Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
July 19th, 2019

First Name and Family Name:
ID Number:

Problem n.1

The file `pines.txt` collects the GPS coordinates of 324 pines distributed in the *Pineta Dannunziana*, a protected natural area in Pescara. Assume the data to be independent realizations from a bivariate Gaussian distribution.

a) Perform a statistical test of level $\alpha = 0.01$ to verify if the centre of the Pineta Dannunziana can be assumed to be located in position $Long = 14.2350$, $Lat = 42.4520$. Report the p-value of the test and verify its assumptions.

b) Estimate an elliptical region $\mathcal{A}$ that contains 99% of the pines. Report: the analytical expression of the region, its centre, the direction and the length of the principal axes of the ellipse. Report a qualitative plot of the region.

Problem n.2

The file `buoys.txt` collects the GPS coordinates of 752 buoys located in the Adriatic sea and used to measure seawater quality. The file also reports the dissolved oxygen in water [mg/l] at each measured location, recorded on July 15th, 2019.

a) Cluster the buoys *based only on the GPS coordinates* by using a hierarchical clustering method (Euclidean distance and Ward linkage). Report a qualitative plot of the dendrogram and evaluate the number of clusters you deem appropriate for the data. Report the GPS coordinates of the centers of the clusters and their numerosity.

b) Assume the observations at different buoys to be independent. Perform a statistical test to verify if there is a significant difference between the mean dissolved oxygen in the groups identified at point (a). Formulate an appropriate model, state and verify the corresponding assumptions. Comment the results.

Problem n.3

Piadeina is a new successful sandwich franchise mainly based in the Adriatic coastline. The dataset `piadeina.txt` collects data from different shops along the coastline. The variables included in the dataset refer to the number of item sold for several types of food and drinks, along with total sales and a variable indicating the local weather condition (maximum daily temperature).

a) Formulate a linear regression model for the total sales, as a function of all the other variables. Include in the model a possible dependence of the total sales on the categorical variable 'day of the week', but **only in the intercept**. Report the estimates of the 15 parameters of the model (the coefficients $\beta_0, ..., \beta_{13}$ and the errors' standard deviation σ). Analyse the model residuals, and verify the assumptions of the model.

b) Perform a variable selection through a Lasso method, by setting the parameter controlling the penalization to $\lambda = 5$. Report the significant coefficients.

c) Optimize the parameter λ within the range $[0, 100]$ via cross-validation. Report the optimal λ and the corresponding estimated coefficients.

Problem n.4

File `watertemp.txt` contains the mean daily water temperature registered at 132 monitoring stations in the Adriatic Sea, during the 365 days of 2017. The dataset also reports the zone of the measurement (*Deep*, *Medium* or *Surface* water).

a) Perform a smoothing of the data through a projection over a Fourier basis with 45 basis elements. Report the first 3 Fourier coefficients obtained at the Stations 1 and 2.

b) Perform a functional principal component analysis of the smoothed data obtained at point (a). Report the variance explained along the first 5 functional principal components, a qualitative plot of the first 3 eigenfunctions and the screeplot. Interpret the principal components.

c) Having reported a qualitative plot of the scores along the first 2 functional principal components, use the categorical variable *zone* to further enhance the interpretations.

d) Propose a possible dimensionality reduction for the data and discuss the results.