Análisis de correspondencias canónicas

viernes, 28 de abril de 2023 14:03

Buscar la asociación de conjuntos de variables.

Ej: Evaluar la relación entre el desempeño escolar y desempeño universitario de alumnos de MACC.

Variables colegio → Notas.

Variables en la universidad → Notas

Variables colegio ⟷ Variables en la universidad

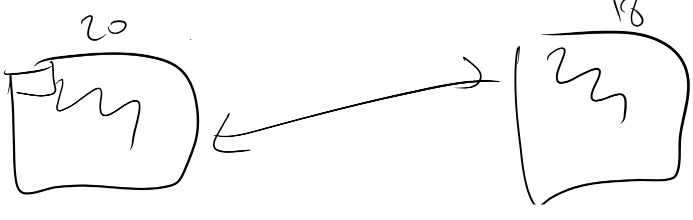
Ej: Notas en asignaturas ciencias sociales ⟷ Notas en asignaturas ciencias naturales.

Metodología general

Estudiar la correlación entre comb. lineales de variables del grupo 1 con comb. lineales de variables del grupo 2.

- Buscamos la combinación lineal con correlación más alta.
- Luego, buscamos la segunda comb. lineal no correlacionada con la anterior con correlación más alta.

Idealmente buscamos reducir información de muchas dimensiones de relaciones entre variables de dos grupos a pocos pares de variantes canónicas.



$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} \\ & \rho_{22} \end{pmatrix}$$

Suponga que el primer grupo tiene p variables y lo representamos con el vector $\mathbb{X}^{(1)}$

Suponga que el segundo grupo tiene q variables y lo representamos con $\mathbb{X}^{(2)}$

Supondremos siempre $p \leq q$

Definimos

$$E(\mathbb{X}^{(1)}) = \mu^{(1)} \qquad cov(\mathbb{X}^{(1)}) = \Sigma_{11}$$

$$E(\mathbb{X}^{(2)}) = \mu^{(2)} \qquad cov(\mathbb{X}^{(2)}) = \Sigma_{22}$$

$$cov(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}) = \Sigma_{12} = \Sigma_{21}'$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad \rho = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix}$$

Obs: Las $p \cdot q$ asociaciones lineales entre $\mathbb{X}^{(1)}$ y $\mathbb{X}^{(2)}$ están en Σ_{12}

Sean

$$U = a'\mathbb{X}^{(1)}, \quad V = b'\mathbb{X}^{(2)}$$

a y b son vectores de coeficientes

$$var(U) = a'\Sigma_{11} a$$

$$\mathrm{Var}(U) = a' \Sigma_{11} a$$
$$\mathrm{Var}(V) = b' \Sigma_{22} b$$
$$\mathrm{Cov}(U,V) = a' \Sigma_{12} b$$

TAREA.

Buscar a, b tal que

$$\mathrm{cov}(U,V) = \frac{a' \Sigma_{12} b}{\sqrt{a'\Sigma_{11} a}\sqrt{b' \Sigma_{22} b}} \quad \text{(pepe)}$$

Sea máxima.

i) El primer par de variantes canónicas son las comb. lineales U_1, V_1 con varianza 1 que maximiza a (pepe)

ii) El segundo par de variantes canónicas son comb. las lineales U_2, V_2 con varianza 1, que maximiza a (pepe) y no están correlacionadas con U_1 y V_1

iii) El K-ésimo par de variantes canónicas son las comb. lineales U_K, V_K que maximizan a (pepe) y no están correlacionadas con los K-1 pares anteriores.

Suponga $p \leq q$ y $\mathbb{X}^{(1)}, \mathbb{X}^{(2)}$ como lo definimos anteriormente tq. $\Sigma = \mathrm{cov}(\mathbb{X})$

Si $U = a' \mathbb{X}^{(1)}$ $V = b' \mathbb{X}^{(2)}$

0.75

Si $U = a' \mathbb{X}$ $\quad V = b' \mathbb{X}$ '

Entonces

1) $$\max_{a,b} \operatorname{cor}(U,V) = \rho_1^{*} \rightarrow 0.75$$

se obtiene con $U_1 = e_1' \Sigma_{11}^{-1/2} \mathbb{X}^{(1)}$

$$V_1 = f_1' \Sigma_{22}^{-1/2} \mathbb{X}^{(2)}$$

El K-ésimo par de variantes canónicas, $K = 2, 3, \ldots, p$

$$U_K = e_K' \Sigma_{11}^{-1/2} \mathbb{X}^{(1)} \qquad V_K = f_K' \Sigma_{22}^{-1/2} \mathbb{X}^{(2)}$$

no correlacionados con los $K-1$ pares anteriores,

donde

$$\rho_1^{*2} \geq \rho_2^{*2} \geq \ldots \geq \rho_p^{*2}$$ son los valores propios

de la matriz

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

Con vectores propios asociados $e_1, e_2, \ldots, e_p$.

