Reducción y Visualización de Datos (2023) - Práctica 1

1.1 En una muestra de $n = 26$ países, se observa un vector $\mathbf{x} \in \mathbb{R}^3$ cuyas componentes son % del PBI del país destinado a los rubros de Agricultura (X_1), Industria (X_2) y Energía (X_3). Los datos correspondientes están en el archivo `PBI.csv`.

a. Volcar los datos a una matriz $\mathbf{X}$ cuyas filas correspondan a las observaciones y sus columnas a las variables.
b. Realizar *scatterplots* de a pares e interpretar la relación observada entre las variables.
c. Calcular el promedio $\overline{\mathbf{x}}$ y la matriz $\tilde{\mathbf{X}} = \mathbf{HX}$, donde $\mathbf{H}$ es la matriz de centrado.
d. Calcular la matriz $\mathbf{Q} = \tilde{\mathbf{X}}^t\tilde{\mathbf{X}} = \mathbf{X}^t\mathbf{HX}$ e interpretar la información que contiene. A partir de ella, obtener la función de varianzas y covarianzas $\mathbf{S}$.
e. Se construye una nueva matriz $\mathbf{Z} = \mathbf{HXD}^{-1/2}$, donde $\mathbf{D}^{-1/2} = \operatorname{diag}(s_{11}^{-1/2}, s_{22}^{-1/2}, s_{33}^{-1/2})$. ¿Qué información tiene $\mathbf{Z}$?
f. Calcular la matriz de varianzas y covarianzas de $\mathbf{Z}$.

1.2 Se observará el largo y ancho de ciertas placas rectangulares de acero, variables que forman de un vector aleatorio $\mathbf{x}$ de media $\mathsf{E}(\mathbf{x}) = \boldsymbol{\mu} = (10\ 4)^t$ y covarianza

$$\mathsf{Cov}\,(\mathbf{x}) = \boldsymbol{\Sigma} = \begin{pmatrix} 5 & -1 \\ -1 & 1/2 \end{pmatrix}$$

Sea $\mathbf{y}$ el vector aleatorio que contiene la información del costo (Y_1) y el precio de venta (Y_2) de una placa, dichas componentes se calculan en función de las dimensiones como $Y_1 = 4(X_1 + X_2)$ e $Y_2 = 5X_1 + 2X_2$.

a. Escribir la transformación lineal de $\mathbf{x}$ a $\mathbf{y}$ con una matriz $\mathbf{A} \in \mathbb{R}^{2\times 2}$.
b. Calcular $\mathsf{E}(\mathbf{y})$ y $\mathsf{Cov}(\mathbf{y})$, ¿Son Y_1 e Y_2 variables independientes?
c. En el archivo `placas.txt` se encuentran los datos de ancho y largo observados para $n = 14$ placas. Armar la matriz de datos $\mathbf{X}$ y obtener la matriz de datos $\mathbf{Y}$, con precios y costos, usando un producto matricial que involucre a la matriz $\mathbf{A}$.
d. A partir de expresiones matriciales en las que aparezca $\mathbf{X}$, estimar las matrices $\mathsf{Cov}(\mathbf{x})$, $\mathsf{Corr}(\mathbf{x})$, $\mathsf{Cov}(\mathbf{y})$ y $\mathsf{Corr}(\mathbf{y})$.

1.3 Dada una matriz de datos $\mathbf{X}$ con n filas dadas por los vectores $\mathbf{x}_i^t \in \mathbb{R}^k$ y además un vector $\mathbf{m} \in \mathbb{R}^k$, programar en `R` una función que calcule la siguiente sumatoria:

$$\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^t.$$

Luego, aplicando la función a los datos de los ejercicios anteriores:

a. ¿Con qué vector $\mathbf{m}$ la función arroja el mismo resultado que `cov`?
b. ¿Cómo debe usarse la función para que devuelva el mismo resultado que `cor`?

