Practicals single block analyses

Context and dataset introduction

```
library(CCA)
data("nutrimouse")
```

The dataset comes from a nutrigenomic study in the mouse (*Martin et al., 2007*, https://doi.org/10.1002/hep.21510) in which the effects of five regimens with contrasted fatty acid compositions on liver lipids and hepatic gene expression in mice were considered.

Objective: Investigate the impact of five diets with distinct fatty acid profiles on liver lipids and gene expression.

Design: 40 mice, cross-classified by:

Genotype: WT vs PPARα -/-

Diet: corn and colza oils (50/50) for a reference diet (REF), hydrogenated coconut oil for a saturated fatty acid diet (COC), sunflower oil for an Omega6 fatty acid-rich diet (SUN), linseed oil for an Omega3-rich diet (LIN) and corn/colza/enriched fish oils for the FISH diet (43/43/14)

Variables:

Gene expression: 120 selected genes (among about 30,000) as potentially relevant in the context of the nutrition study (macroarray)

Lipid composition: 21 hepatic fatty acids (GC analysis)

Single Block Analysis – PCA

Goals

Explore structure and variability in each block (genes and lipids)

Tasks

1. Data exploration

- Visualize distributions
- Check normalisation

2. Principal Component Analysis (PCA)

- Compute PCA for genes and lipids
- Plot scree plots (explained variance)

```
library(ropls)
opls()
PCA_res@pcaVarVn
```

3. Sample space visualization

- Plot scores: Dim.1 vs Dim.2 and Dim.3 vs Dim.4
- Identify main sources of variation

```
PCA_res@scoreMN
```

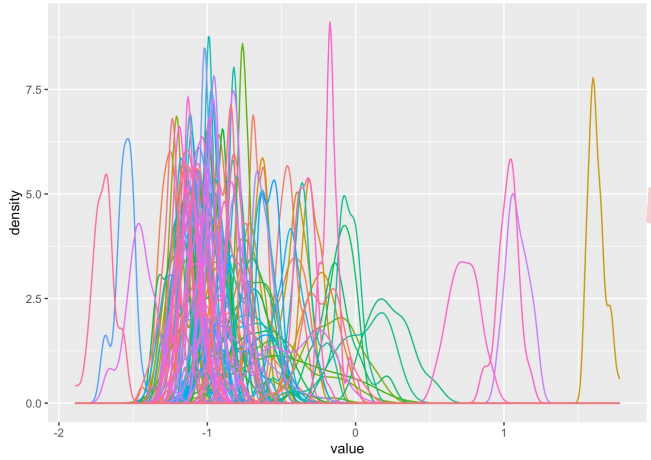
4. Variable contributions

- Plot loadings: Dim.1 vs Dim.2 and Dim.3 vs Dim.4
- Interpret biological relevance

