CODE ▾

Supervised Multiblock analyses

Florence Mehl

March 07, 2025

- Nutrimouse dataset
- Discriminant analysis of genotypes
    - Question 1: based on lipids and genes data, can we discriminate wt vs ppar samples ?
    - Question 2: Choose optimal number of latent variables?
    - Question 3: Is the model statistically significant?
    - Question 4: what is the variance explained for each block by each latent variable and globally?
    - Question 5: observe the samples distributions in the space of the latent variables.
    - Question 6: which genes and lipids are discriminant for genotype?
- Consensus OPLS Discriminant analysis of genotypes
    - Question 1: based on lipids and genes data, can we discriminate wt vs ppar samples ?
    - Question 2: Is the model statistically significant?
    - Question 3: What is the contribution of each data block?
    - Question 4: Show the distribution of samples in the space of the predictive and orthogonal latent variables?
    - Question 5: Show the loadings of variables in the space of the predictive and orthogonal latent variables?
    - Question 6: Show the importance of variables in the model?
    - Question 7: train a model to discriminate between wt and ppar with 30 observations and test it with 10 observations?

Nutrimouse dataset

The data sets come from a nutrigenomic study in the mouse (Martin et al., 2007) in which the effects of five regimens with contrasted fatty acid compositions on liver lipids and hepatic gene expression in mice were considered.

Two sets of variables were acquired on forty mice: - genes: expressions of 120 genes measured in liver cells, selected (among about 30,000) as potentially relevant in the context of the nutrition study. These expressions come from a nylon macroarray with radioactive labelling - lipids: concentrations (in percentages) of 21 hepatic fatty acids measured by gas chromatography

Biological units (mice) were cross-classified according to two factors experimental design (4 replicates): - genotype: 2-levels factor, wild-type (WT) and PPARalpha -/- (PPAR) - diet: 5-levels factor. Oils used for experimental diets preparation were 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)

HIDE

```
data("nutrimouse")
genes <- nutrimouse$gene
lipids <- nutrimouse$lipid
metadata <- data.frame(genotype = nutrimouse$genotype, diet = nutrimouse$diet)
metadata$sample_name <- paste0(rownames(metadata), "_", metadata$genotype, "_", metadata$diet)
rownames(genes) <- metadata$sample_name
rownames(lipids) <- metadata$sample_name
```

Discriminant analysis of genotypes

Question 1: based on lipids and genes data, can we discriminate wt vs ppar samples ?

Run block.plsda analysis with block.plsda() from mixomics package

HIDE

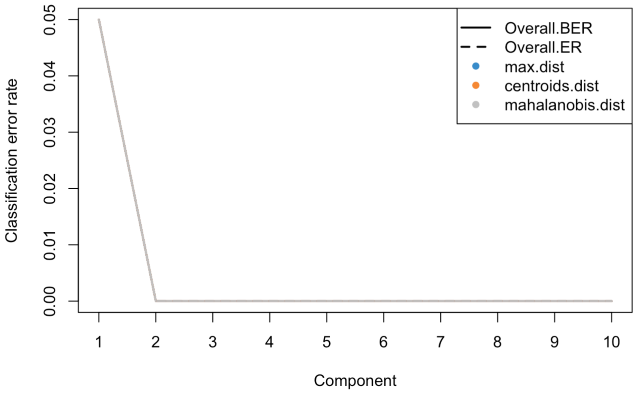
```
# prepare data
blockPLS_data <- list(genes=genes, lipids=lipids)
genotype <- as.factor(metadata$genotype)

# run analysis
blockPLS_res <- block.plsda(X = blockPLS_data, Y = genotype, design = "full", all.outputs = T, ncomp = 10)
```

Question 2: Choose optimal number of latent variables?

Run perf() plot the results with plot() Run the analysis with optimal number of latent variables

HIDE

```
blockPLS_perf <- perf(blockPLS_res, validation = 'Mfold', folds = 7, nrepeat = 1, auc = TRUE, cpus=2, progressBar = FALSE)

plot(blockPLS_perf)
```



HIDE

```
blockPLS_res <- block.plsda(X = blockPLS_data, Y = genotype, design = "full", all.outputs = T, ncomp = 2)
```

Question 3: Is the model statistically significant?

