CCA

CODE

Florence Mehl

March 04, 2025

- Nutrimouse dataset
  - 1. Perform CCA (mixOmics::rcc) between 20 genes and all lipids. Investigate correlations, variable relationship and sample distribution with plots.
    - Plot the scores
    - Plot the loadings
  - 2. Perform CCA with scaled datasets and observe the difference
    - Plot the scores
    - Plot the loadings
  - 3. Perform regularized CCA with all genes and lipids.
    - Plot the scores
    - Plot the loadings

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
```

1. Perform CCA (mixOmics::rcc) between 20 genes and all lipids. Investigate correlations, variable relationship and sample distribution with plots.

The gene expression data is reduced to 20 genes so that the number of variables is less than the number of samples, to perform an unregularized CCA.

HIDE

```
nutrimouse$gene_selected <- nutrimouse$gene[, 1:20]
str(nutrimouse$gene_selected)
```

```
## 'data.frame':    40 obs. of  20 variables:
##  $ X36b4: num  -0.42 -0.44 -0.48 -0.45 -0.42 -0.43 -0.53 -0.49 -0.36 -0.5 ...
##  $ ACAT1: num  -0.65 -0.68 -0.74 -0.69 -0.71 -0.69 -0.62 -0.69 -0.66 -0.62 ...
##  $ ACAT2: num  -0.84 -0.91 -1.1 -0.65 -0.54 -0.8 -1 -0.91 -0.74 -0.79 ...
##  $ ACBP : num  -0.34 -0.32 -0.46 -0.41 -0.38 -0.32 -0.44 -0.37 -0.39 -0.36 ...
##  $ ACC1 : num  -1.29 -1.23 -1.3 -1.26 -1.21 -1.13 -1.22 -1.29 -1.15 -1.21 ...
##  $ ACC2 : num  -1.13 -1.06 -1.09 -1.09 -0.89 -0.79 -1 -1.06 -1.08 -0.82 ...
##  $ ACOTH: num  -0.93 -0.99 -1.06 -0.93 -1 -0.93 -0.94 -1.05 -0.88 -0.92 ...
##  $ ADISP: num  -0.98 -0.97 -1.08 -1.02 -0.95 -0.97 -0.94 -1.02 -0.98 -0.99 ...
##  $ ADSS1: num  -1.19 -1 -1.18 -1.07 -1.08 -1.07 -1.05 -1.16 -1.05 -1 ...
##  $ ALDH3: num  -0.68 -0.62 -0.75 -0.71 -0.76 -0.75 -0.67 -0.75 -0.66 -0.69 ...
##  $ AM2R : num  -0.59 -0.58 -0.66 -0.65 -0.59 -0.55 -0.66 -0.66 -0.53 -0.62 ...
##  $ AOX  : num  -0.16 -0.12 -0.16 -0.17 -0.31 -0.23 -0.09 -0.22 -0.06 -0.23 ...
##  $ BACT : num  -0.22 -0.32 -0.32 -0.32 -0.31 -0.29 -0.25 -0.21 -0.15 -0.2 ...
##  $ BIEN : num  -0.89 -0.88 -0.89 -0.77 -0.97 -0.84 -0.86 -0.9 -0.74 -0.76 ...
##  $ BSEP : num  -0.69 -0.6 -0.7 -0.67 -0.68 -0.55 -0.67 -0.66 -0.6 -0.58 ...
##  $ Bcl.3: num  -1.18 -1.07 -1.17 -1.12 -0.93 -1.08 -1.03 -1.01 -1.01 -1.1 ...
##  $ C16SR: num  1.66 1.65 1.57 1.61 1.66 1.7 1.58 1.62 1.72 1.55 ...
##  $ CACP : num  -0.92 -0.87 -1.02 -0.89 -0.93 -0.97 -0.97 -0.96 -0.85 -0.95 ...
##  $ CAR1 : num  -0.97 -0.92 -0.98 -0.97 -1.06 -1.03 -0.91 -1.11 -0.85 -0.99 ...
##  $ CBS  : num  -0.26 -0.36 -0.4 -0.39 -0.35 -0.31 -0.32 -0.4 -0.26 -0.39 ...
```

HIDE

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

Plot the scores

The sample distribution plot can be performed with **variates**, sample coordinates in the new reference (rotated axes) for each of the two blocks.

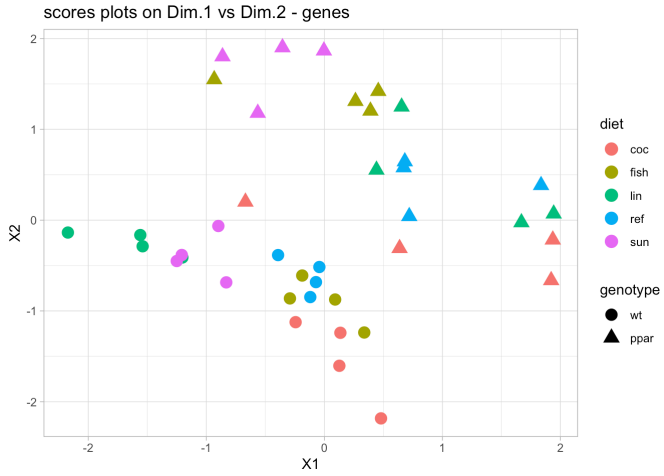
HIDE