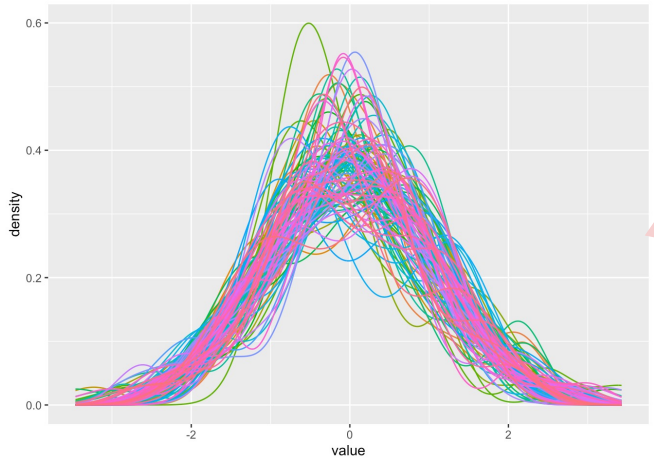
```
PCA_res@loadingsMN
```

Data distribution

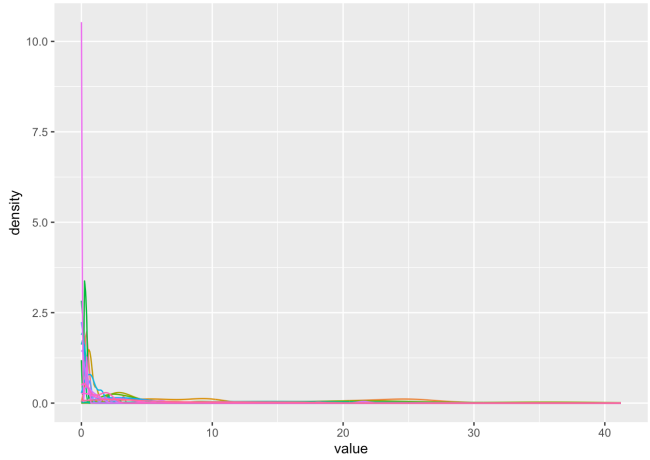
genes



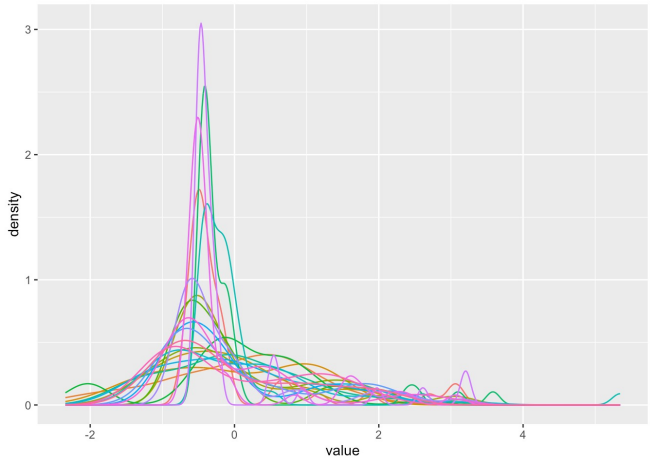
scaling



lipids

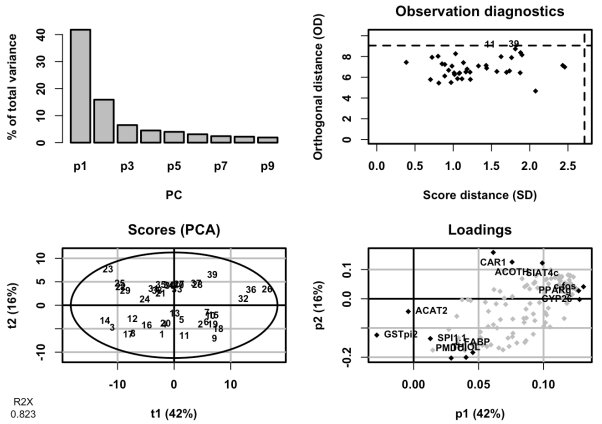


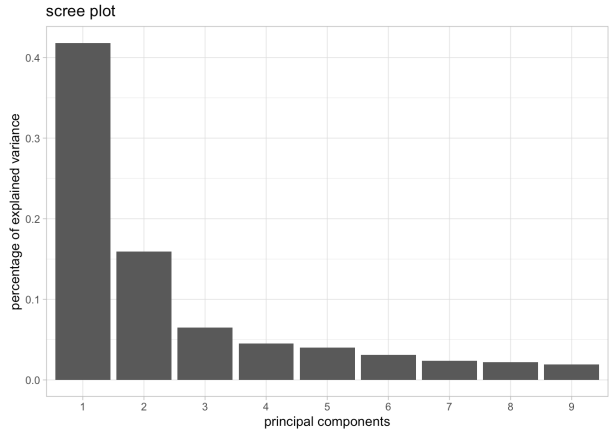
scaling



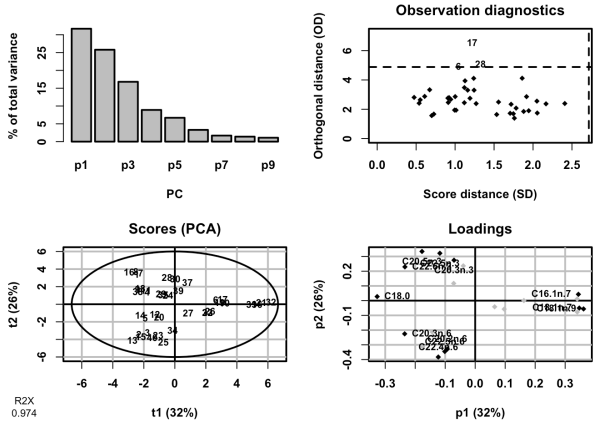
PCA models

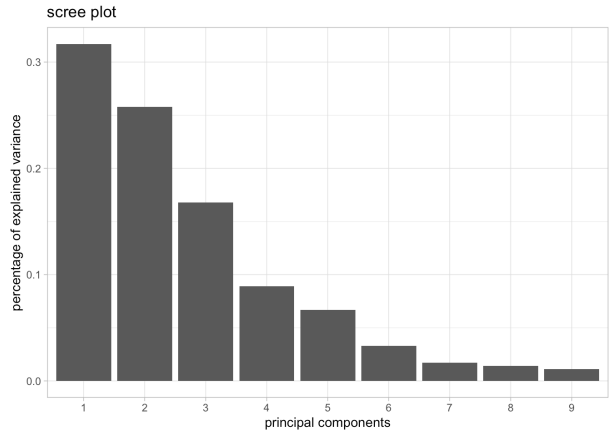
genes





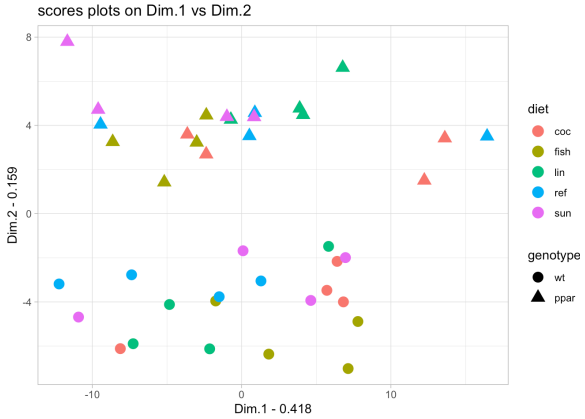
lipids

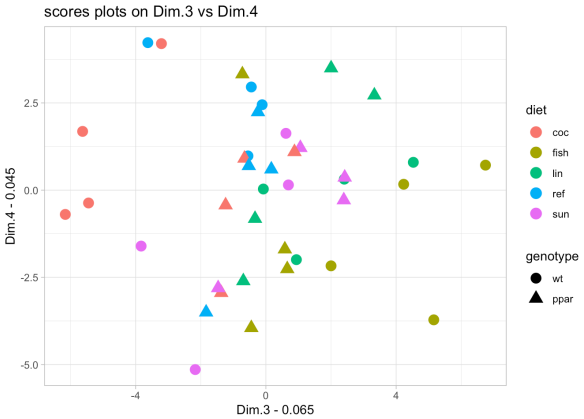


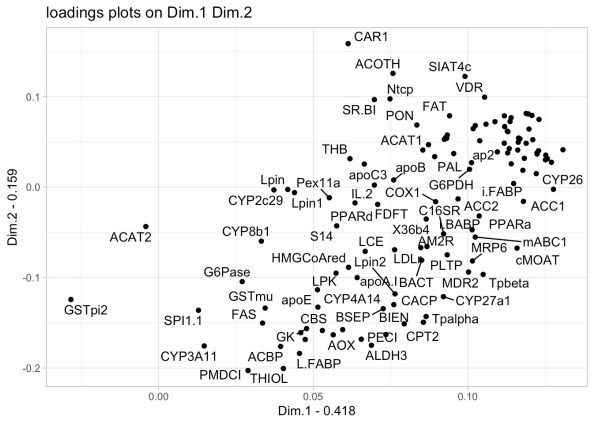


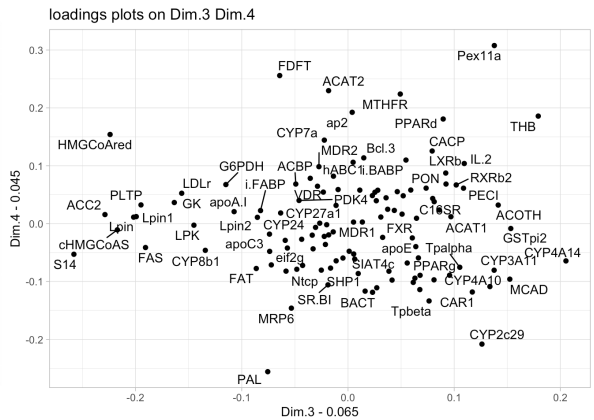
Sample space visualisation and variable contributions

