Universidad Nacional Autónoma de México

Facultad de Ciencias

Análisis Canónico y técnicas relacionadas

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
A C T U A R I O

PRESENTAN:
Carlos González Munguía
Lany Muñoz Mota

DIRECTOR DE TESIS:
Mat. Margarita E. Chávez Cano

2012

Datos de los alumnos:

González
Munguía
Carlos
Teléfono: 26-33-62-11
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
Número de cuenta: 303249272

Muñoz
Mota
Lany
Teléfono: 56-85-58-50
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
Número de cuenta: 303179580

Datos del tutor:
Matemática
Margarita Elvira
Chávez
Cano

Datos del sinodal 1:
Doctora
María del Pilar
Alonso
Reyes

Datos del sinodal 2:
Mestro en Ciencias
José Antonio
Flores
Díaz

Datos del sinodal 3:
Mestro en Ciencias
José Salvador
Zamora
Muñoz

Datos del sinodal 4:
Actuario
Harim
García
Lamont

Datos del trabajo escrito:
Título: Análisis Canónico y técnicas relacionadas
Número de páginas: 169
Año: 2012

Índice general

Introducción

El Análisis Canónico **(AC)** forma parte de un conjunto de técnicas estadísticas multivariadas, considerado como un método que es a la vez descriptivo y exploratorio, así como de inferencia y de predicción, destinado a clasificar individuos si es que no existen o a explicar clasificaciones ya hechas; es caracterizado por poder considerar todos los tipos de variables, es decir, categóricas o numéricas. Dichas clasificaciones se realizan a partir de la relación que existe entre las características de los individuos entre dos o más grupos de variables.

Se dice que el **AC** es una técnica que permite estudiar la estructura de varios grupos de individuos con respecto a un conjunto de variables que impone como condición que la variabilidad entre los diferentes grupos sea máxima y la variabilidad dentro de los grupos sea mínima, además permite resumir una relación compleja y proveer un método útil de reducción de dimensionalidad de un problema.

El **AC** está relacionado con diferentes técnicas multivariadas, considerándose el caso general o bien puede decirse que: el Análisis de Regresión Múltiple, MANOVA*, Análisis Discriminante, Análisis de Componentes Principales y Factores se muestran como un caso particular del **AC**. Es por esto que se necesita desarrollar la teoría matemática para poder sustentar todo lo que involucra el **AC**,

* por sus siglas en ingles: Multivariate Analysis of Variance

como las variables y correlaciones canónicas, definición correcta de los supuestos estadísticos requeridos, encontrar regiones de confianza y realizar diferentes pruebas de significancia para tener la interpretación adecuada de los resultados bajo el diseño de recomendaciones que permita evitar posibles errores.

Cabe señalar que al ser el caso general de una "familia" de técnicas multivariadas, es ampliamente utilizado en distintas áreas de investigación, tales como son estudios biológicos, geográficos, económicos, actuariales, psicológicos, educacionales, entre otros. Hoy en día éstas diferentes ramas de la investigación tienen la necesidad de encontrar una relación entre más de dos conjuntos de variables, al tener sustentado el desarrollo del **AC**, es fácil encontrar una combinación de herramientas que permita analizar y describir esta relación. Dicho análisis lleva al caso general del **AC** conocido como Análisis Canónico Generalizado $C(2n+1)$.

Objetivos

El **AC** está desarrollado con álgebra de matrices, por lo que uno de los objetivos es: *formalizar la derivación del **AC**, explicando todo el procedimiento*, en sentido en que se entiendan todos los resultados que son necesarios en el análisis.

Al tener las variables y correlaciones canónicas, se desea *saber si los resultados son estadísticamente significativos, utilizar diferentes pruebas de significancia estadística* y *explicar su desarrollo.* Inmediatamente, al derivar las pruebas estadísticas, surge la pregunta acerca de la estimación por intervalos de confianza, pero al ser el **AC** una técnica multivariada, otro de los objetivos es: *desarrollar las regiones de confianza*, así como, *hacer las interpretaciones correspondientes.*

El **AC** utiliza el menor número de restricciones sobre los conjuntos de variables que son utilizadas o que necesitan ser estudiadas, ya que no requiere de un gran

número de supuestos, es por esto que se *desea identificar los posibles errores, causa por los que ocurren y su solución*; por lo que se pretende *dar a conocer recomendaciones para llevar a cabo el* $\boldsymbol{AC}$ *adecuadamente.*

Al ser el **AC** una técnica multivariada general, se espera *mostrar las técnicas multivariadas relacionadas al* $\boldsymbol{AC}$.

Otro de los objetivos es: *justificar las ventajas que posee el* $\boldsymbol{AC}$ *ante otras técnicas multivariadas.*

Todos los resultados del presente trabajo, se derivan utilizando dos conjuntos de variables, X y Y. Por lo que se desea *derivar un Análisis Canónico Generalizado*, es decir, derivar el análisis utilizando más de dos conjuntos de variables.

El último objetivo es: *ejemplificar el* $\boldsymbol{AC}$ por medio de un problema de investigación utilizando todos los resultados obtenidos.

En el capítulo 1, se plantea el principal objetivo del **AC**, que es encontrar si existe relación entre los distintos grupos de variables de un conjunto de datos. En un análisis simple se tienen dos conjuntos de variables y posteriormente se realiza una comparación con el Análisis de Componentes Principales; el análisis de componentes principales es una técnica multivariada descriptiva, que tiene la finalidad de reducir la dimensión de un conjunto de variables multivariadas mediante el estudio de la relación entre las variables originales, geométricamente es una rotación y traslación de los ejes de las variables originales en un sistema coordenado.

Posteriormente se presenta el desarrollo de las variables y correlaciones canónicas a partir de las transformaciones lineales $\boldsymbol{u} = \boldsymbol{Lx}$ y $\boldsymbol{v} = \boldsymbol{My}$. Una vez que se desarrollen las variables y las correlaciones canónicas es necesario entender que los procedimientos estadísticos siempre requieren de ciertos supuestos para cumplir los objetivos inferenciales.

Una vez que se tienen todos los supuestos necesarios, se puede entonces desarrollar la teoría para las correlaciones y variables canónicas muestrales a partir del desarrollo de las transformaciones lineales anteriores.

En el capítulo 2, se desarrollan las pruebas de significancia estadística para el **AC** partiendo de suponer la existencia de al menos una relación entre los conjunto de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$. Las pruebas se utilizan para contrastar cuántas relaciones son estadísticamente significativas, para después probar que el resto de las relaciones que no se consideran, sean no significativas para la construcción de las combinaciones lineales del **AC**. Después se plantea la pregunta: *¿Qué tanta variabilidad del conjunto de variables dependientes puede ser representada por el conjunto de variables independientes?*, ya que es de suma importancia para medir la variabilidad entre conjuntos; ésta pregunta es contestada definiendo el coeficiente de traza, desarrollado por Hopper (1959,1962), que se obtiene a partir de la matriz correlaciones canónicas muestrales.

En el capítulo 3, se muestra el desarrollo matemático para la construcción de las regiones de confianza. Se parte de una muestra de g poblaciones de las cuales se han tomado g muestras de tamaños $n_1, n_2, n_3, \ldots, n_g$ respectivamente y entonces se utilizan los n_i individuos del $i - ésimo$ grupo y se define a x_{ij} el vector de las p mediciones tomadas sobre el $j - ésimo$ individuo en el $i - ésimo$ grupo para entonces encontrar las regiones de confianza.

Posteriormente se obtienen las regiones de confianza alternativas que son casos particulares de las regiones de confianza y se crean principalmente por los inconvenientes gráficos que llegan a ocurrir. Las regiones pueden tener como centroides a los *individuos medios* y las regiones de confianza para éstos, se desarrollan de la misma manera que las regiones de confianza antes desarrolladas, pero centradas en los individuos medios de sus respectivos grupos; por lo que también se busca que la dispersión entre los grupos sea máxima con relación a la

dispersión dentro de cada grupo.

En el capítulo 4, se muestran los principales errores y causas posibles por las que ocurren dichos errores al utilizar el ***AC***, algunos de los errores se relacionan con hipótesis mal planteadas, errores de calculos o aplicación de otra técnica multivariada no adecuada para el análisis a realizar. Algunos otros errores son causados por el mal manejo de los paquetes estadísticos, ya que no arrogan estadísticas necesarias para la interpretación adecuada del ***AC***, esto ocurre principalmente porque el ***AC***, es una técnica multivariada no tan popular.

En capítulo 5, se plantea una serie de recomendaciones para llevar a cabo la selección de variables y la verificación de los objetivos en el **AC**. Llevar a cabo el diseño del **AC** considerando, el *error de medición* de las variables, su *aportación en el análisis*, los *datos faltantes* y los *valores extremos*. Al contrario de como se vio anteriormente en la verificación de los supuestos requeridos, algunos de éstos no son estrictamente necesarios, pero la interpretación de los resultados en el **AC**, mejora si se consideran los supuestos de Linealidad y Normailidad Multivariada.

Se muestra como realizar adecuadamente la interpretación de las variables canónicas con ayuda de los diferentes *coeficientes canónicos*; pero el uso de estos coeficientes adicionales es un tanto problemática por la aceptación de los términos o definiciones que deben emplearse para hacer referencia a conceptos diferentes, entonces, para utilizar un mismo concepto se interpretan y se muestran los coeficientes. Sin embargo se debe tener presente que el análisis canónico tiene ciertas limitaciones y es necesaria la validación de los resultados para asegurar que sean específicos y correctos.

En el capítulo 6, se describen varias técnicas estadísticas para estudiar las diferentes preguntas y objetivos que se plantean dadas las relaciones entre dos o más conjuntos de variables y especificar las diferencias en que cada técnica estadística es más apropiada, argumentar cómo identificar aquellos problemas en los cuales el **AC** es adecuado para el problema en cuestión, pero que es generalmente ignorado. Se muestra una breve comparación de la construcción y similitudes que tiene el Análisis Canónico con el Análisis de Varianza Multivariado, el Análisis de Regresión Múltiple, el Análisis Discriminante, el Análisis de Componentes Principales, el Biplot Canónico y por último se presenta al Análisis Canónico como el caso general de todas las técnicas antes mencionadas.

En el capítulo 7, dada la construcción de las variables canónicas para dos conjuntos de variables, se muestra el desarrollo inductivo que lleva a generalizar el Análisis Canónico, se presenta el problema de no solo tener dos conjuntos de variables, ahora se tienen tres conjuntos de variables aleatorias: $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ y $\boldsymbol{X_3}$ de manera que la dimensión de cada variable sea: n_1, n_2 y n_3 respectivamente, con $n_1 \leqslant n_2$; se define el Análisis Canónico Parcial, partiendo del supuesto de que existe relación entre el tercer conjunto con los otros dos y que se puede realizar una regresión adecuada para considerar los dos conjuntos de residuales como los conjuntos de variables para realizar el Análisis Canónico. Posteriormente se consideran cinco conjuntos de variables para desarrollar el Análisis Canónico Biparcial y siete conjuntos para el Análisis Canónico Biparcial G_2 (**C7**); mediante un proceso inductivo, se desarrolla el Análisis Canónico Generalizado ($C2n+1$).

En el capítulo 8, se muestra a través de un ejemplo de un análisis en donde un investigador ha recopilado 600 datos sobre tres variables psicológicas, cuatro variables académicas y el género; se busca conocer cómo el conjunto de variables psicológicas se relacionan con las variables académicas y de género, en particular, determinar el número de variables canónicas estadísticamente significativas.

Se realiza el análisis descriptivo ya que es una ayuda para comprender mejor todos los resultados. A continuación se realiza el análisis canónico y se desarrolla toda la parte teórica. Se calcula la matriz de correlación, se obtienen todos los coeficientes descritos con detalle así como su interpretación, se realizan pruebas estadísticas y se concluye.

Capítulo 1

Desarrollo del Análisis Canónico

1.1. Ideas generales en el análisis canónico

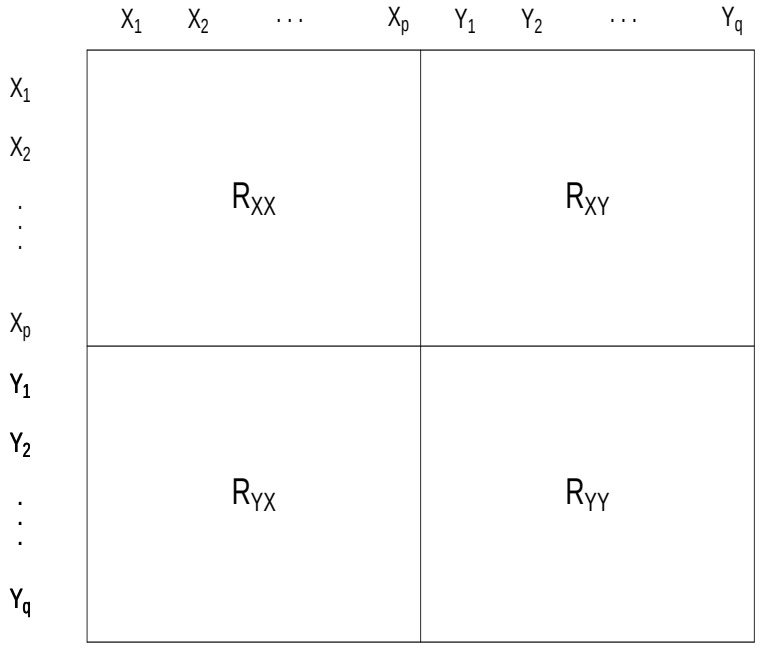
El objetivo del **AC** es encontrar si existe relación entre los distintos grupos de variables de un conjunto de datos, es decir, se determinan combinaciones lineales de las variables de tal manera que la dispersión sea máxima entre los grupos y a su vez sea mínima dentro de estos. Es de esperar que las variables sean adecuadamente representadas en subespacios de dos o tres dimensiones definidas por los primeros dos o tres vectores canónicos; es importante resaltar que la distancia euclidiana es apropiada para interpretar todo el desarrollo que conlleva el **AC**, ya que los vectores canónicos son ortogonales entre sí.

En un análisis simple se tienen dos conjuntos de variables, predictoras o independientes ($\boldsymbol{X}$) de dimensión p y otro conjunto de variables respuesta o dependientes ($\boldsymbol{Y}$) de dimensión q, donde estas últimas pueden ser o no aleatorias, entonces el conjunto formado por las variables de X y de Y, es un conjunto de $p+q$ variables. Se supone que cada una de las variables está correlacionada con cada una de las $p+q-1$ variables restantes, dichas correlaciones están representadas en una matriz cuadrada simétrica de dimensión $p+q$, llamada matriz de correlación, denotada por $\boldsymbol{R}$.

La matriz de correlación, puede verse como una matriz particionada en cuatro submatrices; donde $\boldsymbol{R_{XX}}$ es la submatriz que corresponde a las correlaciones dentro del conjunto $\boldsymbol{X}$, $\boldsymbol{R_{YY}}$ es la submatriz correspondiente a las correlaciones dentro del conjunto $\boldsymbol{Y}$ y las submatrices $\boldsymbol{R_{XY}}$ y $\boldsymbol{R_{YX}}$ son las correlaciones entre las variables del conjunto $\boldsymbol{X}$ y $\boldsymbol{Y}$. No es difícil deducir que la submatriz $\boldsymbol{R_{XY}}$ es igual a $\boldsymbol{R'_{YX}}$ y recíprocamente, así se tienen dos tipos de información dentro de $\boldsymbol{R}$, el patrón de relación dentro de cada conjunto y el patrón de correlaciones entre los dos conjuntos.

La figura 1 representa la estructura de la matriz $\boldsymbol{R}$. Suponiendo que $\boldsymbol{R_{XY}}$ contiene al menos un coeficiente distinto de cero, el problema es entonces determinar el número de correlaciones significativas entre conjuntos.

Figura 1: ***Representación gráfica de la matriz de correlación particionada.***



Una vez que se encuentra al menos una correlación distinta de cero entre el conjunto de variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ y $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_q)$, se determina que existe una relación entre los conjuntos de interés, se calcula la correlación

de todas las combinaciones de las Xs dada una combinación lineal de las Ys, de tal manera que la primera combinación lineal resultante de $\boldsymbol{X}$ tiene la máxima correlación con la combinación lineal de $\boldsymbol{Y}$. Esta correlación es *la primera correlación canónica* (ρ_1) y la correspondiente pareja de combinaciones de Xs y Ys es *la primera variable canónica.*

La segunda correlación y variables canónicas están definidas como la primera, pero esta segunda pareja no está correlacionada con la primera pareja; cada una de las siguientes correlaciones y variables canónicas, están definidas análogamente y por la construcción de las variables, el procedimiento garantiza que no existen más parejas que el número de variables en el conjunto más pequeño, esto es, a lo más habrá tantas parejas como el mínimo entre p y q.

Sea $r = min\{p, q\}$, entonces estas r parejas de variables canónicas cumplen con las siguientes propiedades:[1]

i. Todas las correlaciones entre ellas son cero, a excepción de las que existen entre las correspondientes parejas.

ii. Las correlaciones entre las correspondientes parejas forman una sucesión decreciente.

Una condición inicial que se le pide a las variables para realizar el procedimiento, es que éstas estén estandarizadas con media cero y varianza uno.

La relación entre los dos conjuntos de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$ está dada por estas *variables canónicas*, esto es posible si al realizarse pruebas de hipótesis hay suficiente evidencia para concluir que existe al menos una relación entre las variables de $\boldsymbol{X}$ y $\boldsymbol{Y}$ distinta de cero, entonces al saber que existe al menos una relación entre las variables se tiene que analizar que esta relación está comprendida por la primera o por las pocas primeras variables canónicas, y si alguna de las

[1]THOMPSON, Bruce. (1984) "Canonical correlation analysis: Uses and interpretation". [21] pp. 10-18

variables originales pueden quedar fuera de consideración sin efecto de significancia a la conclusión del **AC**.

1.2. Análisis de componentes principales y vectores propios en el estudio del Análisis Canónico

El análisis de componentes principales es una técnica multivariada descriptiva, que tiene la finalidad de reducir la dimensión de un conjunto de variables multivariadas mediante el estudio de la relación entre las variables originales, es decir, se trata de formar un nuevo conjunto de variables no correlacionadas, con la característica que cada una de estas nuevas variables sea una combinación lineal de las variables originales; al ser una técnica descriptiva, el nuevo conjunto de variables, que serán llamadas componentes principales, tiene la particularidad de describir la variabilidad que existe entre el conjunto de variables multivariadas original, de tal manera que las componentes principales sean deducidas en un cierto "orden" de importancia, es decir, que la variabilidad explicada de la primera componente principal, explique tanto como sea posible la variabilidad de las variables originales, a su vez, la segunda componente explique tanto como sea posible la variabilidad que resta por ser explicada, y así sucesivamente hasta formar un nuevo conjunto de variables. Con el objetivo principal de saber si las pocas primeras componentes principales explican la mayor parte de variabilidad, lo cual implicará una reducción significativa de la dimensionalidad de las variables originales, y con esto una simplificación de análisis posteriores.

Geométricamente, el análisis de componentes principales es una rotación y traslación de los ejes de las variables originales en un sistema coordenado, que ayudan a la identificación de grupos en los datos o sujetos con la característica que dichos grupos aglomeran la mayor parte de la variabilidad de las variables y dando como resultado nuevos ejes coordenados, de tal forma que los nuevos ejes

coinciden con la dirección donde se maximiza la dispersión de los datos, a estos nuevos ejes coordenados se les conoce como componentes principales.

Cabe mencionar, que si las variables a estudiar son muy pocas o son independientes no se recomienda realizar dicho análisis, ya que el objetivo principal es reducir la dimensión del conjunto de datos y en estos casos, se recomienda realizar el estudio de los datos con las variables originales.

Sea $\boldsymbol{X}$ un vector con p variables aleatorias correlacionadas que siguen una distribución desconocida, pero que se puede suponer:

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \text{ y } Var(\boldsymbol{X}) = \boldsymbol{\Sigma},$$

donde $\boldsymbol{\Sigma}$ es la matriz de varianzas y covarianzas de las variables.

El primer paso es encontrar la primera componente principal que sea una combinación lineal de las variables que tenga varianza máxima, es decir:

Sea Y_1 la primera combinación lineal, entonces,

$$Y_1 = \boldsymbol{u_1'X} = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p = \sum_{i=1}^{p} u_{1i}X_i$$

donde $\boldsymbol{u_1'}$ es un vector de p constantes, es decir, $\boldsymbol{u_1'} = (u_{11}, u_{12}, \ldots, u_{1p})$ y $\boldsymbol{X'} = (X_1, X_2, \ldots, X_p)$,
de tal forma que:

$$Var(Y_1) = Var(\boldsymbol{u_1'X}) = \boldsymbol{u_1'\Sigma u_1},$$

por lo tanto el objetivo es **maximizar** $(\boldsymbol{u_1'\Sigma u_1})$, en sentido que Y_1 acumule la máxima variabilidad posible, pero se puede maximizar la varianza sin límite, aumentando el módulo del vector $\boldsymbol{u_1}$, es decir, en las ecuaciones de las componentes existe un factor de escala arbitraria (existen infinitas soluciones

en las mismas direcciones del espacio), por lo tanto, conviene que los vectores directores tengan módulo 1.

Para que la maximización de la varianza de Y_1 tenga solución, se debe imponer la restricción de que $\boldsymbol{u_1'u_1} = 1$ para garantizar la unicidad de la solución. Utilizando multiplicadores de Lagrange, se tiene

$$\eta = \boldsymbol{u_1'\Sigma u_1} - \lambda(\boldsymbol{u_1'u_1} - 1),$$

para maximizar la función lagrangeana se obtiene la derivada de η con respecto a $\boldsymbol{u_1}$ y con respecto a λ: [2]

$$\frac{\delta\eta}{\delta u_1} = 2\boldsymbol{\Sigma u_1} - 2\lambda\boldsymbol{u_1}$$

$$\frac{\delta\eta}{\delta\lambda} = -(\boldsymbol{u_1'u_1} - 1)$$

Igualando a cero las ecuaciones anteriores, se obtiene:

$$2\boldsymbol{\Sigma u_1} - 2\lambda\boldsymbol{u_1} = 0 \Leftrightarrow \boldsymbol{\Sigma u_1} - \lambda\boldsymbol{u_1} = 0 \Leftrightarrow (\boldsymbol{\Sigma} - \lambda\boldsymbol{I})\boldsymbol{u_1} = 0 \tag{1.1}$$

$$\boldsymbol{u_1'u_1} - 1 = 0$$

La ecuación (1.1) tiene solución no trivial si $\boldsymbol{\Sigma} - \lambda\boldsymbol{I}$ es una matriz singular y por lo tanto su determinante es igual a cero, es decir:

$$|\boldsymbol{\Sigma} - \lambda\boldsymbol{I}| = 0.$$

También de la ecuación (1.1) se obtiene que

$$\boldsymbol{\Sigma u_1} - \lambda\boldsymbol{u_1} = 0 \Leftrightarrow \boldsymbol{\Sigma u_1} = \lambda\boldsymbol{u_1},$$

[2] MORRISON, D. F. (1978): "Multivariate Statistical Methods".[18] pp. 110-125

multiplicando por la izquierda de ambos lados de la ecuación por $\boldsymbol{u_1'}$, se obtiene

$$\boldsymbol{u_1'\Sigma u_1} = \boldsymbol{u_1'}\lambda\boldsymbol{u_1} = \lambda\boldsymbol{u_1'Iu_1} = \lambda.$$

Con esto, el problema de **maximizar** $(\boldsymbol{u_1'\Sigma u_1})$ se reduce a encontrar los vectores y valores propios de $\boldsymbol{\Sigma}$.

Al calcular la varianza de la primera componente principal, se tiene que

$$Var(Y_1) = Var(\boldsymbol{u_1'X}) = \boldsymbol{u_1'\Sigma u_1} = \lambda_1.$$

Por lo tanto, al elegir el valor propio más grande, que se denota por λ_1, y su correspondiente vector propio $\boldsymbol{u_1}$ se maximiza la varianza de Y_1.

Continuando con el análisis, se obtiene de manera análoga la segunda componente principal, es decir, una nueva combinación lineal $Y_2 = \boldsymbol{u_2'X}$ tal que

$$Var(Y_2) = Var(\boldsymbol{u_2'X}) = \boldsymbol{u_2'\Sigma u_2} = \lambda_2,$$

sujeta a las condiciones de

$$\boldsymbol{u_2'u_2} = 1 \text{ y } Cov(Y_1, Y_2) = 0$$

Es decir, que la segunda componente principal sea no correlacionada con la primera Y_1 y a su vez, maximizando el resto de la varianza que no es explicada por la primera; dicho procedimiento se repite sucesivamente hasta obtener la $s-ésima$ combinación lineal $Y_s = \boldsymbol{u_s'X}$, que también está no correlacionada con $Y_1, Y_2, \ldots, Y_{s-1}$; la $s-ésima$ combinación lineal Y_s es la $s-ésima$ componente principal. Además por construcción se tiene que la $s-ésima$ componente principal

es el vector propio asociado con el $s-ésimo$ valor propio, y por construcción cumple que

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s,$$

donde

$$Var(Y_i) = \lambda_i.$$

Nótese que se obtienen s componentes principales, pero en general, es de esperarse que la mayor parte de la variabilidad en $\boldsymbol{X}$ pueda ser explicada por k componentes principales con $k < s$.

El vector de componentes principales puede ser expresado como

$$\boldsymbol{Y} = \boldsymbol{U'X}$$

donde $\boldsymbol{U}$ es una matriz ortogonal cuya $i-ésima$ columna $\boldsymbol{u_i}$, es el $i-ésimo$ vector propio de $\boldsymbol{\Sigma}$. De esta expresión se puede ver que las componentes principales son una transformación lineal ortogonal de $\boldsymbol{X}$.

Sea $\boldsymbol{\Lambda}$ la matriz diagonal, cuyo $i-ésimo$ elemento sobre la diagonal es λ_i. Por las características de estas matrices, se tiene que

$$\boldsymbol{\Sigma} = \boldsymbol{U\Lambda U'}$$

que corresponde a la descomposición espectral de la matriz $\boldsymbol{\Sigma}$.

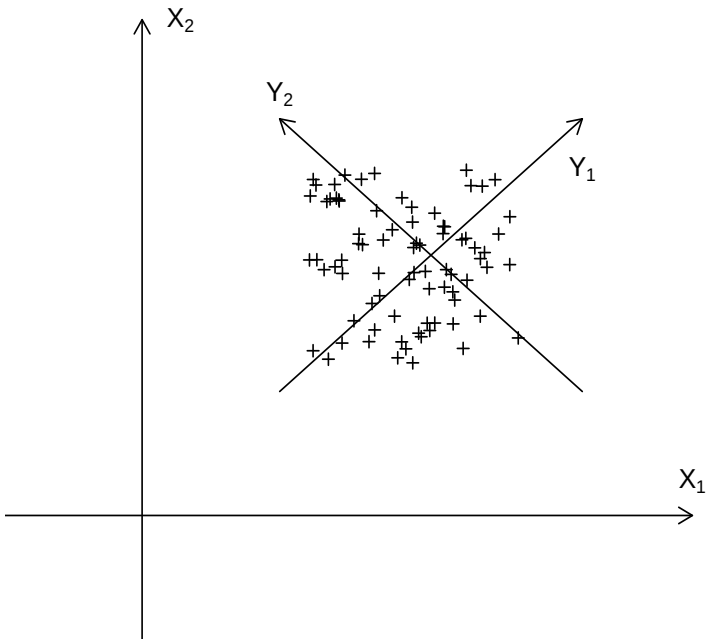
De la ecuación anterior se obtiene

$$\boldsymbol{U'\Sigma U} = \boldsymbol{\Lambda}.$$

El análisis de componentes principales consiste en encontrar los valores propios λ_i y los vectores propios $\boldsymbol{u_i}$ de la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$.

El valor propio es la varianza usual de los puntos proyectados, y gráficamente la componente del vector propio es el coseno del ángulo entre el eje de la variable y el eje principal correspondiente.

Figura 2: Representación de dos ejes principales



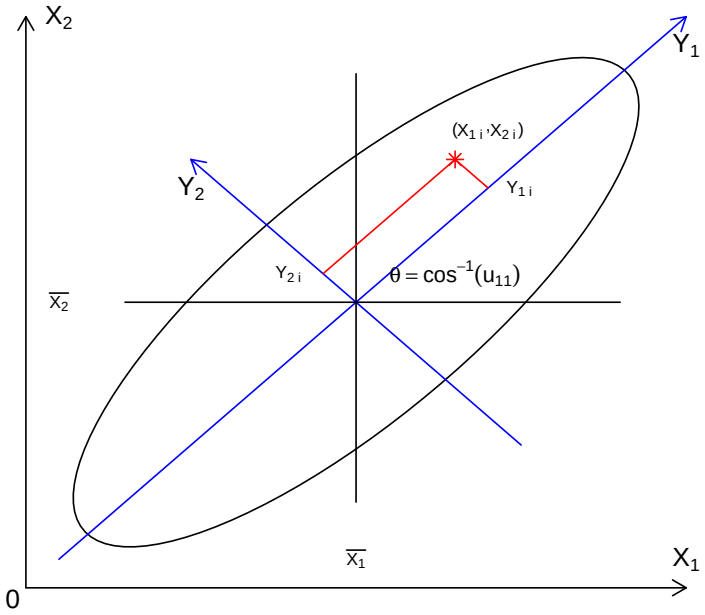
Idealización en dos variables de la rotación y translación de los ejes coordenados X_1 ***y*** X_2 ***hacia donde se encuentra la mayor dispersión de los datos, dando como resultado los ejes principales*** Y_1 ***y*** Y_2

Para tener una idea de la relación que existe entre el **AC** y el análisis de Componentes Principales; dada una muestra $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, de tamaño N, donde x_i es un punto en el plano $\forall i = 1, ..., N$; se tiene que realizar una proyección ortogonal de los puntos de la muestra sobre el eje principal, el punto y_{1i} es la proyección del punto $x_i = (x_{1i}, x_{2i})$ sobre el eje definido por la dirección Y_1, dicho eje tiene la propiedad que la dispersión de los puntos proyectados y_1, con $i = 1, \ldots, N$, es mayor que la dispersión de los mismos puntos pero proyectados sobre cualquier otro eje que pasa por $(\overline{x}_1, \overline{x}_2)$, donde $\overline{x}_1 = \frac{\sum_{i=1}^{N} x_{1i}}{N}$ y $\overline{x}_2 = \frac{\sum_{i=1}^{N} x_{2i}}{N}$.

La dirección en la cual es máxima la dispersión de los puntos proyectados define el primer eje principal, este eje también es llamado la línea de mejor ajuste para la muestra.

En la figura 3 se tiene la representación gráfica en dos variables, donde se muestra que el coseno del ángulo entre la variable X_1 y el primer eje principal Y_1 determina la primera componente u_{11} del primer vector propio $\boldsymbol{u_1}$, mientras que el coseno del ángulo entre la variable X_2 y Y_1 determina u_{12}. Similarmente, los cosenos de los ángulos entre el segundo eje principal Y_2 y los ejes coordenados X_1 y X_2, determinan las componentes u_{21} y u_{22} de $\boldsymbol{u_2}$.

Figura 3: Representación del ángulo entre la variable X_1 y el primer eje principal



1.3. Desarrollo de las variables y correlaciones canónicas

El cálculo de las variables y correlaciones canónicas está basado en el álgebra de matrices, la cual es utilizada para la transformación de matrices de las formas cuadráticas a las canónicas. Las definiciones y propiedades utilizadas a

continuación serán discutidas en el Apéndice, ya que se parte del conocimiento del álgebra de matrices y se procede a obtener resultados e interpretaciones.

Teorema 1: Descomposición singular

Sea $\boldsymbol{A}$ una matriz de dimensión $p \times q$ con $p \leq q$ de rango k, entonces dicha martiz puede ser expresada como

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{M'},$$

donde

$\boldsymbol{L}$ es una matriz ortogonal de $p \times p$

$\boldsymbol{M}$ es una matriz ortogonal de $q \times q$

$\boldsymbol{\Lambda}$ es una matriz de $p \times q$ con la diagonal de valores propios distintos de cero de $\boldsymbol{AA'}$ o $\boldsymbol{A'A}$ y cero en todo lo demás, como se muestra a continuación

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & & & & \cdots & 0 \\ 0 & \lambda_2 & 0 & 0 & & & & & \vdots \\ 0 & 0 & \lambda_3 & 0 & & & & & \\ \vdots & 0 & 0 & \ddots & & & & & \vdots \\ & & & & \lambda_k & & & & \\ \vdots & & & & & 0 & & & \vdots \\ & & & & & & \ddots & & \\ \vdots & & & & & & & 0 & \vdots \\ 0 & \cdots & & \cdots & & \cdots & & \cdots & 0 \end{pmatrix}$$

y es denotada por $\boldsymbol{\Lambda} = \big(diag\,(\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0) \mid 0_{p\times(q-p)}\big)$.

Demostración

Sean $\lambda_1^2, \lambda_2^2, \ldots, \lambda_k^2$ valores propios distintos de cero de $\boldsymbol{AA'}$ y sea

$$\boldsymbol{\Lambda^2 = \Lambda\Lambda'} = diag\,(\lambda_1^2, \lambda_2^2, \ldots, \lambda_k^2, 0, 0, \ldots, 0) \quad \text{matriz de } p \times p.$$

Por el análisis de componentes principales se sabe que existe una matriz ortogonal $\boldsymbol{L}$ de orden p tal que

$$\boldsymbol{AA'} = \boldsymbol{L}\; diag\left(\lambda_1^2, \lambda_2^2, \ldots, \lambda_k^2, 0, 0, \ldots, 0\right) \boldsymbol{L'} \tag{1.2}$$

haciendo una partición de $\boldsymbol{L}$ se puede ver a la matriz de la siguiente manera

$$\boldsymbol{L} = [l_1\; l_2 \cdots l_p] = (\boldsymbol{L_1}, \boldsymbol{L_2})$$

con $\boldsymbol{L_1}$ de orden $p \times k$ y $\boldsymbol{L_2}$ de orden $p \times (p-k)$. Entonces por la ecuación (1.2) se tiene

$$\boldsymbol{AA'L_1} = \boldsymbol{L_1\Lambda_1^2} \tag{1.3}$$

donde $\boldsymbol{\Lambda_1} = diag\,(\lambda_1, \lambda_2, \cdots, \lambda_k)$;
esto es posible ya que $\boldsymbol{L}$ es una matriz ortogonal $\Rightarrow \boldsymbol{L_1'L_1} = \boldsymbol{I_{k\times k}}$, nótese que $\boldsymbol{\Lambda_1^2}$ es una matriz cuadrada de $k \times k$.

Multiplicando la ecuación anterior por la izquierda por $\boldsymbol{A'}$ y por la derecha por $\boldsymbol{\Lambda_1^{-1}}$ se obtiene

$$\boldsymbol{A'AA'L_1\Lambda_1^{-1}} = \boldsymbol{A'L_1\Lambda_1^2\Lambda_1^{-1}}$$

por ser $\boldsymbol{\Lambda_1^2}$ y $\boldsymbol{\Lambda_1^{-1}}$ matrices diagonales de $k \times k$, se tiene lo siguiente

$$\boldsymbol{A'AA'L_1\Lambda_1^{-1} = A'L_1\Lambda_1^{-1}\Lambda_1^2} \tag{1.4}$$

Sea $\boldsymbol{M_1} = \boldsymbol{A'}_{q\times p}\boldsymbol{L_{1p\times k}\Lambda_1^{-1}}_{k\times k}$ y se observa que $\boldsymbol{M_1}$ es de tamaño $q \times k$, sustituyendo a $\boldsymbol{M_1}$ en (1.4)

$$\boldsymbol{A'AM_1 = M_1\Lambda_1^2}$$

Se sabe que $\boldsymbol{A}$ tiene q columnas y sólo k de éstas son linealmente independientes, entonces con este resultado siempre se puede encontrar, como se hizo con la matriz $\boldsymbol{L}$, el complemento de la matriz $\boldsymbol{M}$, en este caso se definirá como la matriz $\boldsymbol{M_2}$, es claro que dicha matriz es de tamaño $q \times (q-k)$, por ser el complemento de $\boldsymbol{M_1}$, la cual cumple

$$\boldsymbol{AM_2 = 0} \tag{1.5}$$

por la definición de $\boldsymbol{L_2}$.

