Tema 3.- Análisis de componentes principales (ACP)

Asignatura: ESTADÍSTICA MULTIVARIANTE (Prácticas)

Grados en: Física y Matemáticas; Ingeniería Informática y Matemáticas; Matemáticas

(4º Curso - 1er semestre 2023-2024)

©Prof. Dr. José Luis Romero Béjar

Departamento de Estadística e Investigación Operativa
Facultad de Ciencias (Despacho 10)

Periodo de docencia: 11/09/2023 a 22/12/2023

Objetivo y utilidad

- El **objetivo** del análisis de componentes principales (ACP) **es condensar la información** aportada por múltiples variables en unas pocas de ellas o en unas pocas **combinaciones lineales** de ellas (con **máxima variabilidad**).
- Su **utilidad** principal es como **análisis preliminar** antes de aplicar otras técnicas estadísticas previas como regresión, clustering, etc.

Limitaciones e inconvenientes

- El principal inconveniente que presentan este tipo de métodos es la **dificultad para validar los resultados**.
- Es un **método muy sensible a valores extremos o atípicos** (outliers).

Requisitos previos

- **Variables correladas**.
- **Ausencia de outliers**.

  Los datos extremos o atípicos (outliers) en alguna de las variables **requieren de un análisis pormenorizado** dado que influyen en el resultado final de la reducción de la dimensión.
- **Datos estandarizados** (media 0 y desviación estándar 1).

  Así se evita que aquellas variables cuya escala sea mayor dominen al resto.

¿Qué es una componente principal?

- Las **componentes principales** son combinaciones lineales de las variables originales con máxima varianza y perpendiculares entre sí.
- En la sección siguiente se probará que los **coeficientes** de estas combinaciones lineales son los **vectores propios de la matriz de covarianzas** y que sus **varianzas** serán los **valores propios asociados** a estos vectores propios.
- Es decir el ACP **identifica las direcciones en las que la varianza es mayor**.

Cálculo práctico de componentes principales

- Para obtener la **primera componente principal** se resuelve un **problema de optimización** para encontrar el valor de los pesos (loadings) con los que **se maximiza la varianza**.
- Una vez **calculada la primera** componente **se calcula la segunda** repitiendo el mismo proceso, pero añadiendo la condición de que la combinación lineal no puede estar correlacionada con la primera componente. Esto equivale a decir que **tienen que ser perpendiculares**. **El proceso se repite de forma iterativa** hasta calcular todas las posibles componentes o hasta que se decida detener el proceso.
- En la siguiente sección se justificará que una forma de resolver este problema de optimización es mediante el cálculo de **vectores y valores propios** de la matriz de covarianzas.
- El orden de **importancia de las componentes** vendrá dado por la magnitud del valor propio asociado a cada vector propio.

Interpretación de las componentes principales

- El vector que define la **primera componente principal** sigue la **dirección en la que las observaciones varían más**.
- La **segunda componente** sigue la dirección en la que los datos muestran **mayor varianza** y **no está correlacionada con la primera** (son direcciones perpendiculares), y así sucesivamente con la tercera y sucesivas componentes principales.

Proporción de varianza explicada

- ¿Qué cantidad de información original se pierde al proyectar las observaciones en un espacio de dimensión inferior?, es decir, ¿qué cantidad de información es capaz de capturar cada una de las componentes principales obtenidas?

  Esta información viene dada por **la proporción de varianza explicada**, así como **la proporción de varianza explicada acumulada**.
- Estas cantidades son muy **importantes** a la hora de **decidir el número adecuado** de componentes principales.

Número adecuado de componentes principales

No hay un criterio o método único que permita identificar el número óptimo de componentes principales a utilizar. Distintas formas de proceder pueden ser:

- **Evaluar la proporción de varianza explicada acumulada** y seleccionar el **número de componentes mínimo a partir del cual el incremento deja de ser sustancial**.
- Obtener el **promedio de las varianzas explicadas por cada componente principal y quedarnos tantas componentes principales como número de varianzas superen este promedio**.
- etc.

Planteamiento del problema

En esta sección se justifica de manera formal como las **las componentes principales se identifican con los vectores propios de la matriz de covarianzas** así como que **sus varianzas se identifican con el valor propio** asociado a dicho vector.

