eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

TAREA 4

EJERCICIOS

1. En clase ilustramos la construcción de una dataframe en el formato N filas (observaciones) $\times p$ columnas (variables), a partir de ficheros obtenidos de fuentes públicas. Utiliza la primera parte de ése código (en el fichero `prostituto.R`) para generar una dataframe de resultados electorales (y variables añadidas sobre paro, población autóctona y tamaño de municipio).

2. Utilizando sólo las variables de resultados electorales,

   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 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.

   c) Representa las observaciones 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) 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.

   e) Explica qué análisis te parece más descriptivo (datos brutos, normalizados por columnas, en porcentajes por filas). ¿Puedes aventurar una interpretación para alguna o algunas componentes principales?

3. Considera la afirmación: "Los partidos PSE-PSOE e IU obtienen mejores resultados en municipios grandes".

   a) ¿Cuál es el coeficiente de correlación estimado entre el porcentaje de votos sumados de ambos partidos y el tamaño del municipio?

   b) ¿Cuál es el coeficiente de correlación *parcial* estimado entre el porcentaje de votos sumados de ambos partidos y el tamaño del municipio, cuando controlas el efecto de la población autóctona y el nivel de paro?

   c) ¿Cuál es tu conclusión?

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], [10], [8], [7], [1] o [5].

   Todo está en Biblioteca.

2. Lss funciones `princomp` ó `prcomp` (hay pequeñas diferencias entre ellas) te será de utilidad; puedes ver la documentacion *on line* de R. 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`.

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. (Para utilizarla, has de tener en tu sesión un `library(MASS)` previo.) Mira también [11], 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. 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 compsonente principal y la matriz A de vectores propios de la matriz de covarianza (o correlación).

7. En clase se mostró un modo de servirse de R para representar sobre un mapa variables de interés, que pueden ser las variables originales, valores de componentes principales, o cualquier otra cosa. Puedes imitar el código en la segunda parte del fichero `prostituto.R` para hacer lo que te convenga.

   Adicionalmente, encontrarás información sobre modos alternativos de hacer mapas en [9], [12], [2] o (mucho más avanzado, en realidad un libro sobre Geoestadística con R) [4]. En el LEC dispones además de QGIS, un sistema de información geográfica (documentado en `www.qgis.org`) y GRASS (documentado en `http://grass.osgeo.org`) y en [6].

Referencias

[1] A. Basilevsky. *Statistical Factor Analysis and Related Methods*. Wiley, 1992.

[2] Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer Verlag, 2008.

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

[4] Tomislav Hengl. *A Practical Guide to Geostatistical Mapping*. Lulu, 2009.

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

[6] M. Neteler and H. Mitasova. *Open Source GIS: A GRASS GIS Approach*. Springer Verlag, 2007.

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

[8] A. C. Rencher. *Methods of Multivariate Analysis*. Wiley, 1995.

[9] Deepayan Sarkar. *Lattice. Multivariate Data Visualization with R*. Springer, 2008.

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

[11] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, third edition, 1999.

[12] H. Wickham. *ggplot2 : elegant graphics for data analysis*. Springer-Verlag, 2009.