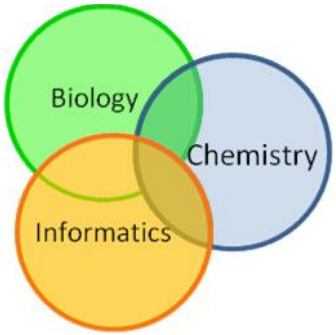


Partial Least Squares (O-/PLS/-DA) modeling of metabolomic sample processing methods

Partial Least Squares (O-/PLS/-DA)

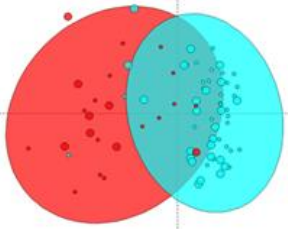
Goal:

Use PLS to identify metabolites which best discriminate (most different between) sample processing methods
(Used DATA: Pumpkin data 1.csv)

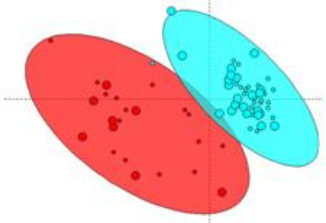
Topics:

1. Model Selection
2. Results visualization
3. Feature Selection
4. Validation

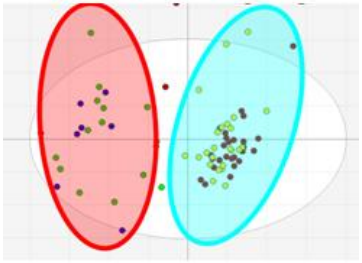
PCA

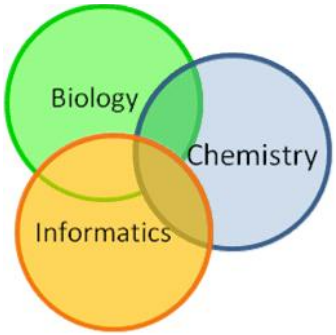


PLS-DA



OPLS-DA





Partial Least Squares Modeling Discriminant Analysis (PLS-DA)

Partial Least Squares (O-/PLS/-DA)

Used DATA: Pumpkin data 1.csv

Steps

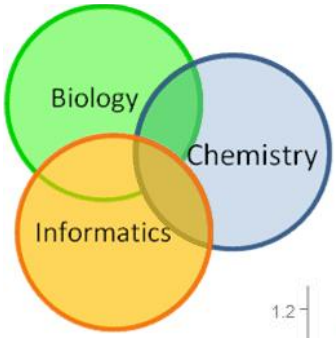
1.Calculate a PLS model to discriminate between extraction_treatment methods
2.Select optimal scaling and model latent variable (LV) number
3.Overview PLS scores and loadings plots
4.Validate model
5.Repeat steps 1-4 for an O-PLS model

Visualize:

1.Sample scores annotated by extraction and treatment
2.Variable loadings plot

Exercise:

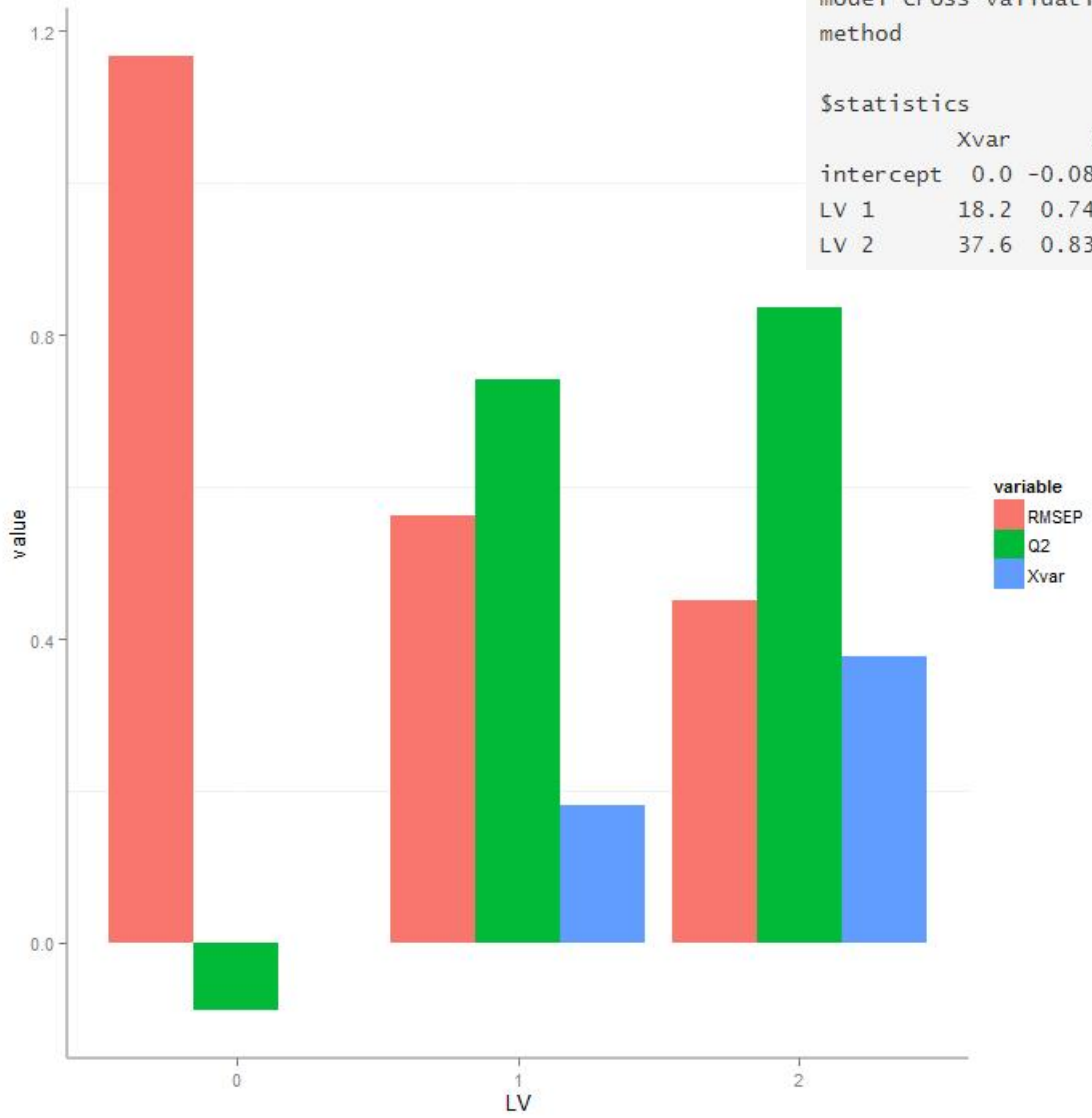
1.How are scores different between PLS and O-PLS?
2.Are there any moderate or extreme outliers?
3.What variables contribute most to the differences between treatments?



Model Performance Statistics

Partial Least Squares (O-/PLS/-DA)

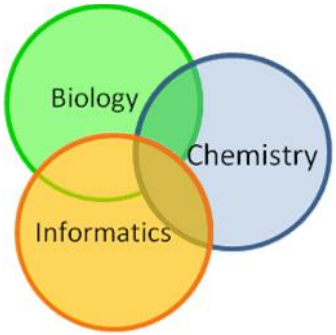
```
                                                    selection
Dependent Variables                   (1)Extraction_Treatment
Latent variables (LVs)                                      2
Orthogonal latent variables (OLVs)                          0
model cross-validation                                    LOO
method                                              oscorespls

$statistics
          Xvar      Q2 RMSEP
intercept  0.0 -0.0888 1.167
LV 1      18.2  0.7415 0.563
LV 2      37.6  0.8367 0.450
```



RMSEP - root mean squared error of prediction (fit to left out set)