Sean $X_1, X_2, \ldots, X_p$ un conjunto de p **variables aleatorias correladas**. Denotemos por $X = (X_1, X_2, \ldots, X_p)^t$ al vector aleatorio que forman. Asumimos que X es **centrado**, $E[X] = 0$ y denotamos por $R = E[XX^t]$ su **matriz de covarianzas**.

Consideramos **(no más de p)** variables de la forma: $U_1 = a_1^t X, \ldots, U_q = a_q^t X$. El objetivo que se persigue es obtener los $a_1, \ldots, a_q \in \mathbb{R}^p, q \leq p$, adecuados.

Requerimientos previos:

- $U_1 = a_1^t X, \ldots, U_q = a_q^t X$ deben ser **incorreladas**. De esta forma **se eliminará información redundante**.
- La **varianza** de cada $U_i, i \in \{1, \ldots, q\}$ es **máxima**. De esta forma las nuevas variables **proporcionarán información significativa**.

Enunciado del problema

En las condiciones anteriores, el objetivo es encontrar $U_1 = a_1^t X, \ldots, U_q = a_q^t X$, mutuamente incorreladas, teniendo cada U_i máxima varianza entre todas las combinaciones lineales de X incorreladas con $U_1 = a_1^t X, \ldots, U_{i-1} = a_{i-1}^t X$.

- Las variables $U_1 = a_1^t X, \ldots, U_q = a_q^t X$ solución del problema anterior reciben el nombre de **componentes principales**.

Resolución del problema

Tal y como se ha indicado anteriormente, la **resolución de este problema es secuencial**:

- **En primer lugar se obtiene** U_1 imponiendo que tiene máxima varianza.
- A continuación **se obtiene** U_2 imponiendo que es la de mayor varianza entre todas las combinaciones lineales incorreladas (perpendiculares) a U_1.
- Se procede del mismo modo para **obtener** U_3 imponiendo ahora que es la de mayor varianza entre todas las combinaciones lineales perpendiculares a U_1 y U_2.
- Para el resto de componentes principales **hasta** U_q **se procede del mismo modo**.

A continuación se va a justificar **cómo los coeficientes de las componentes principales son los vectores propios de la matriz de covarianzas**, asociados a los valores propios de mayor módulo en cada paso.

Resolución del problema - Paso 1

En este primer paso se obtiene **la primera componente principal**, U_1, **maximizando su varianza**. Para garantizar la existencia de este máximo han de imponerse condiciones de acotación sobre el vector de pesos, en este caso que a_1 **es un vector unitario**.

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

Teniendo en cuenta que X es un vector aleatorio centrado, $E[X]$ =0, se tiene que $E[a_1^t X] = 0$, lo que implica que:

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

y por tanto el problema queda como sigue,

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

Finalmente, aplicando el Teorema de los multiplicadores de Lagrange para la obtención de extremos condicionados, el problema se reduce a,

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

Resolución del problema - Paso 1 (Continuación)

Derivando la expresión anterior respecto a_1 (matricialmente y teniendo en cuenta que R es simétrica) e igualando a cero, $\frac{\partial(a_1^t R a_1 - \lambda(a_1^t a_1 - 1))}{\partial a_1} = 0$, se obtiene,

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

Observación:

- La derivada de una forma cuadrática es: $\frac{\partial(x^t Ax)}{\partial x} = (A + A^t)x$ tal que $x \in \mathbb{R}^n, A \in M_n(\mathbb{R})$.

Es fácil darse cuenta que a_1 es vector propio asociado a λ, valor propio de R, ya que la expresión anterior se escribe como,

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

que determina el subespacio propio asociado a λ. Finalmente, es fácil deducir que λ es la varianza de U_1 ya que,

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

multiplicando (1) a la izquierda por a_1^t y porque a_1 es unitario.

En conclusión, la primera componente principal es $U_1 = a_1^t X$ con a_1 el vector propio asociado al valor propio de R con mayor módulo.

Resolución del problema - Paso 2

En este segundo paso se obtiene **la segunda componente principal**, U_2, incorrelada con la primera componente principal calculada anteriormente, **maximizando su varianza**.

Para garantizar la existencia de este máximo también han de imponerse condiciones de acotación sobre el vector de pesos, en este caso que a_2 es también **un vector unitario**.

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

Teniendo en cuenta que X es un vector aleatorio centrado, $E[X] = 0$, se tiene que $E[a_2^t X] = 0$, lo implica que, como antes, $Var[U_2] = a_2^t R a_2$.