```
str(cca.res$variates)
```

```
## List of 2
##  $ X: num [1:40, 1:2] -1.203 -1.25 -0.831 0.338 -0.119 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:40] "1" "2" "3" "4" ...
##   .. ..$ : NULL
##  $ Y: num [1:40, 1:2] -1.203 -1.25 -0.831 0.338 -0.119 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:40] "1" "2" "3" "4" ...
##   .. ..$ : NULL
```

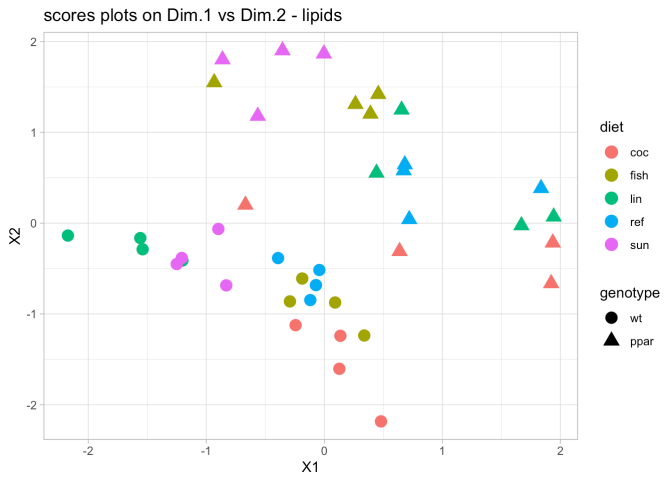
HIDE

```
cca.res_scores_genes <- data.frame(metadata, cca.res$variates$X)

ggplot(cca.res_scores_genes, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - genes") +
  theme_light()
```



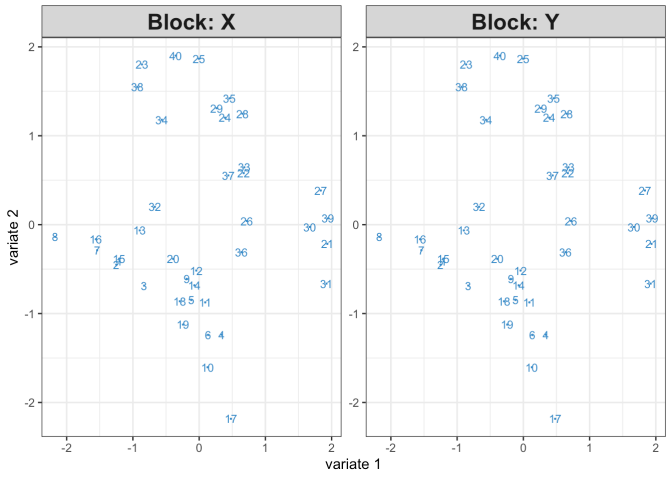
HIDE

```
cca.res_scores_lipids <- data.frame(metadata, cca.res$variates$Y)

ggplot(cca.res_scores_lipids, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - lipids") +
  theme_light()
```



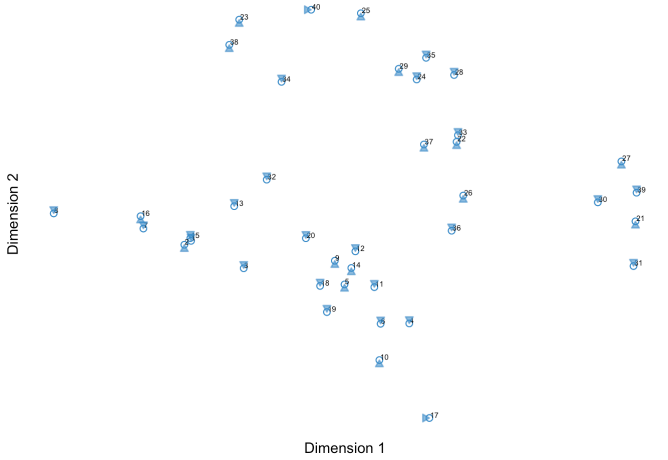
HIDE