Con estos resultados se tiene que

$$\boldsymbol{M_1'M_2 = \Lambda_1^{-1}L_1'AM_2 = 0}$$

por (1.5) y porque $\boldsymbol{M}$ es una matriz ortogonal de $q \times q$ de la forma

$$\boldsymbol{M} = (\boldsymbol{M_1M_2}),$$

multiplicando (1.3) por la derecha por $\boldsymbol{\Lambda_1^{-1}}$ y sabiendo que $\boldsymbol{M_1} = \boldsymbol{A'L_1\Lambda_1^{-1}}$, se obtiene

$$\boldsymbol{AM_1} = \boldsymbol{L_1\Lambda_1}$$

entonces, si se multiplica por la derecha ambos lados de la igualdad por $\boldsymbol{M_1'}$, se tiene

$$\boldsymbol{AM_1M_1'} = \boldsymbol{L_1\Lambda_1M_1'}$$

y sabiendo que $\boldsymbol{M_1M_1'} = \boldsymbol{I - M_2M_2'}$ y $\boldsymbol{AM_2} = \boldsymbol{0}$ se tiene que

$$\boldsymbol{A} = \boldsymbol{L_1\Lambda_1M_1'}$$

lo cual puede ser escrito también como

$$\boldsymbol{A} = \boldsymbol{L\Lambda M'} \tag{1.6}$$

lo cual prueba el teorema.■

De manera análoga se puede hacer una partición de la matriz $\boldsymbol{M}$ tal que se tengan q columnas

$$\boldsymbol{M} = [m_1\ m_2 \cdots m_q]$$

por lo tanto se tiene que la matriz $\boldsymbol{A}$ es la combinación lineal:

$$\boldsymbol{A} = \lambda_1 l_1 m_1' + \lambda_2 l_2 m_2' + \cdots + \lambda_k l_k m_k'.$$

De (1.6), se sigue que $\boldsymbol{AA'}=\boldsymbol{L\Lambda\Lambda'L'}$ y $\boldsymbol{A'A}=\boldsymbol{M\Lambda'\Lambda M'}$; como $\boldsymbol{L}$ y $\boldsymbol{M}$ son ortogonales, entonces también se sigue que las columnas de $\boldsymbol{L}$ son los vectores propios de $\boldsymbol{AA'}$, y $\boldsymbol{M}$ de $\boldsymbol{A'A}$. Las primeras k columnas corresponden a los valores propios $\lambda_1^2, \lambda_2^2, \ldots, \lambda_k^2$ y el resto son ceros.[3]

Desarrollo de las variables y correlaciones canónicas

Sean dos vectores $\boldsymbol{x}$ y $\boldsymbol{y}$ de p y q variables aleatorias respectivamente, con $p \leq q$ y sea la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ de todas las $p+q$ variables aleatorias que se ve de la siguiente manera

$$\boldsymbol{\Sigma} = Var\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix} \begin{matrix} p \\ q \end{matrix}$$

$$\begin{matrix} p & q \end{matrix}$$

Ahora se prueba el siguiente teorema:

Teorema 2: Variables y correlaciones canónicas

Sean las transformaciones lineales

$$\boldsymbol{u} = \boldsymbol{Lx} \quad \text{y} \quad \boldsymbol{v} = \boldsymbol{My},$$

de $\boldsymbol{x}$ y $\boldsymbol{y}$ en las nuevas variables $\boldsymbol{u}$ y $\boldsymbol{v}$. Entonces:

$$Var\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I_p} & \boldsymbol{P} \\ \boldsymbol{P'} & \boldsymbol{I_q} \end{pmatrix},$$

[3]KSHIRSAGAR, Anant M. (1972) "Multivariate Analysis". Statistics: textbooks and monographs Vol. 2, New York : Marcel Dekker, pp.247-249

donde

$$\boldsymbol{P} = \underbrace{\left(\begin{array}{ccccccccc} \rho_1 & 0 & 0 & \cdots & & \cdots & & \cdots & 0 \\ 0 & \rho_2 & 0 & & & & & & \vdots \\ 0 & 0 & \rho_3 & & & & & & \\ \vdots & 0 & 0 & \ddots & & & & & \vdots \\ & & & & \rho_r & & & & \\ \vdots & & & & & 0 & & & \vdots \\ & & & & & & 0 & & \\ \vdots & & & & & & & \ddots & \vdots \\ 0 & \cdots & & \cdots & & \cdots & & \cdots & 0 \end{array}\right)}_{q} p$$

que se denota como $\boldsymbol{P} = \big(diag\,(\rho_1, \rho_2, \cdots, \rho_r, 0, \cdots, 0) \mid 0_{p\times(q-p)}\big)$, $r = min(p,\ q)$ el rango de $\boldsymbol{\Sigma_{12}}$; $\boldsymbol{\Sigma_{11}}$ y $\boldsymbol{\Sigma_{22}}$ son matrices no-singulares y $\boldsymbol{L}$ y $\boldsymbol{M}$ son soluciones de los sistemas:

$$\mid \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} - \boldsymbol{\rho^2\Sigma_{11}} \mid \boldsymbol{L} = \boldsymbol{0}\,;\ \boldsymbol{L'\Sigma_{11}L} = \boldsymbol{I_p}$$

$$\mid \boldsymbol{\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}} - \boldsymbol{\rho^2\Sigma_{22}} \mid \boldsymbol{M} = \boldsymbol{0}\,;\ \boldsymbol{M'\Sigma_{22}M} = \boldsymbol{I_q}$$

donde $\boldsymbol{\rho^2} = \boldsymbol{PP'}$ para el caso en que se toma la matriz $\boldsymbol{L}$ y $\boldsymbol{\rho^2} = \boldsymbol{P'P}$ para $\boldsymbol{M}$.

Demostración

Sea la matriz $\boldsymbol{C}$ de $p \times q$

$$\boldsymbol{C} = \boldsymbol{\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}},$$

donde $\boldsymbol{\Sigma_{11}^{\frac{1}{2}}}$ y $\boldsymbol{\Sigma_{22}^{\frac{1}{2}}}$ son raíces cuadradas de $\boldsymbol{\Sigma_{11}}$ y $\boldsymbol{\Sigma_{22}}$ respectivamente.

Como $\boldsymbol{\Sigma_{11}}$ y $\boldsymbol{\Sigma_{22}}$ son definidas positivas, entonces $\boldsymbol{\Sigma_{11}^{\frac{1}{2}}}$ y $\boldsymbol{\Sigma_{22}^{\frac{1}{2}}}$ existen, son simétricas y definidas positivas. Se sabe que el rango de una matriz permanece inalterado si es multiplicada por la derecha o por la izquierda por una matriz no-singular, entonces el rango de $\boldsymbol{C}$ es el mismo que el de $\boldsymbol{\Sigma_{12}}$, por lo tanto el rango de $\boldsymbol{C}$ es r. Ahora bien, usando el resultado del teorema anterior, existen $\boldsymbol{A}$ y $\boldsymbol{B}$ ortogonales de orden p y q respectivamente tales que

$$\boldsymbol{C} = \boldsymbol{APB'} \tag{1.7}$$

donde $\boldsymbol{P}$ es la matriz anteriormente definida y $\rho_1^2, \rho_2^2, \cdots, \rho_r^2$ son los valores propios distintos de cero de

$$\boldsymbol{CC'} = \boldsymbol{\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}}$$

o de

$$\boldsymbol{C'C} = \boldsymbol{\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}},$$

es decir, ρ_i^2 con $i = 1, 2, \cdots, r$, son raíces distintas de cero de alguna de las ecuaciones siguientes:

$$\mid \boldsymbol{CC' - \rho^2 I} \mid = 0 \quad \text{o} \quad \mid \boldsymbol{C'C - \rho^2 I} \mid = 0.$$

Haciendo el desarrollo algebraico, las ecuaciones pueden quedar en términos del producto de $\boldsymbol{\Sigma_{11}}$ y $\boldsymbol{\Sigma_{22}}$ con $\boldsymbol{\rho^2}$, sin pérdida de generalidad se muestra la explicación para la primera ecuación:

$$| \boldsymbol{CC'} - \boldsymbol{\rho^2 I} | = 0$$

$$| \boldsymbol{\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}} - \boldsymbol{\rho^2 I} | = 0$$

multiplicando la ecuación anterior por la derecha y por la izquierda por $| \boldsymbol{\Sigma_{11}^{\frac{1}{2}}} |$ se tiene

$$| \boldsymbol{\Sigma_{11}^{\frac{1}{2}}} || \boldsymbol{\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}} - \boldsymbol{\rho^2 I} || \boldsymbol{\Sigma_{11}^{\frac{1}{2}}} | = 0$$

por propiedades de los determinantes y como $\boldsymbol{\Sigma_{11}}$ es no-singular, el producto de determinantes resulta

$$| \boldsymbol{\Sigma_{11}^{\frac{1}{2}} \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}} \Sigma_{11}^{\frac{1}{2}}} - \boldsymbol{\Sigma_{11}^{\frac{1}{2}} \rho^2 \Sigma_{11}^{\frac{1}{2}}} | = 0$$

simplificando la ecuación, queda de la siguiente manera

$$| \boldsymbol{\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}} - \boldsymbol{\rho^2 \Sigma_{11}} | = 0$$

Un desarrollo análogo para la segunda ecuación se realiza de tal manera que ρ_i^2 con $i = 1, 2, \cdots, r$, son raíces distintas de cero de la ecuación

$$| \boldsymbol{\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}} - \boldsymbol{\rho^2 \Sigma_{11}} | = 0 \tag{1.8}$$

o

$$| \boldsymbol{\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}} - \boldsymbol{\rho^2 \Sigma_{22}} | = 0 \tag{1.9}$$

Sean

$$\boldsymbol{L} = \boldsymbol{A'}\boldsymbol{\Sigma}_{\boldsymbol{11}}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}} \quad \text{y} \quad \boldsymbol{M} = \boldsymbol{B'}\boldsymbol{\Sigma}_{\boldsymbol{22}}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}},$$

dado que $\boldsymbol{A}$ y $\boldsymbol{B}$ son matrices ortogonales, entonces $\boldsymbol{A'A} = \boldsymbol{I}$ y $\boldsymbol{B'B} = \boldsymbol{I}$, entonces:

$$\boldsymbol{L}\boldsymbol{\Sigma}_{\boldsymbol{11}}\boldsymbol{L'} = \boldsymbol{I_p} \quad \text{y} \quad \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{22}}\boldsymbol{M'} = \boldsymbol{I_q}$$

además

$$\begin{aligned} \boldsymbol{L}\boldsymbol{\Sigma}_{\boldsymbol{12}}\boldsymbol{M'} &= \boldsymbol{A'}\boldsymbol{\Sigma}_{\boldsymbol{11}}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}}\boldsymbol{\Sigma}_{\boldsymbol{12}}\boldsymbol{\Sigma}_{\boldsymbol{22}}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}}\boldsymbol{B} \\ &= \boldsymbol{A'CB} \\ &= \boldsymbol{A'APB'B} \\ &= \boldsymbol{P} \end{aligned}$$

Por lo tanto la matriz de varianzas y covarianzas de las variables transformadas $\boldsymbol{u} = \boldsymbol{Lx}$ y $\boldsymbol{v} = \boldsymbol{My}$ es

$$Var\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}\boldsymbol{\Sigma}_{\boldsymbol{11}}\boldsymbol{L'} & \boldsymbol{L}\boldsymbol{\Sigma}_{\boldsymbol{12}}\boldsymbol{M'} \\ \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{21}}\boldsymbol{L'} & \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{22}}\boldsymbol{M'} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I_p} & \boldsymbol{P} \\ \boldsymbol{P'} & \boldsymbol{I_q} \end{pmatrix} \tag{1.10}$$

lo cual prueba el teorema. ■

Así $\boldsymbol{u}$ es combinación lineal de $x_1, \cdots, x_p$ elementos de $\boldsymbol{x}$, y $\boldsymbol{v}$ es combinación lineal de $y_1, \cdots, y_q$ elementos de $\boldsymbol{y}$. Por lo tanto $\boldsymbol{u}$ es llamada variable canónica del espacio $\boldsymbol{x}$, y $\boldsymbol{v}$ es llamada variable canónica del espacio $\boldsymbol{y}$.

Las columnas de $\boldsymbol{L'}$ y $\boldsymbol{M'}$ son los vectores canónicos. Dadas $\rho_1^2, \rho_2^2, \cdots, \rho_r^2$ que por construcción cumplen que $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_r^2$, entonces $\boldsymbol{u_1}$ es llamada la primera variable canónica del espacio $\boldsymbol{x}$, y $\boldsymbol{v_1}$ es la primera variable canónica del espacio $\boldsymbol{y}$, $\boldsymbol{u_2}$ es la segunda variable canónica del espacio $\boldsymbol{x}$ y así sucesivamente. La

relación entre las p variables de $\boldsymbol{x}$ y las q variables de $\boldsymbol{y}$ es expresada únicamente en términos de r raíces: $\rho_1^2, \rho_2^2, \cdots, \rho_r^2$.

Por lo tanto, de (1.10) se concluye que la submatriz $\boldsymbol{P}$, de la matriz de varianzas y covarianzas de las combinaciones $\boldsymbol{u}$ y $\boldsymbol{v}$, corresponde a la matriz de correlación de las variables $\boldsymbol{x}$ y $\boldsymbol{y}$, o bien:

$$corr(\boldsymbol{u_i}, \boldsymbol{v_j}) = \begin{cases} \rho_i, & \text{si } i = j \\ 0, & \text{en otro caso} \end{cases}$$

$$corr(\boldsymbol{u_i}, \boldsymbol{u_j}) = 0, \;\; corr(\boldsymbol{v_i}, \boldsymbol{v_j}) = 0 \quad \text{para toda } i \neq j.$$

Es decir, la primera variable canónica del espacio $\boldsymbol{x}$ está correlacionada únicamente con la primera variable canónica del espacio $\boldsymbol{y}$, la segunda con la segunda y así sucesivamente. Estas correlaciones son $\rho_1, \rho_2, \cdots, \rho_r$, por lo tanto $\rho_1^2, \rho_2^2, \cdots, \rho_r^2$, son las raíces de (1.8) o (1.9) y $0 \leq \rho_i^2 \leq 1 \;\forall\; i = 1, \cdots, r$ son conocidas como las correlaciones canónicas.

Multiplicando por la izquierda la expresión (1.7) por $\boldsymbol{A'}$ y como $\boldsymbol{A'A} = \boldsymbol{I}$, se tiene

$$\boldsymbol{A'C} = \boldsymbol{A'APB'}$$

lo cual implica

$$\boldsymbol{A'C} = \boldsymbol{PB'}$$

Expresando $\boldsymbol{A}$ y $\boldsymbol{B}$ en términos de $\boldsymbol{L}$ y $\boldsymbol{M}$, entonces la expresión anterior queda de la siguiente manera:

$$\begin{aligned} \boldsymbol{L}\boldsymbol{\Sigma}_{\mathbf{11}}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{-\frac{1}{2}} &= \boldsymbol{P}\boldsymbol{M}\boldsymbol{\Sigma}_{\mathbf{22}}^{\frac{1}{2}} \\ \boldsymbol{L}\boldsymbol{\Sigma}_{\mathbf{11}}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\frac{1}{2}} &= \boldsymbol{P}\boldsymbol{M}\boldsymbol{\Sigma}_{\mathbf{22}} \\ \boldsymbol{L}\boldsymbol{\Sigma}_{\mathbf{12}} &= \boldsymbol{P}\boldsymbol{M}\boldsymbol{\Sigma}_{\mathbf{22}} \\ \boldsymbol{L}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{-\mathbf{1}} &= \boldsymbol{P}\boldsymbol{M} \end{aligned}$$

transponiendo ambos lados de la ecuación anterior, se tiene

$$\boldsymbol{\Sigma}_{\mathbf{22}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{21}}\boldsymbol{l}_{\boldsymbol{i}} = \begin{cases} \rho_i \boldsymbol{m}_{\boldsymbol{i}}, & \text{si } i = 1, 2, \cdots, r \\ 0, & \text{si } i = r+1, \cdots, p \end{cases} \tag{1.11}$$

donde $\boldsymbol{l}_{\boldsymbol{i}}$ es la $i - ésima$ columna de $\boldsymbol{L}$; de manera similar si se multiplica por la derecha por $\boldsymbol{B}$ ambos lados de la ecuación y tomando en cuenta que $\boldsymbol{B'B} = \boldsymbol{I}$, se obtiene:

$$\begin{aligned} \boldsymbol{CB} &= \boldsymbol{APB'B} \\ \boldsymbol{CB} &= \boldsymbol{AP} \end{aligned}$$

que desarrollando se tiene:

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\frac{1}{2}}\boldsymbol{M'} &= \boldsymbol{\Sigma}_{\mathbf{11}}^{\frac{1}{2}}\boldsymbol{L'P} \\ \boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{M'} &= \boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{11}}^{\frac{1}{2}}\boldsymbol{L'P} \\ \boldsymbol{\Sigma}_{\mathbf{11}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{M'} &= \boldsymbol{L'P} \end{aligned}$$

Análogamente al caso anterior, si se transpone ambos lados de la ecuación se tiene

$$\boldsymbol{\Sigma}_{\mathbf{11}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{m}_{\boldsymbol{i}} = \begin{cases} \rho_i \boldsymbol{l}_{\boldsymbol{i}}, & \text{si } i = 1, 2, \cdots, r \\ 0, & \text{si } i = r+1, \cdots, p \end{cases} \tag{1.12}$$

Así $\boldsymbol{\Sigma}_{\mathbf{11}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{12}}$ es la matriz de coeficientes de regresión del vector $\boldsymbol{y}$ sobre el vector $\boldsymbol{x}$. Entonces la ecuación (1.12) muestra que $\boldsymbol{l}_{\boldsymbol{i}}$ es la "proyección de $\boldsymbol{m}_{\boldsymbol{i}}$ en el espacio

$\boldsymbol{x}$" y similarmente la ecuación (1.11) muestra que $\boldsymbol{m_i}$ es la "proyección de $\boldsymbol{l_i}$ en el espacio $\boldsymbol{y}$". Entonces, conjuntamente las ecuaciones (1.11) y (1.12) son expresadas de la siguiente manera:

$$\begin{aligned} \boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}m_i} &= \rho_i \boldsymbol{l_i} \\ \boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}l_i} &= \rho_i \boldsymbol{m_i} \end{aligned} \quad \text{para } i = 1, 2, \cdots, r$$

o bien

$$\begin{aligned} \boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}m_i} - \rho_i \boldsymbol{l_i} &= \boldsymbol{0} \\ \boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}l_i} - \rho_i \boldsymbol{m_i} &= \boldsymbol{0} \end{aligned} \quad \text{para } i = 1, 2, \cdots, r$$

que es equivalente a

$$\begin{aligned} \boldsymbol{\Sigma_{12}m_i} - \boldsymbol{\Sigma_{11}}\rho_i \boldsymbol{l_i} &= \boldsymbol{0} \\ \boldsymbol{\Sigma_{21}l_i} - \boldsymbol{\Sigma_{22}}\rho_i \boldsymbol{m_i} &= \boldsymbol{0} \end{aligned} \quad \text{para } i = 1, 2, \cdots, r$$

Por lo tanto la expresión en forma matricial es

$$\begin{bmatrix} -\rho_i \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & -\rho_i \boldsymbol{\Sigma_{22}} \end{bmatrix} \begin{bmatrix} \boldsymbol{l_i} \\ \boldsymbol{m_i} \end{bmatrix} = \boldsymbol{0} \quad \text{para} \quad i = 1, 2, \ldots, r \tag{1.13}$$

considerando que

$$\begin{aligned} \boldsymbol{\Sigma_{21}l_i} &= \boldsymbol{0} \qquad i = r+1, \cdots, p, \\ \boldsymbol{\Sigma_{12}m_i} &= \boldsymbol{0} \qquad i = r+1, \cdots, q. \end{aligned}$$

Utilizando las ecuaciones (1.11) y (1.12), se puede eliminar $\boldsymbol{m_i}$; además, se sabe que por construcción, $\rho_i > 0$ y por lo tanto se puede dividir en ambos lados de la ecuación para obtener una expresión para $\boldsymbol{m_i}$. Así $\boldsymbol{m_i} = \boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}l_i}\frac{1}{\rho_i}$, ahora sustituyendo el valor de $\boldsymbol{m_i}$ en la ecuación (1.12) se tiene

$$\begin{aligned} \boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}l_i}\frac{1}{\rho_i} &= \rho_i \boldsymbol{l_i} \\ \boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}l_i} &= \rho_i^2 \boldsymbol{l_i} \end{aligned}$$

multiplicando por la izquierda ambos lados de la ecuación por $\boldsymbol{\Sigma_{11}}$:

$$\boldsymbol{\Sigma_{11}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}l_i} = \boldsymbol{\Sigma_{11}}\rho_i^2\boldsymbol{l_i} \qquad i = 1, \cdots, r$$

$$\boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}l_i} = \boldsymbol{\Sigma_{11}}\rho_i^2\boldsymbol{l_i}$$

$$\boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}l_i} - \boldsymbol{\Sigma_{11}}\rho_i^2\boldsymbol{l_i} = \mathbf{0}$$

$$\left[\boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} - \boldsymbol{\Sigma_{11}}\rho_i^2\right]\boldsymbol{l_i} = \mathbf{0} \quad i = 1, \cdots, r \tag{1.14}$$

De forma análoga si se sustituye el valor de $\boldsymbol{l_i}$ en la ecuación (2.10), se obtiene

$$\left[\boldsymbol{\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}} - \boldsymbol{\Sigma_{22}}\rho_i^2\right]\boldsymbol{m_i} = \mathbf{0} \quad i = 1, \cdots, r \tag{1.15}$$

Aunque otra manera de obtener el resultado directamente de (2.6) es multiplicar la ecuación por la izquierda por $\boldsymbol{C'A}$ o multiplicarla por la izquierda por $\boldsymbol{C'}$ y por la derecha por $\boldsymbol{B}$, para el primer caso se tiene

$$\boldsymbol{CC'A = APB'BP'A'A}$$

se sabe que $\boldsymbol{A}$ y $\boldsymbol{B}$ son ortogonales, por lo tanto la expresión queda reducida como sigue

$$\boldsymbol{CC'A = APP'}$$

para el segundo caso la expresión es

$$\boldsymbol{C'CB = BP'A'APB'B}$$

simplificando

$$\boldsymbol{C'CB = BP'P}$$

Utilizando las igualdades de $\boldsymbol{L}$ y $\boldsymbol{M}$, definidas en la página 31 y si se multiplica por la izquierda a la expresión (1.14) por $\boldsymbol{l_i'}$ se obtiene:

$$\boldsymbol{l_i'\left[\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{11}\rho_i^2\right]l_i = 0}$$

$$\boldsymbol{l_i'\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}l_i - l_i'\Sigma_{11}\rho_i^2 l_i = 0}$$

$$\boldsymbol{l_i'\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}l_i = l_i'\Sigma_{11}\rho_i^2 l_i}$$

$$\boldsymbol{\frac{l_i'\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}l_i}{l_i'\Sigma_{11}l_i}} = \rho_i^2 \qquad i = 1, 2, \cdots, r$$

Por lo tanto ρ_i es el coeficiente de correlación múltiple entre $\boldsymbol{u_i = l_i'x}$ y $\boldsymbol{v_i = m_i'y}$.

Ahora bien, dado que $\boldsymbol{C = APB'}$ y de la igualdad del producto de matrices $\boldsymbol{CC'}$:

$$\begin{aligned}
\boldsymbol{CC'} &= \boldsymbol{APP'A'} \\
\boldsymbol{\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}} &= \boldsymbol{APP'A'} \\
\boldsymbol{\Sigma_{11}^{\frac{1}{2}}\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}\Sigma_{11}^{\frac{1}{2}}} &= \boldsymbol{\Sigma_{11}^{\frac{1}{2}}APP'A'\Sigma_{11}^{\frac{1}{2}}} \\
\boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} &= \boldsymbol{\left(\Sigma_{11}^{\frac{1}{2}}A\right)(PP')\left(\Sigma_{11}^{\frac{1}{2}}A\right)'}
\end{aligned}$$

por la definición de $\boldsymbol{L}$ y tomando en cuenta que $\boldsymbol{A}$ es ortogonal, se tiene:

i. $\boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} = (\boldsymbol{L^{-1}})\, diag\,(\rho_1^2, \ldots, \rho_r^2, 0, \ldots, 0)\,(\boldsymbol{L^{-1}})'$

ii. $\boldsymbol{\Sigma_{11}} = (\boldsymbol{L^{-1}})\,(\boldsymbol{L^{-1}})'$

No es difícil observar, que de (1.10), el determinante de la matriz de varianzas y covarianzas de las variables canónicas $\boldsymbol{u}$ y $\boldsymbol{v}$ es

$$\begin{vmatrix} \boldsymbol{I_p} & \boldsymbol{P} \\ \boldsymbol{P'} & \boldsymbol{I_q} \end{vmatrix} = \left|\boldsymbol{I_p}\right| \left|\boldsymbol{I_q} - \boldsymbol{P'I_pP}\right| = \left|\boldsymbol{I} - \boldsymbol{P'P}\right| = \prod_{i=1}^{r}\left(1-\rho_i^2\right)$$

El resultado de este determinante es importante ya que después será utilizado en el desarrollo de las pruebas de hipótesis. [4]

1.4. Algunos resultados acerca de las correlaciones canónicas

Cuando $p = 1$, la ecuación $\mid \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} - \boldsymbol{\rho^2\Sigma_{11}} \mid = 0$ se reduce a

$$\boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} - \boldsymbol{\rho^2\Sigma_{11}} = 0$$

cuya solución es

$$\boldsymbol{\rho^2} = \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}}$$

mostrando que es una sola correlación canónica, y además es el coeficiente de correlación múltiple entre $\boldsymbol{x}$ y $\boldsymbol{y}$. De manera similar, pero ahora considerando la ecuación $\mid \boldsymbol{\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}} - \boldsymbol{\rho^2\Sigma_{22}} \mid = 0$ y sí $q = 1$, sólo existe una correlación canónica distinta de cero, y es el coeficiente de correlación múltiple entre $\boldsymbol{x}$ y $\boldsymbol{y}$.

[4] KSHIRSAGAR, "Multivariate Analysis" [13] pp.247-258

Cuando p y q son iguales a 1, la correlación canónica se reduce al coeficiente de correlación ordinario y el resultado en cuestión se obtiene de cualquiera de las ecuaciones antes empleadas.

Además las correlaciones canónicas son invariantes para una transformación lineal no-singular de $\boldsymbol{x}$ y para una transformación no-singular de $\boldsymbol{y}$. En otras palabras, las correlaciones canónicas entre $\boldsymbol{x}$ y $\boldsymbol{y}$ son las mismas que entre $\boldsymbol{Px}$ y $\boldsymbol{Qy}$, donde $\boldsymbol{P}$ y $\boldsymbol{Q}$ son cualesquiera dos matrices no-singulares de orden p y q respectivamente. Esto se sigue del hecho de que la matriz de varianzas y covarianzas de $\boldsymbol{Px}$ y $\boldsymbol{Qy}$ es

$$\begin{bmatrix} \boldsymbol{P\Sigma_{11}P'} & \boldsymbol{P\Sigma_{12}Q'} \\ \boldsymbol{Q\Sigma_{21}P'} & \boldsymbol{Q\Sigma_{22}Q'} \end{bmatrix},$$

Por tanto, por $\mid \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} - \boldsymbol{\rho^2\Sigma_{11}} \mid = 0$, las correlaciones canónicas entre $\boldsymbol{Px}$ y $\boldsymbol{Qy}$ son las raíces de la ecuación

$$\left| -\boldsymbol{\rho^2}\left(\boldsymbol{P\Sigma_{11}P'}\right) + \left(\boldsymbol{P\Sigma_{12}Q'}\right)\left(\boldsymbol{Q\Sigma_{22}Q'}\right)^{-1}\left(\boldsymbol{Q\Sigma_{21}P'}\right) \right| = 0$$

la cual se reduce a:

$$\left|\boldsymbol{P}\right| \left| -\boldsymbol{\rho^2\Sigma_{11}} + \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} \right| \left|\boldsymbol{P'}\right| = 0,$$

$$\left| -\boldsymbol{\rho^2\Sigma_{11}} + \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} \right| = 0,$$

ya que $\boldsymbol{P}$ es no-singular.

1.5. Supuestos estadísticos requeridos

Los procedimientos estadísticos están basados en supuestos que deben ser conocidos para ser correctamente aplicados, es importante identificarlos cuando se tienen propósitos inferenciales como es el caso del **AC** y cuando se tienen situaciones en las que es necesario hacer pruebas estadísticas para realizar un análisis adecuado.

El **AC** tiene tres supuestos principales:

i) El error de medición de las variables es mínimo.

ii) Las varianzas de las variables son no restringidas.

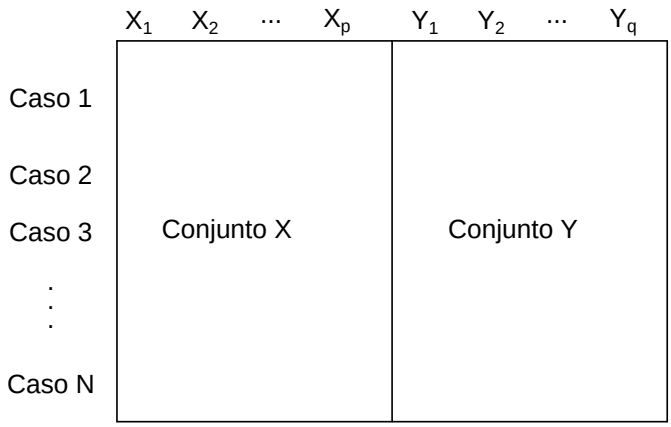
iii) Las variables tienen distribución normal multivariada.

Por las propiedades del Teorema Central del Límite, cuando el número de variables aleatorias independientes de la muestra es grande y siguen una misma distribución, la suma de las variables se aproxima a una distribución normal multivariada aún cuando su distribución original no sea normal multivariada. Cabe mencionar que sí las variables están medidas en diferentes escalas, es imposible que sigan la misma distribución.

1.6. Variables y correlaciones canónicas muestrales

Dada una muestra de N observaciones de $p+q$ variables de $\boldsymbol{x}$ y $\boldsymbol{y}$, y si se denotan estas observaciones muestrales sobre $\boldsymbol{x}$ y $\boldsymbol{y}$ como dos matrices $\boldsymbol{X}$ y $\boldsymbol{Y}$ de orden $N \times p$ y $N \times q$ respectivamente, entonces se tiene para cada una de las N observaciones sus correspondientes $p+q$ valores sobre cada variable de los conjuntos $\boldsymbol{X}$ y $\boldsymbol{Y}$, de los cuales todas las variables están contenidas en una sola matriz particionada como se muestra en la siguiente figura.

Figura 4: ***Representación gráfica de los dos conjuntos muestrales de variables X y Y en una sola matriz de*** $p+q$ ***variables***



Entonces, para la construcción de varianzas y covarianzas muestrales, se define la matriz $\boldsymbol{S}$,

$$\boldsymbol{S} = \begin{array}{c} \\ \left(\begin{array}{cc} \boldsymbol{S_{11}} & \boldsymbol{S_{12}} \\ \boldsymbol{S_{21}} & \boldsymbol{S_{22}} \end{array} \right) \\ \begin{array}{cc} p & q \end{array} \end{array} \begin{array}{c} \\ p \\ q \end{array}$$

donde

$$\boldsymbol{S_{11}} = \boldsymbol{X'} \left(\boldsymbol{I} - \tfrac{1}{N} \boldsymbol{E_{NN}} \right) \boldsymbol{X}$$

$$\boldsymbol{S_{12}} = \boldsymbol{S'_{21}} = \boldsymbol{X'} \left(\boldsymbol{I} - \tfrac{1}{N} \boldsymbol{E_{NN}} \right) \boldsymbol{Y}$$

$$\boldsymbol{S_{22}} = \boldsymbol{Y'} \left(\boldsymbol{I} - \tfrac{1}{N} \boldsymbol{E_{NN}} \right) \boldsymbol{Y}$$

con $\boldsymbol{E_{NN}}$ una matriz de tamaño $N \times N$ con todos sus elementos unitarios.

Además, suponiendo que $\boldsymbol{x}$ y $\boldsymbol{y}$ siguen una distribución normal multivariada, $\frac{1}{N}\boldsymbol{S}$ es el estimador por máxima verosimilitud de la matriz de varianzas y covarianzas de las $p+q$ variables aleatorias

$$\frac{1}{N}\boldsymbol{S} = \widehat{Var\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}} = \begin{pmatrix}\widehat{\boldsymbol{\Sigma_{11}}} & \widehat{\boldsymbol{\Sigma_{12}}}\\ \widehat{\boldsymbol{\Sigma_{21}}} & \widehat{\boldsymbol{\Sigma_{22}}}\end{pmatrix}$$

Las variables y correlaciones canónicas muestrales siguen la misma construcción y supuestos de las poblacionales, entonces es necesario estimar las raíces de los coeficientes de correlación canónica con la información de la muestra, es decir, estimar los valores de $\boldsymbol{\Sigma_{11}}$, $\boldsymbol{\Sigma_{12}}$, $\boldsymbol{\Sigma_{22}}$ por máxima verosimilitud, o bien, encontrar $\frac{1}{N}\boldsymbol{S_{11}}$, $\frac{1}{N}\boldsymbol{S_{12}}$, $\frac{1}{N}\boldsymbol{S_{22}}$ respectivamente, para obtener los coeficientes muestrales.

Sea la matriz $\boldsymbol{C^*}$ de orden $p \times q$

$$\boldsymbol{C^*} = \left(\frac{1}{N}\boldsymbol{S_{11}}\right)^{-\frac{1}{2}}\left(\frac{1}{N}\boldsymbol{S_{12}}\right)\left(\frac{1}{N}\boldsymbol{S_{22}}\right)^{-\frac{1}{2}},$$

por el **Teorema 1**, existen dos matrices ortogonales $\boldsymbol{A^*}$ y $\boldsymbol{B^*}$ de orden p y q respectivamente tales que

$$\boldsymbol{C^* = A^* F B^{*\prime}}$$

donde $\boldsymbol{F}$ es la matriz estimada de $\boldsymbol{P}$ *(matriz de correlaciones poblacionales).*

Sean las transformaciones $\boldsymbol{u^* = Gx}$ y $\boldsymbol{v^* = Hy}$, donde $\boldsymbol{G}$ y $\boldsymbol{H}$ son de orden $p \times p$ y $q \times q$ respectivamente y cuya matriz de varianzas y covarianzas es:

$$Var \begin{pmatrix} \boldsymbol{u^*} \\ \boldsymbol{v^*} \end{pmatrix} = \begin{pmatrix} \boldsymbol{GS_{11}G'} & \boldsymbol{GS_{12}H'} \\ \boldsymbol{HS_{21}G'} & \boldsymbol{HS_{22}H'} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I_p} & \boldsymbol{F} \\ \boldsymbol{F'} & \boldsymbol{I_q} \end{pmatrix}$$

donde

$$\boldsymbol{G = A^{*\prime}} \left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{-\frac{1}{2}},$$
$$\boldsymbol{H = B^{*\prime}} \left(\tfrac{1}{N}\boldsymbol{S_{22}}\right)^{-\frac{1}{2}}$$

multiplicando por la izquierda la igualdad de $\boldsymbol{C^*}$ por $\boldsymbol{A^{*\prime}}$, se tiene:

$$\begin{aligned} \boldsymbol{A^{*\prime}C^*} &= \boldsymbol{A^{*\prime}A^*FB^{*\prime}} \\ &= \boldsymbol{FB^{*\prime}} \end{aligned}$$

Sustituyendo $\boldsymbol{A^*}$ y $\boldsymbol{B^*}$ en términos de $\boldsymbol{G}$ y $\boldsymbol{H}$ respectivamente, la expresión anterior queda de la siguiente manera:

$$\boldsymbol{G}\left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{-\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{12}}\right)\left(\tfrac{1}{N}\boldsymbol{S_{22}}\right)^{-\frac{1}{2}} = \boldsymbol{FH}\left(\tfrac{1}{N}\boldsymbol{S_{22}}\right)^{\frac{1}{2}}$$

multiplicando por la derecha ambos lados de la expresión anterior por $\left(\frac{1}{N}\boldsymbol{S_{11}}\right)^{\frac{1}{2}}$ se obtiene

$$\boldsymbol{G}\left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{-\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{12}}\right)\left(\tfrac{1}{N}\boldsymbol{S_{22}}\right)^{-\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{22}}\right)^{\frac{1}{2}} = \boldsymbol{FH}\left(\tfrac{1}{N}\boldsymbol{S_{22}}\right)$$

de donde

$$\begin{aligned} \boldsymbol{GS_{12}} &= \boldsymbol{FHS_{22}} \\ \boldsymbol{GS_{12}S_{22}^{-1}} &= \boldsymbol{FH} \end{aligned}$$

Transponiendo ambos lados de la segunda ecuación anterior se obtiene:

$$\boldsymbol{S_{22}^{-1}S_{21}G' = H'F}$$

que se expresa de la siguiente manera

$$\boldsymbol{S_{22}^{-1}S_{21}g_i} = \begin{cases} \boldsymbol{h_i}r_i & \text{para } i = 1, 2, \ldots, p \\ 0 & \text{para } i = p+1, \ldots, q \end{cases} \tag{1.16}$$

donde $\boldsymbol{g_i}$ es la $i-ésima$ columna de $\boldsymbol{G}$, $\boldsymbol{h_i}r_i$ es el $i-ésimo$ elemento del vector de tamaño $q \times 1$ distinto de cero y todos los demás elementos iguales a cero.