Run a permutation test with DIABLO.test() from RVAideMemoire package

HIDE

```
blockPLS_permtest <- DIABLO.test(blockPLS_res, progress = FALSE)

blockPLS_permtest
```

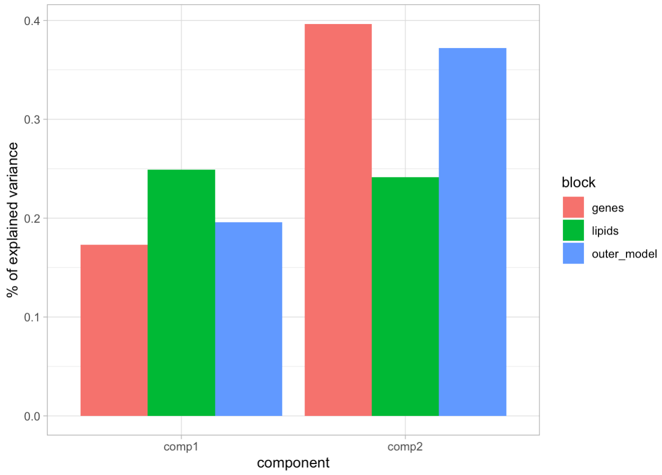
```
##
##  Permutation test based on cross-validation
##
## data:  blockPLS_res
## DIABLO (2 components)
## 999 permutations
## CER = 0.005, p-value = 0.001
```

Question 4: what is the variance explained for each block by each latent variable and globally?

- for each block: AVE_X
- global: AVE[["AVE_outer"]]

HIDE

```
blockPLS_expl <- do.call("rbind",blockPLS_res$AVE$AVE_X[1:2])
blockPLS_expl <- rbind(blockPLS_expl, blockPLS_res$AVE[["AVE_outer"]])
rownames(blockPLS_expl)[3] <- "outer_model"
blockPLS_expl <- melt(blockPLS_expl)
colnames(blockPLS_expl) <- c("block", "comp", "value")

ggplot(blockPLS_expl, aes(x=comp, y=value, fill=block)) +
  geom_bar(stat="identity", position=position_dodge()) +
  labs(x="component",
       y="% of explained variance") +
  theme_light()
```

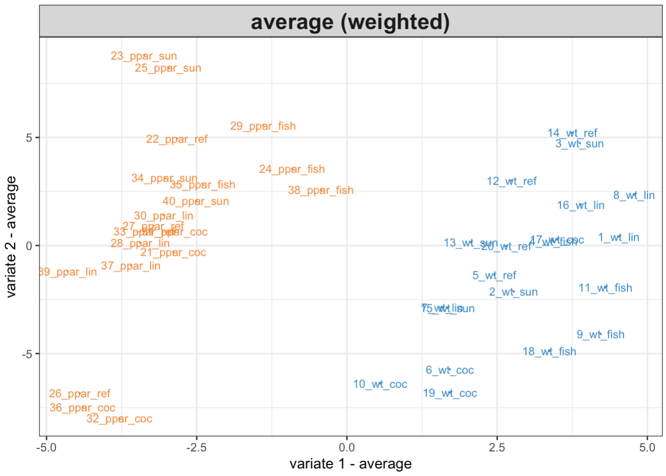


Question 5: observe the samples distributions in the space of the latent variables.

- plot scores with plotIndiv()

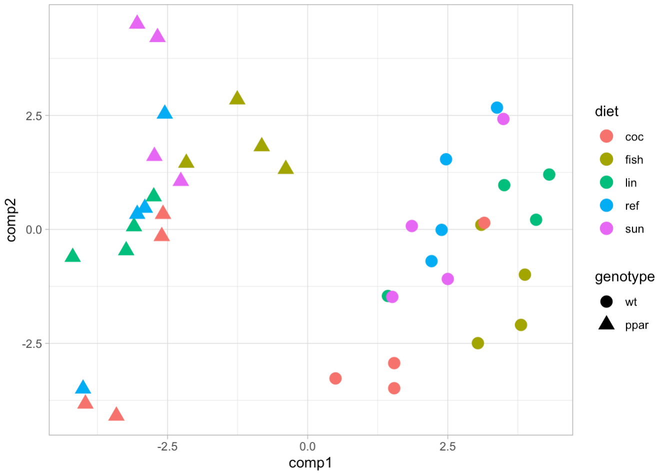
HIDE

```
plotIndiv(blockPLS_res, block = "weighted.average")
```



