Bioreactor time-course analysis

Olivier Chapleur

Contents

1 Data

In this data set, **three bioreactors** with similar performances were considered as replicates. Different parameters were measured accross time in the three bioreactors.

Performance data: Based on chemical measurement, the time course evolution of a set of parameters was measured (CH4, C02, acetate, propionate).

Metabolites data: The time course evolution of 20 selected metabolites was measured with GCMS.

Microbial data: DNA from samples taken across time was extracted and sequenced. (16S metabarcoding).

2 Data preprocessing

Metabolites (GCMS) data are log transformed.

Microbial data

1) are filtered (only OTUs with at least 1% of abundance in at least 1 sample are kept = 51 OTUs).
2) a count of 1 sequence is added to each sample/OTU (to avoid 0 in the datamatrix)
3) relative abundance is calculated
4) obtain data is clr transformed

Performance data is not transformed.

```
## There are 51 OTUs after 0.01 % filter
```

3 Spline smoothing

All the data are modelled with spline smoothing with the Linear Mixed Model Splines framework (package `lmms`).

As a reminder, the LMMS modeling step tests 4 different models for each OTUs. 0 = linear model, 1 = linear mixed effect model spline (LMMS) with defined basis, 2 = LMMS taking subject-specific random intercept, 3 = LMMS with subject specific intercept and slope.

3.1 Metabolites

```
## Data-driven Linear Mixed-Effect Model Splines
##  Profiles were modelled for 20 features with  48 time points.
##
##  Basis:
##  [1] "p-spline"
##
##  Knots:
##
##  [1] 17.57143 26.14286 32.71429 38.28571 44.14286 50.57143
##
##  Time points:
##
##   [1] 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
## [24] 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
## [47] 56 57
##
##  Table of models used to model profiles:
##  0  2  3
## 10  4  6
##
##  Profiles not modelled:
## [1] "All features were modelled"
```

For the metabolites data, 10 molecules were modelled with a straight line, 4 with LMMS with subject-specific random intercept, 6 with LMMS with subject specific intercept and slope.

3.2 OTUs

```
## Data-driven Linear Mixed-Effect Model Splines
##  Profiles were modelled for 51 features with  48 time points.
##
##  Basis:
##  [1] "p-spline"
##
##  Knots:
##
##  [1] 17.57143 26.14286 32.71429 38.28571 44.14286 50.57143
##
##  Time points:
##
##   [1] 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
## [24] 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
```

```
## [47] 56 57
##
##  Table of models used to model profiles:
##  0  1  2
## 30 19  2
##
##  Profiles not modelled:
## [1] "All features were modelled"
```

For the microbiome data, 30 OTUs were modelled with a straight line, 19 with LMMS and 2 with LMMS with subject specific random intercept.

3.3 Performance data

```
## Data-driven Linear Mixed-Effect Model Splines
##  Profiles were modelled for 2 features with  48 time points.
##
##  Basis:
##  [1] "p-spline"
##
##  Knots:
##
##  [1]   5.428571  13.000000  27.142857  44.142857  66.000000 113.000000
##
##  Time points:
##
##   [1] 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
## [24] 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
## [47] 56 57
##
##  Table of models used to model profiles:
## 1
## 2
##
##  Profiles not modelled:
## [1] "All features were modelled"
```

```
## Data-driven Linear Mixed-Effect Model Splines
##  Profiles were modelled for 2 features with  48 time points.
##
##  Basis:
##  [1] "p-spline"
##
##  Knots:
##
##  [1]  3.5 11.0 19.5 29.0 40.5 57.0 92.5
##
##  Time points:
##
##   [1] 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
## [24] 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
## [47] 56 57
##
##  Table of models used to model profiles:
```

```
## 3
## 2
##
##  Profiles not modelled:
## [1] "All features were modelled"
```

Acetate and Propionate were modelled with LMMS with subject specific random intercept. CH4 and CO2 were modelled with LMMS with subject specific intercept and slope.

4 Filtering of the obtained profiles

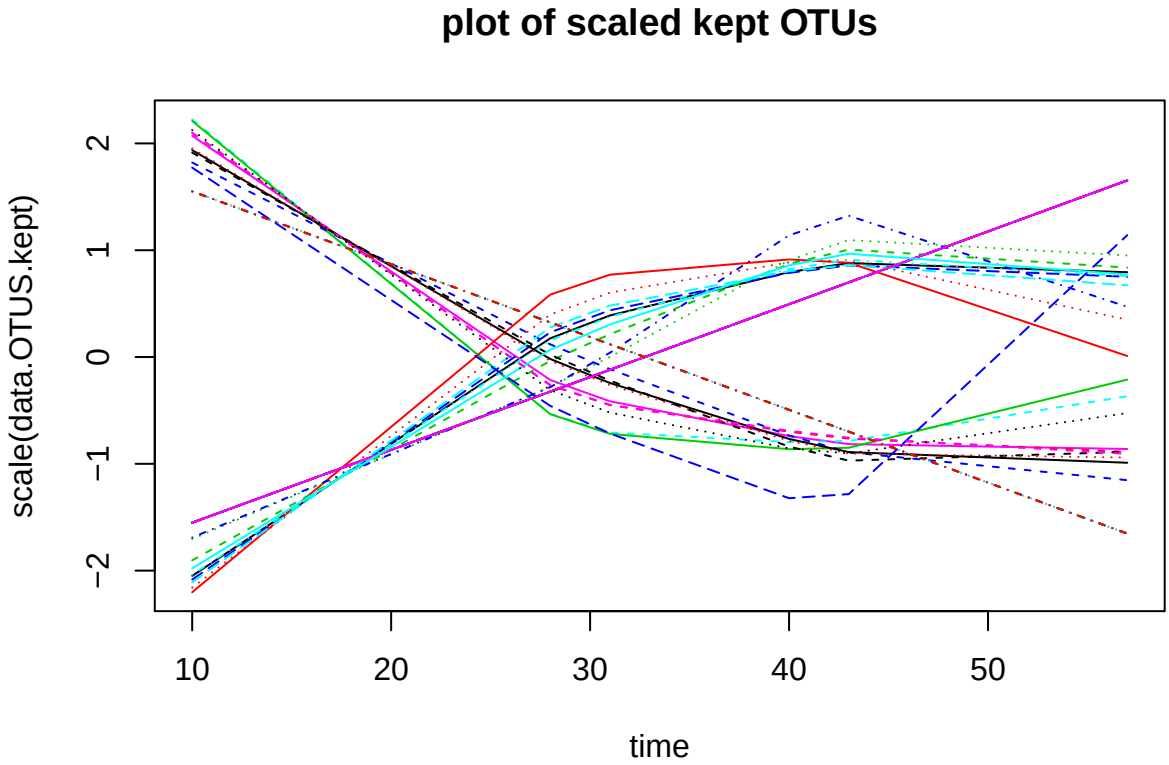
Straight line modelling can occur when the inter-individual variation is too high. To remove the noisy profiles, we first use the Breusch-Pagan test, which tests the homo-sedasticity of the residues. We then add a filter on the mean squared error to reduce the dispersion of the residues around the line.

