Institución Universitaria
Acreditada en Alta Calidad

Innovación Tecnológica con Sentido Humano

Análisis de Correlación Canónica

Cristian Guarnizo-Lemus
cristianguarnizo@itm.edu.co

Maestria en Automatización y Control Industrial

Contenido

Introducción

- El análisis de correlación canónica (CCA) busca la cantidad de relación (lineal) entre dos conjuntos de variables. También se puede decir que busca identificar y cuantificar las asociaciones entre dos conjuntos de variables.

Contenido

Correlación Múltiple

Asumimos que los 2 conjuntos de variables $\mathbf{y}^\top = [y_1, y_2, \ldots, y_p]$ y $\mathbf{x}^\top = [x_1, x_2, \ldots, x_q]$ son medidos con la misma unidad de muestreo. Consideramos la medida de la correlación entre $\mathbf{y}$ y $\mathbf{x}$. De acuerdo a la correlación múltiple, la covarianza y correlacion muestral entre $y, x_1, x_2, \ldots, x_q$ puede ser resumido en las matrices

$$\mathbf{S} = \begin{bmatrix} s_y^2 & \mathbf{s}_{yx}^\top \\ \mathbf{s}_{yx} & \mathbf{S}_{xx} \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1 & \mathbf{r}_{yx}^\top \\ \mathbf{r}_{yx} & \mathbf{R}_{xx} \end{bmatrix},$$

donde $\mathbf{s}_{yx}^\top = [s_{y1}, s_{y2}, \ldots, s_{yq}]$ contiene las covarianzas muestrales de y con $x_1, x_2, \ldots, x_q$ y $\mathbf{S}_{xx}$ es la matriz de covarianza muestral de x. Las correlaciones se construyen de igual forma.

Correlación Múltiple

La correlación múltiple cuadratica entre y y los x's se puede calcular de las matrices covarianza y correlación particionadas como

$$R^2 = \frac{\mathbf{s}_{yx}^\top \mathbf{S}_{xx}^{-1} \mathbf{s}_{yx}}{s_{yy}} = \mathbf{r}_{yx}^\top \mathbf{R}_{xx}^{-1} \mathbf{r}_{yx}.$$

La correlación múltiple R puede ser definido como la correlación máxima entre y y una combinación lineal de x, $R = \max_{\mathbf{b}} \ r_{y,\mathbf{b}^\top \mathbf{x}}$.

Correlación Múltiple

Para el caso de varios y's y x's, la estructura de la covarianza con dos vectores $\mathbf{y}$ $(y_1, \ldots, y_p)$ y $\mathbf{x}$ $(x_1, \ldots, x_q)$, esta dada como

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{bmatrix},$$

donde $\mathbf{S}_{yy} \in \mathbb{R}^{p \times p}$ es la matriz de covarianza muestral, $\mathbf{S}_{yx} \in \mathbb{R}^{p \times q}$ es la matriz de covarianza muestral entre y y x, y $\mathbf{S}_{xx} \in \mathbb{R}^{q \times q}$ es la matriz de covarianza de x.

Correlación Múltiple

La correlación múltiple cuadratica entre y's y x's se puede calcular de las matrices covarianza y correlación particionadas como

$$R_M^2 = \frac{\left|\mathbf{S}_{yx}^\top \mathbf{S}_{xx}^{-1} \mathbf{S}_{yx}\right|}{\left|\mathbf{S}_{yy}\right|},$$

el cual puede ser reescrito como

$$R_M^2 = \left|\mathbf{S}_{yy}^{-1} \mathbf{S}_{yx}^\top \mathbf{S}_{xx}^{-1} \mathbf{S}_{yx}\right| = \prod_{i=1}^{s} r_i^2,$$

donde $s = \min(p, q)$ y r_i^2 son los valores propios de $\mathbf{S}_{yy}^{-1} \mathbf{S}_{yx}^\top \mathbf{S}_{xx}^{-1} \mathbf{S}_{yx}$. Los r_i se conocen como las correlaciones canónicas. Recordar que

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|, \quad |\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}.$$

Correlación Múltiple

- La mejor medida de asociación es la correlación canónica cuadrática más grande (máximo valor propio) r_1^2 de $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}^{\top}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}$.
- Los otros valores propios proveen medidas de dimensiones suplementarias de relacion lineal entre **y** y **x**.

