Unit 3.- Principal component analysis (PCA)

Course: MULTIVARIATE STATISTICS

©Prof. José Luis Romero Béjar - Carlos Francisco Salto Díaz

November, 2023

Objective and usefulness

- The **objective** of Principal Component Analysis (PCA) **is to condense the information** provided by multiple variables into a few of them or a few **linear combinations** of them (with **maximum variability**).
- Its **main utility** is as a **preliminary analysis** before applying other statistical techniques such as regression, clustering, etc.

Limitations

- The main drawback of these methods is the **difficulty in validating the results**.
- It is a **method highly sensitive to outliers** (extreme or atypical values).

Prior assumptions

- **Correlated variables**.
- **Absence of outliers**.
  Outliers in any of the variables **require a detailed analysis** as they influence the final outcome of dimensionality reduction.
- **Standardized data** (mean 0 and standard deviation 1).
  This prevents variables with a larger scale from dominating the others.

What is a principal component?

- **Principal components** are linear combinations of the original variables with maximum variance and are perpendicular to each other.
- In the following section, it will be demonstrated that the **coefficients** of these linear combinations are the **eigenvectors of the covariance matrix**, and their **variances** will be the **associated eigenvalues** of these eigenvectors.
- In other words, PCA **identifies the directions in which the variance is higher**.

Practical calculation of principal components

- To obtain the **first principal component**, an **optimization problem** is solved to find the weights (loadings) that **maximize the variance**.
- Once the **first component is calculated**, the **second component** is calculated by repeating the same process, but adding the condition that the linear combination cannot be correlated with the first component. This is equivalent to saying they **must be orthogonal**. **The process is iteratively repeated** until all possible components are calculated or until it is decided to stop the process.
- In the next section, it will be justified that one way to solve this optimization problem is by calculating **eigenvectors and eigenvalues** of the covariance matrix.
- The order of **importance of the components** will be determined by the magnitude of the eigenvalue associated with each eigenvector.

Interpretation of the principal components

- The vector defining the **first principal component** follows the **direction in which the observations vary the most**.
- The **second component** follows the direction in which the data exhibit **the highest variance** and **is uncorrelated with the first one** (they are orthogonal directions), and so on for the third and subsequent principal components.

Proportion of explained variance

- How much of the original information is lost when projecting the observations into a lower-dimensional space? In other words, how much information does each of the obtained principal components capture?

  This information is provided by **the proportion of explained variance**, as well as **the cumulative proportion of explained variance**.
- These quantities are very **important** when **deciding on the appropriate number** of principal components.

Appropriate number of principal components

There is no single criterion or method that allows for identifying the optimal number of principal components to use. Different approaches can include:

- **Evaluate the cumulative proportion of explained variance** and select the **minimum number of components beyond which the increase ceases to be substantial**.
- Calculate the **average of the variances explained by each principal component**, and take as many principal components as the number of variances that exceed this average.
- etc.

Notations

In this section, a formal justification is provided for how **the principal components are identified with the eigenvectors of the covariance matrix**, as well as **their variances being identified with the associated eigenvalues** of said vector.

Let $X_1, X_2, \ldots, X_p$ be a set of p **correlated random variables**. Let's denote $X = (X_1, X_2, \ldots, X_p)^t$ the random vector they define. It is assumed X is **centred**, $E[X] = 0$, and let's denote by $R = E[XX^t]$ its **covariance matrix**.

Let's consider **(no more of p)** variables defined by: $U_1 = a_1^t X, \ldots, U_q = a_q^t X$. The objective pursued is to obtain the suitable $a_1, \ldots, a_q \in \mathbb{R}^p, q \leq p$.

Prior requirements:

- $U_1 = a_1^t X, \ldots, U_q = a_q^t X$ must be **uncorrelated**. This way, **redundant information will be eliminated**.
- The **variance** of each $U_i, i \in 1, \ldots, q$ is **maximum**. This way, the new variables **will provide meaningful information**.

Problem statement

Under the aforementioned conditions, the goal is to find $U_1 = a_1^t X, \ldots, U_q = a_q^t X$, mutually uncorrelated, with each U_i having maximum variance among all linear combinations of X that are uncorrelated with $U_1 = a_1^t X, \ldots, U_{i-1} = a_{i-1}^t X$.

- The variables $U_1 = a_1^t X, \ldots, U_q = a_q^t X$ solution to the previous problem are referred to as the **principal components**.

