Statistical Methods 2

Portfolio 1 – Principal component analysis

Complete the following tasks and submit your work on Blackboard by 4pm on Friday 27/01/2023.

Task 1 (40 marks)

In `R` you can find the dataset `USArrests` which contains the numbers per 100 000 residents of murder arrests (variable "Murder"), assault arrests (variable "Assault") and rape arrests (variable "Rape") for 50 US states in 1975.

Using only the above three variables, compute the principal components (PCs) for this dataset, interpret them and make a plot showing the resulting two dimensional reduction of the dataset as well as the projection of the variables onto the space spanned by the first two PCs. (Note that there are some `R` commands that allow to easily make this type of plots.)

While performing this task you must respect the following constraints:

- PCA should be performed using both the covariance matrix $\boldsymbol{S}$ and the correlation matrix $\boldsymbol{R}$.
- For the matrix $\boldsymbol{S}$, PCA should be performed "manually", that is, by using only `R` commands for computing the eigenvalues and eigenvectors of a matrix.
- For the matrix $\boldsymbol{R}$, PCA should be performed using the command `prcomp`. Explain what are the components `sdev`, `rotation` and `x` (for `retx=TRUE`) returned by this function (i.e. express these components in term of the notation used in the lectures). In particular, explain how the matrix $\boldsymbol{Y}$ and the eigenvalues of $\boldsymbol{R}$ can be obtained with `prcomp`.
- Discuss whether the correlation matrix or the covariance matrix should be used for this dataset.
- For the PCA that you decide to keep (i.e. the one based on $\boldsymbol{S}$ or the one based on $\boldsymbol{R}$), make a scree plot, compute q_{K} (the number of PCs to keep according to Kaiser' criterion) and $q_{\mathrm{H}}^{(M)}$ (the number of PCs to keep according to Horn's parallel analysis). Decide of the appropriate number q of PCs to keep, and justify your choice.
- All the results you obtain should be interpreted. In particular, discuss how well your two dimensional reduction of the dataset is expected to represent the data.

Task 2 (20 marks)

Choose a dataset which contains different groups of observations (e.g. the `IRIS` dataset, where the groups represent the different kinds of flowers or the `Wine` dataset where the groups represent the type of wines, etc..). Then,

- Choose whether PCA should be applied using the covariance matrix $\boldsymbol{S}$ or the correlation matrix $\boldsymbol{R}$. Justify your choice.

- Plot the two dimensional reduction of observations. On the same plot, represent the orthogonal projection of the variables onto the space spanned by the first two principal components and color the observations according to their group.
- Interpret the results. In particular, has the two dimensional reduction of the dataset separated the different groups of observations? (This should be the case for the `IRIS` and the `Wine` dataset.)

Task 3 (40 marks)

Choose a dataset on which you can fit a regression model of the form $Z_i \sim f(\alpha + \beta^\top x_i^0)$, and estimate the model parameters using principal component regression (PCR). Decide whether PCA should be applied on the covariance matrix $\boldsymbol{S}$ or on the correlation matrix $\boldsymbol{R}$ of the covariates, and justify your choice. Justify also your choice for the number q of principal components to retain.

In addition, if possible compare for different values of $\eta \in (0, 1)$ the PCR estimate of the model parameter obtained with q_η principal components with the estimate obtained when all the variables are used. Otherwise, split the dataset into a training and a test set and, letting $q = q_\eta$, compute the prediction error on the test set for different values of $\eta \in (0, 1)$.

For this task you can for instance fit a linear regression model on the `Communities and Crime` dataset (available from the UCI Machine Learning repository or from the `R` package `mogavs`).