University of Granada

VOLUNTARY EXERCISE - 1

Multivariate Statistics

Author:

Quintín Mesa Romero

1 Introduction to the exercise

The **Principal Components Analysis (PCA)** is a technique that tries to condense the information of multiple variables by a linear dimensionality reduction, into a few of them or a few linear combinations of them.

So, **Principal Components** are linear combinations of the original variables that capture the most variance and are mutually orthogonal. The coefficients of these linear combinations are the eigenvectors of the covariance matrix, whose variances are the associated eigenvalues of the eigenvectors.

The process to obtain the first principal component consists of trying to find the weights that maximize the variance. Then, repeating the same process, the second principal component is calculated, but with the restriction that the linear combination must not be correlated with the fisrt p.c.

With regard to the **third principal component**, we have been proposed to solve the following statement:

Task 1

Deduce the third principal component as the third eigenvector of the covariance matrix with maximum explained variance rate and orthogonal to the first and second principal components.

2 Solution

Let $X = (X_1, X_2, ..., X_p)^t$ be a random vector defined by a set of p correlated random variables. We assume that $E[X] = 0$ (centered) and we denote $R = E[XX^t]$ its covariance matrix.

The variables $U_1 = a_1^t X, ..., U_q = a_q^t X$, where $a_i \in \mathbb{R}^q, q \leq p, i = 1, ..., q$ is the eigenvector asociated to the ith eigenvalue with the highest module of R, and must be suitable obtained, are referred to as the **principal components**.

In this section, the **thrid principal component**, U_3, is obtained, following a process similar to the one followed to obtain the U_2, but imposing that it has the highest variance among all linear combinations orthogonal to U_1 and U_2.

To guarantee the existance of this maximum, we have to impose some bounding conditions on the weights vector. In this case, a_3 is a unit vector.

Indeed, the problem to be solved is the following:

$$\begin{aligned} &\max(\mathrm{Var}[U_3]) \\ \text{s.a.} \quad &||a_3|| = a_3^t a_3 = 1 \\ &Cov(U_1, U_3) = 0 \\ &Cov(U_2, U_3) = 0 \end{aligned}$$

Taking into consideration that X is a centered random vector, $E[X] = 0$, so then, $E[a_3^t X] = 0$, which implies that

$$Var[U_3] = E[U_3^2] = E[U_3 U_3] = E[a_3^t X a_3^t X] = E[a_3^t X X^t a_3] = a_3^t E[XX^t] a_3 = a_3^t R a_3$$

Likewise,

$$Cov(U_1, U_3) = E[a_1^t X a_3^t X] = E[a_1^t X X^t a_3] = a_1^t E[XX^t] a_3 = a_1^t R a_3$$

$$Cov(U_2, U_3) = E[a_2^t X a_3^t X] = E[a_2^t X X^t a_3] = a_2^t E[XX^t] a_3 = a_2^t R a_3$$

Therefore, the problem is as follows:

$$\begin{aligned} &\max_{a_3} (a_3^t R a_3) \\ \text{s.a} \quad &||a_3|| = a_3^t a_3 = 1 \\ &a_1^t R a_3 = 0 \\ &a_2^t R a_3 = 0 \end{aligned}$$

Applying the Lagrange Multiplier Theorem for the calculation of constrained extremes, the problem boils down to the following:

$$\max_{a_3} \{a_3^t R a_3 - \lambda(a_3^t a_3 - 1) - \mu a_1^t R a_3 - \gamma a_2^t R a_3\}$$

Derivating the last expression with respect to a_3, considering that R is symmetric and setting it to 0, we obtain,

$$2Ra_3 - 2\lambda a_3 - \mu R a_1 - \gamma R a_2 = 0$$

If we multiply the last expression on the left by a_1^t, it is obtained,

$$2a_1^t R a_3 - 2\lambda a_1^t a_3 - \mu a_1^t R a_1 - \gamma a_1^t R a_2 = 0$$

Taking into account that $a_1^t R a_3 = 0$ (second constraint), $a_1^t R a_2 = 0$ $(Cov(U_1, U_2)) = 0$, and $a_1^t R a_1 \neq 0$, the expression becomes $\mu a_1^t R a_1 = 0$, from which we deduce that $\mu = 0$.

Now, having proved that $\mu = 0$, if we multiply the initial expression by a_2^t, on the left, is obtained,

$$2a_2^t R a_3 - 2\lambda a_2^t a_3 - \gamma a_2^t R a_2 = 0$$

Taking into account that $a_2^t R a_3 = 0$ (third constraint), $a_2^t R a_3 = 0$ $(Cov(U_2, U_3)) = 0$, and $a_2^t R a_2 \neq 0$, the expression becomes $\gamma a_2^t R a_2 = 0$, from which we deduce that $\gamma = 0$.

Thus, the equation to be solved is as follows:

$$2Ra_3 - 2\lambda a_3 = 0 \Longleftrightarrow (R - \lambda I)a_3 = 0$$

from which we deduce that a_3 is the eigenvector associated to the eigenvalue λ of the matrix R.

Finally, from that result, we deduce that $Var[U_3] = a_3^t R a_3 = \lambda a_3^t a_3 = \lambda$.

In conclussion, the **third principal component is $U_3 = a_3^t X$, where a_3 is the eigenvector associated to the third eigenvalue with the highest module of the covariance matrix, R.**