

# 16S and cytokines combination

```
#token 35f0903098e0c9ef30faf7514e382e6bbd5c7179
#BiocManager::install("HMP2Data")

#devtools::install_github("jstansfield0/HMP2Data", auth_token = "35f0903098e0c9ef30faf7514e382e6bbd5c7179")

library(HMP2Data)
library(phyloseq)
library(SummarizedExperiment)
library(MultiAssayExperiment)
library(dplyr)
library(magrittr)
library(Hmisc)
library(colorspace)
library(ade4)
source("ScreePlot.R")
source("CIAPlots.R")

pathtoplots <- file.path(getwd(), "Plots/")#change me
```

## MOMS-PI

The MOMS-PI data can be loaded as follows.

### 16S data

Load 16S data as a matrix, rows are Greengene IDs, columns are sample names:

```
data("momspi16S_mtx")
```

Load the Greengenes taxonomy table as a matrix, rows are Greengene IDs, columns are taxonomic ranks:

```
data("momspi16S_tax")
# Check if Greengene IDs match between the 16S and taxonomy data
# all.equal(rownames(momspi16S_mtx), rownames(momspi16S_tax)) # Should be TRUE
```

Load the 16S sample annotation data as a matrix, rows are samples, columns are annotations:

```
data("momspi16S_samp")

# Check if sample names match between the 16S and sample data
# all.equal(colnames(momspi16S_mtx), rownames(momspi16S_samp)) # Should be TRUE
```

The momspi16S function assembles those matrices into a phyloseq object.

```
momspi16S_phyloseq <- momspi16S()
momspi16S_phyloseq

## phyloseq-class experiment-level object
## otu_table() OTU Table: [ 7665 taxa and 9107 samples ]
## sample_data() Sample Data: [ 9107 samples by 9 sample variables ]
## tax_table() Taxonomy Table: [ 7665 taxa by 7 taxonomic ranks ]
```

## Cytokine data

The MOMS-PI cytokine data can be loaded as a matrix, rownames are cytokine names, colnames are sample names:

```
data("momspiCyto_mtx")
dim(momspiCyto_mtx)
```

```
## [1] 29 872
```

Load the cytokine sample annotation data as a matrix, rows are samples, columns are annotations:

```
data("momspiCyto_samp")
dim(momspiCyto_samp)
```

```
## [1] 872 9
```

```
# Check if sample names match between the 16S and sample data
# all.equal(colnames(momspiCyto_mtx), rownames(momspiCyto_samp)) # Should be TRUE
```

The function `momspiCytokines` will make a `SummarizedExperiment` containing cytokine data

```
momspiCyto <- momspiCytokines()
momspiCyto
```

```
## class: SummarizedExperiment
## dim: 29 872
## metadata(0):
## assays(1): ''
## rownames(29): 1 2 ... 28 29
## rowData names(0):
## colnames(872): EP036702_K10_MVAX EP062329_K10_MP1P ...
## EP936022_K10_MVAX EP949081_K10_MVAX
## colData names(9): sample_id subject_id ... project_name file
```

The cytokine data contains data for 29 cytokines over 872 samples.

## Multi-table analysis

Combine 16S and cytokines data

```
#select data collected at the same visit
combined_samp <- merge(momspi16S_samp, momspiCyto_samp,
                      by = c("subject_id", "sample_body_site",
                             "project_name", "study_full_name",
                             "subject_gender", "subject_race",
                             "visit_number"))

table(combined_samp$visit_number)
```

```
##
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
## 115 86 80 99 78 60 51 48 59 39 19 10 3 2 1
```

Select first visit data, this assures that samples we look at were taken at the same time and at the first or second trimester. We don't have trimesters information in the data, but know it from the study design.

```
#select data from first visit only
combined_samp <- combined_samp[combined_samp$visit_number == 1,]
```

```
table(combined_samp$sample_body_site)#all vaginal samples
```

```
##
## vagina
##      115
```