genes

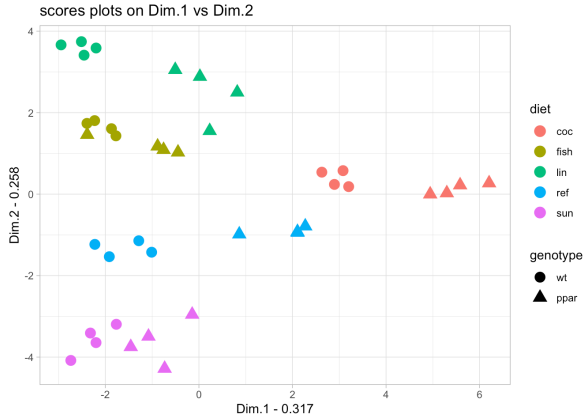


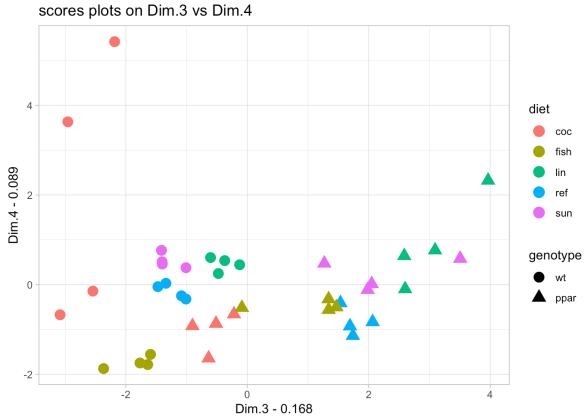


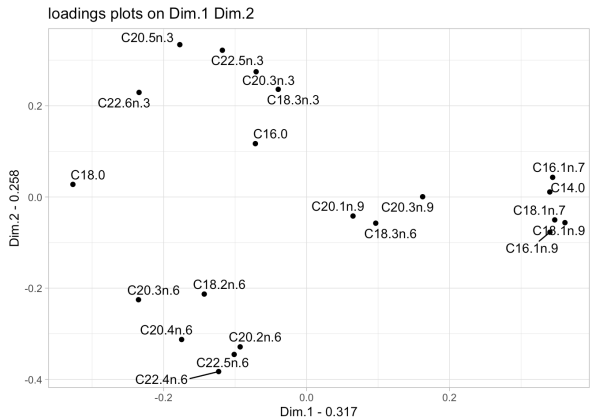


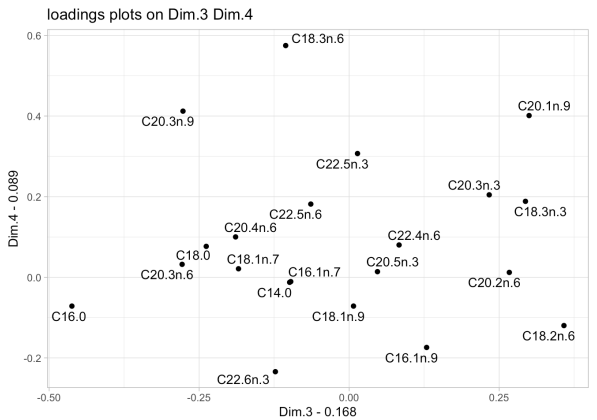


lipids









Principal Component Analysis - summary

Purpose:
Reduce the dimensionality of high-dimensional data while preserving as much variance as possible.

How it works:
PCA constructs new, orthogonal axes called Principal Components (PCs).
These are linear combinations of the original variables.

Key properties:

- PCs are ranked by the amount of variance they explain.
- PC1 captures the most variance, followed by PC2, and so on.
- Each PC is orthogonal to the others.

Why use PCA in omics:

- Reveal dominant patterns and trends. *What overarching patterns or trends can be observed in the dataset?*
- Visualize sample clustering. *Do the samples group or cluster according to relevant biological?*
- Identify variables contributing most to variation. *Which variables play the most significant role in explaining the observed variation across samples?*

PLS between two matrices

Goals:

Explore relationships between genes and lipids

Tasks:

1. PLS Canonical Analysis

Fit a model in canonical mode
Plot scores and loadings for both blocks

```
library(mixOmics)
pls(…,mode=”canonical")
```

```
PLS_res$variates$X and PLS_res$variates$Y
```

```
PLS_res$loadings$X and PLS_res$loadings$Y
```

2. PLS Regression Analysis

Fit model using mode = "regression”
Plot scores and loadings for both blocks

```
pls(…,mode="regression")
```

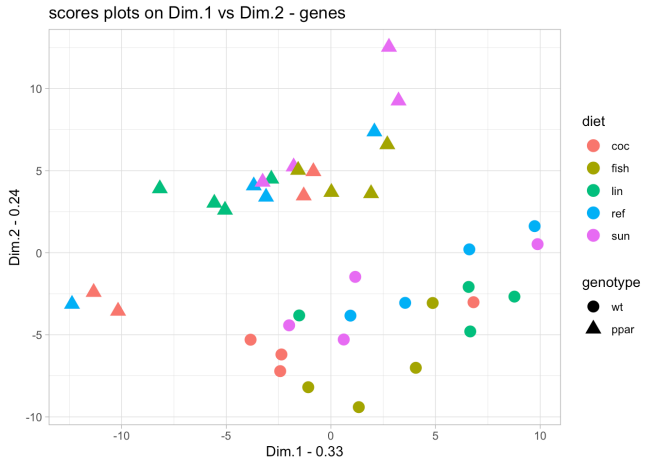
```
PLS_res_reg$variates$X and PLS_res_reg$variates$Y
```