Del mismo modo $cov(U_1, U_2) = E[a_1^t X (a_2^t X)^t] = E[a_1^t X X^t a_2] = a_1^t E[XX^t] a_2 = a_1^t R a_2$ y por tanto el problema queda como sigue,

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

Resolución del problema - Paso 2 (Continuación)

Finalmente, aplicando el Teorema de los multiplicadores de Lagrange para la obtención de extremos condicionados, el problema se reduce a,

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

Derivando la expresión anterior respecto a a_2 (matricialmente y teniendo en cuenta que R es simétrica) e igualando a cero se obtiene,

$$2Ra_2 - 2\lambda a_2 - \mu R a_1 = 0$$

Observación:

- La derivada de una forma lineal es: $\frac{\partial(a^t x)}{\partial x} = a$ tal que $a, x \in \mathbb{R}^n$.

Si multiplicamos esta expresión a la izquierda por a_1^t se obtiene,

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

Teniendo en cuenta que $a_1^t R a_2 = 0$ (es la segunda restricción del problema), que $a_1^t a_2 = 0$ (son perpendiculares) y que $a_1^t R a_1 \neq 0$, la expresión queda como $\mu a_1^t R a_1 = 0$, de donde se deduce que $\mu = 0$.

Resolución del problema - Paso 2 (Continuación)

Así que la ecuación a resolver es

$$2Ra_{\mathbf{2}} - 2\lambda a_{\mathbf{2}} = 0 \rightarrow (R - \lambda I)a_{\mathbf{2}} = 0, \tag{2}$$

de donde nuevamente se deduce que $a_{\mathbf{2}}$ es el vector propio asociado al valor propio λ de la matriz R.

De la misma forma, $Var[U_{\mathbf{2}}] = a_{\mathbf{2}}^t R a_{\mathbf{2}} = \lambda a_{\mathbf{2}}^t a_{\mathbf{2}} = \lambda$, multiplicando la ecuación anterior a la izquierda por $a_{\mathbf{2}}^t$ y teniendo en cuenta que $a_{\mathbf{2}}$ es unitario.

En conclusión, la segunda componente principal es $U_{\mathbf{2}} = a_{\mathbf{2}}^t X$, siendo $a_{\mathbf{2}}$ el vector propio asociado al segundo valor propio de mayor módulo de la matriz R.

Resolución del problema - Paso 3 y sucesivos

En este tercer paso se obtiene **la tercera componente principal**, U_3, incorrelada con la primera y segunda componentes principales calculadas anteriormente, **maximizando su varianza**.

Para garantizar la existencia de este máximo también han de imponerse condiciones de acotación sobre el vector de pesos, en este caso que a_3 es también **un vector unitario**.

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

Tarea voluntaria: justificar que a_3 es el vector propio asociado al tercer valor propio de mayor módulo de la matriz R.

Práctica 1.1 de ACP

En esta práctica se realiza un **primer ejemplo de reducción de la dimensión** en un conjunto de datos. Para la realización de la misma hay que **descargar y ejecutar** el archivo **Practica_1.1_ACP.R** disponible en la plataforma PRADO.

Aspectos tratados:

- Realizar un análisis exploratorio previo de los datos para identificar posibles **datos perdidos** y **valores extremos**.
- **Tomar decisiones y tratar** datos perdidos y valores extremos.
- Realizar un **ACP**.
- Primeros métodos para la **elección del número óptimo** de componentes principales.
- Interpretación de distintas salidas **gráficas de interés** para este método.
- **Lenguaje R:** carga de datos desde un paquete de R, objeto data.frame, tratamiento gráfico de datos y construcción de procedimientos o funciones.

Práctica 1.2 de ACP

En esta práctica se realiza un **segundo ejemplo de reducción de la dimensión** en un conjunto de datos. Para la realización de la misma hay que **descargar y ejecutar** el archivo **Practica_1.2_ACP.R** disponible en la plataforma PRADO.

Se incidirá en:

- Realizar un análisis exploratorio previo de los datos para identificar posibles **datos perdidos** y **valores extremos**.
- **Tomar decisiones y tratar** datos perdidos y valores extremos.
- Realizar un **ACP**.
- **Elección del número óptimo** de componentes principales.
- Interpretación de distintas salidas **gráficas de interés** para este método.
- **Lenguaje R**: notebook **RMarkdown**, carga de ficheros de datos externos, métodos **apply**, **tapply**, **width**, **by**, etc. para la depuración de funciones.
- Hacia el **informe final**.

[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.