Multivariate Statistical Analysis

Pierpaolo De Blasi

University of Torino
email: pierpaolo.deblasi@unito.it
Moodle: math.i-learn.unito.it/msa

Problem set 1

deadline 6 April 2022 11.59pm

Exercise 1

The *air pollution* data (file `usair.txt`) consists of 7 measurements recorded at $n = 41$ cities in the United States. Variables are `SO2`: Sulphur dioxide content in micrograms per cubic meter, `Neg.Temp`: Average annual temperature in F^o (negative values), `Manuf`: Number of manufacturing enterprises employing 20 or more workers, `Pop`: Population size (1970 census) in thousands, `Wind`: Average annual wind speed in miles per hour, `Precip`: Average annual precipitation in inches, `Days`: Average number of days with precipitation per year.

Ignore the `SO2` variable and concentrate on the remaining 6, two of which relate to human ecology (`Manuf, Pop`) and four to climate (`Neg.Temp, Wind, Precip, Days`).

1.1) Compute the $\bar{x}$, **R** and comment on correlations.

1.2) Make a boxplot of each variable and comment about the presence of outliers (no more than two per variable). Identify these observations.

1.3) Construct a normal Q-Q plot for each variable and comment about normality.

1.4) By using scatter plots, comment on whether the outliers at point 1.2) can be detected from them.

1.5) Construct a chi-square Q-Q plot of the squared Mahalanobis distances and comment about normality.

1.6) Identify multivariate outliers, if any, and compare with the answer to point 1.2)

Exercise 2

For $X = (X_1, X_2, X_3)$ distributed as $N_3(\mu, \mathbf{\Sigma})$,

$$\mu = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}, \quad |\rho| < \sqrt{2}/2$$

2.1) Let PC1 and PC2 be the first two (population) principal components of X. Find ρ such that they account for more than 80% of total variation of X.

2.2) Give an interpretation to PC1 and PC2 in terms of the original variables.

2.3) Find the distribution of

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} X_1 - X_2 \\ X_2 - X_3 \end{bmatrix}$$

2.4) Let $\rho = -2/3$ and $\mathbf{\Sigma}_z$ and μ_z be the corresponding covariance matrix and the mean vector of $Z = (Z_1, Z_2)$. Sketch the ellipse

$$(z - \mu_z)^T \mathbf{\Sigma}_z^{-1} (z - \mu_z) = c^2,$$

in the 2 dimensional space $z = (z_1, z_2)$ by setting the constant "c" such that the ellipse contains 0.95 probability with respect to the joint distribution of Z.

2.5) Comment on how the ellipse would change with $\rho = 2/3$ (no need to draw it).

Exercise 3

The *pen digit* data set (file `pendigits.txt`) was created by collecting 250 samples from 44 writers. These writers were asked to write 250 digits in random order inside boxes of 500 by 500 tablet pixel resolution. The raw data on each of $n = 10992$ handwritten digits consisted of a sequence, (x_t, y_t), $t = 1, 2, ..., T$, of tablet coordinates of the pen at fixed time intervals of 100 milliseconds, where x_t and y_t were integers in the range 0-500. These data were then normalized to make the representations invariant to translation and scale distortions. The new coordinates were such that the coordinate that had the maximum range varied between 0 and 100. Usually x_t stays in this range, because most integers are taller than they are wide. Finally, from the normalized trajectory of each handwritten digit, 8 regularly spaced measurements, (x_t, y_t), were chosen by spatial resampling, which gave a total of $p = 16$ variables. The data includes a class attribute, column `digit`, coded $0, 1, \ldots, 9$, about the actual digit.

3.1) Perform a principal component analysis on the standardized variables. Report standard deviations. Decide how many components to retain in order to achieve a satisfactory lower-dimensional representation of the data. Justify your answer.

3.2) Investigate multivariate normality through the first three principal components.

3.3) Make scatter plots with the first three principal components, while color coding the observations according to the `digit` class.

3.4) Comment about outliers with respect to the first three principal components.