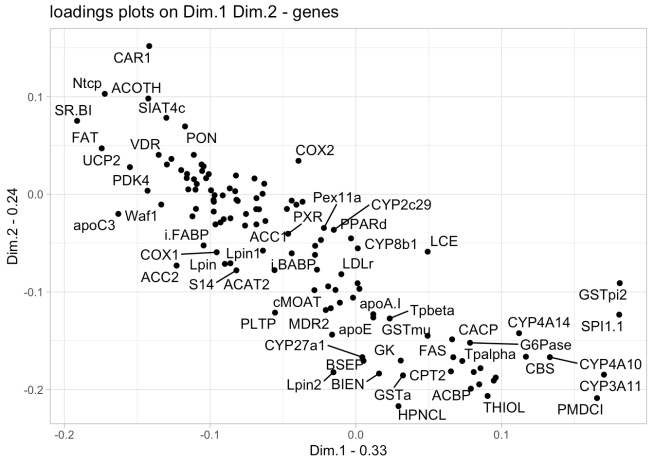
3. Compare both modes

Compare score plots and interpret differences

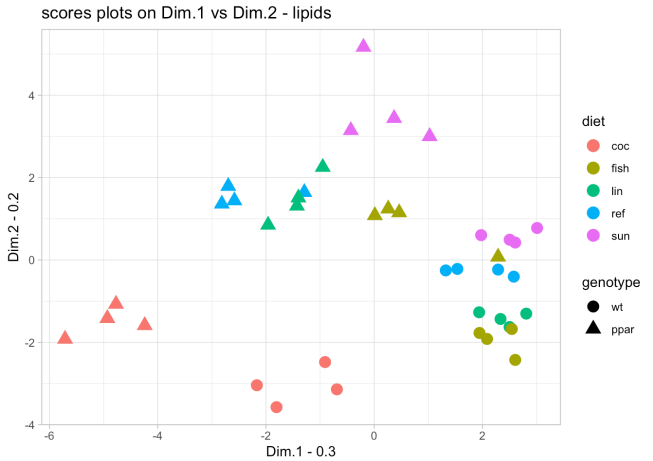
PLS Canonical Analysis

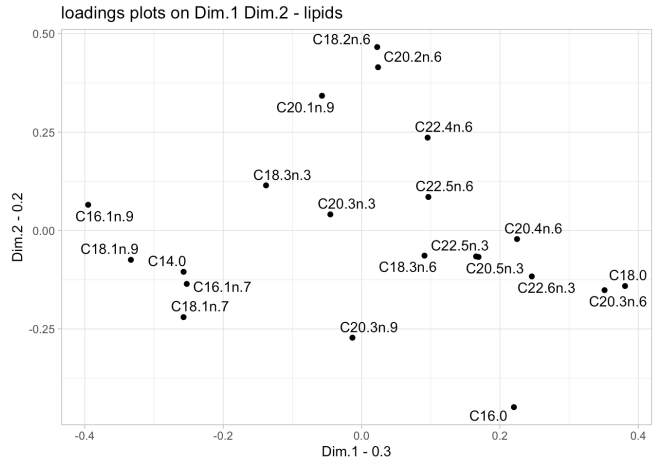
genes





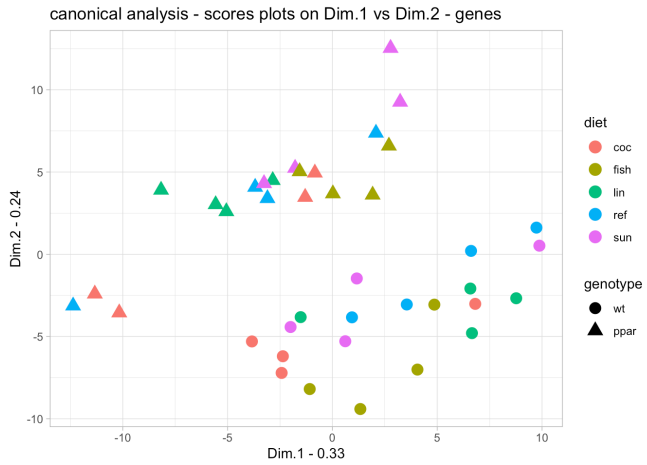
lipids

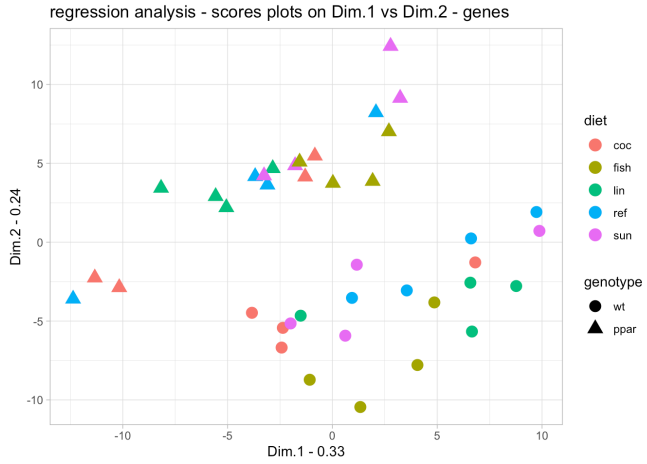




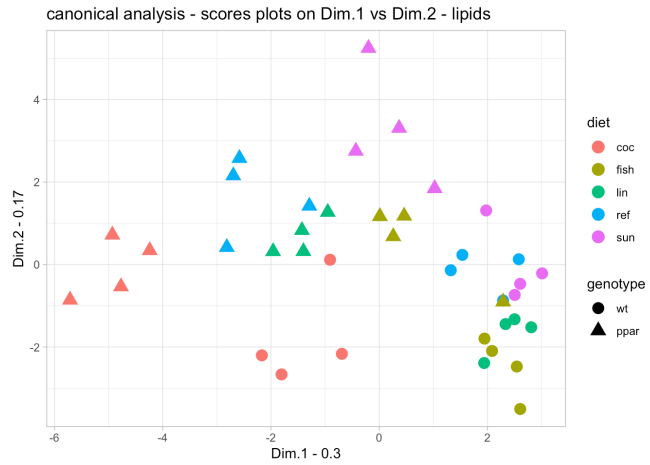
Comparison between PLS in canonical and regression modes

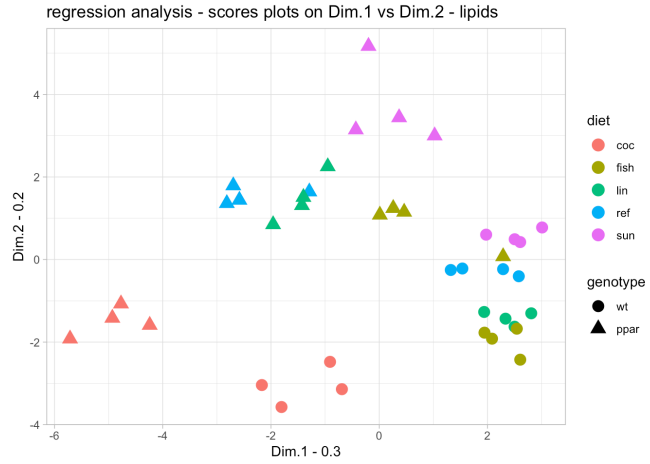
genes (X)





lipids (Y)





Partial Least Square analyses - summary

Purpose:

Explore and model relationships between two continuous datasets by identifying latent components that maximize covariance.

How it works:

PLS constructs new axes (latent variables) as linear combinations of the original variables from each dataset.

These components reduce dimensionality while capturing shared structure between datasets.

Canonical Mode (mode = "canonical")

- Symmetric treatment of X and Y
- No predefined predictor-response roles
- Y is deflated using information from Y

Regression Mode (mode = "regression")

- Asymmetric treatment: X = predictors, Y = responses
- Fits a linear model from X to Y
- Swapping X and Y changes the latent structure
- Y is deflated using information from X

Typical questions

Do both datasets reflect biological conditions of interest?

Can Y be modeled effectively using X?

Which subsets of variables are highly correlated and explain shared variation?

PLS discriminant analysis

Goals:

Discriminate genotypes using gene expression

Tasks:

1. PLS-DA Model

Fit model with ropls::opls()
Evaluate with permutation tests (Q^2, R^2Y)

```
library(ropls)
opls(x = nutrimouse$gene,
     y = metadata$genotype,
     predI = NA,
     permI = 100)
```

```
PLS_res@modelDF
```

2. Interpretation

Plot scores and loadings
Identify discriminant genes (VIP scores)

```
PLS_res@scoreMN
```

```
PLS_res@loadingsMN
```

