# Introduction

\*\*Data\*\*

The data used is the "famous (Fisher's or Anderson's) iris data set gives the measurements in centimetres of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica."

### Q1(b)

In this part new data was generated using factors models. To do so, first a mixture of factor analyser (MFA) was fitted to the 'iris' data. Then using the matrices of the fitted model, the suggested simulation procedure was followed (please see Question pg2).

we can visually analyse the two graphs proposed:

1. QQ-Plots. The plots compare the distribution of the real data versus the simulated data. It can be analysed as we generally use it to compare the data distribution to a theorical distribution. In this case we can see that the distributions are very similar, with only minors departures are the tails of the \*Sepal Width\*.

2. Scatter Plots. Two scatter plots are presented. Each relates the length and width for the sepal and Petal. The black points show the ground truth data, and the red points are the simulated data. It can be seen that the simulated data is almost always contained by the black points, keeping its pattern. Therefore, this method is not creating noise.

In conclusion the simulated data is similar to the real one and stays within the border of the given ground truth.

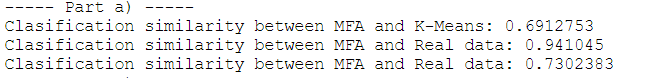
### Q1(c)

There are several ways to select the number of clusters K and the number of dimensions M for this model.

Starting with the number of clusters, K. For this case, from an explanatory point of view, it makes sense to use the same number of factors than the variable 'Species' has, so it easy to understand what the model is doing. But it is also possible to define it by looking at the data and the model. Some possibilities are the Elbow Method and the Silluete method, but for this case, probably the best would be to use Bayesian information criterion (BIC) as it has proven good performance for classification problem, and to be used in mixture model and the likelihood function for factor models is fairly easy to compute for the factor mixture model (as shown in the questions). The methodology to analyse with BIC is, we will have to try different number of K and compute the BIC for each the clustering obtained, and finally select the one with the lowest value.

To select the number of components I would suggest to previously to fit the model, run a Principal Components Analysis (PCA) using the R code 'prcomp' and then see how much each covariable explain the variability of the data using the 'plot' function. This was done during the Lab 8, for this same data (but scaled, as for PCA is normally done), and it was found that only two variables explain 95% of the variance. So I we want to simulate data points using M=2 should also have a good idea. Nevertheless, for prediction purposes we will need to test different M for a given K and compute and error measurement.

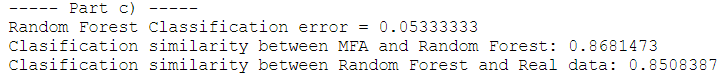
### Q2(b)

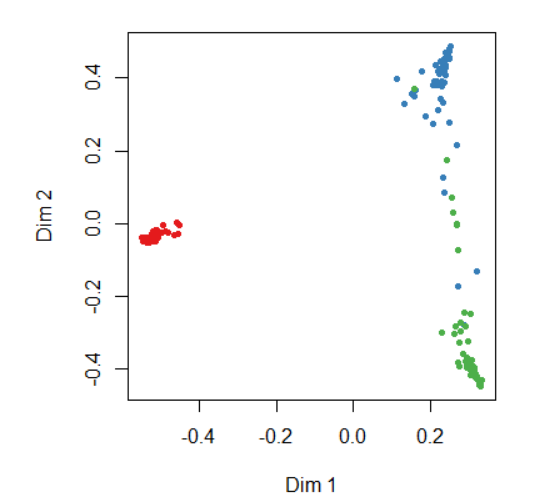


In the console, it is shown the classification similarities obtained using the function 'ari'. We can see that MFA is very similar to the ground truth (~95%) while k-means has greater differences with it (~70%).

One way to improve the k-mean model, is to standardise/scale each variable by dividing by its variance. Doing so, the k-means model can better weight the importance of the different variables to compute the classification, as it would be less affected by the unit of measurement used.

## Q2(c)

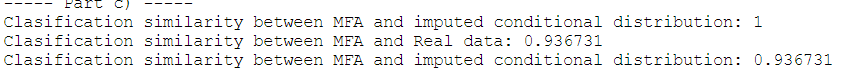




In the result part we can see that the Random Forest model obtained better errors than k-means. Additionally, whilst it only has a ~5% OOB error, the differences with the ground truth are ~85% which is again better than k-means but worse than MFA. Looking at the ‘Proximity plot’ we can see that with some exceptions, the points can be very easily separated with respect the original class, suggesting that the Random Forest model is doing a good work.

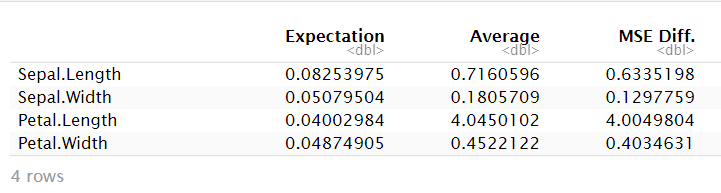
The table above shows the importance that each variable has for the OOB error. We can see that the Random Forest model was able to identify that the original variables are more important than the other ten noise gaussian added. Indeed, the original variables are above 11, while the noise ones are below 3, moreover, six variables have a negative impact on the OOB error, suggesting that they are not adding useful information.

### Q3(c)



We can see that MFA have similarities with the ground truth of ~93%, while the imputed conditional distribution successfully predicted ~77% of the entries. This ~15% difference can be explained by the fact that MFA used all four variables, while the imputation only used three. Additionally, since the similarity between the predictions of the models is ~82%, we can say that there are cases where both models prediction were wrong.

### Q3(d)



The table above summarizes the MSE errors for each variable imputation method, using the expectation imputation and replacing by the mean. We can see that across all the variables, using the mixture of factor analysers to impute missing values has better performance than imputing the missing values with the variable mean. And specially for the variable Petal Length, where the mean replacement MSE is a hundred times bigger than the proposed method.