Reducción y Visualización de Datos (2023) - Práctica 2

Componentes principales

2.1 Dado un vector aleatorio $\mathbf{x} \in \mathbb{R}^3$ de media cero, y matriz de varianzas y covarianzas

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

a. Calcular los autovalores y autovectores de $\mathbf{\Sigma}$ y reescribirla como $\mathbf{\Sigma} = \mathbf{U\Lambda U^t}$, donde $\mathbf{U}$ es ortonormal y $\mathbf{\Lambda}$ es diagonal.
b. Definir las componentes principales ξ_j, con $j \in \{1, 2, 3\}$, en función de las variables originales X_j. Escribir la transformación lineal del vector $\mathbf{x}$ al vector $\boldsymbol{\xi} = (\xi_1\ \xi_2\ \xi_3)^t$ usando un producto matricial de la forma $\boldsymbol{\xi} = \mathbf{A}^t\mathbf{x}$, donde las columnas de $\mathbf{A}$ son las direcciones de proyección.
c. Calcular la matriz de varianzas y covarianzas de $\boldsymbol{\xi}$. ¿Qué observa?
d. Llamamos *variabilidad total* de un vector aleatorio a la traza de su matriz de varianzas y covarianzas. Calcular la variabilidad total de $\mathbf{x}$ y la de $\boldsymbol{\xi}$. ¿Qué observa?
e. ¿Qué porcentaje de la variabilidad total se conserva si definimos a $\boldsymbol{\xi}$ únicamente con las primeras dos dimensiones?
f. Una observación de $\mathbf{x}$ resulta $(2\ 2\ 1)^t$, ¿Qué valores toman las componentes principales?
g. Estimar la matriz $\mathsf{Cov}(\mathbf{x}, \boldsymbol{\xi})$.

2.2 Dado un vector aleatorio $\mathbf{x} \in \mathbb{R}^p$, con media $\boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\mathbf{\Sigma}$, llamamos *estandarización multivariante* al resultado de hacer

$$\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}),$$

lo que produce un vector en $\mathbb{R}^p$, llamémoslo $\mathbf{z}$.

a. Calcular $\mathsf{E}(\mathbf{z})$ y $\mathsf{Cov}(\mathbf{z})$. ¿Qué observa?
b. ¿Qué relación tiene la estandarización multivariante a la estandarización por columnas hecha en el ejercicio **1.1**?
c. ¿Qué relación tiene la estandarización multivariante con la transformación necesaria para obtener las componentes principales de $\mathbf{x}$?
d. Simular una nube de puntos de tamaño $n = 1000$ con distribución Normal en $\mathbb{R}^3$ y los parámetros $\boldsymbol{\mu}$ y $\mathbf{\Sigma}$ del ejercicio **2.1**. Aplicar la estandarización multivariante a estos datos y visualizar la nube antes y después de la transformación mediante *scatterplots* de pares. ¿Qué observa? (Nota: Usar los valores poblacionales conocidos de $\boldsymbol{\mu}$ y $\mathbf{\Sigma}$).

2.3 Se recolecta información sobre proyectos inmobiliarios en un vector aleatorio $\mathbf{x}$ de media $\boldsymbol{\mu} \in \mathbb{R}^3$ y varianza $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times 3}$. Las componentes son X_1 = duración media de la hipoteca (años), X_2 = precio (millones de euros) y X_3 = superficie de la cocina (m^2). Se dispone de observaciones de una muestra de $n = 10$ proyectos en el archivo `constructora.txt`.