The two objects we use for combined 16S and cytokines analysis are: ‘combined\_16S\_mtx’ and ‘combined\_Cyto\_mtx’. Phylogenetic information for those OTUs is available in ‘tax\_table(combined\_16S\_phyloseq)’ object.

```
#select 16S data for those samples
combined_16S_phyloseq <- subset_samples(momspi16S_phyloseq, file %in% combined_samp$file.x)
```

```
#get rid of otus that are not observed in any sample for this subset
combined_16S_phyloseq %<>%
  taxa_sums() %>%
  is_greater_than(0) %>%
  prune_taxa(combined_16S_phyloseq)
```

```
combined_16S_mtx <- otu_table(combined_16S_phyloseq)
```

```
combined_Cyto_mtx <- momspiCyto_mtx[, colnames(momspiCyto_mtx) %in% combined_samp$file.y ]
dim(combined_Cyto_mtx)
```

```
## [1] 29 115
```

We match the samples (contained in columns of both tables) by the file names contained in colnames of each table.

In ‘combined\_samp’ object the names of matched files names for 16S data are recorded in column ‘file.x’ and for cytokines data in column ‘file.y’.

```
#make sure all samples across 3 tables are in the same order
combined_samp <- combined_samp[order(combined_samp$subject_id),]
#reorder cytokines samples
combined_Cyto_mtx <- combined_Cyto_mtx[,combined_samp$file.y]
#reorder taxa samples
combined_16S_mtx <- combined_16S_mtx[,combined_samp$file.x]
```

## Co-inertia analysis

### Basics:

- Let  $Z_1$  and  $Z_2$  be 16S and cytokines tables respectively
- rows: same  $n$  women at first visit
- Columns:  $p_1$  taxa,  $p_2$  cytokines
- To visualize differences/similarities among taxa and cytokines data sets we view samples as rows while taxa and cytokines as columns
- PCA analysis for each table:  $(X, Q_X, D)$  and  $(Y, Q_Y, D)$

