



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

TAREA 5

EJERCICIOS

1. Dispones de estos dos conjuntos de datos:

Bizkaia: Contiene datos municipales sobre varias elecciones en el Territorio Histórico de Vizcaya, así como datos socioeconómicos relacionados con el perfil lingüístico, laboral y demográfico de cada municipio. Es una *dataframe* en formato de exportación. El fichero se llama `Bizkaia.dge` y puede leerse (con un `dget`) tanto desde R como desde S-PLUS.

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* (apareció en el número de la semana 3/9 de Junio de 1997). Disponible como una *dataframe* en `coches.frame`.

todos ellos en el lugar habitual.

Para `coches` y una de las elecciones de las que se proporcionan datos en `Bizkaia` haz lo siguiente:

- a) Estima las matrices de covarianzas y de correlación. Decide —justificando tu decisión— cuál de ellas emplear para hacer un análisis en componentes principales.
 - b) Lleva a cabo un análisis en componentes principales. Decide cuántas 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 desees. ¿Para que conjunto de datos es claramente insostenible la hipótesis de normalidad multivariante?
 - c) 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 `Bizkaia.dge` que has escogido? ¿Ves alguna componente principal interpretable?
Observa que una interpretación mucho más fina y diáfana sólo será posible mediante técnicas de rotación; las estudiaremos en los temas dedicados a Análisis Factorial.
2. Para el caso de los datos electorales en `Bizkaia`, 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.

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], [7], [6] o [4]. Monografías dedicadas a análisis factorial y componentes principales son [1] y [5].
2. La función `princomp` te será de utilidad; puedes ver [2] y la documentación *on line* de S-PLUS. 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`.
Si trabajas sobre R, dispones de las funciones `princomp` (practicamente idéntica a la de S-PLUS) y `prcomp`.
3. Si empleas como base de partida matrices de correlación, has de emplear luego los coeficientes sobre variables tipificadas. Asegúrate de entender por qué.
4. Te puede interesar hacer mapas de componentes principales con las escalas de ordenadas y abscisas iguales. La función `eqscplot` en la librería MASS te será de utilidad. Mira también [8], Sec. 11.1, pág. 330.
5. 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`.
6. Si empleas la función `princomp`, te proporcionará directamente los valores de las componentes principales evaluadas para cada punto muestral. (argumento `scores=T`).
7. Todos los datos en la *dataframe* Bizkaia 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. En particular, tienes en la dirección señalada datos homólogos para otros Territorios Históricos y para el conjunto de la CAPV. Algunos están disponibles en hojas Excel; otros sólo en tablas HTML.

Para éstos últimos, una manera de capturarlos es guardar la tabla HTML (“Save frame as...” o cosa similar en tu navegador) e importarlos como HTML desde Excel (si trabajas en Windows) o programas similares (como la hoja de cálculo de Star Office o Gnumeric, disponibles sobre Linux y varios UNIXes.) Has de eliminar todos los logotipos, gráficos, etc., empalmar los trozos que te interesen en una tabla y leerla luego desde R o S-PLUS.

Cuadro 1: Nomenclatura de las variables en la *dataframe* Bizkaia

Columna	Variable	Descripción
1	Municipio	Nombre del municipio
2	POBTOT96	Población total según censo 1.996
3	ACTIVOS96	Población activa OIT según censo 1.996
4	OCU96	Población ocupada OIT según censo 1.996
5	PARADOS96	Población parada OIT según censo 1.996
6	PARPRIM96	Parados OIT en búsqueda primer empleo
7	PARPREV96	Parados OIT que han trabajado previamente
8	INACT96	Inactivos OIT
9	CONTAP96	Población contada aparte según censo 1.996
10	EUSKMAT96	Población con el euskera como lengua materna según censo 1.996
11	CASTMAT96	Población con el castellano como lengua materna según censo 1.996
12	AMBASMAT96	Población con ambas lenguas como maternas según censo 1.996
13	OTRAMAT96	Población con otra lengua materna según censo 1.996
14	EUSKCASA96	Euskera como lengua hablada en casa (censo 1996)
15	CASTCASA96	Castellano como lengua hablada en casa (censo 1996)
16	DOSCASA96	Castellano y euskera ambas habladas en casa (censo 1996)
17	OTRACASA96	Otra lengua hablada en casa (censo 1996)
18	PAROINEM98	Paro registrado INEM 1998