4.1 OTUs

The 44 selected OTUs are listed below as well as the results of the filtering test.

```
##  [1] "OTU_1"   "OTU_10"  "OTU_11"  "OTU_13"  "OTU_130" "OTU_14"  "OTU_15"
##  [8] "OTU_16"  "OTU_169" "OTU_17"  "OTU_18"  "OTU_19"  "OTU_2"   "OTU_20"
## [15] "OTU_21"  "OTU_24"  "OTU_25"  "OTU_26"  "OTU_28"  "OTU_29"  "OTU_30"
## [22] "OTU_304" "OTU_31"  "OTU_35"  "OTU_38"  "OTU_4"   "OTU_41"  "OTU_44"
## [29] "OTU_45"  "OTU_46"  "OTU_5"   "OTU_50"  "OTU_59"  "OTU_6"   "OTU_60"
## [36] "OTU_61"  "OTU_65"  "OTU_68"  "OTU_7"   "OTU_74"  "OTU_75"  "OTU_8"
## [43] "OTU_82"  "OTU_97"
```

```
##  MSE.filter       BP.test
##  Mode :logical   Mode :logical
##  FALSE:6         FALSE:1
##  TRUE :45        TRUE :50
```



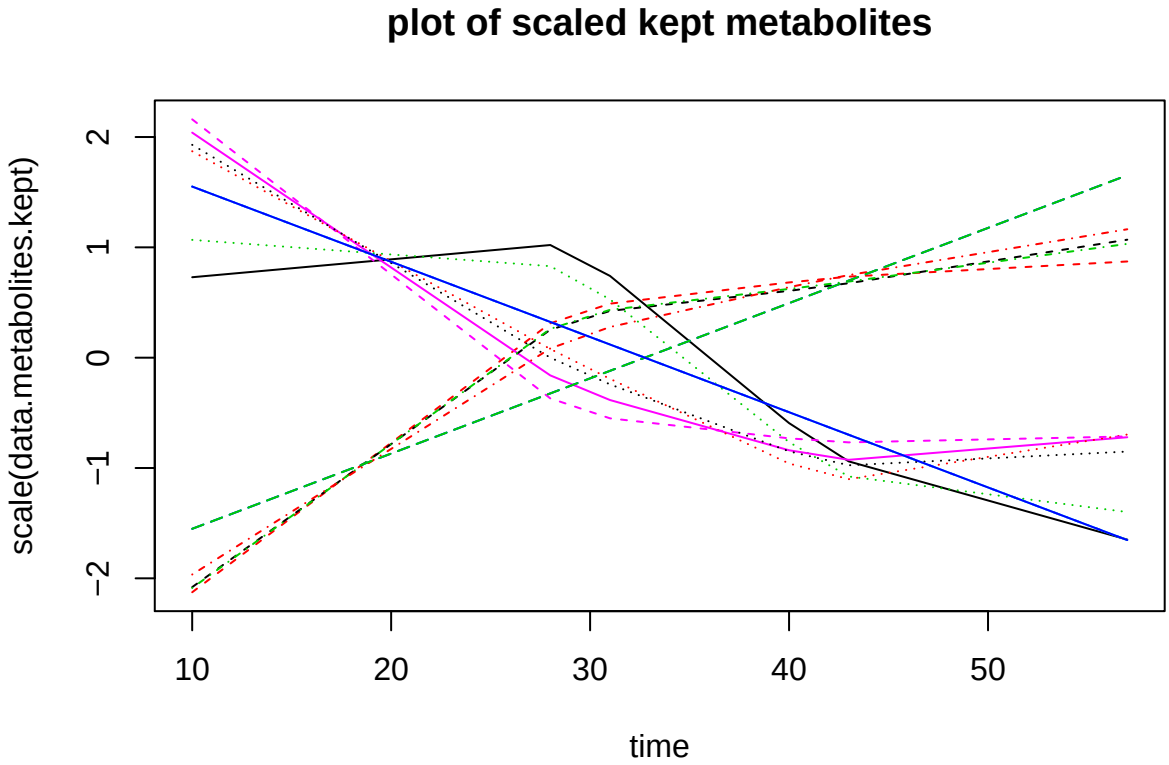
The filtering step removed 7 OTUs. In the above graph, time profiles of the selected OTUs are displayed.

4.2 Metabolites

The 16 selected metabolites are listed below as well as the results of the filtering test.

```
##  [1] "M106T894"  "M179T1018" "M205T1473" "M207T1196" "M229T1227"
##  [6] "M271T1466" "M285T1569" "M290T1524" "M291T1584" "M292T1383"
## [11] "M308T1437" "M310T1500" "M357T2099" "M379T1799" "M398T1643"
## [16] "M415T2220"
```

```
##    molecule           modelsUsed   BP.test
##  Length:20          Min.   :0.0   Mode :logical
##  Class :character   1st Qu.:0.0   FALSE:4
##  Mode  :character   Median :1.0   TRUE :16
##                     Mean   :1.3
##                     3rd Qu.:3.0
##                     Max.   :3.0
```

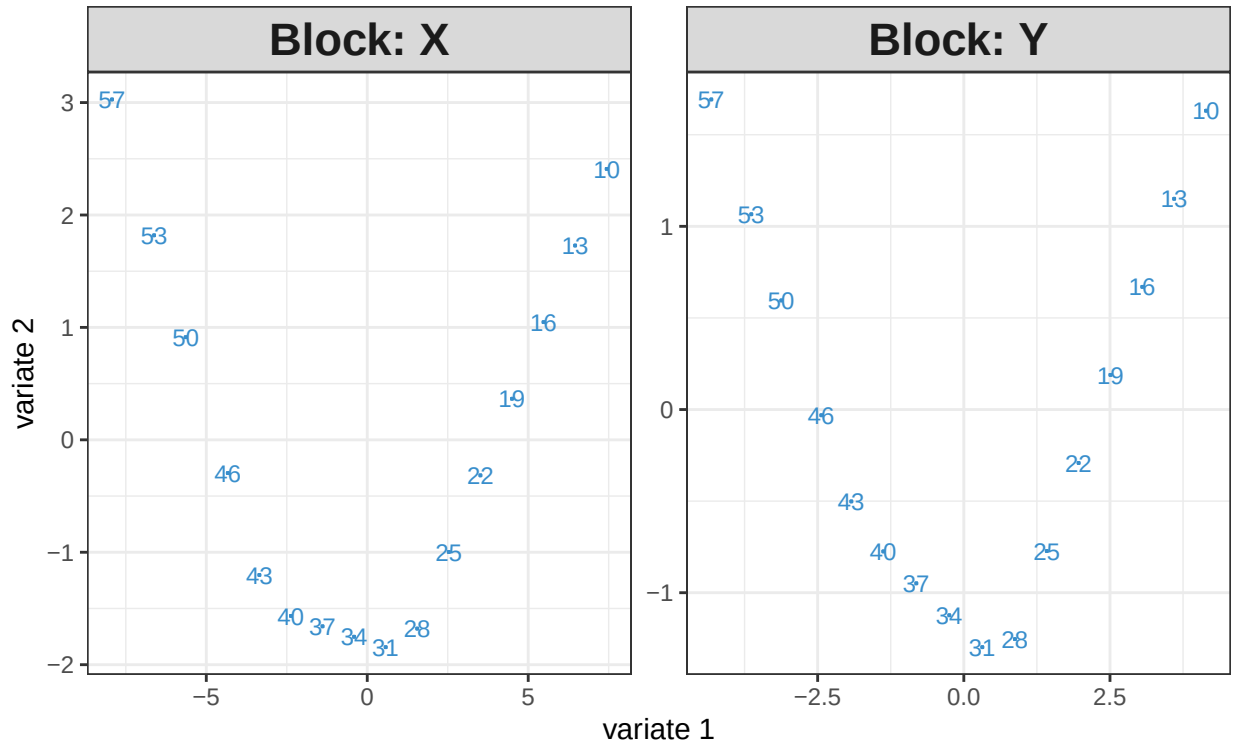


The filtering steps removed 4 metabolites. In the above graph, time profiles of the selected metabolites are displayed.