HIDE

```
# ou:

blockPLS_variates.weighted <- blockPLS_res$variates[c("genes", "lipids")]
for(omic in c("genes", "lipids")){
  for(comp in c("comp1", "comp2")){
      blockPLS_variates.weighted[[omic]][,comp] <- blockPLS_variates.weighted[[omic]][,comp] * blockPLS_res$weights[omi
c, comp]
  }
}
blockPLS_scores.weighted <- abind(blockPLS_variates.weighted[c("genes", "lipids")], along = 3)
blockPLS_scores.weighted <- apply(blockPLS_scores.weighted, c(1,2), mean)

blockPLS_scores.weighted <- data.frame(metadata, blockPLS_scores.weighted)

ggplot(blockPLS_scores.weighted, aes(x=comp1, y=comp2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  theme_light()
```

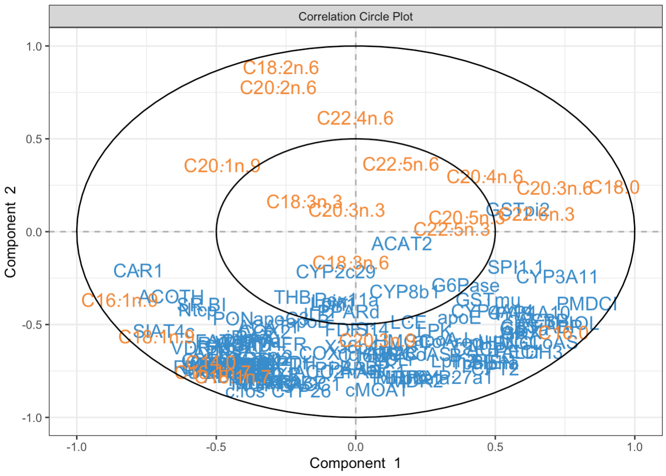


Question 6: which genes and lipids are discriminant for genotype?

- plot loadings with plotVar()

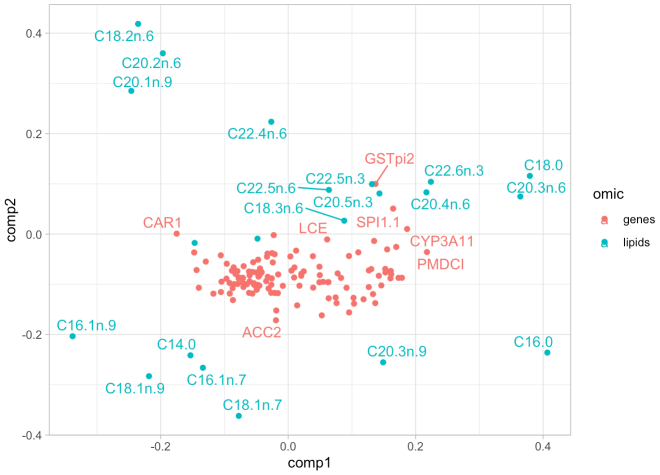
HIDE

```
plotVar(blockPLS_res)
```



HIDE

```
# ou

blockPLS_loadings_genes <- blockPLS_res$loadings$genes
blockPLS_loadings_lipids <- blockPLS_res$loadings$lipids
blockPLS_loadings <- rbind.data.frame(blockPLS_loadings_genes, blockPLS_loadings_lipids)
blockPLS_loadings$omic <- c(rep("genes", dim(genes)[[2]]), rep("lipids", dim(lipids)[[2]]))
blockPLS_loadings$variable <- rownames(blockPLS_loadings)

ggplot(blockPLS_loadings, aes(x=comp1, y=comp2, col=omic, label=variable)) +
  geom_point() +
  geom_text_repel() +
  theme_light()
```



Consensus OPLS Discriminant analysis of genotypes

Question 1: based on lipids and genes data, can we discriminate wt vs ppar samples ?

