Questions workshop 2020

**PCA:**  wines data

* Run the PCA on the data, plot the scores, loadings (or in a bi-plot) and the explained variance
* What does each PC describe?
* How many components should be chosen (how do you figure it out?)
* Why should you scale the data?
* Can you say anything about the different classes in the dataset
* Is PCA sufficient for drawing conclusions?

Run the analysis again, but test for outliers – is the conclusion the same?

**(n)-MDS:** snails

* Run the metric MDS script on the data, plot the results of sammon, classic mds and n-mds
* Which method yields the best results?
* Find the optimal distance measure via the vegan package
* Can you conclude anything from your MDS plots?
* Is MDS sufficient for drawing conclusions? What can you do to enrich the analysis?

**Clustering:**

1) (glass)

* Run the k-means clustering script on the data
* Choose the optimal number of clusters from, gap statistics and “elbow”-plot do the results agree? – Which method do you think is best and why?
* Optional: Try out the NbClust-package: play with different measures of distances and clustering’s – are the results the same?

2) (seed)

* Run the hierchical clustering script on the data, try different dissimilarity measures and clustering methods – any significant differences in the dendrogram plots? – Which method would you prefer?
* Use the assessment function to figure out the best linkage method – do you agree?
* Try to cut the data into subgroups and plot the dendrogram subgroups as clusters
* Find the optimal number of sub-groups (clusters) from the gap statistics or “elbow”-plot

3) (glass)

* Run the Gaussian mixture model script on the data – what were the optimal number of clusters? (How do you identify them?)
* Is the normality assessment always the best to make when generating clusters? – What are the benefits of this assumptions vs the completely unsupervised options in question 1 and 2?

**Multi level data modelling**

**D**ata (dune)

* Run the ANOVA for the data – start with a full analysis including 2 factor interactions – what are your results?
* Calculate a significance level using bonferoni correct and a false discovery rate of q=0.10
* Reduce the model if possible
* Report the final model – what does your p-values state regarding the original hypothesis of the standard ANOVA?
* Do a post-hoc analysis via the Tukey test (only look for the main effects) what can be said about the different factor levels?
* Does ANOVA make sense to use on latent variables?

**Simple OLS, KNN and CV:**

1. OLS (prostate data)

* Preform an ordinary least squares fit both by “hand” (in R) and via the lm function in R containing only main effects.
* Make a model containing all 2nd order interactions and a model containing all 2nd and 3rd order interactions. Which model gives the best RMSE (smallest) when compared to the each other (including the only main effect model).
* Make the same analysis, now using cross-validation – which model now produces the lowest error on the test data? (try different cv methods) – How does this fit with bias variance?

1. KNN (glass data)

* Use the KNN script on the data : set up an appropriate split into training and testing data (try different number than the standard settings)
* How many neighbours provides the lowest error? (try different cv methods to see if the number are the same)

1. CV

* Explain in your own words why CV is needed and how it works, also highlight the differences of loo, repeated CV and non-repeated CV.
* Why is correlation a bad measure to explain a model?

**Imposing sparsity: Lasso, ridge and elastic net**

1. Ridge (prostate )

* Use the ridge script on the data and report the error from the testing set
* Have any of the coefficients become very small? Any of them gone to 0?
* What effect does the parameter have on the model? What happens if it is set to 0?

1. Lasso (prostate)

* Use the Lasso script on the data and report the error from the testing set
* Have any of the coefficients become very small? Any of them gone to 0?
* What effect does the parameter have on the model? What happens if it is set to 0?

1. Elastic net (prostate)

* Use the elastic net script on the data and report the error from the testing set
* Have any of the coefficients become very small? Any of them gone to 0?
* What effect does the and parameter have on the model? What happens if both are set to 0? How can the elastic net be interpreted with regards to the lasso and ridge model?

1. Summary

* Which model gave the smallest error? Which model should you generally use? When should you include sparsity?

**Modelling high dimensional data: PLS and PCR**

1. PCR (sand data )

* How does PCR work – what criteria are we trying to optimize?
* Run the PCR script on the data – report the testing error and optimal number of PCs
* Use Jack-knifing to report the significant parameters in the original domain
* Repeat the process on the categorical dataset using the DA-PCR script
* What happens if number of number of PCs=number of variables?

1. PLS (sand data)

* How does PLS work – what criteria are we trying to optimize?
* Run the PLS script on the data – report the testing error and optimal number of PCs
* Use Jack-knifing to report the significant parameters in the original domain
* Repeat the process on the categorical dataset using the DA-PLS script

Which method is the better PLS or PCR? (In this case or in general?)

Try with the sparsity methods as well for this data set