Técnicas de Análisis Multivariado 2 - Práctica 2

1. Correlación canónica

1. *a*) Consideremos un vector $\mathbf{z} \in \mathbb{R}^d$ tal que $\mathbb{E}[\mathbf{z}] = \mathbf{0}$ y $\text{VAR}[\mathbf{z}] = \boldsymbol{\Sigma}$. 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,

$$\rho^2_{1(23\ldots d)} = \max_{\boldsymbol{\beta}} \frac{\left[\text{COV}\left(z_1, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}\right)\right]^2}{\text{VAR}\left(z_1\right)\text{VAR}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}\right)}\,. \tag{1}$$

Probar que

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

con los parámetros que vienen de la partición

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

Además, probar que el máximo de (1) se realiza en $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$.

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

$$\boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta}} E\left[\left(z_1 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}\right)^2\right].$$

Probar que nuevamente se obtiene $\boldsymbol{\beta}^* = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$.

2. Sea $\mathbf{z} = (\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}})^{\mathrm{T}}$ con $\mathbf{x} \in \mathbb{R}^{d_1}$ e $\mathbf{y} \in \mathbb{R}^{d_2}$. Si $\boldsymbol{\Sigma} = \text{VAR}[\mathbf{z}]$ 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 $\mathbf{Ax}$ y $\mathbf{By}$ si $\mathbf{A}$ y $\mathbf{B}$ son matrices no singulares.

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

SUGERENCIA: Primero mostrar que $\boldsymbol{\alpha}_i = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_i/\rho_i$.

5. Usando los multiplicadores de Lagrange, probar que ρ_1^2 es el máximo de la función $\left(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}\right)^2$ sujeto a las restricciones $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = 1$ y $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{22}\boldsymbol{\beta} = 1$.

SUGERENCIA: Usar A8.1 de Seber.

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

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

donde $|a| \leq 1$, $|b| \leq 1$ y $|c| \leq 1$. Encontrar la primera correlación canónica y las correspondientes variables canónicas.

7. Sea $\mathbf{z} = (\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}})^{\mathrm{T}}$ con $\mathbf{x} \in \mathbb{R}^{d_1}$ e $\mathbf{y} \in \mathbb{R}^{d_2}$. Supongamos que $\mathbb{E}[\mathbf{z}] = \mathbf{0}$ y llamemos $\mathbf{\Sigma} = \textsc{Var}[\mathbf{z}]$ con la partición

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} \mathbb{E}[\mathbf{x}\,\mathbf{x}^{\mathrm{T}}] & \mathbb{E}[\mathbf{x}\,\mathbf{y}^{\mathrm{T}}] \\ \mathbb{E}[\mathbf{y}\,\mathbf{x}^{\mathrm{T}}] & \mathbb{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] \end{pmatrix}.$$

Se quiere predecir $\mathbf{y}$ mediante k combinaciones lineales no correlacionadas de $\mathbf{x}$. Es decir, si definimos $\mathbf{u} = \mathbf{A}\,\mathbf{x}$ con $\mathbf{A} \in \mathbb{R}^{k \times d_1}$ tal que $\textsc{Var}[\mathbf{u}] = \mathbf{\Sigma}_{\mathbf{u}} = \mathbf{A}\mathbf{\Sigma}_{11}\mathbf{A}^{\mathrm{T}} = \mathbf{I}_k$, queremos encontrar un predictor lineal de $\mathbf{y}$ de la forma $\mathbf{B}\mathbf{u}$ (donde $\mathbf{B} \in \mathbb{R}^{d_2 \times k}$).

El criterio será elegir $\mathbf{A}$ y $\mathbf{B}$ que minimicen $\mathbb{E}\left[\|\mathbf{y} - \mathbf{B}\mathbf{u}\|^2\right]$.

a) Probar que, fijada la matriz $\mathbf{A}$, se tiene que

$$\mathbb{E}\left[\|\mathbf{y} - \mathbf{B}\mathbf{u}\|^2\right] \geq \mathbb{E}\left[\|\mathbf{y} - \mathbf{B}^*\mathbf{u}\|^2\right]$$

donde

$$\mathbf{B}^* = \mathbf{\Sigma}_{21}\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{\Sigma}_{11}\mathbf{A}^{\mathrm{T}}\right)^{-1}.$$