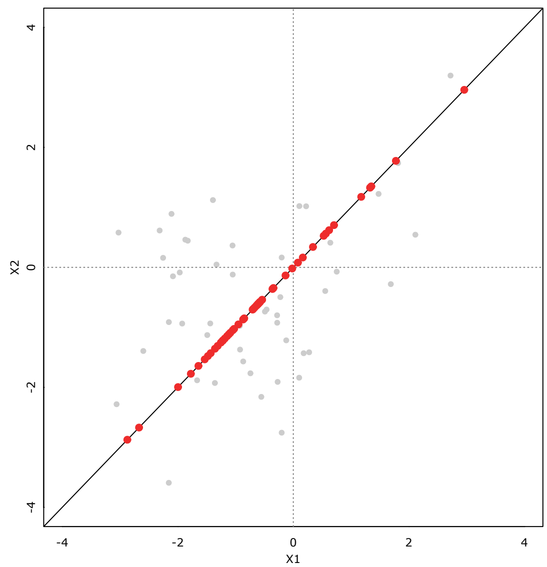
1.4 Dada una matriz de datos $\mathbf{X}$ con n filas dadas por los vectores $\mathbf{x}_i^t \in \mathbb{R}^k$, un vector $\mathbf{m} \in \mathbb{R}^k$ y una matriz $\mathbf{M}$ simétrica, programar en `R` una función que devuelva un vector con los resultados de

$$d_i = \sqrt{(\mathbf{x}_i - \mathbf{m})^t \mathbf{M}^{-1} (\mathbf{x} - \mathbf{m})}$$

con $i \in \{1, 2, ..., n\}$. Luego, aplicando la función a los datos de los ejercicios anteriores:

a. ¿Con qué matriz $\mathbf{M}$ la función arroja las distancias euclídeas al vector $\mathbf{m}$?
b. Analizar qué ocurre si $\mathbf{M}^{\mathbf{-1}} = \mathbf{D}^{-1} = \text{diag}(s_{11}^{-1}, s_{22}^{-2}, ..., s_{kk}^{-1})$.
c. ¿Cómo debe usarse la función para que devuelva las distancias de Mahalanobis de los vectores a la media muestral?

1.5 Reproducir el siguiente gráfico en `R`:



Ayuda: Se trata de n=50 datos que fueron simulados a partir de una distribución Normal bidimensional con la semilla `1234`, con media $(0\ 0)^t$. Se sabe que $\text{Var}(X_1) = \text{Var}(X_2)$ y que $\text{Cov}(X_1, X_2) = {}^1\!/_2\text{Var}(X_1)$. Como primer paso, encontrar la matriz de varianzas y covarianzas $\mathbf{\Sigma}$ que generó los puntos grises y con ella la correspondiente matriz de datos $\mathbf{X}$. Luego, proponer un vector $\mathbf{a}_1 \in \mathbb{R}^2$ que coincida con la dirección de proyección que se ve en el gráfico. ¿Qué condición debe cumplir ese vector para producir la dispersión deseada?

1.6 Sea $\mathbf{x} \in \mathbb{R}^2$ un vector aleatorio con distribución Normal de media $\boldsymbol{\mu} = (0\ 0)^t$ y varianza $\mathbf{\Sigma}$ hallada en el ejercicio anterior.

a. Dado el vector $\mathbf{a}_1$ que se tomó como dirección de proyección en el gráfico, y sea Y_1 la variable aleatoria dada por $\mathbf{a}_1^t\mathbf{x}$, calcular $\mathsf{E}(Y_1)$, $\mathsf{Var}(Y_1)$ y escribir su función de densidad.
b. Mediante un histograma y un gráfico de densidad por núcleos, usar las observaciones del ejercicio anterior para estudiar la bondad de ajuste de la distribución de Y_1.
c. Dada una nueva dirección $\mathbf{a}_2 = {}^1\!/_{\sqrt{2}}(-1\ 1)^t$ y la variable aleatoria $Y_2 = \mathbf{a}_2^t\mathbf{x}$, escribir la función de densidad del vector aleatorio $\mathbf{y} = (Y_1\ Y_2)^t$.
d. Agregar la nueva dirección de proyección al gráfico del ejercicio anterior.
e. Obtener las coordenadas $\mathbf{Y}$ correspondientes a las observaciones de $\mathbf{y}$ con los datos del ejercicio anterior y graficarlas en un par de ejes con escala fija. Estimar las matrices $\mathsf{Cov}(\mathbf{y})$ y $\mathsf{Corr}(\mathbf{y})$. ¿Qué se observa?
f. Repetir los tres incisos anteriores tomando una nueva dirección $\mathbf{a}_3 = {}^1\!/_{\sqrt{5}}(1\ 2)^t$ en lugar de $\mathbf{a}_2$ y analizar las diferencias.