```
plotIndiv(cca.res)
```



HIDE

```
plotArrow(cca.res)
```

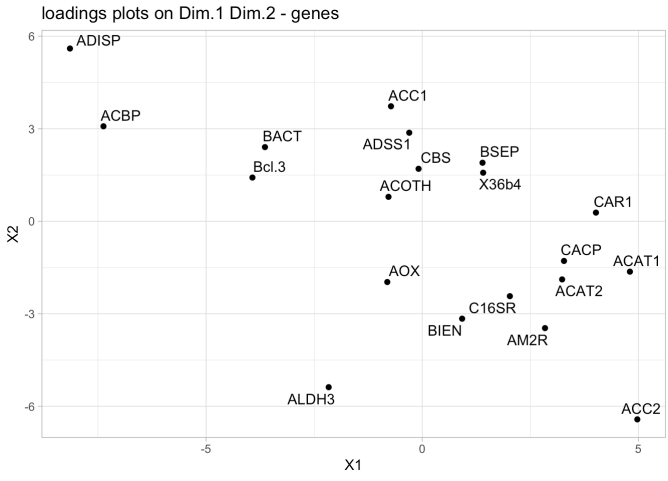


Plot the loadings

Variable relationship is obtained from **loadings** or with `plotVar` .

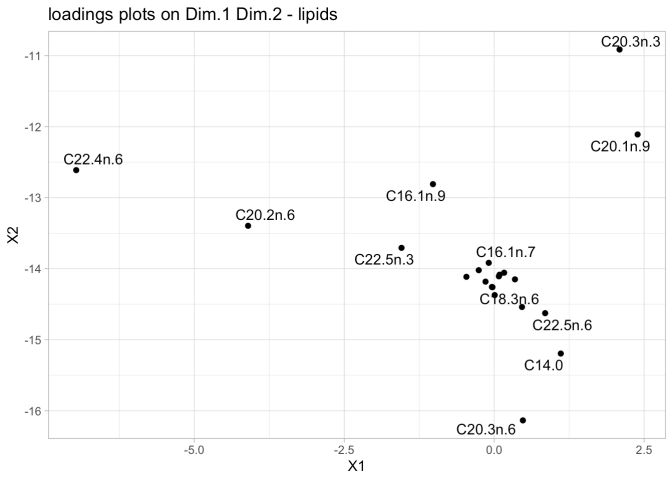
HIDE

```
cca.res_loadings_genes <- data.frame(cca.res$loadings$X)
cca.res_loadings_genes$variable <- rownames(cca.res_loadings_genes)

ggplot(cca.res_loadings_genes, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - genes") +
  theme_light()
```



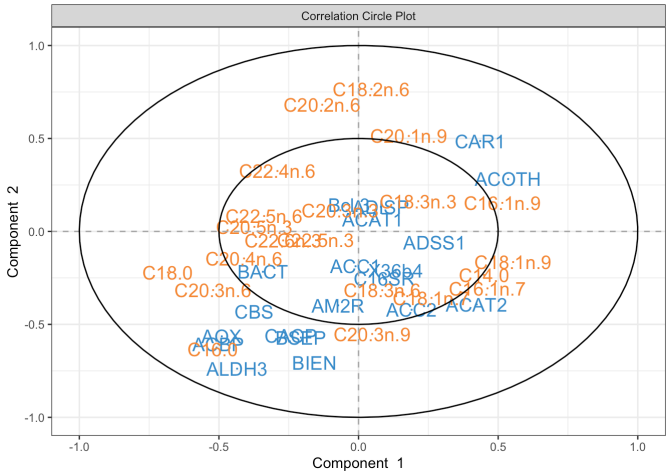
HIDE

```
cca.res_loadings_lipids <- data.frame(cca.res$loadings$Y)
cca.res_loadings_lipids$variable <- rownames(cca.res_loadings_lipids)

ggplot(cca.res_loadings_lipids, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - lipids") +
  theme_light()
```



HIDE