Run ConsensusOPLS-DA analysis with ConsensusOPLS() from ConsensusOPLS package

HIDE

```
COPLS_data <- list(genes=as.matrix(genes), lipids=as.matrix(lipids))
COPLS_data <- lapply(COPLS_data, scale)
genotype <- metadata$genotype
dummy_genotype <- as.matrix(data.frame(wt = ifelse(genotype == "wt", 1, 0),ppar = ifelse(genotype == "ppar", 1, 0)))

COPLS_res <- ConsensusOPLS(
  data = COPLS_data,
  Y = dummy_genotype,
  maxPcomp = 1,
  maxOcomp = 1,
  modelType = "da",
  cvType = "nfold",
  nfold = 40,
  nperm = 100,
  verbose = T,
  kernelParams = list(type='p', params = c(order=1.0))
)
```

Question 2: Is the model statistically significant?

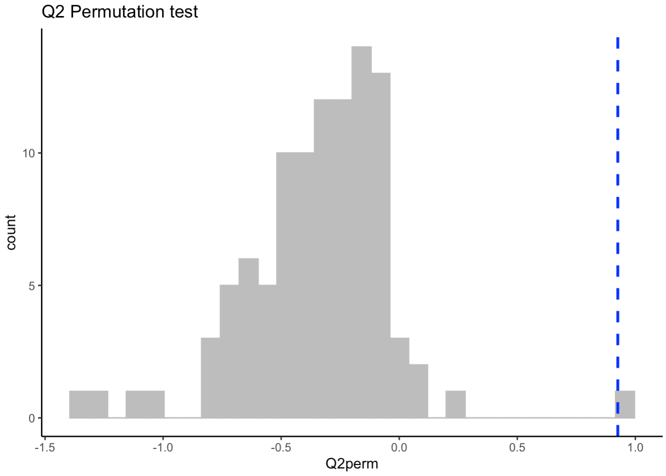
The results of permutations can be found in `COPLS_res$permStats` . The results for the optimal model can be found in `COPLS_res$optimal$modelCV` and `COPLS_res$optimal$modelCV$cv`

- plot Q2 permutations
- plot DQ2 permutations

- plot R2Y permutations

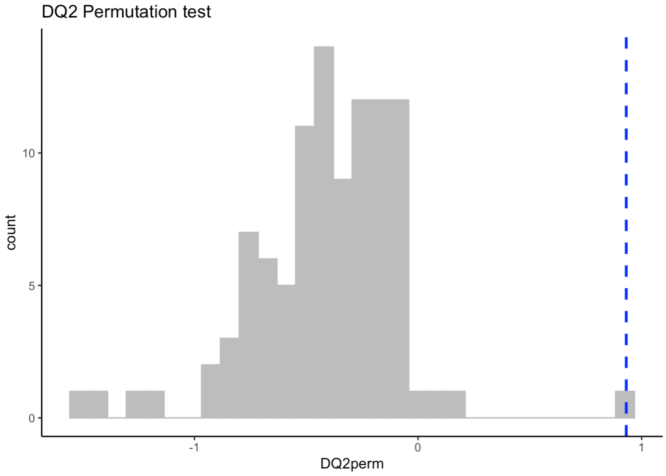
HIDE

```
Q2perm <- data.frame(Q2perm = COPLS_res@permStats$Q2Y)

ggplot(data = Q2perm, aes(x = Q2perm)) +
  geom_histogram(color="grey", fill="grey") +
  geom_vline(aes(xintercept=COPLS_res@Q2["po1"]),color="blue", linetype="dashed", size=1) +
  theme_classic() +
  ggtitle("Q2 Permutation test")
```



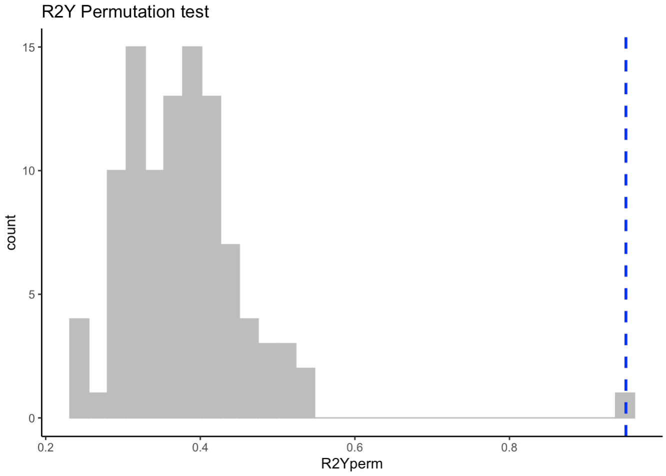
HIDE

```
DQ2perm <- data.frame(DQ2perm = COPLS_res@permStats$DQ2Y)

ggplot(data = DQ2perm, aes(x = DQ2perm)) +
  geom_histogram(color="grey", fill="grey") +
  geom_vline(aes(xintercept=COPLS_res@DQ2["po1"]),color="blue", linetype="dashed", size=1) +
  theme_classic() +
  ggtitle("DQ2 Permutation test")
```



