Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 9: 16/09/24 – 21/09/24

Motivación

- Un **agrónomo** puede tomar p mediciones relativas al rendimiento de las plantas (altura, peso seco, número de hojas, etc) en sitios en una región y **al mismo tiempo**, puede registrar q variables relacionadas con las condiciones climáticas (precipitación diaria promedio, humedad, horas de sol, etc) en estos sitios. Esto con el interés de medir la **asociación** entre "rendimiento" y "clima". Claramente, **la investigación completa envuelve** $p + q$ mediciones en n **unidades (plantas)**.

Motivación

- En **psicología**, un investigador suele tomar p mediciones de variables de aptitud y q variables de rendimiento para **una muestra de n estudiantes** con el objetivo de estudiar la **relación** entre "aptitud" y "rendimiento".

Motivación

- En **psicología**, un investigador suele tomar p mediciones de variables de aptitud y q variables de rendimiento para **una muestra de n estudiantes** con el objetivo de estudiar la **relación** entre "aptitud" y "rendimiento".
- En una **empresa**, n **vendedores** pueden ser evaluados por **dos conjuntos de variables**: desempeño en el trabajo (creciemiento en ventas, ventas rentables, nuevas cuentas, etc) y desempeño psicológico (creatividad,matemática, raciocinio, etc) para estudiar por ejemplo la relación "productividad" y "competencia".

Motivación

- En **psicología**, un investigador suele tomar p mediciones de variables de aptitud y q variables de rendimiento para **una muestra de n estudiantes** con el objetivo de estudiar la **relación** entre "aptitud" y "rendimiento".
- En una **empresa**, n **vendedores** pueden ser evaluados por **dos conjuntos de variables**: desempeño en el trabajo (creciemiento en ventas, ventas rentables, nuevas cuentas, etc) y desempeño psicológico (creatividad,matemática, raciocinio, etc) para estudiar por ejemplo la relación "productividad" y "competencia".

En todos los casos, es evidente que **se quiere establecer una relacion** entre las p variables de un conjunto y las q variables de otro conjunto.

Motivación

En el proceso de selección de una empresa se pueden tomar algunas medidas

- iniciales $(X_1, X_2, \ldots, X_p)$, esto es **antes de ingresar** o cuando está aplicando algún cargo.
- medidas de productividad o de rendimiento $(Y_1, Y_2, \ldots, Y_q)$ cuando ya **asumió el cargo**.

El **interés** es determinar si existe una **relación lineal** fuerte o debil entre las puntuaciones que los candidatos obtuvieron durante el proceso de selección y las puntuación de su rendimiento o productividad en el cargo cuando fueron contratados.

Motivación

Así, se tienen dos conjuntos de variables

$$\begin{aligned} \text{Conjunto 1} &: \{X_1, X_2, \ldots, X_p\} \\ \text{Conjunto 2} &: \{Y_1, Y_2, \ldots, Y_q\} \end{aligned}$$

y el interés es analizar las **relaciones lineales entre éstos dos conjuntos de variables**.

En particular, si se tienen **dos variables** en cada conjunto, esto es

$$\begin{aligned} \text{Conjunto 1} &: \{X_1, X_2\} \\ \text{Conjunto 2} &: \{Y_1, Y_2\} \end{aligned}$$

se desea analizar las **relaciones lineales entre los dos conjuntos de variables**.

Motivación

Seguramente usted pensó en una **matriz de correlación** para las **cuatro** variables, pues ésta contiene toda la información sobre las **asociaciones lineales entre pares de variables** en los dos conjuntos.

Suponga que se tiene la siguiente matriz de correlación

$$\begin{array}{c} \\ Y_1 \\ Y_2 \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{cccc} \quad Y_1 \quad & \quad Y_2 \quad & \quad X_1 \quad & \quad X_2 \quad \end{array} \\ \begin{pmatrix} 1 & -0.370 & 0.221 & \mathbf{0.445} \\ -0.370 & 1 & 0.316 & 0.168 \\ 0.221 & 0.316 & 1 & -0.176 \\ \mathbf{0.445} & 0.168 & -0.176 & 1 \end{pmatrix} \end{array}$$

¿Qué puede decir con respecto a la **relación lineal entre los dos conjuntos** de datos?

Motivación

En este caso, **la mayoría de las correlaciones lineales son bajas**, de hecho, **la más alta (0.445)** se observa entre X_2 e Y_1; aunque ésta es **moderada**. Así que, aún para conjuntos de **apenas 2 variables**, queda un poco **dificil evaluar la correlación entre ellos** a partir de esta matriz. Imagine cuando los conjuntos están conformados por mas de dos variables!

Motivación

En general, intentar extraer de la matriz de correlación alguna idea de la asociación entre los dos conjuntos de variables no es sencillo o casi imposible!. Esto se debe principalmente a que las correlaciones entre los dos conjuntos:

- pueden **no tener un patrón consistente**,
- deben **ajustarse** de alguna manera para las **correlaciones dentro del conjunto**.