```
plotVar(cca.res)
```



2. Perform CCA with scaled datasets and observe the difference

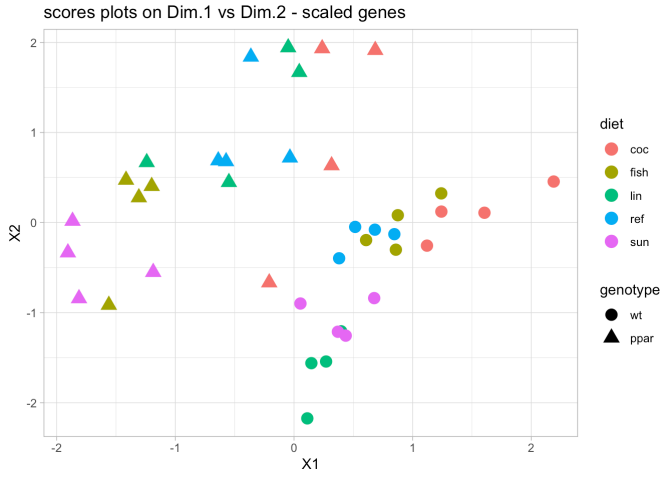
HIDE

```
cca.res.scale <- rcc(X=scale(nutrimouse$gene_selected, center=T, scale=T),
                     Y=scale(nutrimouse$lipid, center=T, scale=T), ncomp=2)
```

Plot the scores

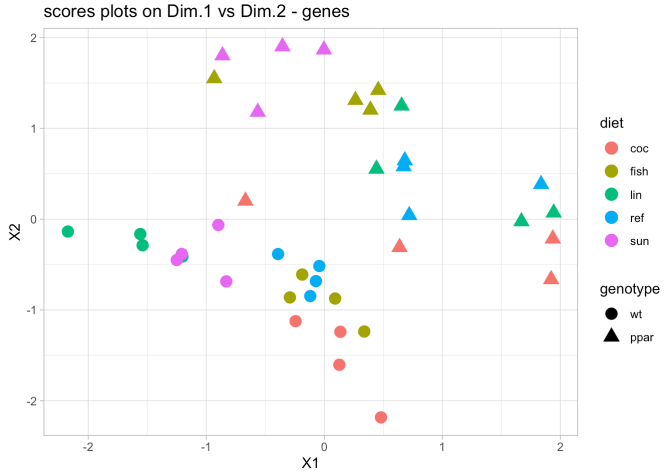
HIDE

```
cca.res.scale_scores_genes <- data.frame(metadata, cca.res.scale$variates$X)

ggplot(cca.res.scale_scores_genes, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - scaled genes") +
  theme_light()
```



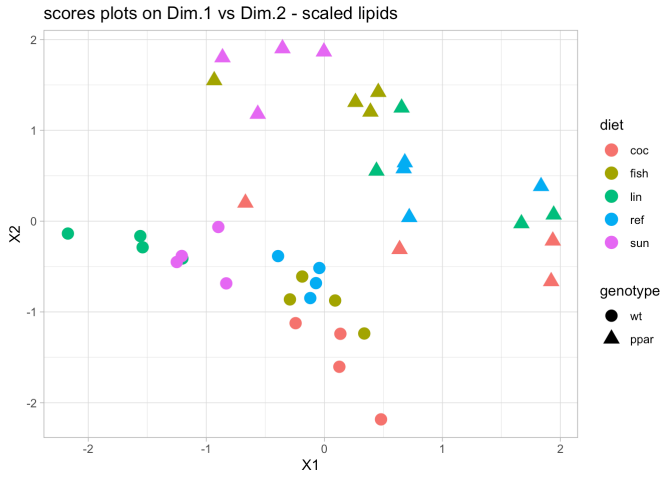
HIDE

```
ggplot(cca.res_scores_genes, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - genes") +
  theme_light()
```



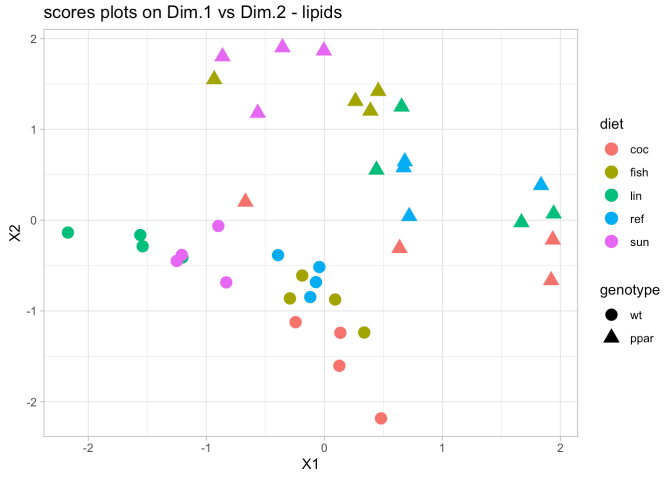
HIDE

