Institut National des Sciences Appliquées de Toulouse
Mathematical and Modeling Engineering Department

October 2017 - January 2018

Fifth-year project

Identification and caracterisation of virome and bacteriome in calves suffering from infectious bronchopneumonia

Authors :

Gicu Stratan
Soizick Magon de La Giclais

Tutor :

Nathalie Villa-Vialaneix

Abstract

Infectious bronchopneumonia represents, with diarrhea, 30% to 90% of calf breeding mortality but there is almost no information in biology litterature on these infections. The purpose of this project was to establish links between bacteria found from calf noses and those found in calf lungs, and also to explain the presence of a list of viruses using the abundance of these bacteria.

Hence, the data set under analysis gives information about infectious agents: some samples were taken from 23 calf breeding farms where the calves were suffering from infectious bronchopneumonia. We implemented statistical methods in order to analyse this data set, such as principal component analysis, sparse partial least squares discriminant analysis, partial least squares in regression mode and random forest. Then, healthy calf breeding data was compared with these results.

[...]

Contents

Acknowledgment

Nathalie Villa-Vialaneix, [...]
Elias Salem, [...]

Introduction

1 Context and description of the dataset

1.1 Context

The dataset provided comes from biological studies perfomed on calf breeding farms during the acute phases of the respiratory disease complex (BRD). Bovine respiratory disease complex is considered one of the most common and economically important diseases in calves, it is displayed as pneumonia with different states of severity and respiratory signs depending on the causative agents and stressors (host immune reaction, environment. . .). Furthermore, a control group of healthy calves was sampled as well.

The dataset is divided in three files containing microbiota data; we call microbiota all the microbes that are found in a particular niche or habitat. In our case, the niche is the respiratory tract of the calves, especially in the nasal cavities (upper respiratory tract) and in the lungs (lower respiratory tract).

First part: SSamples originating from 23 breeding farms taken from calves suffering from BRD, the calves where sampled during the acute phase (less than 3 days since the onset of symptoms) and were not vaccinated nor treated with antibiotics (in order not to intervene with the present microbiota). In each breeding farm, a group of 4-5 calves were studied, from which 2 kinds of samples were taken:

- A nasal swab, taken from the nasal cavity and representing the upper respiratory tract (called "EN");
- Bronchoalveolar lavage fluid, taken from the lungs and representing the lower respiratory tract (called "LBA")..

For every pool of calves in the breeding farms and each of the sample types, we have the abundance of each of the detected bacteria i.e how many times the sequence of specific bacteria was detected.

The set contains also data indicating the presence/absence of seven major pathogens known to cause the respiratory disease; these data were generated by RT-PCR (specific and direct test).

Second part: The same kind of samples was taken from 6 breeding farms where the calves did not have any sign of respiratory disease.

In the next section will be seen the details of each one of these data.

1.2 Dataset description

As explained previously, three text files are available for the study:

- `abundances`: bacteria counts where calves are suffering from infectious bronchopneumonia;
- `pathogenes`: presence/absence of seven pathogens targets;
- `abundances_ctrl`: bacteria counts where calves are healthy.

They are all under *Comma-Separated Values* (CSV) format. In this chapter will be seen what they contain and what type of transformations have been perfomed, in order to make them suitable for further study.

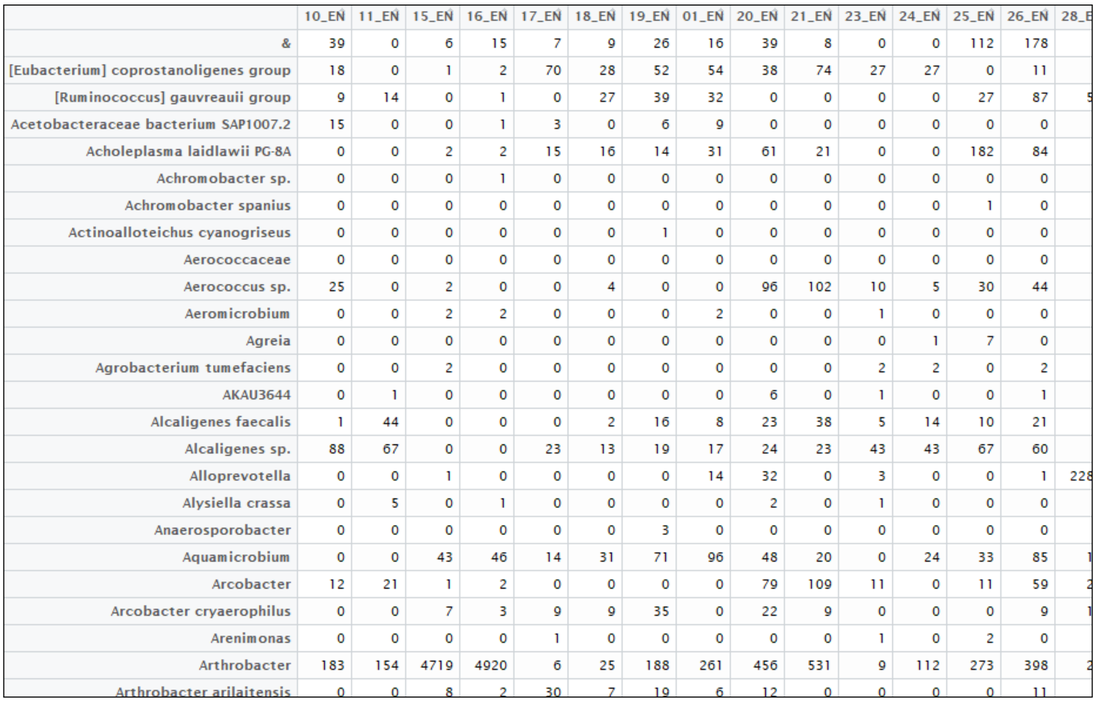
The `abundances` file

In this file, there is information (counts) about every bacteria found in every samples.

