

Análisis Estadístico Multivariado | UTDT

Problem Set 2

Alejandra Clemente

Fiona Franco Churruarín

Segundo Trimestre 2023

Ejercicio 1

El archivo *ine_dta* es una base de datos que contiene los gastos promedio, en euros, de los hogares españoles, por grandes rubros y comunidad autónoma, correspondientes a los relevamientos de la encuesta de presupuestos familiares del año 2005 realizada por el Instituto Nacional de Estadísticas de España (INE). La descripción de las variables y de la base puede consultarse en las etiquetas asociadas a cada una.

- (a) Realice un análisis descriptivo de los datos.
- (b) Realice un análisis de componentes principales.
- (c) Efectúe la selección de los componentes principales de acuerdo con los siguientes modos:
 - i. Búsqueda del “codo”.
 - ii. Búsqueda por umbral de varianza a explicar, considerando un 80%.
 - iii. Búsqueda por tope mínimo al valor de los eigenvalores, considerando la varianza media.
- (d) Interprete los componentes seleccionados. Para ello, puede emplear el archivo *renta.csv*.

Aclaración: el archivo *mmta_csv* contiene información referida al ingreso promedio por hogar, por comunidad; relevado por el INE para el año 2005.

Ejercicio 2

En relación al ejercicio anterior, realice el *biplot* asociado a las dos dimensiones principales, considerando un parámetro $c = 0.5$.

Ejercicio 3

Teniendo en cuenta la misma base de datos del primer punto, efectúe los *biplots* correspondientes a las dos dimensiones principales y a la primera y tercera, considerando en ambos casos, un parámetro $c = 0$. Comente brevemente el resultado en relación a lo obtenido mediante el análisis de componentes principales efectuado en el primer ejercicio.

Ejercicio 4

La base *wb.dta* contiene información relevada por el Banco Mundial en 77 países durante el año 2007 (los detalles de las variables relevadas se pueden observar en las etiquetas asignadas a las mismas). A partir de estos datos, se pide:

- (a) Realice un análisis normado de componentes principales a partir de los datos.
- (b) ¿Cuántos componentes sugiere extraer?
- (c) ¿Cuál es el porcentaje de variabilidad total explicado por las componentes seleccionadas?
- (d) ¿Qué interpretación sugiere de las componentes, con arreglo a las correlaciones existentes con las variables originales? Para este punto puede ayudarse con la estructura de análisis presente en el libro *P2.xls*.
- (e) Clasifique a los países de acuerdo con las dos componentes principales.
- (f) Realice un biplot considerando las dos dimensiones principales de análisis para un parámetro $c = 0.5$. ¿Qué puede decir sobre la posición de los Estados Unidos y China, y su influencia en el análisis, a la luz de los resultados obtenidos?

Ejercicio 5

La base *hspendusa2009.dta* contiene información relevada durante los años 2008-2009 por el Instituto de Estadísticas de los Estados Unidos, correspondiente a los gastos medios de los hogares por capítulos y por área metropolitana (los detalles de las variables se pueden observar en el archivo *usa.xls*). Efectúe un análisis de componentes principales.