De manera similar, multiplicando por la derecha por $\boldsymbol{B^*}$ en la igualdad de $\boldsymbol{C^*}$, se tiene:

$$\begin{aligned} \boldsymbol{C^*B^*} &= \boldsymbol{A^*FB^{*\prime}B^*} \\ \boldsymbol{C^*B^*} &= \boldsymbol{A^*F} \end{aligned}$$

Sustituyendo $\boldsymbol{A^*}$ y $\boldsymbol{B^*}$ en términos de $\boldsymbol{G}$ y $\boldsymbol{H}$ respectivamente, la expresión anterior queda de la siguiente manera:

$$\begin{aligned} \left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{-\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{12}}\right)\left(\tfrac{1}{N}\boldsymbol{S_{22}}\right)^{-\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{22}}\right)^{\frac{1}{2}}\boldsymbol{H'} &= \left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{\frac{1}{2}}\boldsymbol{G'F} \\ \left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{-\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{-\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{12}}\right)\boldsymbol{H'} &= \left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{-\frac{1}{2}}\left(\tfrac{1}{N}\boldsymbol{S_{11}}\right)^{\frac{1}{2}}\boldsymbol{G'F} \\ \boldsymbol{S_{11}^{-1}S_{12}H'} &= \boldsymbol{G'F} \end{aligned}$$

la última expresión anterior se escribe de la siguiente manera

$$\boldsymbol{S_{11}^{-1}S_{12}h_i} = \begin{cases} \boldsymbol{g_i}r_i & \text{para } i = 1, 2, \ldots, p \\ 0 & \text{para } i = p+1, \ldots, q \end{cases} \tag{1.17}$$

donde $\boldsymbol{h_i}$ es la $i-ésima$ columna de $\boldsymbol{H}$, $\boldsymbol{g_i}r_i$ es el $i-ésimo$ elemento del vector de tamaño $p \times 1$ distinto de cero y todos los demás elementos iguales a cero.

Por lo tanto las ecuaciones muestrales son expresadas en forma matricial de la siguiente manera

$$\begin{bmatrix} -r_i\boldsymbol{S_{11}} & \boldsymbol{S_{12}} \\ \boldsymbol{S_{21}} & -r_i\boldsymbol{S_{22}} \end{bmatrix} \begin{bmatrix} \boldsymbol{g_i} \\ \boldsymbol{h_i} \end{bmatrix} = 0 \quad i = 1, 2, \ldots, p$$

Por construcción se sabe que $r_i \neq 0$, en la ecuación (2.15), dividiendo ambos lados de la ecuación entre r_i se tiene

$$\boldsymbol{h_i} = \boldsymbol{S_{22}^{-1}S_{21}g_i}\frac{1}{r_i}$$

sustituyendo $\boldsymbol{h_i}$ en la ecuación (2.16):

$$\begin{aligned} \boldsymbol{S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}g_i}\frac{1}{r_i} &= r_i\boldsymbol{g_i} \\ \boldsymbol{S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}g_i} &= r_i^2\boldsymbol{g_i} \end{aligned}$$

en un desarrollo análogo en la obtención de las correlaciones canónicas poblacionales se tiene:

$$\left[\boldsymbol{S_{12}S_{22}^{-1}S_{21}} - \boldsymbol{S_{11}}r_i^2\right]\boldsymbol{g_i} = 0 \quad i = 1, \ldots, p$$
$$\left[\boldsymbol{S_{21}S_{11}^{-1}S_{12}} - \boldsymbol{S_{22}}r_i^2\right]\boldsymbol{h_i} = 0 \quad i = 1, \ldots, p$$

multiplicando la primera de las dos ecuaciones anteriores por $\boldsymbol{g_i'}$

$$\boldsymbol{g_i'S_{12}S_{22}^{-1}S_{21}g_i} - \boldsymbol{g_i'S_{11}}r_i^2\boldsymbol{g_i} = 0$$
$$\boldsymbol{g_i'S_{12}S_{22}^{-1}S_{21}g_i} = \mathrm{r}_i^2\boldsymbol{g_i'S_{11}g_i}$$

de donde

$$\frac{\boldsymbol{g_i'S_{12}S_{22}^{-1}S_{21}g_i}}{\boldsymbol{g_i'S_{11}g_i}} = r_i^2 \quad i = 1, \ldots, p.$$

De manera similar despejando a $\boldsymbol{g_i}$ de la expresión (1.17) y sustituyendola en la ecuación (1.16), se tiene a r_i^2 en términos de $\boldsymbol{h_i}$, la cual está dada por

$$\frac{\boldsymbol{h_i'S_{21}S_{11}^{-1}S_{12}h_i}}{\boldsymbol{h_i'S_{22}h_i}} = r_i^2 \quad i = 1, \ldots, q.$$

Por lo tanto, r_i es el coeficiente de correlación múltiple muestral entre $\boldsymbol{g_i'x}$ y $\boldsymbol{h_i'y}$ que satisface las siguientes propiedades:

i. $\boldsymbol{S_{12}S_{22}^{-1}S_{21} = DFD'}$

ii. $\boldsymbol{S_{11} = DD'}$

donde

$$\boldsymbol{D^{-1}} = \boldsymbol{G} = [g_1 g_2 \cdots g_p]$$
$$\boldsymbol{F} = diag\left(r_1^2, r_2^2, \ldots, r_p^2\right)$$

Nótese que la $i-ésima$ columna de D^{-1} es el vector de coeficientes $\boldsymbol{g_i}$ de la $i-ésima$ variable canónica $\boldsymbol{g_i'x}$ de la muestra.

Entonces, por propiedades de determinantes se tiene que

$$\frac{\left|\boldsymbol{S}\right|}{\left|\boldsymbol{S_{11}}\right|\left|\boldsymbol{S_{22}}\right|} = \frac{\left|\boldsymbol{S_{22}}\right|\left|\boldsymbol{S_{11}} - \boldsymbol{S_{12}S_{22}^{-1}S_{21}}\right|}{\left|\boldsymbol{S_{11}}\right|\left|\boldsymbol{S_{22}}\right|}$$

$$= \frac{\left|\boldsymbol{DD'} - \boldsymbol{DFD'}\right|}{\left|\boldsymbol{DD'}\right|}$$

$$= \left|\boldsymbol{I} - \boldsymbol{F}\right| = \prod_{i-1}^{p}\left(1 - r_i^2\right) \tag{1.18}$$

El resultado del cociente de determinantes corresponde a la relación entre los dos vectores originales $\boldsymbol{x}$ y $\boldsymbol{y}$; lo importante de éste, es que puede ser expresado en función de los coeficientes de las correlaciones muestrales.

Este cociente es conocido como la estadística $\boldsymbol{\Lambda}$ *de Wilks*, que se desarrolla en el próximo capítulo.

Capítulo 2

Pruebas de Significancia Estadística

2.1. Pruebas

El **AC** parte de suponer la existencia de al menos una relación entre el conjunto de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$ ya que es de suma importancia para la derivación de las variables y correlaciones canónicas.

Las pruebas que se desarrollan a continuación son utilizadas para contrastar cuántas relaciones son estadísticamente significativas, para después probar que el resto que no se considera, sean no significativas para la construcción de las combinaciones lineales del **AC**.

Una vez realizadas las pruebas anteriores, se procede a calcular la proporción de la variabilidad total explicada por las combinaciones lineales que resultan estadísticamente significativas del **AC**.

2.1.1. Prueba de Wilks usando la aproximación de Bartlett

Existen pruebas estadísticas para contrastar la hipótesis nula definida como la ausencia de relación entre los conjuntos de variables; una de éstas, es la aproximación a una distribución ji-cuadrada, atribuida a Bartlett (1941).

La prueba consiste en encontrar el número exacto de combinaciones lineales estadísticamente significativas que hay entre los conjuntos de variables para el **AC**, es decir, probar cuántas parejas de variables están relacionadas dentro de la submatriz de correlaciones $\boldsymbol{R_{XY}}$.

Esta prueba se realiza en dos partes; la primera consiste en verificar si existe relación significativa entre los conjuntos de variables y saber cuántas relaciones son significativas; posteriormente se realiza la segunda parte de la prueba que está basada en mostrar que el resto de las relaciones sean no significativas para el **AC**. Esta prueba está basada en las propiedades de determinantes y fue desarrollada por Samuel Stanley Wilks, por tanto, es conocida como la $\boldsymbol{\Lambda}$ de Wilks.

Desarrollo:

Primera parte

Se postula la hipótesis nula:

$\boldsymbol{H_o}$: Las correlaciones canónicas $\rho_1, \rho_2, \ldots, \rho_k$ con k igual al $min\{p, q\}$, entre los vectores $\boldsymbol{x}$ y $\boldsymbol{y}$ son nulas, es decir

$$\rho_1 = \rho_2 = \cdots = \rho_k = 0$$

contra la hipótesis alternativa:

$\boldsymbol{H_a}$: Las primeras s correlaciones canonicas son no nulas y el resto lo es, es decir

$$\rho_1 \neq 0, \rho_2 \neq 0, \ldots, \rho_s \neq 0, \ \rho_{s+1} = \rho_{s+2} = \ldots = \rho_k = 0$$

Para realizar la prueba se utiliza el resultado del cociente de determinantes de la ecuación (1.18) que es la estadística de Wilks, denotada por

$$\prod_{i-1}^{p}\left(1-r_i^2\right)$$

donde $r_1^2 > r_2^2 > \cdots > r_k^2$ son el cuadrado de los coeficientes de las correlaciones canónicas muestrales, a partir de la cual, Bartlett desarrolló una aproximación de la distribución de una función de la estadística dada por:

$$\boldsymbol{B_1} = -\left[n - \frac{1}{2}\left(p+q+1\right)\right] ln \prod_{i=1}^{s}\left(1-r_i^2\right) \sim \boldsymbol{\chi}_{\nu}^2$$

donde $\nu = pq-(p-s)(q-s)$ son los grados de libertad y n es el número de casos.

Para valores grandes de $\boldsymbol{B_1}$, se rechaza la hipótesis nula $\boldsymbol{H_o}$, es decir, dado un nivel de significancia α, la región de rechazo es la que cumple

$$\boldsymbol{B_1} > \boldsymbol{\chi}_{\nu}^{2\,(1-\alpha)}.$$

Si se rechaza $\boldsymbol{H_o}$, se concluye que existen s relaciones linealmente significativas entre los conjuntos, donde $s \geq 1$, es decir, se tiene que ρ_1 es la correlación canónica más grande, ρ_2 la siguiente más grande y así sucesivamente hasta llegar a la ρ_s significativas para el modelo; con esto se concluye la primera parte de la prueba y se procede a derivar la segunda.

Segunda parte

Cuando existen s relaciones significativas, esto es, que existen s coeficientes de correlación distintos de cero, se tiene que probar que el resto de las $k-s$ relaciones sean no significativas para el **AC**, ahora bien, de la expresión $\prod_{i=1}^{k}(1-r_i^2)$, se toman los $k-s$ últimos elementos del producto:

$$\prod_{i=s+1}^{k}\left(1-r_i^2\right)$$

en este caso la aproximación es

$$\boldsymbol{B_2} = -\left[n-\frac{1}{2}(p+q+1)\right] ln \prod_{i=s+1}^{k}\left(1-r_i^2\right) \sim \boldsymbol{\chi}_{\nu}^2$$

donde $\nu=(p-s)(q-s)$ son los grados de libertad y $s=1,2,\ldots,k-1$.

Se postula entonces, la hipótesis nula:

$\boldsymbol{H_{o_2}}$: Las correlaciones canónicas $\rho_{s+1},\rho_{s+2},\ldots,\rho_k$ con $s<k$, entre los vectores $\boldsymbol{x}$ y $\boldsymbol{y}$ son nulas, es decir

$$\rho_{s+1}=0,\rho_{s+2}=0,\ldots,\rho_k=0$$

contra la hipótesis alternativa:

$\boldsymbol{H_{a_2}}$: La siguiente $s+1$ correlación canónica es no nula, es decir $\rho_{s+1}\neq 0$

Así con esto, se concluye bajo $\boldsymbol{H_{a_2}}$ que hay evidencia suficiente para decir que aumenta una correlación significativa, es decir, que entonces hay $s+1$ dependencias

lineales significativas entre los dos conjuntos. Bartlett afirma que la prueba es más exacta para tamaños de la muestra n mayores que 50. [5]

En la siguiente tabla se muestran los grados de libertad y la aproximación $\boldsymbol{\chi}^2_\nu$ correspondientes a cada número de raíces canónicas muestrales.

Raíces	**Grados de libertad**	**Aproximación a una $\boldsymbol{\chi}^2_\nu$**
Primeras s	$pq-(p-s)(q-s)$	$-\left[n-\frac{1}{2}(p+q+1)\right] ln \prod_{i=1}^{s}\left(1-r_i^2\right)$
Restantes $k-s$	$(p-s)(q-s)$	$-\left[n-\frac{1}{2}(p+q+1)\right] ln \prod_{i=s+1}^{k}\left(1-r_i^2\right)$
Total k	pq	$-\left[n-\frac{1}{2}(p+q+1)\right] ln\boldsymbol{\Lambda}(n,p,q)$

En teoría no hay un procedimiento específico para identificar el valor exacto de s, pero en la práctica se utilizan valores para $s=1,2,3,\ldots$, y así aplicar la prueba sucesivamente hasta llegar a un punto en el cual $\boldsymbol{H_{o_2}}$, no se rechace; con esto, se puede concluir que sólo $\rho_1,\rho_2,\ldots,\rho_s$ correlaciones canónicas son distintas de cero, y el resto son nulas.

Por lo tanto el número total de variables entre $\boldsymbol{X}$ y $\boldsymbol{Y}$ pueden ser explicadas con s combinaciones lineales con $s \ll p+q$ y las s combinaciones lineales tienen las siguientes propiedades:

i) La predicción de regresión múltiple de cualquier variable $\boldsymbol{Y}$ con las variables $\boldsymbol{X}$ es una función lineal de las s combinaciones.

ii) Cuando $k=s$ entonces no hay correlación nula entre los conjuntos $\boldsymbol{X}$ y $\boldsymbol{Y}$.

iii) Las s correlaciones canónicas no están mutuamente correlacionadas entre sí.

[5] DARLINGTON, "Canonical Variate Analysis and Related Techniques",[7] pp. 10-11

Además el **AC** es completamente simétrico entre los conjuntos $\boldsymbol{X}$ y $\boldsymbol{Y}$, por lo que es posible encontrar s combinaciones lineales de las variables del conjunto $\boldsymbol{Y}$ con las mismas propiedades cuando el papel de los conjuntos $\boldsymbol{X}$ y $\boldsymbol{Y}$ son invertidos.

2.1.2. Estadística de Rao en relación a la lambda de Wilks con base en una aproximación a una distribución F

Es una prueba alternativa a la estadística de Wilks pero más potente por ser una mejor aproximación; entonces las hipótesis son como la prueba anterior.

Se postula la hipótesis nula como:

$\boldsymbol{H_o}$: Las correlaciones canónicas $\rho_1, \rho_2, \ldots, \rho_k$ con k igual al $min\{p, q\}$, entre los vectores $\boldsymbol{x}$ y $\boldsymbol{y}$ son cero, es decir

$$\rho_1 = \rho_2 = \cdots = \rho_k = 0$$

contra la hipótesis alternativa:

$\boldsymbol{H_a}$: $\rho_1 \neq 0, \rho_2 \neq 0, \ldots, \rho_s \neq 0 \quad \rho_{s+1} = \rho_{s+2} = \cdots = \rho_k = 0$

Para llevar a cabo la prueba se utiliza la estadística:

$$\boldsymbol{R_w} = \frac{1 - W^{\frac{1}{t}}}{W^{\frac{1}{t}}} \cdot \frac{mt - \left(\frac{pq}{2}\right) + 1}{pq} \sim \boldsymbol{F}_{(\upsilon,\nu)}$$

donde

$$m = n - \frac{1}{2}(p + q + 1)$$

$$t = [(p^2q^2 - 4)/(p^2 + q^2 - 5)]^{\frac{1}{2}}$$

n es el número de casos

con W como en la prueba anterior, es decir, $W = \prod_{i=1}^{k} (1 - r_i^2)$

Entonces $\boldsymbol{R_w}$ sigue una distribución $\boldsymbol{F}$ con (υ, ν) grados de libertad, donde $\upsilon = pq$ y $\nu = mt - \left(\frac{pq}{2}\right) + 1$.

Para valores grandes de $\boldsymbol{R_w}$, se rechaza la hipótesis nula $\boldsymbol{H_o}$ a un nivel de significancia α, sí

$$\boldsymbol{R_w} > \boldsymbol{F}_{(\upsilon,\nu)}^{1-\alpha}$$

Si se rechaza $\boldsymbol{H_o}$, se concluye que existen s correlaciones significativas entre los conjuntos, donde $s \geq 1$.

2.2. Análisis de variabilidad bajo la hipótesis alternativa

Dado el número de dependencias lineales estadísticamente significativas que existen entre los conjuntos $\boldsymbol{X}$ y $\boldsymbol{Y}$ bajo la hipótesis alternativa y utilizándo los resultados matriciales derivados del desarrollo del **AC**; surge la siguiente pregunta:

¿Qué tanta variabilidad del conjunto de variables dependientes puede ser representada por el conjunto de variables independientes?

Es importante este criterio debido a que se puede conocer la variabilidad total que representa cada variable del conjunto $\boldsymbol{Y}$. Para medir la variabilidad entre conjuntos, se utiliza el coeficiente de traza, desarrollado por Hopper (1959,1962), denotado por $\overline{t^2}$ y se obtiene a partir de la matriz de correlaciones canónicas muestrales, definida anteriormente como $\boldsymbol{F} = diag(r_1^2, r_2^2, \ldots, r_p^2)$, de la siguiente manera

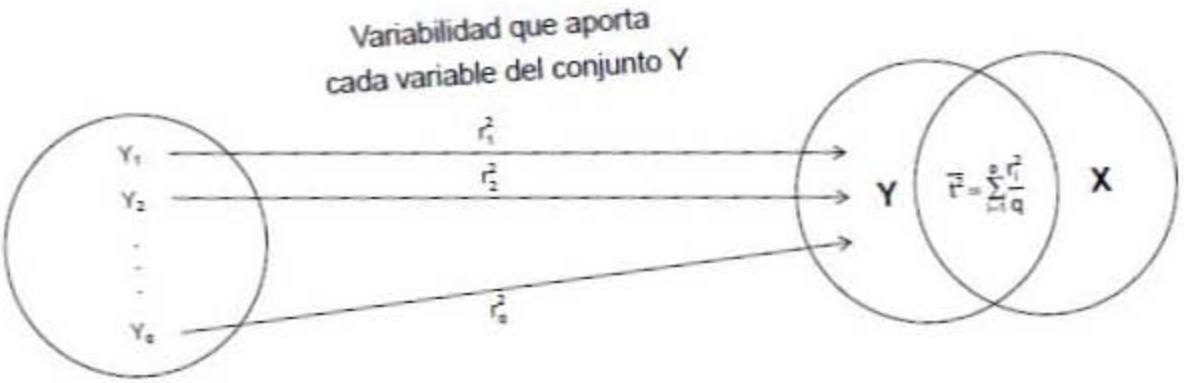
$$\overline{t^2} = \frac{\sum_{i=1}^{p} r_i^2}{q}$$

que es la suma de los cuadrados de los coeficientes de las correlaciones canónicas muestrales entre el número de variables del conjunto $\boldsymbol{y}$, entonces, la traza es interpretada como la proporción de variabilidad de $\boldsymbol{Y}$.

Por ser una proporción, el recorrido de valores de la traza está entre cero y uno; es cercano a cero cuando la submatriz de correlación muestral entre $\boldsymbol{x}$ y $\boldsymbol{y}$ tiene todos sus valores cercanos a cero, y se aproxima a uno cuando cada una de las variables de $\boldsymbol{y}$ es combinación lineal de las variables de $\boldsymbol{x}$.

Dicho cociente es fácil de calcular pero difícil de interpretar ya que se tiene que analizar la proporción que cada combinación lineal de $\boldsymbol{x}$ y $\boldsymbol{y}$ aporta a la variabilidad total.

Figura 5: ***Representación gráfica del coeficiente de traza.***



Capítulo 3

Regiones de Confianza en el Análisis Canónico

3.1. Preámbulo

Para las regiones de confianza en el **AC**, se tienen diferentes grupos muestrales de diferentes poblaciones con parámetros desconocidos, por ende, la derivación de las regiones de confianza se realiza con parámetros estimados de la población; las regiones de confianza se construyen a partir de aproximaciones de la distribución muestral, pero el sustituir parámetros por estimadores crea un serio problema con la interpretación de las regiones de confianza ya que al tener muestras grandes y/o muchos grupos, los datos a menudo suelen ser parecidos y esto implica parámetros estimados repetidos.

Para hacer una gráfica de una región del $(1-\alpha)\times 100\,\%$ de confianza es necesario tener el **AC** ya desarrollado, ya que se necesita tener la interpretación o la referencia de los vectores canónicos para que cada región esté basada en estos; sin embargo, en la práctica no es posible analizar a toda la población, por lo tanto, la construcción de los ejes de referencia deben ser obtenidos forzosamente de las

variables canónicas muestrales. El inconveniente es que si se utilizan diferentes muestras de la población en estudio y se lleva a cabo un análisis de variables canónicas sobre cada nueva muestra, se obtendrán cada vez nuevos y diferentes vectores canónicos, que se sabe, estos vectores canónicos serán los ejes para construir las regiones de confianza.

La derivación de todos los resultados que se presentan a continuación, se encuentra con mayor detalle en el artículo de KRZANOWSKI [12] y por ende se presentan únicamente los resultados.

3.2. Desarrollo matemático

Sean $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \ldots, \boldsymbol{P}_g$, g poblaciones de las cuales se han tomado g muestras de tamaños $n_1, n_2, n_3, \ldots, n_g$ respectivamente; correspondientes a p variables, es decir:

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix},$$

para cada n_i individuos de cada grupo, la matriz $\boldsymbol{X}$ es de la siguiente manera

$$\boldsymbol{X} = \begin{pmatrix} X_{1,11} & X_{1,12} & \cdots & X_{1,1p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1,n_1 1} & X_{1,n_1 2} & \cdots & X_{1,n_1 p} \\ \hline X_{2,11} & X_{2,12} & \cdots & X_{2,1p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{2,n_2 1} & X_{2,n_2 2} & \cdots & X_{2,n_2 p} \\ \hline \\ \vdots & \vdots & \vdots & \vdots \\ \\ \hline X_{g,11} & X_{g,12} & \cdots & X_{g,1p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{g,n_g 1} & X_{g,n_g 2} & \cdots & X_{g,n_g p} \end{pmatrix} \begin{matrix} \boldsymbol{P}_1 \\ \\ \boldsymbol{P}_2 \\ \\ \\ \\ \vdots \\ \\ \boldsymbol{P}_g \end{matrix},$$

Para simplificar el número de subíndices de los elementos de la matriz $\boldsymbol{X}$, se utilizan los n_i individuos en el $i-ésimo$ grupo y se define a x_{ij} el vector de las p mediciones tomadas sobre el $j-ésimo$ individuo en el $i-ésimo$ grupo, así $j=1,\ldots,n_i$, $i=1,\ldots,g$ y $n=\sum_{i=1}^{g} n_i$.

Debido al tamaño de la matriz $\boldsymbol{X}$, el vector de medias del $i-ésimo$ grupo es una expresión difícil de entender, debido a que el vector se centra en el $i-ésimo$ grupo y una vez centrado se van sumando entrada a entrada los x_{ij} vectores del $i-ésimo$ grupo, por ejemplo, para el primer grupo, la suma de las entradas de los vectores es:

$$(x_{1,11}, x_{1,12}, \ldots, x_{1,1p}) + (x_{1,21}, x_{1,22}, \ldots, x_{1,2p}) + \cdots + (x_{1,n_1 1}, x_{1,n_1 2}, \ldots, x_{1,n_1 p})$$

$$=$$

$$\begin{aligned}((\mathrm{x}_{1,11} + x_{1,21} + \cdots + x_{1,n_1 1}),& \\ (x_{1,12} + x_{1,22} + \cdots + x_{1,n_1 2}), \cdots ,& \\ (x_{1,1p} + x_{1,2p} + \cdots + x_{1,n_1 p}))&\end{aligned}$$

y cada entrada del vector de medias tiene n_i observaciones, sólo se divide cada entrada entre n_i.

Entonces el ***vector de medias*** del $i-ésimo$ grupo es:

$$\overline{x_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij};$$

la **media total** está dada por

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{g} n_i\overline{x_i} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij};$$

la **matriz de dispersión dentro de los grupos** está dada por

$$\boldsymbol{W} = \frac{1}{n-g} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i}) (x_{ij} - \overline{x_i})' ;$$

y la **matriz de covarianza entre grupos** está dada por

$$\boldsymbol{B} = \frac{1}{g-1} \sum_{i=1}^{g} n_i (\overline{x_i} - \overline{x}) (\overline{x_i} - \overline{x})' .$$

La matriz $\boldsymbol{W}$ mide que tan dispersos están los datos dentro de los grupos y la matriz $\boldsymbol{B}$ mide la dispersión entre los grupos. El desarrollo de las matrices es necesario porque a partir de éstas, se construyen las regiones de confianza.

Por ejemplo, si las variables canónicas son representadas en dos dimensiones, las regiones de confianza son círculos o elipses, y esferas o elipsoides si se encuentran en tres dimensiones, en caso de que sea mayor a tres no se tiene manera adecuada de representarlas gráficamente.

Utilizando los resultados algebraicos del **Capítulo 1**, en particular la ecuación $(\boldsymbol{B} - l\boldsymbol{W})\, a = 0$, y dado que $\boldsymbol{A} = (a_1, \ldots, a_s)$ es el conjunto de vectores canónicos y $\boldsymbol{L} = diag(l_1, \ldots, l_s)$ el conjunto de raíces canónicas que satisfacen la siguiente igualdad

$$\boldsymbol{BA} = \boldsymbol{WAL};$$

Para seguir con la derivación de las regiones de confianza, lo primero es encontrar el centro de cada región.

Sean s variables canónicas $\boldsymbol{Z'} = (Z_1, \ldots, Z_s)$, por cada una de las variables X_i, definidas como combinaciones lineales $Z_k = a_k'(\boldsymbol{X} - \overline{x})$, entonces los valores

de las variables canónicas son z_{ijk} y *las medias grupales canónicas* están dadas por $\overline{z}_{jk}$ y se obtienen a partir de la sustitución de x_{ij} y $\overline{x}_j$ en la ecuación de las combinaciones lineales.

Supóngase que las variables en el $i-ésimo$ grupo provienen de una muestra aleatoria de su correspondiente población $(i = 1, \ldots, g)$, y que todas las variables siguen una distribución normal; específicamente, los x_{ij} son independientes e idénticamente distribuidos $\boldsymbol{N}(\boldsymbol{\mu_i}, \boldsymbol{\Sigma})$ para $i = 1, \ldots, g$ y $j = 1, \ldots, n_i$; con $\boldsymbol{\mu_i}$ vector de medias de todas las poblaciones.

En consecuencia:

$$a_i \sim \boldsymbol{N}\left(v_i, n_i^{-1} I\right) \text{ con } i = 1, \ldots, s \text{ y}$$
$$\overline{z_i} = (\overline{z}_{i1}, \ldots, \overline{z}_{is})' \sim \boldsymbol{N}\left(v_i, n_i^{-1} I\right);$$

donde v_i son las posiciones de las *medias* en los vectores canónicos a_i como ejes, así, $v_i = \boldsymbol{A'}(\mu_i - \boldsymbol{\mu})$, entonces:

$$\Phi = n_i (\overline{z_i} - v_i)' (\overline{z_i} - v_i) \sim \chi_s^2$$

por lo tanto, las regiones de confianza basadas en Φ están centradas en $\overline{z_i}$.

La distribución exacta de $\overline{z}_{ij}$ es difícil de obtener, aun con la hipótesis de normalidad de las a_i, ya que las $\overline{z}_{ij}$ son sumas de p productos de parejas de variables normales, por lo que un solo producto tiene una expresión complicada, sin embargo, para la práctica con un número p grande, se utiliza el Teorema Central del Límite, entonces los $\overline{z}_{ij}$ tienen una aproximación normal asintótica.

Sea $\Phi^* = \boldsymbol{\Sigma} n_i (\mu_i - \boldsymbol{\mu})(\mu_i - \boldsymbol{\mu})'$, con los valores de $\boldsymbol{l_i}$ y a_i estimados por λ_i y α_i respectivamente, con $i = 1, \ldots, s$ valores y vectores propios que satisfacen la ecuación

$$(\Phi^* - \lambda \boldsymbol{\Sigma})\, a = 0.$$

Anderson, T. W. (1984) dio distribuciones asintóticas de los valores α_i y λ_i, cuando $(n-g)\boldsymbol{W}$ sigue una distribución *Wishart* con $f_1 = (n-g)$ grados de libertad y parámetro $\boldsymbol{\Sigma}$, mientras $(g-1)\boldsymbol{B}$ sigue una distribución *Wishart* con $f_2 = (g-1)$ grados de libertad y parámetro Φ^*.

Anderson afirma que si $f_1, f_2 \to \infty$ de tal manera que se puede encontrar una $\gamma > 0$ tal que $\dfrac{f_2}{f_1} \to \gamma$, entonces:

$$E(l_k) = \lambda_k, E(a_k) = \alpha_k, Var(l_k) = \frac{2\lambda_k^2(1+\gamma)}{\gamma f_1} \tag{3.1}$$

$$\Omega = Var(a_k) = \frac{1}{f_1}\sum_{j\neq k=1}^{s} \frac{\lambda_k(\lambda_j + \gamma\lambda_k)}{\gamma(\lambda_j - \lambda_k)^2}\alpha_j\alpha_j' + \frac{1}{2f_1}\alpha_k\alpha_k' \tag{3.2}$$

$$Cov(a_i, a_j) = -\frac{\lambda_i\lambda_j(1+\gamma)}{f_1\gamma(\lambda_i - \lambda_j)^2}\alpha_i\alpha_j' \text{con } i \neq j \tag{3.3}$$

Los valores λ_i y a_i se distribuyen asintóticamente normal y por simplicidad se define a $\dfrac{f_2}{f_1} = \gamma$.

3.3. Regiones de confianza

Sea $\overline{z}_{ik}$ la media del $i-$*ésimo* grupo sobre la $k-$*esima* variable canónica, es decir,

$$\overline{z}_{ik} = a_k'(\overline{x}_i - \overline{x}) = a_{k1}(\overline{x}_{i1} - \overline{x}_1) + \cdots + a_{kp}(\overline{x}_{ip} - \overline{x}_p).$$

Para a_k fijo, se sabe que $\overline{z}_{ik} \sim \boldsymbol{N}(v_{ik}, n_i^{-1})$, donde $v_{ik} = a_k'(\mu_i - \boldsymbol{\mu})$ y $a_k = (a_{k1}, \ldots, a_{kp})' \sim \boldsymbol{N}(\alpha_k, \Omega)$, con Ω definido en la ecuación (3.2) y sea

$\zeta_{ik} = \alpha_k'(\mu_i - \boldsymbol{\mu})$ y $\zeta_i = (\zeta_{i1}, \zeta_{i2}, \ldots, \zeta_{is})'$.

Utilizando las propiedades de la esperanza condicional para $\overline{z}_{ik}$ se tiene

$$E(\overline{z}_{ik}) = E\left[E(\overline{z}_{ik} \mid a_k)\right] = E\left[a_k'(\mu_i - \boldsymbol{\mu})\right] = \alpha_k'(\mu_i - \boldsymbol{\mu})$$

$$\therefore E(\overline{z}_{ik}) = \zeta_{ik} \tag{3.4}$$

El valor ζ_{ik} es la posición de la $i - ésima$ media poblacional con la $k - ésima$ variable canónica poblacional (α_k), como eje.

Dada $E(\overline{z}_{ik})$ y que $\overline{z}_{ik}$ sigue una distribución normal, es necesario encontrar el valor de la $Var(\overline{z}_{ik})$, de esta forma, se tienen los parámetros de la distribución normal de $\overline{z}_{ik}$.

Utilizando propiedades de la esperanza y varianza, y la ecuación (3.4) se tiene

$$\begin{aligned} Var(\overline{z}_{ik}) &= E_{a_k}\{Var(\overline{z}_{ik} \mid a_k)\} + Var_{a_k}\left[E(\overline{z}_{ik} \mid a_k)\right] \\ &= E_{a_k}\left(\frac{1}{n_i}\right) + Var_{a_k}\left[a_k'(\mu_i - \boldsymbol{\mu})\right] \\ &= \frac{1}{n_i} + (\mu_i - \boldsymbol{\mu})'\Omega(\mu_i - \boldsymbol{\mu}) \end{aligned} \tag{3.5}$$

La *matriz de dispersión* de las regiones de confianza, se obtiene utilizando la ecuación (3.3),

$$Cov(\overline{z}_{ij}, \overline{z}_{ik}) = -\frac{\theta_j\theta_k(1+\gamma)}{(n-g)\gamma(\theta_j - \theta_k)^2}\alpha_j\alpha_k' \text{ con } k \neq j \tag{3.6}$$

donde $\theta_i = \dfrac{(1+\boldsymbol{l_i})}{(g-1)}$ son valores que se sustituyen por λ_j, a_i por α_i que cumplen con la ecuación $(\Phi^* - \lambda\boldsymbol{\Sigma})\alpha = 0$ y cada uno sigue una distribución *Wishart* con $f_1 = (n-g)$ grados de libertad.

Por lo tanto $\overline{z}_{ij}$ sigue una distribución normal con

$$E(\overline{z}_{ik}) = \zeta_{ik}$$

$$Var(\overline{z}_{ik}) = \frac{1}{n_i} + (\mu_i - \boldsymbol{\mu})'\Omega(\mu_i - \boldsymbol{\mu})$$

$$Cov(\overline{z}_{ij}, \overline{z}_{ik}) = -\frac{\theta_j\theta_k(1+\gamma)}{(n-g)\gamma(\theta_j - \theta_k)^2}\alpha_j\alpha_k' \text{ con } k \neq j$$

3.3.1. Ejemplo teórico sobre dos ejes de variables canónicas

Para ilustrar el desarrollo de las regiones de confianza, supóngase que se tiene el grupo de medias y los primeros dos ejes de variables canónicos, es decir, dado el vector de medias canónico para el $i-ésimo$ grupo que está denotado por $(\overline{z}_{i1}, \overline{z}_{i2})'$, este vector tiene una distribución normal bivariada con media $(\zeta_{i1}, \zeta_{i2})'$ y matriz de dispersión:

$$\begin{bmatrix} v_{i1} & c_i \\ c_i & v_{i2} \end{bmatrix}$$

donde v_{i1} y v_{i2} están dados por (3.5) con $k = 1, 2$ respectivamente, mientras c_i está dada por (3.6) con $k, j = 1, 2$ respectivamente; entonces se busca una región de confianza para el vector $(\zeta_{i1}, \zeta_{i2})'$, sobre la $i-ésima$ posición de las medias con las primeras dos variables canónicas como ejes.

Una vez que se tiene la derivación de las regiones de confianza, entonces:

$$(\overline{z}_{i1} - \zeta_{i1}, \overline{z}_{i2} - \zeta_{i2}) \begin{bmatrix} v_{i1} & c_i \\ c_i & v_{i2} \end{bmatrix}^{-1} \begin{bmatrix} \overline{z}_{i1} - \zeta_{i1} \\ \overline{z}_{i2} - \zeta_{i2} \end{bmatrix} \sim \chi^2_2$$

Dado que sigue una distribución χ^2_2 se encuentra la región al nivel $(1-\alpha) \times 100\,\%$ de confianza para el vector $(\zeta_{i1}, \zeta_{i2})'$ que es una elipse centrada en $(\overline{z}_{i1}, \overline{z}_{i2})'$, con límite en los puntos críticos $\chi^2_{2,(1-\alpha)}$ de una distribución χ^2_2 con 2 grados de libertad.

Sean y_1 y y_2, los valores de las coordenadas a lo largo de los dos primeros ejes respectivamente, el desarrollo de la ecuación de la elipse se muestra a continuación:

$$(y_2 - \overline{z}_{i2}, y_1 - \overline{z}_{i1}) \frac{\begin{bmatrix} v_{i1} & -c_i \\ -c_i & v_{i2} \end{bmatrix}}{v_{i1}v_{i2} - c_i^2} \begin{bmatrix} y_2 - \overline{z}_{i2} \\ y_1 - \overline{z}_{i1} \end{bmatrix} = \chi^2_{2,(1-\alpha)}$$

$$\Rightarrow (y_1 - \overline{z}_{i1})\left[v_{i1}(y_1 - \overline{z}_{i1}) - c_i(y_2 - \overline{z}_{i2})\right]$$
$$+ (y_2 - \overline{z}_{i2})\left[-c_i(y_1 - \overline{z}_{i1}) + v_{i2}(y_2 - \overline{z}_{i2})\right] = (v_{i1}v_{i2} - c_i^2)\chi^2_{2,(1-\alpha)}$$

$$\Rightarrow v_{i1}(y_1 - \overline{z}_{i1})^2 - 2c_i(y_1 - \overline{z}_{i1})(y_2 - \overline{z}_{i2}) + v_{i2}(y_1 - \overline{z}_{i1})^2 = (v_{i1}v_{i2} - c_i^2)\chi^2_{2,(1-\alpha)}$$

La elipse tiene parámetros estimados, por lo que tiene un parámetro que determina el grado de desviación de la elipse con respecto a una circunferencia, mejor conocido como excentricidad dada por

$$(v_{i1}v_{i2} - c_i^2)^{\frac{1}{2}}/v_{i2}$$

y el área está dada por

$$\pi(v_{i1}v_{i2} - c_i^2)^{\frac{1}{2}}\chi^2_{2,(1-\alpha)}.$$

La ecuación de la elipse no debe ser usada directamente con los valores de v_{i1}, v_{i2} y c_i ya que son desconocidos; todo este desarrollo es para sustituir a θ_i por $(1 + \boldsymbol{l_i})/(g-1)$, a_i por α_i y ζ_{ij} por $\overline{z}_{ij}$; y una vez sustituidos, se evalúan en la ecuación de la elipse y entonces la excentricidad y el área son calculadas.