- Co-inertia axes:  $Y^T DX = K\Lambda^{1/2}A^T$  of decomposition  $(Y^T DX, Q_X, Q_Y)$
- Plot  $F_X = XA$  and  $F_Y = YK$

```
combined_16S_mtx <- t(combined_16S_mtx)
combined_16S_mtx <- combined_16S_mtx/apply(combined_16S_mtx, 1, sum)
combined_Cyto_mtx <- t(combined_Cyto_mtx)

#cut the last 5 characters that correspond to the -omics type identifier
rownames(combined_Cyto_mtx) <- substr(
  rownames(combined_Cyto_mtx), 1, nchar(rownames(combined_Cyto_mtx))-5)

rownames(combined_16S_mtx) <- substr(
  rownames(combined_16S_mtx), 1, nchar(rownames(combined_16S_mtx))-5)

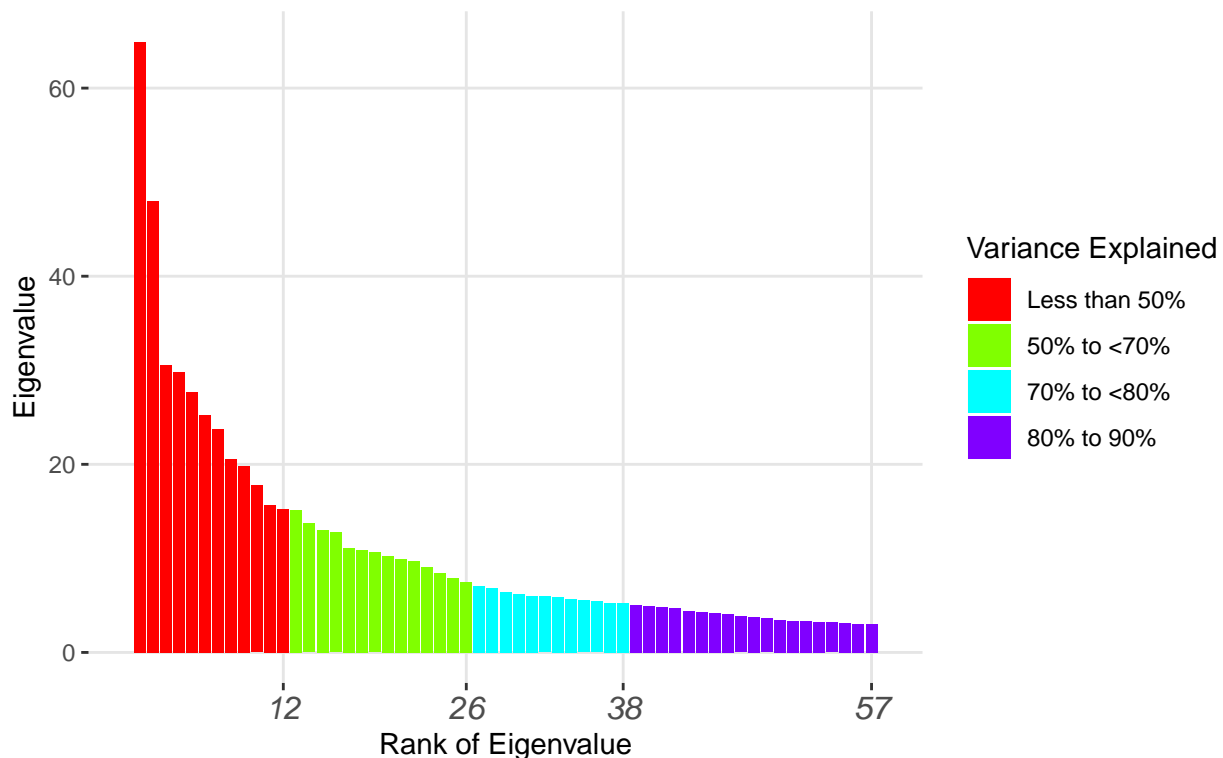
#make sure all rownames match
all(rownames(combined_16S_mtx) == rownames(combined_16S_mtx))

## [1] TRUE

16S PCA on centered and scaled data

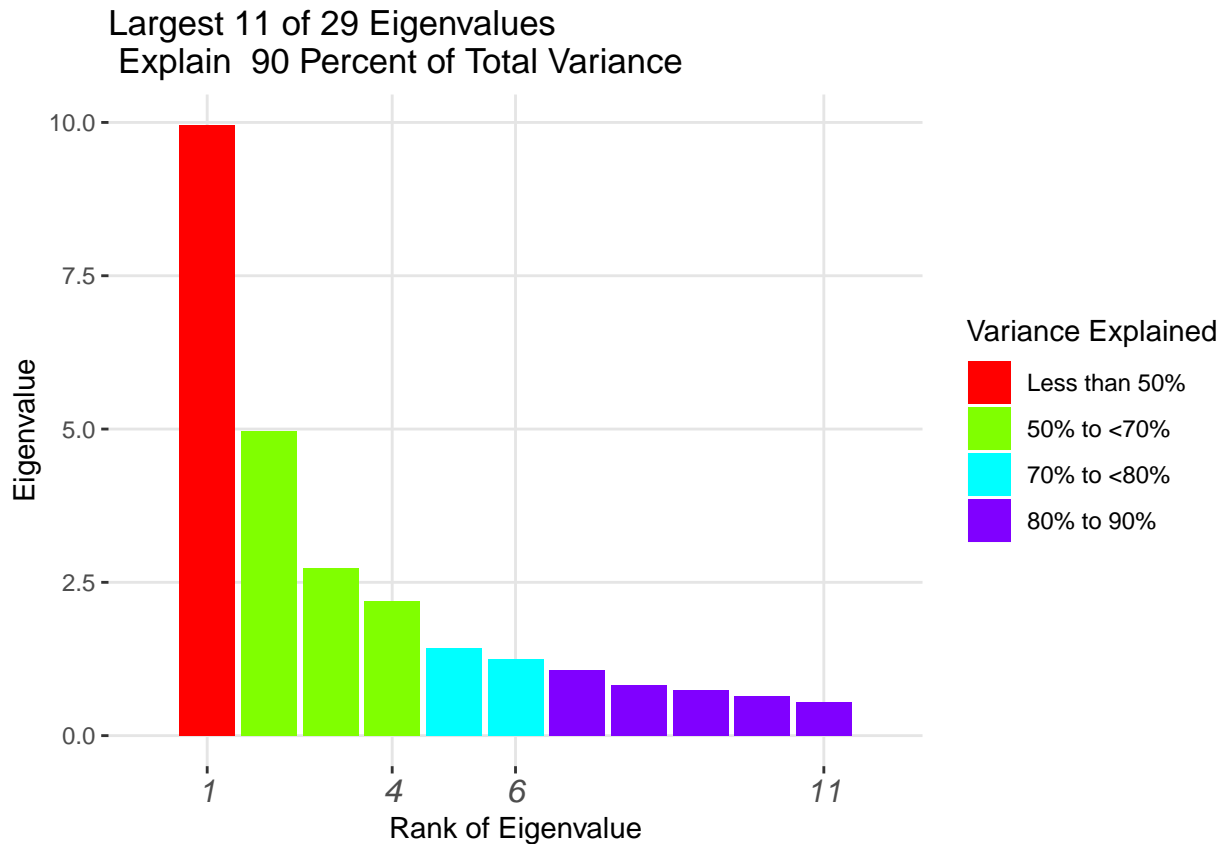
taxa.pca <- dudi.pca(combined_16S_mtx, scannf=FALSE, nf =61,
                    center = TRUE, scale = TRUE)
breaks <- c(50, 70, 80, 90)
scree.taxa.pca <- screePlot(taxa.pca, breaks) +
  theme( panel.background = element_rect(fill = "white"),
        panel.grid.major = element_line(colour = "grey90"))
scree.taxa.pca
```