HIDE

```
R2Yperm <- data.frame(R2Yperm = COPLS_res@permStats$R2Y)

ggplot(data = R2Yperm, aes(x = R2Yperm)) +
  geom_histogram(color="grey", fill="grey") +
  geom_vline(aes(xintercept=COPLS_res@R2Y["po1"]),color="blue", linetype="dashed", size=1) +
  theme_classic() +
  ggtitle("R2Y Permutation test")
```

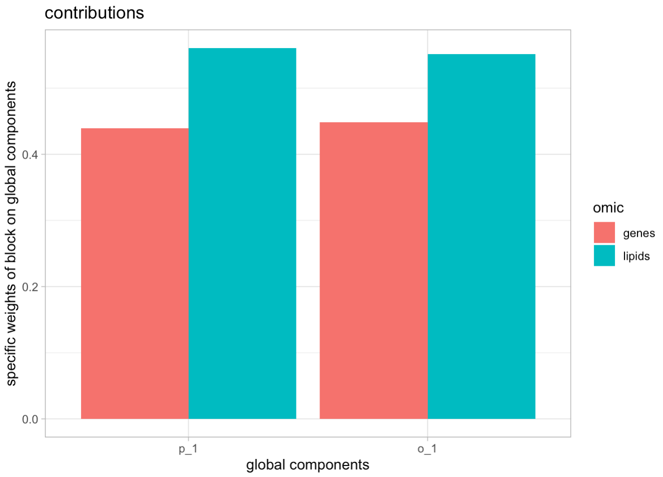


Question 3: What is the contribution of each data block?

- plot blockContribution of the optimal model

HIDE

```
contributions <- COPLS_res@blockContribution
contributions <- melt(contributions)
colnames(contributions) <- c("dataset", "Dim", "value")

ggplot(contributions, aes(x=Dim, y=value, fill=dataset)) +
  geom_bar(stat = "identity", position=position_dodge()) +
  theme_light() +
  labs(x = "global components", y = "specific weights of block on global components", fill = "omic",
       title = "contributions")
```

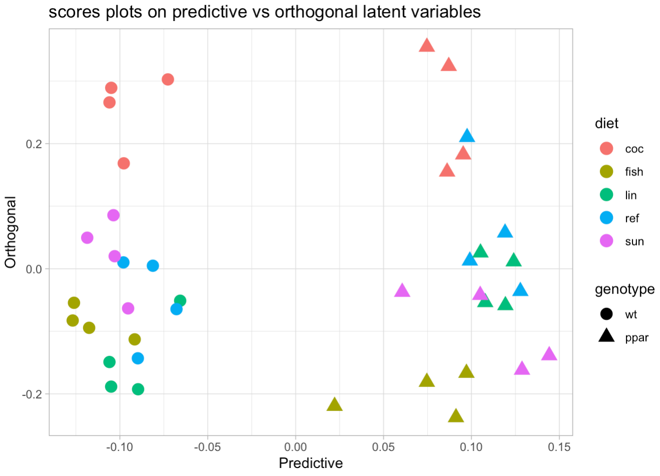


Question 4: Show the distribution of samples in the space of the predictive and orthogonal latent variables?

- plot scores of the optimal model

HIDE

```
scores <- data.frame(metadata, COPLS_res@scores)

ggplot(scores, aes(x=p_1, y=o_1, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(x="Predictive",
       y="Orthogonal",
       title = "scores plots on predictive vs orthogonal latent variables") +
  theme_light()
```

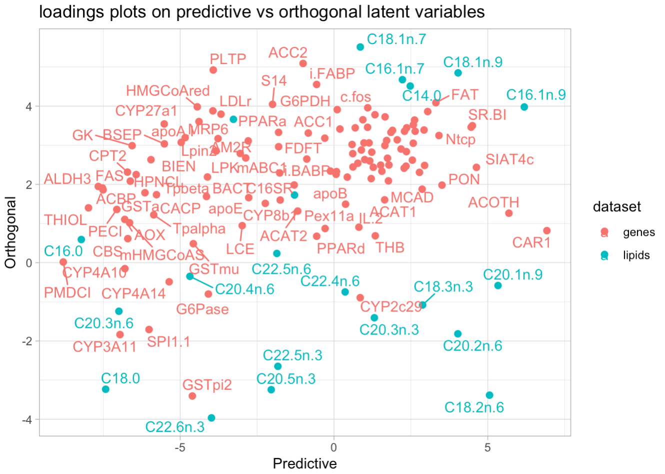


Question 5: Show the loadings of variables in the space of the predictive and orthogonal latent variables?

