1 Análisis Multivariado - Práctica 4

1.1 Correlación canónica

1. (a) Consideremos un vector $\mathbf{z} \in \mathbb{R}^d$ tal que $\mathcal{E}[\mathbf{z}] = \mathbf{0}$ y $\mathcal{D}[\mathbf{z}] = \Sigma$. Para medir la asociación lineal entre la primera componente z_1 y las demás, $\mathbf{y} = (z_2, \ldots, z_d)'$, se define el coeficiente de correlación múltiple al cuadrado $\rho^2_{1(23\ldots d)}$ como la mayor correlación (al cuadrado) entre z_1 y cualquier combinación lineal de $\mathbf{y}$. Es decir,

$$\rho^2_{1(23\ldots d)} = \max_{\beta} \frac{[\operatorname{cov}(z_1, \beta'\mathbf{y})]^2}{\operatorname{var}(z_1)\operatorname{var}(\beta'\mathbf{y})}. \tag{1}$$

Probar que

$$\rho^2_{1(23\ldots d)} = \frac{\sigma_{12}\Sigma_{22}^{-1}\sigma_{21}}{\sigma_{11}}$$

con los parámetros que vienen de la partición

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Además, probar que el máximo (1) se realiza en $\beta = \Sigma_{22}^{-1}\sigma_{21}$.

(b) Supongamos ahora que queremos predecir z_1 mediante una combinación lineal de $\mathbf{y}$. Entonces se busca

$$\beta^* = \arg\min_{\beta} E\left[(z_1 - \beta'\mathbf{y})^2\right].$$

Probar que nuevamente se obtiene $\beta^* = \Sigma_{22}^{-1}\sigma_{21}$.

2. Sea $\mathbf{z} = (\mathbf{x}', \mathbf{y}')'$ con $\mathbf{x} \in \mathbb{R}^{d_1}$ e $\mathbf{y} \in \mathbb{R}^{d_2}$. Si $\Sigma = \mathcal{D}[\mathbf{z}]$ es definida positiva, probar que la primera correlación canónica ρ_1 es estrictamente menor que 1.

   Sugerencia: Usar A3.2 de Seber.

3. Probar que las correlaciones canónicas son invariantes por transformaciones afines. Es decir, las correlaciones canónicas entre $\mathbf{x}$ e $\mathbf{y}$ son las mismas que entre $A\mathbf{x}$ y $B\mathbf{y}$ si A y B son matrices inversibles.

4. Dadas dos variables canónicas u_i y v_j con $i \neq j$, demostrar que $\operatorname{cov}[u_i, v_j] = 0$.

   Sugerencia: Primero mostrar que $\mathbf{a}_i = \Sigma_{11}^{-1}\Sigma_{12}\mathbf{b}_i$.

5. Usando multiplicadores de Lagrange, probar que ρ_1^2 es el máximo de $(\alpha'\Sigma_{12}\beta)^2$ sujeto a $\alpha'\Sigma_{11}\alpha = 1$ y $\beta'\Sigma_{22}\beta = 1$.

   Sugerencia: Usar A8.1 de Seber.

6. Sea $\mathbf{z} = (\mathbf{x}', \mathbf{y}')'$ con $\mathbf{x} \in \mathbb{R}^2$ e $\mathbf{y} \in \mathbb{R}^2$ y supongamos que

$$\mathcal{D}[\mathbf{z}] = \sigma^2 \begin{pmatrix} 1 & a & b & b \\ a & 1 & b & b \\ b & b & 1 & c \\ b & b & c & 1 \end{pmatrix}$$

donde a, b y c tienen módulo menor que 1. Encontrar la primera correlación canónica y las correspondientes variables canónicas.

7. Sea $\mathbf{z} = (\mathbf{x}', \mathbf{y}')'$ con $\mathbf{x} \in \mathbb{R}^{d_1}$ e $\mathbf{y} \in \mathbb{R}^{d_2}$. Supongamos que $\mathcal{E}[\mathbf{z}] = \mathbf{0}$ y consideramos $\Sigma = \mathcal{D}[\mathbf{z}]$ con la partición
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \mathcal{E}[\mathbf{x}\mathbf{x}'] & \mathcal{E}[\mathbf{x}\mathbf{y}'] \\ \mathcal{E}[\mathbf{y}\mathbf{x}'] & \mathcal{E}[\mathbf{y}\mathbf{y}'] \end{pmatrix}.$$
Se quiere predecir $\mathbf{y}$ mediante k combinaciones lineales no correlacionadas de $\mathbf{x}$. Es decir, si definimos $\mathbf{u} = A\mathbf{x}$ con A de $k \times d_1$ tal que $\mathcal{D}[\mathbf{u}] = \Sigma_{\mathbf{u}} = A\Sigma_{11}A' = I_k$, queremos encontrar un predictor lineal de $\mathbf{y}$ de la forma $B\mathbf{u}$ (donde B será de $d_2 \times k$). El criterio será elegir A y B que minimicen $E\left[\|\mathbf{y} - B\mathbf{u}\|^2\right]$.