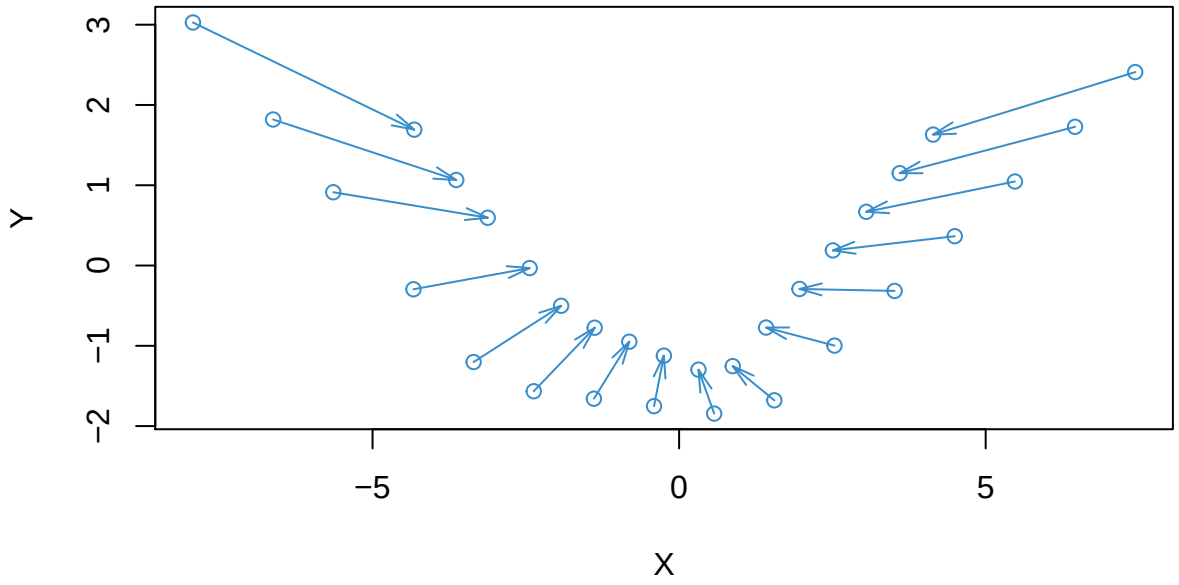
5 sPLS

In the following section, we use *Projection on Latent Structures (PLS)* to cluster both OTU and metabolite time profiles. We also use *sparse PLS* to identify a biological signature per cluster.

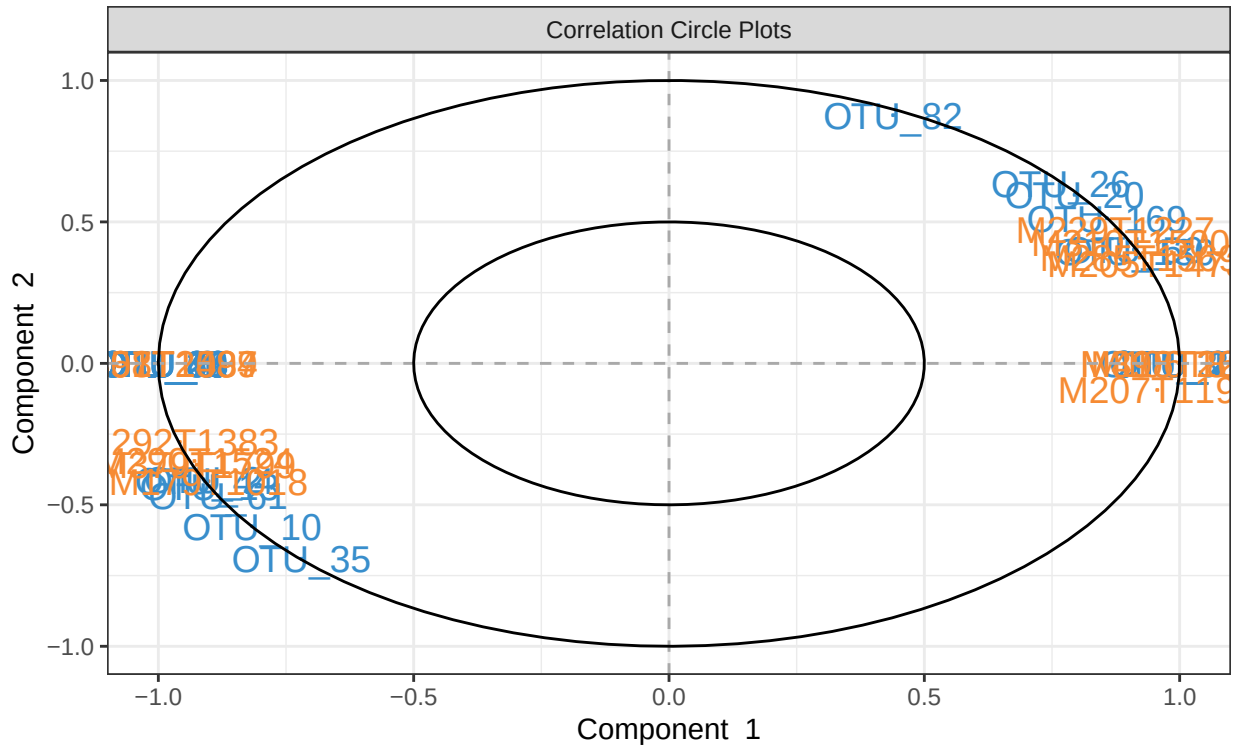
In the graphs below, time points are represented as points placed according to their projection in the smaller subspace spanned by the components of the sPLS. They allow to visualize the similarities (the points are grouped together) and the dissimilarities between the times.



We also use an *Arrow plot* to represent the similarity between the 2 datasets. Each arrow corresponds to one time. The start of the arrow indicates the location of the time in X (OTUs) in one plot, and the tip the location of the same time in Y (metabolites) in the other plot. Short arrows indicate if both data sets strongly agree and long arrows a disagreement between the two data sets.

