Statistical Methods 2

Portfolio 3 – Kernel principal component analysis

Complete the following task and submit your work on Blackboard by 4pm Friday 10/02/2023

Task

Using `R` and a dataset of your choice, illustrate the use of kernel principal component analysis (KPCA) for data visualization or classification (or for something else!). For this task you can use a simulated dataset if you want (but do not simulate the data as in the lecture notes; be creative!) and you can perform KPCA using e.g. the `R` package `kernlab`.

The following constraints must be satisfied:

- Your example should illustrate the fact that KPCA can be useful in situations where principal component analysis (PCA) does not work well. This means that you should perform your analysis using both KPCA and PCA, and that the results obtained with the former method should be significantly "better" than those obtained using the latter.
- You must perform the analysis for different types of kernels.
- You must "demonstrate" that your choice for the bandwidth parameter γ of the kernel is suitable for your application. In addition, your results must include results obtained when the "median trick" is used to choose γ.
- You must justify the number of principal components that you keep.
- You must comment all your results.

Remark: If you choose a classification task it could be interesting to compare the results obtained when PCA is applied on $\boldsymbol{\Phi}$ (first approach discussed in the lecture notes) with those obtained when PCA is applied on $\boldsymbol{C}_n\boldsymbol{K}^0$ (second approach discussed in the lecture notes). Note that this latter approach is simply principal component regression (PCR) applied to $\boldsymbol{C}_n\boldsymbol{K}^0$, and that PCR can be performed in `R` using the command `pcr`.