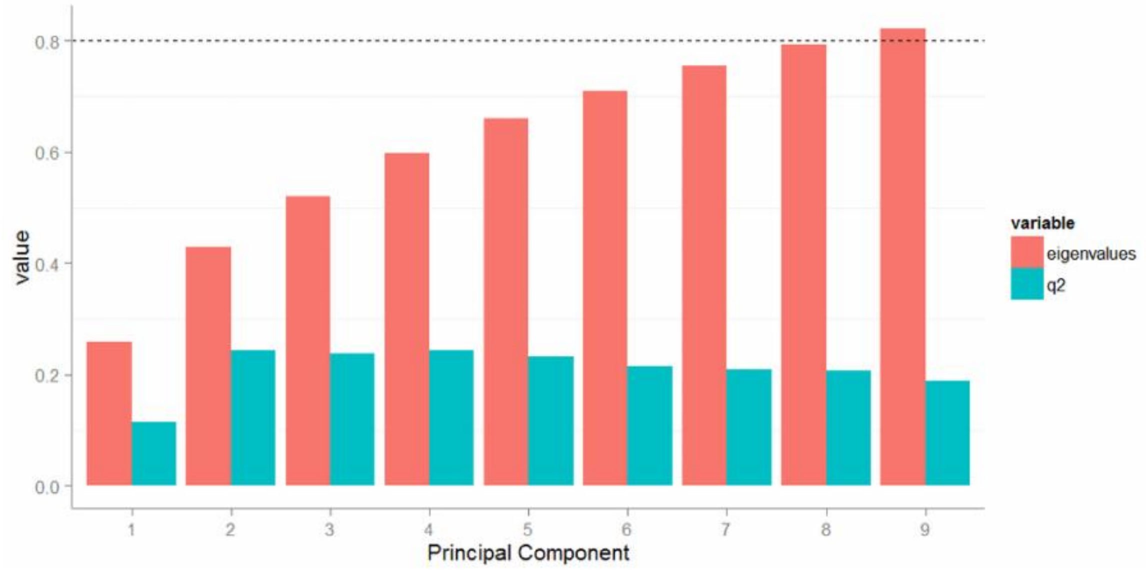
Sample with both <u>high leverage and DmodX</u> are likely <u>outliers</u> and can negatively effect statistical tests and predictive modeling

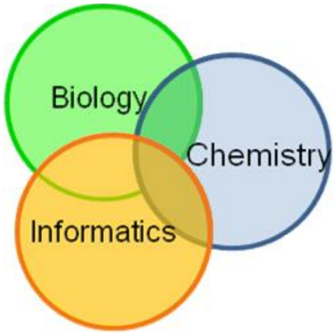


PCA Variance Explained (autoscaled)

