Prueba de evaluación continua 2
Estadística Multivariante

Alex Sánchez y Francesc Carmona

2 de enero de 2021
Fecha límite de entrega: 17-01-2021

Ejercicio 1 (65 pt.)

En el trabajo de Hunt et al.[1] se estudió la capacidad reproductiva de cinco especies de aves marinas en dos colonias en el sureste del mar de Bering. Además, el apéndice de este estudio resume las colonias y los tamaños de las poblaciones de otros trabajos. El archivo `seabirds.csv` recoge los datos (número de pájaros) de 23 especies en 9 colonias en el área del norte polar y subpolar.
El principal interés de este ejercicio es representar las colonias de diversas formas y estudiar posibles conglomerados.

(a) Calcular las frecuencias relativas, las frecuencias relativas marginales y la matriz de perfiles. El resultado debería ser la tabla 12.6 del libro de Krebs[2] y que reproducimos al final de este documento.

(b) Calcular la matriz de distancias ji-cuadrado entre los perfiles de las columnas y su inercia total.

(c) Con la matriz de distancias ji-cuadrado entre los perfiles realizar un escalado multidimensional. Dibujar las coordenadas principales para las columnas.

(d) Realizar un análisis de correspondencias y calcular las inercias principales (en %) y la inercia total con los valores propios.

Dibujar una representación simétrica del CA. A pesar de la confusión de nombres, ¿cuales son las especies que caracterizan a la colonia SI (Skomer Island, Irish Sea)?

(e) Dada la gran cantidad de ceros en la tabla 12.6, en el libro de Krebs[2] se sugiere la utilización de la distancia de Canberra entre las columnas de la tabla 12.6. La distancia de Canberra no tiene una única definición y, además, ha cambiado a lo largo de la historia. Una posible definición entre dos vectores $\mathbf{p} = (p_1, p_2, \ldots, p_k)'$ y $\mathbf{q} = (q_1, q_2, \ldots, q_k)'$ de la misma longitud es

$$d_C(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Cuando el denominador es cero, el cociente es NaN, y el sumando se elimina.

Comprobar que esta definición no sirve para calcular la matriz de similaridades de la tabla 12.7 del libro de Krebs[2] y que se reproduce al final de este documento.

Una modificación de la distancia anterior es considerar la distancia

$$d_C(\mathbf{p}, \mathbf{q}) = \frac{1}{k} \sum_{i=1}^{k} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Igual que antes, cuando el denominador es cero, el sumando se elimina.

Comprobar que con esta definición se puede obtener la tabla 12.7[1].

Finalmente, se puede comprobar que ésta tampoco es la definición que utiliza **R** para calcular la distancia de Canberra. Tras una ardua investigación, se comprueba que la definición de **R** es

$$d_C(\mathbf{p}, \mathbf{q}) = \frac{k}{k - n_z} \sum_{i=1}^{k} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

donde n_z es el número de denominadores cero.

Comprobar que ésta es la definición de la distancia de Canberra según **R**.

(f) Realizar un MDS con la distancia de Canberra de **R**. Comprobar que se trata de una distancia euclídea. Dibujar el mapa.

Comparar el resultado con el obtenido con la distancia ji-cuadrado. Utilizar la función `procrustes()` del paquete `vegan`.

(g) Realizar un análisis de conglomerados jerárquico con el método de Ward[2] de la distancia de Canberra según **R**. Dibujar el dendograma resultante.

Dibujar también un *heatmap* de este análisis con la función `heatmap()`. Para ello, hay que elegir bien los parámetros `distfun=` y `hclustfun=` y una escala de colores. ¿Para qué sirve el heatmap?

Nota: No nos interesa reordenar las variables (especies) y tampoco un dendograma sobre ellas.

(h) El siguiente paso es estudiar por algún criterio el número óptimo de conglomerados para el análisis jerárquico. Con la distancia de Canberra según **R** en particular, lo más sencillo es utilizar el criterio de las siluetas.

(i) Estudiar con la misma distancia el número óptimo de conglomerados con el método PAM.

(j) De los apartados anteriores se deduce que hay un número razonable de conglomerados, aunque no sea óptimo. Dibujar el dendograma del apartado (g) con esa partición.

Ejercicio 2 (35 pt.)

Un estudio contiene dos medidas de los anillos de crecimiento en la escala del salmón de Alaska y de Canadá. Los datos se pueden obtener[3] en el libro de Johnson et al.[3] y se adjuntan en el archivo `salmon.txt`.

(a) Realizar una estadística descriptiva univariante y multivariante según el factor `Origin`. Añadir algunos gráficos ilustrativos, en particular el de dispersión.

(b) Realizar un análisis discriminante lineal. La función `lda()` del paquete MASS puede servir.

(c) Clasificar una observación con un *Freshwater* de 120 y un valor de *Marine* de 380.

(d) Comparar las matrices de covarianzas de las dos poblaciones con el test de la razón de verosimilitudes.

También se puede aplicar el test M de Box.