Contenido

Análisis de Correlación Canónica

- Aquí asumimos que se tienen transformaciones lineales de x y y, esto es $u = \mathbf{a}^\top \mathbf{x}$ y $v = \mathbf{b}^\top \mathbf{y}$.
- Las varianzas de las transformaciones lineales son $\mathbf{S}_{uu} = \mathbf{a}^\top \mathbf{S}_{xx}\mathbf{a}$, $\mathbf{S}_{vv} = \mathbf{b}^\top \mathbf{S}_{yy}\mathbf{b}$, y la covarianza esta dada por $\mathbf{S}_{uv} = \mathbf{a}^\top \mathbf{S}_{xy}\mathbf{b}$.
- De lo anterior la correlación queda definida como

$$R^2 = \frac{\mathbf{a}^\top \mathbf{S}_{xy}\mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{S}_{xx}\mathbf{a}}\sqrt{\mathbf{b}^\top \mathbf{S}_{yy}\mathbf{b}}}$$

Análisis de Correlación Canónica

Se requiere maximizar la correlación

$$\max_{\mathbf{a},\mathbf{b}} \mathbf{a}^\top \mathbf{S}_{xy} \mathbf{b}$$

sujeto a

$$\mathbf{a}^\top \mathbf{S}_{xx} \mathbf{a} = 1, \quad \mathbf{b}^\top \mathbf{S}_{yy} \mathbf{b} = 1.$$

Asumimos que se obtienen los vectores $\mathbf{a}_1$ y $\mathbf{b}_1$ que maximizan el problema anterior. Entonces r_1 es la correlación entre u_1 y v_1. Se puede demostrar que los vectores $\mathbf{a}_1$ y $\mathbf{b}_1$ son los vectores propios.

Análisis de Correlación Canónica

Empleando multiplicadores de Lagrange se tiene

$$\mathcal{L}(\mathbf{a}, \mathbf{b}, \lambda_1, \lambda_2) = \mathbf{a}^\top \mathbf{S}_{xy}\mathbf{b} + \lambda_1 \left(\mathbf{a}^\top \mathbf{S}_{xx}\mathbf{a} - 1\right) + \lambda_2 \left(\mathbf{b}^\top \mathbf{S}_{yy}\mathbf{b} - 1\right)$$

Derivando e igualando a cero se obtienen las siguientes expresiones:

$$\left(\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx} - \lambda^2\mathbf{I}\right)\mathbf{a} = \mathbf{0}, \quad \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{a} = \lambda^2\mathbf{a}$$
$$\left(\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} - \lambda^2\mathbf{I}\right)\mathbf{b} = \mathbf{0}, \quad \mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{b} = \lambda^2\mathbf{b}$$

Finalmente, realizando SVD se obtienen

$$r_i, \quad u_i = \mathbf{a}_i^\top \mathbf{x}, \quad v_i = \mathbf{b}_i^\top \mathbf{y}.$$

Propiedades

- Las correlaciones canónicas son invariantes a los cambios de escala en y o x. Por ejemplo, si la medida de escala es cambiada de pulgadas a centímetros, las correlaciones canónicas no cambiaran.
- La primera correlación canónica r_1 es la correlaci'on máxima entre funciones lineales de $\mathbf{y}$ y $\mathbf{x}$. Entonce, r_1 excede el valor de la correlación entre cualquier x y y.

Contenido

Aplicaciones

- Análisis de relación lineal entre variables.
- Test de significancia.