Principal Components Analysis

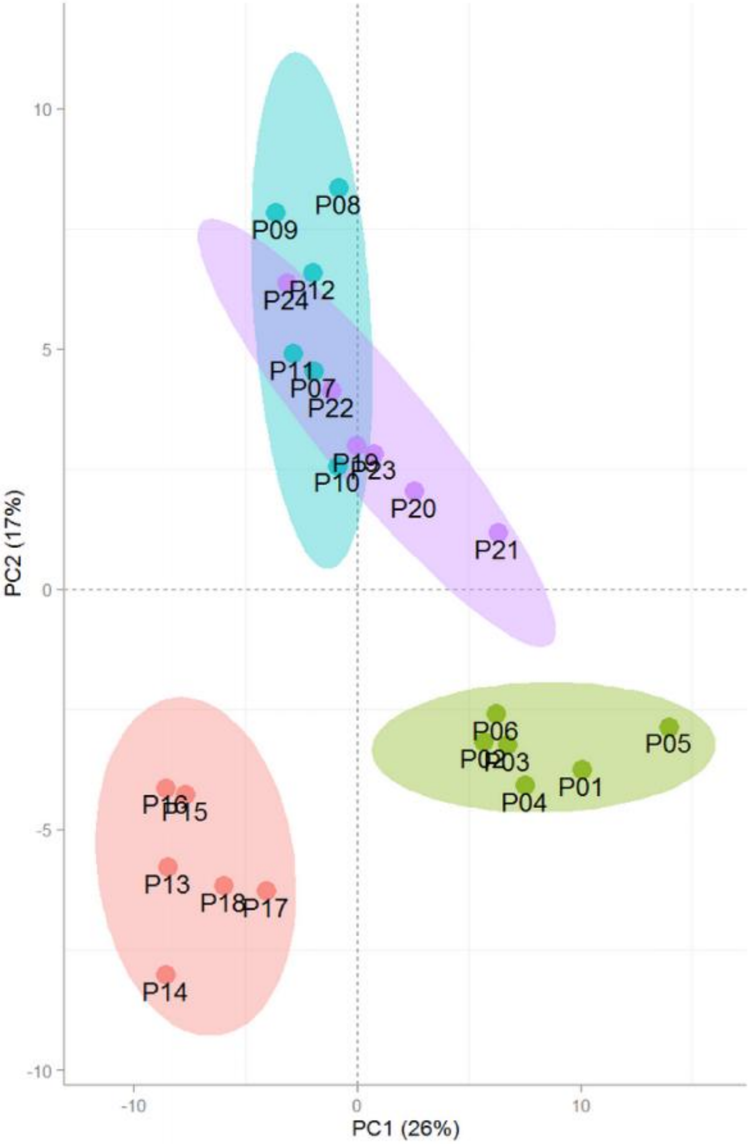


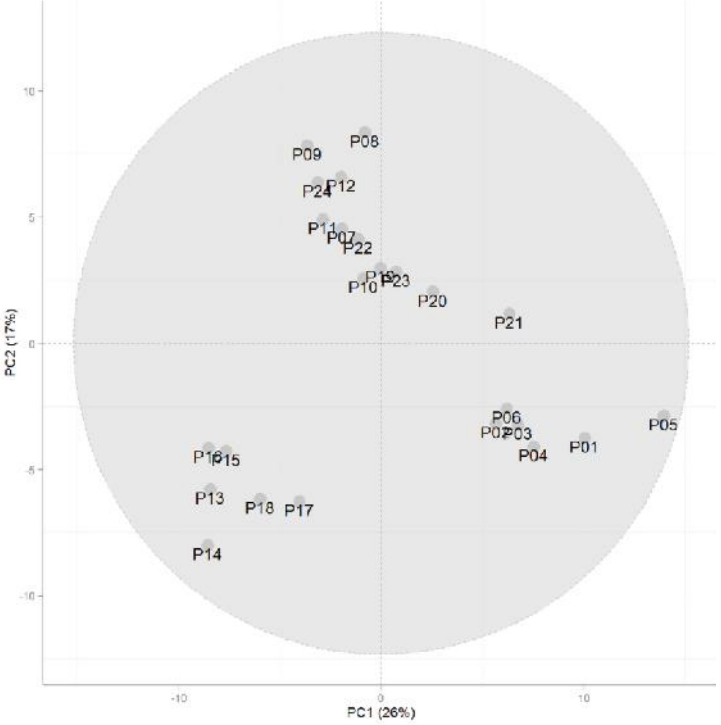




Principal Components Analysis

PCA Scores (autoscaled)





Extraction|Treatment

- 🟥 100% MeOH | fresh frozen
- 🟩 ACN:IPA:H2O (3:3:2) | fresh frozen
- 🟦 MeOH:CHCl3:H2O (5:2:2) | fresh frozen
- 🟪 MeOH:CHCl3:H2O (5:2:2) | lyophilized

- Loadings on PC1 describe differences due to extraction
- Loadings on PC2 describe differences due to drying

