Relaciones entre dos conjuntos de variables

Básicamente se tienen dos situaciones cuando queremos relacionar dos conjuntos de variables:

1. **Situación simétrica:** se desea tratar ambos grupos de variables del mismo modo, donde no se le da ninguna preferencia a ninguno de los dos conjuntos de variables para explicar al otro. Se busca investigar globalmente la relación entre ambos conjuntos de variables.
2. **Situación asimétrica:** corresponde al caso donde un conjunto de variables explican a las variables del segundo conjunto y no al revés.

Relaciones entre dos conjuntos de variables

- **Ejemplo de la situacion simétrica**: cuando se considera que el primer grupo de variables es de rendimiento escolar y el segundo de uso del tiempo de ocio entre los estudiante. No existe claramente un conjunto de variables que sea la causa del otro conjunto.
- **Ejemplo del la situacion asimétrica**: el primer grupo de variables mide el rendimiento en secundaria y el segundo en la universidad. Claramente las primeras pueden causar las segundas pero no al revés. Los modelos de regresion multivariada son una herramienta para analizar los datos asimétricos.

La situación simétrica la estudio inicialmente Hotelling en 1936, como una extensión de los componenentes principales y se denomina **análisis de correlación canónica**

Análisis de Correlación Canónica

- El análisis de correlación canónica busca identificar y cuantificar las asociaciones entre dos conjuntos de variables.
- El análisis de correlación canónica se enfoca en la correlación entre una combinación lineal de las variables en un conjunto y una combinación lineal de las variables en otro conjunto.
- La idea es determinar primero el par de combinaciones lineales que tienen la mayor correlación. A continuación, se determina un segundo par de combinaciones lineales que tienen la mayor correlación entre todos los pares no correlacionados con el primer par seleccionado, y así sucesivamente.
- Los pares de combinaciones lineales se denominan **variables canónicas**, y sus correlaciones se llaman **correlaciones canónicas**.

Análisis de Correlación Canónica: Variables canónicas y correlaciones Canónicas

- Las **correlaciones canónicas** miden la fuerza de asociación entre los dos conjuntos de variables.
- El aspecto de maximización de la técnica representa un intento de concentrar una relación de alta dimensionalidad entre dos conjuntos de variables en unos cuantos pares de **variables canónicas**
- El primer grupo, de p variables, se representa por el vector aleatorio $\boldsymbol{x}^{(1)}$, $(p \times 1)$. El segundo grupo, de q variables, se representa por el vector aleatorio $\boldsymbol{x}^{(2)}$, $(q \times 1)$.
- Se asumirá en el desarrollo teórico, que $\boldsymbol{x}^{(1)}$ representa el conjunto *más pequeño*, de modo que $p \leq q$.

Análisis de Correlación Canónica: Variables canónicas y correlaciones Canónicas

Para los vectores aleatorios $\boldsymbol{x}^{(1)}$ y $\boldsymbol{x}^{(2)}$, sea

$$E(\boldsymbol{x}^{(1)}) = \mu^{(1)}, \quad Cov(\boldsymbol{x}^{(1)}) = \boldsymbol{\Sigma}_{11}$$

$$E(\boldsymbol{x}^{(2)}) = \mu^{(2)}, \quad Cov(\boldsymbol{x}^{(2)}) = \boldsymbol{\Sigma}_{22}$$

$$Cov(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}'$$

Será conveniente considerar conjuntamente $\boldsymbol{x}^{(1)}$ y $\boldsymbol{x}^{(2)}$, por lo que usando resultados anteriores y la relación anterior, se tiene que

$$\boldsymbol{X}_{(p+q)\times 1} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \hline \boldsymbol{x}^{(2)} \end{bmatrix} = \left[\begin{array}{c} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_p^{(1)} \\ \hline x_1^{(2)} \\ x_2^{(2)} \\ \vdots \\ x_q^{(2)} \end{array}\right]$$