1.7 En un estudio sobre la contaminación del agua, se tomaron muestras de $n = 54$ lagos en Estados Unidos. En cada observación se registró:

- X_1 = alcalinidad (mg/l de carbonato de calcio).
- X_2 = clorofila e(mg/l).
- X_3 = concentración media de mercurio (en ppm) del tejido muscular de un grupo de peces tomados al azar.

En el archivo `mercurio.csv` se encuentran los datos, en los que figura una columna adicional con una variable `etiqueta` en la que se marcaron algunas observaciones.

a. Explorar la visualización en tres dimensiones de los datos usando la librería `plotly`. Marcar las observaciones etiquetadas con distinto color.
b. Se proponen transformaciones no lineales, definiendo nuevas variables a partir de las originales como:
$$W_1 = \sqrt{X_1}, \quad W_2 = \sqrt{X_2} \ \text{ y } \ W_3 = \log(X_3).$$
Aplicar estas transformaciones a los datos e interpretar qué efecto producen en la visualización del inciso anterior. ¿Es posible deducir por qué fueron marcadas algunas observaciones?
c. Comparar la matriz de correlaciones de los datos originales con la de los datos transformados.
d. Se propone realizar una transformación lineal al vector $\mathbf{w} = (W_1 \ W_2 \ W_3)^t$, cuyas direcciones de proyección están en las columnas de
$$\mathbf{A} = \begin{pmatrix} -3/4 & 2/3 \\ -2/3 & -4/5 \\ 1/5 & -1/20 \end{pmatrix}$$
¿Qué dimensión tienen los vectores resultantes? Aplicar esta transformación a los datos y graficar usando la misma escala en el eje horizontal y el vertical. ¿Qué observa en relación a los puntos marcados? ¿Qué observa en relación a la covarianza muestral de los nuevos puntos?
e. Realizar un ranking con las distancias de Mahalanobis de las observaciones de $\mathbf{w}$ a su media muestral. Identificar la observación que se encuentra a mayor distancia. ¿Qué ocurre si se usan distancias euclídeas?

Bonus

1.8 Si $\mathbf{x}$ es un vector aleatorio y $\mathbf{a}$ un vector no aleatorio, probar que $\mathsf{Var}(\mathbf{x} - \mathbf{a}) = \mathsf{Var}(\mathbf{x})$.

1.9 Sea $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$ una muestra aleatoria de vectores en $\mathbb{R}^d$ tales que $\mathsf{E}(\mathbf{x}_i) = \boldsymbol{\mu}$ y $\mathsf{Var}(\mathbf{x}_i) = \boldsymbol{\Sigma}$, probar que

(a) $\mathsf{E}(\overline{\mathbf{x}}) = \boldsymbol{\mu}$ y $\mathsf{Var}(\overline{\mathbf{x}}) = \boldsymbol{\Sigma}/n$.

(b) $\mathsf{E}(\mathbf{Q}) = (n-1)\boldsymbol{\Sigma}$ donde $\mathbf{Q} = \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^t$.

1.10 Supongamos que $\mathbf{x}$ es un vector aleatorio en $\mathbb{R}^d$ tal que $\mathsf{E}(\mathbf{x}) = \boldsymbol{\mu}$ y $\mathsf{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$. Sea $\mathbf{A} \in \mathbb{R}^{d\times d}$ una matriz fija simétrica, probar que $\mathsf{E}(\mathbf{x}^t\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^t\mathbf{A}\boldsymbol{\mu}$. ¿Cuánto vale esta esperanza si $\boldsymbol{\mu}$ es el origen y $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$?

1.11 En el ejercicio 1.7, visualizar en un gráfico tridimensional el resultado de aplicar a las observaciones de $\mathbf{w}$ la transformación lineal dada por la matriz $\mathbf{B} = \mathbf{A}\mathbf{A}^t$. ¿Cuál es el rango de $\mathbf{B}$? Interpretar el resultado.