Existen argumentos que sugieren la utilización de las elipses para la representación de las regiones de confianza; uno de estos, es que la variabilidad entre grupos está mejor representada, otro argumento es que al obtener los datos por un muestreo estadísticamente incorrecto, las elipses no muestran una gran diferencia en la dirección de los ejes de cada una. Además, al utilizar elipses en lugar de círculos, se sabe cuál de los ejes es el que aporta mayor variabilidad.

El desarrollo de las regiones de confianza en forma de elipses, es recomendado cuando el número de grupos es *pequeño*, ya que encierra más variabilidad dentro de cada grupo, y si se tiene un número grande de grupos, es muy probable que las elipses se traslapen entre sí.

Al graficar las regiones de confianza en forma de elipses, y notar que las elipses se traslapen, es necesario recurrir a un desarrollo alternativo de las regiones de confianza, conocido como *regiones de confianza alternativas*.

3.4. Regiones de confianza alternativas

Las regiones de confianza alternativas son casos particulares de las regiones de confianza (RC) ya desarrolladas, éstas son creadas a partir de los inconvenientes gráficos que llegan a ocurrir, dichas regiones alternativas se desarrollan debido a que en el **AC** pueden existir muchos grupos y/o el número de individuos dentro de cada grupo es grande. Al realizar el desarrollo de las RC y encontrar uno o varios "inconvenientes", es necesario tener claros los objetivos del estudio, para saber qué desarrollo alternativo se debe utilizar.

3.4.1. Individuos medios

El *individuo medio* está definido por la observación más cercana a la media dentro de cada grupo. Sin pérdida de generalidad, si existen dos observaciones que tienen la misma distancia respecto a la media dentro de cada grupo, se utiliza cualquiera de éstas dos.

Las regiones de confianza para los individuos medios se desarrollan de la misma manera que en las RC, con la diferencia de que dichas regiones están centradas en los individuos medios de sus respectivos grupos; por lo que también se busca que la dispersión entre los grupos sea máxima con relación a la dispersión dentro de cada grupo.

A diferencia del desarrollo de las RC, se debe cumplir el supuesto de normalidad, con la misma matriz de covarianza para todos los grupos, es decir, se tiene que cumplir que $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \cdots = \boldsymbol{\Sigma_g}$, entonces se tiene una hiperesfera centrada en cada individuo medio para los g grupos; cuando la hipótesis anterior es rechazada, no se puede determinar los ejes comunes de representación. En la práctica, la hipótesis de igualdad de covarianzas raramente se cumple, a pesar de esto, si los signos de los elementos de las matrices de covarianzas muestrales de cada grupo no cambian de un grupo a otro, la orientación de las hiperesferas no es significativamente distinta y aún es posible realizar el análisis.

Dado un nivel de confianza $(1-\alpha)$, las regiones alternativas están dadas por

$$\frac{R_{(1-\alpha)}}{\sqrt{n_i}} \quad \text{donde} \quad R^2_{(1-\alpha)} = F^{(1-\alpha)}\frac{(N-g)p}{(N-g-p+1)}$$

por lo tanto, los radios son:

$$R_{(1-\alpha)} = \sqrt{F^{(1-\alpha)}\frac{(N-g)p}{(N-g-p+1)n_i}} \quad \text{con} \quad i = 1, \ldots, g$$

donde $F^{(1-\alpha)}$ es el cuantil $(1-\alpha)\times 100\,\%$ de la distribución F de *Fisher* con $(p, N-g-p+1)$ grados de libertad, p es el número de variables observadas, g el de grupos, N el total de individuos y n_i los individuos en el grupo $i-ésimo$.

Si todos los grupos tienen el mismo número de individuos, los radios son los mismos para todas las regiones.

Capítulo 4

Errores y posibles soluciones en el Análisis Canónico

4.1. Principales errores y causas por las que ocurren

Al utilizar técnicas estadísticas para saber si existe o no algún tipo de relación entre dos conjuntos de variables, muchas veces se llegan a cometer errores en el análisis, desarrollo, derivación o interpretación de los resultados; en el presente capítulo se ejemplifican los errores más comunes, así como sus posibles soluciones.

Algunos errores son originados por:

i) Prácticas de predicción.
Predecir cierto rendimiento en dos baterías.

ii) Teoría acerca del contenido de la relación entre dos conjuntos.
Considerar independencia cuando existe dependencia entre las variables.

iii) Por las variables medidas en dos ocasiones.
Estudios longitudinales, Tiempo 1 contra Tiempo 2.

iv) Temas de relaciones personales.
Esposo y esposa, estudiante y profesor.

Los errores anteriores no son mutuamente excluyentes entre ellos, es decir, se puede dar el caso de cometer los cuatro en un mismo análisis y las tres principales causas por las que ocurren son las siguientes:

La *primera*, es cuando no se postulan adecuadamente las hipótesis para probar la relación que existe entre los conjuntos de variables y por ende existen diferentes pruebas estadísticas que pueden ser utilizadas para cada hipótesis diferente.

La *segunda* y la más frecuente en el **AC**, son los errores en los cálculos matriciales, esto ocurre por tener una teoría matemática inadecuada o mal entendida dada la hipótesis postulada y/o por una mala utilización en los paquetes computacionales.

La *tercera* razón incluso cuando el **AC** es adecuado dadas las hipótesis y el objetivo del problema, es que existen varias teorías y análisis diferentes que se puede llegar a pensar que son adecuados cuando en realidad no lo son.

4.2. Errores causados por la utilización de los paquetes estadísticos

Los grandes y complejos cálculos matriciales que involucra el **AC**, han llevado al desarrollado dentro de los paquetes estadísticos de bibliotecas (*libraries*) o herramientas específicamente para el **AC**, pero a pesar de esto, dicho análisis no es tan conocido como otras técnicas estadísticas, es por esto que en la práctica surgen problemas que conllevan varias complicaciones, y por ende un mal análisis y resultados erróneos; entonces, el primer problema, es que los programas y paquetes estadísticos tienen algoritmos para métodos más sencillos que no requieren de grandes y complejos cálculos o simplemente no cuentan con estos para realizar el **AC**.

El segundo error, es causado por algunos paquetes estadísticos ya que estos, no arrojan estadísticas o resultados auxiliares que son necesarios para la validación de los supuestos, debido a que se necesitan descargar bibliotecas específicas como en el caso de **R** y con ésto, poder concluir correctamente.

Capítulo 5

Recomendaciones para evitar errores en el Análisis Canónico

5.1. Selección del conjunto de variables y verificación de objetivos en el Análisis Canónico

El **AC** tiene un sustento matemático sólido en el sentido en que los conjuntos de variables a analizar, no tienen un gran número de supuestos que deban ser validados respecto a otras técnicas estadísticas, es decir, al utilizar el **AC** se puede tomar el menor número de restricciones sobre los tipos de variables que pueden ser utilizadas o que necesitan ser estudiadas, pero si se utiliza otra técnica que no es la adecuada por los supuestos que deben seguir las variables, la dificultad en la interpretación de los resultados o de los objetivos a lograr aumenta. Por lo tanto, es importante tener los objetivos del estudio claramente definidos antes de realizar el análisis, para asegurar que el **AC** sea aplicado correctamente, es por esto que el analista tiene que considerar dos puntos muy importantes; el primero es la *selección del conjunto de variables*, y el segundo es la *verificación de los objetivos en el análisis.*

5.1.1. Selección del conjunto de variables

El **AC** es adecuado para datos que pueden ser agrupados en conjuntos de variables y dado que el objetivo se encuentra claramente definido, se tiene que dar a cada conjunto de variables un valor teórico (peso), del por qué están siendo relacionadas entre sí. Los resultados que se obtienen del **AC**, son sensibles a los cambios de las variables de un conjunto a otro, es decir, los resultados se derivan tanto de las correlaciones entre las variables en cada conjunto como de la correlación entre variables canónicas entre conjuntos, es por esto que al realizar un cambio de alguna de las variables de un conjunto a otro, cambia notablemente la estructura de las variables canónicas.

5.1.2. Importancia de la matriz de correlación R en la selección de variables

Al utilizar el **AC** es útil pensar sólo en el conjunto de variables a analizar y no querer saber en primera instancia, si son independientes o dependientes, sin embargo, ya que se deduce que pueden o no compartir una relación, es factible hacer referencia a los conjuntos de variables múltiples que se correlacionan, es por esto que sólo se debe de tomar en cuenta la relación que existe entre los dos conjuntos, es decir, la matriz de correlación $\boldsymbol{R}$.

Para corregir errores que se ocasionan en la derivación del análisis de los dos conjuntos de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$, se deben realizar y analizar las siguientes preguntas respecto a $\boldsymbol{R}$:

i) Preguntarse acerca del número de relaciones independientes entre los dos conjuntos de variables.

ii) Acerca de los grados de redundancia entre los dos conjuntos.

iii) Sobre la similitud o semejanza de la matriz de correlación $\boldsymbol{R}$ y la matriz de covarianza $\boldsymbol{\Sigma}$.

Dado el objetivo y los conjuntos de variables a analizar, es importante considerar la matriz $\boldsymbol{R}$, pero en la mayoría de los casos, realmente lo que se analiza son las submatrices de ésta, es decir, $\boldsymbol{R_{XY}}$, $\boldsymbol{R_{XX}}$ y/o $\boldsymbol{R_{YY}}$, ya que la hipótesis indica cuál de éstas hay que analizar con mayor detenimiento. En la práctica el analista da cierto "valor" o "ponderación" a las variables originales y con esto, crea un nuevo conjunto entre las variables originales y las ponderaciones, pero estos conjuntos "forzosamente" relacionados pueden ser muy diferentes, es decir, se puede dar el caso en el cual varias variables tengan altas ponderaciones y baja correlación, mientras otras variables pueden tener baja ponderación y una alta correlación. En teoría es preferible utilizar altas correlaciones, aunque en la práctica se consideran las condiciones del problema.

5.1.3. Verificación de los objetivos en el análisis

El **AC** puede tener una gran cantidad de objetivos debido a las pocas restricciones en las variables; en la mayoría de los casos, son de investigación y se puede incluir uno o todos los siguientes:

i. Establecer si los dos conjuntos de variables son independientes o determinar la importancia de las relaciones que puedan existir entre los conjuntos.

ii. La obtención de una matriz de pesos o ponderaciones dadas por las condiciones del problema para cada conjunto de variables dependientes e independientes de manera que las combinaciones lineales de cada conjunto sean lo más altamente correlacionadas.

iii. Explicar el porqué de las relaciones existentes entre los conjuntos de variables dependientes e independientes y a su vez, medir la contribución de cada variable.

5.2. Diseño del Análisis Canónico

Para llevar a cabo el diseño del **AC** es necesario considerar un *tamaño de muestra* adecuado, el *error de medición* de las variables, su *aportación en el análisis*, los *datos faltantes* y los *valores extremos*.

5.2.1. Tamaño de muestra y error de medición

Antes de empezar con el tamaño de muestra, se tiene que dejar claro que el principal problema se encuentra en la medición de los datos, ya que se sabe que la medición es un proceso esencial; existen algunas variables que son relativamente sencillas de medir, por ejemplo la edad o el número de hijos, mientras que otras tienden a tener cierto grado de subjetividad que hace especialmente difícil su medición, como la intensidad de dolor o el concepto de calidad de vida; en cualquier caso, el proceso de medición conlleva siempre algún grado de error, es por esto, que existen factores asociados a los individuos, al observador o al instrumento de medición que pueden influir en la variabilidad de los datos, como es el caso de la temperatura corporal en la que pueden aparecer errores en el registro debido tanto al estado del paciente, como a defectos en el termómetro utilizado.

Las diversas cuestiones relacionadas con el impacto del tamaño de muestra y la necesidad de un número suficiente de observaciones por variable, se encuentra con frecuencia en los análisis multivariados. En la mayoría de los análisis estadísticos, se tiende a incluir muchas variables independientes por una variable dependiente, sin darse cuenta de las implicaciones para el tamaño de la muestra, es decir, se sabe que sí son pequeños no son representativos y tienden a arrojar correlaciones engañosas o relaciones significativas cuando no lo son y viceversa; en el otro caso, muestras muy grandes dan una tendencia a que las variables en un modelo, sean estadísticamente significativas cuando no lo son.

Entonces, el tamaño de muestra adecuado se relaciona con la "confiabilidad" de los resultados, el error de medida durante el intervalo de recopilación de los datos y la subjetividad o importancia que le dé el analista a esta recopilación; pero es recomendable al menos 25 observaciones por **cada** variable independiente para evitar "sobre ajuste" en el modelo.

5.2.2. Las variables y su aportación en el análisis

El **AC** es el menos restrictivo de todos los análisis multivariados, en el sentido en que las variables pueden ser métricas o no métricas y dependientes o independientes y aun así tener poca importancia para la estimación estadística dentro del **AC**, esto ocurre, porque el método se concentra en calcular los pesos de las variables originales maximizando de esta forma las correlaciones canónicas y eligiendo las variables canónicas; además no pone especial énfasis en alguna variable. Sin embargo, debido a que la técnica produce variables para las cuales está maximizada la correlación entre ellas, para alguna en cualquiera de los conjuntos se refiere a todas las demás en ambos conjuntos por lo que la eliminación o exclusión de una sola variable puede afectar a todos los resultados.

La solución de estos inconvenientes está en manos de los especialistas, ya que ellos son los que deben estar conceptualmente familiarizados con los conjuntos de variables antes de la aplicación del **AC**; esto hace que la relación entre las dependientes e independientes tenga una base conceptual sólida para todas las demás.

5.2.3. Datos faltantes y valores extremos (outliers)

Al igual que en otras técnicas multivariadas, existen diferentes procedimientos para la imputación de datos faltantes, los más conocidos son: la estimación de

los valores por medio de una media aritmética, por una interpolación en el caso longitudinal, por una regresión polinomial entre otros; el procedimiento que se utilice para la imputación de los datos, depende del contexto de los datos faltantes, ya que si estos surgen aleatoriamente, pueden ser imputados con los métodos más comunes; pero si no provienen de eventos aleatorios, sino que faltan por errores de instrumentos de medición, por ejemplo, una máquina que mide la temperatura de cierto lugar que deja de funcionar por un largo periodo de tiempo, la imputación de los datos requiere de un análisis específico para estos.

Los valores extremos o outliers, se pueden detectar mediante métodos de análisis univariados y multivariados, y en todos estos, lo más común es la eliminación de las observaciones que los contengan, pero si se eliminan un gran número de observaciones, se pueden producir cambios sustanciales en los resultados del **AC**.

5.3. Supuestos

Algunos supuestos no son estrictamente necesarios, pero la interpretación de los resultados en el **AC**, mejora si se consideran los siguientes supuestos:

5.3.1. Linealidad

La linealidad es importante para el **AC** y afecta a dos aspectos de los resultados de dicho análisis; en primer lugar, las correlaciones canónicas entre el par de variables está basado en una relación lineal, si las variables se relacionan de una manera no lineal, la relación no será coherente con el coeficiente de correlación canónica; en segundo lugar, el **AC** maximiza la correlación lineal entre las variables y aunque el **AC** es el método multivariado general, sigue estando obligado a la

identificación de relaciones lineales, es por esto que, si la relación es no lineal entonces una o ambas variables tendrán que ser transformadas si es posible.

5.3.2. Normalidad Multivariada

El supuesto de normalidad en las variables, es el más deseable, ya que implica que se tienen todos los supuestos necesarios para derivar las correlaciones canónicas *muestrales*; esta hipótesis puede obviarse cuando se dispone de un número suficiente de datos por cada variable, aunque siempre se tiene que probar dicho supuesto.

5.4. Significado de las variables canónicas y criterios para la elección

Una vez que los conjuntos de variables se han identificado argumentados por los objetivos del análisis, el siguiente paso del **AC** es obtener las variables canónicas (Capítulo 1). Cada variable canónica consta de un par de variables, una que representa a las independientes y la otra que representa a las dependientes; el número máximo de variables canónicas que pueden ser obtenidas de los dos conjuntos de variables es igual al número de variables en el conjunto más pequeño, ya sean independientes o dependientes. Por ejemplo, cuando cierto problema de investigación consta de cinco variables independientes y tres variables dependientes, el número máximo de variables canónicas que se pueden obtener es tres.

5.4.1. Significado de las variables canónicas y su rotación

El **AC** está estrechamente relacionado con el análisis de componentes principales, ya que la rotación de las variables canónicas se considera como una interpretación entre los conjuntos de variables, es decir, ya que la rotación no cambia la suma de los cuadrados de coeficientes de correlación canónica, simplemente da lugar a un modelo estadístico más sencillo; cabe señalar, que cada par de variables es ortogonal (independiente) de todas las variables que se deriven. La variabilidad de la primera variable canónica que se deriva es la máxima de todo el conjunto de variables, la segunda variable canónica se construye maximizando la variabilidad restante, y así sucesivamente, hasta que todas las variables canónicas han sido derivadas.

Algunos autores no recomiendan la rotación para la interpretación del **AC**, ya que puede reducir la optimización de las correlaciones canónicas cuando cada par de variables canónicas es derivado, es decir, al aplicar una rotación, se tienen resultados diferentes en la matriz $\boldsymbol{R}$, por lo que se puede llegar a concluir que existe relación entre pares de variables cuando en realidad no lo hay, y viceversa; por lo tanto, a pesar de que la rotación puede mejorar la interpretación de los resultados, se recomiendan ser cautelosos al utilizar la rotación en el **AC**.

5.4.2. Criterios importantes para elegir las variables canónicas a interpretar

Al igual que en la mayoría de las técnicas estadísticas, la práctica más común es analizar que variables canónicas son estadísticamente significativas considerando un *p-value* menor de 0.05.

Se cree que un único criterio como el nivel de significancia, es demasiado superficial o carece de "fuerza" para poder tener suficiente evidencia para tomar

una decisión, es por esto que, se recomiendan tres criterios, que son: *el nivel de significancia estadístico*, una *alta correlación entre las variables canónicas*, y *el nivel de redundancia para el porcentaje de variabilidad* explicada entre los dos conjuntos de variables.

Otro criterio importante para saber si las variables y correlaciones canónicas son adecuadas y tienen importancia para el análisis, son las diferentes pruebas de significancia que se trataron en el capítulo 3, que son:

i. *La Prueba de Wilks usando la aproximación de Bartlett.*

ii. *La estadística de Rao en relación a la lambda de Wilks con base en una distribución F.*

iii. *Análisis de variabilidad bajo la hipótesis alternativa*

5.5. Interpretación de las variables canónicas

El paso 5 es el más importante de las recomendaciones y antes de continuar con el último paso, se debe tener en cuenta que las correlaciones canónicas sean estadísticamente significativas; es por esto, que se deben realizar las interpretaciones correctamente y si es posible analizar de nuevo todos los resultados para tener fundamentos importantes y necesarios para la interpretación de las variables canónicas.

Las interpretaciones en el **AC** son de gran importancia, porque cada una de las variables originales tiene una relativa relación con las variables canónicas, es por esto, que al realizar la interpretación de las variables canónicas hay que tener en cuenta, *la importancia de las variables canónicas, la estructura de las correlaciones, las cargas canónicas cruzadas y los coeficientes canónicos.*

5.5.1. La importancia de las variables canónicas

La interpretación de las variables canónicas implica el análisis de la magnitud o peso asignados a cada variable original por el especialista, las variables con mayor peso relativamente tienen una contribución mayor a las variables canónicas, por ejemplo, un peso pequeño puede significar que su correspondiente variable es irrelevante para determinar una relación o que ha sido excluida de la relación debido a un alto grado de colinealidad; un problema con el uso de los pesos, es que estos están sujetos a una considerable inestabilidad y todo esto es debido al tamaño de una muestra a otra, esta inestabilidad se produce porque el procedimiento del desarrollo en el **AC** maximiza las correlaciones canónicas para una muestra en particular de la variable observada ya sea dependiente o independiente, es por esto que se recomienda cautela en la utilización de pesos para interpretar los resultados.

5.5.2. La estructura de las correlaciones

Las correlaciones canónicas han sido cada vez más utilizadas como base para la interpretación debido a las deficiencias inherentes o problemas que traen consigo los pesos canónicos, las correlaciones canónicas sirven para medir la correlación lineal o la relación entre una variable original en el conjunto de variables dependientes o independientes con el conjunto de variables canónicas, es decir, refleja la variación que existe entre las variables observadas con las variables canónicas y se puede interpretar como un factor de peso en la evaluación de la contribución relativa de cada variable para cada combinación canónica; dicha interpretación es interesante, en el sentido en que los criterios que se utilizan para determinar la interpretación de las correlaciones canónicas son los mismos que con el análisis de factores. Las correlaciones canónicas como pesos a considerar, también pueden estar sujetos a una considerable variabilidad de una muestra

a otra; este tipo de variabilidad sugiere que las correlaciones y por lo tanto las relaciones que se les atribuye, pueden deberse no tanto al tamaño de muestra, sino a factores externos. Sin embargo, las correlaciones canónicas se consideran más aceptables que el peso como un medio de interpretación por la natural interacción entre las variables.

5.5.3. Cargas canónicas cruzadas

Otra forma de interpretación de las variables canónicas es mediante el cálculo de las *cargas canónicas cruzadas*, que es simplemente una forma alternativa sugerida para dar un enfoque diferente a las correlaciones canónicas.

Este procedimiento consiste directamente en analizar si existe relación entre las correlaciones de las variables originales y las de las canónicas. Las cargas convencionales son las correlaciones de las variables originales con sus respectivas variables canónicas después de haber máximado las correlaciones entre los dos conjuntos de variables canónicas (dependientes e independientes) entre sí. Como se señaló anteriormente, también esto puede parecer similar a la regresión múltiple pero se diferencia en que cada variable independiente, se correlaciona con todo el conjunto dependiente en lugar de una sola variable, así, las cargas canónicas cruzadas proporcionan una medida directa de interpretación de las relaciones que existen entre las variables dependientes e independientes mediante la eliminación de un paso intermedio que se realiza en las cargas convencionales. En algunos análisis, no se calculan las cargas canónicas cruzadas entre las variables originales y las canónicas, en tales casos, los pesos canónicos son considerados semejantes.

5.5.4. Coeficientes canónicos

Lo más común es que al identificar las correlaciones canónicas estadísticamente significativas, se les llame *coeficientes canónicos*; pero existen varios coeficientes adicionales que pueden ser calculados para aumentar la variabilidad explicada, pero el uso de estos coeficientes adicionales es un tanto problemática, sin embargo para algunos investigadores es necesario tomarlos en cuenta por el origen del problema.

Cabe mencionar que han surgido un sin fin de problemas en la aceptación de los términos o definiciones que deben emplearse para hacer referencia a conceptos diferentes, por ejemplo, para algunos autores los pesos canónicos son los coeficientes canónicos, o la dimensión canónica son las variables canónicas, o la correlación canónica es la relación entre los conjuntos de variables; entonces, para utilizar un mismo concepto se interpretan los coeficientes de la siguiente manera:

- **Coeficientes de estructura**

Los coeficientes de estructura canónicos son particularmente importantes. De esta forma, Meredith (1964) sugiere que, si las variables dentro de cada conjunto están intercorrelacionadas moderadamente, entonces la posibilidad de interpretar a las variables canónicas por inspección de los pesos (coeficientes) de regresión apropiados, es prácticamente nula.

Kerlinger y Pedhazur (1973) argumentan: Un **AC** también produce pesos, los cuales, teóricamente al menos, son interpretados como pesos de regresión. Estos pesos parecen ser la relación débil en la cadena del análisis de correlación canónica.

Levine (1977) sugiere que se deben interpretar los coeficientes de estructura si se desea obtener información acerca de la naturaleza de la relación que existe

entre las correlaciones canónicas y el conjunto de variables originales y no quedarse únicamente con el cálculo de los puntajes.

Es decir, el cuadrado de los coeficientes de estructura canónicos representa la proporción de variabilidad linealmente compartida por una variable dentro del conjunto de las variables canónicas. Este hecho no es intuitivamente obvio por su forma matricial.

$$\boldsymbol{RW} = \boldsymbol{D} \tag{5.1}$$

donde $\boldsymbol{W}$ es la matriz de coeficientes de la función de las dos variables criterio y $\boldsymbol{D}$ es la matriz de coeficientes de estructura.

El caso más sencillo de (5.1) para el criterio del coeficiente de estructura es cuando la función canónica y los coeficientes de estructura son los mismos, es decir, que las variables de los conjuntos están no correlacionadas unas con otras, esto es, cuando $\boldsymbol{R_{XX}}$ o $\boldsymbol{R_{YY}}$ constituyen una matriz identidad, entonces:

$$\boldsymbol{W} = \boldsymbol{D}$$

- **Coeficientes de comunalidad y adecuación**

Como los coeficientes de estructura canónica, parten de otros dos coeficientes canónicos, como en el análisis de componentes principales o factores, las variables son no correlacionadas unas con otras y la suma del cuadrado de todos los coeficientes de estructura proporciona el valor del coeficiente de comunalidad según Thompson(1980) y Thorndike (1977). Este valor indica qué proporción de la

variabilidad de cada variable es reproducible a partir de los resultados canónicos, esto es, que tan útil es cada variable en la representación canónica. El promedio de todos los cuadrados de los coeficientes de estructura para las variables en un conjunto con respecto a una función, representa el coeficiente de adecuación canónica. Es decir, estos coeficientes indican qué tan adecuada, en promedio, es la representación de la variabilidad de un conjunto de variables canónicas con respecto al total de la variabilidad del conjunto original.

- **Coeficientes de redundancia y análisis**

Stewart y Love (1968) definen el llamado índice de redundancia que es la media de los coeficientes al cuadrado de la correlación múltiple para predecir las variables de un conjunto por el otro conjunto y hace referencia a la proporción promedio de variabilidad en las variables de un conjunto que está representado por las variables en otro conjunto. Una correlación canónica relativamente fuerte puede obtenerse entre dos funciones lineales, aun cuando estas funciones lineales no expliquen una proporción de variabilidad significativa para sus respectivas variables, por como está construido el índice de redundancia a diferencia de las correlaciones canónicas, es una medida en general no simétrica, en el sentido en que dadas las combinaciones lineales, los coeficientes de redundancia no son los mismos coeficientes.

Wollenberg (1977) argumenta que el análisis de redundancia obtiene combinaciones lineales no correlacionadas de las variables predictoras con varianza uno, de tal forma que la media de las correlaciones al cuadrado con las variables del otro grupo sea máxima. Es decir, se maximiza el índice de redundancia entre la combinación lineal de un conjunto de variables y el otro conjunto.

El coeficiente de redundancia, en el sentido estricto, no es multivariado ya que no se ve afectado por la correlación dentro de las variables dependientes; sin

embargo, es multivariado en el sentido de que involucra las variables dependientes para su cálculo.

Si las funciones canónicas son calculadas para optimizar las correlaciones canónicas, entonces surge la siguiente pregunta: *¿por qué se querría consultar un coeficiente que no es multivariado y que no es utilizado como parte del análisis?*. Wollenberg (1977) notó que el análisis de correlaciones canónicas sólo maximiza una parte del coeficiente de redundancia en sí, es decir, las correlaciones canónicas maximizan la relación entre cada una de las combinaciones lineales que pueden no estar relacionadas con los conjuntos de variables originales, de manera inversa, la maximización del coeficiente de redundancia puede estar relacionado con los conjuntos de variables originales. Entonces, se utiliza la técnica que permite ver cuánta relación debe estar dada para optimizar la redundancia o la correlación canónica, esta técnica tiene por nombre "factorización redundancia-canónica" presentada por DeSarbo (1981).

Los coeficientes de redundancia en el **AC** tienen un valor limitado; Thompson sugiere que: un coeficiente de redundancia puede tomar el valor de 1 sólo cuando las dos variables comparten el 100 % de la variabilidad, y una variable canónica representa perfectamente a las variables originales en su dominio. El coeficiente de redundancia para explicar el conjunto de la variable $\boldsymbol{y}$ con la variable $\boldsymbol{x}$ como el valor promedio del cuadrado de las correlaciones entre $\boldsymbol{x}$ y $\boldsymbol{y}$, se denota como:

$$\boldsymbol{R_d}(\boldsymbol{y} \mid \boldsymbol{x'}\alpha) = \frac{1}{q}\alpha' \boldsymbol{R_{yx}} \boldsymbol{R_{xy}} \alpha$$

donde α es el vector de los coeficientes para la combinación lineal de $\boldsymbol{x}$.

Entonces la redundancia total para explicar el conjunto de la variable $\boldsymbol{y}$ con el conjunto de la variable $\boldsymbol{x}$ a través de las combinaciones lineales es:

$$\boldsymbol{R(y \mid x)} = \sum_{i=1}^{k} \boldsymbol{R_d(y \mid x'\alpha)},$$

por la construcción del coeficiente, se puede demostrar que:

$$\boldsymbol{R(y \mid x)} = \frac{t_r(\boldsymbol{R_{yx} R_{xx}^{-1} R_{xy}})}{t_r(\boldsymbol{R_{yy}})},$$

y se denota como:

$$\boldsymbol{R(y \mid x)} = \sum_{j=1}^{k} \frac{1}{q} \boldsymbol{R_j^2} = \boldsymbol{\overline{R^2}}$$

donde $\boldsymbol{R_j^2}$ es el coeficiente de correlación múltiple al cuadrado en una regresión entre la variable $\boldsymbol{x}$ y $\boldsymbol{y}$. Por lo tanto, el coeficiente de redundancia puede ser interpretado como índice, el cual evalúa qué tan bien están representadas las relaciones dentro y entre los conjuntos de variables por las k-variables canónicas.

El análisis de redundancia en el **AC**, tiene más sentido cuando el objetivo principal es la derivación de las funciones que "capturan" la mayor variabilidad de las variables originales. Stewart y Love (1968) publican la presentación del índice, considerando la utilidad de dicho índice $\boldsymbol{R_j^2}$ como un índice de resumen y no como una herramienta analítica.

- **Coeficientes de índice**

Los coeficientes definidos "coeficientes de índice" por Thompson y Frankiewicz (1979), son también considerados como coeficientes de correlación. El cálculo de la matriz para el caso en que el primer conjunto tenga n variables y el segundo m es:

$$\boldsymbol{R_{11(n\times n)}W_{1(n\times m)}} = \boldsymbol{S_{1(n\times m)}}$$

$$\boldsymbol{R_{12(n\times m)}W_{2(m\times m)}} = \boldsymbol{I_{1(n\times m)}}$$

$$\boldsymbol{R_{22(m\times m)}W_{2(m\times m)}} = \boldsymbol{S_{2(m\times m)}}$$

$$\boldsymbol{R_{21(m\times n)}W_{1(n\times m)}} = \boldsymbol{I_{2(m\times m)}}$$

Donde $\boldsymbol{R}$ representa los distintos cuadrantes de la matriz de correlación, $\boldsymbol{W}$ representa las "ponderaciones" o las matrices de los coeficientes de funciones (que es el $min(n, m)$), $\boldsymbol{S}$ e $\boldsymbol{I}$, representan las matrices de coeficientes de estructura canónica y coeficientes de índice respectivamente. Los coeficientes de índice representan la correlación entre los resultados originales, las variables ponderadas y sin ponderar.

Estos índices, según Kuylen y Verhallen (1981), son más conservativos, menos exagerados que los coeficientes de estructura y forman una base más sólida para una interpretación adecuada.

5.6. Limitaciones y validación de los resultados

Los resultados del **AC** deben ser validados con diferentes procedimientos para asegurar que los resultados sean específicos y correctos, es decir, que la validación de los resultados de la muestra se generalice a la población.

Un procedimiento, es la obtención de dos muestras (si el tamaño de la muestra, los costos y el tipo de estudio lo permite) y realizar el análisis en cada muestra por separado; con esto, se puede comparar la similitud de todos los resultados del **AC**; si se encuentran diferencias marcadas en la mayoría de los resultados, se debe considerar una investigación adicional para garantizar que los resultados finales sean representativos.

Aunque son pocos los procedimientos de validación para los resultados que se han desarrollado específicamente para el **AC**, se debe tomar en cuenta las limitaciones de los resultados que puede tener dicha técnica.

Entre las limitaciones que pueden tener mayor impacto en los resultados y su interpretación son las siguientes:

1. Las correlaciones canónicas representan la varianza explicada por las combinaciones lineales de los conjuntos de variables y no la varianza explicada por cada una de las variables.

2. La interpretación de las variables canónicas puede ser difícil porque es calculada para maximizar la relación, y las "ayudas" para la interpretación como la rotación de variables en el análisis de factores o en componentes principales son limitadas.

3. En algunas ocasiones, es difícil identificar las relaciones estadísticamente significativas entre los conjuntos de variables dependientes e independientes ya que los cálculos y estadísticas precisas aún no han sido desarrolladas en su totalidad para interpretar un **AC**, y por ende se llega a recurrir a pruebas estadísticas inadecuadas.

Estas limitaciones no tienen por objeto desalentar la utilización del **AC**, más bien, se hacen notar para aumentar la eficacia del análisis y además en la mayoría de los casos no repercute en los resultados, aunque es mejor contemplarlos.

Capítulo 6

Técnicas Multivariadas relacionadas con el Análisis Canónico

6.1. Relación del AC con otras técnicas multivariadas

El **AC** es una técnica que permite estudiar la estructura de varios grupos de individuos con respecto a un conjunto de variables observadas. La condición que se impone es que la separación entre los distintos grupos sea máxima con respecto a la variabilidad dentro de los grupos. A continuación se describen varias técnicas estadísticas para estudiar las diferentes preguntas y objetivos que se llegan a plantear dadas las relaciones entre dos conjuntos de variables y especificar las diferencias en que cada técnica estadística es más apropiada; y a su vez, argumentar cómo identificar aquellos problemas en los cuales el **AC** es adecuado para el problema en cuestión, pero que es generalmente ignorado.

El análisis de varianza univariado y multivariado, la regresión múltiple, análisis discriminante, análisis de factores, análisis de componentes principales y cualquier técnica univariada o multivariada que tenga efecto sobre el análisis de varianza, pueden considerarse como un caso particular del **AC**, es decir, el análisis canónico es la técnica general abarcando todas las otras como casos particulares. Dado el tipo de variables entre los dos conjuntos $\boldsymbol{X}$ y $\boldsymbol{Y}$, se muestra la relación en la siguiente tabla:

Características de los métodos	Sólo una variable dependiente	Más de una variable dependiente
Sólo variables independientes nominales	Análisis de varianza univariado	Análisis de varianza multivariado
Variables independientes nominales o de intervalo	Regresión Múltiple	Análisis de variables canónicas

6.2. Análisis de varianza multivariado

El análisis de varianza multivariado o MANOVA es un método utilizado para contrastar la hipótesis de igualdad entre los vectores de medias de los distintos grupos de individuos. Si los datos corresponden a distribuciones con k dimensiones, éstos son normales con matriz de covarianzas dentro de los grupos en común. En forma general se puede plantear como contrastes entre filas o combinaciones lineales de las columnas de la matriz de medias.

En un MANOVA las medias de la muestra además de ser centradas están ponderadas por los tamaños muestrales de cada grupo y por la inversa de la

raíz cuadrada de la matriz de covarianzas dentro de los grupos; también busca la combinación lineal que sea máxima, dicha combinación lineal es la misma que se obtiene en el análisis canónico.

El MANOVA se obtiene calculando la descomposición en valores singulares de la matriz $\boldsymbol{Y}$, por tanto la relación de esta descomposición y el MANOVA está en que las pruebas para contrastar la igualdad de los vectores de medias de los grupos son funciones de los valores singulares r_i, con $i = 1, 2, \ldots, p$ donde $p = rango(\boldsymbol{Y})$ que cumplen lo siguiente:

i. r_1^2 es la mayor raíz característica.

ii. $t^2 = \sum_{i=1}^{p} r_i^2$ es la traza.

Estas pruebas son válidas para muestras de poblaciones normales o para muestras suficientemente grandes, además de que se tiene que cumplir la condición de homogeneidad de matrices de covarianzas dentro de los grupos.

Por tanto el MANOVA permite realizar pruebas de igualdad de los vectores de medias y el análisis canónico construye los ejes de mayor separación de los grupos representados por esos vectores de medias, pero ambos métodos utilizan la misma matriz de datos transformada para realizar el análisis. En el primero se tienen suposiciones respecto a la distribución normal multivariada y a la homogeneidad de las matrices de covarianzas de los datos y en el segundo, desde el punto de vista descriptivo, se puede realizar sin tener en cuenta dichas suposiciones.

