4.7 Análisis de correlación canónica 367

4.7.1 Determinación de las variables canónicas 369

4.7.2 Cálculo de las correlaciones canónicas para los conjuntos de variables de la prueba de lenguaje y de la prueba de matemáticas 373

## 4.7 Análisis de correlación canónica

Las correlaciones canónicas constituyen una generalización de las correlaciones simples y múltiples. Las correlaciones simples estiman la relación lineal existente entre dos variables aleatorias X y Y. Las correlaciones múltiples estiman la relación lineal existente entre dos conjuntos de variables aleatorias artificiales Uk y Vk las cuales se construyen a partir de dos grupos de variables aleatorias observables.

Sean **X**(1) = (X1, X2, ..., Xp) y **X**(2) = (X1, X2, ..., Xq) dos vectores observables aleatorios p-variado y q-variado respectivamente con p<q, y siendo Σ12  la matriz de varianzas y covarianzas de estos vectores, se construyen las variables artificiales U y V de la siguiente manera:

U **= aT X(1)**

V **= bT X(2)**

Las variables artificiales U y V deben satisfacer las siguientes condiciones:

Var ( U ) = **a**T Cov ( **X**(1) ) **a** = **a**T **Σ**11 **a**

Var ( V ) = **a**T Cov ( **X**(2) ) **a** = **b**T **Σ**11 **b**

Cov ( U , V ) = **a**T Cov ( **X**(1) , **X**(2) ) **b** = **a**T **Σ**12 **b**

 Los vectores a y b se deben determinar de tal forma que maximicen la correlación entre las variables canónicas U y V.



Entonces en base a lo descrito anteriormente se puede definir a la variables canónicas como sigue:

 El primer par de variables canónicas, es el par de combinaciones lineales U1 y V1 que tienen varianza igual a uno y que maximizan la correlación Corr (U1, V1).

El segundo par de variables canónicas es el par de combinaciones lineales U2 y V2 que tienen varianza igual a uno y que maximizan la correlación Corr (U2, V2) y que no están correlacionadas con el primer par de variables aleatorias canónicas

El k-ésimo par de variables canónicas es el par de combinaciones lineales Uk y Vk que tienen varianza igual a uno y que maximizan la correlación Corr (Uk, Vk) y que no están correlacionadas con ninguno de los pares anteriores.

### 4.7.1 Determinación de las variables canónicas

Dados los vectores aleatorios **X**(1) ∈ Rp y **X**(2) ∈ Rq, donde p < q, entonces se define lo siguiente:

1. Las matrices de varianza y covarianza

Cov ( **X**(1) ) = **Σ**11

Cov ( **X**(2) ) = **Σ**22

1. Los vectores de media

E ( **X**(1) ) = **μ**(1)

E ( **X**(2 ) ) = **μ**(2)

1. La covarianza entre **X**(1) y **X**(2)

Cov ( **X**(1), **X**(2) ) = **Σ**12 = **Σ**21T

Siendo los vectores de coeficientes **a** ∈ Rp y **b** ∈ Rq de las combinaciones lineales U = **a**T **X**(1) y V = **b**T **X**(2) entonces,

maxa b Corr (U, V) = ρ1\*

a partir de lo cual se construye el primer par de variables canónicas

U1 =  **X**(1) y V1 =  **X**(2),

El k-ésimo par de variables canónicas para k = 2, 3, ... , p es

Uk =  **X**(1) y Vk =  **X**(2)

que maximiza Corr (Uk , Vk) = ρk\*, donde las k-ésimas variables canónicas no están correlacionadas con las anteriores k-1 variables canónicas.

El propósito es encontrar un par de vectores Uk y Vk, tal que el coeficiente de correlación entre estos vectores sea el mayor posible para k, l = 1,2, ... ,p. Donde , representan los valores propios de la matriz con vectores propios e1,e2,...,ep de tamaño (px1), los cuales son iguales a los valores propios de la matriz  con vectores propios asociados f1,f2,...,fp. de tamaño (qxp).

Donde  =  y  =  ; los  y son los vectores propios de la matriz Σ ∈ M(p+q)\*(p+q) que como se explica posteriormente se forma por la unión de las matrices **Σ**11, **Σ**22, **Σ**12 y **Σ**21T, y  y  son las matrices de varianza y covarianza de cada uno de los grupos de variables observables elevados a la -1/2.