Mostrar además que

$$\mathbb{E}\left[\|\mathbf{y} - \mathbf{B}^*\mathbf{u}\|^2\right] = \textsc{tr}\left(\mathbf{\Sigma}_{22}\right) - \textsc{tr}\left(\mathbf{\Sigma}_{21}\,\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{\Sigma}_{11}\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{A}\mathbf{\Sigma}_{12}\right),$$

por lo que el problema se reduce a encontrar

$$\mathbf{A}^* = \underset{\mathbf{A}\mathbf{\Sigma}_{11}\mathbf{A}^{\mathrm{T}} = \mathbf{I}_k}{\operatorname{argmax}} \textsc{tr}\left(\mathbf{\Sigma}_{21}\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{\Sigma}_{11}\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{A}\mathbf{\Sigma}_{12}\right). \tag{2}$$

b) Mostrar que la matriz $\mathbf{A}^*$ que cumple (2) tiene como filas a los autovectores correspondientes a los primeros k autovalores de $\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{21}$ y tales que son ortogonales con la distancia dada por $\mathbf{\Sigma}_{11}$, donde entendemos que $\mathbf{u}$ y $\mathbf{v}$ son ortogonales con la distancia dada por $\mathbf{\Sigma}_{11}$ si $\mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}_{11}\mathbf{v} = 0$.

Sugerencia: Usar el lema que sigue: Sea $\mathbf{Q} \in \mathbb{R}^{d \times d}$, $\mathbf{Q} \geq 0$, $k < d$ y $\mathbf{C} \in \mathbb{R}^{k \times d}$ tal que $\mathbf{C}\mathbf{C}^{\mathrm{T}} = \mathbf{I}_k$ entonces

$$\sum_{i=1}^{k} \lambda_i\left(\mathbf{C}\mathbf{Q}\mathbf{C}^{\mathrm{T}}\right) \leq \sum_{i=1}^{k} \lambda_i\left(\mathbf{Q}\right)$$

donde $\lambda_1(\mathbf{Q}) \geq \cdots \geq \lambda_d(\mathbf{Q})$ son los autovalores de $\mathbf{Q}$ y $\lambda_1(\mathbf{C}\mathbf{Q}\mathbf{C}^{\mathrm{T}}) \geq \cdots \geq \lambda_k(\mathbf{C}\mathbf{Q}\mathbf{C}^{\mathrm{T}})$ los autovalores de $\mathbf{C}\mathbf{Q}\mathbf{C}^{\mathrm{T}}$.

8. En un estudio de 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

$$\mathbf{R} = \left[\begin{array}{c|c} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \hline \mathbf{R}_{21} & \mathbf{R}_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 1 & 0{,}615 & -0{,}111 & -0{,}266 \\ 0{,}615 & 1 & -0{,}195 & -0{,}085 \\ \hline -0{,}111 & -0{,}195 & 1 & -0{,}269 \\ -0{,}266 & -0{,}085 & -0{,}269 & 1 \end{array}\right]$$

a) Hallar las correlaciones canónicas muestrales.

b) Determinar el primer par canónico $\widehat{u}_1$ y $\widehat{v}_1$ 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):

X_1 = estructura del núcleo
X_2 = peso
X_3 = núcleos dañados
X_4 = material extraño
X_5 = proteínas

Y_1 = trigo por barril de harina
Y_2 = ceniza presente en la harina
Y_3 = proteínas en la harina
Y_4 = índice de calidad del glúten

Si particionamos a la matriz de correlación como

$$\mathbf{R} = \left[\begin{array}{c|c} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \hline \mathbf{R}_{21} & \mathbf{R}_{22} \end{array}\right]$$

entonces resulta:

$$\mathbf{R} = \left[\begin{array}{ccccc|cccc} 1 & & & & & & & & \\ 0.754 & 1 & & & & & & & \\ -0.690 & -0.712 & 1 & & & & & & \\ -0.446 & -0.515 & 0.323 & 1 & & & & & \\ 0.692 & 0.412 & -0.444 & -0.334 & 1 & & & & \\ \hline -0.605 & -0.722 & 0.737 & 0.527 & -0.383 & 1 & & & \\ -0.479 & -0.419 & 0.361 & 0.461 & -0.505 & 0.251 & 1 & & \\ 0.780 & 0.542 & -0.546 & -0.393 & 0.737 & -0.490 & -0.434 & 1 & \\ -0.152 & -0.102 & 0.172 & -0.019 & -0.148 & 0.250 & -0.079 & -0.163 & 1 \end{array}\right]$$

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

b) Interpretar las variables canónicas muestrales $\widehat{u}_1$ y $\widehat{v}_1$. 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 $\widehat{u}_1$? Qué proporción de la varianza muestral total de las variables Y queda explicada por $\widehat{v}_1$?