6.3. Análisis de regresión múltiple

El análisis de regresión múltiple, predice una sola variable dependiente con una función lineal de un conjunto de variables independientes, en el **AC** se forma una variable aleatoria constituida de varias variables dependientes; pero para algunos problemas de investigación el interés no puede centrarse en una sola variable independiente, en su lugar, se puede estar interesado en las diferentes relaciones entre conjuntos de múltiples variables tanto independientes como dependientes, es por esto, que el análisis canónico es la solución para este tipo de problema de investigación por ser un método que permite la evaluación de la relación entre dos o más conjuntos de variables.

6.4. Análisis discriminante

El análisis discriminante es un caso particular del **AC** debido a su característica de determinar dimensiones independientes para cada conjunto de variables, es decir, el análisis discriminante es un método para estimar la relación entre una sola variable dependiente no métrica y un conjunto de variables independientes métricas y se sabe también, que el análisis discriminante puede incluir el uso de "falsas" variables sobre el lado dependiente y es por esto, que el **AC** puede también ser utilizado para desempeñar el análisis discriminante o un análisis de tablas de contingencia.

También se tiene el problema de considerar que un individuo proviene de alguna de g poblaciones distintas, pero se desconoce cuál es dicha población, entonces se desea asignar el individuo a la población correcta, teniendo como base una serie de medidas, las cuales se expresan como un vector de variables y el interés es obtener una regla de asignación de los individuos, es decir, una asignación que minimice cualquier error de mala clasificación para todos los individuos.

Si se compara con el **AC**, se construyen los vectores de tal forma que maximice la variabilidad entre los grupos corregida por la variabilidad dentro de éstos; en cambio en el análisis discriminante lo que se encuentran son reglas para asignar individuos a una población dada a priori y calcula las coordenadas discriminantes que son las obtenidas en el **AC**.

6.5. Análisis de componentes principales y análisis de factores

El **AC** está estrechamente relacionado con el análisis de componentes principales y éste a su vez, se encuentra relacionado con el análisis de factores, cuyo principal objetivo es definir una estructura subyacente entre las variables, es decir, el análisis de componentes principales y el análisis de factores son un caso particular del **AC**, esto es debido a la creación de la estructura óptima de cada conjunto de variables que maximiza la relación entre los conjuntos de variables independientes y dependientes, es decir, dichos análisis explican la relación lineal entre un conjunto de variables observadas y un número indeterminado de factores o variables, según sea el caso.

La relación lineal entre los conjuntos de variables, está ligada al análisis de correlación canónica y el análisis de factores, destacando que la rotación de las componentes principales puede facilitar la interpretación de los resultados canónicos. Es de aquí donde los factores son "extraídos" de la matriz de correlación sujeto a que cada factor, como se sabe, debe ser no correlacionado con el resto de los factores; además cada factor se sujeta a la restricción de que cada uno de ellos debe reproducir la matriz de correlación. Así los factores tienden a disminuir el valor de variabilidad que representan.

El hecho de que cada nuevo factor deba reproducir al máximo a la estructura de dependencia es una característica importante para este procedimiento pues tiene la propiedad de obtener el número mínimo de factores posible.

Por otro lado la interpretación de la primera componente principal derivada a partir de la matriz de correlación, tiende a estar correlacionada con sus respectivos factores, sin embargo los factores se pueden rotar para tener una mejor interpretación; este tipo de rotación se puede hacer si se muestra que un conjunto dado de factores rotados y un conjunto correspondiente de factores no rotados deben reproducir exactamente a la matriz de donde fueron extraídos.

El análisis de factores permite generar varias soluciones para un mismo conjunto de datos, cada solución es generada por un esquema de rotación de factores, es decir, cada rotación tiene una interpretación diferente. Cuando no se realiza ninguna rotación, el análisis que se emplea es el de componentes principales, ya que en el análisis de factores subyace un nuevo modelo.

6.6. Biplot canónico

Es una técnica estadística multivariada utilizada en situaciones experimentales donde se tiene un conjunto de variables de respuesta y se quiere buscar las diferencias entre varios grupos; es definido como Biplot Canónico por Vicente-Villardón (1992) y Gower Y. Hand (1996). El Biplot Canónico considera la variabilidad de los ejes y utiliza los desarrollos matemáticos de Krzanowski (1989) en el contexto del Análisis de Variables Canónicas.

A diferencia del **AC**, en el Biplot Canónico, además de interpretar las diferencias y semejanzas entre los grupos; se interpretan las relaciones entre las variables; y las relaciones entre grupos y variables. Vicente y Gabriel proponen el

Biplot Canónico como herramienta de presentación de los resultados de técnicas más formales como el análisis de varianza, análisis discriminante, el **AC**, entre otras.

Lo que se pretende con este método, es obtener una representación simultánea de las filas y las columnas, de las cuales éstas últimas son las variables de la matriz $\boldsymbol{X}$, y representa las medias muestrales de cada uno de los grupos para cada una de las variables. Para considerar el efecto de la dispersión de los individuos y de las escalas de medida de las variables, se introduce una ponderación con respecto a la matriz de covarianzas dentro de los grupos y otra con relación a los tamaños muestrales. Entonces el Biplot se obtiene de construir una matriz que está en función de los valores singulares de $\boldsymbol{Y}$, dichas ponderaciones son las medias de los g grupos y de las N variables.

Las ponderaciones tienen varias propiedades entre las que destacan:

- Para las ponderaciones de medias:
  - La matriz puede interpretarse como la proyección de $\overline{\boldsymbol{X}}$ sobre el espacio de máxima separación de los grupos.
  - Se pueden construir círculos o esferas de confianza alrededor de las medias, construyendo un intervalo de confianza univariado sobre la proyección de cada una de las medias en cada una de las variables.

- Para las ponderaciones de los grupos
  - La matriz de dichas ponderaciones representa la estructura de covarianza de las variables y los ejes canónicos, ponderados por las inversas de los valores propios de $\overline{Y}'\,\overline{Y}$.
  - La longitud de las ponderaciones dentro de la columna, es proporcional a la variabilidad dentro de los grupos.

6.7. Análisis canónico como caso general

Debido a que las otras técnicas multivariadas requieren de más restricciones, en general se cree que las técnicas antes mencionadas, conllevan limitaciones que producen resultados robustos estadísticamente que pueden presentarse de una manera menos interpretable; sin embargo, en situaciones con conjuntos de múltiples variables dependientes e independientes, el **AC**, es la técnica multivariada más apropiada y de gran eficacia, debido a que cuenta con el menor número de restricciones estadísticas sobre los tipos de datos que pueden utilizar, y el análisis que se realiza, se elabora con todas las posibles relaciones entre ambos conjuntos de variables.

Por lo tanto, el **AC** se convierte en una alternativa lógica y precisa para muchos de los problemas más complejos en donde estén involucrados conjuntos de varias variables.

6.8. Ventajas del Análisis Canónico

El **AC** tiene varias ventajas, una de éstas se refiere al planteamiento de las hipótesis, los límites de la probabilidad de cometer errores de tipo I son menores, es decir, se sabe que el riesgo de cometer un error tipo I se relaciona con la probabilidad de encontrar un resultado estadísticamente significativo cuando no lo es, lo cual implica un aumento del riesgo de error tipo I; dicha incongruencia en el error, se ve reflejada a partir de los resultados y esto se debe a que las mismas variables en un conjunto de datos se utilizan para muchas pruebas estadísticas; por ejemplo en regresión múltiple, si se pretende analizar si seis variables del conjunto $\boldsymbol{X}$ pueden predecir ocho variables del conjunto $\boldsymbol{Y}$, implica que son necesarias ocho ecuaciones de regresión, es decir, una para cada variable dependiente, pero esto no es el verdadero problema, el verdadero problema es consecuencia del uso de

pruebas de significancia estadística por separado para cada ecuación, por lo que aumenta la probabilidad de cometer el error tipo I.

Es por esto que una de las ventajas del **AC**, es que puede considerar todas las relaciones existentes entre ambos conjuntos de variables, aunque dichas variables sean dependientes e independientes y lo mejor sucede porque dicha consideración, se efectúa en una sola relación y no a través de relaciones por separado para cada variable dependiente. Por ejemplo, en estudios psicológicos se observan resultados que reflejan mejor a la realidad, esto es debido a la complejidad en investigaciones en las cuales intervienen seres humanos y comportamiento organizacional, ya que pueden surgir múltiples variables que representan un concepto y por lo tanto crear problemas cuando las variables son examinadas por separado, es decir, en el **AC** se representaría una relación entre los conjuntos de variables, en lugar de entre varias variables individuales. Por construcción, el **AC** permite analizar los conjuntos de variables e identificar dos o más relaciones entre los conjuntos y ser teóricamente consistente.

Otra de las ventajas al utilizar el **AC** , es que se puede desarrollar el análisis con una muestra de la población original, es decir, lo más interesante es mostrar que el número de relaciones canónicas en la población es igual al número de relaciones canónicas muestrales (como se vio en capítulos anteriores), por la prueba de Bartlett, que se aplica sobre la muestra, nos demuestra que las correlaciones muestrales deben ser controladas en orden para así elegir o admitir las correlaciones que sean significativas y con esto se eliminan todas las correlaciones que no tengan significancia entre los conjuntos $\boldsymbol{X}$ y $\boldsymbol{Y}$ en la población, en otras palabras, ya que al aplicar la prueba de Bartlett, las correlaciones muestrales que sean exactamente cero, entonces serán también cero dentro de la población.

Capítulo 7

Análisis Canónico Generalizado

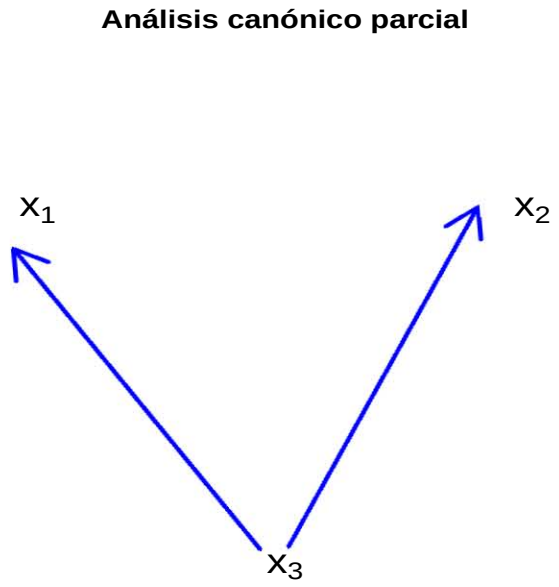
7.1. Análisis canónico parcial

Dada la construcción de las variables canónicas para dos conjuntos de variables, el siguiente paso para una generalización del **AC** consiste en analizar e interpretar no solamente dos conjuntos de variables, ahora se tienen tres conjuntos de variables aleatorias : $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ y $\boldsymbol{X_3}$ de manera que la dimensión de cada variable sea: n_1, n_2 y n_3 respectivamente, con $n_1 \leqslant n_2$.

Además, se tienen que distribuir conjuntamente con parámetros $(0, \boldsymbol{\Sigma})$; donde $[\boldsymbol{\Sigma ij}], coni,j= 1,2,3; que cumplan con los supuestos requeridos para el caso de dos conjuntos de variables.$

Supóngase que cada conjunto $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$, está relacionado linealmente con $\boldsymbol{X_3}$. Entonces se realiza una regresión sobre $\boldsymbol{X_3}$ de cada una de las otras variables, para así obtener ϵ_{13} y ϵ_{23}, donde ϵ_{12} son los residuos correspondientes a la regresión de $\boldsymbol{X_1}$ sobre $\boldsymbol{X_3}$ representada por $\boldsymbol{X_1^*}$ y ϵ_{23} son los residuos correspondientes a la regresión de $\boldsymbol{X_2}$ sobre $\boldsymbol{X_3}$ representada por $\boldsymbol{X_2^*}$.

Figura 6: ***Representación gráfica de la relación lineal entre las variables.***



En términos matriciales, la representación de los residuos es:

$$\epsilon_{13} = \boldsymbol{X_1} - \boldsymbol{X_1^*} = \boldsymbol{X_1} - \boldsymbol{\Sigma_{13}\Sigma_{33}^{-1}X_3}$$

$$\epsilon_{23} = \boldsymbol{X_2} - \boldsymbol{X_2^*} = \boldsymbol{X_2} - \boldsymbol{\Sigma_{23}\Sigma_{33}^{-1}X_3}$$

Donde la matriz de covarianzas conjunta es:

$$\boldsymbol{\Sigma_{\epsilon_{13},\epsilon_{23}}} = \begin{bmatrix} \boldsymbol{\Sigma_{11} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{31}} & \boldsymbol{\Sigma_{12} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{32}} \\ \boldsymbol{\Sigma_{21} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{31}} & \boldsymbol{\Sigma_{22} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{32}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma_{11\cdot 3}} & \boldsymbol{\Sigma_{12\cdot 3}} \\ \boldsymbol{\Sigma_{21\cdot 3}} & \boldsymbol{\Sigma_{22\cdot 3}} \end{bmatrix}$$

Ahora se definen correlaciones y variables canónicas parciales entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ como las correlaciones y variables canónicas entre ϵ_{13} y ϵ_{23}.

De acuerdo a la definición anterior, se deriva que las variables canónicas parciales son obtenidas como en el caso más simple de las variables canónicas a través de las transformaciones lineales de ϵ_{13} y ϵ_{23}, es decir:

$$\boldsymbol{U} = \boldsymbol{L'\epsilon_{13}}$$

$$\boldsymbol{V} = \boldsymbol{M'\epsilon_{23}},$$

Por lo tanto la matriz de covarianza conjunta particionada con base en las dimensiones de $\boldsymbol{U}$ y $\boldsymbol{V}$ adopta la forma de la matriz (1.10) siendo ahora $\boldsymbol{P}$ una matriz de orden $n_1 \times n_2$, con $r = rango(\boldsymbol{\Sigma_{12\cdot 3}})$.

Como en el caso de dos variables, se puede derivar que los elementos no nulos de $\boldsymbol{P}$ son justamente las correlaciones entre las parejas de componentes de $\boldsymbol{U}$ y $\boldsymbol{V}$: (u_1, v_1),(u_2, v_2),...,(u_r, v_r). A tales elementos no nulos de $\boldsymbol{P}$ se les define como correlaciones canónicas parciales entre los vectores $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ y son una medida de interdependencia entre dichos vectores después de haber sido eliminada la atribución lineal de $\boldsymbol{X_3}$.

De la descomposición singular de la matriz $\boldsymbol{C} = \boldsymbol{\Sigma_{11\cdot 3}^{-\frac{1}{2}} \Sigma_{12\cdot 3} \Sigma_{22\cdot 3}^{-\frac{1}{2}}}$ se deriva que las matrices $\boldsymbol{L}$ y $\boldsymbol{M}$ de las ecuaciones anteriores se obtienen como solución de los sistemas:

$$[\boldsymbol{\Sigma_{12\cdot 3}\Sigma_{22\cdot 3}^{-1}\Sigma_{21\cdot 3}} - \lambda^2 \boldsymbol{\Sigma_{11\cdot 3}}]\boldsymbol{L} = 0$$

$$[\boldsymbol{\Sigma_{21\cdot 3}\Sigma_{11\cdot 3}^{-1}\Sigma_{12\cdot 3}} - \lambda^2 \boldsymbol{\Sigma_{22\cdot 3}}]\boldsymbol{M} = 0$$

Desarrollando de una manera análoga al caso en dos variables, se tiene:

$$\boldsymbol{L'\Sigma_{11\cdot 3}L} = \boldsymbol{I_{n_1}}$$

$$\boldsymbol{M'\Sigma_{22\cdot 3}M} = \boldsymbol{I_{n_2}}$$

Por lo tanto, la matriz de varianzas y covarianzas de los vectores $\boldsymbol{U}$ y $\boldsymbol{V}$ tiene la forma de (1.10) en donde los elementos no nulos de $\boldsymbol{P}$ son las raíces cuadradas con signo positivo de las soluciones comunes a las ecuaciones de los determinantes, es decir:

$$| \boldsymbol{\Sigma_{12\cdot 3}\Sigma_{22\cdot 3}^{-1}\Sigma_{21\cdot 3}} - \lambda^2 \boldsymbol{\Sigma_{11\cdot 3}} | = 0$$

$$| \boldsymbol{\Sigma_{21\cdot 3}\Sigma_{11\cdot 3}^{-1}\Sigma_{12\cdot 3}} - \lambda^2 \boldsymbol{\Sigma_{22\cdot 3}} | = 0$$

Una característica a destacar del **AC** parcial es la siguiente:

Si $(\boldsymbol{X_1^T}, \boldsymbol{X_2^T}, \boldsymbol{X_3^T})$ tienen una distribución conjunta normal, las correlaciones canónicas parciales entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ coinciden con las correlaciones canónicas según el sentido de los vectores $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ condicionado a $\boldsymbol{X_3}$, es decir, respecto a la regresión sobre $\boldsymbol{X_3}$.

7.2. Análisis canónico biparcial (G_2)

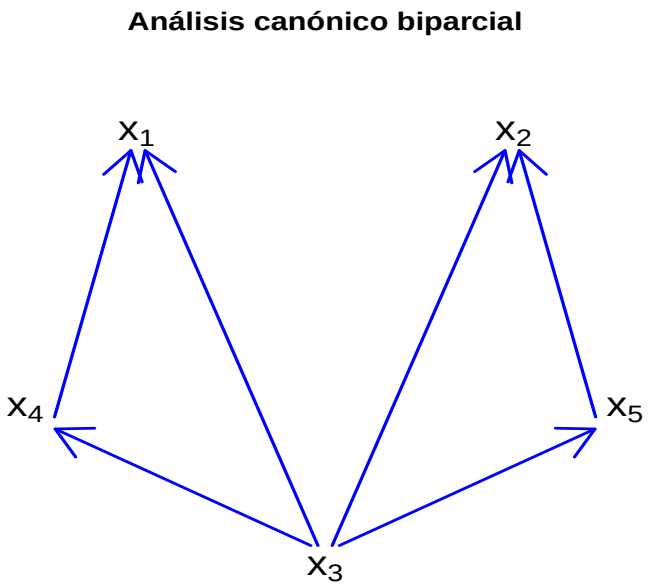
Considerando el caso en donde se tienen cinco conjuntos de variables aleatorias, se tiene entonces que realizar el análisis canónico para $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$ y $\boldsymbol{X_5}$, las cuales tienen dimensiones n_1, n_2, n_3, n_4 y n_5 respectivamente; con la condición $n_1 \leqslant n_2$. Con distribución conjunta con parámetros $(0, \boldsymbol{\Sigma})$; donde, $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma ij}], con$i,j= 1,2,3,4,5

Supóngase que la variable $\boldsymbol{X_3}$ está relacionada linealmente con cada una de las otras variables y que $\boldsymbol{X_4}$ y $\boldsymbol{X_5}$ lo están con $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ respectivamente.

Así, el análisis tiene como objetivo determinar la interdependencia entre las variables $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ después de eliminar la relación lineal de las otras variables. Es decir, primero se encuentran los residuos de las regresiones de $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ sobre $\boldsymbol{X_3}$,

entonces se toman los residuos como representación de las variables y se denota como ϵ_{13} y ϵ_{23} respectivamente, entonces se elimina el efecto lineal que aportan las variables $\boldsymbol{X_4}$ y $\boldsymbol{X_5}$ sobre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$, pero también el efecto que aportan sobre $\boldsymbol{X_3}$.

Figura 7: ***Representación gráfica de la relación lineal entre las variables.***



Entonces los vectores de residuos son:

$$\epsilon_{134} = \epsilon_{13} - \boldsymbol{\Sigma_{14\cdot 3}\Sigma_{44\cdot 3}^{-1}}\epsilon_{43}$$

$$\epsilon_{235} = \epsilon_{23} - \boldsymbol{\Sigma_{25\cdot 3}\Sigma_{55\cdot 3}^{-1}}\epsilon_{53}$$

Es decir, ϵ_{134} y ϵ_{235} son independientes por cómo se construyeron $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$, y expresan a su vez a los vectores $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ después de que se eliminaron las relaciones lineales supuestas y su matriz de varianzas y covarianzas conjunta es de la forma:

$$\boldsymbol{\Sigma_{\epsilon_{13},\epsilon_{23}}} = \begin{bmatrix} \boldsymbol{\Sigma^*_{11}} & \boldsymbol{\Sigma^*_{12}} \\ \boldsymbol{\Sigma^*_{21}} & \boldsymbol{\Sigma^*_{22}} \end{bmatrix}$$

donde:

$$\boldsymbol{\Sigma^*_{11}} = \boldsymbol{\Sigma_{11\cdot 3}} - \boldsymbol{\Sigma_{14\cdot 3}\Sigma^{-1}_{44\cdot 3}\Sigma_{41\cdot 3}}$$

$$\boldsymbol{\Sigma^*_{22}} = \boldsymbol{\Sigma_{22\cdot 3}} - \boldsymbol{\Sigma_{25\cdot 3}\Sigma^{-1}_{55\cdot 3}\Sigma_{52\cdot 3}}$$

$$\boldsymbol{\Sigma^*_{12}} = \boldsymbol{\Sigma_{12\cdot 3}} - \boldsymbol{\Sigma_{14\cdot 3}\Sigma^{-1}_{44\cdot 3}\Sigma_{42\cdot 3}} - \boldsymbol{\Sigma_{15\cdot 3}\Sigma^{-1}_{55\cdot 3}\Sigma_{52\cdot 3}} + \boldsymbol{\Sigma_{14\cdot 3}\Sigma^{-1}_{44\cdot 3}\Sigma_{45\cdot 3}\Sigma^{-1}_{55\cdot 3}\Sigma_{52\cdot 3}}$$

Por la estructura que tienen las primeras dos matrices, se define:

$$\boldsymbol{\Sigma^*_{11}} = \boldsymbol{\Sigma_{11\cdot 3\cdot 4}}$$

$$\boldsymbol{\Sigma^*_{22}} = \boldsymbol{\Sigma_{22\cdot 3\cdot 5}}$$

Dichas ecuaciones son definidas como correlaciones y variables canónicas G_2-Biparciales entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$, con residuales ϵ_{134} y ϵ_{235} respectivamente.

Por la construcción de las variables canónicas G_2-Biparciales entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ que son obtenidas a través de la transformación lineal que se derivó como en el caso parcial; en el caso de dos variables se tiene:

$$\boldsymbol{U = L'\epsilon_{134}}$$

$$\boldsymbol{V = M'\epsilon_{235}},$$

cuya matriz de varianzas y covarianzas tiene las características que se habían desarrollado para los casos anteriores.

Para este caso $r = rango(\boldsymbol{\Sigma^*_{12}})$, con elementos no nulos de $\boldsymbol{P}$ y ahora se definen como correlaciones canónicas G_2-Biparciales entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ y son una media de interdependencia entre dichos vectores después de eliminar los efectos lineales supuestos de $\boldsymbol{X_3}$, $\boldsymbol{X_4}$ y $\boldsymbol{X_5}$.

La descomposición singular de una matriz apropiada, que por construcción para este caso es $\boldsymbol{C} = \boldsymbol{\Sigma^{-\frac{1}{2}}_{11\cdot 3\cdot 4}\Sigma^*_{12}\Sigma^{-\frac{1}{2}}_{22\cdot 3\cdot 5}}$, permite desarrollar que si las matrices $\boldsymbol{L}$ y $\boldsymbol{M}$ anteriores cumplen las condiciones para los sistemas de ecuaciones siguientes:

$$[\boldsymbol{\Sigma^*_{12}\Sigma^{-1}_{22\cdot 3\cdot 5}\Sigma^*_{21}} - \lambda^2\boldsymbol{\Sigma_{11\cdot 3\cdot 4}}]\boldsymbol{L} = 0$$

$$[\boldsymbol{\Sigma^*_{21}\Sigma^{-1}_{11\cdot 3\cdot 4}\Sigma^*_{12}} - \lambda^2\boldsymbol{\Sigma_{22\cdot 3\cdot 5}}]\boldsymbol{M} = 0$$

Entonces, como en el caso de dos variables, se tiene:

$$\boldsymbol{L'\Sigma_{11\cdot 3\cdot 4}L} = \boldsymbol{I_{n_1}}$$

$$\boldsymbol{M'\Sigma_{22\cdot 3\cdot 5}M} = \boldsymbol{I_{n_2}}$$

Así, los vectores $\boldsymbol{U}$ y $\boldsymbol{V}$ tienen matriz de varianzas y covarianzas conjunta con las características de la matriz de varianzas y covarianzas para el caso de dos variables, en donde los elementos no nulos de $\boldsymbol{P}$ son las raíces cuadradas con signo positivo de las soluciones comunes de las ecuaciones de los determinantes:

$$|\ \boldsymbol{\Sigma^*_{12}\Sigma^{-1}_{22\cdot 3\cdot 5}\Sigma^*_{21}} - \lambda^2\ \boldsymbol{\Sigma_{11\cdot 3\cdot 4}}\ | = 0$$

$$|\ \boldsymbol{\Sigma^*_{21}\Sigma^{-1}_{11\cdot 3\cdot 4}\Sigma^*_{12}} - \lambda^2\ \boldsymbol{\Sigma_{22\cdot 3\cdot 5}}\ | = 0$$

Por lo tanto el análisis canónico G_2-Biparcial toma como caso particular al análisis canónico parcial, esto es, si el vector $\boldsymbol{X_3}$ está no correlacionado con $\boldsymbol{X_4}$

y $\boldsymbol{X_5}$ y estos últimos lo están con $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ respectivamente, las particiones de la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ pueden verse como $\boldsymbol{\Sigma_{34}}$, $\boldsymbol{\Sigma_{35}}$, $\boldsymbol{\Sigma_{14}}$ y $\boldsymbol{\Sigma_{25}}$ que son matrices de elementos nulos y los residuos ϵ_{134} y ϵ_{235} coinciden con ϵ_{13} y ϵ_{23} vectores del análisis canónico parcial.

7.3. Generalización del análisis canónico biparcial (G_2)

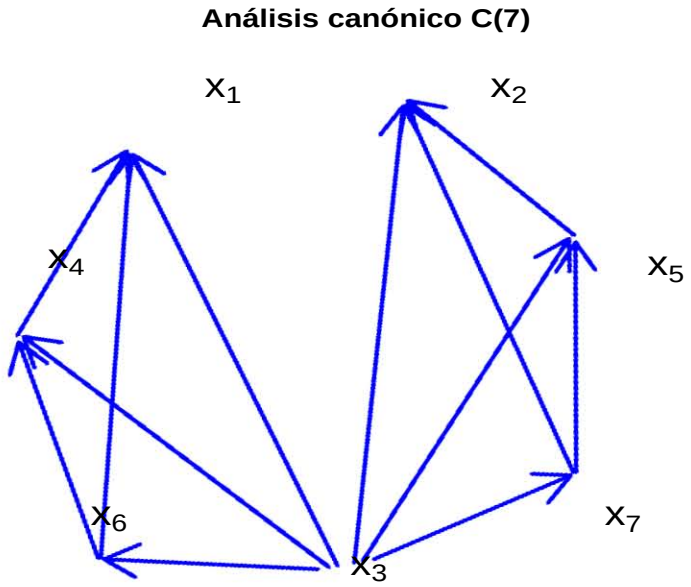
La generalización del **AC** biparcial, se define como análisis canónico $\boldsymbol{C(2n+1)}$; con $n \in \boldsymbol{N}$. Siguiendo la notación anterior, se denotará al análisis canónico parcial como análisis canónico $C(3)$ y al modelo G_2-Biparcial como análisis canónico $C(5)$.

A continuación se desarrolla el caso del análisis canónico C(7), que a su vez incluye como caso particular el modelo G_2-Biparcial y se generaliza este proceso por inducción para modelos de cualquier número impar de variables.

7.3.1. Análisis canónico $C(7)$

Sean $\boldsymbol{X_i}$ variables aleatorias de dimensiones n_i con $i = 1, 2, \ldots, 7$; con la condición $n_1 \leqslant n_2$, que se distribuyen conjuntamente con parámetros $(0, \boldsymbol{\Sigma})$; donde $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma ij}]$, con $i, j = 1, 2, 3, ..., 7$ y se supone una dependencia lineal entre ellos de manera tal, que la variable $\boldsymbol{X_3}$ está asociada linealmente a cada una de las otras variables, que las variables $\boldsymbol{X_6}$ y $\boldsymbol{X_7}$ están asociados a $\boldsymbol{X_4}$ y $\boldsymbol{X_5}$ y a $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ respectivamente; que $\boldsymbol{X_4}$ y $\boldsymbol{X_5}$ lo están con $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ respectivamente, lo citado anteriormente se muestra en la siguiente gráfica.

Figura 8: ***Representación gráfica de la relación lineal entre las variables.***



Por lo tanto el objetivo del análisis canónico $C(7)$ es determinar las interdependencias entre las variables $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ suponiendo que son eliminadas anteriormente las relaciones lineales supuestas, como se muestra en la Figura 8; la primera relación a eliminar es sobre la variable $\boldsymbol{X_1}$.

Pasos para eliminar el efecto lineal sobre $\boldsymbol{X_1}$

i) *Eliminación del efecto lineal de* $\boldsymbol{X_3}$ *sobre* $\boldsymbol{X_1}$*:*

$$\epsilon_{13} = \boldsymbol{X_1} - \boldsymbol{\Sigma_{13}\Sigma_{33}^{-1}X_3}$$

ii) *Eliminación del efecto lineal de* $\boldsymbol{X_3}$ *sobre* $\boldsymbol{X_6}$*:*

$$\epsilon_{63} = \boldsymbol{X_6} - \boldsymbol{\Sigma_{63}\Sigma_{33}^{-1}X_3}$$

iii) *Eliminación del efecto lineal de* ϵ_{63} *sobre* ϵ_{13}*:*

$$\epsilon_{136} = \epsilon_{13} - \boldsymbol{\Sigma_{16\cdot 3}\Sigma_{66\cdot 3}^{-1}\epsilon_{63}}$$

iv) *Eliminación del efecto lineal de* $\boldsymbol{X_3}$ *sobre* $\boldsymbol{X_4}$*:*

$$\epsilon_{43} = \boldsymbol{X_4} - \boldsymbol{\Sigma_{43}\Sigma_{33}^{-1}X_3}$$

v) *Eliminación del efecto lineal de* ϵ_{63} *sobre* ϵ_{43}*:*

$$\epsilon_{436} = \epsilon_{43} - \boldsymbol{\Sigma_{46\cdot 3}\Sigma_{66\cdot 3}^{-1}\epsilon_{63}}$$

vi) *Eliminación del efecto lineal de* ϵ_{436} *sobre* ϵ_{136}*:*

$$\epsilon_{1364} = \epsilon_{136} - \boldsymbol{\Sigma_{14\cdot 3\cdot 6}\Sigma_{44\cdot 3\cdot 6}^{-1}\epsilon_{436}}$$

Por lo tanto la expresión en el paso vi muestra cuando han sido eliminados todos los efectos lineales de las demás variables sobre $\boldsymbol{X_1}$.

Siguiendo un desarrollo similar para eliminar los efectos lineales sobre $\boldsymbol{X_2}$ de las demás variables se tiene:

Pasos para eliminar el efecto lineal sobre $\boldsymbol{X_2}$

i) *Eliminación del efecto lineal de* $\boldsymbol{X_3}$ *sobre* $\boldsymbol{X_2}$*:*

$$\epsilon_{23} = \boldsymbol{X_2} - \boldsymbol{\Sigma_{23}\Sigma_{33}^{-1}X_3}$$

ii) *Eliminación del efecto lineal de* $\boldsymbol{X_3}$ *sobre* $\boldsymbol{X_7}$*:*

$$\epsilon_{73} = \boldsymbol{X_7} - \boldsymbol{\Sigma_{73}\Sigma_{33}^{-1}X_3}$$

iii) *Eliminación del efecto lineal de* ϵ_{73} *sobre* ϵ_{23}*:*

$$\epsilon_{237} = \epsilon_{23} - \boldsymbol{\Sigma_{27\cdot3}\Sigma_{77\cdot3}^{-1}\epsilon_{73}}$$

iv) *Eliminación del efecto lineal de* $\boldsymbol{X_3}$ *sobre* $\boldsymbol{X_5}$*:*

$$\epsilon_{53} = \boldsymbol{X_5 - \Sigma_{53}\Sigma_{33}^{-1}X_3}$$

v) *Eliminación del efecto lineal de* ϵ_{73} *sobre* ϵ_{53}*:*

$$\epsilon_{537} = \epsilon_{53} - \boldsymbol{\Sigma_{57\cdot3}\Sigma_{77\cdot3}^{-1}\epsilon_{73}}$$

vi) *Eliminación del efecto lineal de* ϵ_{537} *sobre* ϵ_{237}*:*

$$\epsilon_{2375} = \epsilon_{237} - \boldsymbol{\Sigma_{25\cdot3\cdot7}\Sigma_{55\cdot3\cdot7}^{-1}\epsilon_{537}}$$

Por lo tanto la expresión en el paso vi muestra cuando han sido eliminados todos los efectos lineales de las demás variables sobre $\boldsymbol{X_2}$.

De las ecuaciones anteriores, se tiene la matriz de varianzas y covarianzas conjunta de tales vectores de residuos:

$$\boldsymbol{\Sigma_{\epsilon_{1364},\epsilon_{2375}}} = \begin{bmatrix} \boldsymbol{\Sigma_{11}^{**}} & \boldsymbol{\Sigma_{12}^{**}} \\ \boldsymbol{\Sigma_{21}^{**}} & \boldsymbol{\Sigma_{22}^{**}} \end{bmatrix}$$

donde:

$$\boldsymbol{\Sigma_{11}^{**} = \Sigma_{11\cdot3\cdot6\cdot4}}$$

$$\boldsymbol{\Sigma_{12}^{**} = \Sigma_{12\cdot3\cdot6} + \Sigma_{12\cdot3\cdot7} + \Sigma_{16\cdot3}\Sigma_{66\cdot3}^{-1}\Sigma_{67\cdot3}\Sigma_{77\cdot3}^{-1}\Sigma_{72\cdot3}}$$

$$\boldsymbol{-\Sigma_{14\cdot3\cdot6}\Sigma_{44\cdot3\cdot6}^{-1}[\Sigma_{42\cdot3\cdot6}\Sigma_{42\cdot3\cdot7} + \Sigma_{46\cdot3}\Sigma_{66\cdot3}^{-1}\Sigma_{67\cdot3}\Sigma_{77\cdot3}^{-1}\Sigma_{72\cdot3} - \Sigma_{42\cdot3}]}$$

$$-[\boldsymbol{\Sigma_{15\cdot3\cdot6}\Sigma_{15\cdot3\cdot7}+\Sigma_{16\cdot3}\Sigma_{66\cdot3}^{-1}\Sigma_{67\cdot3}\Sigma_{77\cdot3}^{-1}\Sigma_{75\cdot3}}]\boldsymbol{\Sigma_{55\cdot3\cdot7}^{-1}\Sigma_{52\cdot3\cdot7}}$$

$$+\boldsymbol{\Sigma_{14\cdot3\cdot6}\Sigma_{44\cdot3\cdot6}}[\boldsymbol{\Sigma_{45\cdot3\cdot6}\Sigma_{45\cdot3\cdot7}+\Sigma_{46\cdot3}\Sigma_{66\cdot3}^{-1}\Sigma_{67\cdot3}\Sigma_{77\cdot3}^{-1}\Sigma_{75\cdot3}-\Sigma_{45\cdot3}}]\boldsymbol{\Sigma_{55\cdot3\cdot7}^{-1}\Sigma_{52\cdot3\cdot7}}$$

$$\boldsymbol{\Sigma_{22}^{**}=\Sigma_{22\cdot3\cdot7\cdot5}}\ ,$$

tienen una estructura similar a las obtenidas para el caso parcial y el caso biparcial (G_2).

Las correlaciones y variables canónicas $C(7)$ entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ son las correlaciones y variables entre ϵ_{1364} y ϵ_{2375}. Por lo tanto las variables canónicas $C(7)$ serán los vectores transformados $\boldsymbol{U}$ y $\boldsymbol{V}$:

$$\boldsymbol{U=L'\epsilon_{1364}}$$

$$\boldsymbol{V=M'\epsilon_{2375}},$$

cuya matriz de varianzas y covarianzas tiene la forma de la matriz para el caso parcial.

Con $r=rango(\boldsymbol{\Sigma_{12}^{**}})$ y los elementos no nulos de la matriz $\boldsymbol{P}$ se definen como correlaciones canónicas $C(7)$ entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$.

