The University of New South Wales

Department of Statistics
MATH5855 - Multivariate Analysis

Assignment 3
Due 12th October 2018, 5pm

1. (Use SAS as a software package.) The file `bank.dat` contains the measurements of 100 genuine and 100 forged swiss 1000-frank bills. The columns correspond to the following variables:

- X_1: Length of the bill, X_2 : Height of the bill, measured on the left,
- X_3: Height of the bill, measured on the right, X_4 : Distance of inner frame to lower border,
- X_5: Distance of inner frame to upper border, X_6 : Length of image diagonal.

Perform principal component analysis by working with the **covariance matrix.** Answer the following questions:

i) Estimate the first and the second principal component using the variables X_i. Give a meaningful "interpretation" of these components having in mind the magnitudes and the signs of the component weights.

ii) Perform the same analysis using the **standardized** variables Z_i (i.e., by using the correlation matrix). How many principal components do you need to explain at least 90% of the variability in each case (i.e., when the analysis is performed on the covariance matrix and when it is performed on the correlation matrix). Having in mind the nature of the variables $X_1 - X_6$, why could you state that for this particular data set the analysis using the covariance matrix is superior. Explain your answer.

iii) Create an indicator variable called `forge` with values 1 and 2 for the first and second set of hundred measurements. Then plot the values of the second against the first principal component's value for each of the observations. Label the points by the value of `forge`. Do the first two principal components deliver a good way of separating the forged and the genuine banknotes?

iv) Perform a linear discriminant analysis using the given data set and evaluate the accuracy of the classification by using the `crosslisterr` option. Report your findings.

2. For the covariance matrix $\Sigma \in \mathscr{M}_{p,p}$ it is known that all its elements $\sigma_{ij} > 0$ for every $i, j = 1, 2, \ldots, p$. Using first principals and definitions, prove that:

a) Coefficients of the first principal component are all of the same sign,
b) Coefficients of each other principal component cannot be all of the same sign.

3. Data on $n = 20$ consecutive years has been collected reflecting annual average prices of beef steers X_1 and of hogs X_2 and the annual per capita consumption of beef X_3 and of pork X_4. We are interested in the relation of livestock prices to meat production. The file `price-cons.dat` contains the variables Y (year index) and X_1, X_2, X_3, X_4. We

could proceed by calculating $U = (X_1+X_2)/2, V = X_3+X_4$ and then regressing U on V. A perhaps better procedure would be to construct a (weighted) price index $U = a_1X_1+a_2X_2$ and consumption index $V = b_3X_3 + b_4X_4$ and to look at the maximal correlation between U and V. This is the canonical correlation analysis approach.

i) Find and list both canonical correlations and the related canonical variates. Express the canonical variates using the **raw** coefficients and also by using the **standardized** coefficients. Since the prices are in dollar units but the consumption is in pounds, does it make sense to standardize here?

ii) Formulate the hypothesis of independence of the price index and of the consumption index (intuition shows that it must be rejected). Using the output, explain precisely how the Wilks statistic has been calculated using the roots form the output. Also, explain precisely how the degrees of freedom for the F-approximation have been calculated.

iii) Is one only canonical variable pair enough (i.e., is the second canonical correlation also significant)?

4. In Lecture 8, we formulated a result stating how to calculate the weights of a variance-efficient portfolio of p stocks $X_1, X_2, \ldots, X_p$.

a) Prove the following general linear algebra result: for a non-singular $p \times p$ matrix A and for p-dimensional vectors U and V: $(A + UV')^{-1} = A^{-1} - \frac{1}{1+V'A^{-1}U}A^{-1}UV'A^{-1}$.

b) Using a) (or otherwise) show that for a portfolio of equally-correlated assets whose returns have the **same** variances (that is, when

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \ldots & \ldots & 1 \end{pmatrix} , -\frac{1}{p-1} < \rho < 1)$$

the components in the variance-efficient portfolio have equal weights of $1/p$.

c) Calculate the determinant of Σ. Using the value of the determinant (or otherwise) explain why the restriction $-\frac{1}{p-1} < \rho < 1$ on the common correlation must hold.