¿Cómo **cuantificar la asociación lineal** entre los conjuntos de variables

$$\mathbf{X} = \{X_1, X_2, \ldots, X_p\}$$

e

$$\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_q\}?$$

Introducción

Hotelling (1935) propuso hacer una combinación lineal con las variables del primer conjunto (U), la cuál llamó variable canónica. Así mismo, hacer una combinación lineal con las variables del segundo conjunto (V), la cuál también llamó variable canónica. Entonces, **la correlación entre** las variables canónicas U y V ($\rho_{U,V}$) se llama **correlación canónica** y mide el grado de **asociación lineal existente entre dos conjuntos** de variables.

Introducción

- El **objetivo principal** en la correlación canónica es estudiar las **relaciones lineales entre dos conjuntos de variables** (no de variable a variable).
- La **idea básica** de la correlación canónica es resumir la información de cada conjunto de variables en combinaciones lineales de las variables de cada conjunto(variables canónicas), cuyos **coeficientes son seleccionados a partir del criterio de maximización de la correlación entre los conjuntos de variables**.

Pontificia Universidad JAVERIANA Bogotá

Introducción

Observaciones:

- *El análisis de correlación canónica* ***generaliza al análisis de correlación entre dos variables****. Aquí la diferencia es que se tienen dos conjuntos de variables y el objetivo es entender cómo éstos conjuntos se relacionan de forma lineal.*
- *Así como en el ACP, el análisis de correlación canónica busca* ***reducir el (los) conjunto(s) de datos****.*
- *Los* ***resultados*** *de este tipo de análisis suelen ser* ***difíciles de interpretar****, en consecuencia, esta técnica se utiliza menos que otras técnicas multivariadas.*

Objetivos de la Correlación Canónica

Principales objetivos de la correlación canónica:

- ✓ Compreender y cuantificar la asociación lineal entre dos conjuntos de variables.
- ✓ Resumir la información de cada conjunto de variables en las variables canónicas, para evaluar la correlación canónica entre ellas.

¿Cuándo usar análisis de correlación canónica?

- ✓ Cuando el interés principal es la correlación entre dos conjuntos de variables que, al principio, parecen estar moderadamente correlacionadas. Por lo tanto, no tiene sentido tratarlos por separado, ignorando la interdependencia entre ellos.
- ✓ Cuando se quiere predecir el resultado de un indicador basándose en un conjunto de variables a partir de otro conjunto de variables.

Análisis de Correlación Canónica

Sea

$$\mathbf{X}_{(p+q)\times 1} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \\ X_{p+1} \\ \vdots \\ X_{p+q} \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \\ \cdots \\ X_{p+1} \\ \vdots \\ X_{p+q} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{p\times 1} \\ \cdots \\ \mathbf{Y}_{q\times 1} \end{bmatrix}$$

Análisis de Correlación Canónica

donde

- $\mathbf{X}_{p\times 1} = [X_1 \quad X_2 \quad \dots \quad X_p]^\top$
- $\mathbf{Y}_{q\times 1} = [X_{p+1} \quad X_{p+2} \quad \dots \quad X_{p+q}]^\top$

Observe que

$$\mathbf{X}_{p\times 1} \qquad \text{e} \qquad \mathbf{Y}_{q\times 1}$$

son **dos conjuntos de variables**.

El interés es medir la relación lineal entre éstos dos conjuntos.

donde

- $\mathbf{X}_{p\times 1} = [X_1 \quad X_2 \quad \ldots \quad X_p]^\top$
- $\mathbf{Y}_{q\times 1} = [X_{p+1} \quad X_{p+2} \quad \ldots \quad X_{p+q}]^\top$

Observe que

$$\mathbf{X}_{p\times 1} \qquad \text{e} \qquad \mathbf{Y}_{q\times 1}$$

son **dos conjuntos de variables**.

El interés es medir la relación lineal entre éstos dos conjuntos.

Considere

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{XX} & \vdots & \mathbf{\Sigma}_{XY} \\ \cdots\cdots & \cdots & \cdots\cdots \\ \mathbf{\Sigma}_{YX} & \vdots & \mathbf{\Sigma}_{YY} \end{bmatrix}_{(p+q)\times(p+q)}$$

Análisis de Correlación Canónica

donde

- $\mathbf{\Sigma}_{XX} = \text{Var}[\mathbf{X}]$ es la matriz $(p \times p)$ de varianzas y covarianzas de las variables del conjunto $\mathbf{X}$.
- $\mathbf{\Sigma}_{YY} = \text{Var}[\mathbf{Y}]$ es la matriz $(q \times q)$ de varianzas y covarianzas de las variables del conjunto $\mathbf{Y}$.
- $\mathbf{\Sigma}_{XY} = \text{Cov}[\mathbf{X}, \mathbf{Y}]$ es la matriz $(p \times q)$ de covarianzas de las variables de los conjuntos $\mathbf{X}$ e $\mathbf{Y}$.
- $\mathbf{\Sigma}_{YX} = \text{Cov}[\mathbf{Y}, \mathbf{X}]$ es la matriz $(q \times p)$ de covarianzas de las variables de los conjuntos $\mathbf{Y}$ e $\mathbf{X}$.

