**Questions about project 1**

* *Define the Mahalanobis distance and give its intuitive interpretation:*

The Mahalanobis distance measures the distance between a new observation and the mean of the other observations which is weighted by the variance-covariance matrix. It measures the number of standard deviations from a point to the mean of a distribution in a multi-dimensional way.

* *Explain how to use Mahalanobis distance for outlier detection:*

We can use the Mahalanobis distance to detect outliers derivating their Mahalnobis distance. Then we rebuild the model without the ouliers, by setting an appropriate threshold to the outliers’ Mahalanobis distance already calculated.

**Questions about project 2**

* *What is the principle of K-fold cross-validation for model selection:*

The principle of K-fold cross validation is to partition the original sample by K sub-samples equally distributed. On these K sub-samples, 1 is defined as the validation data and the K-1 others are defined as the training data. Then one uses the result on the training data to test the validation data. One repeats this operation for each sub-sample as validation data.

* *What is the principle of random-x resampling for a regression model:*

The random-x method (or resampling) allows us to generate random sample of size N based on the empirical distribution of the initial sample instead of the population (which is unknown). Then we can compute a regression on each sample and then obtain different estimators (thus we can compute statistics about the estimators).

**Questions about project 3**

* *Explain the difference between the evaluation of root mean square error of an estimator in a simulation framework versus with a bootstrap approach:*

The root mean square error measures the difference between the actual value of the parameter and the estimator whereas a bootstrap approach allows to approximate function of an estimator based on repeated samples.

* *In a Bayesian framework, what is the posterior predictive distribution:*

The posterior predictive distribution is the distribution of possible unobserved values conditional on the observed values from the experiment (values we will predict if we run the experiment again).