- plot loadings of the optimal model

HIDE

```
loadings <- rbind.data.frame(COPLS_res@loadings$genes, COPLS_res@loadings$lipids)
loadings$dataset <- c(rep("genes", nrow(COPLS_res@loadings$genes)), rep("lipids", nrow(COPLS_res@loadings$lipids)))
loadings$variable <- rownames(loadings)

ggplot(loadings, aes(x=p_1, y=o_1, col=dataset, label = variable)) +
  geom_point(size=2) +
  labs(x="Predictive",
       y="Orthogonal",
       title = "loadings plots on predictive vs orthogonal latent variables") +
  geom_text_repel() +
  theme_light()
```

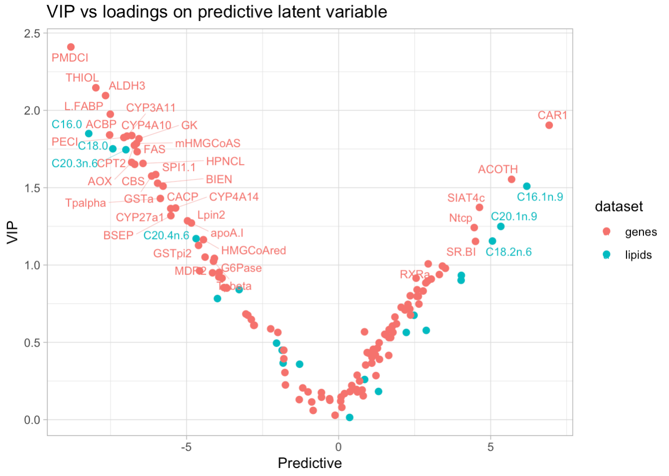


Question 6: Show the importance of variables in the model?

- plot loadings and VIP of the optimal model

HIDE

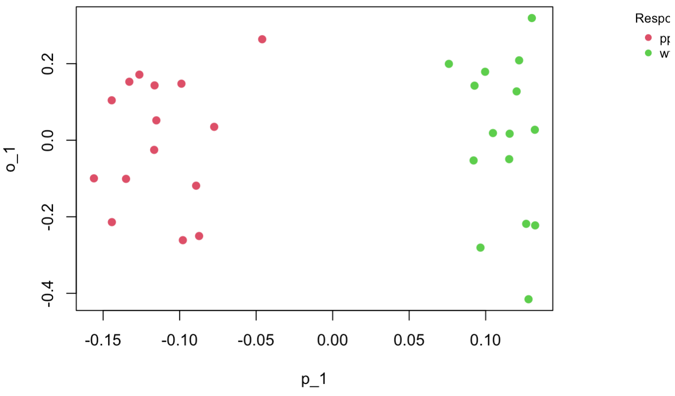
```
VIP <- data.frame(VIP = c(COPLS_res@VIP$genes$p, COPLS_res@VIP$lipids$p), variable = c(rownames(COPLS_res@VIP$genes), ro
wnames(COPLS_res@VIP$lipids)))

loadings_VIP <- merge(loadings, VIP, by="variable")
loadings_VIP$label <- ifelse(loadings_VIP$VIP > 1, loadings_VIP$variable, NA)

ggplot(loadings_VIP, aes(x=p_1, y=VIP, col=dataset, label = label)) +
  geom_point(size=2) +
  labs(x="Predictive",
       y="VIP",
       title = "VIP vs loadings on predictive latent variable") +
  geom_text_repel(size=3, max.overlaps = 50, segment.size=.1) +
  theme_light()
```



Question 7: train a model to discriminate between wt and ppar with 30 observations and test it with 10 observations?