Largest 57 of 114 Eigenvalues  
Explain 90 Percent of Total Variance



Cytokines PCA on centered and scaled data

```
cyto.pca <- dudi.pca(combined_Cyto_mtx, scannf=FALSE, nf =61,  
                    center = TRUE, scale = TRUE)  
breaks <- c(50, 70, 80, 90)  
scree.cyto.pca <- screePlot(cyto.pca, breaks) +  
  theme( panel.background = element_rect(fill = "white"),  
        panel.grid.major = element_line(colour = "grey90"))  
scree.cyto.pca
```



## Combine the tables using co-inertia

```
coin <- coinertia(taxa.pca, cyto.pca, scannf = FALSE, nf = 2)
```

RV coefficient – measure of similarity between 16S and cytokines tables

```
RV<- coin$RV  
RV
```

```
## [1] 0.1046239
```

Sample scores plots. Length of the arrows indicates the samples that have larger differences across two data sets.

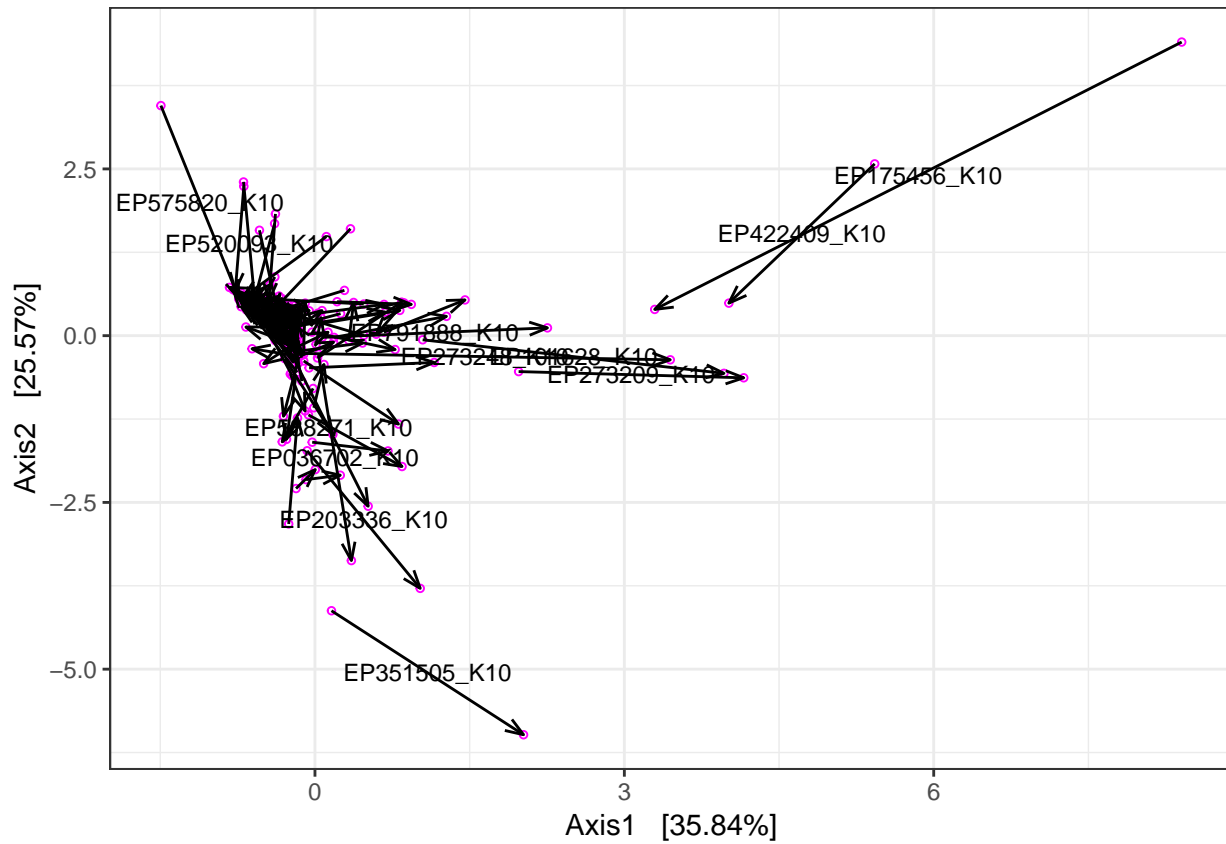
