Introducción.

Mediante las técnicas de regresión lineal univariantes se trata de explicar el comportamiento de una variable *Y*, llamada *variable dependiente, endógena o explicada,* a partir del comportamiento de una variable *X* (en el caso de la regresión simple) o varias variables X_1, X_2, ..., X_p (en el caso de la regresión múltiple), llamadas *variables independientes, exógenas o explicativas*, mediante una relación lineal, en el caso general, de la forma:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

o, en el caso en que las variables estuviesen centradas en el origen, de la forma:

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

donde la nueva variable $\hat{Y}$, llamada *variable predicha, estimada o teórica* aproxime lo más posible el comportamiento observado de *Y*.

En el sentido de los mínimos cuadrados, los coeficientes β_i se obtienen de forma que se minimice la varianza de los residuos $e = Y - \hat{Y}$, llamada *varianza residual*, lo que equivale a hacer máximo el cuadrado del coeficiente de correlación (simple o múltiple, según el número de variables explicativas), ya que sabemos que se cumple que:

$$Var(Y) = Var(\hat{Y}) + Var(e) \quad \text{y} \quad \rho^2 = \frac{Var(\hat{Y})}{Var(Y)}$$

y, por tanto:

$$Min\{Var(e)\} \Leftrightarrow Max\left\{\frac{Var(\hat{Y})}{Var(Y)}\right\} = Max\{\rho^2\}$$

Así pues, de entre todas las posibles combinaciones lineales de las variables explicativas X_i, la solución de mínimos cuadrados obtiene aquélla que haga máximo el cuadrado del coeficiente de correlación múltiple entre las variables *Y* y X_1, X_2, ..., X_p o, lo que es lo mismo, aquélla que haga máximo el cuadrado del coeficiente de correlación lineal simple entre *Y* e $\hat{Y}$.

El análisis de Correlación Canónica trata de extender esta idea para la extracción de posibles relaciones entre dos conjuntos distintos de variables centradas en el origen, $\{Y_1, Y_2, ..., Y_q\}$ y $\{X_1, X_2, ..., X_p\}$, observadas sobre unos mismos individuos.

Esas posibles relaciones, las concreta en las relaciones de proximidad existente entre unas

nuevas variables, $Y^{(k)}$ y $X^{(k)}$, que construye como combinaciones lineales resúmenes de sus respectivos conjuntos de variables observadas, de la forma:

$$Y^{(k)} = \alpha_1^{(k)} Y_1 + \alpha_2^{(k)} Y_2 + ... + \alpha_q^{(k)} Y_q$$
$$X^{(k)} = \beta_1^{(k)} X_1 + \beta_2^{(k)} X_2 + ... + \beta_P^{(k)} X_p$$

Extendiendo el anterior resultado expuesto para la regresión, como relación de proximidad entre estas nuevas variables, $Y^{(k)}$ y $X^{(k)}$, toma el cuadrado de su coeficiente de correlación simple.

Por lo que, el Análisis de Correlación Canónica, tratará de encontrar unos coeficientes $\alpha_1^{(k)}, \alpha_2^{(k)}, ..., \alpha_q^{(k)}$ y $\beta_1^{(k)}, \beta_2^{(k)}, ..., \beta_P^{(k)}$, que proporcionen las mejores relaciones de aproximación del tipo

$$\alpha_1^{(k)} Y_1 + \alpha_2^{(k)} Y_2 + ... + \alpha_q^{(k)} Y_q = Y^{(k)} \cong X^{(k)} = \beta_1^{(k)} X_1 + \beta_2^{(k)} X_2 + ... + \beta_P^{(k)} X_p$$

en el sentido de que hagan máximo el cuadrado de su coeficiente de correlación simple entre $Y^{(k)}$ y $X^{(k)}$.

Nótese que ambas nuevas variables, $X^{(k)}$ e $Y^{(k)}$, pueden interpretarse como variables latentes o factores explicativos de cierta faceta común del comportamiento de sendos conjunto de variables $\{Y_1, Y_2, ..., Y_q\}$ y $\{X_1, X_2, ..., X_p\}$; tanto más común cuanto mayor correlación (absoluta) presenten.