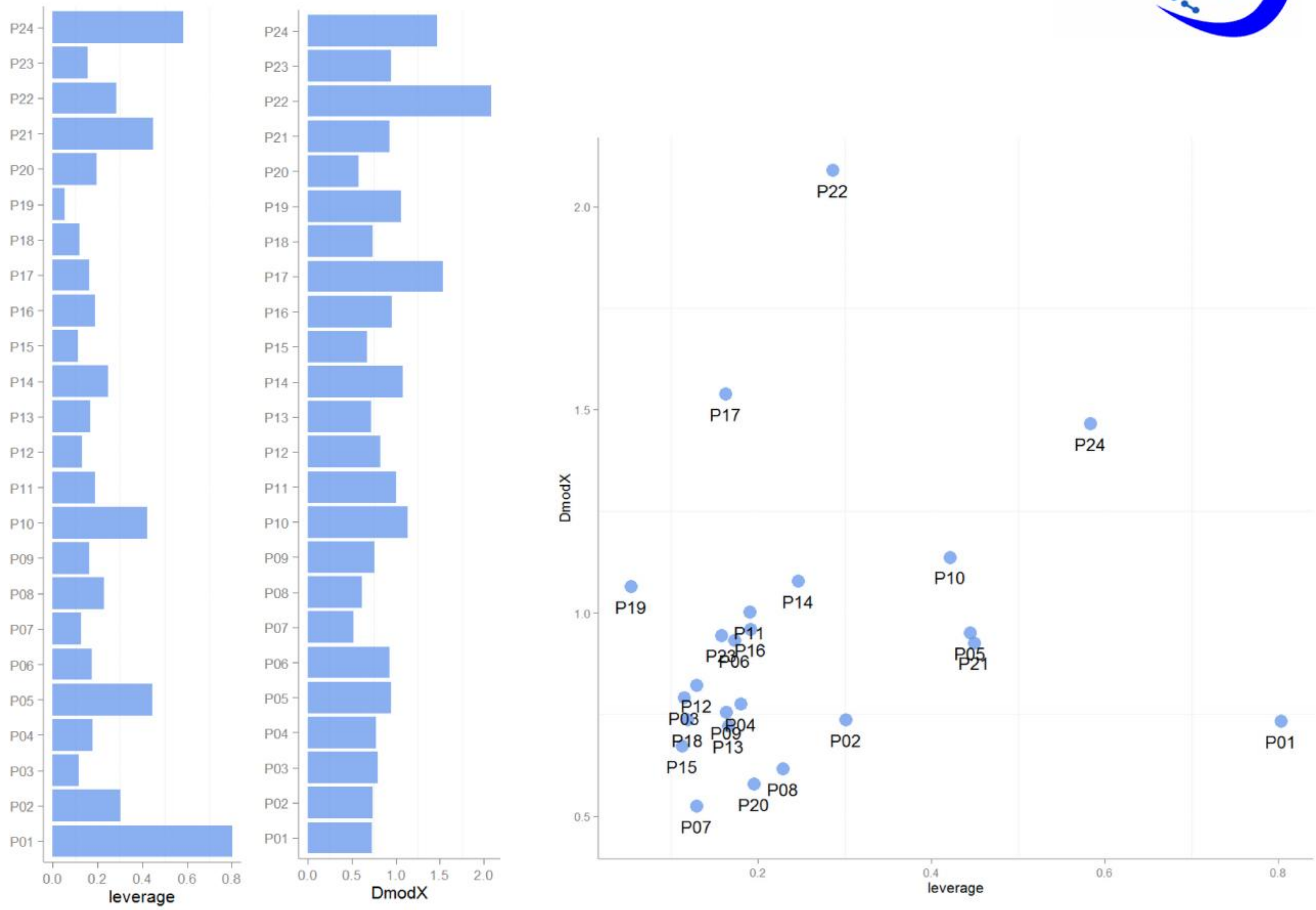
Biology

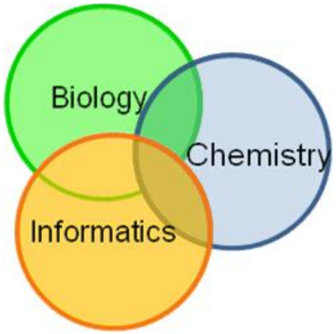
Chemistry

Informatics

PCA Leverage and DmodX (autoscaled)