```
Samp.coin <- CoinertiaPlot(coin = coin,  
                          Quant = 0.9, Prop.Var = 0.9,  
                          Env_Var=NULL,
```

```

Env_Var2 = NULL,
color=NULL, shape=NULL,
PtColor= "magenta",PtShape=1, PtSize=1,
linetype=1, LblSize=3,
LabelsOpt = NULL,
ArrLen=0.10, ArrAngle=20)

```

```
Samp.coin$p + theme_bw()
```



Samples with largest difference across two data sets. Samples with arrow lengths in 0.9 quantile are chosen.

```
#Taxa with major differences across two sets
```

```
rownames(Samp.coin$Dissimilarity[Samp.coin$Dissimilarity$Quantile >= 0.9, ])
```

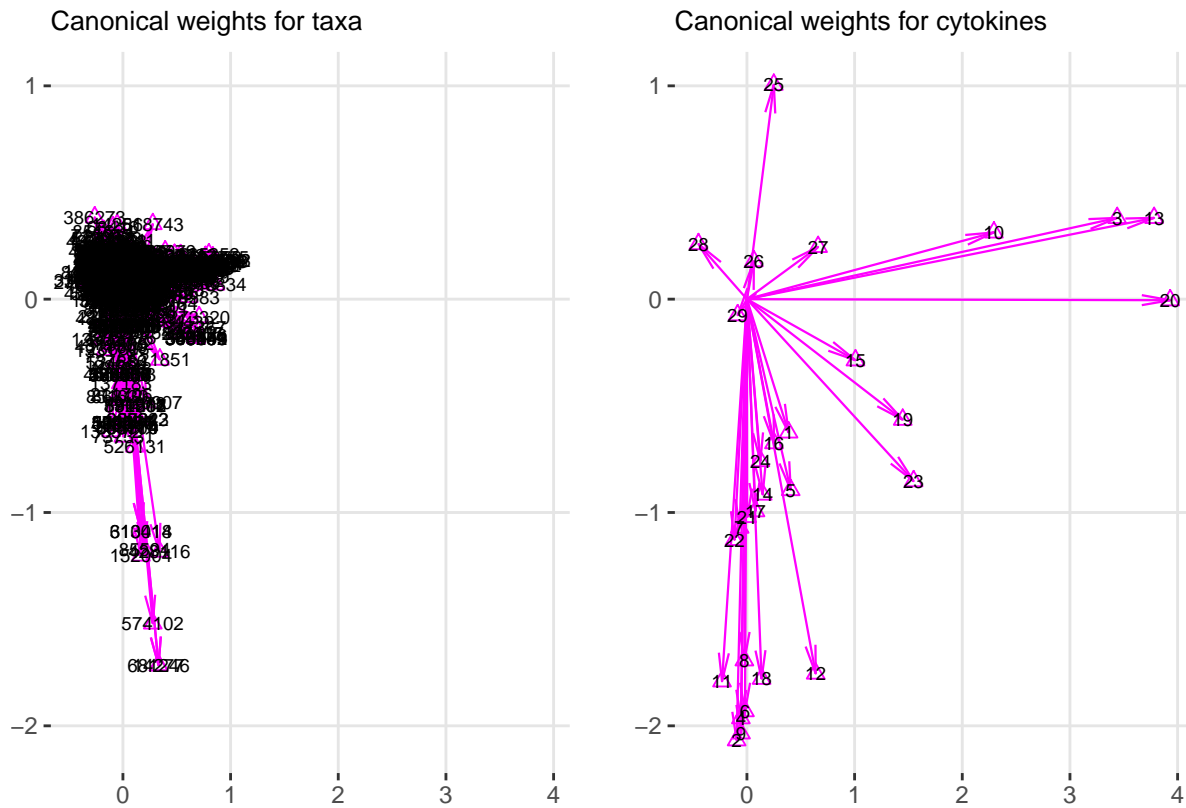
```
## [1] "EP520093_K10" "EP273209_K10" "EP791888_K10" "EP203336_K10"
## [5] "EP588271_K10" "EP422409_K10" "EP351505_K10" "EP101628_K10"
## [9] "EP575820_K10" "EP036702_K10" "EP273248_K10" "EP175456_K10"
```