Problem resolution

As previously mentioned, the **resolution of this problem is sequential**:

- First, U_1 is obtained by imposing that it has maximum variance.
- Next, U_2 **is obtained** by imposing that it has the highest variance among all uncorrelated (orthogonal) linear combinations with U_1.
- The same procedure is followed to **obtain** U_3 by now imposing that it has the highest variance among all linear combinations orthogonal to U_1 and U_2
- For the rest of the principal components **up to** U_q**, the procedure is the same**.

Next, it will be justified **how the coefficients of the principal components are the eigenvectors of the covariance matrix**, associated with the eigenvalues of largest magnitude at each step.

Problem resolution - Step 1

In this first step, we obtain **the first principal component**, U_1, **maximizing its variance**. To ensure the existence of this maximum, bounding conditions on the weight vector must be imposed, in this case, that a_1 **is a unit vector**.

$$\max Var[U_1]$$
$$\text{s.a.} ||a_1|| = a_1^t a_1 = 1$$

Considering that X is a centred random vector, $E[X]$ =0, then $E[a_1^t X] = 0$, which implies that:

$$Var[U_1] = E[U_1^2] = E[a_1^t X a_1^t X] = E[a_1^t X X^t a_1] = a_1^t E[XX^t] a_1 = a_1^t R a_1,$$

and therefore, the problem becomes as follows,

$$\max_{a_1} a_1^t R a_1$$
$$\text{s.a. } a_1^t a_1 = 1$$

Finally, applying the *Lagrange Multiplier Theorem* for the calculation of constrained extremes, the problem is reduced to,

$$\max_{a_1} \{a_1^t R a_1 - \lambda(a_1^t a_1 - 1)\}$$

Problem resolution - Step 1 (continuation)

Taking the derivative of the above expression with respect to a_1 (matrix-wise and considering that R is symmetric) and setting it to zero, $\frac{\partial(a_1^t R a_1 - \lambda(a_1^t a_1 - 1))}{\partial a_1} = 0$, is obtained,

$$2Ra_1 - 2\lambda a_1 = 0 \tag{1}$$

Remark:

- The derivative of a quadratic form is: $\frac{\partial(x^t A x)}{\partial x} = (A + A^t)x$ such that $x \in \mathbb{R}^n, A \in M_n(\mathbb{R})$.

It is easy to realize that a_1 **is the eigenvector associated with** λ, the eigenvalue of R, since the previous expression is written as,

$$(R - \lambda I)a_1 = 0,$$

which determines the **eigensubspace associated with** λ. Finally, it is easy to deduce that λ is the variance of U_1 since,

$$Var[U_1] = a_1^t R a_1 = \lambda a_1^t a_1 = \lambda,$$

multiplying (1) on the left by a_1^t and because a_1 is unitary.

In conclusion, the first principal component is $U_1 = a_1^t X$ **with** a_1 **the eigenvector associated with the eigenvalue of** R **with the highest magnitude.**

Problem resolution - Step 2

In this second step, **the second principal component**, U_2, is obtained, uncorrelated with the first principal component calculated previously, **maximizing its variance**.

To guarantee the existence of this maximum, bounding conditions must also be imposed on the weight vector, in this case a_2 is also **a unit vector**.

$$\begin{aligned} &\max Var[U_2] \\ \text{s.a.} &||a_2|| = a_2^t a_2 = 1 \\ &cov(U_1, U_2) = 0 \end{aligned}$$

Taking into account that X is a centred random vector, $E[X] = 0$, then $E[a_2^t X] = 0$, which implies that, as before, $Var[U_2] = a_2^t R a_2$.

Likewise $cov(U_1, U_2) = E[a_1^t X a_2^t X] = E[a_1^t X X^t a_2] = a_1^t E[XX^t] a_2 = a_1^t R a_2$ and therefore the problem is as follows,

$$\begin{aligned} &\max_{a_2} a_2^t R a_2 \\ \text{s.a.} \;\; &a_2^t a_2 = 1 \\ &a_1^t R a_2 = 0 \end{aligned}$$

Finally, applying the *Lagrange Multiplier Theorem* for the calculation of constrained extremes, the problem is reduced to,

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

Problem resolution - Step 2 (continuation)

Taking the derivative of the above expression with respect to a_2 (matrix-wise and considering that R is symmetric) and setting it to zero, is obtained,

$$2Ra_2 - 2\lambda a_2 - \mu Ra_1 = 0$$

Remark:

- The derivative of a linear form is: $\frac{\partial(a^t x)}{\partial x} = a$ such that $a, x \in \mathbb{R}^n$.

