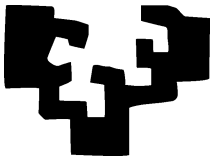

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 6

EJERCICIOS

Los ejercicios siguientes tienen por objeto suministrar práctica en la aplicación de reescalado multidimensional, completando el abanico de técnicas de visualización y reducción de dimensionalidad vistas hasta ahora en el curso.

1. Considera los datos ya manejados en `coches.frame`. Cada fila puede verse como un vector en R^{25} (la última columna es redundante, al ser combinación lineal de las precedentes). Si consideras como medida de distancia entre los diferentes modelos de coche la euclídea ordinaria,

$$d^2(i,j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_E^2 \tag{1}$$

   es claro que en R^{25} puedes representar el modelo de coche i-ésimo mediante un vector $\boldsymbol{y}_i \in R^{25}$ de modo que $d^2(i,j) = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_E^2$; ¡bastaría que tomases $\boldsymbol{y}_i = \boldsymbol{x}_i$! Pero ¿es posible representar cada coche mediante un punto en $\boldsymbol{y}_i \in R^p$ con $p < 25$ de manera que

$$d^2(i,j) = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_E^2$$

   exactamente? En general, no, salvo que los vectores $\boldsymbol{x}_i$ originales estén realmente en un subespacio de dimensión p.

   Si estamos dispuestos a tolerar que la igualdad anterior se verifique sólo de modo aproximado (es decir, $d^2(i,j) \approx ||\boldsymbol{y}_i - \boldsymbol{y}_j||_E^2$, podremos no obstante obtener una representación en un espacio de dimensión, en general, mucho menor que la de R^{25}.

   a) Define $d^2(i,j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_E^2$ y emplea la función `cmdscale` para hacer un análisis multidimensional métrico, fijando $p = 2$. Obtendrás una solución idéntica —salvo giros, reflexiones y traslaciones— a la que obtenías mediante componentes principales.

   b) Define cualquier otra medida de disimilaridad entre modelos de coche y haz el análisis de reescalado multidimensional métrico correspondiente ($p = 2$). Entre las ayudas tienes algunas sugerencias.

   c) ¿Qué ocurriría si como distancia entre modelos emplearas la diferencia entre las medias de sus calificaciones (la última columna, que hemos desechado antes)?

2. En el lugar habitual dispones de la *dataframe* `morse.dge` con los datos del experimento de Rothkopf. Consiste en lo siguiente: a varios sujetos se les hace oír en rápida sucesión una pareja de caracteres en código Morse, preguntándoles si se trata del mismo o de dos caracteres diferentes. El tanto por ciento de respuestas "son iguales" se toma como medida bruta de la similaridad entre los dos caracteres.

   La diagonal principal contiene números próximos a 100, porque parejas de caracteres iguales son reconocidos como tales la mayoría de las veces. Fuera de la diagonal principal hay números en general menores, pero también algunos elevados porque hay caracteres que suenan de modo muy similar al oído no muy entrenado.

Observa que, a diferencia de lo que sucedía con el ejemplo anterior, aquí el punto de partida son distancias: no hay una configuración de puntos previa de la que dichas distancias deriven. Este es la situación típica del reescalado multidimensional. Lleva a cabo sendos análisis, métrico y no métrico, e interpreta los resultados.

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

- **Bibliografía.** Los libros [5], [7], [8], [4], [6], y [2] son manuales de Análisis Multivariante en general. Libros específicos sobre Reescalado Multidimensional son [1] y [3].

  Todo está en Biblioteca.

- **Sugerencias sobre disimilaridades a emplear.** ¿Qué distancias o disimilaridades cabe utilizar en el primer ejercicio? Hay multitud de posibilidades: puedes emplear una distancia ponderada dando más peso a las características que creas más importantes. Puedes considera sólo algunas característica (mecánica, estética, etc.). Puedes emplear una distancia como la de Mahalanobis (que surgirá abundantemente y de modo natural en lo que resta de curso), definida como:

$$d^2(i,j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)\,' C^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{2}$$

  en que C es la matriz de covarianzas muestral.

  Puedes emplear disimilaridades como $d(i,j) = \mathrm{máx}_k(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})$; puedes hacer cualquier cosa que se te ocurra y tenga sentido.

- **Funciones en R.** Aunque podrías programar reescalado multidimensional métrico a partir de primeros principios (y sería bueno que lo hicieras), tienes funciones como `cmdscale`. Otras funciones que hacen diferentes variedades de reescalado: `isoMDS` (biblioteca `MASS`) realiza reescalado no métrico, al igual que la función `sammon` (en la misma biblioteca).

Referencias

[1] I. Borg and P. Groenen. *Modern Multidimensional Scaling. Theory and Applications*. Springer-Verlag, New York, 1997.

[2] C. Chatfield and A.J. Collins. *Introduction to Multivariate Analysis*. Chapman & Hall, London, 1980.

[3] T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman and Hall, 1994.

[4] C.M. Cuadras. *Métodos de Análisis Multivariante*. Eunibar, Barcelona, 1981.

[5] W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications*. Wiley, New York, 1984.

[6] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.

[7] D. Peña. *Análisis de Datos Multivariantes*. McGraw-Hill, 2002.

[8] G.A.F. Seber. *Multivariate Observations*. Wiley, New York, 1984.