```
cca.res.scale_scores_lipids <- data.frame(metadata, cca.res.scale$variates$Y)

ggplot(cca.res_scores_lipids, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - scaled lipids") +
  theme_light()
```



HIDE

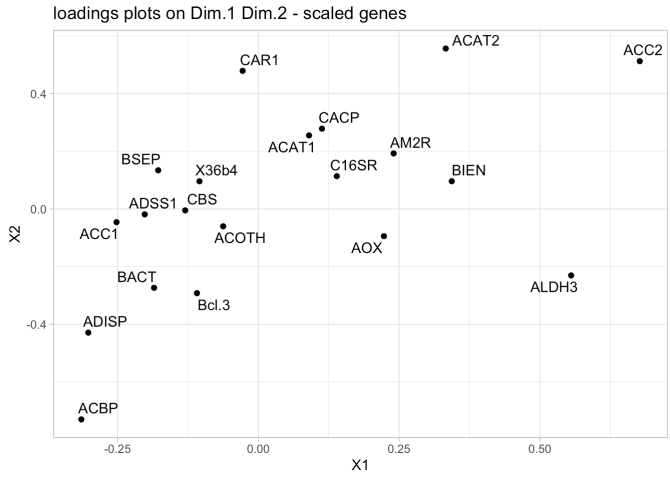
```
ggplot(cca.res_scores_lipids, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - lipids") +
  theme_light()
```



Plot the loadings

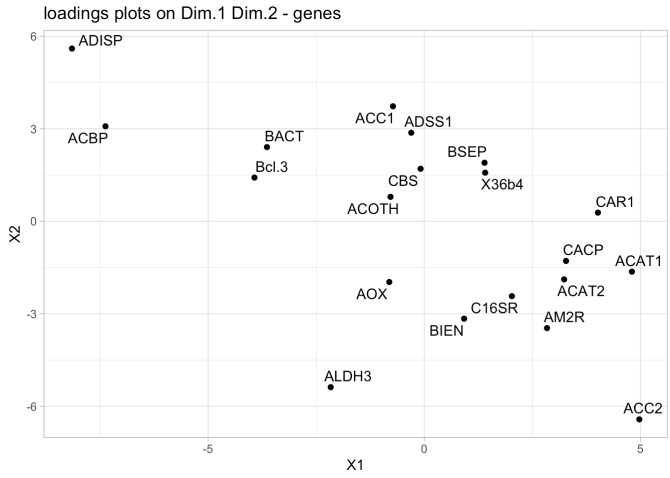
HIDE

```
cca.res.scale_loadings_genes <- data.frame(cca.res.scale$loadings$X)
cca.res.scale_loadings_genes$variable <- rownames(cca.res.scale_loadings_genes)

ggplot(cca.res.scale_loadings_genes, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - scaled genes") +
  theme_light()
```



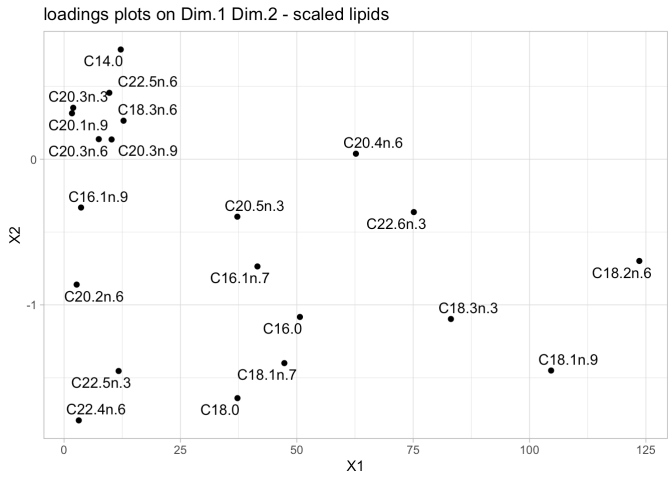
HIDE