Ejemplos de la utilidad de esta técnica, pueden ser los siguientes:

- Análisis de la relación entre tipo de ingresos de las personas (cuenta propia, cuenta ajena, mixtos, prestaciones sociales, ...) con sus características socio-demográficas (edad, nivel de estudio, ocupación, ...)
- Análisis de la relación entre hábito de consumo de las familias (gastos en alimentación, transporte, educación, ocio, ...) con las características estructurales y demográficas de las mismas (número de miembros, composición de la familia por sexos, edades, estudios,...; nivel de renta, número de perceptores, ...)
- Análisis de la relación entre las respuestas a dos cuestionarios recogidos sobre un mismo conjunto de individuos en dos instantes de tiempo diferentes para estudiar las causas de una posible alteración de la conducta ante los mismos; o entre las respuestas a dos cuestionarios diferentes sobre un mismo tema desarrollados por distintos investigadores, para tratar de ver sus posibles correspondencias.

En este planteamiento se aprecia una simetría entre el tratamiento de las X's y de las Y's, ya que ambas pueden interpretarse, al mismo tiempo, como explicativas y explicadas recíprocamente. Por ello este enfoque recibe el nombre de Análisis Simétrico de Correlación Canónica. Así, en términos similares a éstos, fue introducida esta técnica, por Hotelling, en 1936, y cuyos aspectos principales se revisarán a continuación.

De otra forma, cuando se establecen concretamente los papeles de cada grupo de variables como explicativo o como explicado, imprimiendo con ello una dirección de causalidad de

las relaciones, se habla de Análisis Asimétrico de Correlación Canónica.

Correlación canónica y variables canónicas: cálculo e interpretación geométrica.

Supongamos dos conjuntos distintos de variables normales multivariantes, centradas en el origen, $y'=\{Y_1, Y_2, ..., Y_q\}$ y $x'=\{ X_1, X_2, ..., X_p \}$, donde notamos por x e y sus correspondientes variables vectoriales (columnas) q y p dimensionales con distribuciones respectivas $N(\mathbf{0},\Sigma_y)$ y $N(\mathbf{0},\Sigma_x)$.

Concatenando los vectores x e y, podemos obtener fácilmente la matriz Σ de varianzas y covarianzas de todas las variables X's e Y's, la cual será una matriz r-dimensional son $r=p+q$, que puede expresarse de la siguiente forma:

$$\Sigma = E[(x;y)(x;y)'] = E\left[\begin{pmatrix} x \\ y \end{pmatrix}\begin{pmatrix} x' & y' \end{pmatrix}\right] = E\begin{bmatrix} xx' & xy' \\ yx' & yy' \end{bmatrix} = \begin{pmatrix} E[xx'] & E[xy'] \\ E[yx'] & E[yy'] \end{pmatrix} = \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{pmatrix}$$

Notando por $\alpha^{(k)}$ y $\beta^{(k)}$ los correspondientes vectores columna de coeficientes que intervienen abajo, el problema del Análisis de Correlación Canónica será encontrar una serie de nuevas variables tipificadas e incorrelacionadas en cada uno de los dos grupos, $Y^{(k)}$ y $X^{(k)}$, que llamaremos *variables canónicas,*

$$Y^{(k)} = \alpha^{(k)'} y = \alpha_1^{(k)} Y_1 + \alpha_2^{(k)} Y_2 + ... + \alpha_q^{(k)} Y_q$$
$$X^{(k)} = \beta^{(k)'} x = \beta_1^{(k)} X_1 + \beta_2^{(k)} X_2 + ... + \beta_P^{(k)} X_p$$

de forma que sea máxima su correlación en términos absolutos (el cuadrado del coeficiente de correlación lineal) .

Al ser centradas las variables $Y_1, Y_2, ..., Y_q$ y $X_1, X_2, ..., X_p$, lo serán en consecuencia las nuevas variables canónicas $Y^{(k)}$ y $X^{(k)}$, ya que son combinaciones lineales sin términos independientes de las anteriores.

