Analysis of High Dimensional Data

Wilson Tendong, Luis CampoverdeReinoso, Omkar Kulkarni

21 April 2016

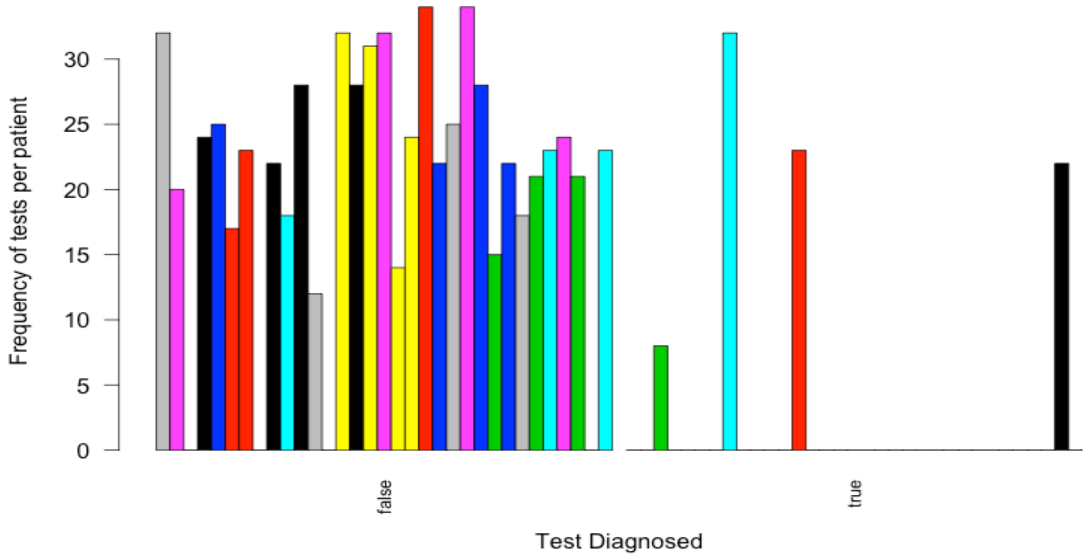
Part I

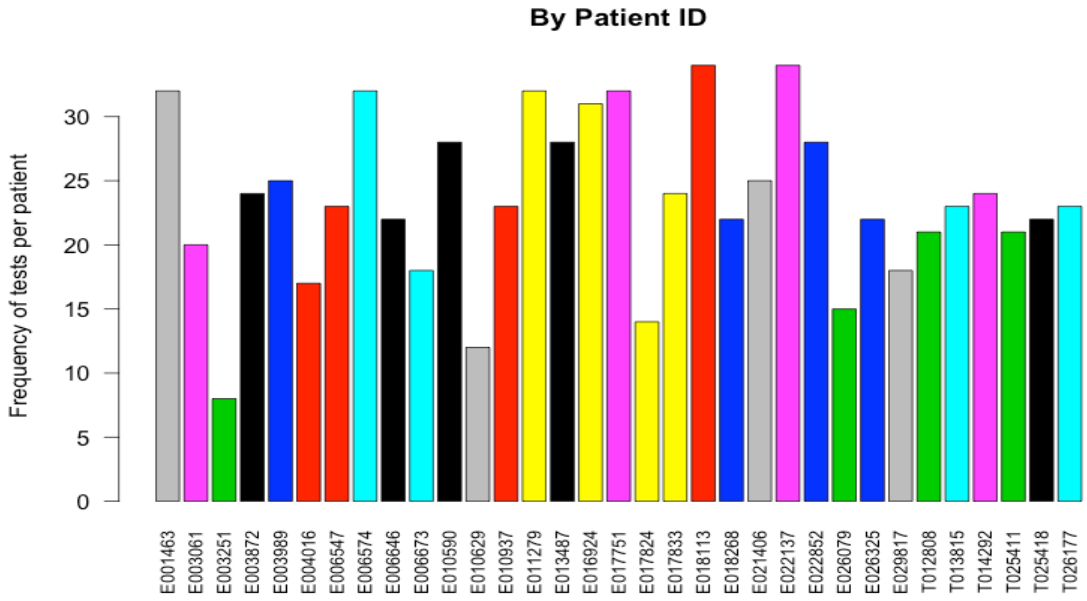
1. Data exploration

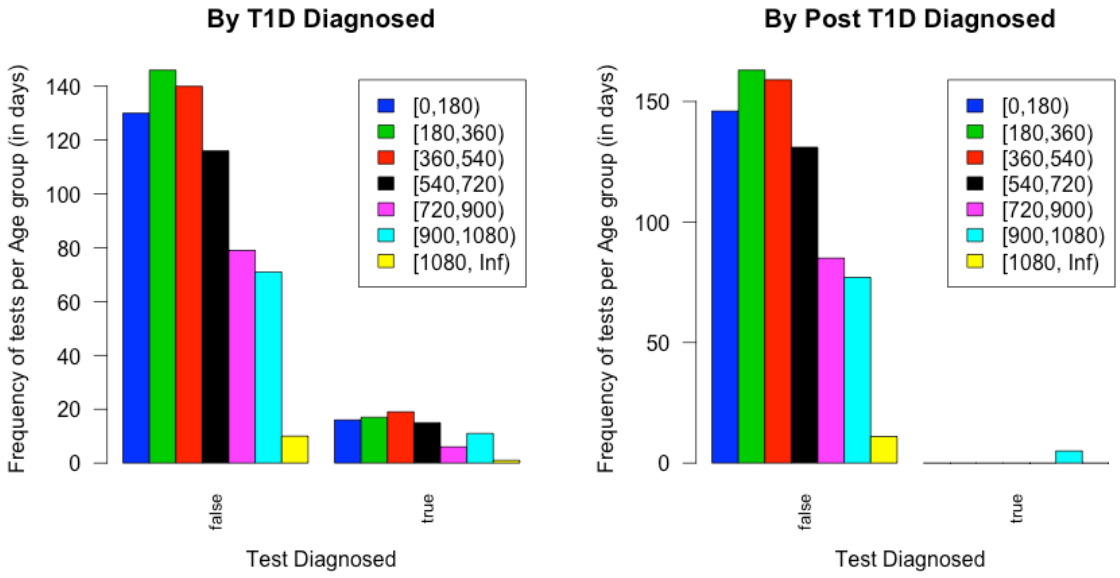
We can see the frequency of each test per child. There are seven children who were tested 30 times. Also there are four children with negative tests. There is just one case where a child with negative and positive tests. The mean frequency of blood sample tests is 23. We also call your attention that mean age is 483 days but with a standard deviation is 295.

Continuous Variables

	median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
Total_Reads	55740.0	74614.6	2300.6	4516.1	4112496000.0	64128.7	0.9
T1D_Diagnosed	1.0	1.1	0.0	0.0	0.1	0.3	0.3
IAA_Level	0.4	2.4	0.2	0.4	30.1	5.5	2.3
GADA_Level	0.0	5.6	0.9	1.8	619.3	24.9	4.5
IA2A_Level	0.1	9.3	3.1	6.1	7623.0	87.3	9.3
ZNT8A_Level	0.1	0.3	0.0	0.1	1.8	1.3	4.6
ICA_Level	0.0	10.6	3.1	6.1	7557.1	86.9	8.2
Age(days)	452.0	482.9	10.6	20.8	86862.6	294.7	0.6
Freq. (ID)	23.0	23.5	0.2	0.4	39.9	6.3	0.3