```
ggplot(cca.res_loadings_genes, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - genes") +
  theme_light()
```



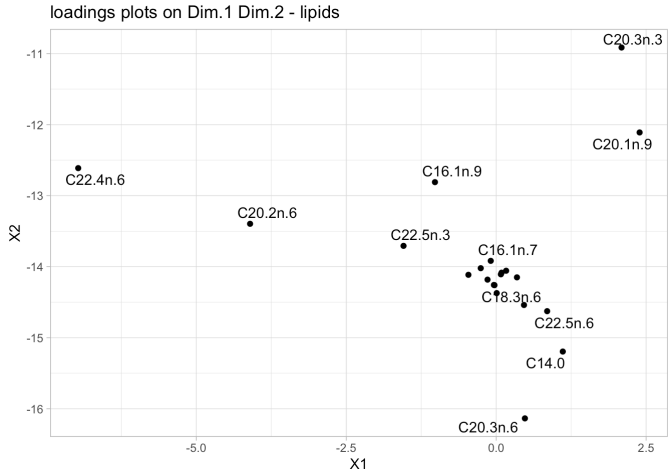
HIDE

```
cca.res.scale_loadings_lipids <- data.frame(cca.res.scale$loadings$Y)
cca.res.scale_loadings_lipids$variable <- rownames(cca.res.scale_loadings_lipids)

ggplot(cca.res.scale_loadings_lipids, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - scaled lipids") +
  theme_light()
```



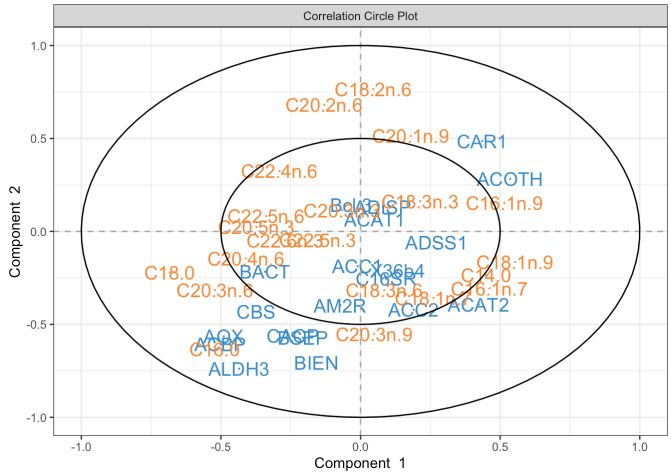
HIDE

```
ggplot(cca.res_loadings_lipids, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - lipids") +
  theme_light()
```



HIDE

```
plotVar(cca.res)
```



3. Perform regularized CCA with all genes and lipids.

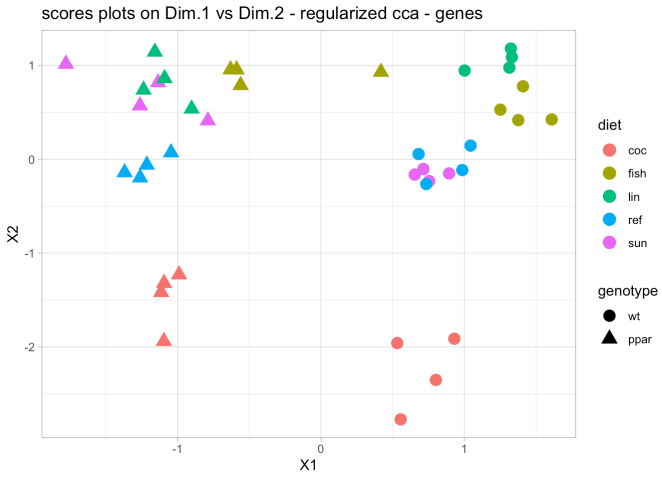
HIDE

