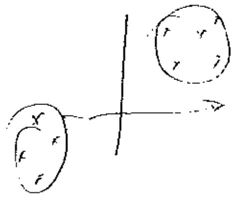


# Tema 36 Daniel Peña

1. Análisis de Correlación Canónica. Introducción.

~~2. Introducción.~~

2. Correlación canónica y variables canónicas: cálculo e interpretación geométrica.

3. Propiedades.

4. Contrastación del modelo y análisis de la dimensionalidad.

5. Relación con otras técnicas de análisis multivariante.

E. Uriel pg. 7
D. Peña pg 486
C. Pérez pg 277

1. ANÁLISIS DE CORRELACIÓN CANÓNICA. INTRODUCCIÓN.

Es un método para relacionar las variables en dos grupos simétricos, es decir, se trata ambos grupos de variables del mismo modo.
(sin relac. causa-efecto)
Este estudio fué iniciado por Hotelling en 1936 como extensión de la idea de componentes principales mediante el análisis de correlaciones canónicas.
La correlación canónica se utiliza cuando un conjunto de variables multivariante puede dividirse en dos grupos homogéneos (por criterios económicos, demográficos, sociales...), y se desea estudiar la relación entre ambos conjunto de variables. En particular, los dos grupos pueden corresponder a las mismas variables medidas en dos momentos distintos ~~en~~ el tiempo, el espacio, etc...

2. CORRELACIÓN CANÓNICA Y VARIABLES CANÓNICAS: CÁLCULO E INTERPRETACIÓN GEOMÉTRICA.

Suponemos que se tiene un conjunto de datos de n individuos y K variables que pueden subdividirse en dos grupos: el primero incluye p var. y el segundo q, donde $p+q=K$. Llamaremos:

$X_{n\times p}$: matriz que contiene los valores de las p primeras variables en los n elementos.

$Y_{n\times q}$: matriz que contiene los valores de las q segundas variables en los n elementos.

La matriz de covarianzas conjunta es:

$$V_{xy} = E\left(\begin{pmatrix}x\\y\end{pmatrix}(x'\,y')\right) = \begin{bmatrix}V_{11} & V_{12}\\V_{21} & V_{22}\end{bmatrix}$$

Para investigar la relación entre ambos grupos de variables, se buscan dos variables ~~resumen~~ indicadoras, una de cada conjunto, que tengan correlación máxima.
Es posible, que una vez encontrada esta primera relación no exista más relación entre ambos conjuntos de variables y entonces toda la relación entre los conjuntos se resume en una dimensión.
Para comprobarlo, se busca una segunda variable indicadora del primer conjunto, que esté incorrelada con la primera, y que tenga correlación máxima con otra variable indicadora del segundo conjunto.
Procediendo de esta manera, se pueden obtener $r = \min(p,q)$ relaciones entre variables indicadoras de ambos conjuntos que pueden ordenarse según su importancia. Determinar el número de relaciones entre las variables permite juzgar cuántas dimensiones distintas tiene la relación.
El proceso para obtener las $2r$ combinaciones lineales, $(x_1^*, x_2^*, \ldots x_q^*)$, $(y_1^*, y_2^*, \ldots y_q^*)$ donde $r = \min(p,q)$, que se llamarán variables canónicas, consiste en obtener los valores y vectores propios de las matrices:

$$A_{p\times p} = V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}$$

$$B_{q\times q} = V_{22}^{-1} V_{21} V_{11}^{-1} V_{12}$$

Estas matrices tienen un rango igual al min(p, q), y si se extraen sus r valores propios no nulos y los vectores propios unidos a dichos valores propios, se pueden formar r combinaciones lineales de las variables de ambos grupos que:

a) Tienen correlación máxima cuando provienen del mismo valor propio

b) Están incorreladas dentro de ~~mismo~~ cada grupo

c) Están incorreladas si corresponden a distintos vectores propios.