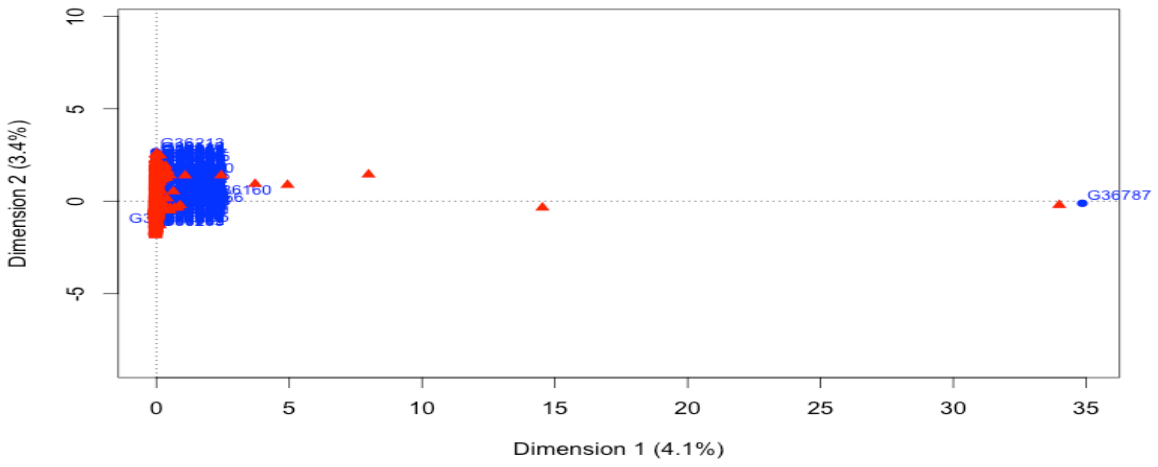
Locking the data, we want to know if bacteria population OTU related with the group age.

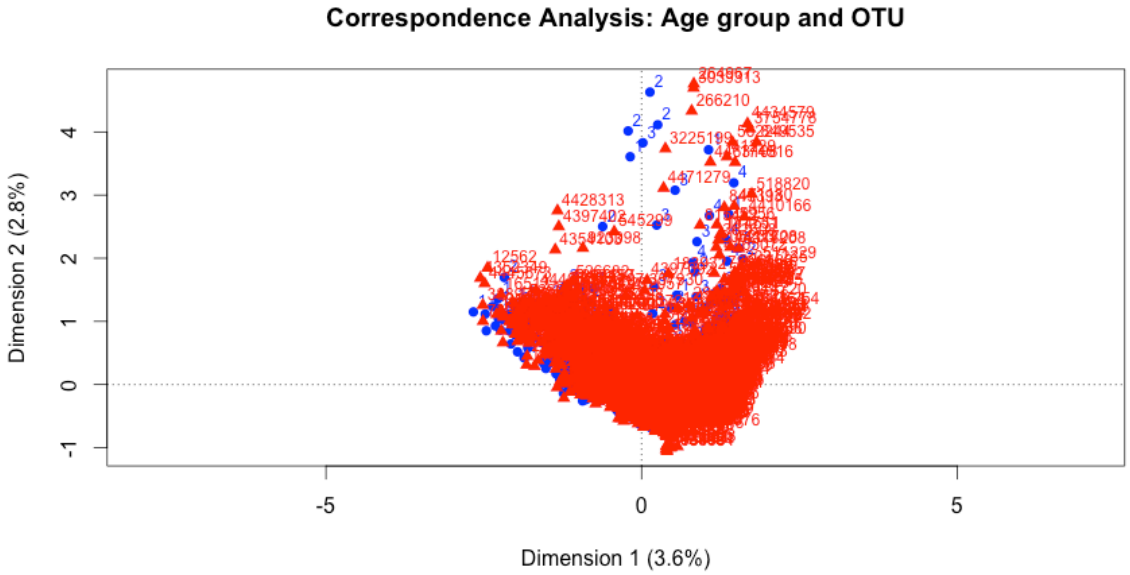
During our analysis, we also need statistical tools to analyze cross-tabulations in order, for instance, to detect and measure the strength of the patterns of association between nominal variables. A number of statistical approaches are used for these purposes, encompassing hypothesis testing, logistic regression, and log-linear modelling. Besides these approaches, Correspondence Analysis (CA) is an exploratory statistical technique frequently applied to contingency tables. CA is now widely used in fields as diverse as archaeology, biology, marketing research, analysis of food preferences, textual analysis and crime studies.

The objectives of CA are quite the same as PCAs. However, the concept of similarity between rows or columns is different. Here, similarities between two rows or two columns are completely symmetric. Two rows (resp.columns) will be close to each other if they associate with the columns (resp.rows) in the same way.

We are looking for the rows (resp.columns) whose distribution is the most different from the populations; those ones that looks the most simmilar, or those which less alike.
Each group of rows (resp. columns) is characterized by the columns (resp. rows) to which it is too much or too little associated.