El objetivo de la correlación canónica es medir la fuerza de la relación entre dos conjuntos de variables. Para el análisis es conveniente definir el vector **X** que esta formado por la unión de dos vectores de dos grupos de variables observables consideradas en el análisis como se muestra a continuación:



El vector de medias del vector **X**∈ Rp+q está formada por el valor esperado de cada uno de los vectores componentes como se muestra a continuación:



La matriz de varianzas y covarianzas del vector **X** se calcula de la siguiente manera:



La expresión anterior se puede expresar como la unión de las matrices de varianzas y covarianzas de los vectores **X**(1) y **X**(2) respectivamente y la matriz de covarianzas entre estos mismos vectores como se muestra a continuación:



La matriz **Σ** ∈ M(p+q) x( p+q) es simétrica y definida positiva, se puede abreviar de la siguiente manera:



### 4.7.2 Correlación canónica para los conjuntos de variables de la prueba de lenguaje y de la prueba de matemáticas

El procedimiento descrito en la sección 4.6.1 para la determinación de las variables canónicas, se aplicó a los conjuntos de variables de la prueba de lenguaje y de la prueba de matemáticas. Siendo el vector **X**(1) ∈ Rp el que corresponde a las variables de lenguaje, donde p=26 y el vector **Y**(2) ∈ Rq el que corresponde a las variables de matemáticas, donde q=30. Los resultados de las correlaciones canónicas y los coeficientes de las variables canónicas se los obtuvo utilizando el software de aplicaciones estadísticas SPSS (Social purpose statistical system) versión 8.0, el procedimiento es el siguiente se abre la matriz de datos, se va al menú principal, se abre la ventana de New y se entra a la ventana Sintax en la que se realiza la siguiente función:

 INCLUDE 'Canonical Correlation.sps'

CANCORR SET 1 = Nombres de las variables de la prueba de lenguaje (ó códigos de las variables) /Set 2 = Nombres de las variables de la prueba de matemáticas (ó códigos de las variables). Luego desde el menú inicial se corre esta función con la que se obtiene los coeficientes de las variables canónicas y los coeficientes de correlación canónica.

Las correlaciones canónicas ρk\* obtenidas entre cada par de las variables canónicas Uk , Vk para k = 1, 2, ... ,26 son mostradas en la tabla CXLVII. A continuación se analizan los cuatro primeros pares de variables canónicas, que son los que tienen los mayores coeficientes de correlación canónica.

**Tabla CXLVII**

**Correlaciones canónicas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Correlación** | **Variable** | **Correlación** |
| 1 | 0,745 | 14 | 0,362 |
| 2 | 0,706 | 15 | 0,35 |
| 3 | 0,657 | 16 | 0,288 |
| 4 | 0,636 | 17 | 0,268 |
| 5 | 0,617 | 18 | 0,253 |
| 6 | 0,615 | 19 | 0,245 |
| 7 | 0,553 | 20 | 0,231 |
| 8 | 0,546 | 21 | 0,182 |
| 9 | 0,516 | 22 | 0,168 |
| 10 | 0,5 | 23 | 0,143 |
| 11 | 0,441 | 24 | 0,088 |
| 12 | 0,428 | 25 | 0,058 |
| 13 | 0,383 | 26 | 0,03 |

**El primer par de variables canónicas es U1 y V1**

U1 = -0.517X1 + -0.419X2 + 0.066X3 -0.237 X4 -0.361X5 -0.261X6 -0.218X7 -0.314X8 -0.478X9 -0.31X10 -0.274X11 -0.502X12 -0.452X13 -0.477X14 -0.307X15 -0.309X16 -0.429X17 -0.425X18 -0.161X19 -0.216X20 - 0.418X21 -0.478X22 -0.414X23 -0.385X24 -0.442X25 - 0.487X26

V1 = -0.449Y1 -0.18Y2 + 0.236Y3 -0.27 Y4 - 0.419Y5 - 0.172Y6 - 0.469Y7 - 0.303Y8 - 0.171Y9 - 0.033Y10 - 0.249Y11 - 0.093Y12 - 0.504Y13 - 0.249Y14 - 0.446Y15 - 0.402Y16 - 0.332Y17 - 0.518Y18 - 0.409Y19 + 0.083Y20 - 0.255Y21 - 0.335Y22 -0.255Y23 - 0.253Y24 - 0.319Y25 - 0.458Y26 -0.317 Y27 - 0.074Y28 +0.039Y29 - 0.308Y30

Si se analizan los coeficientes (de las variables observables del conjunto de la prueba de lenguaje y de la prueba de matemáticas) canónicos asociados a la primera correlación canónica resulta que X1 (identificación de diptongos)= -0.517 y X12 (identificación de palabras diminutivas) = -0.502, Y13 (ejercicio de divisibilidad) = -0.504 y Y18 (propiedades de conjuntos) = -0.518 son las variables que más aportación ofrecen a la correlación existente ρ\*1 =0.745. Tal que la Var (U1 ) = 1, Var (V1 ) = 1 y la Cov (U1 , V1 ) = 0.745.