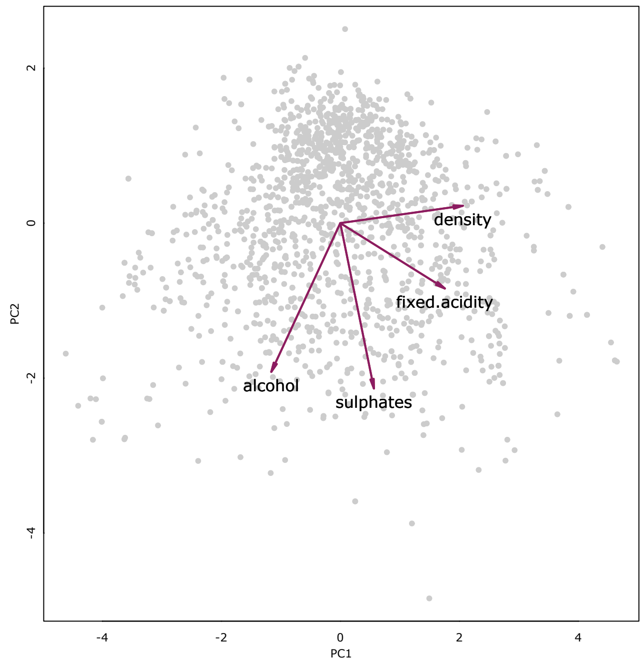
a. Centrar los datos y realizar *scatterplots* de pares.
b. Obtener las componentes principales muestrales a partir de una estimación de $\mathsf{Cov}(\mathbf{x})$. Escribir la transformación lineal necesaria a los datos para obtener los *scores* en dimensión 2.
c. Graficar los *scores* e interpretar la posición de los puntos. ¿Qué variable o variables tienen más peso en cada componente?
d. Estimar la matriz de covarianzas del vector completo de componentes principales y compararla con la estimación disponible de $\boldsymbol{\Sigma}$. ¿Qué observa? ¿Qué proporción de la variabilidad total capturan las primeras dos componentes?
e. Repetir el análisis obteniendo los *scores* a partir de una estimación de $\mathsf{Corr}(\mathbf{x})$ y comentar las diferencias.

2.4 El archivo `paises_mundo.csv` tiene indicadores económicos y sociales de 96 países en algún momento de la década de los 90. Las variables son la tasa de mortalidad infantil cada 1000 nacidos vivos (`mortinf`), producto nacional bruto (`PNB`), producción de electricidad (`prod_elec`), consumo de energía per cápita (`cons_energia`) y emisión de CO2 per cápita (`CO2`). Se tiene la siguiente hoja de ruta para realizar un análisis de los datos:

1 Realizar un *scatterplot* de pares. Comentar sobre la linealidad y de los datos y su ajuste con una distribución Normal.
2 Obtener y graficar los *scores* de PCA en sus dos primeras coordenadas. ¿Cómo se puede interpretar la ubicación de los países?
3 Calcular la proporción de variabilidad total acumulada por las dos primeras coordenadas de los *scores*.
4 Realizar un *heatmap* de la correlación muestral entre los *scores* y los datos. ¿Qué variables son las que inciden más en cada componente principal? ¿Tiene sentido calcular coeficientes de correlación?

a. Se propone recorrer esta ruta partiendo de tres escenarios distintos:
   i. Los datos originales, sin estandarizarlos por columnas.
   ii. Tomar logaritmo natural de los datos y usarlos sin estandarizar por columnas.
   iii. Tomar logaritmo natural a los datos y luego estandarizarlos por columnas.
b. ¿Cómo se podría detectar si hay observaciones atípicas? ¿Afectan sensiblemente al análisis de componentes principales?

2.5 (Construcción de un *biplot* de componentes principales) El archivo `vinos.csv` contiene información sobre 4 variables medidas a 1580 botellas de vino tinto. Reproducir la siguiente representación en dos dimensiones:



Ayuda: Las coordenadas de los puntos corresponden a los *scores* de PCA de los datos estandarizados por columnas, mientras que las direcciones de las flechas corresponden a las coordenadas de las variables. En la librería base de `R` usar las funciones `arrows` y `text`.

Coordenadas discriminantes

En construccion.

Correlación canónica

En construccion.

Projection pursuit

En construccion.

Bonus

En construccion.