Práctica 2: Análisis de Componentes Principales

Realiza cada una de las tareas que se presentan a continuación para practicar los contenidos de la asignatura.

Debes presentar un informe con los resultados e interpretaciones del análisis realizado en la TAREA 3 e incluye en el informe todo el código de R que has utilizado para llevar a cabo este trabajo.

Tareas

Tarea 1.

Utiliza el siguiente código para generar el banco de datos que necesitas para realizar la primera tarea:

```
set.seed(seed=21121)
x1=rnorm(100)
set.seed(seed=21122)
x2=rnorm(100,sd=2)
set.seed(seed=21123)
x3=rnorm(100,sd=0.5)

dat.t1<-data.frame(x1,x2,x3)
```

- Realiza un gráfico de dispersión para cada par de variables (por ejemplo con la función *pairs*) y observa el comportamiento de las tres varibles del banco de datos. Realiza un Análisis de Componentes Principales de este banco de datos que acabas de generar (*dat.t1*), utilizando directamente su matriz de varianzas-covarianzas, y responde a estas preguntas concretas:
  - 1. ¿Qué proporción de varianza explica la primera componente principal? ¿y la segunda? ¿y la tercera?
  - 2. Interpreta cómo se forma (con qué variables originales) la primera componente principal, y la segunda y la tercera.
  - 3. ¿Qué te parecen los resultados obtenidos? ¿Por qué crees que las componentes principales se han construido de esta manera?

Tarea 2

El archivo adjunto de R *EPF2.Rdata* contiene el objeto **EPF**, que contiene información sobre la **Encuesta de Presupuestos Familiares de España**. En dicho banco de datos los individuos se corresponden con cada una de las provincias españolas (las ciudades de Ceuta y Melilla aparecen juntas). Por otro lado el banco de datos consta de 9 variables correspondientes a los gastos de las familias encuestadas en:

- **X1**: Alimentación
- **X2**: Vestido y calzado
- **X3**: Vivienda
- **X4**: Mobiliario doméstico
- **X5**: Gasto sanitario

- **X6**: Transporte
- **X7**: Enseñanza y cultura
- **X8**: Turismo y ocio
- **X9**: Otros gastos

Las variables están expresadas en escala logarítmica para favorecer su simetría. Lleva a cabo un Análisis de Componentes Principales de este banco de datos intentando responder a las siguientes cuestiones:

- 1. Explica si consideras más adecuado hacer un Análisis de Componentes Principales basado en la matriz de covarianza de los datos o basado en su matriz de correlaciones.
- 2. Explora cuántas componentes principales necesitas para explicar el 80% de la varianza original del banco de datos.
- 3. Explora e intenta interpretar las dos primeras componentes.
- 4. Realiza un gráfico en el que se muestren los individuos (provincias) según las dos primeras componentes principales.
- 5. Caracteriza el comportamiento de las provincias de **Madrid** y **Barcelona** en función de estas componentes principales.

Tarea 3 (TAREA A ENTREGAR)

Recuperamos el banco de datos de la práctica 1 que encontrarás corregido en el fichero **datos_prac1_ok.RData**.

Recuerda que las dos primeras variables simplemente identifican cada provincia con su código y su nombre.

- a. Calcula las desviaciones típicas de las variables cuantitativas del banco de datos. Responde, a la vista del significado de las variables y del resultado anterior, si consideras que se debería realizar el análisis de componentes principales con la matriz de varianzas-covarianzas o con la de correlaciones.
- b. Realiza un Análisis de Componentes Principales sobre este banco de datos en el modo en el que hayas justificado en el primer apartado y responde las siguientes preguntas:
  - 1. ¿Qué porcentaje de varianza del banco de datos original explica la primera componente principal? ¿Y la segunda?
  - 2. ¿Con cuántas componentes principales nos deberíamos quedar si queremos mantener al menos el 90% de la varianza del banco de datos orignal?
  - 3. Intenta interpretar de forma breve el significado de la primera componente principal.
  - 4. Intenta interpretar de forma breve el significado de la segunda componente principal.
  - 5. Representa todas las provincias en un gráfico según las dos primeras componentes principales y comenta el resultado, resaltando las que tengan algún comportamiento que te llame la atención (por ejemplo, las más extremas en alguna de las dos primeras componentes). *Puedes utilizar las Comunidades Autónomas a las que pertencen (la relación está disponible entre los ficheros de la práctica 1) para marcar en diferentes colores, por ejemplo, las provincias de cada comunidad, y así comprobar si aparecen cercanas en el gráfico.*