Observación: $\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_{YX}^\top$.

Análisis de Correlación Canónica

De acuerdo con la **propuesta de Hotelling**, para determinar el grado de **relación lineal entre los dos conjuntos de datos $\mathbf{X}$ e $\mathbf{Y}$**, definimos las **variables canónicas:**

- $U = \mathbf{a}^\top \mathbf{X}$, combinaciones lineales de $\mathbf{X}$
- $V = \mathbf{b}^\top \mathbf{Y}$, combinaciones lineales de $\mathbf{Y}$

Para algún par de vectores de coeficientes $\mathbf{a}$ y $\mathbf{b}$.

Análisis de Correlación Canónica

De acuerdo con la **propuesta de Hotelling**, para determinar el grado de **relación lineal entre los dos conjuntos de datos $\mathbf{X}$ e $\mathbf{Y}$**, definimos las **variables canónicas:**

- $U = \mathbf{a}^\top \mathbf{X}$, combinaciones lineales de $\mathbf{X}$
- $V = \mathbf{b}^\top \mathbf{Y}$, combinaciones lineales de $\mathbf{Y}$

Para algún par de vectores de coeficientes $\mathbf{a}$ y $\mathbf{b}$.

Luego, la **correlación canónica** entre U e V

$$\rho_{U,V} = \frac{\text{Cov}[U, V]}{\sqrt{\text{Var}[U]}\sqrt{\text{Var}[V]}}$$

Análisis de Correlación Canónica

mide el **grado de relación lineal** entre los dos conjuntos de variables $\mathbf{X}$ e $\mathbf{Y}$.

Note que:

$$\mathrm{Var}[U] = \mathrm{Var}\left[\mathbf{a}^\top \mathbf{X}\right] = \mathbf{a}^\top \mathrm{Var}[\mathbf{X}]\,\mathbf{a} = \mathbf{a}^\top \boldsymbol{\Sigma}_{XX}\,\mathbf{a}$$

$$\mathrm{Var}[V] = \mathrm{Var}\left[\mathbf{b}^\top \mathbf{Y}\right] = \mathbf{b}^\top \mathrm{Var}[\mathbf{Y}]\,\mathbf{b} = \mathbf{b}^\top \boldsymbol{\Sigma}_{YY}\,\mathbf{b}$$

$$\mathrm{Cov}[U, V] = \mathrm{Cov}\left[\mathbf{a}^\top \mathbf{X}, \mathbf{b}^\top \mathbf{Y}\right] = \mathbf{a}^\top \mathrm{Cov}[\mathbf{X}, \mathbf{Y}]\,\mathbf{b} = \mathbf{a}^\top \boldsymbol{\Sigma}_{XY}\,\mathbf{b}$$

Análisis de Correlación Canónica

Por lo tanto

$$\rho_{U,V} = \frac{\mathbf{a}^\top \boldsymbol{\Sigma}_{XY}\, \mathbf{b}}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_{XX}\, \mathbf{a}}\sqrt{\mathbf{b}^\top \boldsymbol{\Sigma}_{YY}\, \mathbf{b}}}$$

La idea es **determinar** los vectores de coeficientes

$$\mathbf{a} \qquad \text{y} \qquad \mathbf{b}$$

que hagan $\rho_{U,V}$ **sea lo más grande posible.**

Observación: $U = \boldsymbol{a}^\top \boldsymbol{X}$ *y* $V = \boldsymbol{b}^\top \boldsymbol{Y}$ *son todas las combinaciones de* $\boldsymbol{X}$ *e* $\boldsymbol{Y}$*, respectivamente.*

Análisis de Correlación Canónica

Tenga en cuenta que si

$$\text{Var}[U] = 1 = \text{Var}[V]$$

entonces

$$\rho_{U,V} = \frac{\text{Cov}[U, V]}{\sqrt{\text{Var}[U]}\sqrt{\text{Var}[V]}} = \text{Cov}[U, V]$$

Entonces, podemos determinar los vectores de constantes o vectores de coeficientes $\mathbf{a}^\top$ y $\mathbf{b}^\top$ talque

$$\begin{aligned} &\text{Maximizar } \rho_{U,V} \\ &\text{Sujeto a} \quad \text{Var}[U] = 1 = \text{Var}[V] \end{aligned}$$

Análisis de Correlación Canónica

equivalentemente

$$\text{Maximizar } \rho_{U,V} = \frac{\mathbf{a}^\top \boldsymbol{\Sigma}_{XY}\,\mathbf{b}}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_{XX}\,\mathbf{a}}\sqrt{\mathbf{b}^\top \boldsymbol{\Sigma}_{YY}\,\mathbf{b}}}$$

$$\text{Sujeto a } \mathbf{a}^\top \boldsymbol{\Sigma}_{XX}\,\mathbf{a} = 1 = \mathbf{b}^\top \boldsymbol{\Sigma}_{YY}\,\mathbf{b}$$