Por la descomposición singular de $\boldsymbol{C=\Sigma_{11\cdot3\cdot6\cdot4}^{-\frac{1}{2}}\Sigma_{12}^{**}\Sigma_{22\cdot3\cdot7\cdot5}^{-\frac{1}{2}}}$, se pueden encontrar las matrices de transformación $\boldsymbol{L}$ y $\boldsymbol{M}$de las que se derivan los vectores $\boldsymbol{U}$ y $\boldsymbol{V}$ que son soluciones de los sistemas:

$$[\boldsymbol{\Sigma_{12}^{**}\Sigma_{22\cdot3\cdot7\cdot5}^{-1}\Sigma_{21}^{**}}-\lambda^2\boldsymbol{\Sigma_{11\cdot3\cdot6\cdot4}}]\,\boldsymbol{L}=0$$

$$[\boldsymbol{\Sigma_{21}^{**}\Sigma_{11\cdot3\cdot6\cdot4}^{-1}\Sigma_{12}^{**}}-\lambda^2\boldsymbol{\Sigma_{22\cdot3\cdot7\cdot5}}]\,\boldsymbol{M}=0$$

Las cuales cumplen que:

$$\boldsymbol{L'\Sigma_{11\cdot 3\cdot 6\cdot 4}L = I_{n_1}}$$

$$\boldsymbol{M'\Sigma_{22\cdot 3\cdot 7\cdot 5}M = I_{n_2}}$$

donde las correlaciones canónicas $C(7)$ de las r raíces cuadradas con signo positivo de las soluciones comunes de las ecuaciones de los determinantes son:

$$| \boldsymbol{\Sigma^{**}_{12}\Sigma^{-1}_{22\cdot 3\cdot 7\cdot 5}\Sigma^{**}_{21}} - \lambda^2\boldsymbol{\Sigma_{11\cdot 3\cdot 6\cdot 4}} | = 0$$

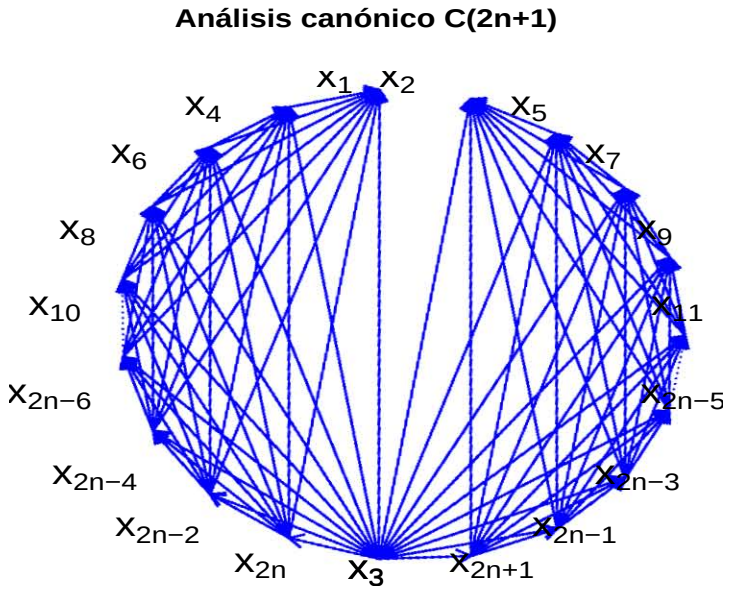
$$| \boldsymbol{\Sigma^{**}_{21}\Sigma^{-1}_{11\cdot 3\cdot 6\cdot 4}\Sigma^{**}_{12}} - \lambda^2\boldsymbol{\Sigma_{22\cdot 3\cdot 7\cdot 5}} | = 0$$

En el caso en el cual $\boldsymbol{X_3}$ está no correlacionada con $\boldsymbol{X_6}$ y $\boldsymbol{X_7}$ y estas últimas lo están con $\boldsymbol{X_1}$ y $\boldsymbol{X_4}$ y con $\boldsymbol{X_2}$ y $\boldsymbol{X_5}$ respectivamente, las particiones de la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ son entonces $\boldsymbol{\Sigma_{36}}$, $\boldsymbol{\Sigma_{37}}$, $\boldsymbol{\Sigma_{64}}$, $\boldsymbol{\Sigma_{61}}$, $\boldsymbol{\Sigma_{75}}$ y $\boldsymbol{\Sigma_{72}}$ que son matrices de elementos nulos y los residuos ϵ_{1364} y ϵ_{2375} coinciden con ϵ_{134} y ϵ_{235} vectores del análisis canónico biparcial $C(5)$.

7.3.2. Análisis canónico $C(2n+1)$

Sean $\boldsymbol{X_i}$ variables aleatorias de dimensiones n_i con $i = 1, 2, \ldots, 2n+1$; con la condición de $n_1 \leqslant n_2$, distribuidas conjuntamente con parámetros $(0, \boldsymbol{\Sigma})$ con la matriz $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma ij}]$, con $i, j = 1, 2, 3, ..., 2n+1$.
Supóngase que existe una dependencia lineal entre ellos como se muestra en la siguiente figura.

Figura 9: ***Representación gráfica de la relación lineal entre las variables.***



Mediante el análisis canónico $\boldsymbol{C(2n+1)}$ se detectan las interdependencias entre las variables $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ después de haber sido eliminada la relación lineal con el resto de las variables.

Siguiendo una analogía para la obtención de los residuos en el caso del análisis canónico parcial, biparcial y el análisis canónico $C(7)$, puede derivarse la estructura de los residuos que representan a $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ después de ser eliminadas las relaciones lineales que existen para las variables.

Por lo tanto, la estructura que representa a las variables $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ por los residuos está dada de la siguiente manera:

$$\epsilon_{13(2n)(2n-2)\cdots 10\,8\,6\,4} = \epsilon_{13(2n)(2n-2)\cdots 10\,8\,6} - [\boldsymbol{\Sigma_{14\cdot 3\cdot 2n\cdots 8\cdot 6}\Sigma^{-1}_{44\cdot 3\cdot 2n\cdots 8\cdot 6}}]\epsilon_{43(2n)\cdots 10\,8\,6}$$

$$\epsilon_{23(2n+1)(2n-1)\cdots 11\,9\,7\,5} = \epsilon_{23(2n+1)(2n-1)\cdots 11\,9\,7} - [\boldsymbol{\Sigma_{2\cdot 5\cdot 3\cdot 2n+1\cdots 9\cdot 7}\Sigma^{-1}_{55\cdot 3\cdot 2n+1\cdots 9\cdot 7}}]\epsilon_{53(2n+1)\cdots 11\,9\,7}$$

donde las particiones $\boldsymbol{\Sigma_{14\cdot3\cdot2n\cdots8\cdot6}}$, $\boldsymbol{\Sigma_{44\cdot3\cdot2n\cdots8\cdot6}}$, $\boldsymbol{\Sigma_{2\cdot5\cdot3\cdot2n+1\cdots9\cdot7}}$, $\boldsymbol{\Sigma_{55\cdot3\cdot2n+1\cdots9\cdot7}}$ tienen una estructura similar a las de las ecuaciones para los casos anteriores, pero generalizado al caso que se está analizando.

De las ecuaciones de los residuos, es inmediato llegar a la matriz de varianzas y covarianzas conjunta:

$$\boldsymbol{\Sigma}_{\epsilon_{13(2n)(2n-2)\cdots10\,8\,6\,4},\epsilon_{23(2n+1)(2n-1)\cdots11\,9\,7\,5}} = \begin{bmatrix} \boldsymbol{\Sigma_{11\cdot3\cdot2n\cdots8\cdot6\cdot4}} & \boldsymbol{\Sigma_{12}^{***}} \\ \boldsymbol{\Sigma_{21}^{***}} & \boldsymbol{\Sigma_{22\cdot3\cdot2n+1\cdots9\cdot7\cdot5}} \end{bmatrix}$$

donde $\boldsymbol{\Sigma_{12}^{***}}$ es de manera análoga al caso del análisis canónico $C(7)$.

De forma similar a la construcción generalizada para el análisis canónico $C(7)$ se sigue un proceso similar al resto de los modelos canónicos anteriores y se denominarán correlaciones y variables canónicas $\boldsymbol{C(2n+1)}$ entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ a las correlaciones y variables canónicas entre los residuos $\epsilon_{13(2n)(2n-2)\cdots10\,8\,6\,4}$ y $\epsilon_{23(2n+1)(2n-1)\cdots11\,9\,7\,5}$, así las variables canónicas son:

$$\boldsymbol{U = L'\epsilon_{13(2n)(2n-2)\cdots10\,8\,6\,4}}$$

$$\boldsymbol{V = M'\epsilon_{23(2n+1)(2n-1)\cdots11\,9\,7\,5}},$$

cuya matriz de varianzas y covarianzas conjunta posee las mismas características que las matrices correspondientes en los casos anteriores y los $r = rango(\boldsymbol{\Sigma_{12}^{***}})$ elementos no nulos de $\boldsymbol{P}$ y se definen como las correlaciones canónicas $\boldsymbol{C(2n+1)}$ entre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$.

Utilizando la descomposición singular de

$$\boldsymbol{C} = \boldsymbol{\Sigma}^{-\frac{1}{2}}_{\boldsymbol{11\cdot 3\cdot (2n)\cdot (2n-2)\cdots 8\ 6\ 4}} \boldsymbol{\Sigma}^{***}_{\boldsymbol{12}} \boldsymbol{\Sigma}^{-\frac{1}{2}}_{\boldsymbol{22\cdot 3\cdot (2n+1)(2n-1)\cdots 11\cdot 9\ 7\ 5}},$$

se demuestra que las matrices $\boldsymbol{L}$ y $\boldsymbol{M}$ dan lugar a los vectores $\boldsymbol{U}$ y $\boldsymbol{V}$ con matriz de varianzas y covarianzas de forma análoga al caso parcial, entonces se tienen las soluciones de los sistemas:

$$[\boldsymbol{\Sigma}^{***}_{\boldsymbol{12}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{22\cdot 3\cdot (2n+1)\cdot (2n-1)\cdots\ 11\cdot\ 9\cdot 7\cdot 5}} \boldsymbol{\Sigma}^{***}_{\boldsymbol{21}} - \lambda^2 \boldsymbol{\Sigma}_{\boldsymbol{11\cdot 3\cdot (2n)\cdots 8\cdot 6\cdot 4}}]\ \boldsymbol{L} = 0$$

$$[\boldsymbol{\Sigma}^{***}_{\boldsymbol{21}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{11\cdot 3\cdot (2n)\cdots\ 8\cdot\ 6\cdot 4}} \boldsymbol{\Sigma}^{***}_{\boldsymbol{21}} - \lambda^2 \boldsymbol{\Sigma}_{\boldsymbol{22\cdot 3\cdot (2n+1)\cdot (2n-1)\cdots 11\cdot 9\cdot 7\cdot 5}}]\ \boldsymbol{M} = 0$$

Las cuales cumplen que:

$$\boldsymbol{L'\Sigma_{11\cdot 3\cdot (2n)\cdots 8\cdot 6\cdot 4} L} = \boldsymbol{I_{n_1}}$$

$$\boldsymbol{M'\Sigma_{22\cdot 3\cdot (2n+1)\cdot (2n-1)\cdots 11\cdot 9\cdot 7\cdot 5} M} = \boldsymbol{I_{n_2}}$$

y las correlaciones canónicas $\boldsymbol{C(2n+1)}$ con las raíces cuadradas con signo positivo de las soluciones comunes de las ecuaciones de los determinantes son:

$$|\ \boldsymbol{\Sigma}^{***}_{\boldsymbol{12}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{22\cdot 3\cdot (2n+1)\cdot (2n-1)\cdots\ 11\cdot\ 9\cdot 7\cdot 5}} \boldsymbol{\Sigma}^{***}_{\boldsymbol{21}} - \lambda^2 \boldsymbol{\Sigma}_{\boldsymbol{11\cdot 3\cdot (2n)\cdots 8\cdot 6\cdot 4}}\ | = 0$$

$$|\ \boldsymbol{\Sigma}^{***}_{\boldsymbol{21}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{11\cdot 3\cdot (2n)\cdots\ 8\cdot\ 6\cdot 4}} \boldsymbol{\Sigma}^{***}_{\boldsymbol{12}} - \lambda^2 \boldsymbol{\Sigma}_{\boldsymbol{22\cdot 3\cdot (2n+1)\cdot (2n-1)\cdots 11\cdot 9\cdot 7\cdot 5}}\ | = 0$$

De manera análoga al resto de los análisis canónicos anteriores, se puede ver que si $\boldsymbol{X_3}$ está no correlacionada con el resto de las variables, las particiones de la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$, son matrices de elementos nulos y los residuos coinciden con los residuos del análisis para $n-1$ vectores, por tanto el análisis

canónico $\boldsymbol{C(2n+1)}$ es entonces el caso del análisis canónico $\boldsymbol{C(2(n-1)+1)}$.

Si se considera el modelo canónico $\boldsymbol{C(2n+1)}$ bajo la hipótesis de normalidad en la distribución conjunta de los vectores iniciales, los vectores de residuos poseen matrices de varianzas y covarianzas que coinciden con las de las distribuciones de las variables condicionadas, es decir:

$$\boldsymbol{X_1 \mid X_3, X_{2n}, \cdots, X_6, X_4} \quad \text{y} \quad \boldsymbol{X_2 \mid X_3, X_{2n+1}, \cdots, X_7, X_5}$$

Capítulo 8

Ejemplo Práctico

Un investigador ha recopilado 600 datos sobre tres variables psicológicas, cuatro variables académicas y el género (las puntuaciones de las pruebas están en términos estándares de la evaluación), para estudiantes del primer año universitario. El interés principal es conocer cómo el conjunto de variables psicológicas se relacionan con las variables académicas y de género, en particular, determinar el número de variables canónicas estadísticamente significativas que son necesarias para entender la relación entre los dos conjuntos de variables. *

La base de datos tiene 600 observaciones sobre ocho variables. Las variables psicológicas son:

Locus: *"ideas principales, que en la personalidad representa una interacción del individuo con su medio ambiente y que no se puede hablar de la personalidad de un individuo"* (Ap.3).

Motivación: *"combinación de procesos intelectuales, fisiológicos y psicológicos que decide, en una situación dada, con qué vigor se actúa y en qué dirección se*

* para saber más sobre el contexto de los datos, visitar la liga http://www.ats.ucla.edu/stat/spss/dae/canonical.htm

encauza la energía" (Ap.4).

Concepto Propio: *"conjunto de percepciones organizado jerárquicamente, coherente y estable, aunque también susceptible de cambios, que se construye por interacción a partir de las relaciones interpersonales."* (Ap.5).

Las variables académicas son: las pruebas de **matemáticas**, **lectura**, **redacción** y **ciencias**. La variable de género es de tipo **dummy**, definida como: cero indica que es *hombre* y uno que indica que es *mujer*.

8.1. Análisis descriptivo

El análisis descriptivo siempre es importante desarrollarlo antes de cualquier otro análisis, debido a que la mayoría de resultados posteriores se pueden suponer desde lo descriptivo, además, para el **AC**, las representaciones gráficas y tablas ayudan a comprender mejor todos los resultados, debido a su gran complejidad en el álgebra matricial.

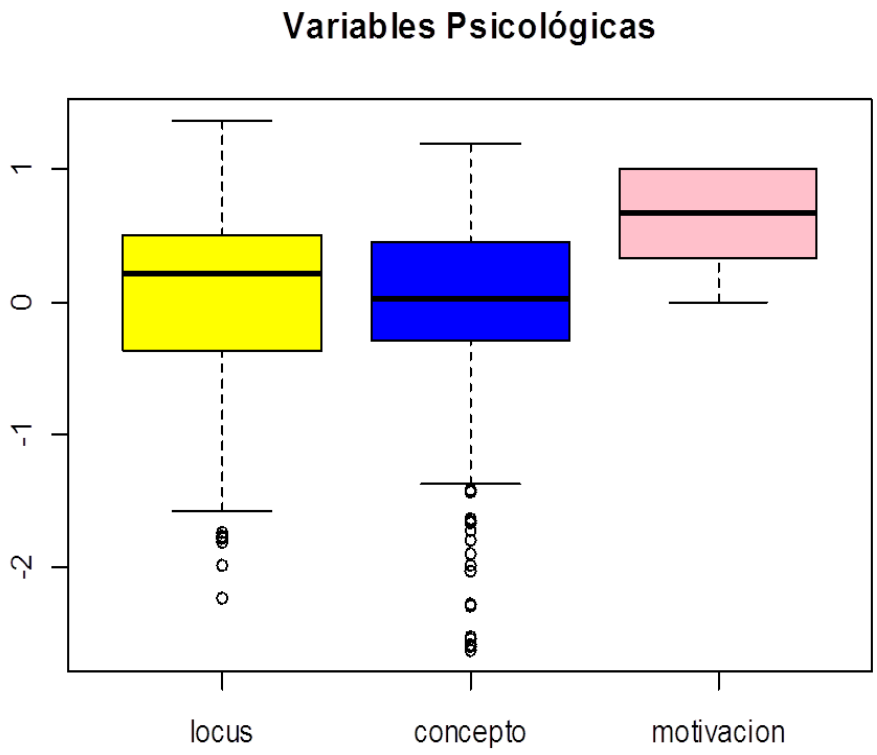
El análisis descriptivo se desarrolla para cada conjunto de variables, es decir, para todas las variables psicológicas, académicas y género.

Variables Psicológicas

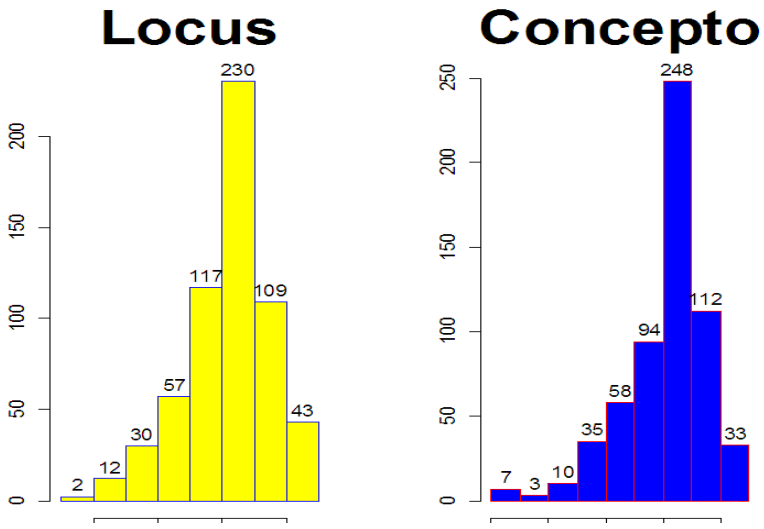
En la siguiente tabla se muestran valores estadísticos importantes de conocer de cada variable como lo es: la media, la desviación estándar, el valor mínimo, el primer cuartil, la mediana, el tercer cuartil, el valor máximo y la cantidad de valores ausentes para cada variable.

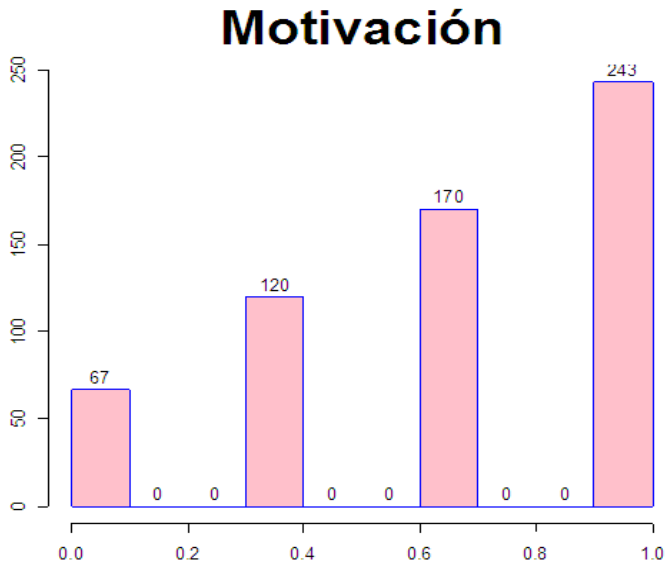
	N	media	Desv. Est.	min	Q1	mediana	Q3	max	VA
locus	600	0.09653	0.67027	-2.23	-0.3725	0.21	0.51	1.36	0
concepto	600	0.00491	0.70551	-2.62	-0.3000	0.03	0.44	1.19	0
motivación	600	0.66083	0.34272	0.00	0.3300	0.67	1.00	1.00	0

La media de las variables psicológicas tiene un rango de separación muy grande respecto una de la otra; los valores de las desviaciones estándar son menores a 1, por lo que se puede decir que las observaciones están a menos de una unidad de distancia respecto a su media, es decir, no están separados; no existe ningún valor perdido para ninguna variable, por lo que no es necesario realizar una imputación de datos. Los valores de la tabla anterior se resumen en la siguiente gráfica.



Una de las gráficas importantes en el análisis descriptivo son las gráficas de frecuencia mejor conocidas como histogramas, son importantes ya que se puede suponer el comportamiento de la función de distribución de cada variable de acuerdo a la forma que tienen; para el ejemplo, se observa que la variable locus y concepto tienen un comportamiento similar, mientras que para la variable motivación, se tiene que sólo toma valores de $\{0, 0.33, 0.67, 1\}$, y de estos cuatro valores se tiene un crecimiento en el comportamiento de las frecuencias.



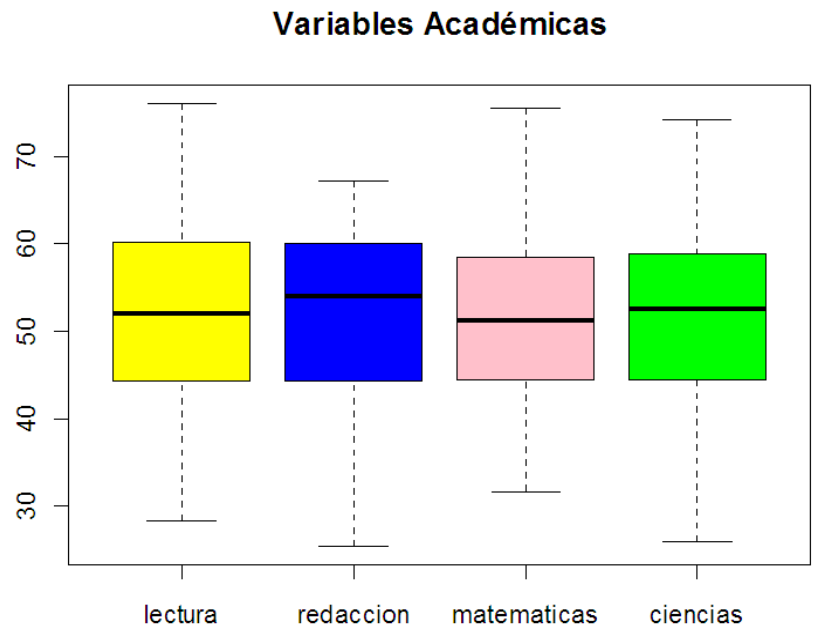


Variables Académicas

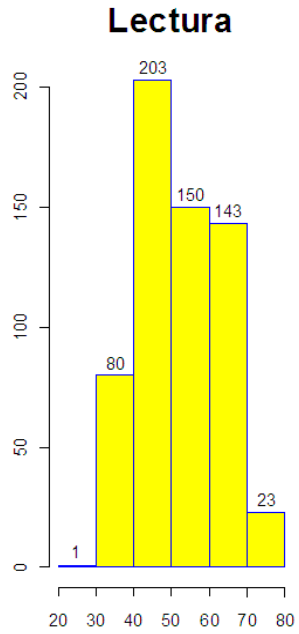
Análogamente, en la siguiente tabla se tiene las estadísticas importantes antes mencionadas.

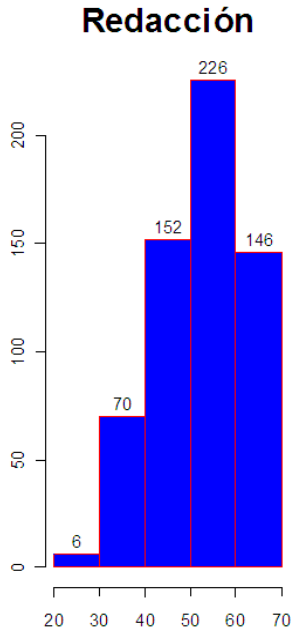
	N	media	Desv. Est.	min	Q1	mediana	Q3	max	VA
lectura	600	51.90183	10.10298	28.3	44.2	52.1	60.10	76.0	0
redacción	600	52.38483	9.72646	25.5	44.3	54.1	59.90	67.1	0
matemáticas	600	51.84900	9.41474	31.8	44.5	51.3	58.38	75.5	0
ciencias	600	51.76333	9.70618	26.0	44.4	52.6	58.65	74.2	0
mujer	600	0.54500	0.49839	0.0	0.0	1.0	1.00	1.0	0

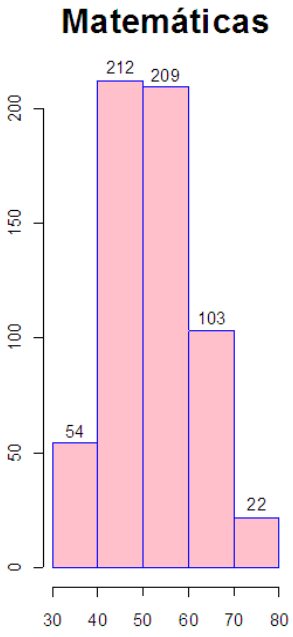
Sin tomar en cuenta la variable de género, las medias de las variables son muy parecidas, es decir, el rango de separación entre cada media no es grande; dado que se tiene un amplio margen de separación entre el valor mínimo y el máximo de cada variable y la desviación estándar es aproximadamente de 10, indica que ésta no es muy grande, por lo que, los datos están aproximadamente a 10 unidades respecto a su media. Lo anterior se muestra en la siguiente gráfica.

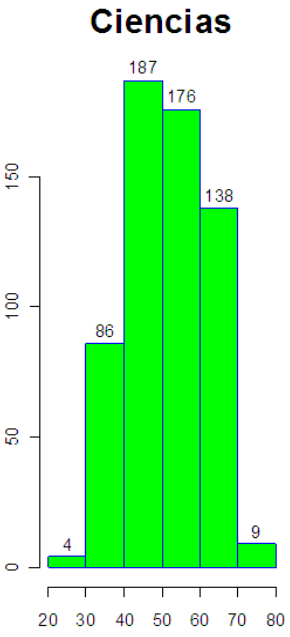


El comportamiento de las variables académicas es muy heterogéneo, en sentido de que es difícil hacer una descripción adecuada de su comportamiento para todas las variables; los histogramas muestran que la distribución de cada variable es unimodal, pero su comportamiento es difícil de interpretar, ya que los sesgos no están bien definidos.



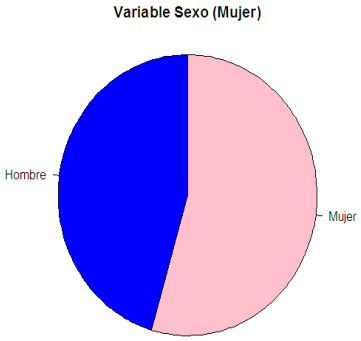






La variable género es una variable categórica y en el objetivo del estudio se desea saber si el género influye en el resto de las variables, por lo que es importante saber si el número de mujeres es semejante al número de hombres, debido a que el género se consideró como una sola variable. En la siguiente gráfica de pastel se observa que la muestra está equilibrada.

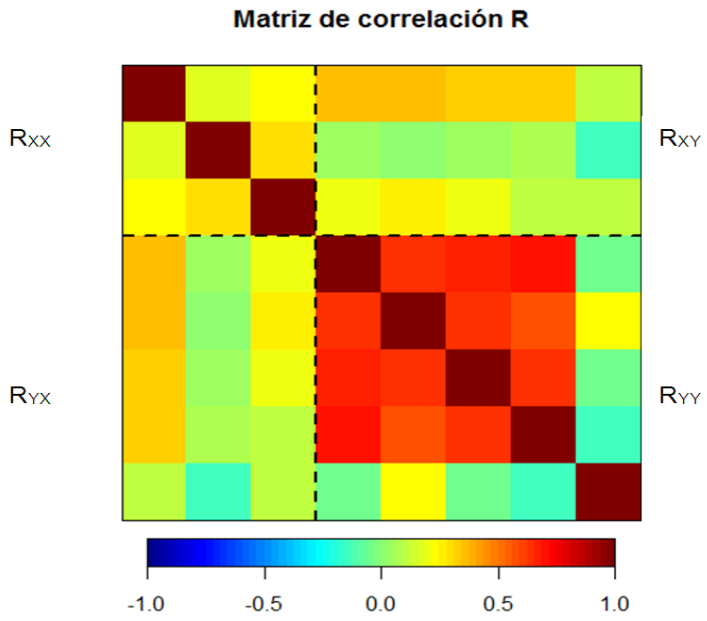
género	Casos	Total %
0	273	45.5
1	327	54.5
Sum	600	100.0



8.2. Análisis canónico

Siguiendo el desarrollo teórico del **AC**, se tiene que la matriz de las variables psicológicas $(\boldsymbol{X})$ es una matriz de rango 3, la matriz de las variables académicas y la variable género $(\boldsymbol{Y})$ es una matriz de rango 5, por lo tanto, el número máximo de variables canónicas que se pueden encontrar es 3.

Para realizar el **AC** es necesario calcular la matriz de correlación $\boldsymbol{R}$, a continuación se muestra una representación gráfica de dicha matriz dividida en las submatrices $\boldsymbol{R_{XX}}$, $\boldsymbol{R_{XY}}$ y $\boldsymbol{R_{YY}}$.



Las entradas de la matriz de correlación son valores de -1 a 1, dichos valores se representan con una escala de colores en donde el color azul marino indica que existe correlación negativa, el color rojo representa la correlación positiva y el color verde representa la no correlación entre variables. A partir de la representación gráfica se tiene que existen relaciones entre variables, es decir, existen correlaciones distintas de cero. Para la matriz $\boldsymbol{R_{XX}}$ se tiene que la correlación entre las variables de ese conjunto es positiva, ya que por la escala de color tiene valores entre 0.15 y 0.3; para la matriz $\boldsymbol{R_{YY}}$, por las tonalidades rojas y verdes, se observa que existen valores de correlación positiva mayor que 0.5, y correlaciones negativas cercanas a 0 respectivamente; para la matriz $\boldsymbol{R_{XY}}$ se tiene que hay correlaciones muy cercanas a cero por los tonos de color verde, sin embargo se observa una franja de tonalidad naranja, que indica correlación positiva no mayor a 0.5.

Al ser el una técnica general, el interés principal es conocer las relaciones que existen dentro y entre los conjuntos $\boldsymbol{X}$ y $\boldsymbol{Y}$, se analizan a continuación las siguientes matrices de correlación:

Submatriz de correlación $\boldsymbol{R_{XX}}$

Matriz R_{XX}	locus	concepto	motivación
locus	1.0000	0.1712	0.24513
concepto	0.1712	1.0000	0.28857
motivación	0.2451	0.2885	1.00000

Se observa que no parecen correlaciones grandes entre las variables del conjunto $\boldsymbol{X}$, por lo que se puede suponer que las variables psicológicas no influyen una conjuntamente con otra para los puntajes de las pruebas, por ejemplo, el locus o concepto o motivación no influye conjuntamente con otra.

Submatriz de correlación $\boldsymbol{R_{YY}}$

Matriz R_{YY}	lectura	redacción	matemáticas	ciencias	mujer
lectura	1.0000	0.6286	0.6793	0.6907	-0.04174
redacción	0.6286	1.0000	0.6327	0.5691	0.24433
matemáticas	0.6793	0.6327	1.0000	0.6495	-0.04822
ciencias	0.6907	0.5691	0.6495	1.0000	-0.13819
mujer	-0.0417	0.2443	-0.0482	-0.1381	1.00000

Las correlaciones altas corresponden a los puntajes de lectura con redacción, matemáticas y ciencias, redacción con matemáticas y matemáticas con ciencias.

Matriz de correlación R

Matriz R	locus	concepto	motivacion	lectura	redaccion	matemáticas	ciencias	mujer
locus	1.0000	0.1712	0.2451	0.3736	0.3589	0.3373	0.3246	0.1134
concepto	0.1712	1.0000	0.2886	0.0607	0.0194	0.0536	0.0698	-0.1259
motivación	0.2451	0.2886	1.0000	0.2106	0.2542	0.1950	0.1157	0.0981
lectura	0.3736	0.0607	0.2106	1.0000	0.6286	0.6793	0.6907	-0.0417
redacción	0.3589	0.0194	0.2542	0.6286	1.0000	0.6327	0.5691	0.2443
matemática	0.3373	0.0536	0.1950	0.6793	0.6327	1.0000	0.6495	-0.0482
ciencias	0.3246	0.0698	0.1157	0.6907	0.5691	0.6495	1.0000	-0.1381
mujer	0.113	-0.125	0.098	-0.041	0.244	-0.048	-0.138	1.000

Las correlaciones más altas dentro de la matriz $\boldsymbol{R_{XY}}$ son: locus con lectura, redacción, matemáticas y ciencias, aunque todo estos son valores muy bajos.

Los coeficientes de las combinaciones lineales de las variables canónicas son los siguientes:

Conjunto X:

	CV 1	CV 2	CV 3
locus	1.2538339	-0.6214775	-0.6616896
concepto	-0.3513499	-1.1876867	0.8267209
motivación	1.2624203	2.0272641	2.0002284

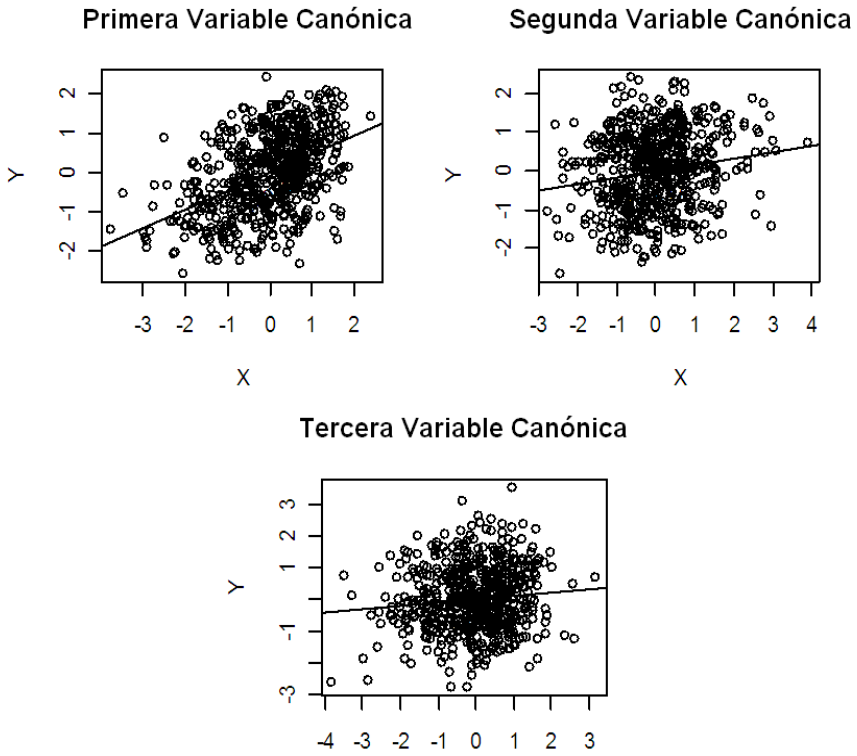
Conjunto Y:

	CV 1	CV 2	CV 3
lectura	0.044620596	-0.004910018	0.02138056
redacción	0.035877112	0.042071471	0.09130733
matemáticas	0.023417185	0.004229472	0.00939821
ciencias	0.005025157	-0.085162175	-0.10983502
mujer	0.632119239	1.084642482	-1.79464692

Dados los coeficientes de las combinaciones lineales, se tiene que la primera variable canónica para el conjunto de variables psicológicas es la combinación $\boldsymbol{u} = \boldsymbol{LX}$, donde los valores del vector $\boldsymbol{L}$ correspondientes a la primera variable canónica son los coeficientes de la primera columna del conjunto $\boldsymbol{X}$, esto es $\boldsymbol{L_1'} = (1.253, -0.351, 1.262)$, el primer vector canónico, así para la segunda variable canónica, el segundo vector canónico es $\boldsymbol{L_2'} = (-0.621, -1.187, 2.027)$ el vector de coeficientes y la tercera combinación de la tercera variable canónica tiene los coeficientes del tercer vector canónico $\boldsymbol{L_3'} = (-0.661, 0.826, 2.000)$.