Ambos son muy sensibles a la no normalidad de los datos y tienden a rechazar la igualdad de covarianzas.

[1] La tabla 12.7 contiene dos o tres erratas.
[2] Se puede utilizar la función `hclust()` o la función `agnes()`.
[3] Estos datos también se pueden hallar en el *data.frame* `salmon` del paquete `rrcov`.

(e) En el caso de poblaciones normales con diferentes matrices de covarianzas se clasificará cada observación en el grupo con máxima probabilidad a posteriori, pero entonces las funciones discriminantes no son lineales, ya que tienen un término de segundo grado.

    Realizar un análisis discriminante cuadrático. La función `qda()` del paquete MASS nos ayudará.

(f) Calcular el número de parámetros que hay que estimar en la discriminación lineal y en la cuadrática.

(g) Calcular los errores de clasificación con ambas reglas utilizando validación cruzada.

    Si son similares, nos quedaremos con el análisis lineal que además es más robusto y de mejor interpretación.

Referencias

[1] George L. Hunt, Zoe A. Eppley and David C. Schneider, Reproductive Performance of Seabirds: The Importance of Population and Colony Size, The Auk 103: 306-317, April 1986.

[2] Krebs, C.J., *Ecological Methodology*, 3rd ed. (in prep) Chapters revised to date (14 March 2014).

[3] Johnson, R.A. and Wichern, D. W., `Applied Multivariate Statistical Analysis` (Prentice Hall, International Editions, 2002, fifth edition)

TABLE 12.6 RELATIVE ABUNDANCES (PROPORTIONS) OF 23 SPECIES OF SEABIRDS ON 9 COLONIES IN NORTHERN POLAR AND SUBPOLAR AREAS[a]

	Cape Hay, Bylot Island	Prince Leopold Island, eastern Canada	Coburg Island, eastern Canada	Norton Sound, Bering Sea	Cape Lisburne, Chukchi Sea	Cape Thompson, Chukchi Sea	Skomer Island, Irish Sea	St. Paul Island, Bering Sea	St. George Island, Bering Sea
Northern fulmar	0	.3422	0	0	0	0	.0007	.0028	.0278
Glaucous-winged gull	.0005	.0011	.0004	.0051	.0004	.0007	0	0	0
Black-legged kittiwake	.1249	.1600	.1577	.1402	.1972	.0634	.0151	.1221	.0286
Red-legged kittiwake	0	0	0	0	0	0	0	.0087	.0873
Thick-billed murre	.8740	.4746	.8413	.0074	.2367	.5592	0	.4334	.5955
Common murre	0	0	0	.7765	.5522	.3728	.0160	.1537	.0754
Black guillemot	.0006	.02200.	.0005	0	.0013	.00001	0	0	0
Pigeon guillemot	0	0	0	0	0	.00003	0	0	0
Horned puffin	0	0	0	.0592	.0114	.0036	0	.0173	.0111
Tufted puffin	0	0	0	.0008	.0002	0	0	.0039	.0024
Atlantic puffin	0	0	0	0	0	0	.0482	0	0
Pelagic cormorant	0	0	0	.0096	.0006	.0001	.0001	0	0
Red-faced cormorant	0	0	0	0	0	0	0	.0099	.0020
Shag	0	0	0	0	0	0	.0001	0	0
Parakeet auklet	0	0	0	.0012	0	0	0	.1340	.0595
Crested auklet	0	0	0	0	0	0	0	.0236	.0111
Least auklet	0	0	0	0	0	0	0	.0906	.0992
Razorbill	0	0	0	0	0	0	.0130	0	0
Manx shearwater	0	0	0	0	0	0	.7838	0	0
Storm petrel	0	0	0	0	0	0	.0389	0	0
Herring gull	0	0	0	0	0	0	.0229	0	0
Great black-backed gull	0	0	0	0	0	0	.0001	0	0
Lesser black backed gull	0	0	0	0	0	0	.0603	0	0

[a] Data from Hunt et al. (1986).

TABLE 12.7 MATRIX OF SIMILARITY COEFFICIENTS FOR THE SEABIRD DATA IN TABLE 12.6. ISLANDS ARE PRESENTED IN SAME ORDER AS IN TABLE 12.6[a]

	CH	PLI	CI	NS	CL	CT	SI	SPI	SGI
CH	1.0	0.88	0.99	0.66	0.77	0.75	0.36	0.51	0.49
PLI		1.0	0.88	0.62	0.70	0.71	0.36	0.51	0.49
CI			1.0	0.66	0.78	0.75	0.36	0.50	0.48
NS				1.0	0.73	0.64	0.28	0.53	0.50
CL					1.0	0.76	0.29	0.51	0.49
CT						1.0	0.34	0.46	0.45
SI							1.0	0.19	0.20
SPI								1.0	0.80
SGI									1.0

[a] The complement of the Canberra metric (1.0 - *C)* is used as the index of similarity. Note that the matrix is symmetrical about the diagonal.