Y al exigirles como condición que se encuentren tipificadas e incorrelacionadas, estaremos imponiendo las siguientes condiciones sobre los respectivos parámetros:

$$Var(Y^{(k)}) = 1 \Leftrightarrow E\left[(\alpha^{(k)'} y)(\alpha^{(k)'} y)'\right] = \alpha^{(k)'} E[yy']\alpha^{(k)} = 1 \Leftrightarrow \alpha^{(k)'} \Sigma_y \alpha^{(k)} = 1$$
$$Cov(Y^{(l)}, Y^{(k)}) = 0 \Leftrightarrow E\left[(\alpha^{(l)'} y)(\alpha^{(k)'} y)'\right] = \alpha^{(l)'} E[yy']\alpha^{(k)} = 0 \Leftrightarrow \alpha^{(l)'} \Sigma_y \alpha^{(k)} = 0, \quad l \neq k$$

y recíprocamente,

$$Var(X^{(k)}) = 1 \Leftrightarrow E\left[(\beta^{(k)'} x)(\beta^{(k)'} x)'\right] = \beta^{(k)'} E[xx']\beta^{(k)} = 1 \Leftrightarrow \beta^{(k)'} \Sigma_x \beta^{(k)} = 1$$
$$Cov(X^{(l)}, X^{(k)}) = 0 \Leftrightarrow E\left[(\beta^{(l)'} x)(\beta^{(k)'} x)'\right] = \beta^{(l)'} E[xx']\beta^{(k)} = 0 \Leftrightarrow \beta^{(l)'} \Sigma_x \beta^{(k)} = 0, \quad l \neq k$$

Cálculo de la primera pareja de variables canónicas

En este caso, los superíndices (k) de las notaciones anteriores tomarían el valor (1). Sin pérdida de rigor, suprimiremos este superíndice en busca de una mayor claridad de las expresiones que derivaremos.

El coeficiente de correlación lineal de Fisher para las primeras variables canónicas será:

$$\rho^2(X,Y) = \frac{Cov^2(Y,X)}{Var(Y)\cdot Var(X)} = \frac{E^2\left[(\alpha' y)(\beta' x)'\right]}{(\alpha'\Sigma_y\alpha)(\beta'\Sigma_x\beta)} = \frac{(\alpha' E[yx']\beta)^2}{(\alpha'\Sigma_y\alpha)(\beta'\Sigma_x\beta)} = \frac{(\alpha'\Sigma_{yx}\beta)^2}{(\alpha'\Sigma_y\alpha)(\beta'\Sigma_x\beta)}$$

por lo que el problema de obtener las primeras variables canónicas puede expresarse como sigue:

* Encontrar α y β tales que : $Var(Y) = Var(X) = 1 \quad y \quad \rho^2(X,Y)$ sea máximo

o lo que es equivalente,

$$\text{Maximizar } (\alpha'\Sigma_{yx}\beta)^2 \quad \text{sujeto a las restricciones : } \begin{cases} \alpha'\Sigma_y\alpha = 1 \\ \beta'\Sigma_x\beta = 1 \end{cases}$$

Como es un problema de optimización con restricciones de igualdad, aplicamos el método de los multiplicadores de Lagrange, para lo que construimos su Lagrangiana L, obtenemos sus puntos estacionarios derivando respecto de las incógnitas e igualando a cero, y seleccionamos los máximos, para los que el Hessiano debe ser definido negativo.

$$L = (\alpha'\Sigma_{yx}\beta)^2 - \lambda\cdot(\alpha'\Sigma_y\alpha - 1) - \mu\cdot(\beta'\Sigma_x\beta - 1)$$