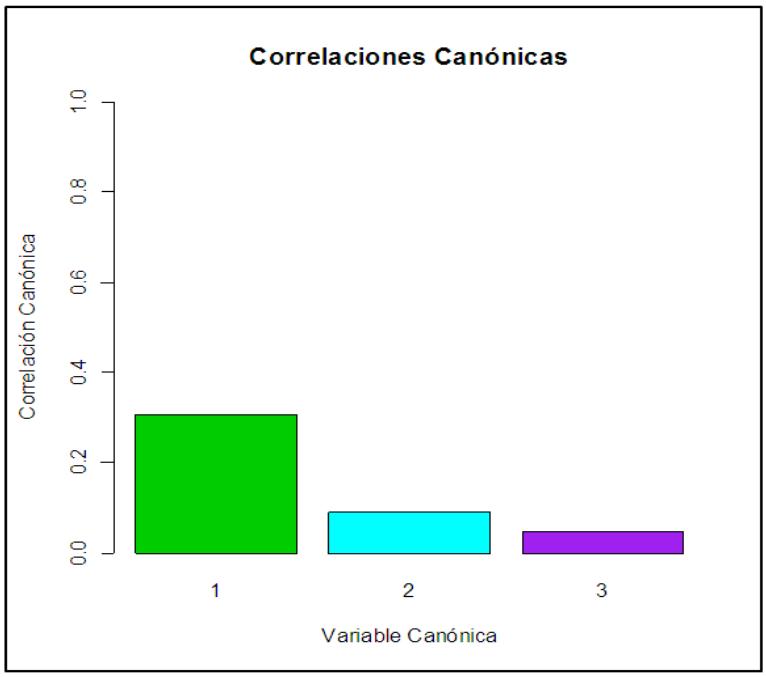
Para el conjunto de variables académicas se tiene que la primera variable canónica es la combinación lineal $\boldsymbol{v} = \boldsymbol{MY}$, donde el primer vector canónico para el conjunto $\boldsymbol{Y}$ es $\boldsymbol{M_1'} = (0.044, 0.035, 0.023, 0.005, 0.632)$ y así sucesivamente para la segunda y tercera variable canónica para el conjunto $\boldsymbol{Y}$.

La interpretación de los coeficientes de los vectores canónicos, es análoga al caso de regresión múltiple para cada variable dadas las demás variables fijas, por ejemplo, el aumento de una unidad en el concepto conduce a una disminución de 0.3513 unidades en la primera variable canónica cuando todas las otras variables se mantienen constantes; por otro lado el ser mujer lleva a un aumento de 0.6321 en la misma variable canónica si el resto se mantiene constante.



Los valores de las correlaciones canónicas son:

CV 1	CV 2	CV 3
rho 1	rho 2	rho 2
0.4640861	0.1675091	0.1039911



Estas correlaciones son calculadas entre las variables canónicas $\boldsymbol{u}$ y $\boldsymbol{v}$, es por ello que reciben el nombre de correlaciones canónicas, por lo tanto la correlación de las primeras variables canónicas $(\boldsymbol{u_1}, \boldsymbol{v_1})$ es 0.464, la correlación de las segundas variables canónicas $(\boldsymbol{u_2}, \boldsymbol{v_2})$ es 0.167 y la correlación de las terceras variables canónicas $(\boldsymbol{u_3}, \boldsymbol{v_3})$ es 0.104 y cumplen con la condición de que la primera correlación es mayor que la segunda correlación y a su vez, la segunda correlación es mayor que la tercera.

Se calculan los **coeficientes de estructura** que son las correlaciones entre las variables observadas y las variables canónicas:

Coeficientes de estructura:

Conjunto X:

	CV 1	CV 2	CV 3
locus	0.90404632	-0.3896883	-0.1756227
concepto	0.02084327	-0.7087386	0.7051632
motivación	0.56715105	0.3508882	0.7451290

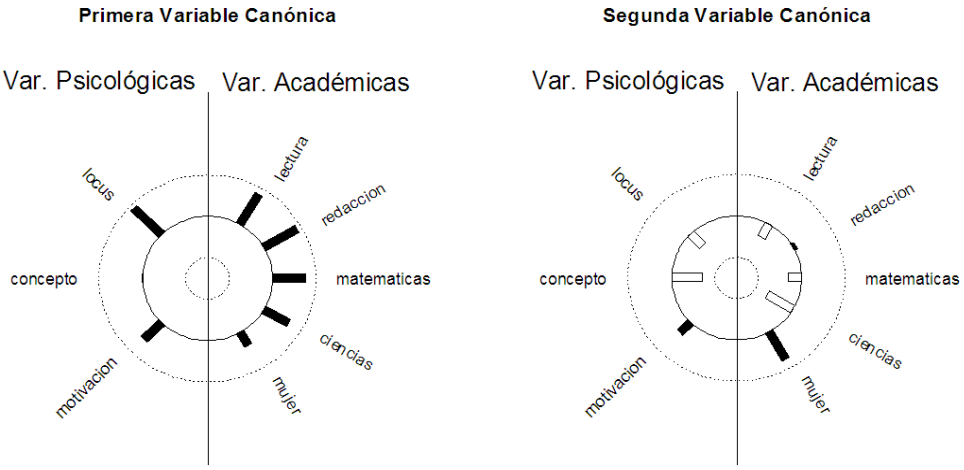
Conjunto Y:

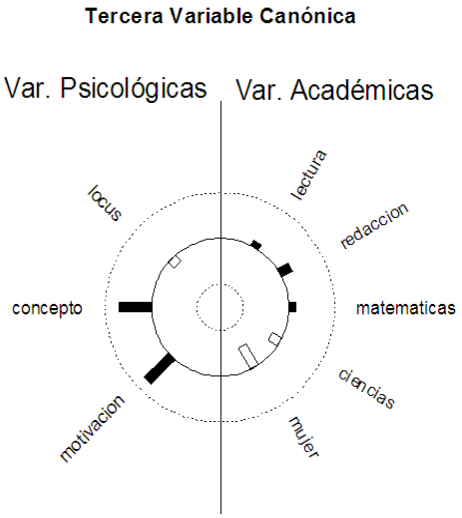
	CV 1	CV 2	CV 3
lectura	0.8404480	-0.35882539	0.1353635
redacción	0.8765429	0.06483672	0.2545609
matemáticas	0.7639483	-0.29794886	0.1477612
ciencias	0.6584139	-0.67679758	-0.2303551
mujer	0.3641127	0.75492816	-0.5434035

A partir de los coeficientes de estructura, se tiene que para la primera variable canónica, para el conjunto $\boldsymbol{X}$ la variable menos correlacionada es la variable concepto con 0.0208, seguida de la variable motivación con 0.567 y por último la variable más correlacionada es locus con una correlación de 0.904; para el conjunto $\boldsymbol{Y}$ la variable menos correlacionada es género con 0.3641, seguida de la variable ciencias con 0.658 y así sucesivamente hasta la variable redacción con una correlación de 0.876; para la segunda variable canónica se tiene que la variable concepto para el conjunto $\boldsymbol{X}$ es la más correlacionada negativamente con -0.708, por otro lado, para el conjunto $\boldsymbol{Y}$ las variables lectura, matemáticas y ciencias tienen correlación negativa con la segunda variable canónica, siendo ciencias la más correlacionada negativamente con -0.676; finalmente para la tercera variable canónica con el conjunto $\boldsymbol{X}$ se tiene que las variables de concepto y motivación tienen altas correlaciones, 0.705 y 0.715 respectivamente, para el conjunto $\boldsymbol{Y}$ se

tiene que la variable género tiene la mayor correlación negativa con -0.543.

Los valores de los coeficientes se representan en las siguientes gráficas para cada una de las variables canónicas:





El cuadrado de los coeficientes de estructura representa la proporción de variabilidad linealmente compartida por una variable observada con el conjunto de las variables canónicas.

Descomposición de la variabilidad en las variables canónicas:

Conjunto X:

	CV 1	CV 2	CV 3
locus	0.8172997408	0.1518569	0.03084332
concepto	0.0004344419	0.5023105	0.49725509
motivación	0.3216603112	0.1231225	0.55521716

Conjunto Y:

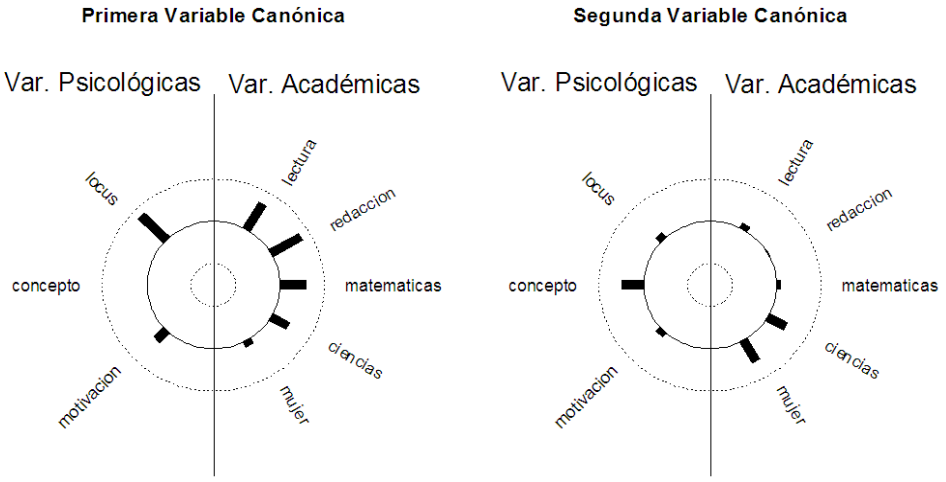
	CV 1	CV 2	CV 3
lectura	0.7063528	0.128755659	0.01832326
redacción	0.7683275	0.004203801	0.06480123
matemáticas	0.5836170	0.088773522	0.02183337
ciencias	0.4335089	0.458054967	0.05306348
mujer	0.1325781	0.569916524	0.29528740

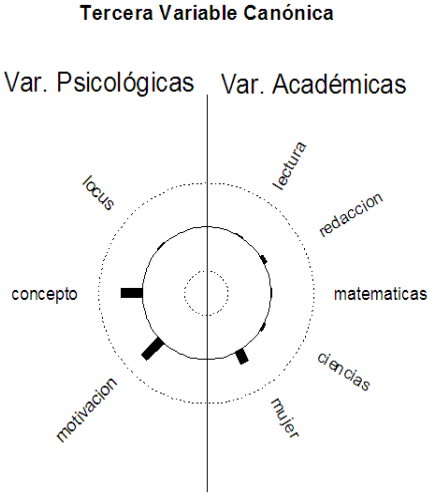
Así, se tiene que para el conjunto $\boldsymbol{X}$, la variable de locus comparte el 81.72 % de la variabilidad con la primera variable canónica, el 15.15 % con la segunda y 3.1 % con la tercera; la variable concepto no comparte variabilidad lineal con la primera variable canónica, sin embargo comparte 50.23 % con la segunda y 49.72 % con la tercera; por último la variable motivación comparte el 32.16 % con la primera variable canónica, 12.31 % con la segunda y 55.52 % con la tercera.

Para el conjunto $\boldsymbol{Y}$ se tiene que la variable lectura comparte 70.63 % de su variabilidad con la primera variable canónica, 12.87 % con la segunda y 1.8 % con la tercera; la variable redacción comparte 76.83 % con la primera variable, 0.42 % con la segunda y 6.48 % con la tercera; la variable matemáticas comparte 58.36 % de su variabilidad con la primera, 8.87 % con la segunda y 2.18 % con la

tercera; la variable ciencias comparte 43.35 % de su variabilidad con la primera variable canónica, 45.81 % y 5.31 % con la segunda y tercera variable canónica respectivamente y por último la variable género comparte 13.25 %, 56.69 % y 29.52 % con cada una de las variables canónicas.

Se presentan las siguientes gráficas para tener una apreciación visual de la variabilidad compartida por las variables observadas con las variables canónicas:





Dados los coeficientes anteriores, se obtiene el valor del coeficiente de comunalidad. Este valor indica la proporción de la variabilidad de cada variable que es reproducible a partir de los resultados de los coeficientes de estructura.

Coeficientes de comunalidad (Proporción del total de varianza explicada por cada variable dentro de cada conjunto):

Conjunto X:

locus	concepto	motivación
1	1	1

Conjunto Y:

lectura	redacción	matemáticas	ciencias	mujer
0.8534317	0.8373326	0.6942239	0.9446273	0.9977820

De los valores de los coeficientes de comunalidad, se interpreta que la variabilidad del conjunto $\boldsymbol{X}$ está representada en su totalidad por las tres variables canónicas, por otro lado, el porcentaje de la variabilidad representada por el conjunto de variables $\boldsymbol{Y}$, donde la proporción mínima representada es para la variable de evaluación en matemáticas con un 69.42 % y la mayor es para la variable de evaluación en ciencias con un 94.46 %.

Para la variable género, se reproduce el 99.77 % de su variabilidad; por lo tanto la construcción de las variables canónicas para estos dos conjuntos reproduce un gran porcentaje de la proporción de variabilidad para las variables.

Otro coeficiente importante para la interpretación del $\boldsymbol{AC}$ es el coeficiente de adecuación que es el promedio de todos los cuadrados de los coeficientes de estructura para las variables de cada conjunto.

Coeficientes de adecuación (proporción del total de variabilidad explicada por cada variable canónica para cada conjunto):

Conjunto X:

CV 1	CV 2	CV 3
0.3797982	0.2590966	0.3611052

Conjunto Y:

CV 1	CV 2	CV 3
0.52487685	0.24994089	0.09066175

Estos coeficientes indican qué tan adecuada, en promedio, es la representación de la variabilidad de un conjunto de variables canónicas con respecto al total de la variabilidad de los conjuntos originales; dado que es el promedio de cuadrados, los valores de los coeficientes son positivos.

En promedio la variabilidad que representa la primera variable canónica para el conjunto $\boldsymbol{X}$ se adecúa 37.97 % y para el conjunto $\boldsymbol{Y}$ se adecúa 52.48 %, la variabilidad que representa la segunda variable canónica se adecúa 25.91 % y 24.99 % para el conjunto $\boldsymbol{X}$ y para el conjunto $\boldsymbol{Y}$ respectivamente, por último la variabilidad que representa la tercera variable canónica es de 36.11 % y 9.06 % para el conjunto $\boldsymbol{X}$ y para el conjunto $\boldsymbol{Y}$ respectivamente. De acuerdo a la derivación de las variables canónicas, la primera pareja de variables canónicas debe ser la que tenga el mayor coeficiente de adecuación para las variables observadas, la segunda pareja, la que tenga el segundo mayor coeficiente y así sucesivamente, por lo tanto, se tiene que las primeras dos parejas de variables canónicas son las que adecúan mejor la representación de la variabilidad de éstas con respecto a las variables observadas de cada conjunto.

Coeficiente de Redundancia (Proporción de la variabilidad total explicada por cada variable canónica entre conjuntos):

X | Y :

CV 1	CV 2	CV 3
0.081799366	0.007270074	0.003905049

Y | X :

CV 1	CV 2	CV 3
0.1130458178	0.0070131703	0.0009804306

El valor del coeficiente de redundancia representa la proporción de la variabilidad total explicada para cada conjunto dada la combinación lineal de la variable canónica del conjunto que se ha fijado, así se tiene que el coeficiente de redundancia para el conjunto $\boldsymbol{X}$ dado $\boldsymbol{Y}$, es la proporción de variabilidad representada por la primera variable canónica que es 8.17 %, 0.72 % y 0.39 % para la segunda y tercera variable canónica respectivamente. Por otra parte, el coeficiente para el conjunto $\boldsymbol{Y}$ dado $\boldsymbol{X}$ es la proporción de la variabilidad representada por la primera variable canónica que es de 11.3 %, 0.7 % por la segunda y 0.09 % por la tercera variable canónica. Así, se tiene que la primera variable canónica para el conjunto $\boldsymbol{Y}$ dado $\boldsymbol{X}$ es la que tiene el mayor coeficiente de redundancia y esto significa que esta relación es la que está mejor representada por esta variable canónica.

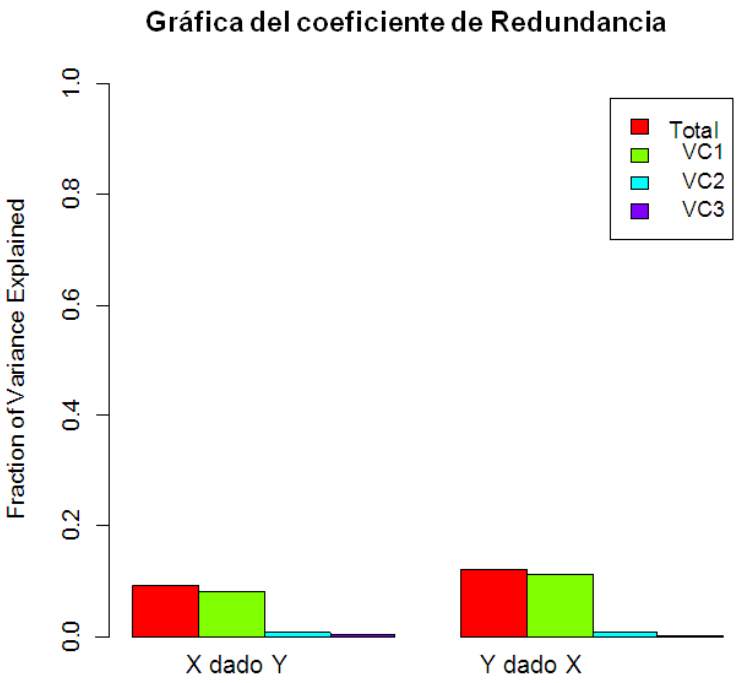
El coeficiente de redundancia agregada, representa el total de la variabilidad de un conjunto que se explica con el otro conjunto, considerando todo el conjunto de las variables canónicas, o bien, la suma de cada proporción de variabilidad representada por cada variable canónica para cada conjunto:

Coeficiente de redundancia agregada (Total de la variabilidad explicada por todas las variables canónicas entre conjuntos):

$$\begin{aligned} \mathrm{X} \mid Y \quad & 0.09297449 \\ \mathrm{Y} \mid X \quad & 0.1210394 \end{aligned}$$

Entonces la proporción de la variabilidad representada por las variables canónicas para el conjunto $\boldsymbol{Y}$ dado $\boldsymbol{X}$ es la que tiene mayor porcentaje, 12.1 %, entonces esta relación es la que mejor se puede representar con las variables canónicas.

Representación gráfica de los coeficientes de adecuación y adecuación total.



Una vez que se han encontrado las variables canónicas y sus coeficientes, se tiene que probar si existe al menos una correlación estadísticamente significativa entre las variables, para ello se postula la hipótesis:

$\boldsymbol{H_o}$: Todas las correlaciones canónicas $\rho_1, \rho_2, \text{y}\rho_3$, son cero, es decir

$$\rho_1 = \rho_2 = \rho_3 = 0$$

contra la hipótesis alternativa:

$\boldsymbol{H_a}$: Al menos una correlación es significativamente distinta de cero
$\rho_i \neq 0,\ p.a. \quad i = 1, 2, 3.$

Se contrasta la hipótesis con la prueba de Wilks a un nivel de significancia $\alpha = 0.05$ y se obtiene la estadística de prueba con la aproximación a la distribución ji-cuadrada dada por Bartlett

Prueba de Bartlett con la aproximación de la Ji-Cuadrada:

	rho^2	Chisq	df	Pr(>X)	
CV 1	0.215376	167.580079	15	<2.2e-16	***
CV2	0.028059	23.383800	8	0.002905	**
CV3	0.010814	6.464032	3	0.091092	.

Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Regla de decisión:

Rechazar $\boldsymbol{H_o}$ al nivel de significancia $\alpha = 0.05$, si la estadística es mayor que el cuantil $(1 - \alpha)$ de una distribución ji-cuadrada con los grados de libertad correspondientes.

Para fines prácticos, se decide comparar el valor del $p - value$ para cada estadística con el nivel de significancia, es decir, se rechaza la hipótesis nula si $p - value < \alpha$.

A partir de los cuadrados de los coeficientes de las correlaciones canónicas se obtiene la estadística de prueba; para cada variable canónica se tiene su estadística

y los grados de libertad de la distribución a la que se aproxima dicha estadística, por lo tanto se tiene que para las dos primeras variables se rechaza la hipótesis nula a un nivel de significancia del 0.05, esto es, la primera y la segunda pareja de variables canónicas tienen correlación significativamente distinta de cero, mientras que la tercera pareja de variables canónicas no tiene correlación significativamente distinta de cero.

Es necesario conocer el número de correlaciones canónicas significativamente distintas de cero, la hipótesis postulada es la misma para la prueba anterior, sin embargo el interés es rechazar la hipótesis nula y bajo la hipótesis alternativa encontrar el número mínimo de variables canónicas significativas con las que se puede representar la relación entre los conjuntos de variables observadas. La prueba es llamada prueba de Rao en relación a la lambda de Wilks y la estadística de prueba tiene una distribución aproximada a una **F**.

Se postula la hipótesis nula:

$\boldsymbol{H_o}$: Todas las correlaciones canónicas $\rho_1, \rho_2, \text{y} \rho_3$, son cero, es decir

$$\rho_1 = \rho_2 = \rho_3 = 0$$

contra la hipótesis alternativa:

$\boldsymbol{H_a}$: Al menos una correlación es significativamente distinta de cero
$\rho_i \neq 0, \ p.a. \quad s = 1, 2, 3.$

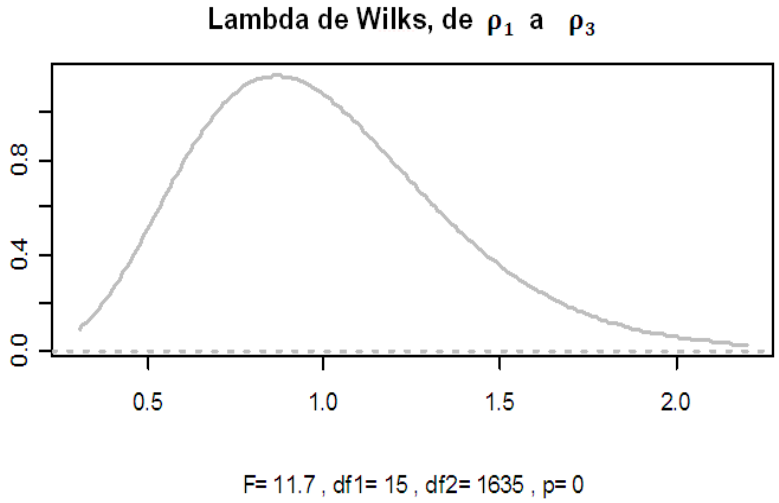
Se contrasta la hipótesis con la prueba de Rao a un nivel de significancia $\alpha = 0.05$ y se obtienen los siguientes valores de la estadística de prueba:

Lambda de Wilks usando una aproximación a una distribución F (Rao's F)

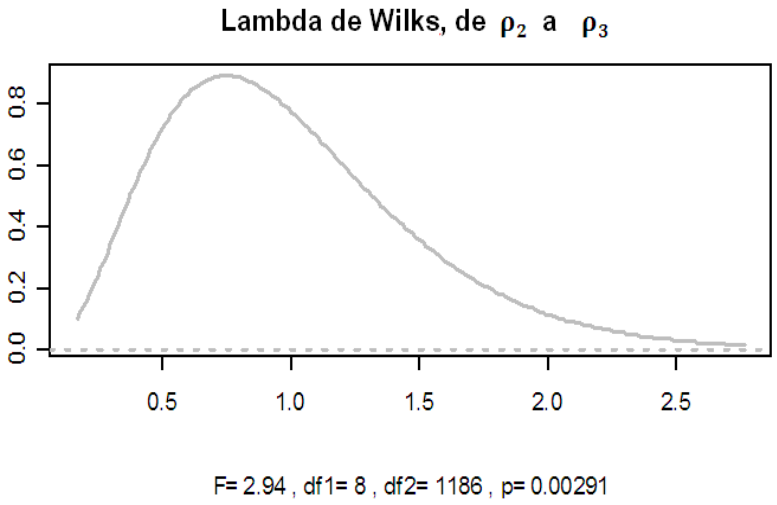
	statapprox	F	df1	df2	p.value
1 to 3:	0.7543611	11.715733	15	1634.653	0.000000000
2 to 3:	0.9614300	2.944459	8	1186.000	0.002905057
3 to 3:	0.9891858	2.164612	3	594.000	0.091092170

Análogo al resultado de la prueba anterior, se tiene que las dos primeras parejas de variables canónicas son las que tienen una correlación significativamente distinta de cero a un nivel de significancia del 0.05, por lo tanto el número de parejas de variables canónicas que se requieren para explicar la relación entre las variables observadas es dos.

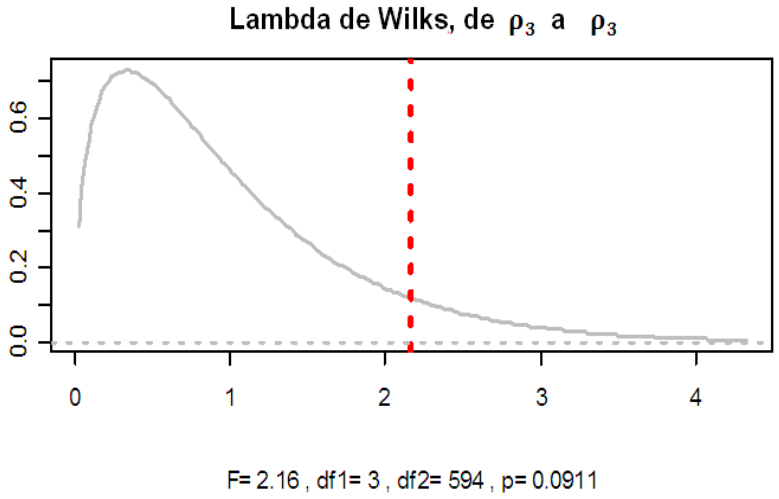
Gráfica 1. Aportación que tiene una variable para la estadística de prueba Lambda de Wilks



Gráfica 2. Aportación que tienen dos variable para la estadística de prueba Lambda de Wilks



Gráfica 3. Aportación que tienen tres variable para la estadística de prueba Lambda de Wilks



En las 3 gráficas anteriores, se muestra la aportación que tiene el aumento de una variable para la estadística de prueba, y se tiene que la tercera variable no representa un aumento significativo ya que la línea vertical de color rojo que se muestra en la tercera gráfica representa el valor de la estadística de prueba que

se obtuvo, así el número de variables canónicas que estén por el lado derecho de ésta, no son significativas, por lo tanto sólo las dos primeras variables canónicas tienen correlación significativa y por lo tanto son las necesarias para describir la relación entre los conjuntos de variables observadas.

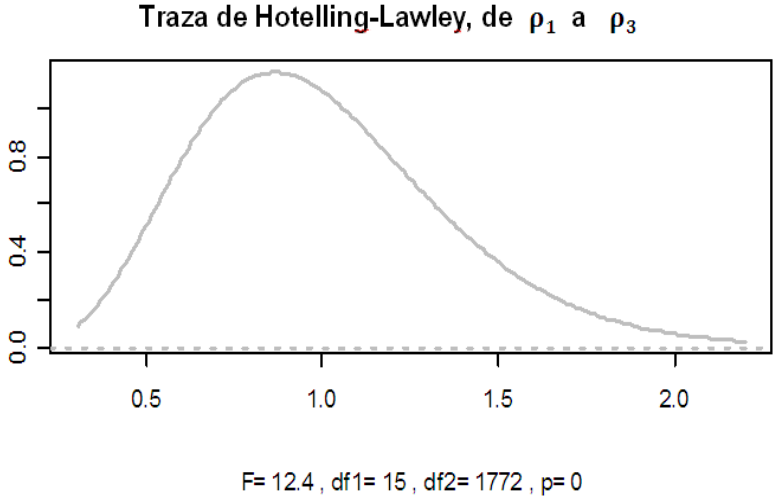
Además de estas dos pruebas estadísticas para contrastar la hipótesis, existen otras dos pruebas análogas a la anterior que también se basan en una aproximación de la estadística de prueba a una distribución F; la primera fue desarrollada por Hotelling y Lawley, y la segunda fue desarrollada por Pillai y Bartlett.

Se presentan las estadísticas de ambas pruebas:

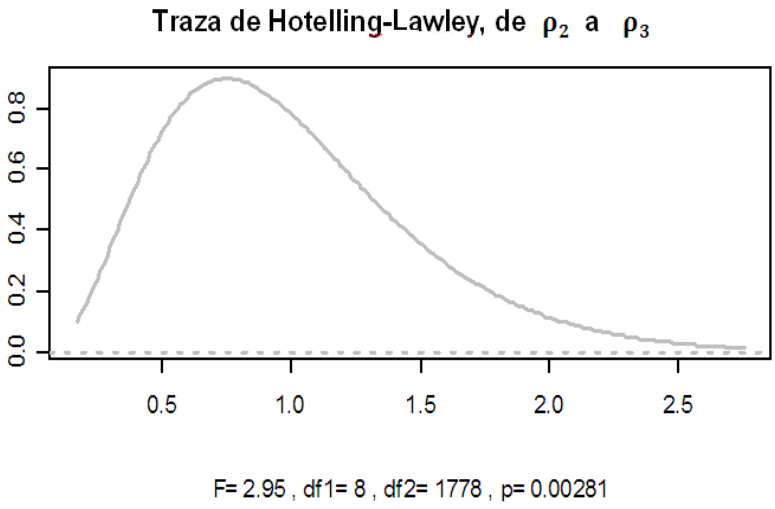
Traza de Hotelling-Lawley usando una aproximación a una distribución F

	statapprox	F	df1	df2	p.value
1 to 3:	0.31429738	12.376332	15	1772	0.000000000
2 to 3:	0.03980175	2.948647	8	1778	0.002806615
3 to 3:	0.01093238	2.167041	3	1784	0.090013166

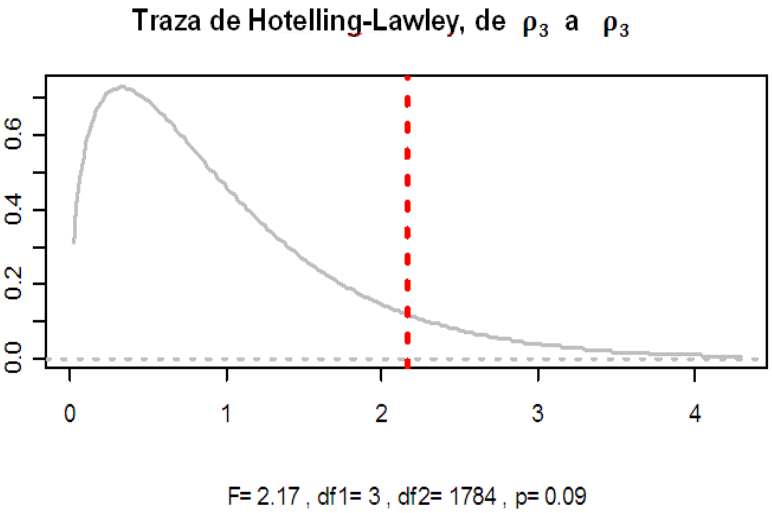
Gráfica 4. Aportación que tiene una variable para la estadística de prueba Hotelling-Lawley



Gráfica 5. Aportación que tienen dos variable para la estadística de prueba Hotelling-Lawley



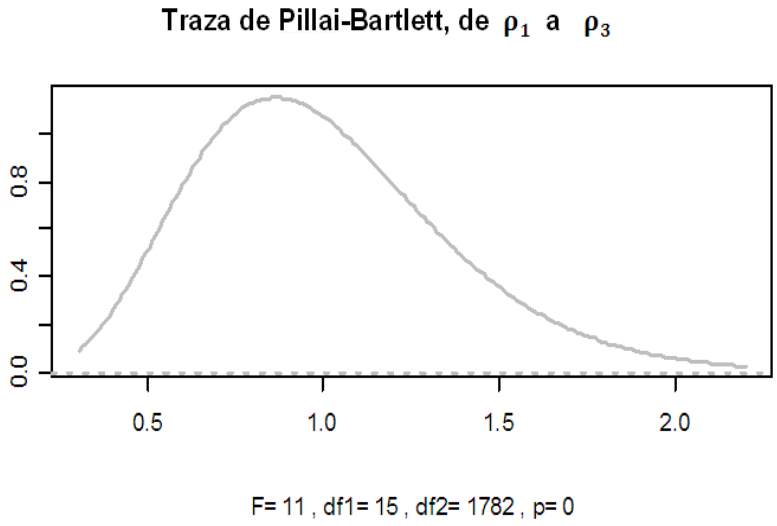
Gráfica 6. Aportación que tienen tres variable para la estadística de prueba Hotelling-Lawley



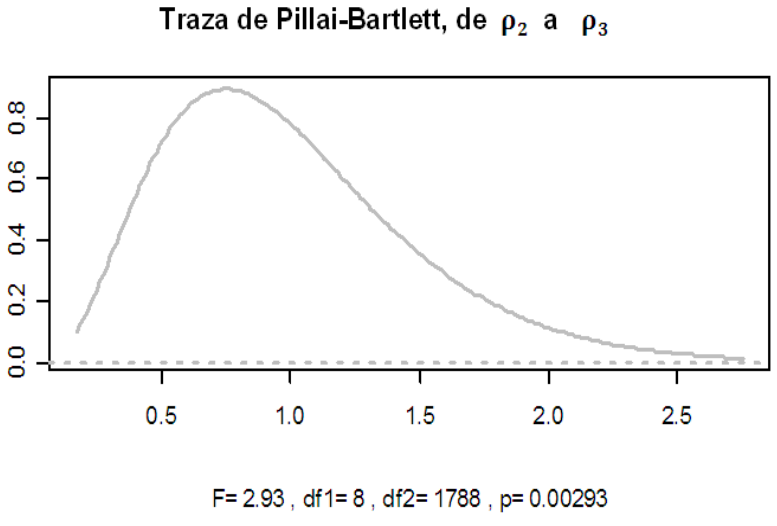
Traza Pillai-Bartlett usando una aproximación a una distribución F:

	statapprox	F	df1	df2	p.value
1 to 3:	0.25424936	11.000571	15	1782	0.000000000
2 to 3:	0.03887347	2.934093	8	1788	0.002932566
3 to 3:	0.01081416	2.163421	3	1794	0.090440464

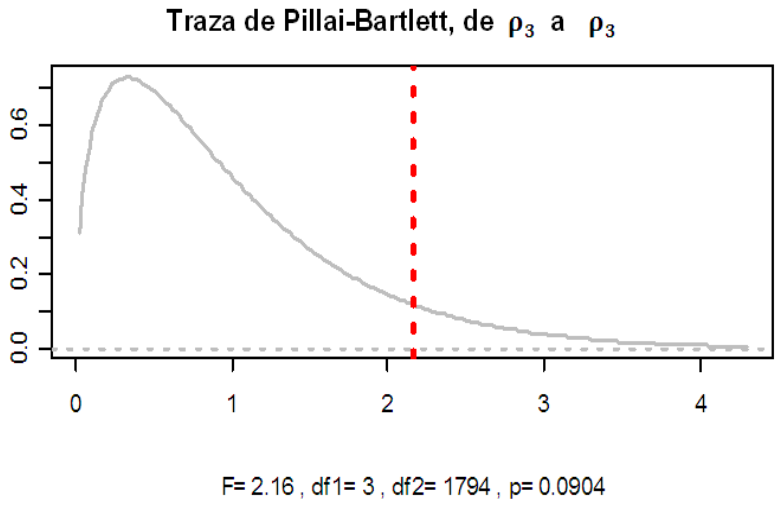
Gráfica 7. Aportación que tiene una variable para la estadística de prueba Pillai-Bartlett



Gráfica 8. Aportación que tienen dos variable para la estadística de prueba Pillai-Bartlett



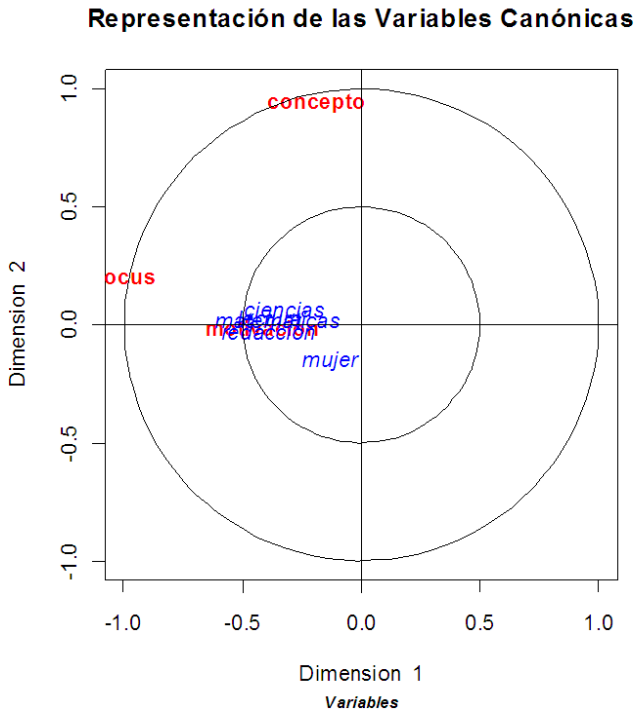
Gráfica 9. Aportación que tienen tres variable para la estadística de prueba Pillai-Bartlett



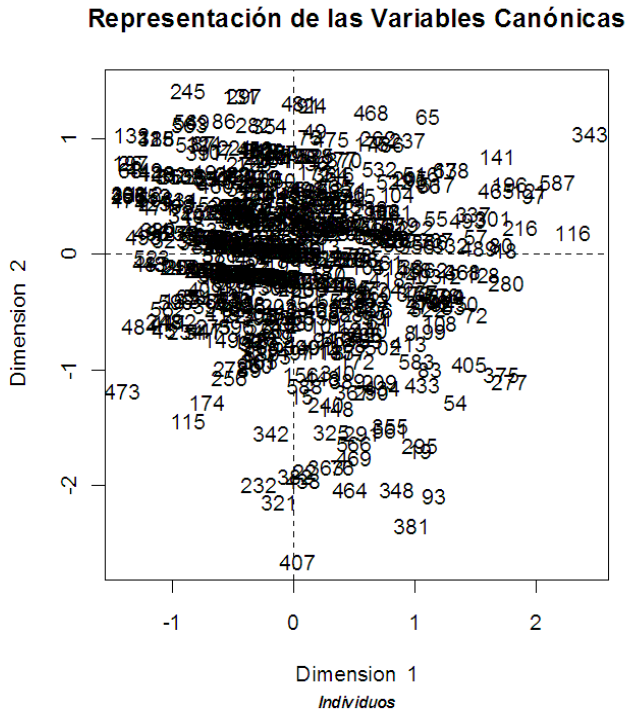
Se tiene para ambas pruebas la misma conclusión, las primeras dos parejas de variables canónicas tienen correlación significativamente distinta de cero a un nivel de significancia α del 0.05 y las gráficas muestran los mismos resultados para la prueba de Rao; en ninguna prueba se tiene que la tercera variable canónica resulte significativa, por lo tanto se tiene que las dos primeras parejas de variables canónicas son significativas para el análisis.