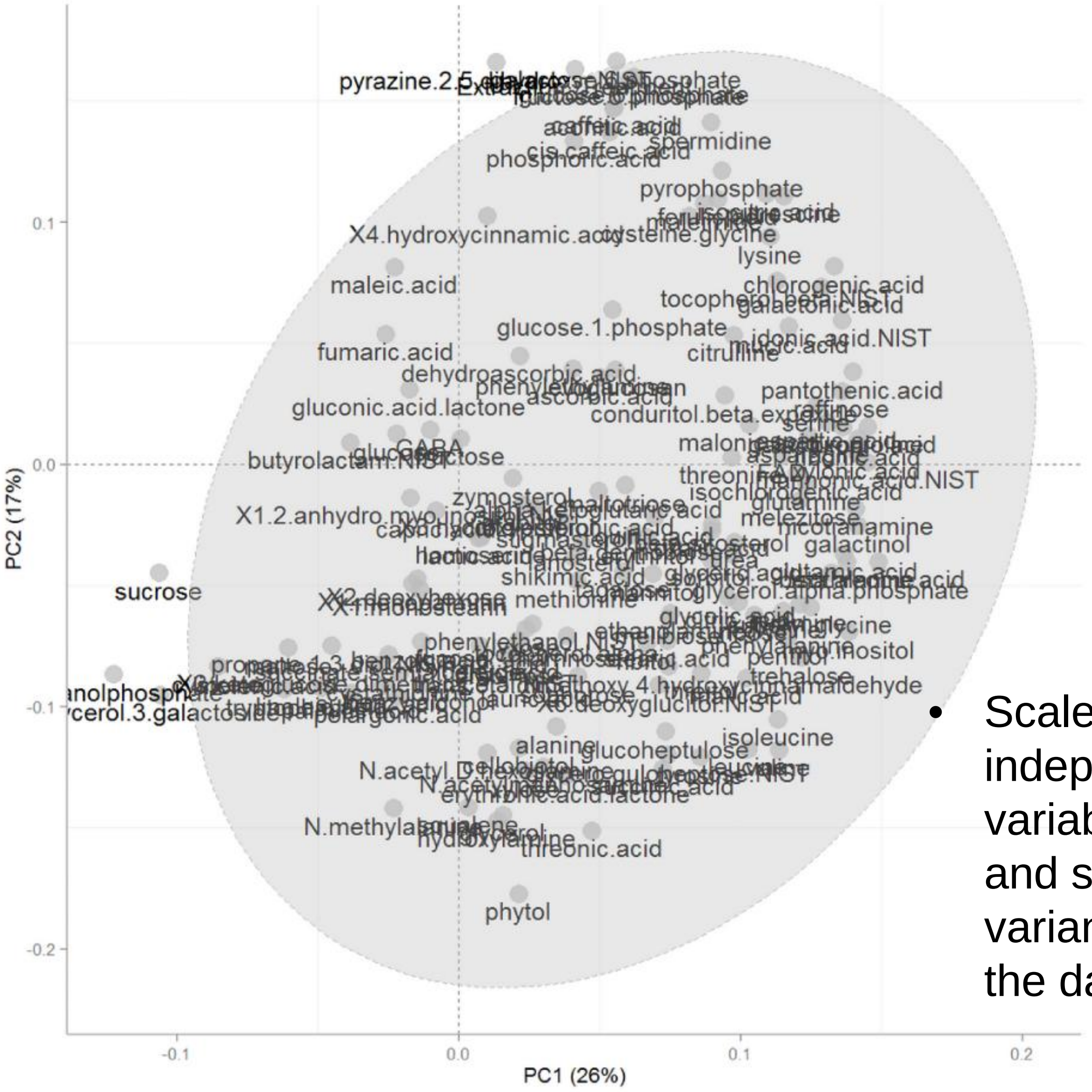
Principal Components Analysis



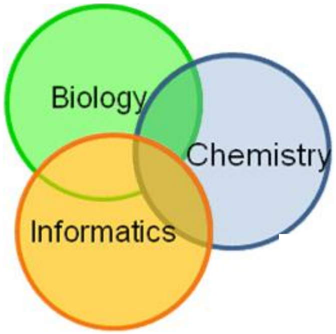


PCA Loadings (autoscaled)