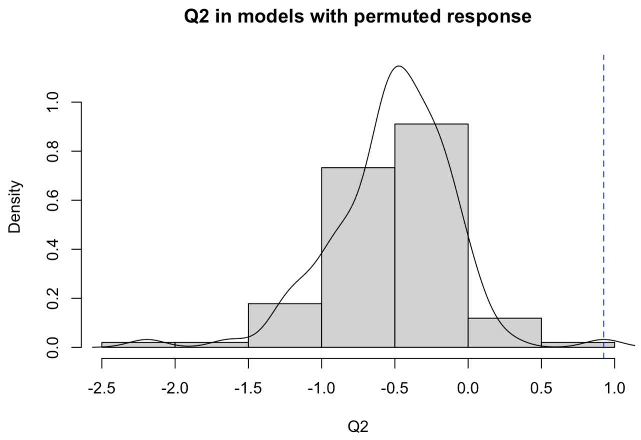
HIDE

```
test.observations <-c(sample(1:20, 5, replace = F),
                      sample(21:40, 5, replace = F))
train.data <- list(genes=as.matrix(genes[-test.observations,]),
                   lipids=as.matrix(lipids[-test.observations,]))
train.data <- lapply(train.data, scale)
test.data <- list(genes=as.matrix(genes[test.observations,]),
                  lipids=as.matrix(lipids[test.observations,]))
test.data <- lapply(test.data, scale)

train.genotype <- metadata$genotype[-test.observations]

train.COPLS_res <- ConsensusOPLS(
  data = train.data,
  Y = train.genotype,
  maxPcomp = 1,
  maxOcomp = 1,
  modelType = "da",
  cvType = "nfold",
  nfold = 30,
  nperm = 100,
  verbose = T,
  kernelParams = list(type='p', params = c(order=1.0))
)

plotScores(train.COPLS_res)
```



