# Introduction

\*\*Data\*\*

The dataset used is the "famous (Fisher's or Anderson's) iris dataset which 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 we generate new data 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, we followed the simulation procedure suggested in (…)(please see Question pg2).

We can visually analyse the two graphs proposed:

1. QQ-Plots. The plots compare the distributions of both, the real and simulated data. This graph is conventionally used in empirical analysis to compare the actual data distribution to a theorical distribution. In this case, similar distributions are observed, showing only minor differences, in particular departures of 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. Ground true data is represented through the black points, while simulated data is in red.. Graphical evidence show little deviation of the simulated data from the pattern created by the black points.. Therefore, this method is not creating noise.

In conclusion there is no significant difference between the simulated and real data, as the simulated data is contains within the borders of the given ground truth.

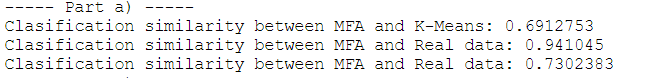
### Q1(c)

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

Starting with the number of clusters, K. In 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. Still, it is also possible to define the number of clusters by looking at the data in conjunction with the model considered. While the Elbow Method and/or the Silluete method are applicable, the Bayesian information criterion (BIC) seems more adequate to this case since its performance for classification problem has already been proven. The BIC has also been used in mixture models and the likelihood function for factor models is fairly easy to compute for the factor mixture model (as shown in the questions). The BIC methodology requires to try different numbers of K and compute the BIC for each of the clustering obtained, and finally select the one with the lowest value.

To select the number of components I would suggest to run a Principal Components Analysis (PCA) using the R code 'prcomp' before fitting the model and then see how much each covariable explains the variability of the data using the 'plot' function. This was done during the Lab 8, for this very data (but scaled, as is conventionally implemented for PCA), and it was found that only two variables explain at least 95% of the variance. So simulating data points using M=2 should yield good results. Nevertheless, for prediction purposes we will need to test different Ms for a given K and compute error measurement.

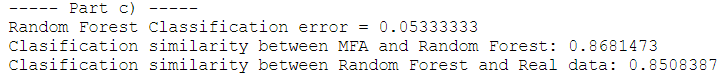
### Q2(b)

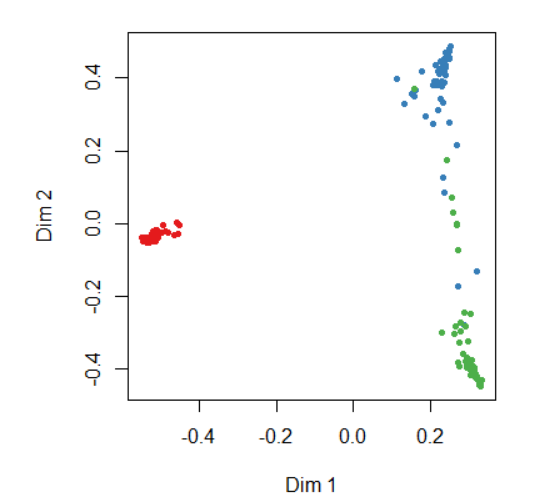


In the console excerpt above, I report the classification similarities obtained using the function 'ari'. We can see that MFA is very similar to the ground truth (~95%) while k-means has greater differences with it (~70%).

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

## Q2(c)

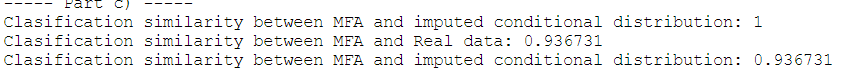




The results are plotted above. We can see that the Random Forest model obtained better errors than k-means. Additionally, whilst it only has a ~5% OOB error, the differences with the ground truth are ~85%, which is again better than k-means but worse than MFA. The ‘Proximity plot’ reveals that, with some exceptions still, the points can be very easily separated with respect to the original class, suggesting that the Random Forest model is good .

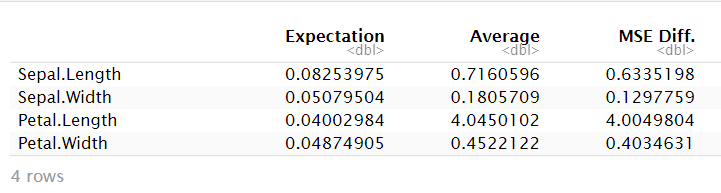
The table above shows the role each variable plays in the computation of the OOB error. We can see that the Random Forest model was able to identify that the original variables are more important than the other ten noise gaussian added. Indeed, the original number of variables is above 11, while the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

### Q3(c)



We can see that MFA has 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 included 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’ predictions were wrong.

### Q3(d)



The table above summarizes the MSE errors for each variable imputation method, using both, the expectation imputation and replacement by the mean. Overall, we can see that, across all variables, using a mixture of factor analysers to impute missing values does yield higher performance than imputing the missing values with the variable mean. This is particularly relevant in the instance of the the variable Petal Length, where the mean replacement MSE is a hundred times higer than the expectation method.