Plots of taxa weights

```
#Plot Canonical weights based on No BV taxa as reference point
```

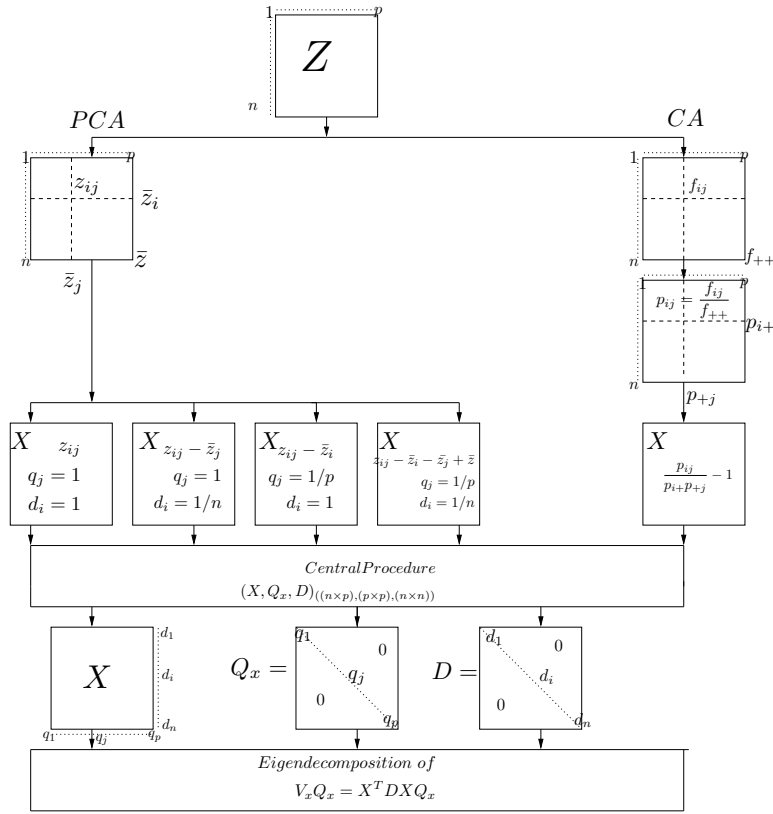
```
CW.coin.ca <- PlotCW(coin = coin, name="Coin", path = pathtoplots, color = "magenta",
  Title1 = "Canonical weights for taxa",
  Title2 = "Canonical weights for cytokines",
  Labels1 = colnames(combined_16S_mtx), Labels2 = colnames(combined_Cyto_mtx),
  scale = TRUE,
  PtShape=2, PtSize=2, linesize=0.4,
  linetype=1, LblSize=2.5,
  ArrLen=0.15, ArrAngle=10,
```

```
TitleSize = 10)
```