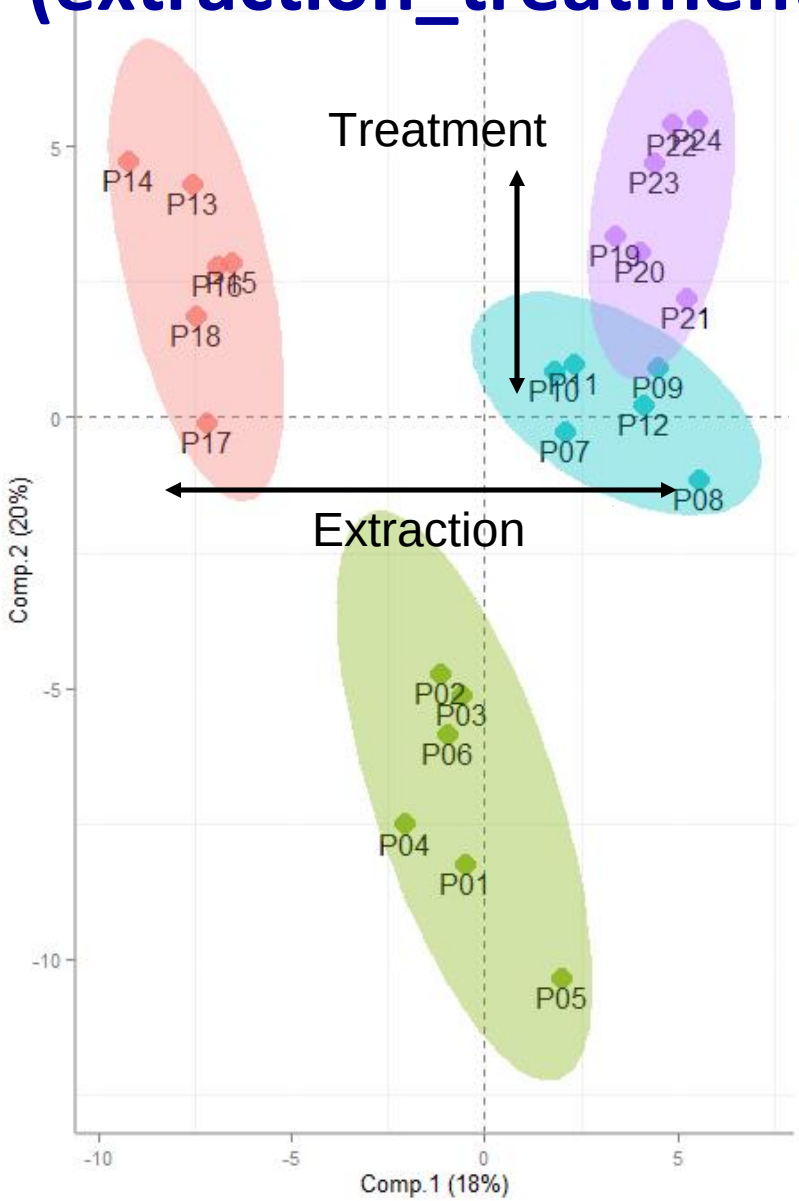
Q^2 - cross-validated fit to the training data

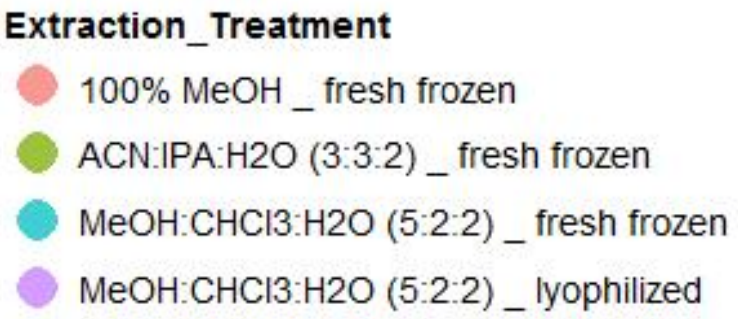
Xvar- explained variance in X (measurements)



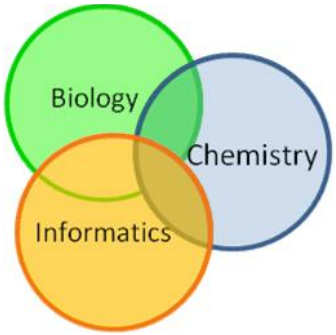
Model Scores (extraction_treatment)

Partial Least Squares (O-/PLS/-DA)





Variance in extraction dominates model



Model Validation

Partial Least Squares (O-/PLS/-DA)

```
$statistics
           Xvar      Q2 RMSEP
intercept   0.0 -0.0888 1.167
LV 1       17.8  0.7042 0.603
LV 2       37.4  0.8034 0.494
```

single model performance estimates

```
Dependent Variables                          (1)Extraction_Treatment
Latent variables (LVs)                                             2
Orthogonal latent variables (OLVs)                                 0
model cross-validation                                           LOO
method                                                     oscorespls
Internal train/test index          20 repetitions generated by random
```