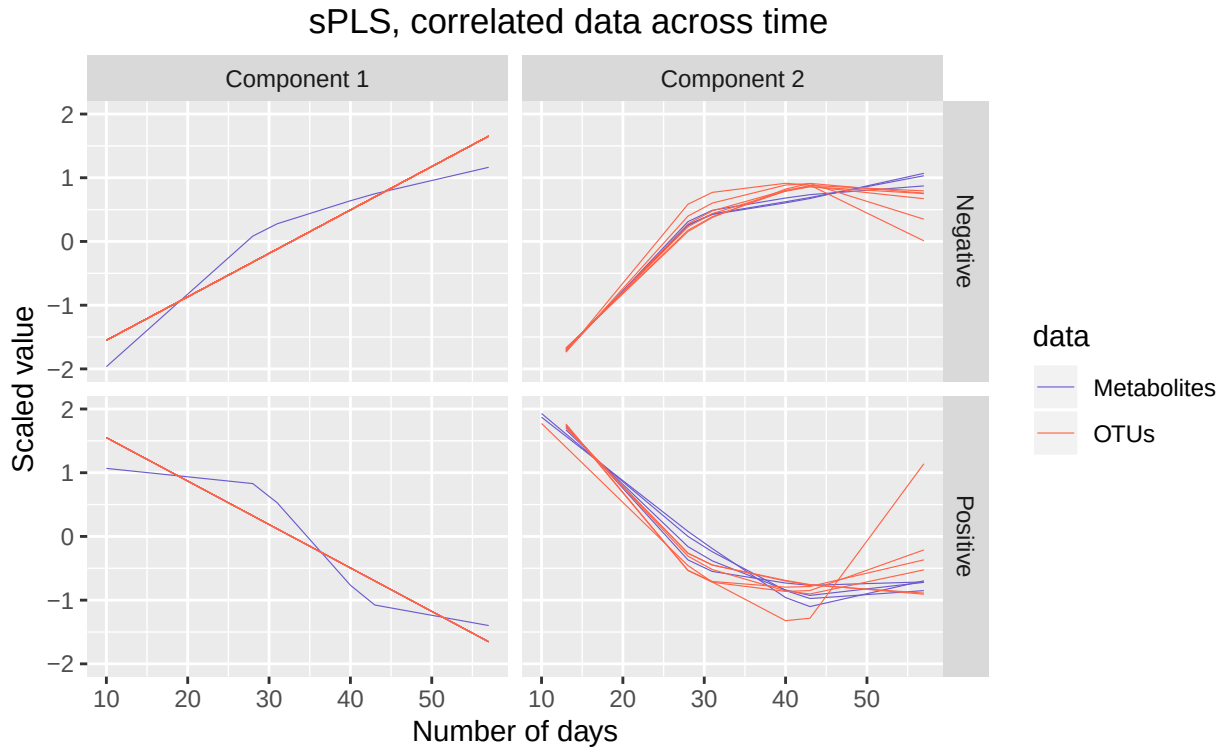


The contribution of each OTU in the construction of the new components can be displayed on the circle of correlations plot. On this graph, the strongly correlated OTUs are projected in the same direction. We use this information to build trajectory clusters.



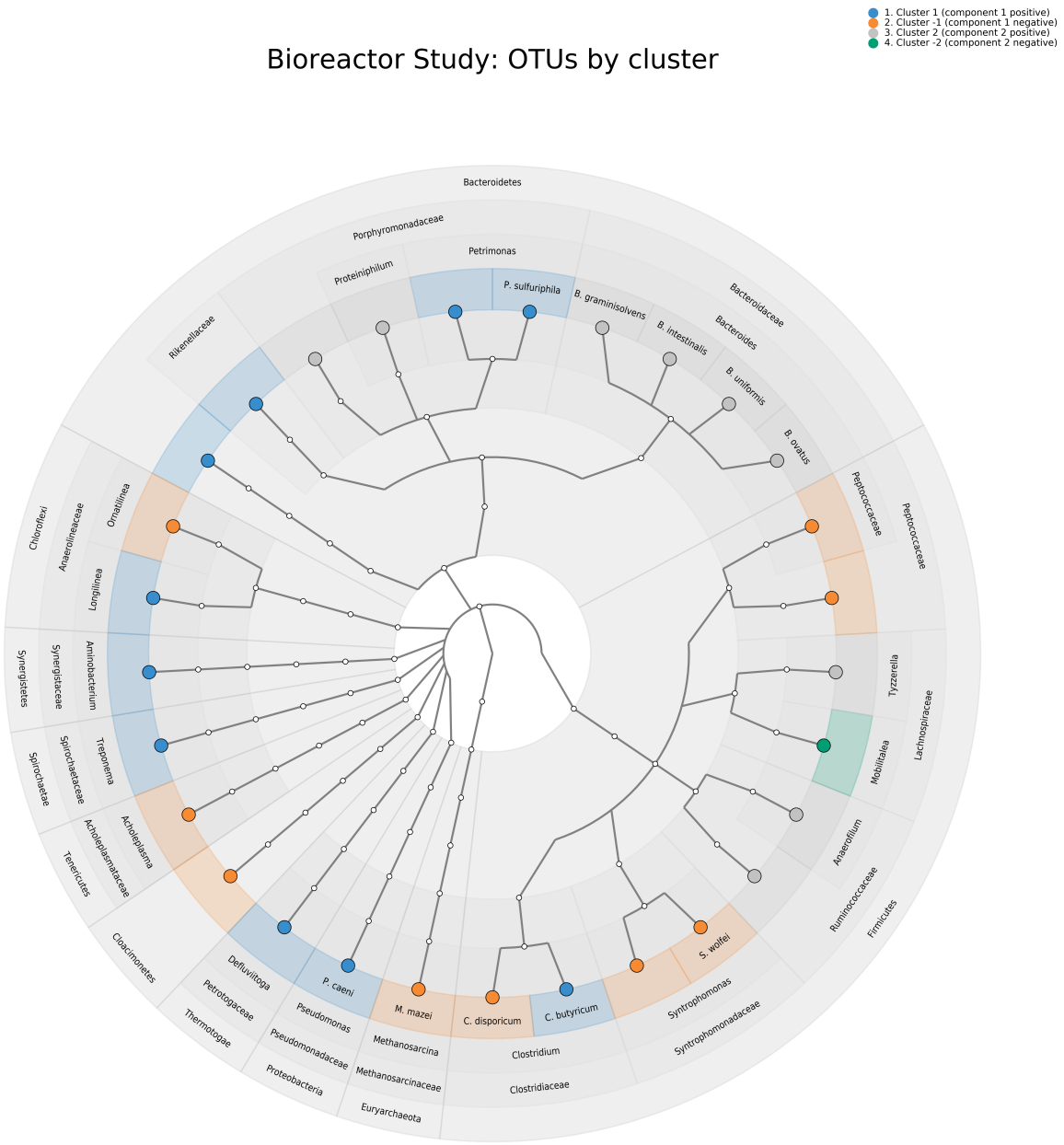
In the graph bellow, we represent the mean profiles for each block.

We have 4 clusters with sPLS. The time profiles for both OTUs and metabolites are represented by cluster. The clusters are separated according to component and contibution. On the first component (first column), OTUs and metabolites with a negative contribution on the component 1 are displayed on the first row. OTUs and metabolites with a positive contribution on the component 1 are displayed on the second row. The same applies to OTUs and metébolites selected on the second component.