### Additional details about the method.

- $Z_{n \times p}$  data matrix
  - $n$  samples
  - $p$  variables (taxa)
- Transformation  $Z_{n \times p} \rightarrow X_{n \times p}$ 
  - Different transformation for each ordination method
  - Column weights:  $Q_{p \times p}$
  - Row weights:  $D_{n \times n}$
- Statistical triplet:  $(X, Q, D)$
- Eigen decomposition of  $(WD)_{n \times n} = XQX^T D$  or  $(VQ)_{p \times p} = X^T D X Q$
- Write:  $X = K\Lambda^{1/2}A^T$ 
  - $K^T D K = \mathbb{I}_r$ ,  $K$  - samples scores
  - $A^T Q A = \mathbb{I}_r$ ,  $A$  - taxa loadings
  - $r$  rank of  $X$



Given statistical triplet  $(X, Q, D)$  the PCA is

- $X_{n \times p}$  - column centered and scaled matrix
  - $n$  number of samples
  - $p$  number of taxa
- $Q = \mathbb{I}_p$
- Correlation PCA:  $Q = \text{diag}\{sd(z_1), \dots, sd(z_p)\}$ , where  $sd(z_j)$  standard deviation for  $j^{th}$  column of original data matrix  $Z$
- $D = \text{diag}\{\frac{1}{n}, \dots, \frac{1}{n}\} = \frac{1}{n} \mathbb{I}_n$
- Then eigendecomposition of

$$VQ = X^T D X Q = \frac{1}{n} X^T X = \Sigma$$

- Inertia:  $\mathcal{I} = \text{tr}(\Sigma) = \sum_{i=1}^p \sigma_i^2 = \sum_{i=1}^r \lambda_i$

Co-inertia analysis idea:

- Analysis of  $Z_1, Z_2, \dots, Z_K$ . Same rows, but different columns
- $Z_1$  – 16S data
- $Z_2$  – cytokines data



## 1. Co-Inertia Analysis

- Seeks a [common structure](#) between two datasets  $Z_1$  and  $Z_2$
- Two tables:  $Z_1$  and  $Z_2$  of dimensions  $n \times p_1$  and  $n \times p_2$
- Use any one table method to obtain triplets:  $(X, Q_X, D)$  and  $(Y, Q_Y, D)$
- [coinertia](#) – coinertia analysis function in ade4
- Co-inertia triplet:  $(Y^T DX, Q_X, Q_Y)$
- Decomposition:  $Y^T DX = K\Lambda^{1/2}A^T$
- Project  $X$  and  $Y$  onto  $A$  and  $K$  respectively
- Let  $A = XQ_X X^T D$  and  $B = YQ_Y Y^T D$
- RV coefficient:  $RV(A, B) = \frac{tr(AB^T)}{\sqrt{tr(AA^T)tr(BB^T)}} = \frac{\mathcal{I}_{XY}}{\sqrt{\mathcal{I}_X}\sqrt{\mathcal{I}_Y}}$
- $RV \in [0, 1]$  with values closer to 1 indicating stronger similarity among  $X$  and  $Y$

Objects in ade4

SVD of a transformed table  $X = K\Lambda^{1/2}A^T$

$k$  – rank of  $X$

matrix	dimensions	ade4	description
$Z$	$n \times p$		raw data; argument for dudi functions
$X$	$n \times p$	\$tab	transformed data
$Q$	$p \times p$	\$cw	diagonal matrix of column weights
$D$	$n \times n$	\$lw	diagonal matrix of row weights
$A$	$p \times k$	\$c1	the principal axes or column normed scores
$K$	$n \times k$	\$l1	the row normed scores
$\Lambda$	$k \times k$	\$eig	diagonal matrix of the $k$ largest eigenvalues
$A\Lambda^{1/2}$	$p \times k$	\$co	column coordinates
$K\Lambda^{1/2}$	$n \times k$	\$li	the row coordinates