Análisis de Correlación Canónica: Variables canónicas y correlaciones Canónicas

$\boldsymbol{X}$ tiene vector de medias

$$\mu_{(p+q)\times 1} = E[\boldsymbol{X}] = \left[\frac{E(\boldsymbol{x}^{(1)})}{E(\boldsymbol{x}^{(2)})}\right] = \left[\frac{\mu^{(1)}}{\mu^{(2)}}\right]$$

y matriz de covarianzas

$$\boldsymbol{\Sigma}_{(p+q)\times(p+q)} = E\left[(\boldsymbol{X}-\mu)(\boldsymbol{X}-\mu)'\right]$$

$$= \left[\begin{array}{c|c} E\left((\boldsymbol{x}^{(1)}-\mu^{(1)})(\boldsymbol{x}^{(1)}-\mu^{(1)})'\right) & E\left((\boldsymbol{x}^{(1)}-\mu^{(1)})(\boldsymbol{x}^{(2)}-\mu^{(2)})'\right) \\ \hline E\left((\boldsymbol{x}^{(2)}-\mu^{(2)})(\boldsymbol{x}^{(1)}-\mu^{(1)})'\right) & E\left((\boldsymbol{x}^{(2)}-\mu^{(2)})(\boldsymbol{x}^{(2)}-\mu^{(2)})'\right) \end{array}\right]$$

$$= \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11\,p\times p} & \boldsymbol{\Sigma}_{12\,p\times q} \\ \hline \boldsymbol{\Sigma}_{21\,q\times p} & \boldsymbol{\Sigma}_{22\,q\times q} \end{array}\right]$$

Análisis de Correlación Canónica: Variables canónicas y correlaciones Canónicas

- Las covarianzas entre pares de variables de diferentes conjuntos, es decir, una variable de $\boldsymbol{x}^{(1)}$ y una variable de $\boldsymbol{x}^{(2)}$, están contenidas en $\boldsymbol{\Sigma_{12}}$, o de forma equivalente, en $\boldsymbol{\Sigma_{21}}$.
- Es decir, los pq elementos de $\boldsymbol{\Sigma_{12}}$ miden la asociación entre los dos conjuntos. Cuando p y q son relativamente grandes, la interpretación de los elementos de $\boldsymbol{\Sigma_{12}}$, colectivamente resulta inutil y complicada.
- Además, a menudo existen combinaciones lineales de variables que son interesantes y útiles para propósitos predictivos o comparativos.
- La principal tarea del análisis de correlación canónica es resumir las asociaciones entre los conjuntos $\boldsymbol{x}^{(1)}$ y $\boldsymbol{x}^{(2)}$ en términos de *algunas pocas covarianzas* (o correlaciones) seleccionadas cuidadosamente en lugar de las pq covarianzas en $\boldsymbol{\Sigma_{12}}$

Análisis de Correlación Canónica: Variables canónicas y correlaciones Canónicas

Las combinaciones lineales proporcionan medidas que resumen de manera sencilla un conjunto de variables. Sean

$$U = \boldsymbol{a}' \boldsymbol{x}^{(1)}$$

$$V = \boldsymbol{b}' \boldsymbol{x}^{(2)}$$

para algun par de vectores de coeficientes $\boldsymbol{a}$ y $\boldsymbol{b}$. Entonces, a partir de lo anterior, obtenemos

$$Var(U) = \boldsymbol{a}' Cov(\boldsymbol{x}^{(1)})\boldsymbol{a} = \boldsymbol{a}' \boldsymbol{\Sigma}_{11} \boldsymbol{a}$$

$$Var(V) = \boldsymbol{b}' Cov(\boldsymbol{x}^{(2)})\boldsymbol{b} = \boldsymbol{b}' \boldsymbol{\Sigma}_{22} \boldsymbol{b}$$

$$Cov(U, V) = \boldsymbol{a}' Cov(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})\boldsymbol{b} = \boldsymbol{a}' \boldsymbol{\Sigma}_{12} \boldsymbol{b}$$