The phylogenetic tree below was produced using GraPhlAn tools. To create such cladograms, GraPhlAn needs a taxonomy file describing the tree structure as well as an annotation file. The latter was generated partly through R-scripts (`./graphlan_bioreactor.R`) and was finalised by hand. The final annotation file is present here (`../Data/annotation_bioreactor.txt`) and below is the bash commands to reproduce the tree.

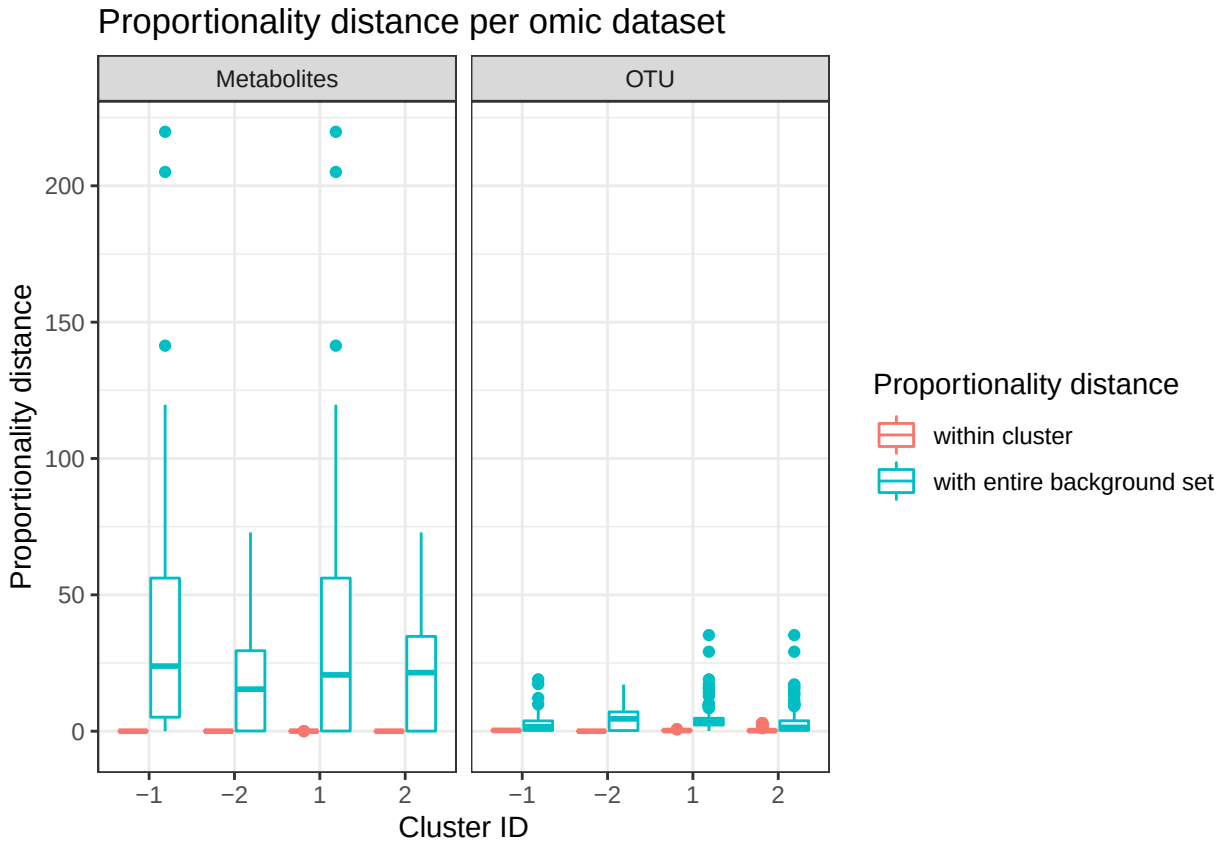
```
graphlan_annotate.py --annot ../Data/annotation_bioreactor.txt \
  ../Data/tree_bioreactor.txt tree_bioreactor.xml
graphlan.py tree_bioreactor.xml tree_bioreactor.png --dpi 600 --size 10
```



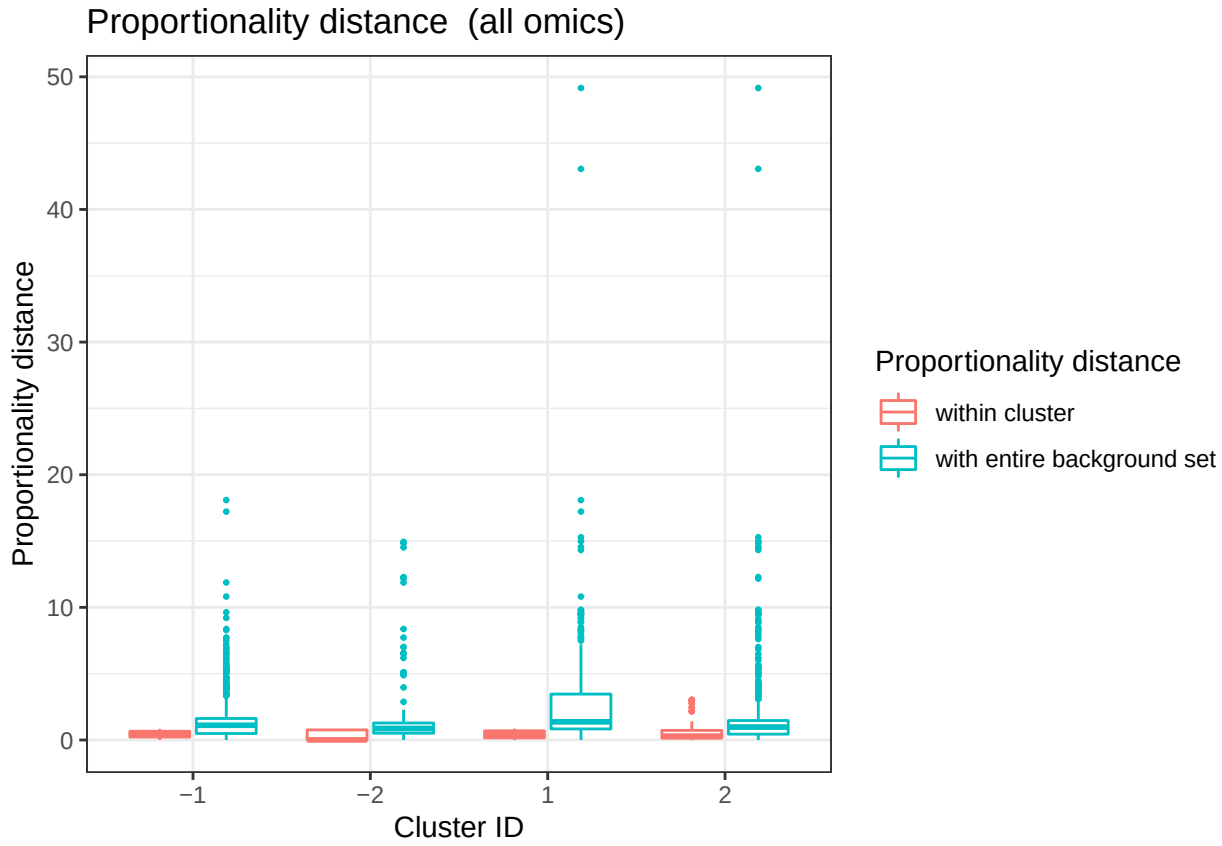
5.1 Proportionality analysis

Interpretation based on correlations between profiles must be made with caution as it is highly likely to be spurious. Proportional distances has been proposed as an alternative to measure association.