**El segundo par de variables canónicas es U2 y V2**

U2 = 0.036X1 -0.035X2 + 0.183X3 +0.011 X4 -0.217X5 -0.12X6 -0.011X7 -0.037X8 -0.282X9 +0.035X10 +0.296X11 - 0.013X12 + 0.017X13 +0.063X14 +0.221X15 +0.286X16 +0.465X17 +0.228X18 +0.173X19 -0.082X20 - 0.108X21 -0.08X22 +0.175X23 +0.117X24 +0.417X25 + 0.13X26

V2 = 0.12Y1 +0.287Y2 + 0.193Y3 +0.07 Y4 + 0.281Y5 - 0.169Y6 - 0.295Y7 - 0.118Y8 + 0.343Y9 + 0.159Y10 - 0.182Y11 + 0.174Y12 + 0.061Y13 - 0.253Y14 - 0.073Y15 - 0.087Y16 - 0.057Y17 + 0.146Y18 + 0.05Y19 + 0.149Y20 + 0.258Y21 + 0.03Y22 -0.061Y23 - 0.003Y24 + 0.002Y25 + 0.157Y26 +0.205 Y27 + 0.302Y28 +0.379Y29 - 0.031Y30

Los coeficientes del segundo par de variables canónicas de la prueba de lenguaje y de la prueba de matemáticas más significativos son X17 (identificación del sujeto de la oración)= 0.465 y X25 (palabras tildadas correctamente) = 0.417, Y9 (ejercicio de valor absoluto) = 0.343 y Y29 (ejercicio de probabilidad) = 0.379, estas son las variables que más aportación ofrecen a la correlación existente entre U2 y V2, donde ρ\*2 =0.706. Tal que la Var (U2 ) = 1, Var (V2 ) = 1 y la Cov (U2 , V2 ) = 0.706.

**El tercer par de variables canónicas es U3 y V3**

U3 = -0.014X1 + 0.008X2 + 0.004X3 +0.354X4 -0.254X5 -0.035X6 +0.015X7 -0.102X8 +0.159X9 +0.083X10 -0.034X11 - 0.13X12 - 0.206X13 -0.206X14 +0.129X15 +0.036X16 +0.142X17 -0.018X18 +0.096X19 +0.051X20 - 0.016X21 +0.032X22 +0.044X23 -0.342X24 +0.004X25 + 0.299X26

V3 = 0.08Y1 -0.078Y2 + 0.191Y3 +0.086 Y4 - 0.119Y5 - 0.002Y6 - 0.007Y7 - 0.038Y8 + 0.051Y9 + 0.016Y10 - 0.215Y11 + 0.451Y12 - 0.106Y13 + 0.099Y14 + 0.043Y15 - 0.135Y16 + 0.104Y17 - 0.054Y18 - 0.054Y19 + 0.264Y20 + 0.188Y21 - 0.024Y22 +0.188Y23 - 0.167Y24 + 0.339Y25 + 0.214Y26 +0.292 Y27 - 0.24Y28 -0.324Y29 - 0.057Y30

En el tercer par de variables canónicas, los coeficientes canónicos más significativos son los asociados a las variables X4 (identificación palabras agudas)= 0.354 y X24 (corrección de sintaxis) = -0.342, Y12 (ejercicio de divisibilidad) = 0.451 y Y25 (factorización de dos polinomios) = 0.339, estas son las variables que más aportación ofrecen a la correlación existente entre U3 y V3, donde ρ\*3 =0.657. Tal que la Var (U3) = 1, Var (V3 ) = 1 y la Cov (U3 , V3 ) = 0.657.

**El cuarto par de variables canónicas es U4 y V4**

U4 = -0.431X1 - 0.488X2 - 0.31X3 -0.158X4 +0.147X5 +0.083X6 +0.082X7 +0.085X8 +0.134X9 -0.074X10 +0.092X11 + 0.115X12 - 0.165X13 -0.279X14 +0.149X15 +0.231X16 -0.03X17 -0.247X18 -0.049X19 +0.296X20 + 0.179X21 +0.059X22 -0.212X23 -0.012X24 +0.132X25 + 0.005X26

V4 = 0.064Y1 +0.25Y2 - 0.337Y3 +0.204 Y4 + 0.071Y5 - 0.103Y6 - 0.166Y7 - 0.129Y8 + 0.112Y9 + 0.112Y10 - 0.096Y11 + 0.185Y12 + 0.096Y13 + 0.26Y14 - 0.002Y15 + 0.026Y16 - 0.36Y17 - 0.081Y18 - 0.166Y19 + 0.139Y20 - 0.127Y21 - 0.309Y22 -0.316Y23 - 0.02Y24 - 0.015Y25 + 0.208Y26 +0.035 Y27 - 0.342Y28 +0.173Y29 - 0.131Y30

Los coeficientes canónicos más altos en el cuarto par de variables canónicas, son los asociados a las variables X1 (identificación diptongos)= -0.431 y X2 (identificación de triptongos) = -0.488, Y3 (resta de números enteros) = -0.337 y Y23 (área del círculo) = -0.316, estas son las variables que más aportación ofrecen a la correlación existente entre U4 y V4, donde ρ\*4 =0.636. Tal que la Var (U4 ) = 1, Var (V4 ) = 1 y la Cov (U4 , V4 ) = 0.657.