Capítulo 9

Análisis de correlación canónica

INTRODUCCIÓN

- Involucra la partición de una colección de variables en dos conjuntos.
- **El objetivo** es encontrar combinaciones lineales del tipo:

$$\boxed{W = a'X} \qquad \text{y} \qquad \boxed{V = b'Y}$$

tal que W y V tienen **correlación máxima**.

- El análisis de correlación canónica puede ser visto como una extensión de la regresión múltiple.

EJEMPLOS

1.- Un investigador médico está interesado en determinar si el estilo de vida y los hábitos alimenticios de individuos tienen un efecto en su salud midiendo variables como la hipertensión, el peso, la ansiedad y el nivel de tensión arterial.

2.- El director comercial de unos grandes almacenes está interesado en determinar si existe relación entre los tipos de productos comprados y las personalidad y el estilo de vida de sus clientes.

CORRELACIÓN CANÓNICA: ACERCAMIENTO ANALÍTICO

Consideremos las ecuaciones canónicas:

$$\boxed{W_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p}$$

$$\boxed{V_1 = b_{11}Y_1 + b_{12}Y_2 + \ldots + b_{1q}Y_q.}$$

OBJETIVO

Estimar $a_{11}, \ldots, a_{1p}$ y $b_{11}, \ldots, b_{1q}$ tal que C_1 es **máximo**.

- C_1 : Es la correlación entre W_1 y V_1. Recibe el nombre de **correlación canónica**.
- W_1 y V_1 son llamadas **variables canónicas**.

PASO (1)

1) Se estiman W_1 y V_1
2) Se identifica un segundo conjunto de variables canónicas W_2 y V_2

$$\boxed{W_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p}$$

$$\boxed{V_2 = b_{21}Y_1 + b_{22}Y_2 + \ldots + b_{2q}Y_q}$$

SE VERIFICA

- La correlación entre W_2 y V_2 es **máxima**
- W_2 y V_2 están **incorreladas** con W_1 y V_1 .

$\boxed{\textbf{PASO } (m)}$

ESTE PROCEDIMIENTO SE REPITE HASTA IDENTIFICAR
EL $m-$ésimo CONJUNTO DE VARIABLES CANÓNICAS W_m y V_m :

$$\boxed{W_m = a_{m1}X_1 + a_{m2}X_2 + \ldots + a_{mp}X_p}$$

$$\boxed{V_m = b_{m1}Y_1 + b_{m2}Y_2 + \ldots + b_{mq}Y_q}$$

de forma que:

- C_m es máxima
- $Cor\,(V_j, V_k) = 0 \quad \forall\, j \neq k$
- $Cor\,(W_j, W_k) = 0 \quad \forall\, j \neq k$
- $Cor\,(W_j, V_k) = 0 \quad \forall\, j \neq k$

EL PROBLEMA DE MAXIMIZACIÓN (PASO 1)

- Sea X un vector aleatorio de dimensión p
- Sea Y un vector aleatorio de dimensión q
- Sea $\sum_{XX}$ la matriz de covarianzas de X
- Sea $\sum_{YY}$ la matriz de covarianzas de Y
- Sean $W = a'X$ y $V = b'Y$ combinaciones lineales de X e Y respectivamente.

OBJETIVO

Estimar a' y b' tal que la correlación entre W y V

$$\boxed{a' \sum_{XY} b}$$

es máxima sujeto a las restricciones: $\left\{ \begin{array}{l} \boxed{a' \sum_{XX} a = 1} \\ \\ \boxed{b' \sum_{YY} b = 1} \end{array} \right.$

PROBLEMA

$\downarrow$

Maximización con restricciones

SOLUCIÓN

$\downarrow$

Multiplicadores de Lagrange

- La solución a' para el primer paso se obtiene:
  - Calculando los vectores propios de la matriz $\sum_{XX}^{-1} \sum_{XY} \sum_{YY}^{-1} \sum_{YX}$
  - Imponiendo la condición $a' \sum_{XX} a = 1$
- La solución b' para el primer paso se obtiene:
  - Calculando los vectores propios de la matriz $\sum_{YY}^{-1} \sum_{YX} \sum_{XX}^{-1} \sum_{XY}$
  - Imponiendo la condición $b' \sum_{YY} b = 1$.