In the following graphs, we represent all the proportionality distance φ_s within clusters and the distance of features inside the clusters with entire background set. We first splited the analysis by omics data type and we computed the distance with the merged data after.



We used a Wilcoxon test to compare the median within the cluster and outside the cluster.

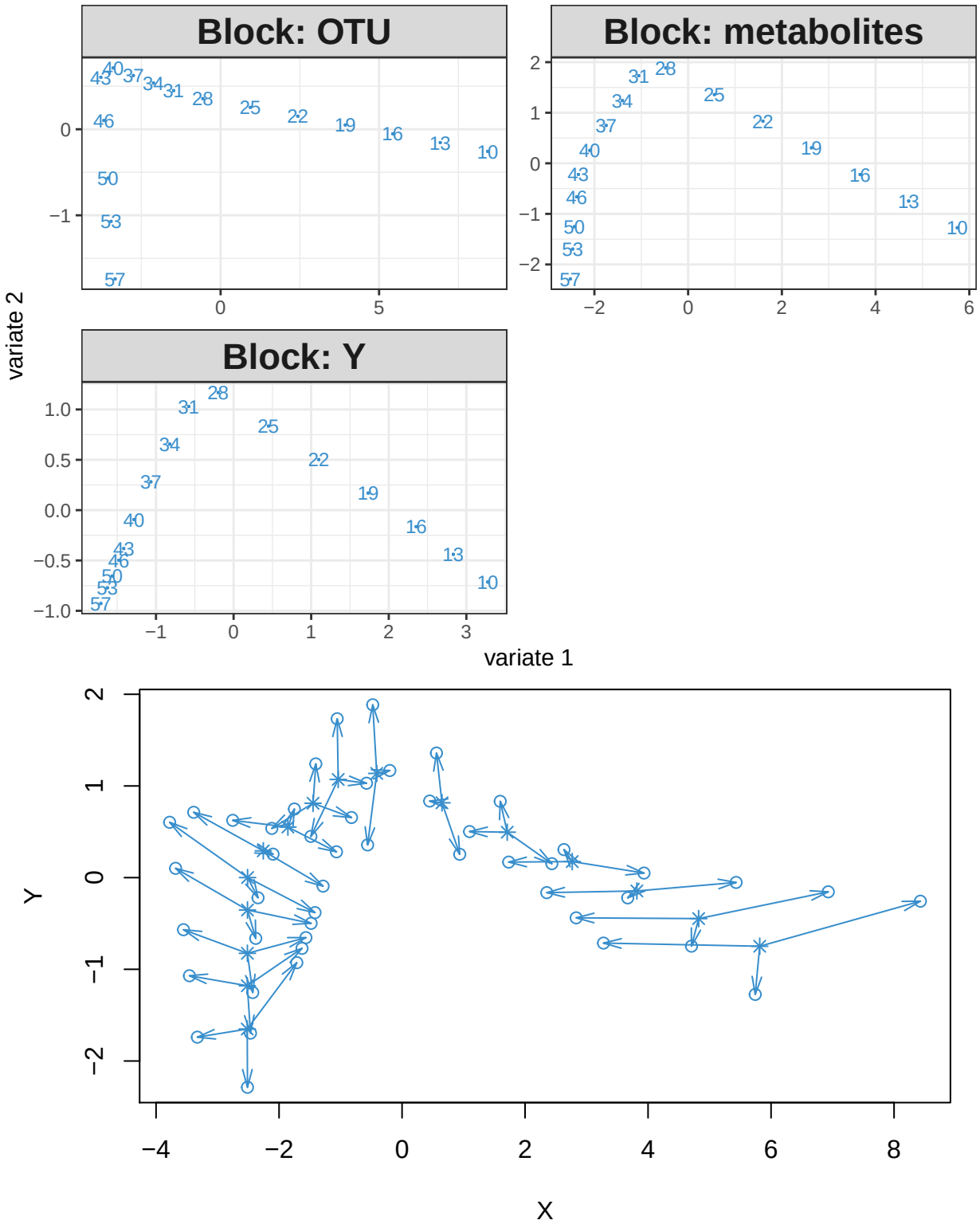


cluster	median inside	median outside	Wilcoxon test Pval
-1	0.42	1.11	1.7561650045667e-28
2	0.29	0.97	5.71081257405136e-24
1	0.43	1.37	9.39576824445502e-57
-2	0.01	0.87	2.81721824093699e-13

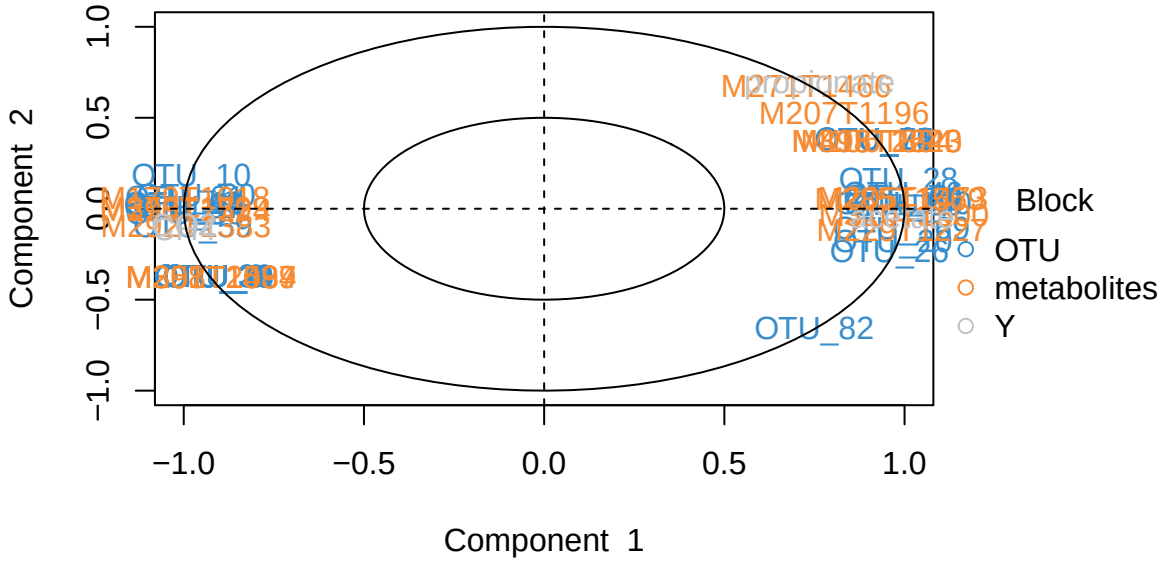
6 block sPLS

The three datasets (OTUs, metabolites and performances) are analysed together. We used a block sparse PLS analysis.