Principal Components Analysis

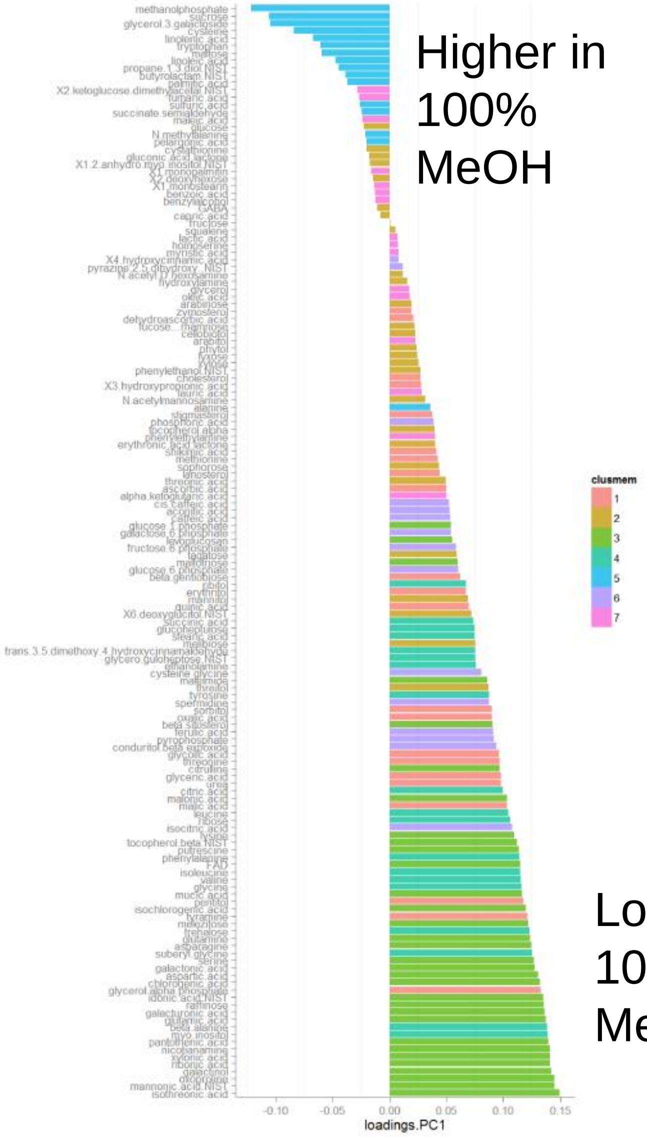


- Scaled loadings are independent of variable magnitude and show a rich variance structure of the data



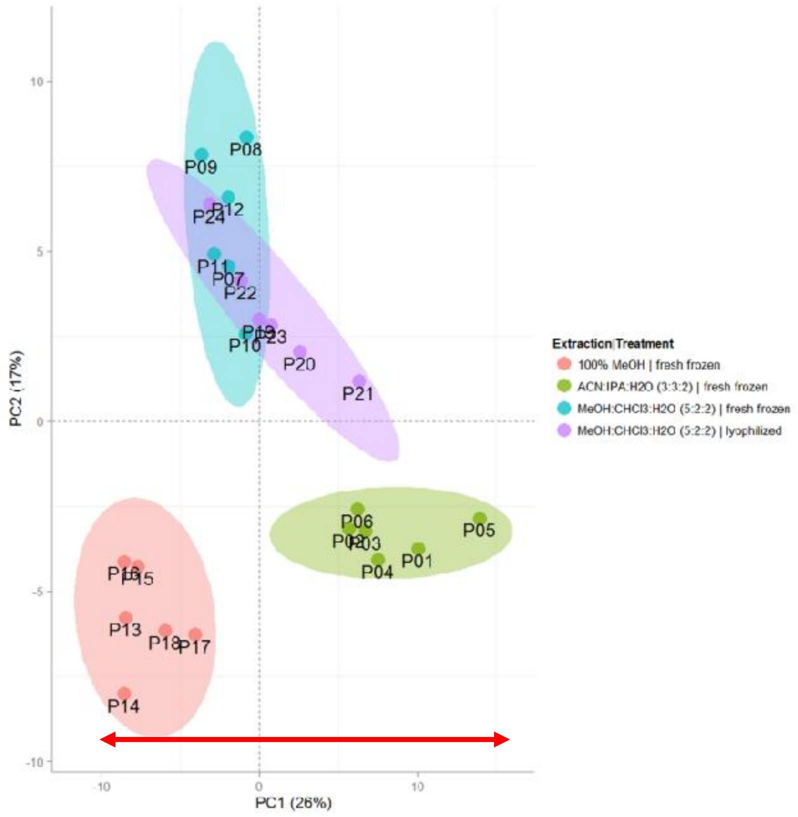
Relationship between scores and loadings (autoscaled)