Relación lineal

Suponga que de la base de datos Marcas de Carros se tienen las matrices de covarianza entre x (Precio, Valor) y y (Economia, Servicio, Diseño, Carro deportivo, Seguridad, Fácil manejo).

$$\mathbf{S}_{xx} = \begin{pmatrix} 1.41 & -1.11 \\ -1.11 & 1.19 \end{pmatrix}, \quad \mathbf{S}_{xy} = \begin{pmatrix} 0.78 & -0.71 & -0.90 & -1.04 & -0.95 & 0.18 \\ -0.42 & 0.82 & 0.77 & 0.90 & 1.12 & 0.11 \end{pmatrix}$$

$$\mathbf{S}_{yy} = \begin{pmatrix} 0.75 & -0.23 & -0.45 & -0.42 & -0.28 & 0.28 \\ -0.23 & 0.66 & 0.52 & 0.57 & 0.85 & 0.14 \\ -0.45 & 0.52 & 0.72 & 0.77 & 0.68 & -0.10 \\ -0.42 & 0.57 & 0.77 & 1.05 & 0.76 & -0.15 \\ -0.28 & 0.85 & 0.68 & 0.76 & 1.26 & 0.22 \\ 0.28 & 0.14 & -0.10 & -0.15 & 0.22 & 0.32 \end{pmatrix}$$

Relación lineal

De las matrices anteriores calculamos

$$\mathbf{K} = \mathbf{S}_{xx}^{1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}^{1/2}.$$

Calculamos la SVD de **K**

$$\mathbf{K} = \mathbf{A}\mathbf{D}\mathbf{B}^\top = [\mathbf{a}_1, \mathbf{a}_2] \begin{bmatrix} \lambda_1^{1/2} & 0 \\ 0 & \lambda_2^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1^\top \\ \mathbf{b}_2^\top \end{bmatrix}$$

donde $r_1 = \lambda_1^{1/2} = 0.98$ y $r_2 = \lambda_2^{1/2} = 0.89$. Usando $\mathbf{a}_1$ y $\mathbf{b}_1$ se tiene

$$u_1 = \mathbf{a}_1^\top \mathbf{x} = 1.602x_1 + 1.686x_2$$

$$v_1 = \mathbf{b}_1^\top \mathbf{y} = 0.568y_1 + 0.544y_2 - 0.012y_3 - 0.096y_4 - 0.014y_5 + 0.915y_6.$$

Test de significancia

En este caso consideramos como hipótesis nula la independencia,

$$H_0 : \Sigma_{yx} = \mathbf{0}.$$

Esto indica que la covarianza de cada y_i con cada x_j es cero, y todas las correlaciones son cero también. El test se puede realizar co la verosimilitud de Wilks:

$$T^{2/n} = \left|\mathbf{I} - \mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\right| = \prod_{i=1}^{s}(1 - r_i).$$

Sin embargo, este estadístico tiene una distribución compleja. Se puede aproximar a

$$-(n - (p + q + 3)/2) \log \prod_{i=1}^{k}(1 - r_i) \sim \chi^2_{\alpha,pq}.$$

Test de significancia

Para el ejemplo de la base de datos Marcas de Carros, se tienen $n = 40$ personas que han calificado los carros de acuerdo a las categorías con $p = 2$ y $q = 6$. Recordemos los coeficientes de correlación canónica fueron $r_1 = 0.98$ y $r_2 = 0.89$. Empleando el estadístico anterior

$$-\{40 - (2+6+3)/2\} \log\left\{\left(1 - 0.98^2\right)\left(1 - 0.89^2\right)\right\} = 165.59 \sim \chi^2_{0.01,12}$$

el cual es altamente significativo (para 99% se tiene un χ^2_{12} de 26.22). La hipótesis de no correlación entre las variables se rechaza.

Institución Universitaria
Acreditada en Alta Calidad

Innovación Tecnológica con Sentido Humano

Referencias

- W. K. Härdle and L. Simar. “Applied Multivariate Statistical Analysis”, 2019. Capitulo 16.
- Rencher. “Methods of Multivariate Analysis”, 2012. Capitulo 11.