Se buscan los vectores de coeficientes $\boldsymbol{a}$ y $\boldsymbol{b}$ tal que

$$Corr(U, V) = \frac{\boldsymbol{a}' \boldsymbol{\Sigma}_{12} \boldsymbol{b}}{\sqrt{\boldsymbol{a}' \boldsymbol{\Sigma}_{11} \boldsymbol{a}} \sqrt{\boldsymbol{b}' \boldsymbol{\Sigma}_{22} \boldsymbol{b}}}$$

sea lo más grande posible

Análisis de Correlación Canónica: Variables canónicas y correlaciones Canónicas

Definimos lo siguiente:

- El primer par de variables canónicas, es el par de combinaciones lineales U_1 y V_1 que tienen varianzas unitarias, que maximizan la correlación anterior.
- El segundo par de variables canónicas es el par de combinaciones lineales U_2 y V_2 que tienen varianzas unitarias, que maximizan la correlación anterior entre todas las elecciones que no están correlacionadas con el primer par de variables canónicas.
- En el k-ésimo paso, el k-ésimo par de variables canónicas, es el par de combinaciones lineales U_k y V_k que tienen varianzas unitarias, que maximizan la correlación anterior entre todas las elecciones que no están correlacionadas con los anteriores $k-1$ pares de variables canónicas. La correlación entre el k-ésimo par de variables canónicas se denomina k-ésima correlación canónica.

Análisis de Correlación Canónica: Variables canónicas y correlaciones Canónicas

Resultado: Suponemos que $p \leq q$ y sean $\boldsymbol{x}^{(1)}$ y $\boldsymbol{x}^{(2)}$ los vectores aleatorios que tienen $Cov(\boldsymbol{x}^{(1)}) = \boldsymbol{\Sigma}_{11_{p\times p}}$, $Cov(\boldsymbol{x}^{(2)}) = \boldsymbol{\Sigma}_{22_{q\times q}}$ y $Cov(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \boldsymbol{\Sigma}_{12_{p\times q}}$, donde $\boldsymbol{\Sigma}$ tiene rango completo. Para los coeficientes $\boldsymbol{a}_{p\times 1}$ y $\boldsymbol{b}_{q\times 1}$, formamos las combinaciones lineales $U = \boldsymbol{a}'\boldsymbol{x}^{(1)}$ y $V = \boldsymbol{b}'\boldsymbol{x}^{(2)}$. Entonces

$$\max_{\boldsymbol{a},\boldsymbol{b}} Corr(U, V) = \rho_1^*$$

se obtiene para las combinaciones lineales (primer par de variables canónicas)

$$U_1 = \boldsymbol{e_1}'\boldsymbol{\Sigma_{11}}^{-1/2}\boldsymbol{x}^{(1)} \quad \text{y} \quad V_1 = \boldsymbol{f_1}'\boldsymbol{\Sigma_{22}}^{-1/2}\boldsymbol{x}^{(2)}$$

Variables canónicas y correlaciones Canónicas

El k-esimo par de variables canónicas, $k = 2, 3, ..., p$,

$$U_k = \boldsymbol{e_k}' \boldsymbol{\Sigma_{11}}^{-1/2} \boldsymbol{x^{(1)}}, \quad V_k = \boldsymbol{f_k}' \boldsymbol{\Sigma_{22}}^{-1/2} \boldsymbol{x^{(2)}}$$

maximiza

$$Corr(U_k, V_k) = \rho_k^*$$

entre aquellas combinaciones lineales no correlacionadas con las anteriores $1, 2, ..., k-1$ variables canónicas.