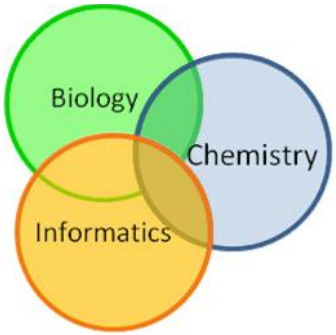
Model properties and validation settings

```
$`Validated Model Performance (Y1)`
                          Xvar               Q2           RMSEP
model           37.16 ± 7.28 0.8772 ± 0.0388 0.5826 ± 0.155
permuted model 30.28 ± 6.41  0.2561 ± 0.265   1.38 ± 0.215
p-value               0.003044       1.904e-09      2.818e-15
```

Based on training/test splitting

Based on training/test splitting and permuted Y

Significance of difference between model performance and permuted NULL distributions



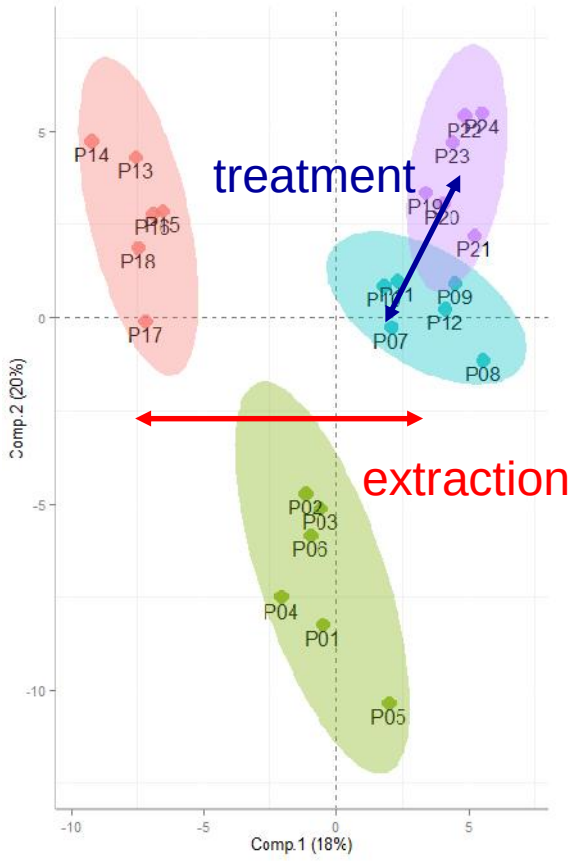
Model Scores Comparison

Extraction_Treatment

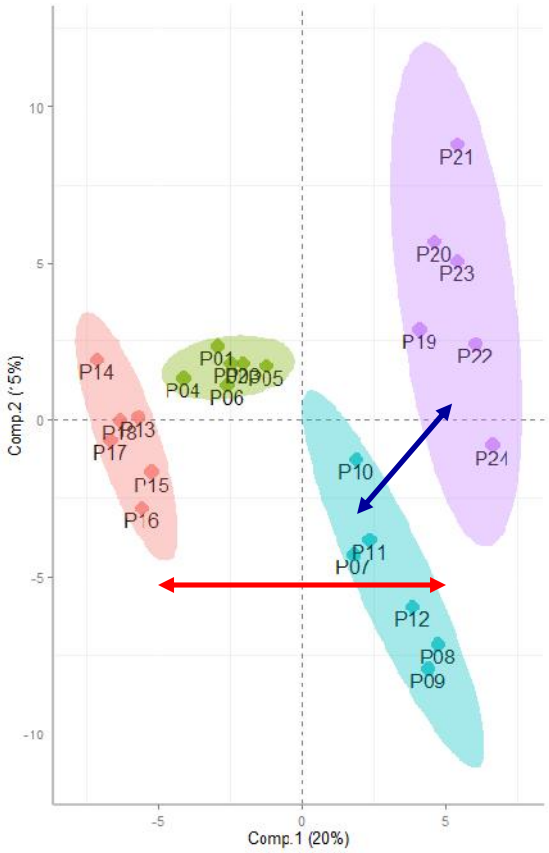
- 100% MeOH _ fresh frozen
- ACN:IPA:H2O (3:3:2) _ fresh frozen
- MeOH:CHCl3:H2O (5:2:2) _ fresh frozen
- MeOH:CHCl3:H2O (5:2:2) _ lyophilized

Partial Least Squares (O-/PLS/-DA)