```
rcca.res <- rcc(X=nutrimouse$gene, Y=nutrimouse$lipid, ncomp=2, method="shrinkage")
```

Plot the scores

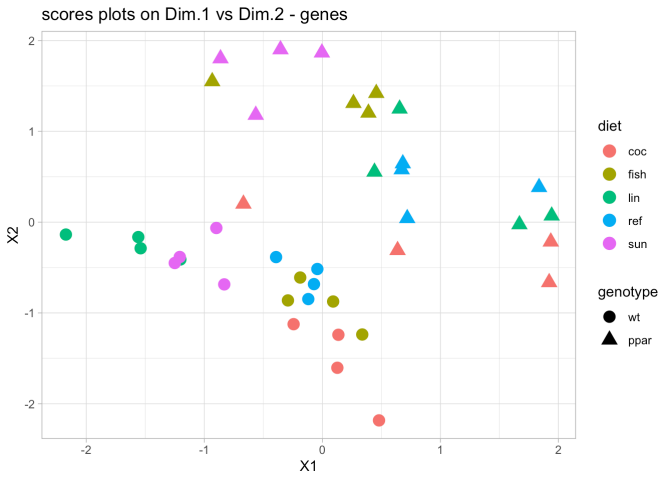
HIDE

```
rcca.res_scores_genes <- data.frame(metadata, rcca.res$variates$X)

ggplot(rcca.res_scores_genes, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - regularized cca - genes") +
  theme_light()
```



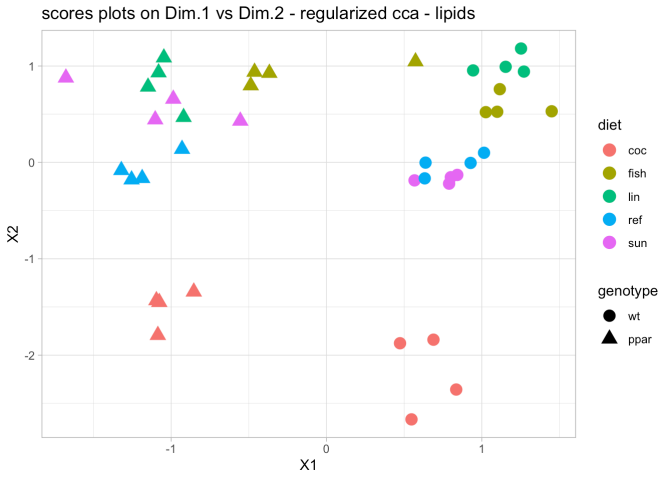
HIDE

```
ggplot(cca.res_scores_genes, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - genes") +
  theme_light()
```



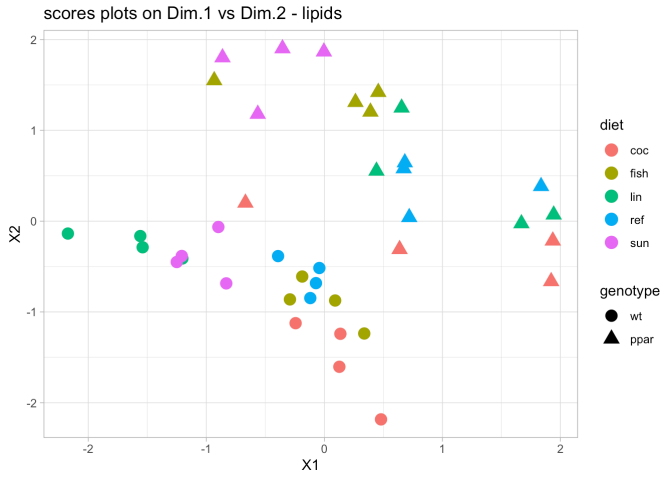
HIDE

```
rcca.res_scores_lipids <- data.frame(metadata, rcca.res$variates$Y)