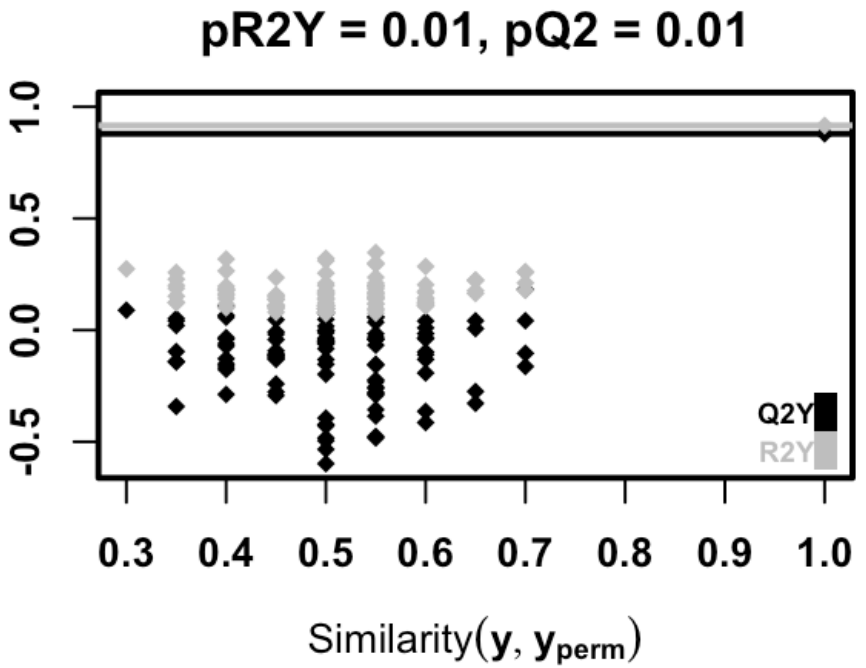
```
PLS_res@vipVn
```

Discriminate genotypes using gene expression

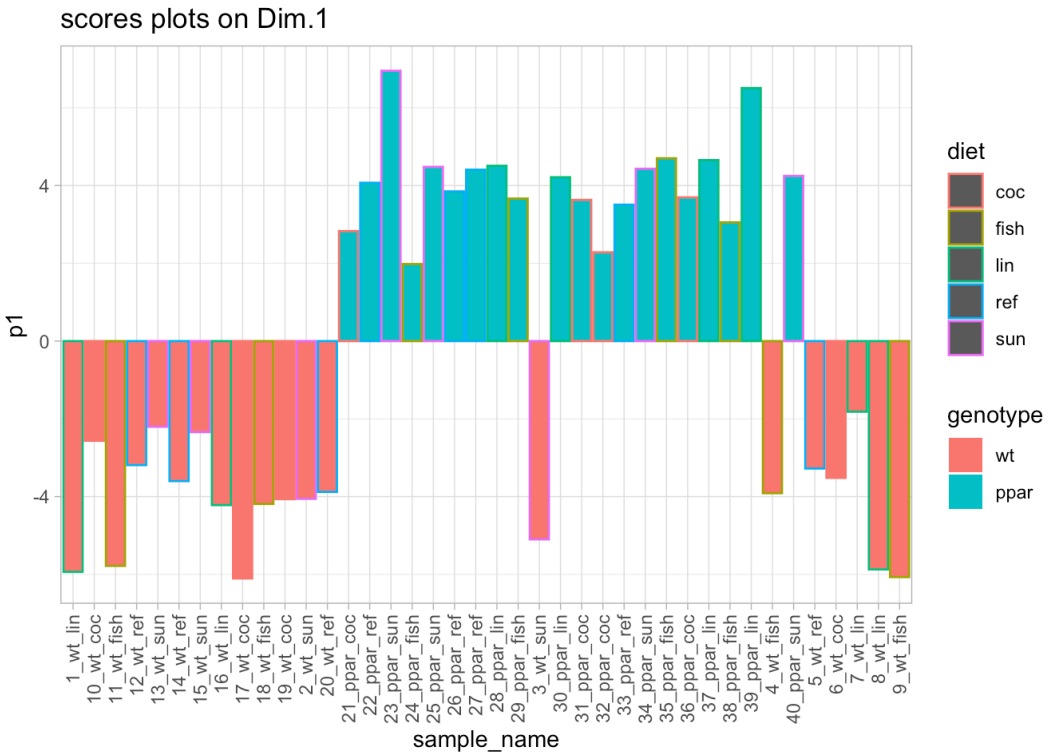
Fit a PLS-DA model with ropls::opls()

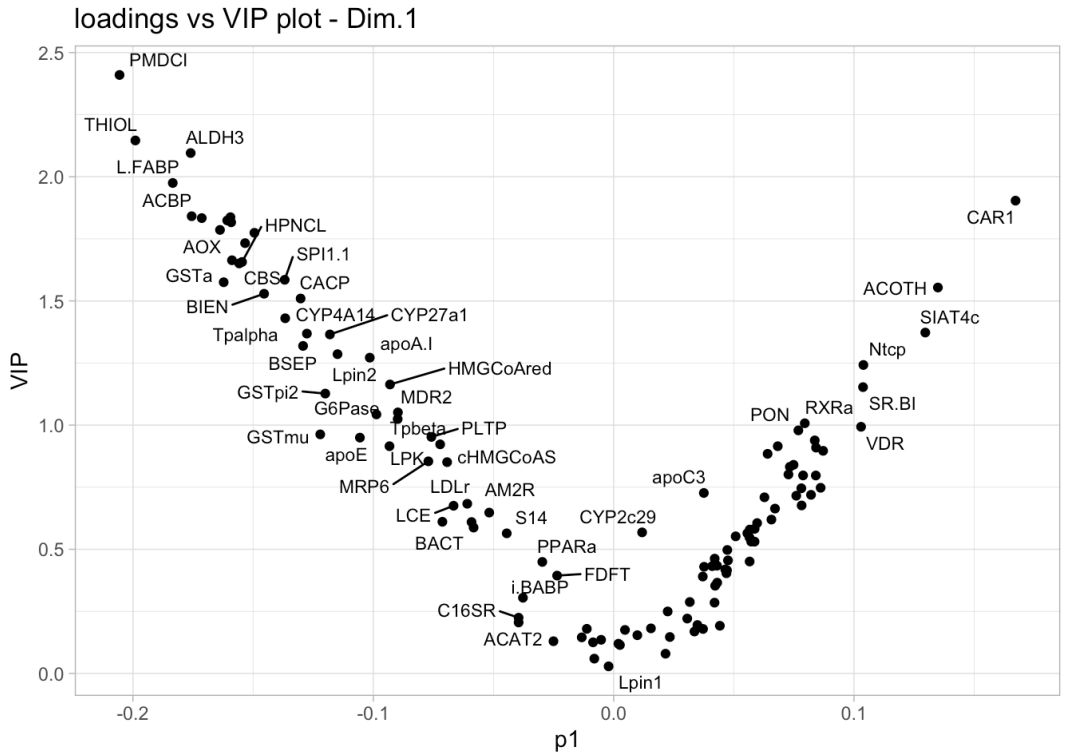
Evaluate the model with metrics and permutation tests

```
## PLS-DA
## 40 samples x 120 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total    0.158    0.916   0.879 0.149   1   0 0.01 0.01
```



Interpretation





OPLS discriminant analysis

Goals:

Improve interpretability by separating predictive and orthogonal variation

Tasks:

1. OPLS-DA Model

Fit model with ropls::opls()
Evaluate with permutation tests (Q^2, R^2Y)

2. Interpretation

Plot scores and loadings
Identify discriminant genes (VIP scores)

```
library(ropls)
opls(x = nutrimouse$gene,
     y = metadata$genotype,
     predI = 1,
     orthoI = 1,
     permI = 100)
```

```
OPLS_res@modelDF
```

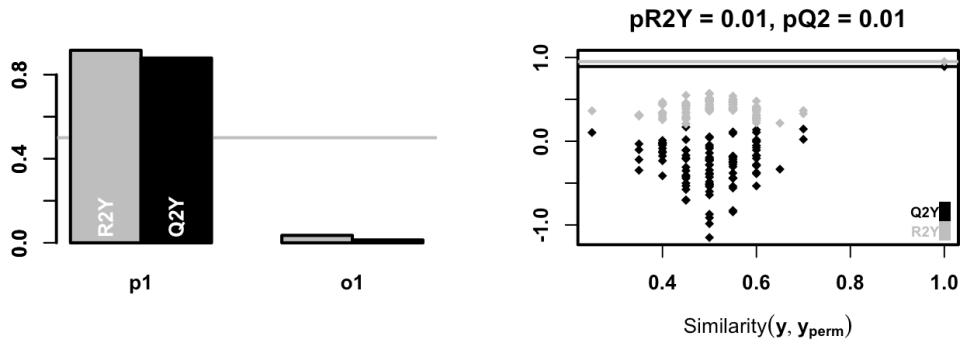
```
OPLS_res@scoreMN
OPLS_res@orthoScoreMN
```