<u>Cálculo de las variables canónicas</u>

Resumido *

Sea,

$$x^* = X\alpha = \sum_{i=1}^{p} \alpha_i x_i$$
$$y^* = Y\beta = \sum_{j=1}^{q} \beta_j y_j$$

<u>hay que encontrar los vectores $\alpha$ y $\beta$ para que $x^*$ e $y^*$ ~~e~~ tengan máx. correlación</u>

Suponemos:

$$X_{p\times 1} \xrightarrow{d} N_p(0, V_{11}) \qquad \text{t.q.} \quad V_{x,y} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$
$$y_{q\times 1} \xrightarrow{d} N_q(0, V_{22})$$

Se maximiza el cuadrado de la correlación entre $(x^*, y^*)$ respecto a $\alpha$ y $\beta$:

$$\max \rho^2 = \frac{(\alpha' V_{12} \beta)^2}{(\alpha' V_{11} \alpha)(\beta' V_{22} \beta)}$$

s.a. $\quad \mathrm{Var}\ x^* = \alpha' V_{11} \alpha = 1$

$\quad \mathrm{Var}\ y^* = \beta' V_{22} \beta = 1$

Resumido *

Se resuelve mediante multiplicadores de Lagrange (ver D.Peña pg 487) y se llega a:

$$\alpha' V_{12}\beta = \lambda\, \alpha' V_{11}\alpha = \lambda$$
$$\beta' V_{21}\alpha = \mu\, \beta' V_{22}\beta = \mu$$

y como $\lambda = \alpha' V_{12}\beta = \beta' V_{21}\alpha = \mu$ entonces:

$$\left.\begin{array}{l} V_{12}\beta = \lambda V_{11}\alpha \\ V_{21}\alpha = \lambda V_{22}\beta \end{array}\right. \Rightarrow \left(V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}\right)\alpha = \lambda^2\alpha \Rightarrow$$

$\Rightarrow \alpha$ es el vector propio ligado al valor propio $\lambda^2$ de la matriz cuadrada de dimensión p $A_{p\times p} = V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}$ con valor propio $\lambda^2$.

Análogamente, se obtiene que $\beta$ debe ser el vector propio ligado al valor propio $\mu^2$ de la matriz $B_{q\times q} = V_{22}^{-1} V_{21} V_{11}^{-1} V_{12}$

Observamos que $\lambda^2 = \mu^2 = \rho^2$, por lo que tendremos que tomar el vector propio ligado al mayor valor propio

* Resumido:

La solución buscada requiere:

1. Construir las dos matrices cuadradas de dimensiones p y q, A y B. El vector propio asociado a su máximo valor propio (que coincide en ambas) proporciona las variables canónicas.
2. Este mayor valor propio es el cuadrado del coefic. de correlación ~~canónica~~ entre las var. canónicas

Para buscar una segunda var. indicadora del primer cjto. de variables que esté incorrelada con la primera y que tenga correlación máx. con otra var. indicadora del seg. conjunto se procede de la siguiente manera (al igual que para obtener la $r = \min(p,q)$ relaciones entre var. indicadoras).

El proceso para obtener las $2r$ comb. lineales $(x_1^*, \dots, x_r^*)$; $(y_1^*, \dots, y_r^*)$, que llamaremos variables canónicas, consiste en obtener los valores y vectores propios de las matrices $A_{p\times p}$ y $B_{q\times q}$.

## Interpretación geométrica

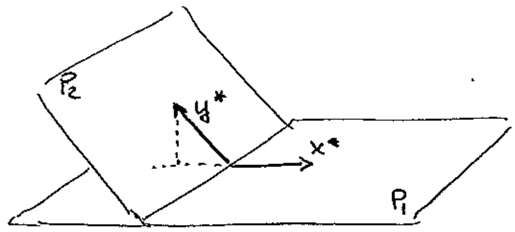
Sea $S(X)$: espacio generado por las columnas de $X_{n\times p}$
$S(Y)$: " " " " " " $Y_{n\times q}$

$$x^* = X\alpha \in S(X)$$
$$y^* = Y\beta \in S(Y)$$

$P_1, P_2$ : matrices proyección sobre $S(X)$, $S(Y)$ dadas por:

$$P_1 = X(X'X)^{-1}X'$$
$$P_2 = Y(Y'Y)^{-1}Y'$$

$\Rightarrow$ la condición exigida es : $\begin{cases} P_1 y^* = \lambda x^* \\ P_2 x^* = \mu y^* \end{cases}$




Represent. de las primeras var. canónicas
($\theta$: ángulo entre ambos subespacios)

$(\cos\theta)^2 = \rho^2 \Rightarrow$ la máx. correlación canónica es el coseno del ángulo que forman los subesp. generados por $X$ y por $Y$.