ggplot(rcca.res_scores_lipids, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - regularized cca - lipids") +
  theme_light()
```



HIDE

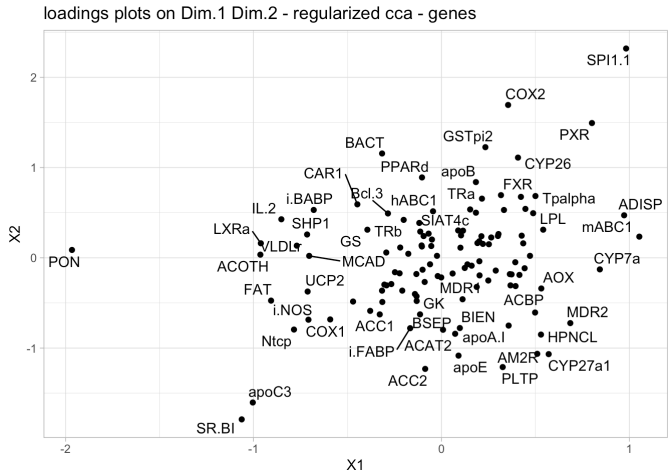
```
ggplot(cca.res_scores_lipids, aes(x=X1, y=X2, col=diet, shape = genotype)) +
  geom_point(size=4) +
  labs(title = "scores plots on Dim.1 vs Dim.2 - lipids") +
  theme_light()
```



Plot the loadings

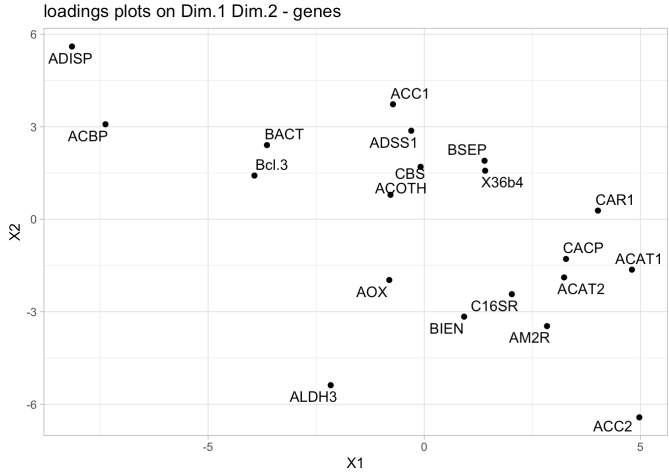
HIDE

```
rcca.res_loadings_genes <- data.frame(rcca.res$loadings$X)
rcca.res_loadings_genes$variable <- rownames(rcca.res_loadings_genes)

ggplot(rcca.res_loadings_genes, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - regularized cca - genes") +
  theme_light()
```



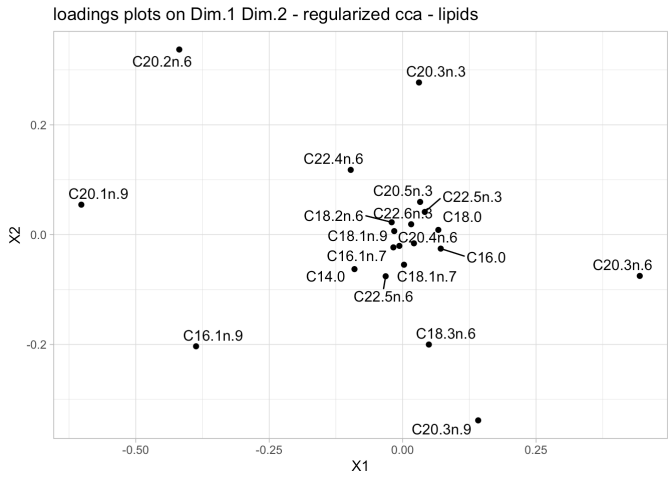
HIDE

```
ggplot(cca.res_loadings_genes, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - genes") +
  theme_light()
```



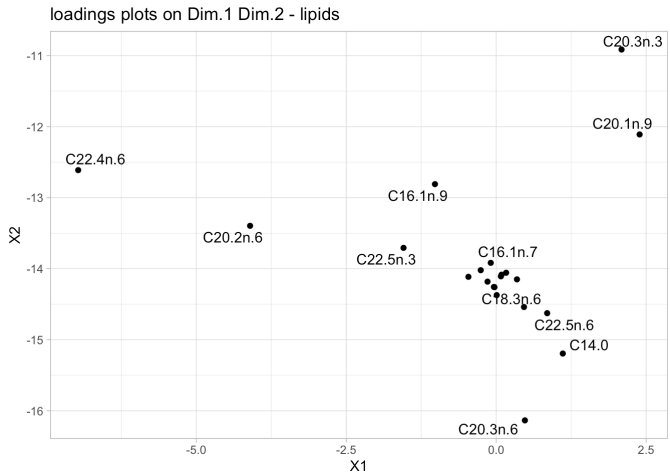
HIDE

```
rcca.res_loadings_lipids <- data.frame(rcca.res$loadings$Y)
rcca.res_loadings_lipids$variable <- rownames(rcca.res_loadings_lipids)

ggplot(rcca.res_loadings_lipids, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - regularized cca - lipids") +
  theme_light()
```



HIDE

```
ggplot(cca.res_loadings_lipids, aes(x=X1, y=X2, label=variable)) +
  geom_point() +
  geom_text_repel() +
  labs(title = "loadings plots on Dim.1 Dim.2 - lipids") +
  theme_light()
```



HIDE

```
plotVar(rcca.res)
```