$$\left.\begin{aligned} \frac{\partial L}{\partial \alpha} &= 2(\alpha'\Sigma_{yx}\beta)\Sigma_{yx}\beta - 2\lambda\cdot\Sigma_y\alpha = 0 \Leftrightarrow (\alpha'\Sigma_{yx}\beta)^2 = \lambda\alpha'\Sigma_y\alpha = \lambda \\ \frac{\partial L}{\partial \beta} &= 2(\alpha'\Sigma_{yx}\beta)\alpha'\Sigma_{yx} - 2\mu\beta'\Sigma_x = 0 \Leftrightarrow (\alpha'\Sigma_{yx}\beta)^2 = \mu\beta'\Sigma_x\beta = \mu \end{aligned}\right\} \Rightarrow \lambda = \mu = (\alpha'\Sigma_{yx}\beta)^2 = \rho^2$$

de donde, sustituyendo los valores de λ y μ en las anteriores ecuaciones y despejando respectivamente los coeficientes α y β, obtenemos:

$$\Sigma_{yx}\beta = (\alpha'\Sigma_{yx}\beta)\Sigma_y\alpha \Leftrightarrow \alpha = \frac{\Sigma_y^{-1}\Sigma_{yx}\beta}{(\alpha'\Sigma_{yx}\beta)}$$

$$\alpha'\Sigma_{yx} = (\alpha'\Sigma_{yx}\beta)\beta'\Sigma_x \Leftrightarrow \Sigma_{xy}\alpha = (\alpha'\Sigma_{yx}\beta)\Sigma_x\beta \Leftrightarrow \beta = \frac{\Sigma_x^{-1}\Sigma_{xy}\alpha}{(\alpha'\Sigma_{yx}\beta)}$$

y sustituyendo éstos en las ecuaciones recíprocas, obtendremos que los coeficientes α y β deben verificar:

$$\beta'\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1} = (\alpha'\Sigma_{yx}\beta)^2\beta' = \rho^2\beta' \Leftrightarrow (\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})\beta = \rho^2\beta$$
$$(\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})\alpha = (\alpha'\Sigma_{yx}\beta)^2\alpha = \rho^2\alpha$$

Recordemos que λ se dice autovalor (o valor propio) de una matriz A asociado a un autovector (o vector propio) u si y sólo si $Au=\lambda u$. Además, si A es una matriz cuadrada de dimensión p, simétrica y semidefinida positiva, tiene p-autovalores no negativos, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, a partir de los cuales se pueden extraer p-autovectores ortogonales.

Por tanto, los posibles vectores de coeficientes α y β, que marcan las respectivas primeras variables canónicas son precisamente autovectores de sendas matrices:

$$\beta \text{ es un autovector de la matriz } \Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}$$
$$\alpha \text{ es un autovector de la matriz } \Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}$$

en ambos casos, asociados a un mismo autovalor asociados ρ^2. (Puede demostrarse que ambas matrices tienen autovalores no negativos, como toda matriz semidefinida positiva).

Así que, en principio, habría hasta p posibles direcciones para las variables X's y q posibles direcciones para las variables Y's candidatas a ser las primeras variables canónicas, ya que ése es número teórico de autovectores asociados a las respectivas matrices anteriormente deducidas. Sin embargo, observamos que cada pareja de autovectores candidatos a ser solución, siempre vienen asociados a un mismo autovalor ρ^2.

Similarmente a como ocurría con los espacios filas y columnas del Análisis Factorial de Correspondencias, puede demostrarse que esas matrices, de dimensiones p·p y q·q respectivamente tienen el mismo número de autovalores no nulos y coinciden dos a dos. Por lo que, en realidad, sólo tendremos que considerar un número de parejas de autovectores candidatas a solución igual al Min(p,q), ya que su autovalor asociado ρ^2 era justamente la cantidad a maximizar en nuestro problema y, obviamente, autovalores nulos en matrices con autovalores no negativos no conducirán a ninguna solución óptima.

Luego, de esas posibles soluciones, las primeras variables canónicas serán justamente aquéllas que toman por coeficientes los autovectores asociados al mayor de los autovalores de sus correspondientes matrices anteriormente deducidas, ρ^2. Éste será pues el coeficiente de determinación entre las primeras variables canónicas. Con base en ello, se define el *primer coeficiente de correlación canónica* como el coeficiente de correlación lineal simple de Fisher, ρ, entre dichas dos primeras variables canónicas.