At the begining, the file contained 99 columns and 406 rows. A certain number of pretreatments have been performed in response to somes findings on the dataset:

- **First nine columns removed**. They contained the name of each bateria, its type, its family, its "blast taxonomy" and technical information that we will not use. We are interested in the name of each bacteria (an identifier) and the measurments made in each sample;
- **Columns A and B summed**. For each sample, there were two columns, identified by the letters *A* and *B*. It corresponds to the two technical replicates of all the farms involved in the study. These columns will be merged (simple sums as the counts have already been normalized to identical library sizes), which leads to 45 columns;
- **Duplicates of bateria merged**. The list of bacteria names is not unique. We could have used the blast identifier but the problem remains the same. First, let us keep only the species name. It is the last one on each row. But sometimes the species is unknown, as we can see above. In this case, the name that appears before will be kept. Then, the replicates are merged by adding together each one of them, which lead us to a list of 270 unique bacteria.
- **Name of one bacteria corrected**. One name of the bacteria was "`&`". It is now replaced by the correct one.
- **29th column removed** (only for paired studies) . The number of columns was odd (45), but on each calf was supposed to be taken two samples: one in the nose and the other in the lungs. It will be removed from the study when a paired analyse will be made.

These pretreatments lead us to a data frame with 270 rows and 45 columns. A part of this data frame can be seen in the Figure 1.1.



	10_EN	11_EN	15_EN	16_EN	17_EN	18_EN	19_EN	01_EN	20_EN	21_EN	23_EN	24_EN	25_EN	26_EN	28_E
&	39	0	6	15	7	9	26	16	39	8	0	0	112	178	
[Eubacterium] coprostanoligenes group	18	0	1	2	70	28	52	54	38	74	27	27	0	11	
[Ruminococcus] gauvreauii group	9	14	0	1	0	27	39	32	0	0	0	0	27	87	5
Acetobacteraceae bacterium SAP1007.2	15	0	0	1	3	0	6	9	0	0	0	0	0	0	
Acholeplasma laidlawii PG-8A	0	0	2	2	15	16	14	31	61	21	0	0	182	84	
Achromobacter sp.	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
Achromobacter spanius	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
Actinoalloteichus cyanogriseus	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
Aerococcaceae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Aerococcus sp.	25	0	2	0	0	4	0	0	96	102	10	5	30	44	
Aeromicrobium	0	0	2	2	0	0	0	2	0	0	1	0	0	0	
Agreia	0	0	0	0	0	0	0	0	0	0	0	1	7	0	
Agrobacterium tumefaciens	0	0	2	0	0	0	0	0	0	0	2	2	0	2	
AKAU3644	0	1	0	0	0	0	0	0	6	0	1	0	0	1	
Alcaligenes faecalis	1	44	0	0	0	2	16	8	23	38	5	14	10	21	
Alcaligenes sp.	88	67	0	0	23	13	19	17	24	23	43	43	67	60	
Alloprevotella	0	0	1	0	0	0	0	14	32	0	3	0	0	1	228
Alysiella crassa	0	5	0	1	0	0	0	0	2	0	1	0	0	0	
Anaerosporobacter	0	0	0	0	0	0	3	0	0	0	0	0	0	0	
Aquamicrobium	0	0	43	46	14	31	71	96	48	20	0	24	33	85	1
Arcobacter	12	21	1	2	0	0	0	0	79	109	11	0	11	59	2
Arcobacter cryaerophilus	0	0	7	3	9	9	35	0	22	9	0	0	0	9	1
Arenimonas	0	0	0	0	1	0	0	0	0	0	1	0	2	0	
Arthrobacter	183	154	4719	4920	6	25	188	261	456	531	9	112	273	398	2
Arthrobacter arilaitensis	0	0	8	2	30	7	19	6	12	0	0	0	0	11	

Figure 1.1: *Abundances data*

The pathogenes file

The second file is composed of 9 columns and 46 rows. This file contains the presence or not of seven viruses for each farm in the two condition. Some pretreatments have been performed:

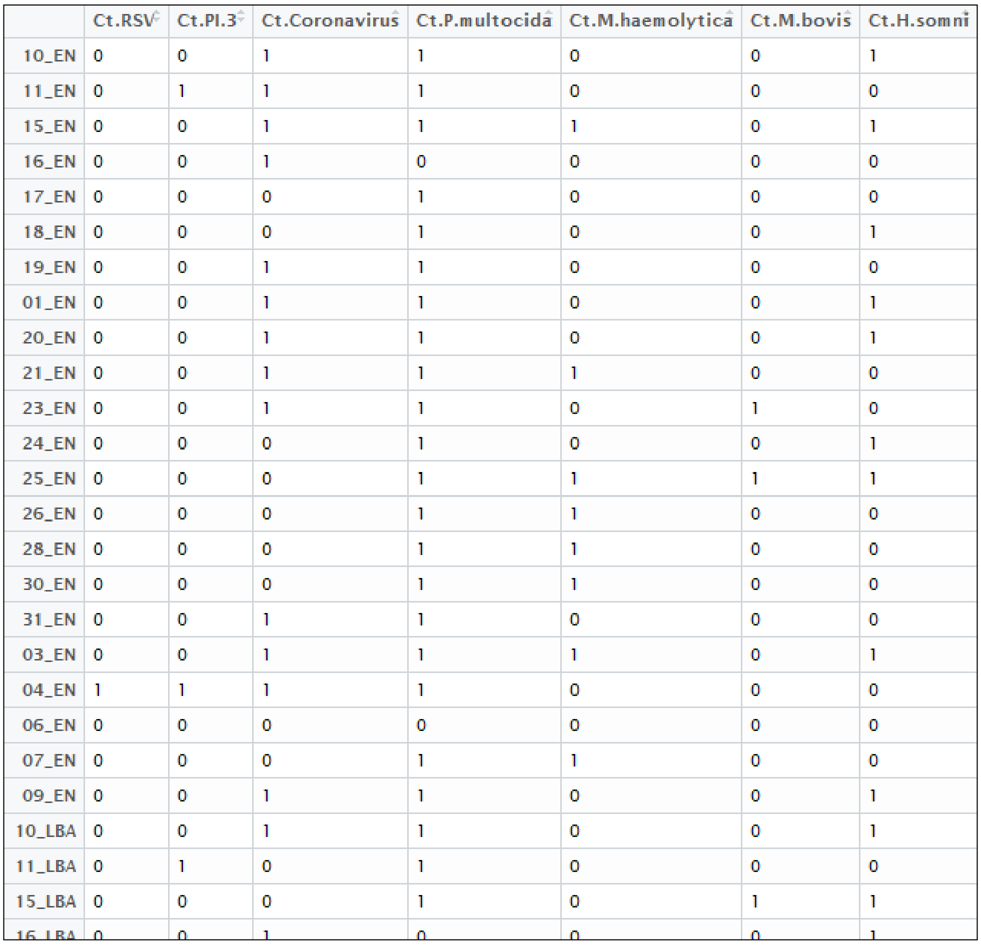
- **Farm identifier and condition merged**. The first and the last columns stand for the identifier of each farm and the sampling condition. A unique identifier is needed, hence these two columns are merged.
- **Binary code for presence/absence of virus**. The presence of a virus was coded with a letter, and the absence with 0. In order to simplify the analysis, it will by replaced as 1 and 0 respectively.

Unlike the previous file, there is no missing calf: there is 23 paired individuals. A part of the data set is displayed in the Figure 1.2.

The viruses observed are the following:

- "Ct.RSV" for respiratory syncytial virus
- "Ct.PI.3" for parainfluenza virus
- "Ct.Coronavirus" for coronavirus
- "Ct.P.multocida" for pasteurella multocida virus
- "Ct.M.haemolytica" for mannheimia haemolytica virus

- "Ct.M.bovis" for mycobacterium bovis virus
- "Ct.H.somni" for histophilus somni virus



	Ct.RSV	Ct.PI.3	Ct.Coronavirus	Ct.P.multocida	Ct.M.haemolytica	Ct.M.bovis	Ct.H.somni
10_EN	0	0	1	1	0	0	1
11_EN	0	1	1	1	0	0	0
15_EN	0	0	1	1	1	0	1
16_EN	0	0	1	0	0	0	0
17_EN	0	0	0	1	0	0	0
18_EN	0	0	0	1	0	0	1
19_EN	0	0	1	1	0	0	0
01_EN	0	0	1	1	0	0	1
20_EN	0	0	1	1	0	0	1
21_EN	0	0	1	1	1	0	0
23_EN	0	0	1	1	0	1	0
24_EN	0	0	0	1	0	0	1
25_EN	0	0	0	1	1	1	1
26_EN	0	0	0	1	1	0	0
28_EN	0	0	0	1	1	0	0
30_EN	0	0	0	1	1	0	0
31_EN	0	0	1	1	0	0	0
03_EN	0	0	1	1	1	0	1
04_EN	1	1	1	1	0	0	0
06_EN	0	0	0	0	0	0	0
07_EN	0	0	0	1	1	0	0
09_EN	0	0	1	1	0	0	1
10_LBA	0	0	1	1	0	0	1
11_LBA	0	1	0	1	0	0	0
15_LBA	0	0	0	1	0	1	1
16_LBA	0	0	1	0	0	0	1

Figure 1.2: *Pathogenes data*

In addition to the first file, this one will allow us to study links between virus and bacteria.

The `abundances_ctrl` file

This file has the same design as the `abundances` file; it will be treaded the same way. The first difference is that this file corresponds to the samples taken from healthy calves in 6 breeding farms, which gives us 12 samples.

The other difference concern the list of bacteria. The bacteria found in healthy individual are not the same as sick individuals. This list is shorter (293 rows) and the bacteria are not the same.

This list is still not composed by unique names of bacteria. It will be fixed by taking the last name of each row as explained before.

Now that this list is composed by unique bacteria names, let us see how much they are different from the `abundances` file in the Figure 1.3.

