**Politecnico di Milano**
**Scuola di Ingegneria Industriale e dell'Informazione**
Applied Statistics
July 18th, 2017

***First Name and Family Name:***
***ID Number:***

## Problem n.1

The file `tourists.txt` collects data on the flow of Italian tourism from outside Lombardy to Milan for the year 2015. Each statistical unit corresponds to a Region of origin and a month of observation. For each unit, the tourists' flow is quantified through the number of nights spent by clients in: '5 stars hotels', '4 stars hotels', '3 stars hotels', '2 stars hotels', '1 star hotels', 'residences', 'B&B' and 'rented flats'.

a) Perform a Principal Component Analysis of the dataset, by only focusing on the quantitative variables of the dataset; here, evaluate whether it is appropriate to use the original variables or the standardized ones and proceed accordingly. Interpret, when possible, the principal components. Report the numerical value of the first 3 loadings and the variance displayed by the data along those directions.

b) Report (qualitatively) the scatter plot of the data along the first two PCs and describe how to interpret the data clouds in the four quadrants. Use the categorical variables 'Month' and 'Region' to further interpret the results at point (a).

c) Propose (with motivations) a dimension reduction of the dataset.

# Problem n.2

The dataset `horsecolic.txt` reports the data about 300 horses that have been affected by a colic and treated with drugs to alleviate their pain. The included quantitative variables are: ‘Rectal.temperature’, ‘Pulse’, ‘Respiratory.rate’, ‘Packed.cell.volume’ (i.e., the number of red cells by volume in the blood). The dataset also reports an additional ‘pain’ variable, that is a qualitative variable indicating whether the horse is still suffering (Pain = ‘Yes’) or not (Pain = ‘No’). The latter variable is collected through a subjective judgement of experts, based on the evaluation of facial expressions and a behavioral assessment.

a) Build 5 Bonferroni confidence intervals (global level 99%) for the mean difference in ‘Rectal.temperature’, ‘Pulse’, ‘Respiratory.rate’ and ‘Packed.cell.volume’ between the horses with pain and without pain. Comment the results and identify the variables along which a significant difference exists. State and verify the appropriate assumptions.

b) Based only on the assumptions you deem appropriate, build a classifier for the condition ‘pain’, based only on the variables along which the groups display a significant difference according to the analysis at point (a). Report the mean within the groups and the prior probabilities estimated from the sample. Report a qualitative plot of the partition induced by the classifier in the space identified by two of the used variables.

c) Estimate the APER of classifier.

# Problem n.3

The file `castle.txt` collects the GPS coordinates of 27 castles in the province of Aosta. Assume the data to be independent realizations from a bivariate Gaussian distribution.

a) Perform a statistical test to verify if the centre of the distribution is located in the centre of Aosta ($Lat = 45.733$, $Long = 7.333$). Verify the assumptions of the test.

b) Consistently with the results at point (a), estimate an elliptical region that contains 95% of the castles. 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.4

The actual landing distance of an albatross can be modeled as

$$Y = \alpha_g + \beta_g \cdot V_a^2 + \gamma_g \cdot V_i^2 + \varepsilon,$$

where $V_a$ [km/h] is the approaching velocity (i.e., the velocity of the bird at a given position before landing), $V_i$ [km/h] the impact velocity (i.e., the velocity of the bird when it touches the land), $g = 1, 2$ denotes the wind conditions ($g = 1$ for *upwind*, $g = 2$ for *downwind*), and $\varepsilon \sim N(0, \sigma^2)$. Based on the data on the landing distances [m] and velocities [km/h] of 100 albatross landed in Australia (file `albatross.txt`), answer the following questions.

a) Estimate the 7 parameters of the model (report $\alpha_g, \beta_g, \gamma_g$ for $g = 1, 2$ and $\sigma$) and verify its assumptions.

b) Based on appropriate test(s), reduce the model.

c) Using model (b), test the hypothesis according to which $\gamma_g = -\beta_g$, for $g = 1, 2$. Possibly propose a constrained model and estimate its parameters.

d) Wilbur arbatross is giving Bianca and Bernie a lift to Australia. Do you deem the landing of Bianca and Bernie to be safe in case of upwind or downwind wind, on a 17 m long runway, if Wilbur is approaching with $V_a = 35$ km/h and $V_i = 25$ km/h? Base your answer on model (c) and on two intervals of global level 99% for Wilbur's landing distance.