Además, de las expresiones

$$\alpha = \frac{\Sigma_y^{-1}\Sigma_{yx}\beta}{(\alpha'\Sigma_{yx}\beta)} = \frac{\Sigma_y^{-1}\Sigma_{yx}\beta}{\rho} \quad \text{y} \quad \beta = \frac{\Sigma_x^{-1}\Sigma_{xy}\alpha}{(\alpha'\Sigma_{yx}\beta)} = \frac{\Sigma_x^{-1}\Sigma_{xy}\alpha}{\rho}$$

deducimos lo innecesario de calcular directamente los dos autovectores, ya que conocido uno, podemos, simplemente sustituyendo, obtener el otro.

Cálculo de las siguientes parejas de variables canónicas

El problema de obtener las posibles sucesivas variables canónicas (como máximo, el Min{p,q}) es similar al de la primera, exigiendo a cada pareja de variables canónicas maximizar su correlación (al cuadrado) con las mismas condiciones de estar tipificadas; pero añadiendo ahora el que las que provengan de un mismo grupo de variables (*X*'s o *Y*'s) deben estar incorrelacionadas entre sí y que, además, las nuevas variables canónicas estén incorrelacionadas con las anteriores variables canónicas del grupo de variables opuesto.

Supongamos que ya disponemos de *r-1* parejas de variables canónicas, que ordenamos matricialmente de la siguiente forma:

$$A_{(r-1)} = \begin{pmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} & \dots & \alpha_1^{(r-1)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} & \dots & \alpha_2^{(r-1)} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ \alpha_q^{(1)} & \alpha_q^{(2)} & \dots & \alpha_q^{(r-1)} \end{pmatrix} \quad , \quad B_{(r-1)} = \begin{pmatrix} \beta_1^{(1)} & \beta_1^{(2)} & \dots & \beta_1^{(r-1)} \\ \beta_2^{(1)} & \beta_2^{(2)} & \dots & \beta_2^{(r-1)} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ \beta_p^{(1)} & \beta_p^{(2)} & \dots & \beta_p^{(r-1)} \end{pmatrix}$$

Estas variables canónicas cumplirán, por las exigencias anteriores que sus correlaciones internas en cada grupo serán nulas; lo que podemos expresar como:

$$A'_{(r-1)}\Sigma_y A_{(r-1)} = I \quad \text{y que} \quad B'_{(r-1)}\Sigma_x B_{(r-1)} = I$$

y que las correlaciones cruzadas con las anteriores sean nulas; lo que podemos expresar como:

$$A'_{(r-1)}\Sigma_{yx} B_{(r-1)} = \begin{pmatrix} \rho_{(1)} & 0 & & 0 \\ 0 & \rho_{(2)} & \dots & 0 \\ \vdots & \vdots & \vdots\vdots & \vdots \\ 0 & 0 & \dots & \rho_{(r-1)} \end{pmatrix}_{(r-1)\cdot(r-1)} = \mathrm{P}_{(r-1)}$$

Ahora, el problema de encontrar la r-ésima pareja de variables canónicas será:

* Encontrar los vectores columnas $\alpha^{(r)}$ y $\beta^{(r)}$ tales que :

$$A'_{(r)}\Sigma_y A_{(r)} = I \,, \quad B'_{(r)}\Sigma_x B_{(r)} = I \,, \; A'_{(r)}\Sigma_{yx} B_{(r)} = \mathrm{P}_{(r)}$$

y que hagan máximo el cuadrado de la correlación lineal ρ^2 entre las variables $X^{(r)}, Y^{(r)}$ siendo :

$$Y^{(r)} = \alpha_1^{(r)} Y_1 + \alpha_2^{(r)} Y_2 + \dots + \alpha_q^{(r)} Y_q \quad \text{y} \quad X^{(r)} = \beta_1^{(r)} X_1 + \beta_2^{(r)} X_2 + \dots + \beta_P^{(r)} X_p$$

o lo que es equivalente,

$$\text{Maximizar } \rho^2_{(\mathrm{r})} = (\alpha^{(r)'}\Sigma_{yx}\beta^{(r)})^2 \quad \text{sujeto a las restricciones : } \begin{cases} A'_{(r)}\Sigma_y A_{(r)} = I \\ B'_{(r)}\Sigma_x B_{(r)} = I \\ A'_{(r)}\Sigma_{yx} B_{(r)} = \mathrm{P}_{(r)} \end{cases}$$