Al tener identificadas a las variables canónicas que son representativas, se presentan las gráficas que muestra la relación entre y dentro de las variables del conjunto $\boldsymbol{X}$ y las variables del conjunto $\boldsymbol{Y}$ con cada una de las variables canónicas.

La siguiente gráfica muestra las dos variables canónicas estadísticamente significativas conjuntamente con las variables observadas, dicha representación ejemplifica que las variables académicas están cercanas entre ellas debido a la alta correlación que existe entre ellas, mientras que las variables psicológicas al no estar correlacionadas, su separación es notable; la variable de género no está tan alejada de las variables académicas, pero si de la variable concepto.



En la gráfica de abajo se muestra la dispersión de los individuos en estudio, conjuntamente con las dos variables canónicas, se observa que la mayoría de los individuos están dentro de los rangos (-2,2) de las variables canónicas, lo cual es comprobable en la tabla de los coeficientes canónicos. Al tener un rango relativamente corto, se dice que se tiene una mejor interpretación de las variables canónicas.



En conclusión, las variables psicológicas no tienen relación una con la otra, por ejemplo, un puntaje alto de motivación no influye en el autoconcepto o el locus del estudiante, mientras que en las variables académicas se tiene que los estudiantes que obtienen una calificación alta en lectura, son estudiantes que también obtienen una alta calificación en redacción, matemáticas y ciencias; este hecho es de esperarse, ya que un estudiante que tiene un régimen de lectura adecuado es capaz de comprender más fácil y rápido cualquier tema o problema que se le presente; también se tiene que los estudiantes con calificación alta en

matemáticas, también obtienen altas calificaciones en ciencias y aquellos que tienen una alta calificación en redacción, tienen alta calificación en matemáticas, por lo tanto, sí existe relación entre las variables académicas, es decir, el desempeño académico de un estudiante que tenga altos puntajes en una materia también tendrá altos puntajes en otras materias.

Se demostró que la variable género no influye con los resultados de las pruebas, ya que la proporción de mujeres y hombres de la muestra esta equilibrada y estadísticamente no tiene relación con las demás varibles, o bien, se puede decir que el puntaje obtenido en las pruebas no depende del sexo del alumno.

Con la matriz $\boldsymbol{R}$ se observa que no hay una alta correlación entre las variables psicológicas y académicas, pero con las variables canónicas y los coeficientes de estructura canónicos, se demuestra que influyen las variables psicológicas locus y motivación con las variables académicas, lo cual es de suponerse ya que si un alumno tiene una gran motivación para seguir estudiando o un objetivo claro, el desempeño en las variables académicas será mejor.

Por último, se realizaron pruebas estadísticas para demostrar que la primera y segunda variable canónica son estadísticamente significativas, es decir, que estas dos variables son necesarias e idóneas para describir adecuadamente la relación entre los conjuntos de variables.

Conclusiones

Dado los desarrollos y resultados del presente trabajo, se ha formalizado y explicado el **AC** y todo lo que conlleva, así como también saber si los resultados son estadísticamente significativos mediante diferentes pruebas estadísticas, por ejemplo: la prueba de Wilks usando la aproximación de Bartlett, la estadística de Rao en relación a la lambda de Wilks con base en una aproximación a una distribución F y el análisis de Variabilidad bajo la Hipótesis Alternativa, cabe mencionar que las pruebas anteriores fueron desarrolladas y explicadas, para el contexto del problema dado saber cuál es la más adecuada y por ende la más potente dependiendo del objetivo del problema. También se desarrollaron las regiones de confianza y sus interpretaciones.

Se propuso un número de recomendaciones para llevar acabo el **AC** y así poder identificar los posibles errores, causa por las que ocurren y su solución; todo esto con el fin de realizar adecuadamente el análisis.

Se justificó porque el **AC** es una generalización de las técnicas multivariadas, como: el análisis de varianza, análisis discriminante, análisis de tablas de contingencia, entre otras; se mostró que una de las principales causas por las que el **AC** es la técnica general es debido a que tiene el menor número de restricciones en cuanto a supuestos, a su vez, los conjuntos de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$ pueden ser vectores aleatorios, uno de ellos ser aleatorio, tener variables dummy o variables no aleatorias al mismo tiempo. Al realizar la justificación del porque el **AC** es

la técnica multivariada general, también se expusieron las principales ventajas que posee el **AC**, por ejemplo: permite extraer la máxima información posible de los conjuntos de datos, realizando trasformaciones de los conjuntos de variables para obtener una manera simplificada y adecuada de representar los datos, es decir, resume relaciones complejas y proporciona un método útil de reducción de dimensión de un problema.

Todos los resultados se desarrollaron en dos conjuntos de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$, pero también se presentó brevemente un *Análisis Canónico Generalizado*, el cual consiste en hacer ver al lector que todos los desarrollos obtenidos en dos conjuntos de variables se puede generalizar para más conjuntos de variables en términos de regresiones múltiples, es decir, aplicar adecuadamente las técnicas estadísticas.

Sin duda, el **AC** requiere considerablemente más trabajo computacional, pero con la evolución de las computadoras y desarrollo de software, derivar un **AC** puede no ser un obstáculo, sin embargo, se ha propuesto y presentado una forma eficiente para llevar a cabo dicho análisis en el paquete estadístico **R** y así obtener los resultados más fácilmente con las interpretaciones adecuadas de los coeficientes.

En conclusión se tiene que el **AC** es la técnica multivariada general y completa que se pueda tener para encontrar las relaciones existentes entre dos o más conjuntos de variables, al no tener gran cantidad de restricciones es eficiente en cualquier cantidad de problemas y tipos de conjuntos, es fácilmente interpretable siguiendo la descripción de los coeficientes que se obtienen y una técnica segura y confiable.

Bibliografía

[1] ANDERSON, T. W. (1984). "Estimating linear restrictions on regression coefficients for multivariate normal distributions". Ann. Math. Statist. 22, 327-51.

[2] AFIFI, A, CLARK, V AND MAY, S. 2004. "Computer-Aided Multivariate Analysis". 4th ed. Boca Raton, Fl: Chapman Hall/CRC.

[3] CAMPBELL, N. A. (1984). "Canonical variate analysis", a general model formulation. Aust. J. Statist 26,86-9.

[4] CAMPBELL, N. A. and ATCHLEY William R. (1981) "The geometry of canonical variate analysis". Syst. Zool. 30(3), pp. 268-280.

[5] CARROLL, J . D., "Generalization of canonical correlation analysis to three or more sets of variables". Proceedings of the American Psychological Association, 1968,76, 227-228.

[6] COOLEY, W. W., LOHNES, P. R., "Multivariate data analysis.", New York: Wiley, 1971.

[7] DARLINGTON, R. B.; WEINBERG, S. L.; WALBERG H.J. "Canonical Variate Analysis and Related Techniques", Review of Educational Research, Vol.43, No. 4. (Autumn, 1973), pp. 433-454.

[8] DEGROOT, M. H.; LI C. C., "Correlations between Similar Sets of Measurements", Biometrics, Vol. 22, No. 4. (Dec., 1966), pp. 781-790.

[9] GRANÉ, AUREA. "Análisis Canónico de Poblaciones (MANOVA)", Departamento de Estadística Universidad Carlos III de Madrid, pp. 1-47.

[10] HORST, P. "Generalized canonical correlations and their application to experimental data". Journal of Clinical Psychology, 1961, 17, 331-347.

[11] KRZANOWSKI, W. J. (1979). "Between-groups comparison of principal components". J. Am. Statist. Assoc74, 703-7.

[12] KRZANOWSKI, W. J. (1989) "On confidence regions in canonical variate analysis". Department of Applied Statistics, University of Reading, Whiteknights. Biometrika (1989), 76, 1, pp. 107-16. Reading, RG62AN, U.K.

[13] KSHIRSAGAR, Anant M. (1972) "Multivariate Analysis" . Statistics: textbooks and monographs Vol. 2, New York : Marcel Dekker, pp.247-258.

[14] HARDOON R. DAVID , SANDOR S. AND SHAWE-T.J.(2003) "Canonical correlation analysis; And overview with application to learning Methods", Department of Computer ScienceEgham, England.

[15] LEVINE, Mark S. (1977) "Canonical analysis and factor comparison". Sage University Paper series on Quantitative Applications in the Social Sciences Series Number 07-006. Beverly Hills and London: SagePublications.

[16] MARDIA, K. V.; KENT, J. T. y BIBBY, J. M. (1979): "Multivariate Analysis".Academic Press, London.

[17] MCKEON, J. J., "Canonical analysis: some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory". Psychometric Monographs, 1966, No. 13.

[18] MORRISON, D. F. (1978): "Multivariate Statistical Methods". McGraw-Hill, London.

[19] PILLAI, K. C. W. (1956) "On the distribution of the largest or the smallest root of a matrix in multivariate analysis". Biometrika, 43 122–127.

[20] SCHATZOFF, M., "Exact Distributions of Wilks's Likelihood Ratio Criterion", Biometrika, Vol. 53, No. 3/4. (Dec., 1966), pp. 347-358.

[21] THOMPSON, Bruce. (1984) "Canonical correlation analysis: Uses and interpretation". Sage University Paper series on Quantitative Applications

in the Social Sciences Series Number 07-047. Beverly Hills and London: SagePublications.

[22] VICENTE VILLARDÓN, José L. "Análisis de un Estudio", Departamento de Estadística. Universidad de Salamanca, pp. 55-83.

[23] VICENTE-VILLARDÓN. J L., GALINDO-VILLARDÓN Mª. P., AMARO I. R.,"Contribuciones al manova-biplot: regiones de confianza alternativas", Departamento de Estadística. Salamanca, España.

Fuentes

http://www.ats.ucla.edu/stat/spss/dae/canonical.htm

http://biplot.usal.es/DOCTORADO/3CICLO/BIENIO-04-06/Canonico/ACPO.pdf

Apéndice

Ap. 1) $\boldsymbol{A}$ es una matriz ortogonal si $\boldsymbol{AA^t = A^tA = I}$

Ap. 2) $\boldsymbol{A_{n\times m}}$ es una matriz cuadrada si $n = m$, es decir, el número de filas es igual al número de columnas.

Ap. 3) CASIQUE G. A., CHANEZ F. J.L., **"El locus de control"**, Revista Panorama Administrativo, Año 1 No. 2 enero-junio 2007, Instituto Tecnológico de Celaya.

Ap. 4) WASHINGTON SANDOVAL ERAZO, **"La motivación"**, Postgrado-ESPE.

Ap. 5) GONZÁLEZ-PINEDA J A., NÚÑEZ P. J., GARCÍA M S. **"Autoconcepto, autoestima y aprendizaje escolar"**, Psicothema, 1997. Vol. 9, nº 2, pp. 271-289, Universidad de Oviedo.

Sintaxis de figuras

```
# Fig. 1.
x<-c(0,40)
y<-c(0,40)
plot(x,y,type="n",axes=FALSE,xlabel=,ylabel=,cex.lab=0.000001)
segments(5,5,35,5);segments(5,35,35,35);segments(5,20,35,20)
segments(5,5,5,35);segments(35,5,35,35);segments(20,5,20,35)
text(7,36.5, expression(X[1]),cex=.8)
text(10,36.5, expression(X[2]),cex=.8)
text(14,36.5, ". . . .",cex=1.5)
text(19,36.5, expression(X[p]),cex=.8)
text(22,36.5, expression(Y[1]),cex=.8)
```

```
text(25,36.5, expression(Y[2]),cex=.8)
text(29,36.5, ". . . .",cex=1.5)
text(34,36.5, expression(Y[q]),cex=.8)
text(2,33, expression(X[1]),cex=.8)
text(2,30, expression(X[2]),cex=.8)
text(2,27, ".",cex=1.5)
text(2,26, ".",cex=1.5)
text(2,25, ".",cex=1.5)
text(2,24, ".",cex=1.5)
text(2,21, expression(X[p]),cex=.8)
text(2,18, expression(Y[1]),cex=.8)
text(2,15, expression(Y[2]),cex=.8)
text(2,12, ".",cex=1.5)
text(2,11, ".",cex=1.5)
text(2,10, ".",cex=1.5)
text(2,9, ".",cex=1.5)
text(2,7, expression(Y[q]),cex=.8)
text(12.5,27.5, expression(R[XX]))
text(12.5,12.5, expression(R[YX]))
text(27.5,12.5, expression(R[YY]))
text(27.5,27.5, expression(R[XY]))

# Fig. 2
x=runif(80,30,70)
y=runif(80,30,70)
plot(x,y,xlim=c(-20,100),ylim=c(-20,100),cex=.7,pch=3)
arrows(-100,0,100,0,length=.1)
arrows(0,-100,0,100,length=.1)
text(100,7,expression(X[1]))
text(7,100,expression(X[2]))
arrows(25,25,80,80,length=.1)
arrows(80,25,25,80,length=.1)
text(80,70,expression(Y[1]))
text(25,85,expression(Y[2]))

# Fig. 3.
library(ellipse)
plot(ellipse((.8),centre = c(10,10),scale = c(3.5, 3.5),
which = c(20, 20),
level = 0.95),xlim=c(0,20),ylim=c(0,20),type = 'l',axes = FALSE)
arrows(0,0,0,20,length=.1);arrows(0,0,20,0,length=.1)
arrows(1,1,20,20,length=.1,col="blue");arrows(15,5,5,15,length=.1,col="blue")
segments(3,10,18,10); segments(10,3,10,18)
```

```
text(-.4,-.4,"0"); text(1,20,expression(X[2]));text(20,1,expression(X[1]))
text(5,14,expression(Y[2])); text(20,19,expression(Y[1]))
text(10,1,bquote(bar(X[1])),cex=.7);text(1,10,bquote(bar(X[2])),cex=.7)
text(13,11,bquote(theta==cos -1*(u[11])),cex =.8)
points(13,15,pch=8,col=2)
segments(13,15,14,14,col=2); segments(13,15,9,11,col=2)
text(c(13,13.5,14,14.5,15),16,expression("(",X["1 i "],","
,X["2 i "],")"),cex=.7);
text(8,11,expression(Y["2 i "]),cex=.7);text(15,14,expression
(Y["1 i "]),cex=.7)
# Referencia de la librería ellipse
# -Bates, D.M. and Watts, D.G. (1988). Nonlinear Regression Analysis
# and its Applications.Wiley.
# -Murdoch, D.J. and Chow, E.D. (1996).A graphical display of large
# correlation matrices. The American Statistician 50, 178-180.

# Fig. 4.
x<-c(0,30)
y<-c(0,30)
plot(x,y,type="n",axes=FALSE,xlabel=,ylabel=,cex.lab=0.000001)
segments(5,5,30,5);segments(5,25,30,25)
segments(5,5,5,25);segments(30,5,30,25);segments(17.5,5,17.5,25)
text(6,26, expression(X[1]))
text(9,26, expression(X[2]))
text(12,26, "...")
text(15,26, expression(X[p]))
text(19,26, expression(Y[1]))
text(22,26, expression(Y[2]))
text(25,26, "...")
text(28,26, expression(Y[q]))
text(2,22, Çaso 1")
text(2,18, Çaso 2")
text(2,15, Çaso 3")
text(2,12, ".")
text(2,11, ".")
text(2,10, ".")
text(2,7, Çaso N")
text(10,15, Çonjunto X")
text(23,15, Çonjunto Y")

# Fig. 5.
symbols(x = c(10,20,21.5), y = c(20,20,20), circles =
c(140,140,140),main = )
```

```
text(10,22, expression(Y[1]),cex=1);text(10,21, expression(Y[2]),cex=1)
text(10,20, ".",cex=1.5);text(10,19.5, ".",cex=1.5)
text(10,19, ".",cex=1.5);text(10,18, expression(Y[q]),cex=1)
text(20,20, expression(bold(Y)) ,cex=1.5)
text(21.5,20, expression(bold(X)) ,cex=1.5)
arrows(10.5,22,19,21,length=.1)
arrows(10.5,21,19,20,length=.1)
arrows(10.5,18,19,19,length=.1)
text(20.75,20, expression(bar(t 2) == sum(frac(r[i] 2, q),
i==1, p)),cex = 1)
text(10,27, "Variabilidad que aporta cada",cex=1.3)
text(10,26, "variable del conjunto Y",cex=1.3)
text(14.5,22, expression(r[1] 2));text(14.5,21, expression(r[2] 2))
text(14.5,18, expression(r[q] 2))

# Fig. 6.
n<-3
a<-rep(1,n)
names(a)<-c("X1","X2","X3")
pie(a,main=.Análisis canónico parcial", radius=1.2,col="white",
clockwise=T,label=, init.angle=210,border=NA)
text(0,-1,expression(x[3]), cex=1.6)
text(sqrt(3)/2,1/2,expression(x[2]), cex=1.6)
text(-(sqrt(3)/2),1/2,expression(x[1]), cex=1.6)
arrows(0.05,-.9,(sqrt(3)/2)-.05 ,(1/2)-.05, col= "blue", lwd=2)
arrows(-0.05,-.9,-(sqrt(3)/2)+.05, 1/2-.05, col= "blue", lwd=2)

# Fig. 7.
m<-2
n<-2*m+1;n
a<-rep(1,n)
pie(a,main=.Análisiscanónico biparcial", radius=1.2,col="white",
clockwise=T,init.angle=90,border=NA,label=)
arrows(0.05,-.9,0.95-.005, -0.31-.1, col= "blue", lwd = 2)
arrows(0.025,-.9,0.58+.05 ,0.81-.05, col= "blue", lwd = 2)
arrows(-0.025,-.9,-0.58-.05 ,0.81-.05, col= "blue", lwd = 2)
arrows(-0.05,-.9,-0.95+.005, -0.31-.1, col= "blue", lwd = 2)
arrows(0.95-.0025, -0.31-.05,0.58+.1 ,0.81-.05, col= "blue",
lwd = 2)
arrows(-0.95+.0025, -0.31-.05,-0.58-.1,0.81-.05, col= "blue",
lwd = 2)
text(0,-1,expression(x[3]),cex=1.6)
text(0.66,0.83,expression(x[2]),cex=1.6)
```

```
text(-0.66 ,0.83,expression(x[1]),cex=1.6)
text(-1.05, -0.33,expression(x[4]),cex=1.6)
text(1.05, -0.33,expression(x[5]),cex=1.6)

# Fig. 8.
m<-3
n<-2*m+1;n
a<-rep(1,n)
pie(a,main=.Análisis canónico C(7)", radius=1.2,col="white",
clockwise=T,init.angle=90,border=NA,label=)
arrows(0.1,-.95,.83,-.75, col= "blue", lwd = 2)
arrows(0.06,-.95,.97,.23, col= "blue", lwd = 2)
arrows(0.03,-.95,.48,.9, col= "blue", lwd = 2)
arrows(.99,.24,.5,.92, col= "blue", lwd = 2)
arrows(.84,-.74,.49,.9, col= "blue", lwd = 2)
arrows(.85,-.74,.99,.21, col= "blue", lwd = 2)
arrows(-0.1,-.95,-.83,-.75, col= "blue", lwd = 2)
arrows(-0.06,-.95,-.97,.23, col= "blue", lwd = 2)
arrows(-0.03,-.95,-.48,.9, col= "blue", lwd = 2)
arrows(-.99,.24,-.5,.92, col= "blue", lwd = 2)
arrows(-.84,-.74,-.49,.9, col= "blue", lwd = 2)
arrows(-.85,-.74,-.99,.21, col= "blue", lwd = 2)
text(0.01,-1.01,expression(x[3]),cex=1.6)
text(.9,-.8,expression(x[7]),cex=1.6)
text(1.05 ,.25,expression(x[5]),cex=1.6)
text(.5, 1,expression(x[2]),cex=1.6)
text(-.9,-.8,expression(x[6]),cex=1.6)
text(-1.05 ,.28,expression(x[4]),cex=1.6)
text(-.5, 1,expression(x[1]),cex=1.6)

# Fig. 9.
m<-9
n<-2*m+1;n
a<-rep(1,n)
pie(a,main=.Análisis canónico C(2n+1)", radius=1.1,col="white",
clockwise=T,init.angle=90),border=NA,label=)
pie(a,main=.Análisis canónico C(2n+1)", radius=1,col="white",
clockwise=T,init.angle=270,border=NA,label=)
a1<-0;a2<-(-.98);a3<-.32;a4<-(-.93);a5<-.61
a6<-(-.78);a7<-.83;a8<-(-.54);a9<-.96;a.<-(-.24)
b1<-.99;b2<-.08;b3<-.91;b4<-.4;b5<-.73
b6<-.67;b7<-.47;b8<-.87;b9<-.16;b.<-.98
arrows(a1,a2,a3,a4, col= "blue",lwd = 1, length=.1)
```

```
arrows(a3,a4,a5,a6, col= "blue",lwd = 1, length=.1)
arrows(a5,a6,a7,a8, col= "blue",lwd = 1, length=.1)
arrows(a7,a8,a9,a., col= "blue",lwd = 1, length=.1)
segments(a9,a.,b1,b2, col= "blue",lwd =1, length=.1, lty=3)
arrows(b1,b2,b3,b4, col= "blue",lwd = 1, length=.1)
arrows(b3,b4,b5,b6, col= "blue",lwd = 1, length=.1)
arrows(b5,b6,b7,b8, col= "blue",lwd = 1, length=.1)
arrows(b7,b8,b9,b., col= "blue",lwd = 1, length=.1)
arrows(a1,a2,a5,a6, col= "blue",lwd = 1, length=.1)
arrows(a1,a2,a7,a8, col= "blue",lwd = 1, length=.1)
arrows(a1,a2,a9,a., col= "blue",lwd = 1, length=.1)
arrows(a1,a2,b1,b2, col= "blue",lwd = 1, length=.1)
arrows(a1,a2,b3,b4, col= "blue",lwd = 1, length=.1)
arrows(a1,a2,b5,b6, col= "blue",lwd = 1, length=.1)
arrows(a1,a2,b7,b8, col= "blue",lwd = 1, length=.1)
arrows(a1,a2,b9,b., col= "blue",lwd = 1, length=.1)
arrows(a3,a4,a7,a8, col= "blue",lwd = 1, length=.1)
arrows(a3,a4,a9,a., col= "blue",lwd = 1, length=.1)
arrows(a3,a4,b1,b2, col= "blue",lwd = 1, length=.1)
arrows(a3,a4,b3,b4, col= "blue",lwd = 1, length=.1)
arrows(a3,a4,b5,b6, col= "blue",lwd = 1, length=.1)
arrows(a3,a4,b7,b8, col= "blue",lwd = 1, length=.1)
arrows(a3,a4,b9,b., col= "blue",lwd = 1, length=.1)
arrows(a5,a6,a9,a., col= "blue",lwd = 1, length=.1)
arrows(a5,a6,b1,b2, col= "blue",lwd = 1, length=.1)
arrows(a5,a6,b3,b4, col= "blue",lwd = 1, length=.1)
arrows(a5,a6,b5,b6, col= "blue",lwd = 1, length=.1)
arrows(a5,a6,b7,b8, col= "blue",lwd = 1, length=.1)
arrows(a5,a6,b9,b., col= "blue",lwd = 1, length=.1)
arrows(a7,a8,b1,b2, col= "blue",lwd = 1, length=.1)
arrows(a7,a8,b3,b4, col= "blue",lwd = 1, length=.1)
arrows(a7,a8,b5,b6, col= "blue",lwd = 1, length=.1)
arrows(a7,a8,b7,b8, col= "blue",lwd = 1, length=.1)
arrows(a7,a8,b9,b., col= "blue",lwd = 1, length=.1)
arrows(a9,a.,b3,b4, col= "blue",lwd = 1, length=.1)
arrows(a9,a.,b5,b6, col= "blue",lwd = 1, length=.1)
arrows(a9,a.,b7,b8, col= "blue",lwd = 1, length=.1)
arrows(a9,a.,b9,b., col= "blue",lwd = 1, length=.1)
arrows(b1,b2,b5,b6, col= "blue",lwd = 1, length=.1)
arrows(b1,b2,b7,b8, col= "blue",lwd = 1, length=.1)
arrows(b1,b2,b9,b., col= "blue",lwd = 1, length=.1)
arrows(b3,b4,b7,b8, col= "blue",lwd = 1, length=.1)
```

```
arrows(b3,b4,b9,b., col= "blue",lwd = 1, length=.1)
arrows(b5,b6,b9,b., col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-a3,a4, col= "blue",lwd = 1, length=.1)
arrows(-a3,a4,-a5,a6, col= "blue",lwd = 1, length=.1)
arrows(-a5,a6,-a7,a8, col= "blue",lwd = 1, length=.1)
arrows(-a7,a8,-a9,a., col= "blue",lwd = 1, length=.1)
segments(-a9,a.,-b1,b2, col= "blue",lwd = 1, length=.1, lty=3)
arrows(-b1,b2,-b3,b4, col= "blue",lwd = 1, length=.1)
arrows(-b3,b4,-b5,b6, col= "blue",lwd = 1, length=.1)
arrows(-b5,b6,-b7,b8, col= "blue",lwd = 1, length=.1)
arrows(-b7,b8,-b9,b., col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-a5,a6, col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-a7,a8, col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-a9,a., col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-b1,b2, col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-b3,b4, col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-b5,b6, col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-b7,b8, col= "blue",lwd = 1, length=.1)
arrows(-a1,a2,-b9,b., col= "blue",lwd = 1, length=.1)
arrows(-a3,a4,-a7,a8, col= "blue",lwd = 1, length=.1)
arrows(-a3,a4,-a9,a., col= "blue",lwd = 1, length=.1)
arrows(-a3,a4,-b1,b2, col= "blue",lwd = 1, length=.1)
arrows(-a3,a4,-b3,b4, col= "blue",lwd = 1, length=.1)
arrows(-a3,a4,-b5,b6, col= "blue",lwd = 1, length=.1)
arrows(-a3,a4,-b7,b8, col= "blue",lwd = 1, length=.1)
arrows(-a3,a4,-b9,b., col= "blue",lwd = 1, length=.1)
arrows(-a5,a6,-a9,a., col= "blue",lwd = 1, length=.1)
arrows(-a5,a6,-b1,b2, col= "blue",lwd = 1, length=.1)
arrows(-a5,a6,-b3,b4, col= "blue",lwd = 1, length=.1)
arrows(-a5,a6,-b5,b6, col= "blue",lwd = 1, length=.1)
arrows(-a5,a6,-b7,b8, col= "blue",lwd = 1, length=.1)
arrows(-a5,a6,-b9,b., col= "blue",lwd = 1, length=.1)
arrows(-a7,a8,-b1,b2, col= "blue",lwd = 1, length=.1)
arrows(-a7,a8,-b3,b4, col= "blue",lwd = 1, length=.1)
arrows(-a7,a8,-b5,b6, col= "blue",lwd = 1, length=.1)
arrows(-a7,a8,-b7,b8, col= "blue",lwd = 1, length=.1)
arrows(-a7,a8,-b9,b., col= "blue",lwd = 1, length=.1)
arrows(-a9,a.,-b3,b4, col= "blue",lwd = 1, length=.1)
arrows(-a9,a.,-b5,b6, col= "blue",lwd = 1, length=.1)
arrows(-a9,a.,-b7,b8, col= "blue",lwd = 1, length=.1)
arrows(-a9,a.,-b9,b., col= "blue",lwd = 1, length=.1)
arrows(-b1,b2,-b5,b6, col= "blue",lwd = 1, length=.1)
```

```
arrows(-b1,b2,-b7,b8, col= "blue",lwd = 1, length=.1)
arrows(-b1,b2,-b9,b., col= "blue",lwd = 1, length=.1)
arrows(-b3,b4,-b7,b8, col= "blue",lwd = 1, length=.1)
arrows(-b3,b4,-b9,b., col= "blue",lwd = 1, length=.1)
arrows(-b5,b6,-b9,b., col= "blue",lwd = 1, length=.1)
text(a1,a2-.051,expression(x[3]),cex=1.6)
text(a3+.12,a4-.07,expression(x["2n+1"]),cex=1.6)
text(a5+.12,a6-.07,expression(x["2n-1"]),cex=1.6)
text(a7+.12,a8-.09,expression(x["2n-3"]),cex=1.6)
text(a9+.15,a.-.05,expression(x["2n-5"]),cex=1.6)
text(b1+.1,b2,expression(x[11]),cex=1.6)
text(b3+.1,b4,expression(x[9]),cex=1.6)
text(b5+.1,b6+.05,expression(x[7]),cex=1.6)
text(b7+.1,b8+.05,expression(x[5]),cex=1.6)
text(b9-.05,b.+.06,expression(x[2]),cex=1.6)
text(a1,a2-.05,expression(x[3]),cex=1.6)
text(-a3-.05,a4-.07,expression(x["2n"]),cex=1.6)
text(-a5-.12,a6-.07,expression(x["2n-2"]),cex=1.6)
text(-a7-.12,a8-.09,expression(x["2n-4"]),cex=1.6)
text(-a9-.15,a.-.05,expression(x["2n-6"]),cex=1.6)
text(-b1-.1,b2,expression(x[10]),cex=1.6)
text(-b3-.1,b4,expression(x[8]),cex=1.6)
text(-b5-.1,b6+.05,expression(x[6]),cex=1.6)
text(-b7-.1,b8+.05,expression(x[4]),cex=1.6)
text(-b9+.05,b.+.06,expression(x[1]),cex=1.6)
```

Sintaxis de ejemplo (Capítulo 10)

```
#Bibliotecas a utilizar
library(catspec)
library(CCA)
library(CCP)
library(fields)
library(foreign, pos=4)
library(yacca)
library(yhat)
library(mvnormtest)

#Se cargan los datos
dat<-read.spss(Ç:/Users/WALTEROSKY/Desktop/TESIS/mmreg.sav",
use.value.labels=TRUE,max.value.labels=Inf,to.data.frame=TRUE)
```

```
attach(dat)#Se define el conjunto de datos
names(dat)#Muestra los nombres de las variables

#Estadísticas descriptivas
#Se definen los dos conjuntos de variables en su forma matricial
V.Psico<-as.matrix(dat[,2:4])#Conjunto de variables X
V.Acad<-as.matrix(dat[,5:9])#Conjunto de variables Y
VP<-(dat[,2:4])#Conjunto de variables X
VA<-(dat[,5:9])#Conjunto de variables Y

#Variables Psicológicas
par(mfrow=c(1,2))
hist(locus,col = ɛellow", border = "blue",labels = TRUE,
main="Locus",xlab=,ylab=,cex.main=3)
hist(concepto,col = "blue", border = red",labels = TRUE,
main=Çoncepto",xlab=,ylab=,cex.main=3)
par(mfrow=c(1,1))
hist(motivacion,col = "pink", border = "blue",labels = TRUE,
main="Motivación",xlab=,ylab=,cex.main=2)
x11()#Se abre nueva ventana de gráfica
boxplot(V.Psico,col=c(ɛellow","blue","pink"))
t(stats(V.Psico))#Resumen del conjunto

#Variables Académicas
par(mfrow=c(1,2))
hist(lectura,col = ɛellow", border = "blue",labels = TRUE,
main="Lectura",xlab=,ylab=,cex.main=2,breaks=5)
hist(redaccion,col = "blue", border = red",labels = TRUE,
main=Redacción",xlab=,ylab=,cex.main=2,breaks=5)
hist(matematicas,col = "pink", border = "blue",labels =TRUE,
main="Matemáticas",xlab=,ylab=,cex.main=2,breaks=5)
hist(ciencias,col = "green", border = "blue",labels = TRUE,
main=Çiencias",xlab=,ylab=,cex.main=2,breaks=5)
x11()
boxplot(V.Acad[,1:4],col=c(ɛellow","blue","pink","green"))
t(stats(V.Acad))#Resumen del conjunto

#Variable género
ctab(table(mujer), addmargins=TRUE)# Número de mujeres y hombres
x11()#gráfica de pastel del porcentaje de hombres y mujeres
pie(table(mujer),radius=1,main="Variable Sexo (Mujer)",col=c("blue","pink"),
init.angle =90,labels=c("Hombre","Mujer"))
```

```
#Análisis Canónico
# correlaciones
correl<-matcor(V.Psico,V.Acad)#Cálculo de la Matriz R
img.matcor(correl, type = 1)
title(main="Matriz de correlación R")
correl #valores de la matriz R
cc1 <- cc(V.Psico,V.Acad)
#Valores de las correlaciones canónicas
rho<- cancor(V.Psico,V.Acad)$cor
res.rcc<- rcc(V.Psico,V.Acad, 0.4640861, 0.1675091)
#Gráfica de las correlaciones canónicas
barplot(res.rcc$cor, col=c(3,5,"purple"),xlab = "Variable Canónica",
main=Çorrelaciones Canónicas", ylab = Çorrelación Canónica",
names.arg = 1:3, ylim = c(0,1))
cca.fit<- cca(V.Psico,V.Acad)#análisis canónico
summary(cca.fit)#resumen de los coeficientes
plot(cca.fit)#gráficas importantes en el AC

#gráfica de los coeficientes de estructura, correlación
helio.plot(cca.fit, x.name="Var. Psicológicas", y.name="Var. Académicas",
main="Primera Variable Canónica",sub=NULL)
helio.plot(cca.fit, cv=2, x.name="Var. Psicológicas",y.name="Var.
Académicas",
main="Segunda Variable Canónica",sub=NULL)
helio.plot(cca.fit, cv=3, x.name="Var. Psicológicas",y.name="Var.
Académicas",
main="Tercera Variable Canónica",sub=NULL)

#gráfica de los coeficientes de estructura, variabilidad
helio.plot(cca.fit,type="variance",x.name="Var. Psicológicas",
y.name="Var. Académicas", main="Primera Variable Canónica",sub=NULL)
helio.plot(cca.fit, cv=2,type="variance", x.name="Var. Psicológicas",
y.name="Var. Académicas", main="Segunda Variable Canónica",sub=NULL)
helio.plot(cca.fit, cv=3, type="variance", x.name="Var. Psicológicas",
y.name="Var. Académicas", main="Tercera Variable Canónica",sub=NULL)

#PRUEBAS PARA ESTIMAR EL VALOR DE S
N = dim(V.Psico)[1] #Se define el número de observaciones
p = dim(V.Psico)[2] #Se define el número de variables dependientes
q = dim(V.Acad)[2] #Se define el número de variables independientes
#Se calcula los p-value's utilizando aproximaciones a la
distribución F,
#de diferentes pruebas estadísticas:
```

```
p.asym(rho, N, p, q, tstat = "Wilks") #Lambda de Wilks
p.asym(rho, N, p, q, tstat = "Hotelling") #Traza de Hotelling-Lawley
p.asym(rho, N, p, q, tstat = "Pillai") #TrazaPillai-Bartlett
# Gráfica, prueba de Lambda Wilks
# considerando 1,2,3 correlaciones canónicas
res1 <- p.asym(rho, N, p, q,tstat = "Wilks")
par(mfrow=c(3,1))
plt.asym(res1,rhostart=1)
plt.asym(res1,rhostart=2)
plt.asym(res1,rhostart=3)
# Gráfica, prueba de Traza de Hotelling-Lawley
#considerando 1,2,3 correlaciones canónicas
res2 <- p.asym(rho, N, p, q,tstat = "Hotelling")
par(mfrow=c(3,1))
plt.asym(res2,rhostart=1)
plt.asym(res2,rhostart=2)
plt.asym(res2,rhostart=3)
# Gráfica, prueba de Traza Pillai-Bartlett
#considerando 1,2,3 correlaciones canónicas
res3 <- p.asym(rho, N, p, q,tstat = "Pillai")
par(mfrow=c(3,1))
plt.asym(res3,rhostart=1)
plt.asym(res3,rhostart=2)
plt.asym(res3,rhostart=3)
#gráfica de las variables canónicas
plt.cc(res.rcc, var.label = T, type="v")
title(main=Representación de las Variables Canónicas",sub="Variables",
cex.sub = 0.75, font.sub = 4)
plt.cc(res.rcc, var.label = F, type=ï")
title(main=Representación de las Variables Canónicas",
sub=Ïndividuos",cex.sub = 0.75, font.sub = 4)
```