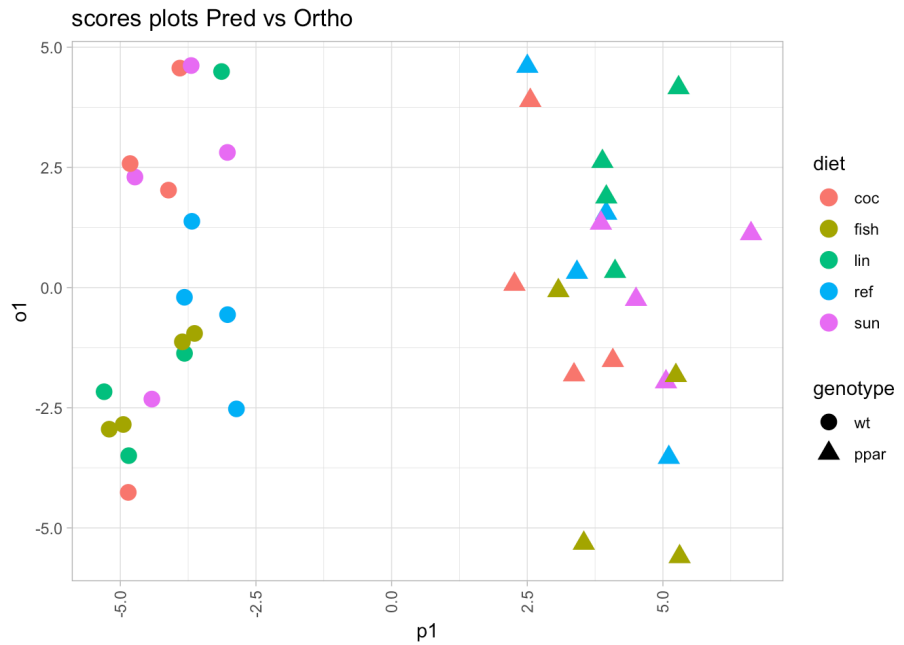
```
OPLS_res@loadingsMN
OPLS_res@orthoLoadingsMN
```

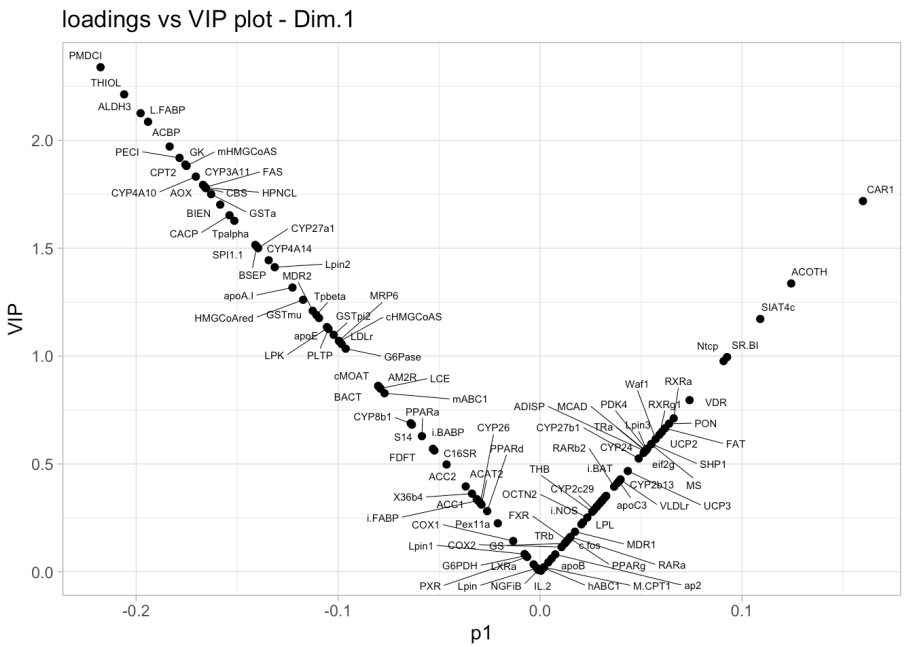
```
OPLS_res@vipVn
```

Discriminate genotypes using gene expression

```
## OPLS-DA
## 40 samples x 120 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total     0.38    0.951   0.892 0.115   1   1 0.01 0.01
```







PLS & OPLS discriminant analysis - summary

Purpose

Identify latent components that maximize covariance between predictors and categorical outcomes, enabling sample discrimination and "biomarker" discovery.

How it works

PLS-DA and OPLS-DA are extensions of PLS regression adapted for classification tasks.
They model the relationship between X (predictors) and Y (response), extracting components that best separate outcome groups.

PLS-DA:

- Uses all variation in X to model Y
- May include noise or irrelevant variation

OPLS-DA:

- Separates predictive variation from orthogonal (non-predictive) variation
- Improves interpretability and model clarity

Typical Questions

Can we distinguish samples based on their outcome category?
Which variables are most discriminant between outcome groups?
Do these variables form a predictive molecular signature for new samples?

Canonical Correlation Analysis (CCA)

Goals:

Investigate linear relationships between genes and lipids

Tasks:

1. **Basic CCA**
   - Use mixOmics::rcc() on 20 genes and all lipids
   - Plot scores and loadings
2. **CCA on scaled data**
   - Repeat with scaled data
   - Compare results
3. **Regularized CCA**
   - Use full gene set
   - Discuss regularization effects

```
library(mixomics)
cca.res <- rcc(X=nutrimouse$gene_selected,
               Y=nutrimouse$lipid)
```

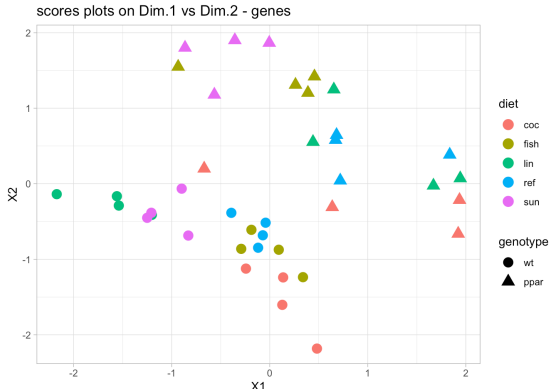
```
cca.res$variates$X
cca.res$variates$Y
```