continúa pág. siguiente

Columna	Variable	Descripción
19	PAROINEM99	Paro registrado INEM 1999
20	MAYORES1096	Población mayor de 10 años según censo 1.996
21	ANALFA96	Población analfabeta según censo 1.996
22	SINEST96	Población sin estudios según censo 1.996
23	PRIMARIOS96	Población con estudios primarios según censo 1.996
24	PROFES96	Población con estudios profesionales según censo 1.996
25	SECUNDA96	Población con estudios secundarios según censo 1.996
26	MEDSUP96	Población con estudios medios según censo 1.996
27	SUPER96	Población con estudios superiores
28	INMIG97	Población inmigrante recibida 1997
29	EMIG97	Población emigrante que ha abandonado el municipio 1997
30	MIGINTDES07	Migrantes internos procedentes de otros lugares 1997
31	MIGINTORI97	Emigrantes internos destinados a otros lugares 1997
32	MIGINTRA97	Migrante intramunicipales 1997
33	MAYORES0596	Población mayor de 5 años censo 1996
34	VASCOF96	Vascófonos censo 1996
35	BILINORIG96	Bilingües de origen censo 1996
36	NEOVASCOF96	Neo-vascófonos censo 1996
37	NEOPAR96	Neo-vascófonos parciales censo 1996
38	PARCAST96	Neo-vascófonos parcialmente castellanizados censo 1996
39	TOTCAST96	Neo-vascófonos totalmente castellanizados censo 1996
40	CASTELLA96	Castellanófonos censo 1996
41	OCUPAD96	Ocupados censo 1996
42	AGRICUL96	Ocupados agricultura censo 1996
43	INDUST96	Ocupados industria censo 1996
44	CONSTR96	Ocupados construcción censo 1996
45	SERVIC96	Ocupados servicios censo 1996
46	VARONES96	Varones censo 1996
47	MUJERES96	Mujeres censo 1996
48	SEDENT96	Población sedentaria censo 1996
49	SEDVAR96	Población sedentaria varones censo 1996
50	SEDMUJ96	Población sedentaria mujeres censo 1996
51	MIGINT96	Migrantes internos 1996
52	MIVAR96	Migrantes internos varones 1996
53	MIMUJ96	Migrantes internos mujeres 1996
54	INM96	Inmigrantes 1996
55	INMVAR96	Inmigrantes varones 1996
56	INMMUJ96	Inmigrantes mujeres 1996
57	CENSO95	Censo electoral en las elecciones de 1.995
58	VOTOS95	Número de votantes en las elecciones de 1.995
59	NULOS95	Número de votos nulos en las elecciones de 1.995
60	BLANCOS95	Número de votos blancos en las elecciones de 1.995
61	PNV95	Número de votos obtenidos por el PNV/EAJ en las elecciones de 1.995
62	PSE95	Número de votos obtenidos por el PSE en las elecciones de 1.995
63	HB95	Número de votos obtenidos por HB en las elecciones de 1.995
64	EA95	Número de votos obtenidos por EA en las elecciones de 1.995
65	PP95	Número de votos obtenidos por el PP en las elecciones de 1.995
66	UA95	Número de votos obtenidos por UA en las elecciones de 1.995
67	IU95	Número de votos obtenidos por IU en las elecciones de 1.995
68	INDCOA95	Número de votos obtenidos por independientes y otros en las elecciones de 1.995
69	CENSO98	Censo electoral en las elecciones de 1.998
70	VOTANTES98	Número de votantes en las elecciones de 1.998
71	ABST98	Número de abstenciones en las elecciones de 1.998
72	BLANNULO98	Número de votos blancos y nulos en las elecciones de 1.998
73	VALIDOS98	Número de votos válidos en las elecciones de 1.998

continúa pág. siguiente

Columna	Variable	Descripción
74	PNV98	Número de votos obtenidos por el PNV/EAJ en las elecciones de 1.998
75	PSE98	Número de votos obtenidos por el PSE en las elecciones de 1.998
76	EH98	Número de votos obtenidos por EH en las elecciones de 1.998
77	EA98	Número de votos obtenidos por EA en las elecciones de 1.998
78	PP98	Número de votos obtenidos por el PP en las elecciones de 1.998
79	UA98	Número de votos obtenidos por UA en las elecciones de 1.998
80	IU98	Número de votos obtenidos por IU en las elecciones de 1.998
81	PH98	Número de votos obtenidos por el PH en las elecciones de 1.998
82	PLN98	Número de votos obtenidos por el PNV/EAJ en las elecciones de 1.998
83	EB98	Número de votos obtenidos por el xxx en las elecciones de 1.998
84	POBEUS96	Población euskaldún mayor de 5 años
85	VASEUS96	Vascófonos activos en euskera
86	VASACTDOS96	Vascófonos activos en ambas lenguas
87	VASPASEUS96	Vascófonos pasivos en euskera
88	BIOREUS96	Bilingües de origen activos en euskera
89	BIACTDOS96	Bilingües de origen activos en ambas lenguas
90	BIPASEUS96	Bilingües de origen pasivos en euskera
91	NEOACTEUS96	Neovascófonos activos en euskera
92	NEOACTDOS96	Neovascófonos activos en ambas lenguas
93	NEOPASEUS96	Neovascófonos pasivos en euskera
94	PARACTEUS96	Neovascófonos parciales activos en euskera
95	PARACTDOS96	Neovascófonos parciales activos en ambas lenguas
96	PARPASEUS96	Neovascófonos parciales pasivos en euskera
97	CASACTEUS96	Castellanizados activos en euskera
98	CASACTDOS96	Castellanizados activos en ambas lenguas
99	CASPASEUS96	Castellanizados pasivos en euskera

Referencias

- [1] A. Basilevsky. *Statistical Factor Analysis and Related Methods*. Wiley, 1992.
- [2] R.A. Becker, J.M. Chambers, and A.R. Wilks. *The New S Language. A Programming Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole, Pacific Grove, California, 1988.
- [3] C.M. Cuadras. *Métodos de Análisis Multivariante*. Eunibar, Barcelona, 1981.
- [4] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.
- [5] I.T. Jolliffe. *Principal Components Analysis*. Springer-Verlag, New York, 1986.
- [6] A.C. Rencher. *Methods of Multivariate Analysis*. Wiley, 1995.
- [7] G.A.F. Seber. *Multivariate Observations*. Wiley, New York, 1984.
- [8] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer-Verlag, New York, third edition, 1999.