Recuerde que

- $U = \mathbf{a}^\top \mathbf{X}$, combinaciones lineales de $\mathbf{X}$
- $V = \mathbf{b}^\top \mathbf{Y}$, combinaciones lineales de $\mathbf{Y}$

así que podemos tener p combinaciones de $\mathbf{X}$ y q combinaciones de $\mathbf{Y}$. En consecuencia, tendremos

Análisis de Correlación Canónica

$$(U_1, V_1), (U_2, V_2), \ldots, (U_k, V_k)$$

pares de variables canónicas.

Particularmente, el **primer par de variables canónicas** denotado (U_1, V_1) es:

$$U_1 = \mathbf{a}_1^\top \mathbf{X}$$
$$V_1 = \mathbf{b}_1^\top \mathbf{Y}$$

para el cual se desea

$$\text{Maximizar } \rho_{U_1, V_1} = \rho_1^*$$
$$\text{Sujeto a } \mathrm{Var}[U_1] = 1 = \mathrm{Var}[V_1]$$

Análisis de Correlación Canónica

equivalentemente

$$\begin{aligned} \text{Maximizar } & \mathbf{a}_1^\top \boldsymbol{\Sigma}_{XY}\, \mathbf{b}_1 \\ \text{Sujeto a } & \mathbf{a}_1^\top \boldsymbol{\Sigma}_{XX}\, \mathbf{a}_1 = 1 = \mathbf{b}_1^\top \boldsymbol{\Sigma}_{YY}\, \mathbf{b}_1 \end{aligned}$$

Análogamente, **el segundo par de variables canónicas** denotado (U_2, V_2) es:

$$U_2 = \mathbf{a}_2^\top \mathbf{X}$$
$$V_2 = \mathbf{b}_2^\top \mathbf{Y}$$

para el cual se desea

$$\begin{aligned}
\text{Maximizar } & \rho_{U_2,V_2} = \rho_2^* \\
\text{Sujeto a } & \mathrm{Var}[U_2] = 1 = \mathrm{Var}[V_2] \\
& \mathrm{Cov}[U_1, U_2] = \mathrm{Cov}[U_1, V_2] = \mathrm{Cov}[U_2, V_1] = \mathrm{Cov}[U_2, V_2] = 0
\end{aligned}$$

Esto significa que U_2 y V_2 sean **no correlacionados** con los anteriores pares (U_1, V_1).

Y así en adelante, hasta **el $k-$ésimo par de variables canónicas** (U_k, V_k):

$$\begin{aligned}
\text{Maximizar } & \rho_{U_k,V_k} = \rho_k^* \\
\text{Sujeto a } & \mathrm{Var}[U_k] = 1 = \mathrm{Var}[V_k] \\
& \mathrm{Cov}[U_k, U_l] = \mathrm{Cov}[V_k, V_l] = \mathrm{Cov}[U_k, V_l] = \mathrm{Cov}[U_l, V_k] = 0 \quad k \neq l
\end{aligned}$$

Observaciones:

- *La técnica de correlación canónica garantiza que las variables canónicas de un par no están correlacionadas con variables canónicas de otro par.*
- *El número de variables canónicas* que puede ser obtenido es

$$k = \text{mín}\{p, q\}$$

Análisis de Correlación Canónica

Claramente se tiene un **problema de optimización con restricción**, así que se pueden aplicar los **multiplicadores de Lagrange:**

$$\mathcal{L} = \mathbf{a}_k^\top \mathbf{\Sigma}_{XY}\, \mathbf{b}_k - \frac{\alpha}{2}\left(\mathbf{a}_k^\top \mathbf{\Sigma}_{XX}\, \mathbf{a}_k - 1\right) - \frac{\beta}{2}\left(\mathbf{b}_k^\top \mathbf{\Sigma}_{YY}\, \mathbf{b}_k - 1\right)$$

donde α y β son los multiplicadores de Lagrange.

Para encontrar para encontrar $\mathbf{a}_k$ y $\mathbf{b}_k$ para $k = \min\{p, q\}$, debemos resolver el sistema:

$$\begin{cases} \frac{d}{d\,\mathbf{a}_k}\mathcal{L} & = 0 \\ \\ \frac{d}{d\,\mathbf{b}_k}\mathcal{L} & = 0 \end{cases}$$

el cual queda dado por

$$\begin{cases} \left[\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{XY}^{\top} - \lambda_k \boldsymbol{\Sigma}_{XX}\right]\mathbf{a}_k & = 0 \\ \left[\boldsymbol{\Sigma}_{XY}^{\top}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} - \lambda_k \boldsymbol{\Sigma}_{YY}\right]\mathbf{b}_k & = 0 \end{cases}$$