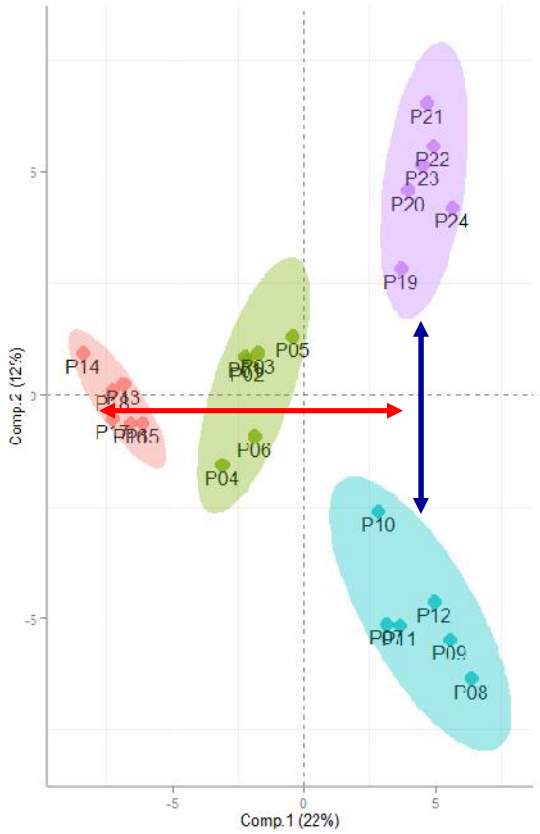
1-Y PLS

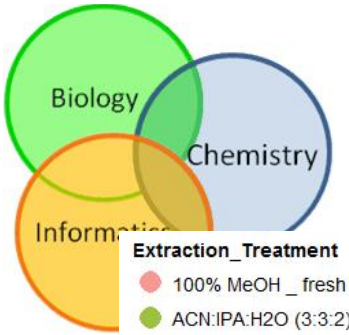


1-Y O-PLS

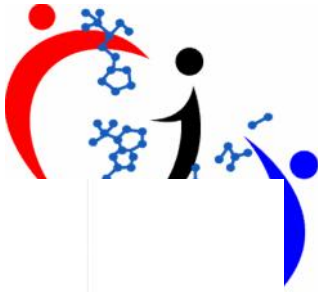


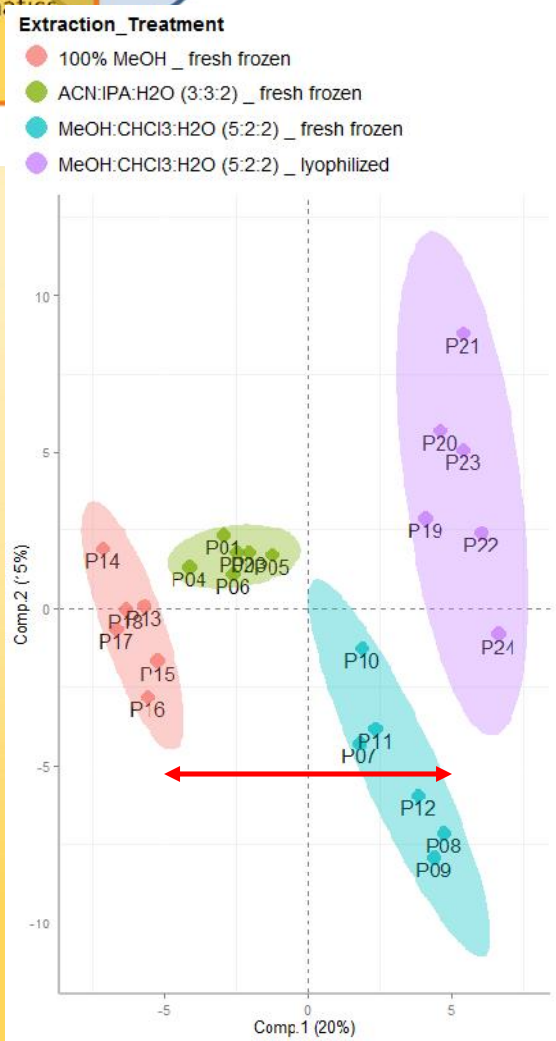
2-Y O-PLS