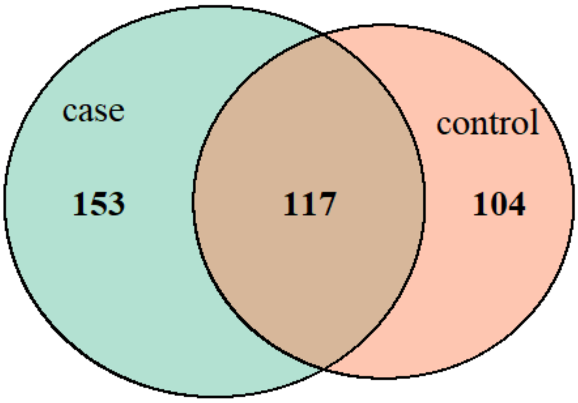


Figure 1.3: *Venn diagram comparing two lists of bacteria: for sick calves and for healthy calves*

2 Methods

2.1 Normalisation

The purpose of the normalisation is to make variable distributions more symmetrical. Furthermore, some statistical methods are sensitive to non-normalised data: most of the classification methods compute the distance between two points using euclidien distance. If one of the variable has high values comparing to the others, it will have a greather influence on the final distance, and consequently will lead to a higher variance on prediction quality.

Let us have a look at the distribution of the first simple, with a log scale for the frequencies.

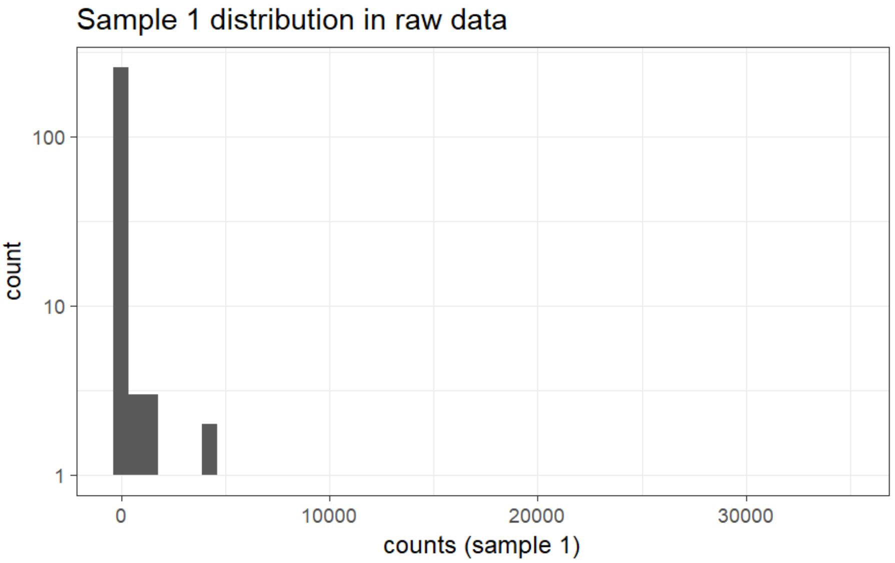


Figure 2.1: *Sample 1 distribution in raw data*

What can be noticed in the Figure 2.1 is that there is a lot of very low values, hence the distribution is highly skewed. It was expected because the observations are counts. In order to analyse this dataset, we will try several transformations known for microbiote datasets [4].

Let $y_{i,j}$ be the counts of a bacteria i found in the sample j, where $i \in [1, .., n]$ and $j \in [1, .., p]$, n and p standing for the number of bacteria and the number of samples respectively. Let $\hat{y}$ be the data after transformation.

Log transformation

$$\tilde{y}_{ij} = \log(y_{ij} + 1)$$

TSS transformation

It is commun in metagenomic datasets to perform TSS (*Total Sum Scaling*) before further normalization. TSS transformation computes relative abundances:

$$y_{ij} = \frac{n_{ij}}{\sum_{k=1}^{p} n_{ik}}$$

for n_{ij} the counts of species j in sample i, p the number of species and n the number of individuals.

TSS+CLR transformation

CLR (*Centered Log Ratio*) transformation:

$$\tilde{y}_{ij} = \log \frac{y_{ij}}{\sqrt[p]{\prod_{k=1}^{p} y_{ik}}}.$$

TSS+ILR transformation

ILR (*Isometric Log Ratio*) transformation:

$$\tilde{\mathbf{Y}}' = \tilde{\mathbf{Y}} \times \mathbf{V}$$

for $\tilde{\mathbf{Y}}$ the matrix of CLR transformed data and a given matrix $\mathbf{V}$ with p rows and $p-1$ columns such that $\mathbf{V}\mathbf{V}^\top = \mathbb{I}_{p-1}$ and $\mathbf{V}^\top\mathbf{V} = \mathbb{I} + a\mathbf{1}$, a being any positive number and $\mathbf{1}$ a vector full of 1.

CSS transformation

The CSS (*Cumulative Sum Scaling*) normalisation is an adaptative extension for metagenomic data of the quantile normalisation used in microarray expression datasets. It is designed so as to account for technical differences between samples.

2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a multidimensional descriptive method that allows to explore the links between variables and the similarities between individuals.

The objective of the PCA is to identify the largest sources of variation and return to a reduced size space (for example 2 or 3) by maximizing the variability of the projection. In other words, PCA reduces the number of variables, in this way the information becomes less redundant.

From a mathematical point of view, the PCA corresponds to the approximation of a matrix (n, p) by a matrix of the same dimensions but of rank $q < p$. In fact, an orthogonal linear transformation is applied on the data to convert a set of observations of possibly correlated variables into uncorrelated principal components. q is often of small value 2, 3 and contributes to the construction of easily understandable graphs. The interpretation of these graphs helps to understand the structure of the analyzed data.

2.3 Partial least squares regression

Partial least squares regression (PLS) [1] is a fast, efficient and optimal statistical method that is widely used to deal with situations with high multicollinearity and when big data are to be taken into account. Its use is recommended in the case where the number of variables p is much greater than the number of individuals n: $p \ll n$, which is relevant for our dataset.