Donde λ_k satisface

$$\begin{cases} \left|\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{XY}^{\top} - \lambda_k \boldsymbol{\Sigma}_{XX}\right| & = 0 \\ \left|\boldsymbol{\Sigma}_{XY}^{\top}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} - \lambda_k \boldsymbol{\Sigma}_{YY}\right| & = 0 \end{cases}$$

Observe que λ_k es el **autovalor** más grande de la matriz

$$\mathbf{A} = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{XY}^{\top}$$

o equivalentemente de la matriz

$$\mathbf{B} = \boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{XY}^{\top}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$$

Ahora bien, apesar que $\boldsymbol{\lambda}_k$ es el **mismo** en **A** y **B**, tiene asociados **diferentes autovectores**:

$$\mathbf{e}_k \text{ :autovector correspondiente a } \lambda_k \text{ obtenida en } \mathbf{A}$$
$$\mathbf{f}_k \text{ :autovector correspondiente a } \lambda_k \text{ obtenida en } \mathbf{B}$$

Análisis de Correlación Canónica

Finalmente,

$$\mathbf{a}_k = \boldsymbol{\Sigma}_{XX}^{-1/2}\,\mathbf{e}_k$$
$$\mathbf{b}_k = \boldsymbol{\Sigma}_{YY}^{-1/2}\,\mathbf{f}_k$$

y por lo tanto, el $k-$ésimo par de variables canónicas queda dado por

$$U_k = \mathbf{a}_k^\top\,\mathbf{X} = \underbrace{\mathbf{e}_k^\top\boldsymbol{\Sigma}_{XX}^{-1/2}}_{\mathbf{a}_k^\top}\mathbf{X}$$

$$V_k = \mathbf{b}_k^\top\,\mathbf{Y} = \underbrace{\mathbf{f}_k^\top\boldsymbol{\Sigma}_{YY}^{-1/2}}_{\mathbf{b}_k^\top}\mathbf{Y}$$

Análisis de Correlación Canónica

La **correlación canónica** es la correlación entre U_k y V_k en valor absoluto:

$$\rho_k^{*2} = \lambda_k = [\mathrm{Cor}(U_k, V_k)]^2 = \frac{\left(\mathbf{a}_k^\top \, \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}} \, \mathbf{b}_k^\top\right)^2}{\left(\mathbf{a}_k^\top \, \boldsymbol{\Sigma}_{XX} \, \mathbf{a}_k\right)\left(\mathbf{b}_k^\top \, \boldsymbol{\Sigma}_{YY} \, \mathbf{b}_k\right)}$$

Observación: $\sqrt{\lambda_k} > 0$.

Las correlaciones entre las variables originales y las variables canónicas son conocidas como **cargas canónicas**.

La **correlación canónica** es la correlación entre U_k y V_k en valor absoluto:

$$\rho_k^{*2} = \lambda_k = [\mathrm{Cor}(U_k, V_k)]^2 = \frac{\left(\mathbf{a}_k^\top \mathbf{\Sigma}_{\mathbf{XY}} \mathbf{b}_k^\top\right)^2}{\left(\mathbf{a}_k^\top \mathbf{\Sigma}_{XX} \mathbf{a}_k\right)\left(\mathbf{b}_k^\top \mathbf{\Sigma}_{YY} \mathbf{b}_k\right)}$$

Observación: $\sqrt{\lambda_k} > 0$.

Las correlaciones entre las variables originales y las variables canónicas son conocidas como **cargas canónicas**.

Existe un criterio (parecido al gráfico de codo de componentes principales) para escoger el número de pares de correlaciones canónicas.

Notas:

1. Si **las variables originales son estandarizadas**, entonces basta substituir

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \vdots & \boldsymbol{\Sigma}_{XY} \\ \cdots\cdots & \cdots & \cdots\cdots \\ \boldsymbol{\Sigma}_{YX} & \vdots & \boldsymbol{\Sigma}_{YY} \end{bmatrix} \quad \text{por} \quad \mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \vdots & \mathbf{P}_{12} \\ \cdots\cdots & \cdots & \cdots\cdots \\ \mathbf{P}_{21} & \vdots & \mathbf{P}_{22} \end{bmatrix}$$

en consecuencia

$$U_k = \mathbf{e}_k^\top \, \mathbf{P}_{11}^{-1/2} \, \mathbf{Z}^{(1)} \qquad \text{y} \qquad V_k = \mathbf{f}_k^\top \, \mathbf{P}_{22}^{-1/2} \, \mathbf{Z}^{(2)}$$

donde **e** y **f** son los respectivos **autovectores** de

Análisis de Correlación Canónica

$$\mathbf{P}_{11}^{-1}\,\mathbf{P}_{12}\,\mathbf{P}_{22}^{-1}\,\mathbf{P}_{12}^{\top} \qquad \text{y} \qquad \mathbf{P}_{22}^{-1}\,\mathbf{P}_{12}^{\top}\,\mathbf{P}_{11}^{-1}\,\mathbf{P}_{12}$$

y

$$\mathrm{Cov}\left[\mathbf{Z}^{(1)}\right] = \mathbf{P}_{11} \qquad \mathrm{Cov}\left[\mathbf{Z}^{(2)}\right] = \mathbf{P}_{22} \qquad \mathrm{Cov}\left[\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}\right] = \mathbf{P}_{12} = \mathbf{P}_{21}^{\top}$$

siendo

- $\mathbf{Z}^{(1)}$ la estandarización de $\mathbf{X}$,
- $\mathbf{Z}^{(2)}$ la estandarización de $\mathbf{Y}$.