Correspondence Analysis
We are going to use the first columns (corresponding to the answers to the second question) as active variables and the last ones (corresponding to the third question) as supplementary variables. In the exes of the plot are the percentages of concentration in each dimension. Since we focus on the variability of the OTU composition between subjects related to the age, we use age (in groups) in the CA.





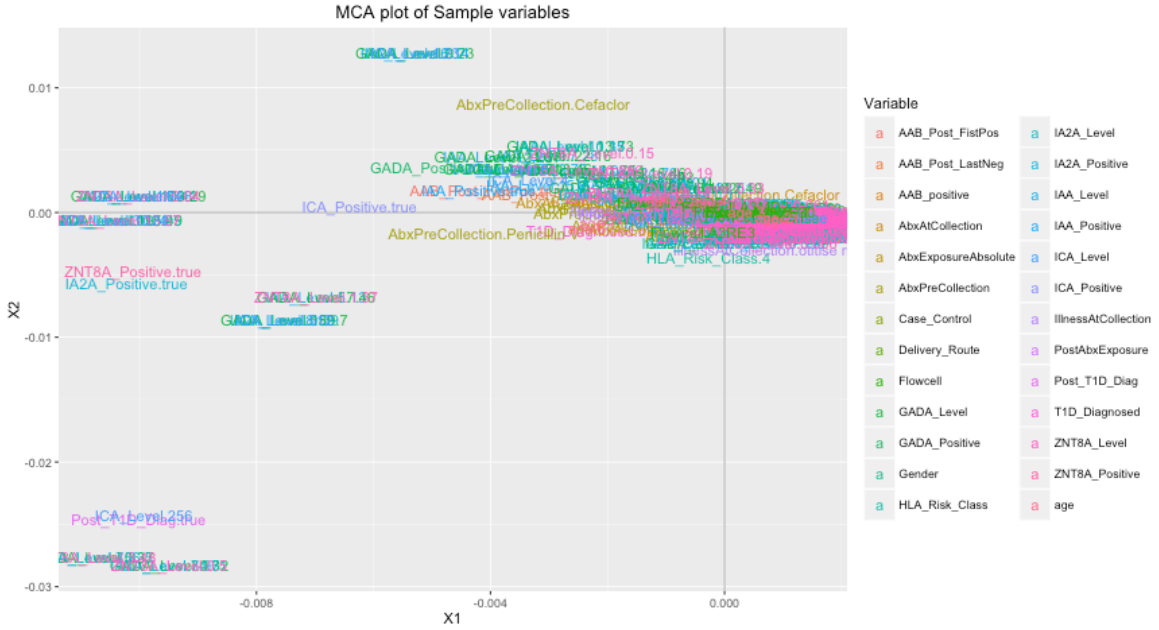
We can see that there are 2 tests that we can qualify as outliers; this is why we use the tests, which are between (-5,5) in both dimensions.
We can see that there are 2 tests that we can qualify as outliers. So, we prefer to move them from our analysis. We can see that by removing this test, we have a group of test related with children between 180 and 360 days since birth. The percentages of concentration in these two dimensions are o greater than 7%, meaning that the variability and concentration is not explained with these two dimensions.
The dispersion is a little better, but qualitatively it's pretty similar to before. This extreme blood sample is from a case control and female with the same diagnose post-test and diagnose from test(false). Our sample is related with Bacteria called Cyanobacteria, class Chloroplast from Streptophyta.

Multiple Correspondence Analysis

In this study, we will use as active variables the ones about consumption behavior and the other variables will be added as supplementary information.

Using the phyloseq dataset, we could find that there are 80 cases that were classified as true T1D but was diagnosed as false after T1D. Also there are 5 that were true diagnosed with T1D after the test.

We can see that the individuals' scatterplot that there is no particular group of individuals. The scatterplot is quite homogeneous.
To interpret the principal components of the MCA, we are going to use extreme individuals (it is easier than using directly groups of individuals). Individuals 777 and 758 are alike to each other. Individuals with a high level of ZNT8A(Zinc transporter 8 Autoantibody) were detected with the same composition than IA2A_Level(insulin autoantibodies) in the T1D test. Also in the sample, IA2A_Level was as high as level of GADA. There are too many blood samples to look at each one by one. That is why we need a representation of the categories.



2. Prediction modeling

This is found in Part two of the document.