There are different versions of PLS regression depending on the objective:

- **PLS1**: A quantitative target variable Y is to be explained, model, predict by p quantitative explanatory variables X^j
- **PLS2**: Canonical version. Relates a set of q quantitative variables Y^k and a set of p quantitative variables X^j.
- **PLS2**: Regression version. Tries to explain, model a set of q variables Y^k by a set of p quantitative explanatory variables X^j.
- **PLS-DA**: Discriminant version. Special case of the previous case. The qualitative variable Y with q classes is replaced by q dummy variables of these classes.

During this project, only the PLS-regression mode and PLS-DA versions were used because they are adapted to our dataset and the objective pursued.

PLS-regression algorithm

```
$X$ is the matrix of the explanatory centered variables
$Y$ is the centered variable to explain, uni or multidimensional
Initialization of the latent variable $\omega_1$ by the first column of $Y$
Set $r$ the number of iterations
For $h = 1$ to $r$ do:
   While Convergence Not Achieved do:
      $u_h = X^{'}\omega_h/\omega_h^{'}\omega_h$
      $u_h = u_h/u_h^{'}u_h$ where $u_h$ is the loading vector associated with X
      $\xi_h = Xu_h$ where $\xi_h$ is the latent variable associated with X
      $v_h = Y^{'}\xi_h/(\xi_h^{'}\xi_h)$
      $v_h = v_h/v_h^{'}v_h$ where $v_h$ is the loading vector associated with Y
      $\omega_h = Y^{'}v_h$ where $\omega_h$ is the latent variable associated with Y
   end while
   $c_h = X^{'}\xi_h/\xi_h^{'}\xi_h$ partial regression of X on $\xi$
   $d_h = Y^{'}\omega_h/\omega_h^{'}\omega_h$ partial regression of Y on $\omega$
   Deflation $X_h = X_{h-1} - \xi_h c_h^{'}$
   Deflation $Y_h = Y_{h-1} - \xi_h v_h^{'}$
end for
```

Convergence is reached when the following vectors verify:

$$YY^{'}XX^{'}u = \lambda u$$

$$Y^{'}XX^{'}Y\omega = \lambda\omega$$

$$XX^{'}YY^{'}v = \lambda v$$

$$X^{'}YY^{'}X\xi = \lambda\xi$$

where u, ω, v, ξ are the respective eigenvectors of the matrices $YY^{'}XX^{'}$, $Y^{'}XX^{'}Y$, $XX^{'}YY^{'}$, $X^{'}YY^{'}X$ associated with the same greater eigenvalue λ.

PLS-Discriminant Analysis algorithm

The only difference compared to the PLS-regression algorithm consists in the variable to explain Y which is this time qualitative with m modalities. It is enough to transform the variable Y in m dummy variables and to apply on this the algorithm above.

2.4 Random Forest

Random Forest is a machine learning method for classification and regression generally. This algorithm is based on the aggregation of decision, regression or classification binary trees, depending on the type of the target variable Y.

It consists of drawing *n_estimators* bootstrap samples. A bootstrap sample is a sample constructed by n random draws with replacement among the n initial observations. The distribution of this sample is the empirical one which gives a $\frac{1}{n}$ weight to each successful draw. A tree is estimated for each bootstrap sample and the prediction is obtained by averaging (regression) or vote (classification) the individual predictions of each tree. During the construction of each tree's node in the learning process, it is selected a random subset of the features that allows to build less correlated trees and make aggregation more efficient [1]. It is generally an effective method, which could give satisfying predictions for presence/absence of viruses.

Let Y be the predictable variable and $z = (x_1, y_1), ..., (x_n, y_n)$ a sample, and B the number of replicates of bootstrap samples each obtained by n draws with replacement.

Random Forest algorithm

```
For $b = 1$ to $B$ do:
    Draw a bootstrap sample ${z_b}^*$
    Estimate a tree on this sample with a random subset of features
Calculate the average estimate $\widehat{f}_B(Y) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}_{z_b}(Y)$ or the result of
the vote.
```

3 Results

3.1 Exploratory analysis

3.1.1 Normalisation

Previously, we saw that the first sample distribution was highly asymmetrical due to the fact that it is from counts data.

If we look at the graphs in the Figure 3.1, the less asymetric distribution seems to be the one obtained with the CSS transformation and the log-transformation, but there is still a lot of very low values. In the following will be used the log-transformation.