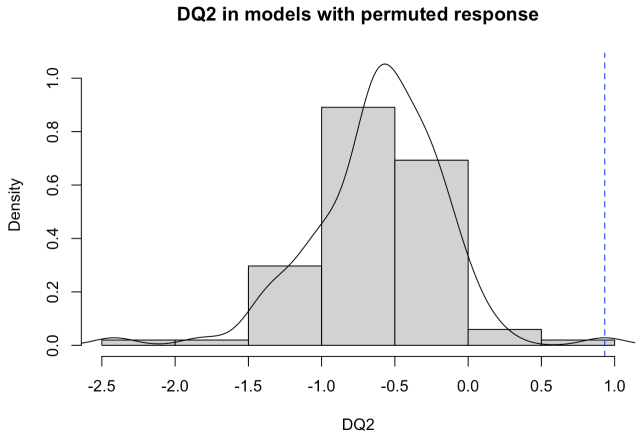
HIDE

```
plotQ2(train.COPLS_res)
```



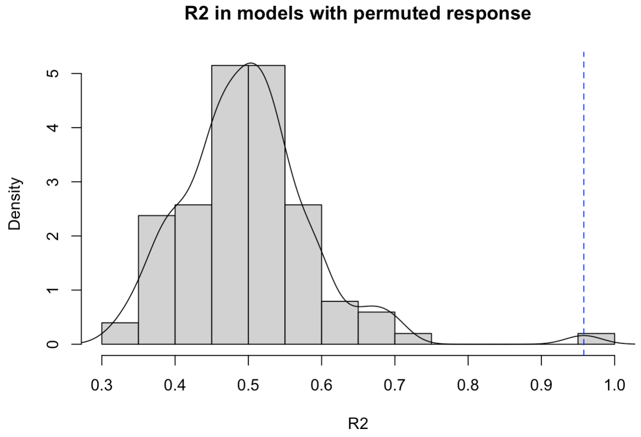
HIDE

```
plotDQ2(train.COPLS_res)
```



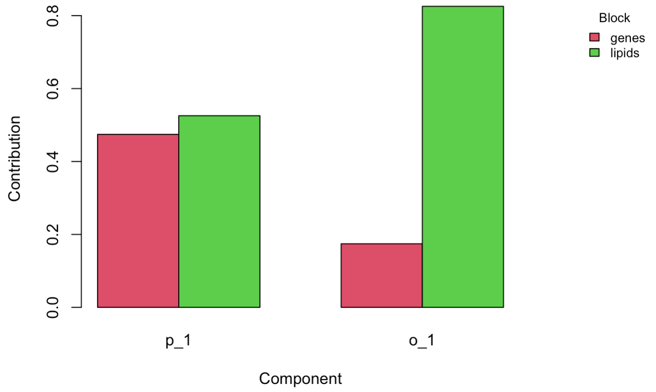
HIDE

```
plotR2(train.COPLS_res)
```



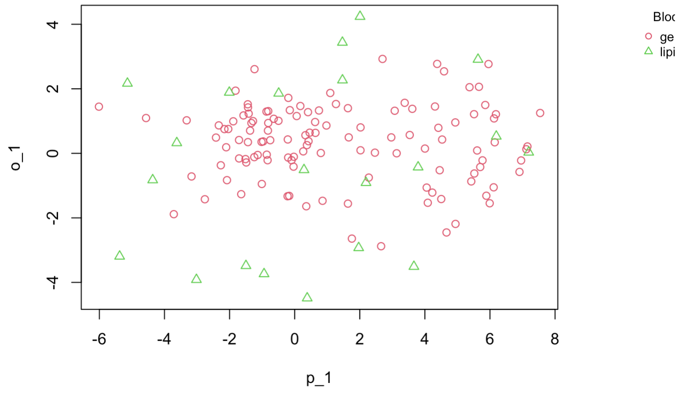
HIDE

```
plotContribution(train.COPLS_res)
```



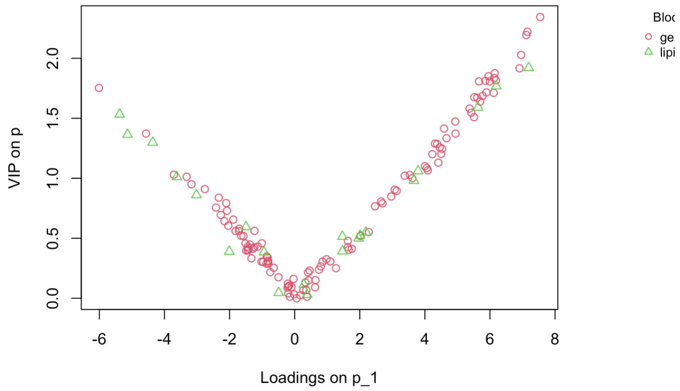
HIDE

```
ConsensusOPLS::plotLoadings(train.COPLS_res) #because mixOmics also has a function called plotLoadings()
```



HIDE

```
plotVIP(train.COPLS_res)
```



HIDE

```
test.COPLS_res <- predict(object = train.COPLS_res,
                          newdata = test.data)


data.frame(test.COPLS_res$class, Y.pred = test.COPLS_res$Y, True.genotype=metadata$genotype[test.observations])
```

```
##             class    margin softmax.ppar softmax.wt Y.pred.ppar  Y.pred.wt
## 18_wt_fish     wt 1.8717180    0.0000000  1.0000000 -0.43585898  1.4358590
## 12_wt_ref      wt 1.0916457    0.0000000  1.0000000 -0.04582286  1.0458229
## 3_wt_sun       wt 1.7609730    0.0000000  1.0000000 -0.38048650  1.3804865
## 1_wt_lin       wt 1.5142481    0.0000000  1.0000000 -0.25712406  1.2571241
## 14_wt_ref      wt 1.3271988    0.0000000  1.0000000 -0.16359938  1.1635994
## 30_ppar_lin  ppar 1.5919554    1.0000000  0.0000000  1.29597771 -0.2959777
## 37_ppar_lin  ppar 1.5319880    1.0000000  0.0000000  1.26599398 -0.2659940
## 32_ppar_coc  ppar 0.5594104    0.9333161  0.0666839  0.77970518  0.2202948
## 34_ppar_sun  ppar 1.6763100    1.0000000  0.0000000  1.33815499 -0.3381550
## 25_ppar_sun  ppar 2.2061198    1.0000000  0.0000000  1.60305991 -0.6030599
##             True.genotype
## 18_wt_fish             wt
## 12_wt_ref              wt
## 3_wt_sun               wt
## 1_wt_lin               wt
## 14_wt_ref              wt
## 30_ppar_lin          ppar
## 37_ppar_lin          ppar
## 32_ppar_coc          ppar
## 34_ppar_sun          ppar
## 25_ppar_sun          ppar
```