Multiplying this expression on the left by a_1^t, it is derived,

$$2a_1^t Ra_2 - 2\lambda a_1^t a_2 - \mu a_1^t Ra_1 = 0$$

Considering that $a_1^t Ra_2 = 0$ (it's the second constraint of the problem), $a_1^t a_2 = 0$ (they are perpendicular), and $a_1^t Ra_1 \neq 0$, the expression becomes $\mu a_1^t Ra_1 = 0$, from which we deduce that $\mu = 0$.
So the equation to solve is

$$2Ra_2 - 2\lambda a_2 = 0 \rightarrow (R - \lambda I)a_2 = 0, \quad (2)$$

from which it is again deduced that a_2 is the eigenvector associated with the eigenvalue λ of the matrix R.

In the same way, $Var[U_2] = a_2^t Ra_2 = \lambda a_2^t a_2 = \lambda$, multiplying the previous equation on the left by a_2^t and considering that a_2 is unitary.

In conclusion, the second principal component is $U_2 = a_2^t X$, where a_2 is the eigenvector associated with the second eigenvalue of largest magnitude of matrix R.

Problem resolution - Step 3 and onwards

In this third step, **the third principal component**, $U_{\mathbf{3}}$, uncorrelated with the previously calculated first and second principal components, is obtained by **maximizing its variance**.

To ensure the existence of this maximum, bounding conditions on the weight vector must also be imposed, in this case, that $a_{\mathbf{3}}$ is also a **unit vector**.

$$\begin{aligned} & \max Var[U_{\mathbf{3}}] \\ \text{s.a.} & ||a_{\mathbf{3}}|| = a_{\mathbf{3}}^t a_{\mathbf{3}} = 1 \\ & cov(U_{\mathbf{1}}, U_{\mathbf{3}}) = 0 \\ & cov(U_{\mathbf{2}}, U_{\mathbf{3}}) = 0 \end{aligned}$$

Voluntary task: to justify that a_3 is the eigenvector associated with the third eigenvalue of largest magnitude of matrix R.

Práctica 1.1 de ACP

In this practice, a **first example of dimensionality reduction** is performed on a dataset. To complete it, you need to **download and run** the file **Practice_1.1_PCA.R** available on the PRADO platform.

Topics covered:

- Conducting a preliminary exploratory analysis of the data to identify possible **missing data (NA - not available)** and **outliers**.
- **Making decisions and handling** missing data and outliers.
- Performing a **PCA (Principal Component Analysis)**
- Initial methods for **choosing the optimal number** of principal components.
- Interpretation of various **interesting graphical outputs** for this method.
- **R language:** loading data from an R package, *data.frame* object, graphical treatment of data, and construction of procedures or functions.

PCA Practice 1.2

In this practice, a **second example of dimensionality reduction** is performed on a dataset. To complete it, you need to **download and run** the file **Practice_1.2_PCA.R** available on the PRADO platform.

The emphasis will be on:

- Conducting a preliminary exploratory analysis of the data to identify possible **missing data** and **outliers**.
- **Making decisions and handling** missing data and outliers.
- Performing a **PCA (Principal Component Analysis)**.
- **Choosing the optimal number** of principal components.
- Interpretation of various **interesting graphical outputs** for this method.
- **R language:** RMarkdown notebook, loading external data files, methods **apply**, **tapply**, **width**, **by**, etc. for function debugging.
- Moving towards the **final report**.

[1] Anderson, T.W. (2003, 3ª ed.). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons.

[2] Gutiérrez, R. y González, A. (1991). Estadística Multivariable. Introducción al Análisis Multivariante. Servicio de Reprografía de la Facultad de Ciencias. Universidad de Granada.

[3] Härdle, W.K. y Simar, L. (2015, 4ª ed.). Applied Multivariate Statistical Analysis. Springer.

[4] Johnson, R.A. y Wichern, D.W. (1988). Applied Multivariate Analysis. Prentice Hall International, Inc.

[5] Rencher, A.C. y Christensen, W.F. (2012, 3ª ed.). Methods of Multivariate Analysis. John Wiley & Sons.

[6] Salvador Figueras, M. y Gargallo, P. (2003). Análisis Exploratorio de Datos. Online en http://www.5campus.com/leccion/aed.

[7] Timm, N.H. (2002). Applied Multivariate Analysis. Springer.

[8] Vera, J.F. (2004). Análisis Exploratorio de Datos. ISBN: 84-688-8173-2.