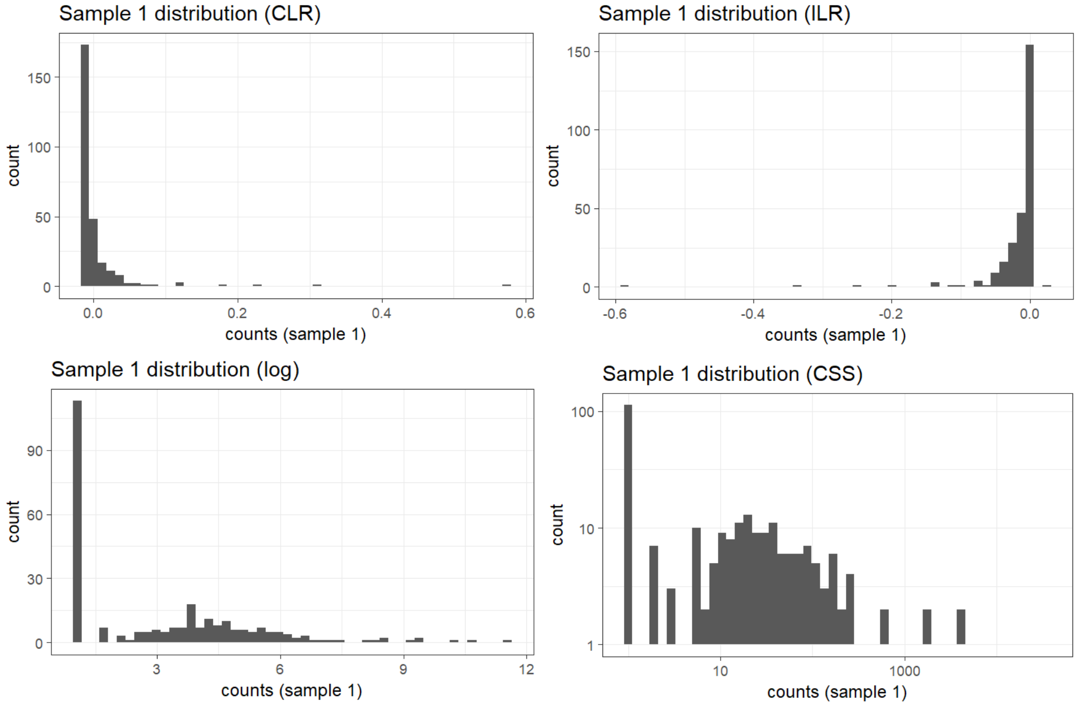


Figure 3.1: *Sample 1 distribution according to the type of transformation*

Let us have a look at the distribution by boxplots of all the samples displayed in the Figure 3.2. It confirms that the best transformations are CSS and logarithmic. It seems that the CSS normalisation treats outliers more efficiently than the log one.

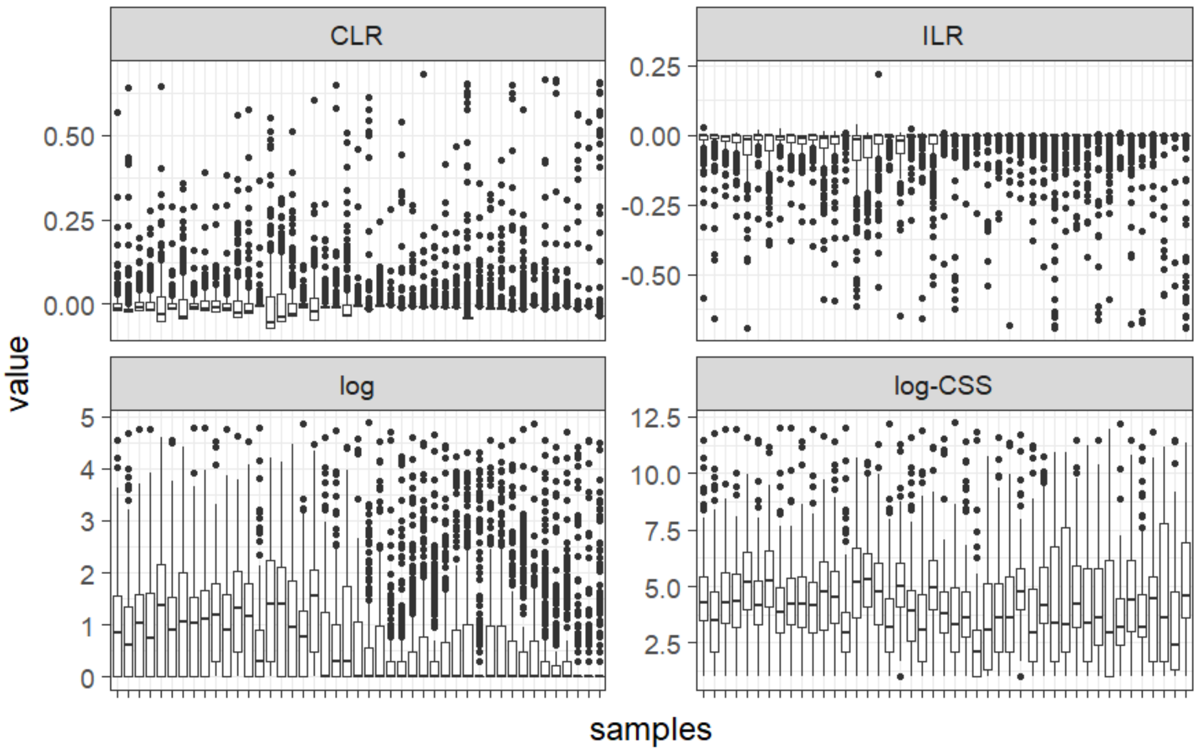


Figure 3.2: *Distributions of all samples according to the type of transformation and the sample*

3.1.2 Principal Component Analysis

It is important to have a clear idea of the data structure. A principal component analysis is adapted to this objective. In the following, we will analyze the results of the PCA on the Log and CSS transformations made on the observation matrix because the two others show very similar results. It will be perfomed in `R`, using the `mixOmics` library [2].