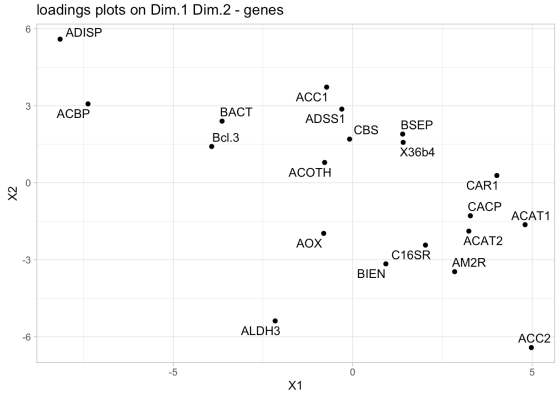
```
cca.res$loadings$X
```

CCA

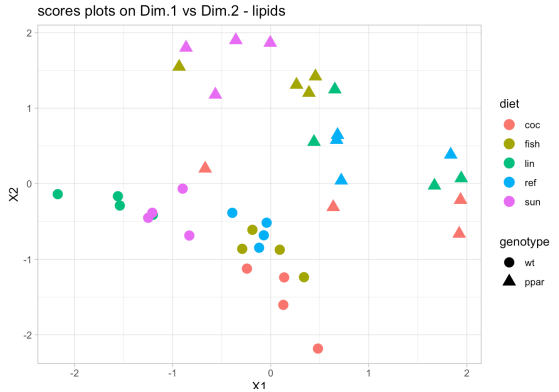
raw data

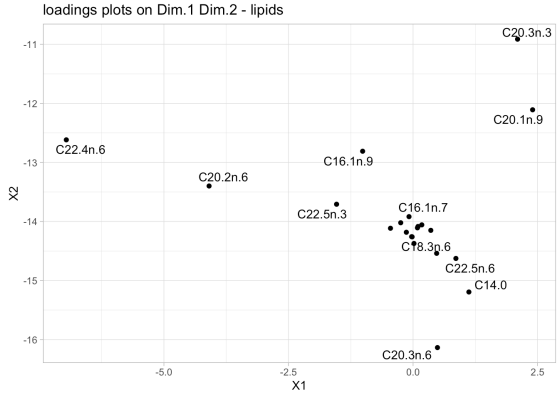
genes (X)

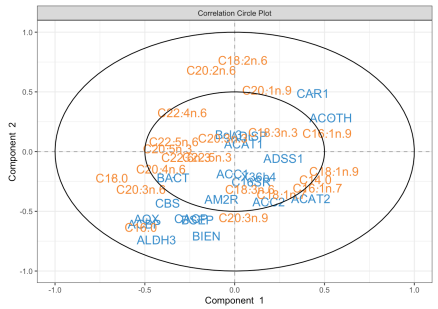




lipids (Y)

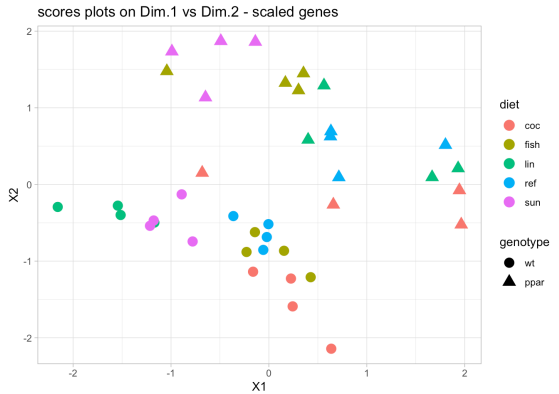


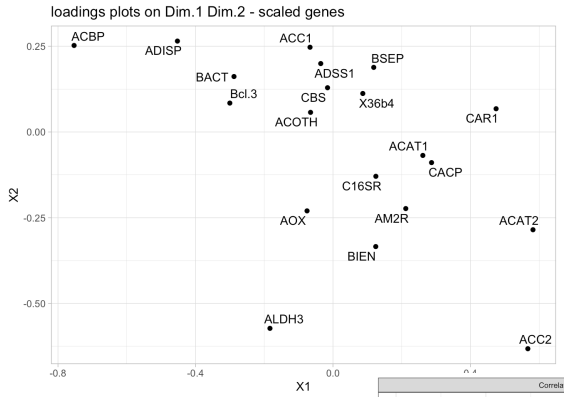




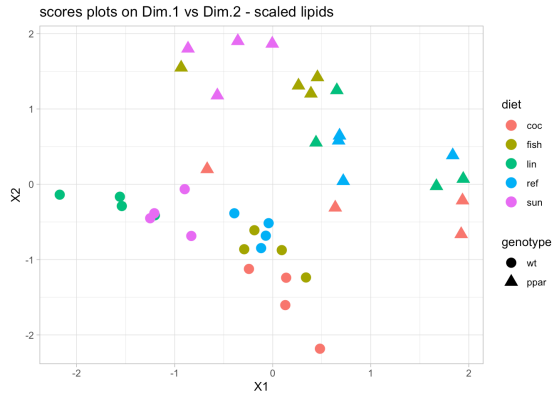
scaled

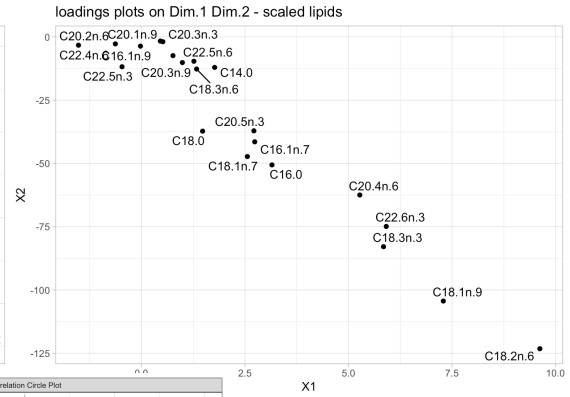
genes (X)

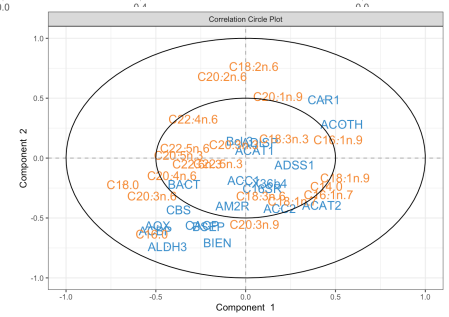




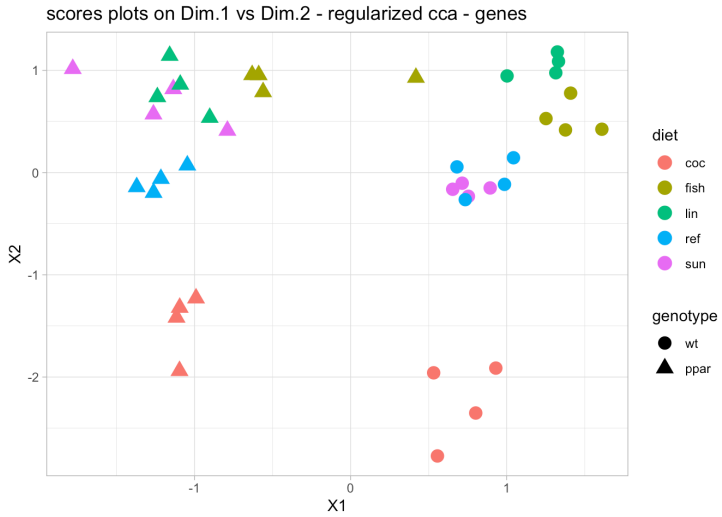
lipids (Y)