3. PROPIEDADES DE LAS VARIABLES Y CORRELACIONES CANÓNICAS

- las var. canónicas son indicadores de los dos conjuntos de variables que se definen por pares, con la condición de máx correlación.
- los coeficientes de las var. canónicas son los vectores propios ligados al mismo valor propio de las matrices $V_{ii}^{-1/2} V_{ij} V_{jj}^{-1} V_{ji}$ ;; $i = 1, 2$ ; $j \neq i$
- Si $\alpha_i' X$ es una var. canónica también lo es $-\alpha_i' X$, y los signos de las var. canónicas suelen tomarse de manera que las correlaciones entre las var. canónicas $\alpha' X$ y $\beta' y$ sean positivas.
- las correlaciones canónicas, $\lambda_i^2$, son el cuadrado del coeficiente de correlación entre las dos variables canónicas correspondientes.
- las correlaciones canónicas son invariantes ante transformaciones lineales de las variables.
- la primera correlación canónica, $\lambda_1^2$, es mayor o igual que el mayor coeficiente de correlación simple al cuadrado entre una variable de cada conjunto.
- El coeficiente de correl. canónica $\lambda_i^2$ es el coeficiente de determinación de una regresión múltiple con respuesta la variable $y_i^* = \beta_i' y$, y variables explicativas las $x$. También es el coef. de determinación entre la regresión múltiple entre $x_i^* = \alpha_i' x$ y el conjunto de las $y$.

## 4. CONTRASTACIÓN DEL MODELO Y ANÁLISIS DE LA DIMENSIONALIDAD.

- Se puede construir un contraste para estudiar si los dos conjuntos de variables están incorrelados ($V_{12}=0$) bajo la hipótesis :

$$x \xrightarrow{d} N_p(0, V_{11})$$
$$y \xrightarrow{d} N_q(0, V_{22})$$

Este contraste equivale a contrastar que todas las correlaciones canónicas son nulas:

$$H_0 : V_{12} = 0$$
$$H_1 : V_{12} \neq 0$$

$$\lambda = -n \sum_{j=1}^{r} \log(1-\lambda_j^2) \underset{\substack{\text{(aproxim.}\\ \text{con corrección}\\ \text{de Barlett)}}}{\Longrightarrow} \lambda' = -m \sum_{j=1}^{r} \log(1-\lambda_j^2)$$

$$\lambda' \xrightarrow{d} \chi^2_{pq}$$

Rechazamos Ho si $\lambda'$ es grande, es decir, cuando los coeficientes de correlación canónica son grandes.

Dimensionalidad

- Este contraste se puede extender para estudiar que los primeros s coeficientes de correlación canónica son distintos de cero y los restantes r-s iguales a cero :