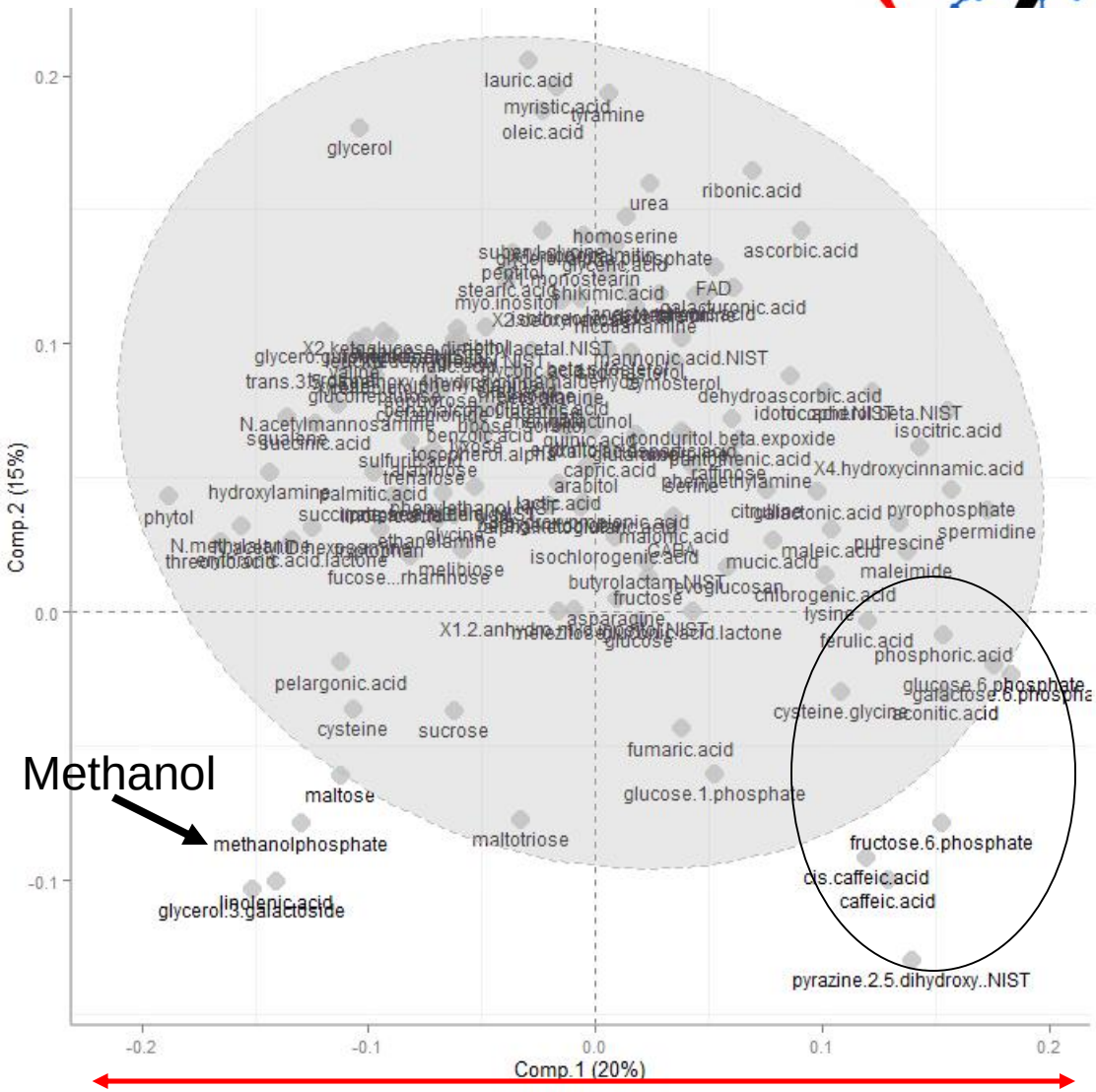




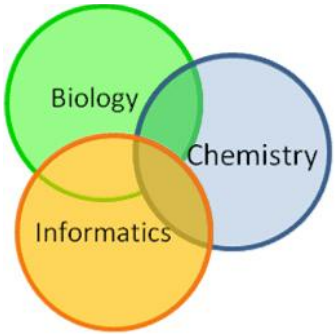
Top Discriminants







Sugar phosphates, polyols



Feature Selection

Partial Least Squares (O-/PLS/-DA)