Principal Components Analysis

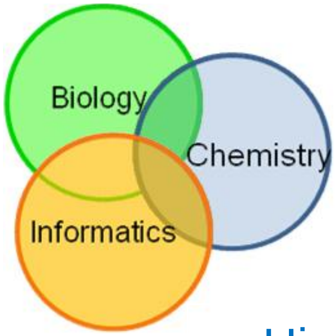


Higher in
100%
MeOH

Lower in
100%
MeOH

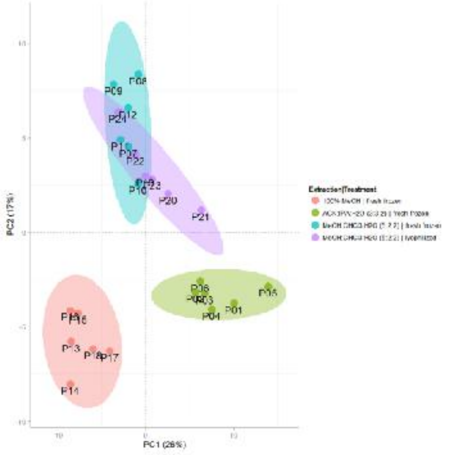


Extraction



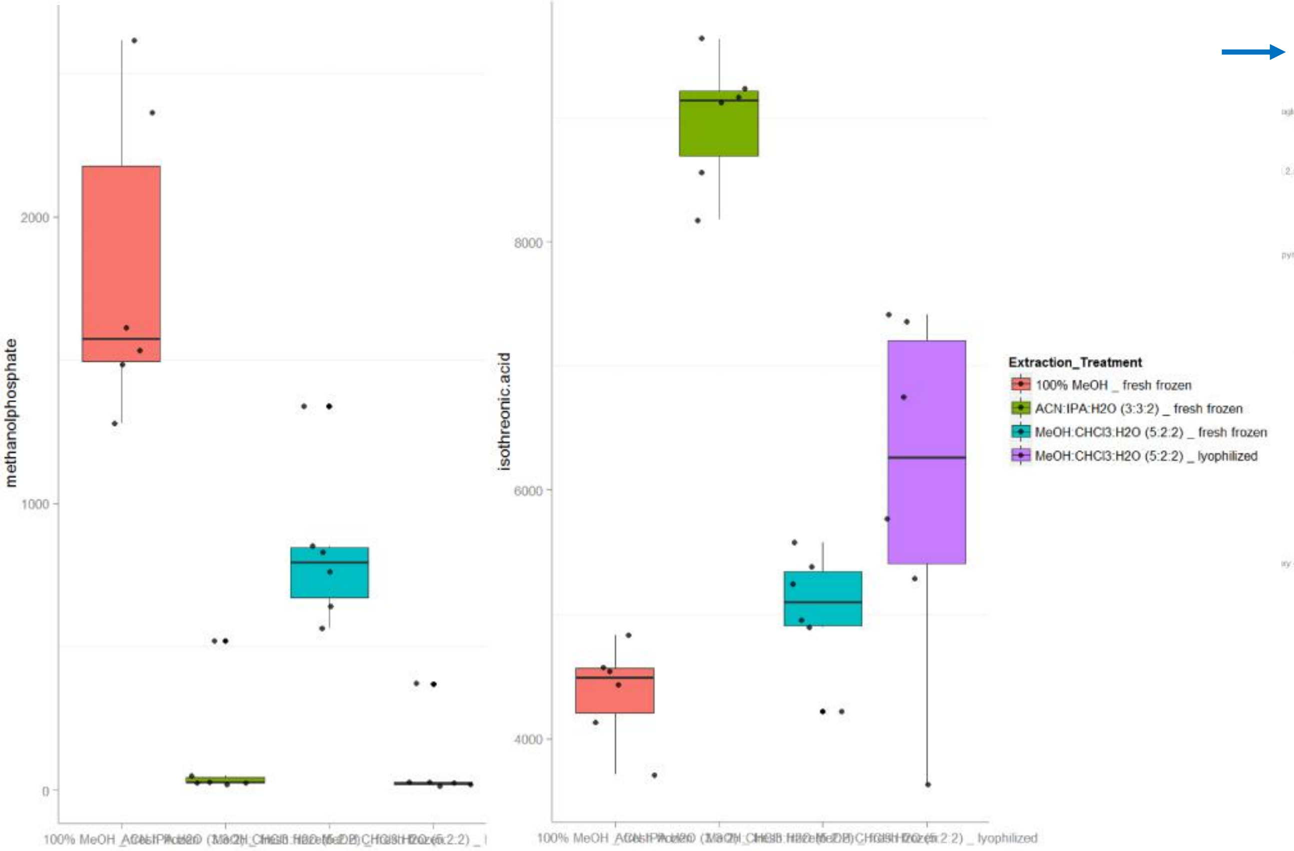
Loadings and Scores

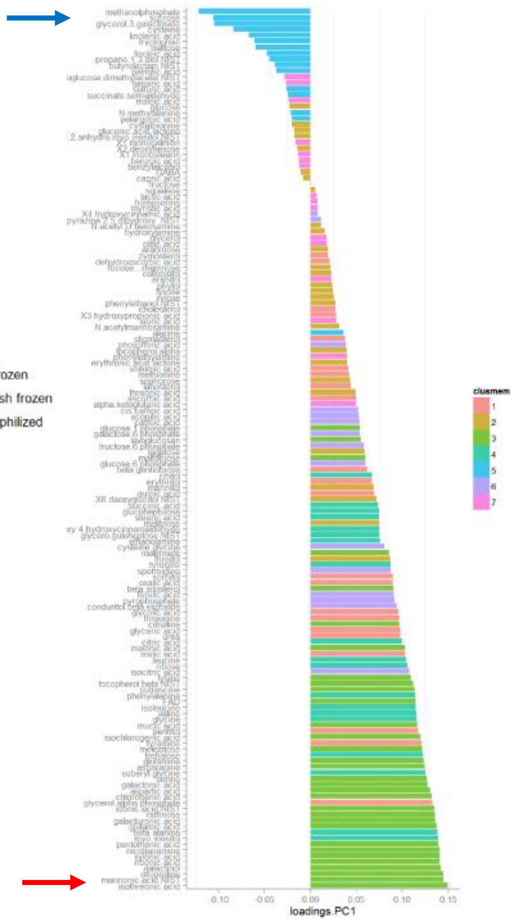
Principal Components Analysis



Highest negative loading on PC1

Highest positive loading on PC1


methanolphosphate
isothreonic.acid
100% MeOH _ fresh frozen
ACN:IPA:H2O (3:3:2) _ fresh frozen
MeOH:CHCl3:H2O (5:2:2) _ fresh frozen
MeOH:CHCl3:H2O (5:2:2) _ lyophilized
Extraction_Treatment
100% MeOH _ fresh frozen
ACN:IPA:H2O (3:3:2) _ fresh frozen
MeOH:CHCl3:H2O (5:2:2) _ fresh frozen
MeOH:CHCl3:H2O (5:2:2) _ lyophilized


methanolphosphate
isothreonic acid
clusmem
loadings.PC1