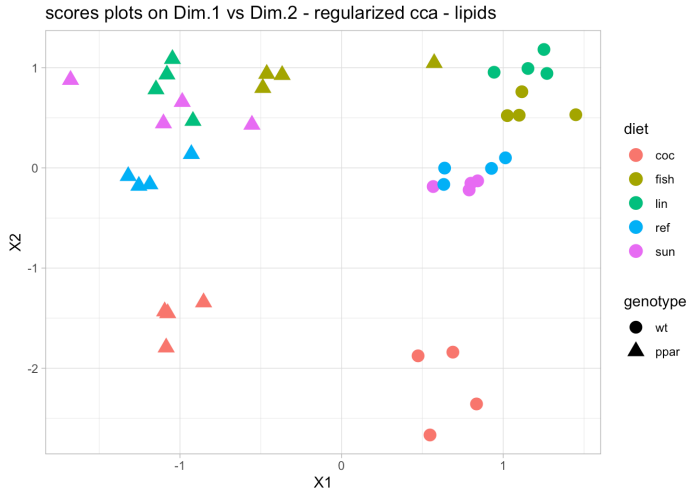


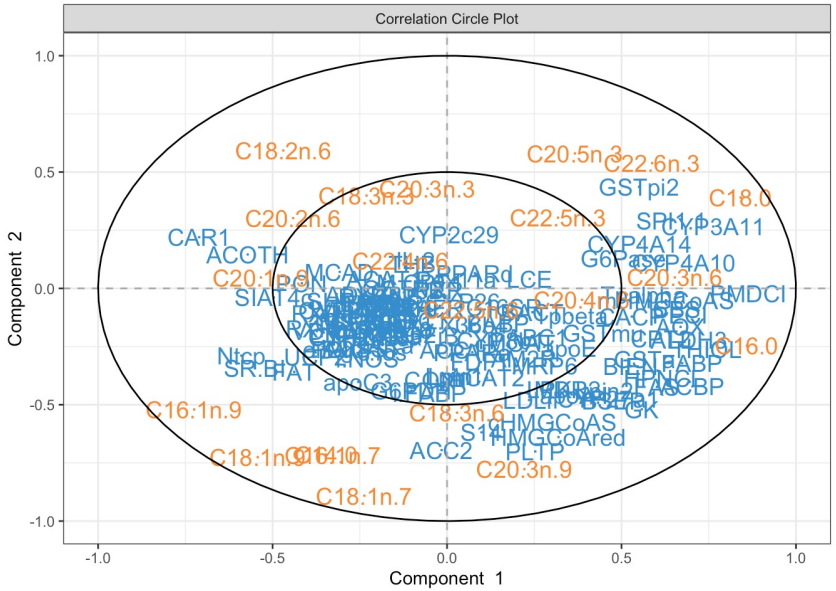


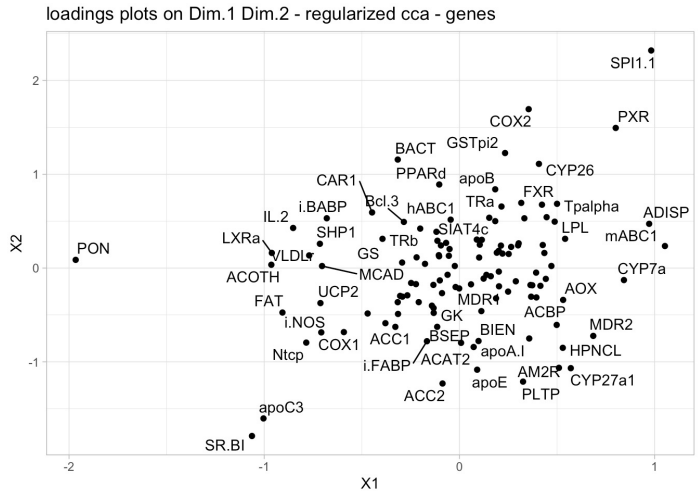


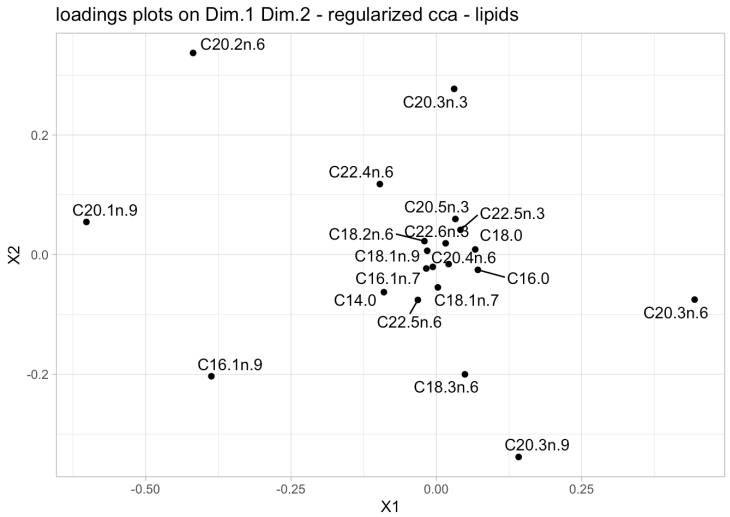
Regularized CCA











CCA - summary

Purpose

Identify linear relationships between two sets of variables, enabling exploration of how changes in one set relate to changes in another, and revealing shared patterns across datasets.

How it works

CCA finds pairs of canonical variables—linear combinations of the original variables from each dataset—that are maximally correlated. It captures shared variation between X and Y, highlighting interdependencies.

- Standard CCA: Finds all pairs of canonical variates that maximize correlation, without distinguishing predictive versus non-predictive variation.
- Regularized / Sparse CCA: Adds constraints to improve interpretability or handle high-dimensional data, selecting only the most relevant variables.

Typical Questions

How strongly are the two sets of variables related?

Which variables contribute most to the shared patterns?

Can we use these relationships to predict or understand associations between datasets?

Single block analyses - summary

PCA

Reduce dimensionality while preserving variance.

What are the dominant patterns in the data?
Do samples cluster by biological conditions?
Which variables explain the most variation?

CCA

Identify linear relationships between two datasets by maximizing correlation between canonical variables.
Symmetric treatment of X and Y, no predefined roles
Reveals shared patterns and associations between datasets

How strongly are the two sets of variables related?
Which variables contribute most to the shared patterns?
Can we use these relationships to predict or understand associations between datasets?

PLS

Explore and model relationships between two continuous datasets by maximizing covariance.
Canonical Symmetric treatment of X and Y, no predefined roles
Regression Asymmetric: X = predictors, Y = responses, fits linear model from X to Y

Do both datasets reflect biological conditions?
Can Y be modeled using X?
Which variables are highly correlated across datasets?

PLS-DA & OPLS-DA

Identify latent components that discriminate outcome categories and enable biomarker discovery.
PLS-DA: Uses all variation in X
OPLS-DA: Separates predictive from orthogonal variation for clearer interpretation

Can we distinguish samples by outcome?
Which variables are most discriminant?
Do they form a predictive molecular signature?