Multivariate Statistical Analysis - Problem Set 1

Pierpaolo De Blasi

deadline 2025-04-08 11.59pm

Instruction: send the report in pdf to pierpaolo.deblasi@unito.it named after your surname; specify in the email whether you have worked in a group and, if so, whom you have worked with (**max 5 per group**).

Exercise 1

For $i = 1, \dots, n$, let Y_i be i.i.d. random variables taking values in $\{1, 2, \dots, p, p+1\}$ with probabilities $\pi_1, \dots, \pi_p, \pi_{p+1} > 0$, $\sum_{j=1}^{p+1} \pi_j = 1$. If we code Y_i via the one-hot vector $Y_i = (Y_{i1}, \dots, Y_{i\,p+1})$ where $Y_{ij} = 1$ when $Y_i = j$, then $\sum_{i=1}^n Y_i$ has multinomial distribution, Multinomial$(n, \pi_1, \dots, \pi_{p+1})$, and the multivariate Central Limit Theorem (CLT) implies

$$\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} N_{p+1}(0, \mathrm{diag}(\pi) - \pi\pi^T)$$

for $\hat{\pi} = n^{-1} \sum_{i=1}^n Y_i$ and $\pi = (\pi_1, \dots, \pi_{p+1})$. Assume n is sufficiently large so that $\sqrt{n}(\hat{\pi} - \pi)$ is normally distributed according to the CLT above and let X be the vector of the first p coordinates.

1. What is the distribution of X? Justify your answer.
2. Let $\mathbf{\Sigma}$ be the $p \times p$ covariance matrix of X. Find the inverse of $\mathbf{\Sigma}$.
3. Let $\pi = (\pi_0, \dots, \pi_0)$ for some $0 < \pi_0 < 1/p$. Find the eigenvalues of $\mathbf{\Sigma}$. How large should p be such that the proportion of variance explained by the last (population) principal component account for less than 20% of total variation of X?
4. Perform a simulation study with $p = 3$, $\pi_0 = 1/4$ and $N = 1000$ Monte Carlo samples of $n = 100$ multinomially distributed Y_i. For $X = (X_1, X_2, X_3)$, make a scatterplot of the N values of X_2 vs X_1 and sketch the ellipse corresponding to the contour of the (theoretical limiting) bivariate density of (X_1, X_2) which contains 95% probability.
5. Find the conditional distributions of $(X_1, X_2)|X_3 = x_3$ and of $X_3|(X_1 = x_1, X_2 = x_2)$.

Exercise 2

The `Boston` data (`MASS` R package) contains housing values in 506 suburbs of Boston. We will work with all variables but `zn`, `chas`, `rad` and `medv`. To find out more about these variables, type `?Boston`.

```
library(MASS)
X<-Boston[,-c(2,4,9,14)]
head(X)
```

```
##      crim indus   nox    rm  age    dis tax ptratio  black lstat
## 1 0.00632  2.31 0.538 6.575 65.2 4.0900 296    15.3 396.90  4.98
## 2 0.02731  7.07 0.469 6.421 78.9 4.9671 242    17.8 396.90  9.14
## 3 0.02729  7.07 0.469 7.185 61.1 4.9671 242    17.8 392.83  4.03
## 4 0.03237  2.18 0.458 6.998 45.8 6.0622 222    18.7 394.63  2.94
```

```
## 5 0.06905  2.18 0.458 7.147 54.2 6.0622 222    18.7 396.90  5.33
## 6 0.02985  2.18 0.458 6.430 58.7 6.0622 222    18.7 394.12  5.21
```

1. Compute the correlation matrix $\boldsymbol{R}$ and comment on the largest 4 correlations.
2. Identify the 3 most extreme univariate outliers.
3. Construct a chi-square Q-Q plot of the squared Mahalanobis distances and comment about normality.
4. Are the univariate outliers identified in point 2. also multivariate outliers? Justify your answer.
5. Perform a principal component analysis on the standardized variables. Decide how many components to retain in order to achieve a satisfactory lower-dimensional representation of the data. Justify your answer.
6. Interpret the first 3 principal components by selecting for each principal component the variables with correlation greater (in absolute value) than 0.4 with that principal component.
7. Describe the 3 outliers identified in point 2. in terms of the first 3 principal components.