



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

## TAREA 4

### EJERCICIOS

1. Dispones de estos dos conjuntos de datos:

**elec2001:** Contiene datos municipales sobre las elecciones autonómicas celebradas en 2001, para los municipios de la CAPV. El fichero se llama `elec2001.dge` y puede leerse directamente desde R mediante un `dget`.

**coches:** Para un conjunto de coches en venta en el mercado español, las valoraciones procedentes de una encuesta realizada a sus lectores por la revista AUTOPISTA. Disponible como una *dataframe* en `coches.frame` (lee con `dget`).

todos ellos en el lugar habitual.

Con cada uno de ellos, haz lo siguiente:

- Estima las matrices de covarianzas y de correlación. Decide —justificando tu decisión— cuál de ellas emplear para hacer un análisis en componentes principales.
- Lleva a cabo un análisis en componentes principales. Decide cuantas componentes principales quieres retener para describir los datos. Puedes emplear tu criterio subjetivo, o ayudarte con un contraste de esfericidad; pero en este caso, casi invariablemente te verás arrastrado a considerar más componentes principales de las que deseas.
- Representa los puntos de la muestra en el plano generado por las dos primeras componentes principales. Rotula los ejes y los puntos. Esta representación —y, si parece oportuno, la homóloga en planos generados por otras parejas de componentes— será habitualmente un paso de rutina en todo análisis multivariante: cosas como agrupamientos, puntos extraños y pautas en los datos suelen emerger en estas representaciones.

- d) Haz —hasta donde puedas— una interpretación de tus resultados. ¿Es “simple” el espacio de valoración de los automóviles (hay simplemente coches “buenos” y “malos”, o hay ineludiblemente que considerar varias dimensiones en la evaluación de un coche)? ¿Con los datos en `elec2001.dge`? ¿Ves alguna componente principal interpretable?
2. Para el caso de los datos electorales en `elec2001`, además de emplear los datos brutos o normalizados por columnas, podrías plantearte el normalizarlos (o reducirlos a porcentajes) *por filas*. Discute las consecuencias que esto tendría.
  3. Sobre el resultado del análisis en componentes principales de las elecciones de 2.001 (el que creas que tiene más sentido, de entre los realizados en los apartados anteriores), realiza un biplot.
  4. Los datos sobre elecciones en `elec2001.dge` pueden verse como una tabla de contingencia: el tipo de datos para los que es útil el Análisis de Correspondencias. Lleva a cabo tal análisis, y representa en un plano de modo simultáneo filas (municipios) y columnas (partidos). Comenta lo que observes.

### AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. Como de costumbre, al margen de tus apuntes de clase puedes mirar las secciones relevantes de cualquiera de los muchos manuales a tu disposición en Biblioteca (clasificados en 519.237, segunda planta). En particular puedes consultar [3], [12], [11], [10], [1] o [7].  
Monografías útiles sobre análisis factorial son [2], [9] y [8].  
Sobre Análisis de Correspondencias simple, en clase seguimos sobre todo a [3] y a [10]. Libros que te interesarán si quieres profundizar son [6], [5] y [4].  
Todo está en Biblioteca.
2. La *dataframe* `e2001` tiene columnas que no son directamente de interés (las tres primeras). Tiene también muchos NA, valores perdidos, que puedes poner a cero así:
 

```
e2001[is.na(e2001)] <- 0
```
3. La función `princomp` te será de utilidad; puedes ver la documentación *on line* de R; forma parte del paquete `stats` que se carga por omisión. Pero es aconsejable que al menos una vez hagas componentes principales “a pelo”, a partir de primeros principios, estimando la matriz de covarianzas (o correlación) y empleando sobre ella la función `eigen`.
4. Si empleas como base de partida matrices de correlación, has de emplear luego los coeficientes sobre variables tipificadas. Asegúrate de entender por qué.
5. Te puede interesar hacer mapas de componentes principales con las escalas de ordenadas y abscisas iguales. La función `eqsplot` en la librería MASS te será de utilidad. (Para utilizarla, has de tener en tu sesión un `library(MASS)` previo.) Mira también [13], Sec. 11.1, pág. 330.

6. El modo “standard” de producir un mapa de componentes principales es hacer un gráfico “mudo” (opción `type="n"`) y a continuación situar las etiquetas correspondientes a los puntos con un `text`.
7. La función `princomp` te proporcionará directamente los valores de las componentes principales evaluadas para cada punto muestral (en el componente `scores`). También un estadillo con la varianza explicada por cada componente principal y la matriz  $A$  de vectores propios de la matriz de covarianza (o correlación).
8. Todos los datos en la *dataframe* `elec2001` proceden del Instituto Vasco de Estadística, EUSTAT, <http://www.eustat.es>. Siéntete libre de reemplazarlos por otros de tu interés de la misma u otras fuentes.
9. Observa: el Análisis de Correspondencias presupone que los datos son brutos, en forma de tabla de contingencia; sobre tablas de contingencia se definió la distancia  $\chi^2$ . La función `ac.R` ya los convierte a frecuencias relativas. No debes reducir los datos a porcentajes ni hacer transformaciones similares.
10. Observa que la función `factanal` incluye un contraste formal de diagonalidad de la matriz de correlación residual. Si el número de factores escogido es “adecuado”, la matriz “residual”  $\hat{\Sigma} - \hat{A}\hat{A}'$  debería ser aproximadamente diagonal. Contrastando esta hipótesis, tenemos un contraste indirecto de que el número de factores es apropiado (si se rechazara la hipótesis, tendríamos motivo para incrementar el número de factores comunes).
11. Tienes una función `biplot` en R que admite como input el resultado de un análisis en componentes principales.

En cualquier caso, podrías escribir tú una función que represente biplots con facilidad, tomando como modelo `ac.R`.

## Referencias

- [1] A. Basilevsky. *Statistical Factor Analysis and Related Methods*. Wiley, 1992.
- [2] R. Cattell. *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*. Plenum Press, 1978.
- [3] C.M. Cuadras. *Métodos de Análisis Multivariante*. Eunibar, Barcelona, 1981.
- [4] B. Escofier and J. Pages. *Análisis Factoriales Simples y Múltiples. Objetivos, Métodos e Interpretación*. Servicio Editorial de la UPV/EHU, Bilbao, 1984.
- [5] M. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, New York, 1984.
- [6] M. Greenacre. *Correspondence Analysis in Practice*. Academic Press, New York, 1993.
- [7] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.
- [8] I.T. Jolliffe. *Principal Components Analysis*. Springer-Verlag, New York, 1986.
- [9] S.A. Mulaik. *The Foundations of Factor Analysis*. McGraw-Hill, 1972.
- [10] D. Peña. *Análisis de Datos Multivariantes*. McGraw-Hill, 2002.
- [11] A.C. Rencher. *Methods of Multivariate Analysis*. Wiley, 1995.
- [12] G.A.F. Seber. *Multivariate Observations*. Wiley, New York, 1984.
- [13] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer-Verlag, New York, third edition, 1999.