- Aquí $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$ son los valores propios de $\boldsymbol{\Sigma_{11}}^{-1/2} \boldsymbol{\Sigma_{12}} \boldsymbol{\Sigma_{22}}^{-1} \boldsymbol{\Sigma_{21}} \boldsymbol{\Sigma_{11}}^{-1/2}$, y $\boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_p}$ de tamaño $(p \times 1)$, son los vectores propios asociados
- Las cantidades $\rho_1^{*2}, \rho_2^{*2}, ..., \rho_p^{*2}$ son también los p valores propios mas grandes de la matriz $\boldsymbol{\Sigma_{22}}^{-1/2} \boldsymbol{\Sigma_{21}} \boldsymbol{\Sigma_{11}}^{-1} \boldsymbol{\Sigma_{12}} \boldsymbol{\Sigma_{22}}^{-1/2}$ con correspondientes vectores propios $(q \times 1)$, $\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_p}$.
- Cada $\boldsymbol{f_i}$ es proporcional a $\boldsymbol{\Sigma_{22}}^{-1/2} \boldsymbol{\Sigma_{21}} \boldsymbol{\Sigma_{11}}^{-1/2} \boldsymbol{e_i}$

Variables canónicas y correlaciones Canónicas

Las variables canónicas tienen las siguientes propiedades:

$$
\begin{aligned}
Var(U_k) &= Var(V_k) = 1 \\
Cov(U_k, U_l) &= Corr(U_k, U_l) = 0, \quad k \neq l \\
Cov(V_k, V_l) &= Corr(V_k, V_l) = 0, \quad k \neq l \\
Cov(U_k, V_l) &= Corr(U_k, V_l) = 0, \quad k \neq l
\end{aligned}
$$

para $k, l = 1, 2, ..., p$

Variables canónicas y correlaciones Canónicas

Si las variables originales se estandarizan a $\mathbf{Z}^{(1)} = \left[Z_1^{(1)}, Z_2^{(1)}, \ldots, Z_p^{(1)}\right]^{'}$ y $\mathbf{Z}^{(2)} = \left[Z_1^{(2)}, Z_2^{(2)}, \ldots, Z_q^{(2)}\right]^{'}$, usando los principios anteriores, el k-esimo par de variables canónicas son de la forma

$$U_k = \mathbf{a}_k^{'}\mathbf{Z}^{(1)} = \mathbf{e}_k^{'}\rho_{11}^{-1/2}\mathbf{Z}^{(1)}$$
$$V_k = \mathbf{b}_k^{'}\mathbf{Z}^{(2)} = \mathbf{f}_k^{'}\rho_{22}^{-1/2}\mathbf{Z}^{(2)}$$

Donde

- $Cov(\mathbf{Z}^{(1)}) = \rho_{11}$
- $Cov(\mathbf{Z}^{(2)}) = \rho_{22}$
- $Cov(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}) = \rho_{12} = \rho^{'}{}_{21}$
- $\mathbf{e}_k$ es el vector propio de $\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2}$
- $\mathbf{f}_k$ es el vector propio de $\rho_{22}^{-1/2}\rho_{21}\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1/2}$