Nuevamente, como es un problema de optimización con restricciones de igualdad, aplicamos el método de los multiplicadores de Lagrange, para lo que construimos su Lagrangiana L, obtenemos sus puntos estacionarios derivando respecto de las incógnitas e igualando a cero, y seleccionamos los máximos, para los que el Hessiano debe ser definido negativo; con algo más de complejidad analítica, por lo que en aras de la brevedad, nos refieriremos a su resultado.

En definitiva, las nuevas parejas de variables canónicas vuelven a estar caracterizadas por los autovectores de las mismas matrices deducidas para la primera variable canónica, presentando siempre el mismo autovalor asociado para cada pareja. Así, si son $\rho^2_{(1)} \geq \rho^2_{(2)} \geq ... \geq \rho^2_{(\min(\mathrm{p,q}))} \geq 0$ los mayores autovalores de ambas matrices $\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}$ y $\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}$, la primera pareja de variables canónicas viene caracterizada por los autovectores en sendas matrices asociados al máximo de los autovalores; la segunda pareja de variables canónicas viene caracterizada por los autovectores en sendas matrices asociados al segundo autovalor en tamaño; la tercera, al tercer autovalor en tamaño; y así sucesivamente, hasta la de la última componente que será la del autovector asociado al menor de los autovalores no nulo.

Luego, las sucesivas parejas de variables canónicas serán justamente aquéllas que toman por coeficientes los autovectores asociados a los autovalores, $\rho^2_{(\mathrm{r})}$, de sus correspondientes matrices anteriormente deducidas, ordenadamente de mayor a menor. Éste será pues el coeficiente de determinación entre las r-ésimas variables canónicas; por lo que se define el *r-ésimo coeficiente de correlación canónica* como el coeficiente de correlación lineal simple de Fisher, $\rho_{(\mathrm{r})}$, entre dichas r-ésimas variables canónicas.

Interpretación Geométrica de los Coeficientes de Correlación Canónica

Como sabemos, todo conjunto de datos admite una interpretación dual, según se realice en el espacio de las variables o, alternativamente, en el espacio de los casos.

El en espacio de las variables, las parejas de variables canónicas $Y^{(k)} = \alpha^{(k)'}y = \alpha_1^{(k)}Y_1 + \alpha_2^{(k)}Y_2 + ... + \alpha_q^{(k)}Y_q$ y $X^{(k)} = \beta^{(k)'}x = \beta_1^{(k)}X_1 + \beta_2^{(k)}X_2 + ... + \beta_P^{(k)}X_p$ clasifican a los casos mediante la generación de hiperplanos cuyos vectores normales son respectivamente los de coordenadas

$$\alpha^{(k)} = (\alpha_1^{(k)}, \alpha_2^{(k)}, ..., \alpha_q^{(k)})' \text{ y } \beta^{(k)} = (\beta_1^{(k)}, \beta_2^{(k)}, ..., \beta_P^{(k)})'$$

Pero también, en el espacio de los casos, generan dos vectores (puntos), cuyas coordenadas son $Y\alpha^{(k)}$ y $X\beta^{(k)}$ respectivamente, sin más que tener en cuenta que la observación de las variables originales sobre los n individuos nos proporciona, con la notación habitual, la

matriz $n \cdot p$ de datos, X, procedente de las observaciones (centradas) de las variables X's y la matriz $n \cdot q$,Y , procedentes de las observaciones (centradas) de las variables Y's.