Análisis de Correlación Canónica

2. Las **variables canónicas** pueden ser **estimadas** substituyendo las matrices poblacionales $\mathbf{\Sigma}_{XX}, \mathbf{\Sigma}_{YY}, \mathbf{\Sigma}_{XY}$ y $\mathbf{\Sigma}_{YX}$ por las correspondientes matrices muestrales $\mathbf{S}_{XX}, \mathbf{S}_{YY}, \mathbf{S}_{XY}$ e $\mathbf{S}_{YX}$.
3. Las variables canónicas estandarizadas u obtenidas apartir de la matriz de correlaciones poblacionales, pueden ser obtenidas substituyendo a las matrices de correlaciones poblacionales por las matrices de correlaciones muestrales.

Análisis de Correlación Canónica

Sobre la **interpretación**:

- En general, las **variables canónicas** son artificiales, **no tienen significado físico**, lo cual hace que esta técnica se utilice menos que otras técnicas multivariandas.
- Se da una **interpretación subjetiva a las variables canónicas** de acuerdo con la magnitud de las correlaciones de las variables originales con las variables canónicas. Así, las correlaciones son medidas para interpretar y analizar las cualidades de variables canónicas.

Análisis de Correlación Canónica

Una evaluación importante de la **calidad del potencial de las variables canónicas** es medir el poder de resumen de la variabilidad contenida en los conjuntos atraves de la proporción de la varianza total

- **explicada por las variables canónicas** es dada por

$$PE_{U_k} = \frac{\sum_{i=1}^{p} [\mathrm{Cor}(U_k, X_i)]^2}{\mathrm{Tr}(\mathbf{S}_{XX})} \quad \text{y} \quad PE_{V_k} = \frac{\sum_{i=1}^{q} [\mathrm{Cor}(V_k, Y_i)]^2}{\mathrm{Tr}(\mathbf{S}_{XX})}$$

Si las variables originales son estandarizadas, $\mathbf{S_Z} = \mathbf{R_X}$, luego

$$PE_{U_k} = \frac{\sum_{i=1}^{p} [\mathrm{Cor}(U_k, Z_i^{(1)})]^2}{p} \quad \text{y} \quad PE_{V_k} = \frac{\sum_{i=1}^{q} [\mathrm{Cor}(V_k, Z_i^{(2)})]^2}{p}$$

Análisis de Correlación Canónica

- explicada por **los primeros r pares canónicos**

$$PVTE_{U_k} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{p}[\mathrm{Cor}(U_k, X_i)]^2}{\mathrm{Tr}(\mathbf{S}_{XX})} \quad \text{y} \quad PVTE_{V_k} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{q}[\mathrm{Cor}(V_k, Y_i)]^2}{\mathrm{Tr}(\mathbf{S}_{XX})}$$

Si las variables originales son estandarizadas, $\mathbf{S_Z} = \mathbf{R_X}$, luego

$$PVTE_{U_k} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{p}[\mathrm{Cor}(U_k, Z_i^{(1)})]^2}{p} \quad \text{y} \quad PVTE_{V_k} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{q}[\mathrm{Cor}(V_k, Z_i^{(2)})]^2}{p}$$

Análisis de Correlación Canónica

- explicada por **los primeros r pares canónicos**

$$PVTE_{U_k} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{p}[\mathrm{Cor}(U_k, X_i)]^2}{\mathrm{Tr}(\mathbf{S}_{XX})} \quad \text{y} \quad PVTE_{V_k} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{q}[\mathrm{Cor}(V_k, Y_i)]^2}{\mathrm{Tr}(\mathbf{S}_{XX})}$$

Si las variables originales son estandarizadas, $\mathbf{S_Z} = \mathbf{R_X}$, luego

$$PVTE_{U_k} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{p}[\mathrm{Cor}(U_k, Z_i^{(1)})]^2}{p} \quad \text{y} \quad PVTE_{V_k} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{q}[\mathrm{Cor}(V_k, Z_i^{(2)})]^2}{p}$$

¿Cuando aplicar (o vale la pena aplicar) análisis de correlación canónica?