Variables canónicas y correlaciones Canónicas

Las correlaciones canonicas ρ_k^* satisfacen

$$Corr(U_k, V_k) = \rho_k^*, \quad k = 1, 2, \ldots, p$$

donde

$$\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$$

son los valores propios distintos de cero de la matriz

$$\rho_{11}^{-1/2} \rho_{12} \rho_{22}^{-1} \rho_{21} \rho_{11}^{-1/2}$$

o equivalentemente los valores propios más grandes de la matriz

$$\rho_{22}^{-1/2} \rho_{21} \rho_{11}^{-1} \rho_{12} \rho_{22}^{-1/2}$$

Variables canónicas y correlaciones Canónicas

Nótese que

$$\begin{aligned}\mathbf{a}_k^{'}(\boldsymbol{x}^{(1)} - \mu^{(1)}) &= a_{k1}(x_1^{(1)} - \mu_1^{(1)}) + a_{k2}(x_2^{(1)} - \mu_2^{(1)}) \\ &\quad + \cdots + a_{kp}(x_p^{(1)} - \mu_p^{(1)}) \\ &= a_{k1}\sqrt{\sigma_{11}}\frac{(x_1^{(1)} - \mu_1^{(1)})}{\sqrt{\sigma_{11}}} + a_{k2}\sqrt{\sigma_{22}}\frac{(x_2^{(1)} - \mu_2^{(1)})}{\sqrt{\sigma_{22}}} \\ &\quad + \cdots + a_{kp}\sqrt{\sigma_{pp}}\frac{(x_p^{(1)} - \mu_p^{(1)})}{\sqrt{\sigma_{pp}}}\end{aligned}$$

donde $Var(x_i^{(1)}) = \sigma_{ii}, i = 1, 2, \ldots, p$.
Por lo tanto, si $\mathbf{a}_k^{'}$ es el vector de coeficientes para la k-ésima variable canónica U_k, construido a partir de las variables originales $x_i^{(1)}$ entonces $\mathbf{a_k^{'}V_{11}^{1/2}}$ es el vector de coeficientes para la k-ésima variable canónica construida a partir de las variables estandarizadas $Z_i^{(1)} = (x_i^{(1)} - \mu_i^{(1)})/\sqrt{\sigma_{ii}}$, donde $\mathbf{V}_{11}^{1/2}$ es la matriz diagonal con i-esimo elemento diagonal $\sqrt{\sigma_{ii}}$.

Variables canónicas y correlaciones Canónicas

- Similarmente $\mathbf{b_k'V_{22}^{1/2}}$ es el vector de coeficientes para la variable canónica construida a partir del conjunto de variables estandarizadas $\mathbf{Z}^{(2)}$.
- En este caso $\mathbf{V}_{22}^{1/2}$ es la matriz diagonal con i-esimo elemento diagonal $\sqrt{\sigma_{ii}} = \sqrt{Var(x_i^{(2)})}$.
- Las correlaciones canónicas *no se modifican* por la estandarización de las variables originales, sin embargo, la elección de los vectores de coeficientes $\mathbf{a}_k, \mathbf{b}_k$ no será única si $\rho_k^{*2} = \rho_{k+1}^{*2}$.

Variables canónicas y correlaciones Canónicas

- La relación entre los coeficientes canónicos de las variables estandarizadas y los coeficientes canónicos de las variables originales se deriva de la estructura especial de la matriz:

$$\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2} \quad o \quad \rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2}$$

- y esto es particular al análisis de correlación canónica. Por ejemplo, en el análisis de componentes principales, si $\mathbf{a}_k^{'}$es el vector de coeficientes para el k-ésimo componente principal obtenido de $\boldsymbol{\Sigma}$ entonces $\mathbf{a}_k^{'}(\boldsymbol{X}-\mu) = a_k^{'}\mathbf{V}^{1/2}\mathbf{Z}$, pero no podemos inferir que $\mathbf{a}_k^{'}\mathbf{V}^{1/2}$ es el vector de coeficientes para el k-ésimo componente principal derivado de ρ.

Ejemplo: Obtención de variables canónicas y correlaciones canónicas para variables estandarizadas

Supongamos que $\boldsymbol{Z^{(1)}} = [Z_1^{(1)}, Z_2^{(1)}]$ y $\boldsymbol{Z^{(2)}} = [Z_1^{(2)}, Z_2^{(2)}]$ son variables estandarizadas y $\boldsymbol{Z} = [\boldsymbol{Z^{(1)}}, \boldsymbol{Z^{(2)}}]$, donde

$$Cov(\boldsymbol{Z}) = \left[\begin{array}{c|c} \boldsymbol{\rho_{11}} & \boldsymbol{\rho_{12}} \\ \hline \boldsymbol{\rho_{21}} & \boldsymbol{\rho_{22}} \end{array}\right] = \left[\begin{array}{cc|cc} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ \hline 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1.0 \end{array}\right]$$

Calcular las variables canónicas y correlaciones canónicas.