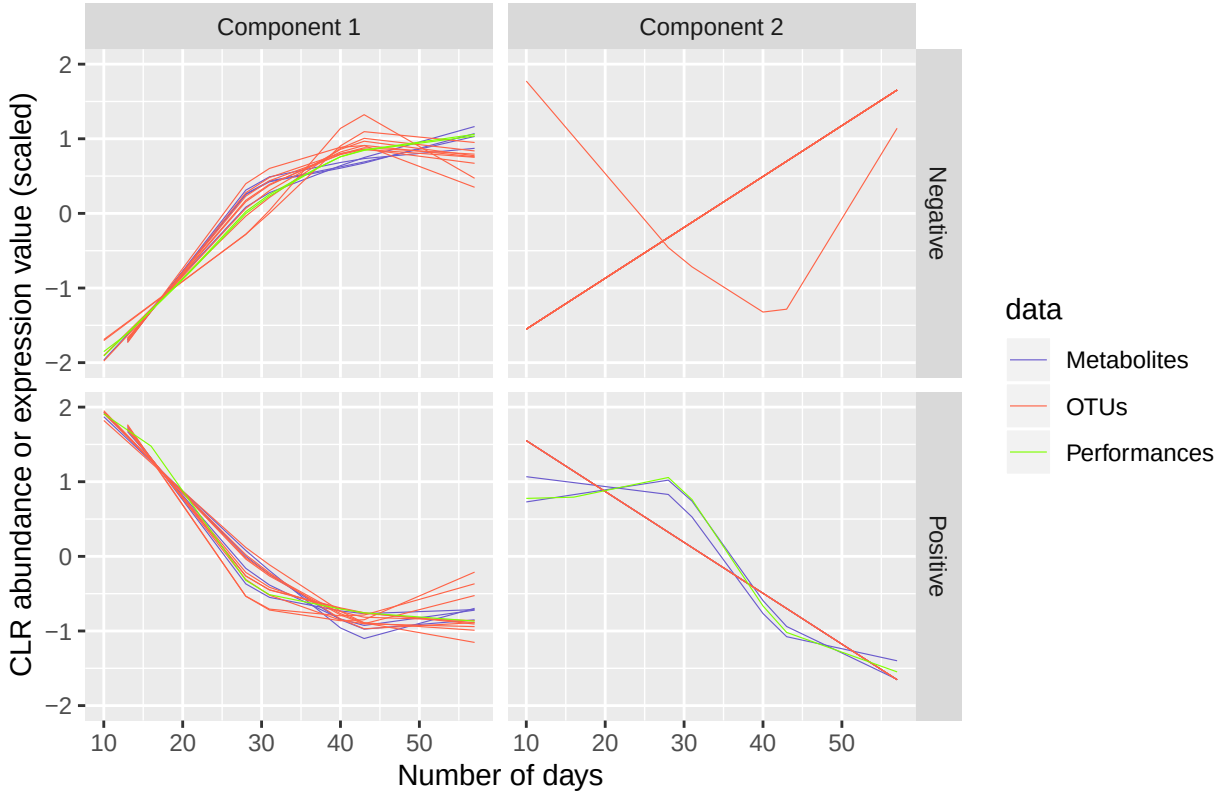
The following graphs have been previously described.