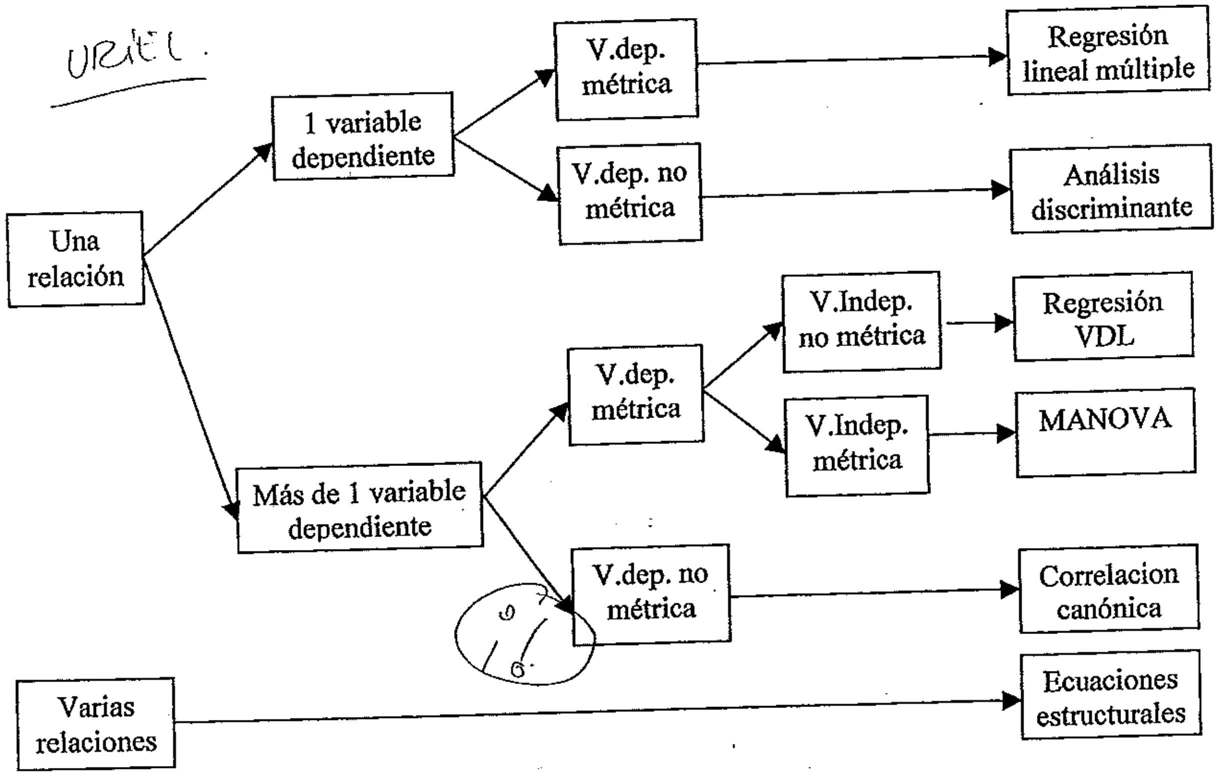
$$H_0 : \lambda_i > 0 \quad i=1,\dots,s; \; \lambda_{s+1} = \dots = \lambda_r = 0$$
$$H_1 : \lambda_i > 0 \quad i=1,\dots,s; \text{ al menos uno } \lambda_j > 0; \; j=s+1,\dots,r$$
$$\lambda = -m \sum_{j=s+1}^{r} \log(1-\lambda_j^2) \xrightarrow{d} \chi^2_{(p-s)(q-s)}$$

Acept. Ho ⇒ la dependencia entre variab. puede expresarse mediante s var. indicadoras

Rechaz. Ho ⇒ no hay reducción de la dimensión posible y describir la dependencia requiere las r dimensiones.

## 5. RELACIÓN CON OTRAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE.

TÉCNICAS DE DEPENDENCIA




Además de su interés propio, el análisis de correlaciònes canónicas cubre como casos particulares las técnicas de regresión y, por extensión, las de análisis discriminante.

Cuando cada uno de los conjuntos tenga una única variable, el análisis de correlación canónica es equivalente al análisis de regresión simple.

Cuando el uno de los conjuntos tenga una variable (p=1) y el otro conjunto tenga varias variables (q>1), el análisis de correlación canónica es equivalente al análisis de regresión múltiple.

El análisis discriminante también se puede abordar desde el análisis de correlación canónica, donde la matriz X (nxp) es la matriz de las p variables explicativas y la matriz Y (nxq) contiene las G-1 = q variables binarias definidas de la siguiente forma:

$$y_i = \begin{cases} 1 & \text{si la observación pertenece al gurpo i, } i = 1, \ldots, G-1 \\ 0 & \text{en otro caso} \end{cases}$$