ILUSTRACIÓN

X_1	X_2	Y_1	Y_2
1.051	−0.435	0.083	0.538
−0.419	−1.335	−1.347	−0.723
1.201	0.445	1.093	−0.112
0.661	0.415	0.673	−0.353
−1.819	−0.945	−0.817	−1.323
−0.899	e 0.375	−0.297	−0.433
3.001	1.495	1.723	2.418
−0.069	−2.625	−2.287	−1.063
−0.919	0.385	−0.547	0.808
−0.369	−0.265	−0.447	−0.543
−0.009	−0.515	0.943	−0.633
0.841	1.915	1.743	1.198
0.781	1.845	1.043	2.048
0.631	−0.495	0.413	−0.543
−1.679	− 0.615	− 1.567	−0.643
−0.229	−0.525	−0.777	−0.252
−0.709	−0.975	0.523	−0.713
−0.519	0.055	− 0.357	0.078
0.051	0.715	0.133	0.328
0.221	0.245	0.403	0.238
−1.399	−0.645	−0.817	−1.133
0.651	0.385	1.063	− 0.633
−0.469	−0.125	−0.557	−0.393
0.421	1.215	−0.017	1.838

PROCEDIMIENTO

OBTENER LA ESTIMACIÓN: $a' = (a_1, a_2)$

- Calcular los vectores propios de la matriz $\sum_{XX}^{-1} \sum_{XY} \sum_{YY}^{-1} \sum_{YX}$
  - Las matrices de covarianzas de las variables X e Y son:

$$\boxed{\sum_{XX} = \begin{pmatrix} 1.0372 & 0.5675 \\ 0.5675 & 1.0221 \end{pmatrix}} \qquad \boxed{\sum_{YY} = \begin{pmatrix} 1.1068 & 0.5686 \\ 0.5686 & 1.0668 \end{pmatrix}}$$

$$\boxed{\sum_{YX} = \begin{pmatrix} 0.7608 & 0.7943 \\ 0.7025 & 0.8452 \end{pmatrix}}$$

- Obtener la matriz

$$\sum_{XX}^{-1} \sum_{XY} \sum_{YY}^{-1} \sum_{YX} = \begin{pmatrix} 0.3417 & 0.3699 \\ 0.5189 & 0.5951 \end{pmatrix}$$

- Obtener $a' = (a_1, a_2)$ diagonalizando la matriz anterior

$$\boxed{a_1 = 0.5358}$$

$$\boxed{a_2 = 0.8443}$$

- Imponer la condición $a' \sum_{XX} a = 1$

$$\boxed{a' \sum_{XX} a = (0.5358,\ 0.8443) \sum_{XX} \begin{pmatrix} 0.5358 \\ 0.8443 \end{pmatrix} = 1.5926}$$

$$\boxed{\text{Imponer la restricción } a' \sum_{XX} a = 1} \quad \Longrightarrow \quad \boxed{\begin{cases} \dfrac{a_1}{\sqrt{1.5926}} = 0.4246 \\ \\ \dfrac{a_2}{\sqrt{1.5926}} = 0.669 \end{cases}}$$

Entonces

$$\boxed{W = 0.4246\ X_1 + 0.669\ X_2}.$$

- La estimación de b' puede obtenerse de forma análoga.

Bibliografía utilizada:

★ **Sharma, Subbash (1996)**. *"Applied Multivariate Techniques". Ed.: Hohn Wiley & Sons, Inc.*

♦ **Temporalización:** Una hora