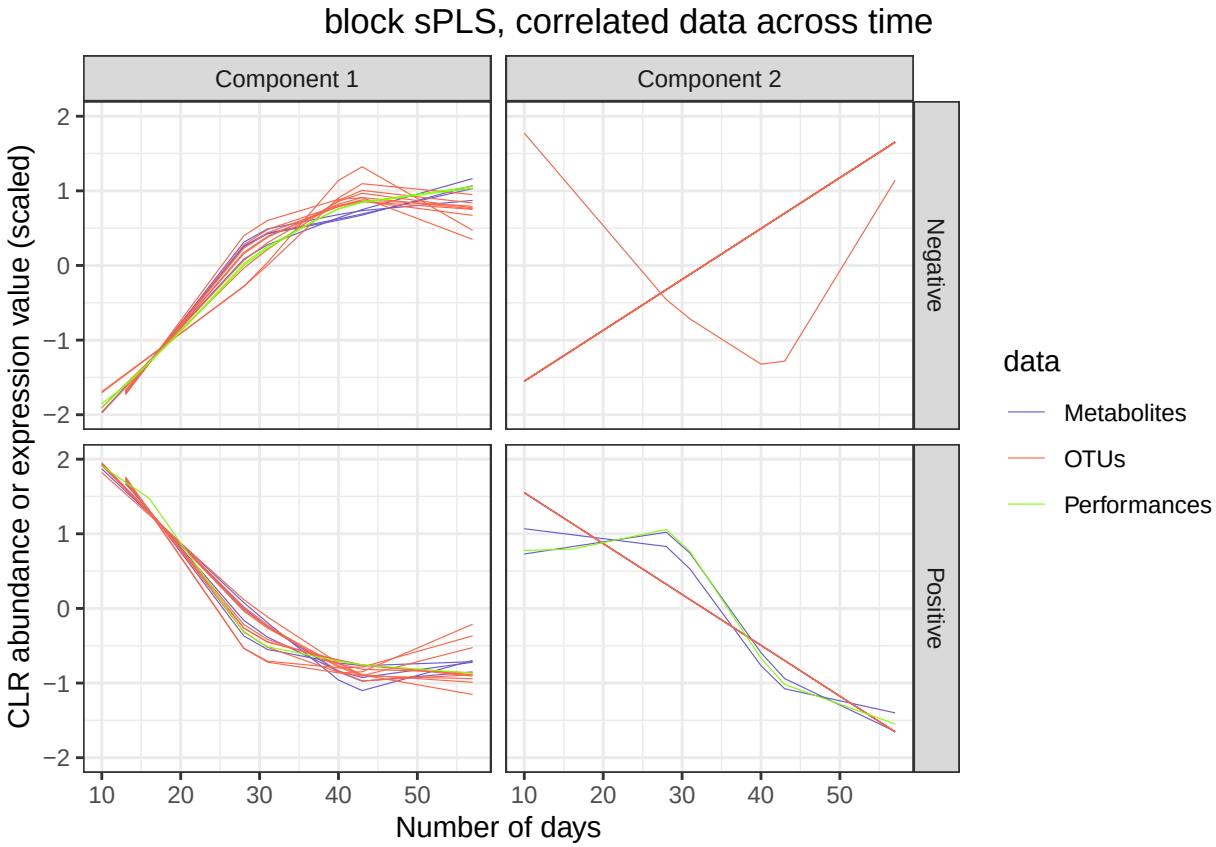

Block: OTU
Block: metabolites
Block: Y
variate 2
variate 1
Y
X

Correlation Circle Plots

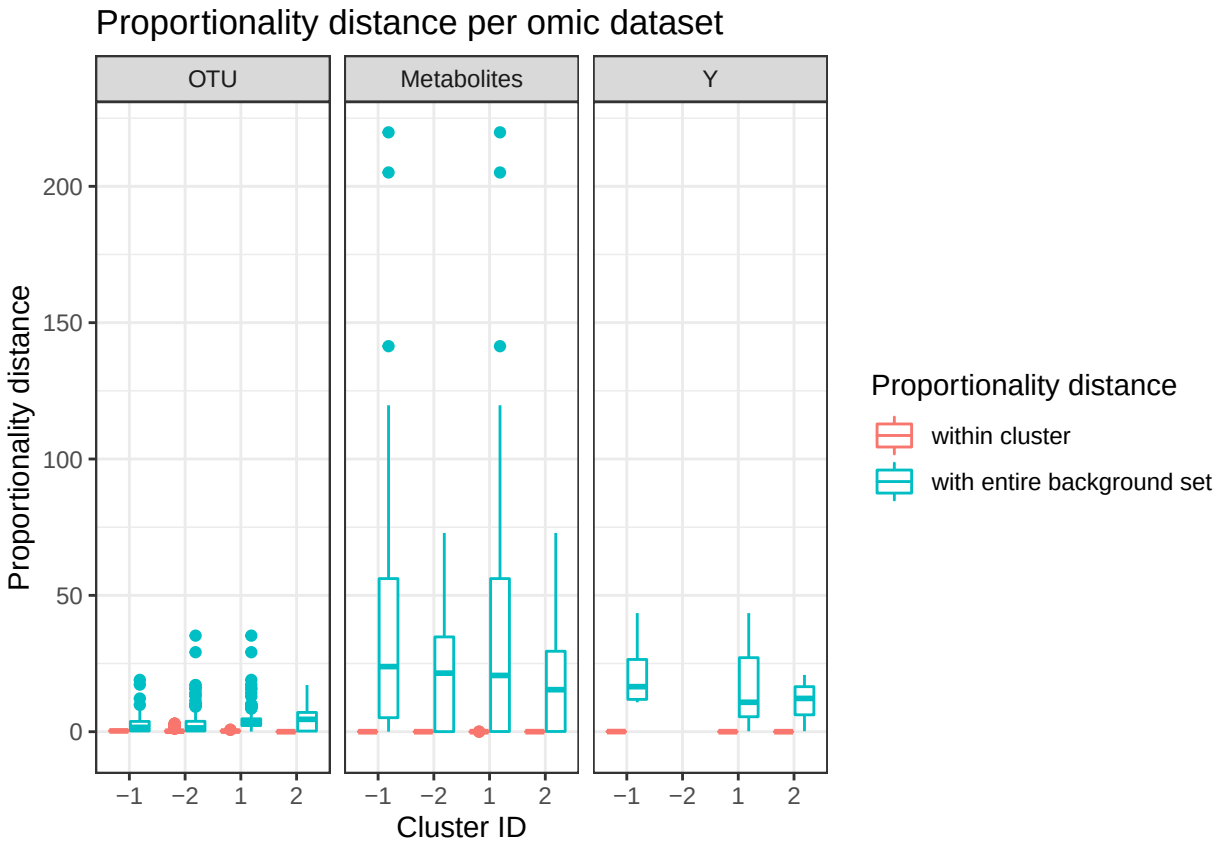


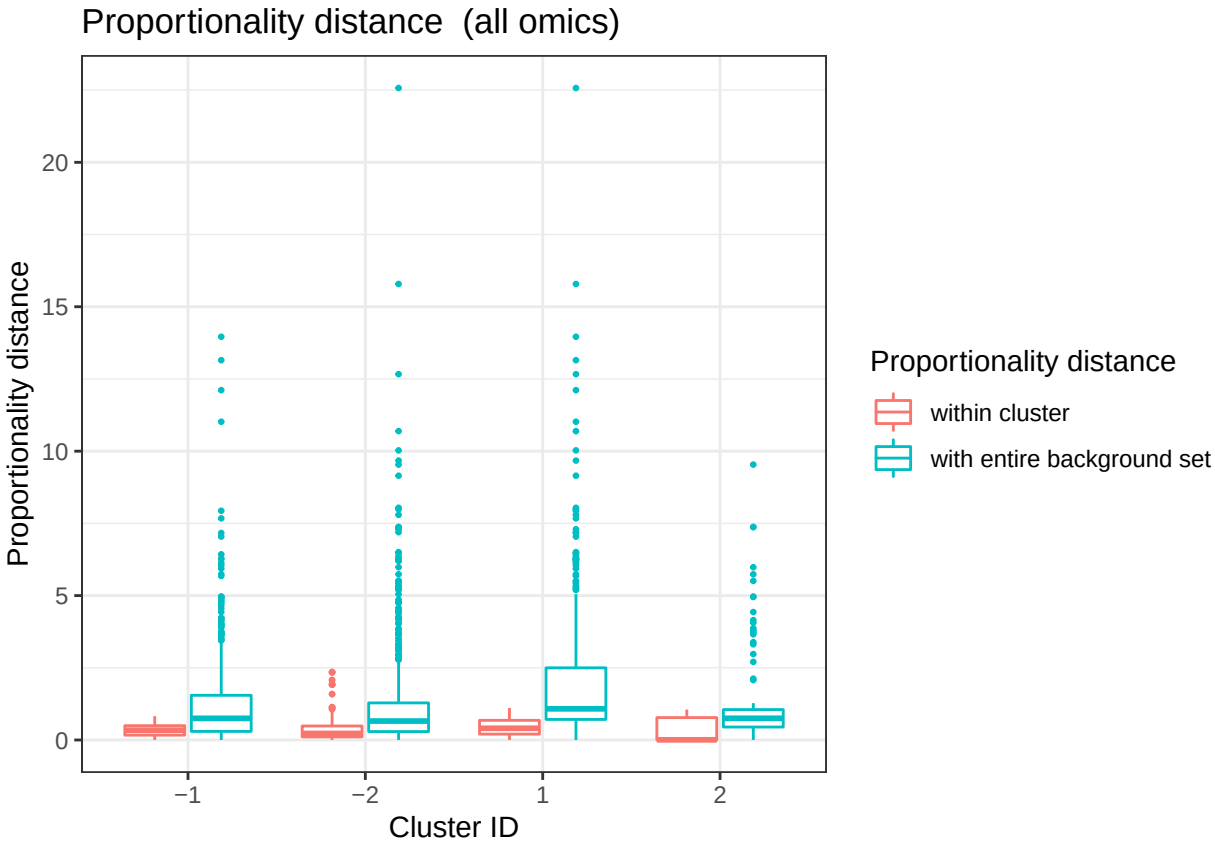
block sPLS, correlated data across time




block sPLS, correlated data across time
Component 1
Component 2
Negative
Positive
CLR abundance or expression value (scaled)
Number of days
data
Metabolites
OTUs
Performances

6.1 Proportionality analysis





cluster	median inside	median outside	Wilcoxon test Pval
-1	0.34	0.75	8.45154678269849e-29
-2	0.23	0.66	2.50553388389403e-21
1	0.41	1.08	3.01601242634991e-45
2	0.01	0.75	2.48100228794962e-06