   (a) Probar que, fijada la matriz A, se tiene que
   $$E\left[\|\mathbf{y} - B\mathbf{u}\|^2\right] \geq E\left[\|\mathbf{y} - B^*\mathbf{u}\|^2\right]$$
   para $B^* = \Sigma_{21}A'(A\Sigma_{11}A')^{-1}$. Mostrar además que
   $$E\left[\|\mathbf{y} - B^*\mathbf{u}\|^2\right] = \operatorname{tr}(\Sigma_{22}) - \operatorname{tr}\left(\Sigma_{21}A'(A\Sigma_{11}A')^{-1}A\Sigma_{12}\right),$$
   por lo que el problema se reduce a encontrar
   $$A^* = \arg\max_{A\Sigma_{11}A' = I_k} \operatorname{tr}\left(\Sigma_{21}A'(A\Sigma_{11}A')^{-1}A\Sigma_{12}\right). \tag{2}$$

   (b) Mostrar que la matriz A^* que cumple (2) tiene como filas a los autovectores correspondientes a los primeros k autovalores de $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{21}$ y tales que son ortogonales con la distancia dada por Σ_{11}.
   Sugerencia: Usar el lema que sigue: Sea $Q \in \mathbb{R}^{d\times d}$, $Q \geq 0$ y $k < d$, $C \in \mathbb{R}^{k\times d}$ tal que $CC' = I_k$ entonces
   $$\sum_{i=1}^{k} \lambda_i(CQC') \leq \sum_{i=1}^{k} \lambda_i(Q)$$
   donde $\lambda_1(Q) \geq \cdots \geq \lambda_d(Q)$ son los autovalores de Q y $\lambda_1(CQC') \geq \cdots \geq \lambda_k(CQC')$ los autovalores de CQC'.

8. En un estudiode pobreza, crimen y disuasión, Parker y Smith reportaron ciertos resúmenes de estadísticas criminales en varios estados de EEUU para los años 1970 y 1973. Una parte de la matriz de correlación muestral aparece más abajo. Las variables son:

   X_1 = homicidios no primarios cometidos durante 1970
   X_2 = homicidios primarios (homicidios que involucran parientes o conocidos) cometidos durante 1970
   Y_1 = severidad de los castigos (mediana de los meses cumplidos en prisión) en 1970
   Y_2 = certeza del castigo (número de admisiones a prisión dividido por el número de homicidios) en 1970

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 1 & 0.615 & -0.111 & -0.266 \\ 0.615 & 1 & -0.195 & -0.085 \\ \hline -0.111 & -0.195 & 1 & -0.269 \\ -0.266 & -0.085 & -0.269 & 1 \end{array}\right]$$

(a) Hallar las correlaciones canónicas muestrales.

(b) Determinar el primer par canónico $\hat{u}_1$ y $\hat{v}_1$ e interpretar estas cantidades.

9. En un estudio realizado por Waugh (1942) basado en $n = 138$ muestras de trigo de variedad Canadian Hard Red Spring y de la harina hecha a partir de él, se midieron las siguientes variables (5 son mediciones en forma estandarizada para el trigo X_i y 4 para la harina Y_i):

$$\begin{aligned} X_1 &= \text{estructura del núcleo} \\ X_2 &= \text{peso} \\ X_3 &= \text{núcleos dañados} \\ X_4 &= \text{material extraño} \\ X_5 &= \text{proteínas} \end{aligned}$$

$$\begin{aligned} Y_1 &= \text{trigo por barril de harina} \\ Y_2 &= \text{ceniza presente en la harina} \\ Y_3 &= \text{proteínas en la harina} \\ Y_4 &= \text{índice de calidad del gluten} \end{aligned}$$

La matriz de correlación fue:

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right] = \left[\begin{array}{ccccc|cccc} 1 & & & & & & & & \\ 0.754 & 1 & & & & & & & \\ -0.69 & -0.712 & 1 & & & & & & \\ -0.446 & -0.515 & 0.323 & 1 & & & & & \\ 0.692 & 0.412 & -0.444 & -0.334 & 1 & & & & \\ \hline -0.605 & -0.722 & 0.737 & 0.527 & -0.383 & 1 & & & \\ -0.479 & -0.419 & 0.361 & 0.461 & -0.505 & 0.251 & 1 & & \\ -0.780 & 0.542 & -0.546 & -0.393 & 0.737 & -0.490 & -0.434 & 1 & \\ -0.152 & -0.102 & 0.172 & -0.019 & -0.148 & 0.250 & -0.079 & -0.163 & 1 \end{array}\right]$$

(a) Hallar las variables canónicas muestrales correspondientes a las correlaciones canónicas.

(b) Interpretar las variables canónicas muestrales $\hat{u}_1$ y $\hat{v}_1$. Representan en algún sentido la calidad del trigo y la harina, respectivamente?

(c) Qué proporción de la varianza muestral total de las variables X queda explicada por $\hat{u}_1$? Qué proporción de la varianza muestral total de las variables Y queda explicada por $\hat{v}_1$?