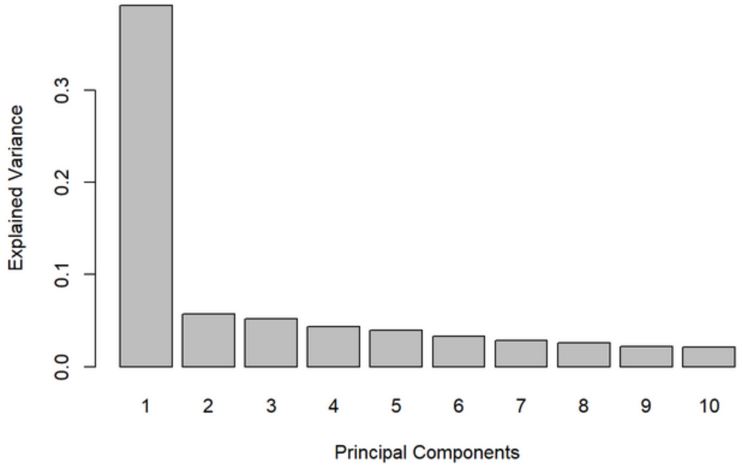


Figure 3.3: *Sample 1: The variance explained by the first 10 components*

The Figure 3.3 represents the decay of the eigenvalues associated to the principal components. It seems that the first component on 406 is sufficient to reconstitute the data. This is explained by the rapid decay of the firsts ten eigenvalues, which are not important beyond that. The first axis, showing the projection of the data on the first principal component seems to illustrate its stability in the construction of the representation subspace.

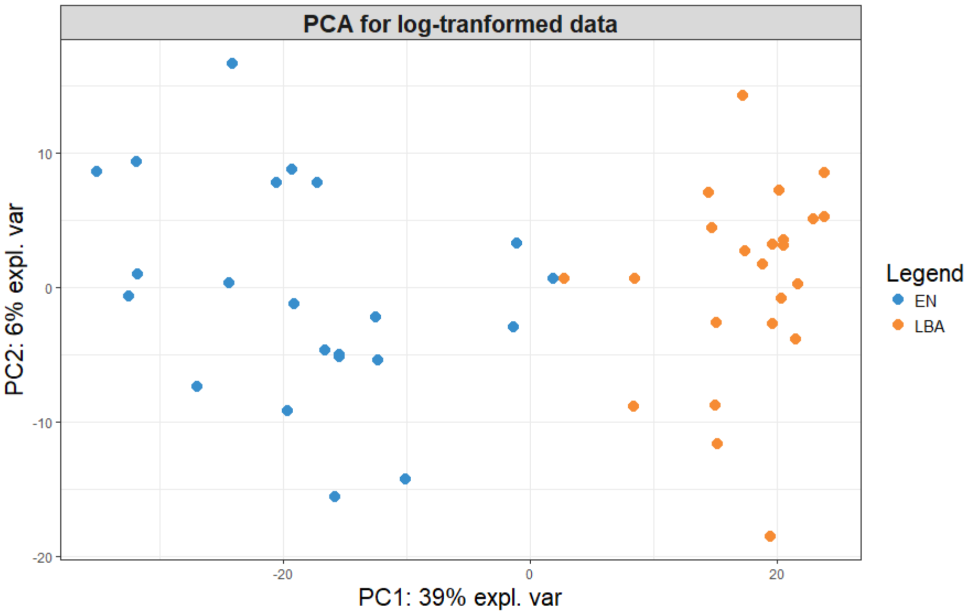


Figure 3.4: *Sample 1: Projection of individuals in the first factorial plane*

The projection of individuals in the first factorial plan displayed in the Figure 3.4 shows that the separation of "EN" and "LBA" classes is marked, and can even be divided linearly. Also, it can be seen that the first axis associated to the first principal component explains 39% of variance and with the second, they contribute mainly to the reconstitution of the data.

3.2 Differences between EN and LBA samples

The purpose of this part is to select bacteria that have an impact on the EN/LBA samples. The methods used are PLS-DA, and more specifically sparse PLS-DA. They will be performed using paired or unpaired analysis, which means indicating the sparse PLS-DA that our samples are paired or not. The library used to implement this method remains `mixOmics` [2].

3.2.1 PLS-DA - paired samples

In this section we will focus on a multilevel analysis, therefore the 29$^{\text{th}}$ column will be removed (section 1.2 page 7) for the multilevel option (in the 4$^{\text{th}}$ step below).

- **Step 1**: Performance of the PLS-DA on the first two components. A first PLS-DA is computed (with 10-fold cross-validation repeated 10 times) to check the efficiency of the method and which type of distance to use in its computation.

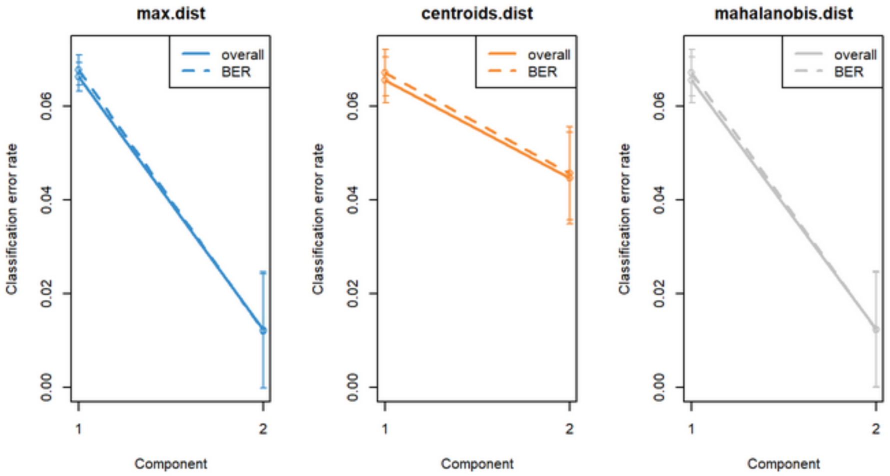


Figure 3.5: *Distances for the first two components*

The Figure 3.5 indicates that the classification error rate decrease from one component to two components in the model, for three different distances: Maximum, Centroids and Mahalanobis. BER stands for Balanced Error Rate, and here the groups are unbalanced because of the 29th column, which explains the small difference between the two lines.