Los $f_1, \ldots, f_p$ son los vectores propios de la matriz

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

Estas variantes satisfacen

$$\operatorname{Var}(U_K) = \operatorname{Var}(V_K) = 1 \quad \forall K$$

$$\operatorname{cov}(U_K, U_\ell) = \operatorname{cor}(U_K, U_\ell) = 0 \quad \forall K \neq \ell$$

$$\operatorname{cov}(V_K, V_\ell) = \operatorname{cor}(V_K, V_\ell) = 0 \quad \forall K \neq \ell$$

$\mathrm{cov}(U_k, V_l) = \mathrm{cor}(U_k, V_l) = 0 \qquad \forall\, k \neq l$

Example 10.1 (Calculating canonical variates and canonical correlations for standardized variables) Suppose $\mathbf{Z}^{(1)} = [Z_1^{(1)}, Z_2^{(1)}]'$ are standardized variables and $\mathbf{Z}^{(2)} = [Z_1^{(2)}, Z_2^{(2)}]'$ are also standardized variables. Let $\mathbf{Z} = [\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}]'$ and

$$\mathrm{Cov}(\mathbf{Z}) = \left[\begin{array}{c|c} \boldsymbol{\rho}_{11} & \boldsymbol{\rho}_{12} \\ \hline \boldsymbol{\rho}_{21} & \boldsymbol{\rho}_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 1.0 & .4 & .5 & .6 \\ .4 & 1.0 & .3 & .4 \\ \hline .5 & .3 & 1.0 & .2 \\ .6 & .4 & .2 & 1.0 \end{array}\right]$$

Then

$$\boldsymbol{\rho}_{11}^{-1/2} = \begin{bmatrix} 1.0681 & -.2229 \\ -.2229 & 1.0681 \end{bmatrix}$$

$$\boldsymbol{\rho}_{22}^{-1} = \begin{bmatrix} 1.0417 & -.2083 \\ -.2083 & 1.0417 \end{bmatrix}$$

and

$$\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2} = \begin{bmatrix} .4371 & .2178 \\ .2178 & .1096 \end{bmatrix}$$

ne eigenvalues ρ_1^{*2}, ρ_2^{*2} of $\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}$ are obtained from

$$0 = \begin{vmatrix} .4371 - \lambda & .2178 \\ .2178 & .1096 - \lambda \end{vmatrix} = (.4371 - \lambda)(.1096 - \lambda) - (2.178)^2$$
$$= \lambda^2 - .5467\lambda + .0005$$

er 10 Canonical Correlation Analysis

yielding $\rho_1^{*2} = .5458$ and $\rho_2^{*2} = .0009$. The eigenvector $\mathbf{e}_1$ follows from the vector equation

$$\begin{bmatrix} .4371 & .2178 \\ .2178 & .1096 \end{bmatrix}\mathbf{e}_1 = (.5458)\mathbf{e}_1$$

$\rho_1^* = 0.73$

Thus, $\mathbf{e}_1' = [.8947, .4466]$ and

$$\mathbf{a}_1 = \boldsymbol{\rho}_{11}^{-1/2}\mathbf{e}_1 = \begin{bmatrix} .8561 \\ .2776 \end{bmatrix}$$

1 grupo 1

$\rho_2^* = 0.3$

From Result 10.1, $\mathbf{f}_1 \propto \boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}\mathbf{e}_1$ and $\mathbf{b}_1 = \boldsymbol{\rho}_{22}^{-1/2}\mathbf{f}_1$. Consequently,

$$\mathbf{b}_1 \propto \boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\mathbf{a}_1 = \begin{bmatrix} .3959 & .2292 \\ .5209 & .3542 \end{bmatrix}\begin{bmatrix} .8561 \\ .2776 \end{bmatrix} = \begin{bmatrix} .4026 \\ .5443 \end{bmatrix}$$

We must scale $\mathbf{b}_1$ so that

$$\mathrm{Var}(V_1) = \mathrm{Var}(\mathbf{b}_1'\mathbf{Z}^{(2)}) = \mathbf{b}_1'\boldsymbol{\rho}_{22}\mathbf{b}_1 = 1$$

The vector $[.4026, .5443]'$ gives

$$[.4026, .5443]\begin{bmatrix} 1.0 & .2 \\ .2 & 1.0 \end{bmatrix}\begin{bmatrix} .4026 \\ .5443 \end{bmatrix} = .5460$$

Using $\sqrt{.5460} = .7389$, we take

$$\mathbf{b}_1 = \frac{1}{.7389}\begin{bmatrix} .4026 \\ .5443 \end{bmatrix} = \begin{bmatrix} .5448 \\ .7366 \end{bmatrix} \longrightarrow 2 \text{ grupo } 2$$

The first pair of canonical variates is

$$U_1 = \mathbf{a}_1'\mathbf{Z}^{(1)} = .86Z_1^{(1)} + .28Z_2^{(1)}$$
$$V_1 = \mathbf{b}_1'\mathbf{Z}^{(2)} = .54Z_1^{(2)} + .74Z_2^{(2)}$$

and their canonical correlation is

$$\rho_1^* = \sqrt{\rho_1^{*2}} = \sqrt{.5458} = .74$$