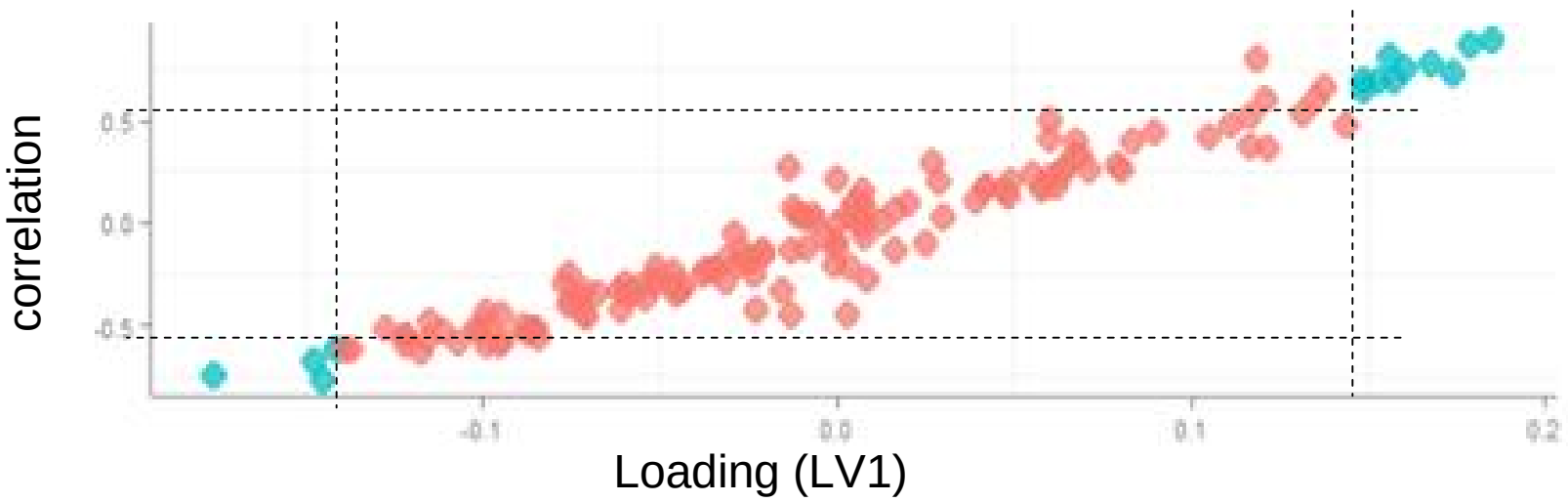
Steps

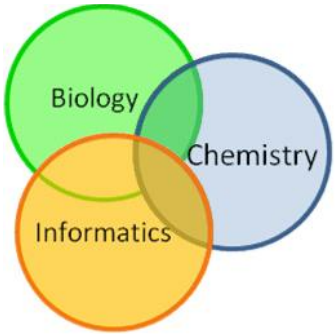
1.Identify top 10% of discriminants for extraction_treatment based on metabolite

- correlation with model scores
- importance (loading, coeffcient, VIP)

Visualize:

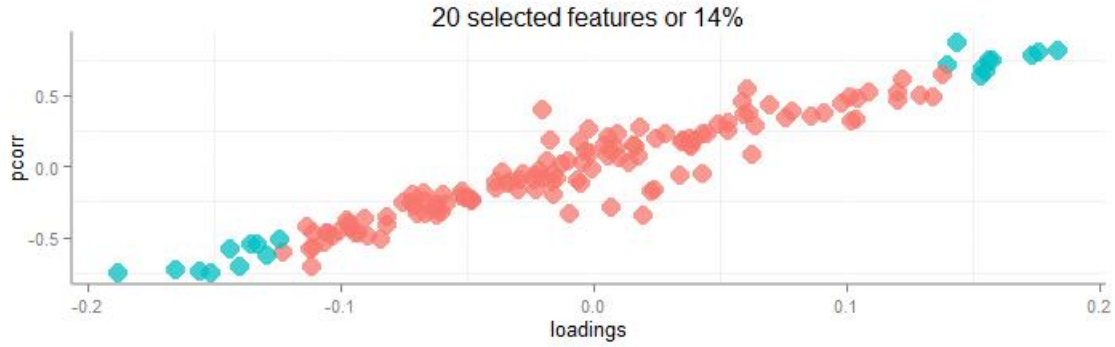
1.Feature selection diagnostics

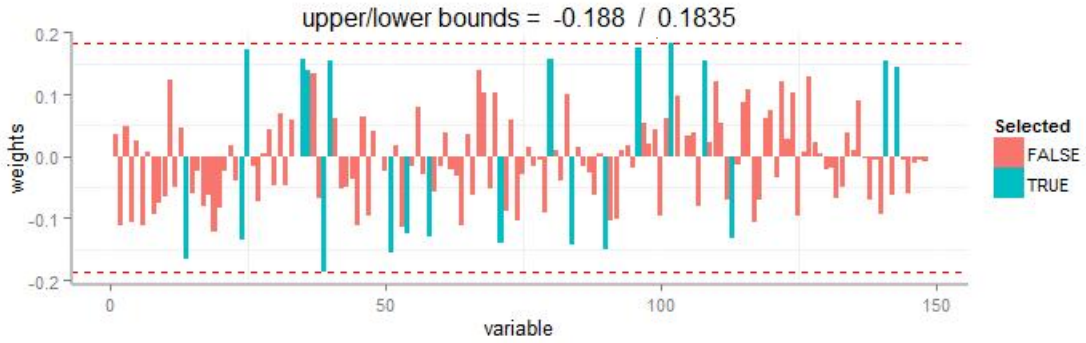




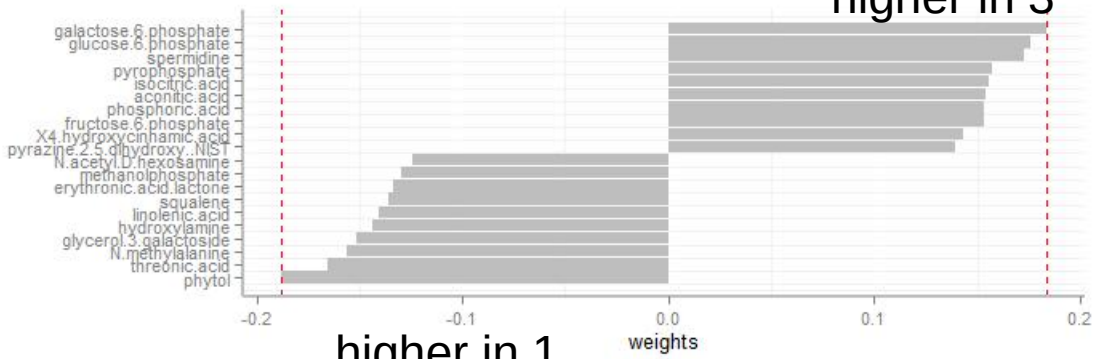
Feature Selection for Extraction

Partial Least Squares (O-/PLS/-DA)



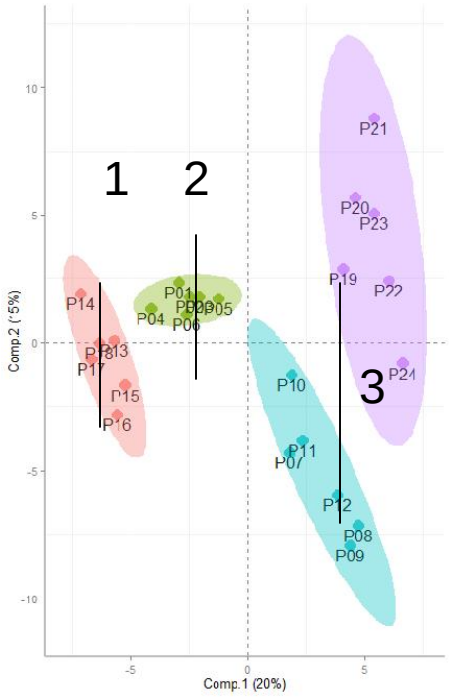


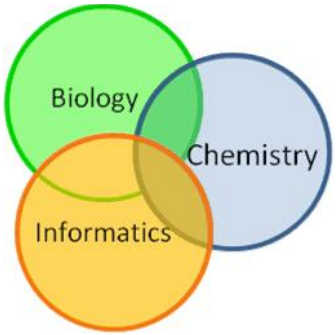
higher in 3



higher in 1

```
$selected.features[[2]]
           options
p.value       0.05
FDR           TRUE
cut.type    number
top             10
separate      TRUE
type      spearman
```





Summary

Partial Least Squares (O-/PLS/-DA)

Modeling Strategy (advanced)

1. Fit preliminary model
2. Evaluate of scores/loadings
3. Remove outliers
4. split data into test and train set (optional)
5. model selection (LV, OLV, etc)
6. internal (training set) model validation (permutation testing, training/testing)
7. Feature selection (optional)
    - Comparison of selected to excluded feature models
8. External model validation (test set)