- **Step 2**: Projection of individuals (left graph in Figure 3.7). The sample plot shows the PLS-DA first two components. The ellipse are 0.95 confidence interval ellipse for each condition type.

- **Step 3**: Tuning sparse PLS-DA.

  In the Figure 3.6, the error rate decreases when 2 components are included in sPLS-DA. The diamonds stand for the optimal number of variables to select on components 1 & 2. In this case, 5 variables will be selected for the first component and 15 for the second.

- **Step 4**: Running sPLS-DA tuned (right graph in Figure 3.7). This step consist of running the sPLS-DA with new parameters such as the number of selected variables and the multilevel option. The multilevel option will indicate that our individuals are paired, *i.e.* that two samples have been taken from the same farm.

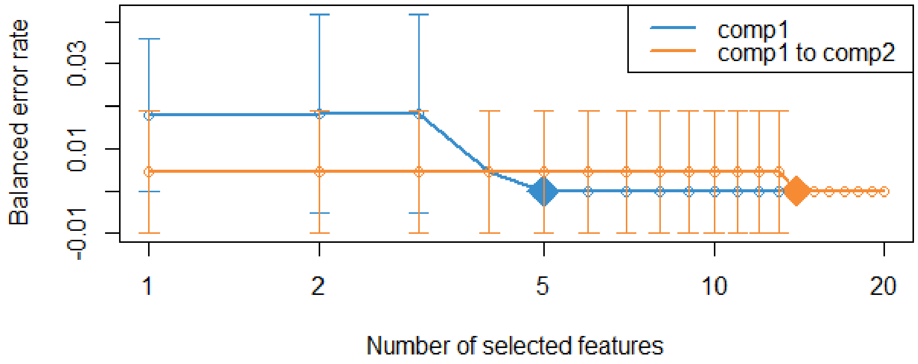


Figure 3.6: *Error rate for the first two components by the number of selected variables*

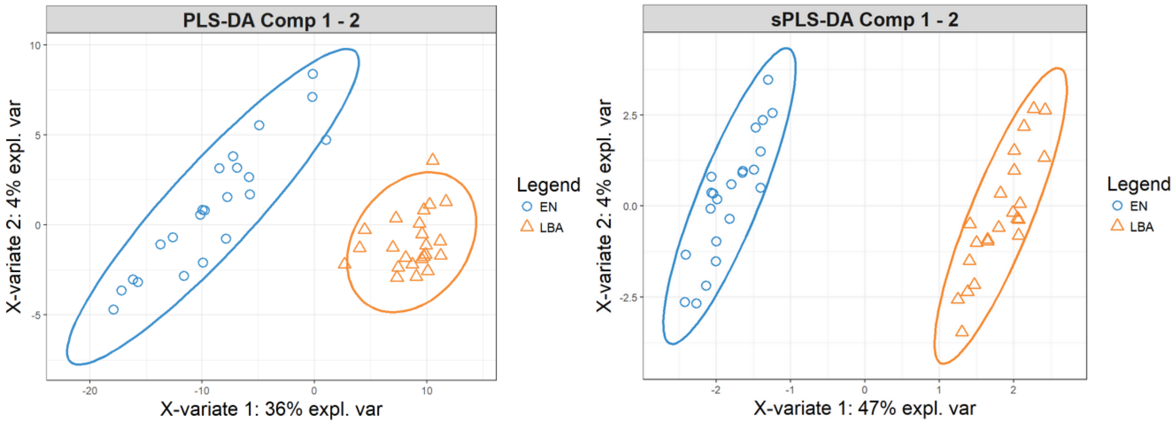


Figure 3.7: *Projection of individuals for PLS-DA and sPLS-DA first two components*

3.2.2 PLS-DA - unpaired samples

3.3 Predicting the presence of viruses

3.3.1 In the nasal cavity (EN)

PLS in regression mode

Student tests

Common bacteria

Random Forest

3.3.2 In the lungs (LBA)

PLS in regression mode

Student tests

Common bacteria

Random Forest

Conclusion

Bibliography

[1] Philippe Besse. *Science des données : Apprentissage Statistique.* 2017. `https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-Intro-ApprentStat.pdf`.

[2] Kim-Anh Le Cao, Florian Rohart, Sebastien Dejean with key contributors Benoit Gautier Ignacio Gonzalez, Francois Bartolo, contributions from Pierre Monget, Jeff Coquery, FangZou Yao, and Benoit Liquet. *Package mixOmics: Omics Data Integration Project.* 2016. `https://CRAN.R-project.org/package=mixOmics`.

[3] Kim-Anh Lê Cao, Mary-Ellen Costello, Vanessa Anne Lakis, François Bartolo, Xin-Yi Chua, Rémi Brazeilles, and Pascale Rondeau. MixMC: A multivariate statistical framework to gain insight into microbial communities. *PLoS ONE*, August 2016.

[4] Jérôme Mariette. *Apprentissage statistique pour l'intégration de données omiques.* PhD thesis, Université Toulouse 3 Paul Sabatier, 2017.