Análisis de Correlación Canónica

1. cuando se tiene interés en estudiar la relación de dos conjuntos de variables.
2. cuando en el planteamiento de hipótesis

$$\begin{aligned} H_0 &: \boldsymbol{\Sigma}_{\mathbf{X}\,\mathbf{Y}} = \mathbf{0} \\ H_1 &: \boldsymbol{\Sigma}_{\mathbf{X}\,\mathbf{Y}} \neq \mathbf{0} \end{aligned}$$

o equivalentemente

$$\begin{aligned} H_0 &: \mathbf{P}_{\mathbf{X}\,\mathbf{Y}} = \mathbf{0} \\ H_1 &: \mathbf{P}_{\mathbf{X}\,\mathbf{Y}} \neq \mathbf{0} \end{aligned}$$

se rechaza H_0.

usamos:

- Para muestras relativamente grandes

$$Q_1 = -n \ln \prod_{i=1}^{p} (1 - \lambda_i) \overset{\mathsf{H}_0 \text{ verd.}}{\sim} \chi^2_{pq}$$

- Para muestras pequeñas

$$Q_2 = -\left[n - 1 - \frac{1}{2}(p + q + 1)\right] \ln \prod_{i=1}^{p} (1 - \lambda_i) \overset{\mathsf{H}_0 \text{ verd.}}{\sim} \chi^2_{pq}$$

Análisis de Correlación Canónica

Example

Los datos del archivo `Hijos.xlsx` corresponden a las **medidas de la cabeza (en milímetros)** de cada uno de los **dos primeros hijos adultos de 25 familias**. La familia es el "individuo" y x_1 (longitud de la cabeza del primer hijo), x_2 (ancho de cabeza del primer hijo), x_3 (longitud de la cabeza del segundo hijo) y x_4 (ancho de cabeza del segundo hijo). Estos datos fueron recopilados por Frets (1921)[a], y la preguntade interés era si existe una **relación entre las medidas de la cabeza de los pares de hijos**.

[a] Frets, G. P. (1921), "Heredity of head form in man", *Genetica*, 3, 193(384)

Análisis de Correlación Canónica

Claramente en este caso tenemos **dos grupos de variables**:

$$\begin{aligned} \text{Hijo 1} &: \{X_1, X_2\} \\ \text{Hijo 2} &: \{Y_1, Y_2\} \end{aligned}$$

Aquí podemos encontrar

$$\mathbf{R} = \begin{bmatrix} 1.00 & 0.73 & \vdots & 0.71 & 0.70 \\ 0.73 & 1.00 & \vdots & 0.69 & 0.71 \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdots\cdots \\ 0.71 & 0.69 & \vdots & 1.00 & 0.84 \\ 0.70 & 0.71 & \vdots & 0.84 & 1.00 \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{XX} & \vdots & \mathbf{R}_{XY} \\ \cdots\cdots & \cdots & \cdots\cdots \\ \mathbf{R}_{YX} & \vdots & \mathbf{R}_{YY} \end{bmatrix}$$

Análisis de Correlación Canónica

Note que **todas las correlaciones son altas**, lo cual nos da **sospecha que vale la pena aplicar análisis de correlación canónica**. Sin embargo, debemos hacer una prueba formal. Hágala!

Para aplicar correlación canónica, cargamos los datos

```
# ----- Datos Hijos ----- #
require(tidyverse)
Hijos <- read_excel("D:/Desktop/Teaching/Dataset/Hijos.xlsx")
Datos<-Hijos %>% as.data.frame()
head(Datos)
  familia  x1  x2  x3  x4
1       1 191 155 179 145
2       2 195 149 201 152
3       3 181 148 185 149
4       4 183 153 188 149
5       5 176 144 171 142
6       6 208 157 192 152
```

Análisis de Correlación Canónica

Extraemos los dos conjuntos de variables y extraemos las matrices de correlaciones muestrales

```
# --- X=(x1,x2) e Y=(x3,x4) --- #
X<-Datos[,2:3]
Y<-Datos[,4:5]

# --- Matrices de Correlación Muestrales --- #
R_XX<-cor(X)  # Mat. de corr. muest. de las var. de X
R_YY<-cor(Y)  # Mat. de corr. muest. de las var. de Y
R_XY<-cor(X,Y) # Mat. de corr. muest. entre X e Y
R_YX<-cor(Y,X) # = R_YX=R^t_XY
```

Calculamos las matrices **A** e **B** y calculamos sus autovalores (λ_k) y autovectores ($\mathbf{e}_k$ y $\mathbf{f}_k$)

```
# --- Matrices A e B --- #
A<-solve(R_XX)%*%(R_XY)%*%solve(R_YY)%*%t(R_XY)
B<-solve(R_YY)%*%t(R_XY)%*%solve(R_XX)%*%(R_XY)

# --- Lambda_k

autoA<-eigen(A)
lambdak<-autoA$values
lambdak
[1] 0.621744734 0.002887956
```

Análisis de Correlación Canónica

```
autoB<-eigen(B)
lambda_k<-autoB$values
lambda_k
[1] 0.621744734 0.002887956
```

Observe que los **autovalores** de **A** e **B** son **exactamente los mismos**.

Calculamos las c**orrelaciones canónicas** $\rho_{U,V}$

```
# --- Correlaciones Canonicas
rho<-sqrt(lambda_k)
round(rho,4)
[1] 0.7885 0.0537
```

Análisis de Correlación Canónica

Tenemos que

$$\rho_{U_1,V_1} = 0.7885 \qquad \text{e} \qquad \rho_{U_2,V_2} = 0.0537$$

el **segundo par de correlaciones canónicas** (estandarizadas) **explican muy poco (0.0537)** de la correlación entre los dos conjuntos de variables.

```
# --- e_k y f_k
e_k<-autoA$vectors
e_k
          [,1]       [,2]
[1,] 0.7269968 -0.7040109
[2,] 0.6866408  0.7101892
```

Análisis de Correlación Canónica

```
f_k<-autoB$vectors
f_k
           [,1]       [,2]
[1,] -0.6837994 -0.7091095
[2,] -0.7296700  0.7050984
```

Extraemos los **coeficientes** para cada par de **variables canónicas**

```
# --- Variables Canonicas
require(expm)

# coeficientes de a
raiz_Rxx<-sqrtm(R_XX)
inv_raiz_Rxx<-solve(raiz_Rxx)
ak_trasp<-t(e_k)%*%inv_raiz_Rxx
```

Análisis de Correlación Canónica

```
a1_trasp<-ak_trasp[1,]
a2_trasp<-ak_trasp[2,]

round(a1_trasp,3)
[1] 0.576 0.498
round(a2_trasp,3)
[1] -1.370  1.375

# coeficientes de b
raiz_Ryy<-sqrtm(R_YY)
inv_raiz_Ryy<-solve(raiz_Ryy)
bk_trasp<-t(f_k)%*%inv_raiz_Ryy
```

Análisis de Correlación Canónica

```
b1_trasp<-bk_trasp[1,]
b2_trasp<-bk_trasp[2,]

round(b1_trasp,3)
[1] -0.464 -0.578
round(b2_trasp,3)
[1] -1.765  1.762
```

Por lo tanto, **la primera variable canónica** (de las variables estandarizadas)

$$\begin{aligned} U_1 &= 0.576\, Z_1 + 0.498\, Z_2 \\ V_1 &= -0.464\, Z_3 - 0.578\, Z_4 \\ \rho_{U_1,V_1} &= 0.7885 \end{aligned}$$

Análisis de Correlación Canónica

son responsables de la **mayor correlación (0.7885)** entre los variables cefálicas de los dos primeros hijos de las familias estudiadas. Las **variables individuales** contribuyen con "pesos/cargas" muy similares.

Observe que en este par de variables canónicas, cada una es una **suma/resta ponderada de las dos medidas de la cabeza** y podría ser etiquetada como la "**circunferencia de la cabeza**".

Análisis de Correlación Canónica

La **segunda variable canónica** (de las variables estandarizadas)

$$\begin{aligned} U_2 &= -1.370\, Z_1 + 1.375\, Z_2 \\ V_2 &= -1.765\, Z_3 + 1.762\, Z_4 \\ \rho_{U_2,V_2} &= 0.0537 \end{aligned}$$

explica muy poco (0.0537) de la correlación entre los dos primeros hijos.

Observe que para este par de variables canónicas, **cada una es una diferencia ponderada de las dos medidas de la cabeza**, así que se puede interpretar (aproximadamente) como "**la forma de la cabeza**".

Análisis de Correlación Canónica

Podemos hacer el análisis de **correlación canónica en** `R` con:

```
 # --- el paquete candisc
require(candisc)
cca<-cancor(X,Y)

cca$cancor   #correlaciones canonicas
[1] 0.7885079 0.0537397

# --- Coeficientes de a y de b
coef(cca, type = "both", standardize = TRUE)
```

Análisis de Correlación Canónica

```
[[1]]
        Xcan1     Xcan2
x1 -0.5521896 -1.366374
x2 -0.5215372  1.378365

[[2]]
        Ycan1     Ycan2
x3 -0.5044484 -1.768570
x4 -0.5382877  1.758566

 # --- el paquete CCA
require(CCA)
z1<-scale(X)
z2<-scale(Y)
acc<-cc(z1,z2)
```

Análisis de Correlación Canónica

```
acc$cor          #correlaciones canonicas
[1] 0.7885079 0.0537397

acc$xcoef          # coeficientes de a
         [,1]      [,2]
x1 -0.5521896 -1.366374
x2 -0.5215372  1.378365

acc$ycoef         # coeficientes de b
         [,1]      [,2]
x3 -0.5044484 -1.768570
x4 -0.5382877  1.758566
```

Tareas

1. Realice el análisis de correlación canónica del ejemplo para las variables originales (sin estandarizar)
2. Replique los ejemplos 10.1 y 10.3 de Johnson and Wichern (2013), Applied Multivariate Statistical Analysis, pp. 543 y 549.