En este espacio, el coseno que forma ambos vectores puede calcularse como:

$$\cos(\phi) = \frac{\left(Y\alpha^{(k)}\right)'\left(X\beta^{(k)}\right)}{\left|Y\alpha^{(k)}\right|\cdot\left|X\beta^{(k)}\right|} \Leftrightarrow \cos^2(\phi) = \frac{\left(\left(Y\alpha^{(k)}\right)'\left(X\beta^{(k)}\right)\right)^2}{\left|Y\alpha^{(k)}\right|^2\cdot\left|X\beta^{(k)}\right|^2} = \frac{\left(\left(Y\alpha^{(k)}\right)'\left(X\beta^{(k)}\right)\right)^2}{\left(Y\alpha^{(k)}\right)'\left(Y\alpha^{(k)}\right)\left(X\beta^{(k)}\right)'\left(X\beta^{(k)}\right)}$$

de donde, al ser las variables originales centradas:

$$\cos^2(\phi) = \frac{\left(\alpha^{(k)'}\frac{1}{n}\left(Y'X\right)\beta^{(k)}\right)^2}{\left(\alpha^{(k)'}\frac{1}{n}\left(Y'Y\right)\alpha^{(k)}\right)\cdot\left(\beta^{(k)'}\frac{1}{n}\left(X'X\right)\beta^{(k)}\right)} = \frac{\left(\alpha^{(k)'}S_{yx}\beta^{(k)}\right)^2}{\left(\alpha^{(k)'}S_y^2\alpha^{(k)}\right)\left(\beta^{(k)'}S_x^2\beta^{(k)}\right)} = \hat{\rho}^2_{(k)}$$

Luego el coeficiente de correlación lineal k-ésimo puede interpretarse también como el coseno del ángulo que forman entre sí las variables canónicas en el espacio de los casos.

Propiedades.

a) Las variables canónicas $(X^{(r)})$ están todas tipificadas y las parejas $(X^{(r)}, X^{(s)})$ se encuentran incorrelacionadas entre sí para r≠s. Recíprocamente, las variables canónicas $(Y^{(r)})$ están todas tipificadas y las parejas $(Y^{(r)}, Y^{(s)})$ se encuentran incorrelacionadas entre sí para r≠s.

b) Las parejas de variables canónicas constituidas por una variable canónica combinación lineal de las X's y otra variable canónica combinación lineal de las Y's, $(X^{(r)}, Y^{(s)})$ presentan una correlación máxima cuando r=s=min(r,s), y una correlación nula cuando r≠s.

c) El cuadrado del coeficiente de correlación lineal, $\rho^2_{(r)}$, que presetan las parejas de variables canónicas $(X^{(r)}, Y^{(r)})$, (cuadrado del coeficiente r-ésimo de correlación canónica, $\rho_{(r)}$), es justamente su autovalor (común) asociado.

d) Ese autovalor (común), $\rho^2_{(r)}$, asociado a la pareja de variables canónicas $(X^{(r)}, Y^{(r)})$, es el coeficiente de determinación de $X^{(r)}$ con respecto de las variables originales $\{Y_1, Y_2, ..., Y_q\}$ y representa, por tanto, la proporción de varianza de aquélla que es explicada por éstas; y, recíprocamente, la proporción de varianza de $Y^{(r)}$ que es explicada por las variables originales $\{X_1, X_2, ..., X_p\}$, por ser igualmente el coeficiente de determinación de aquélla respecto de éstas.

e) los coeficientes $\alpha^{(r)}$ y $\beta^{(r)}$ de la r-ésima pareja de variables canónicas son respectivamente los autovectores ligados al mismo r-ésimo autovalor (una vez

ordenados de mayor a menor) de las matrices:

$$\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy} \quad (\text{sus autovector son los } \alpha^{(r)})$$
$$\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx} \quad (\text{sus autovector son los } \beta^{(r)})$$

f) Si $Y^{(r)} = \alpha_1^{(r)}Y_1 + \alpha_2^{(r)}Y_2 + ... + \alpha_q^{(r)}Y_q = \alpha^{(r)'}y$ es una variable canónica, entonces también los es $-\alpha^{(r)'}y$. Y recíprocamente, si es una variable canónica $X^{(r)} = \beta_1^{(r)}X_1 + \beta_2^{(r)}X_2 + ... + \beta_P^{(r)}X_p = \beta^{(r)'}x$, también lo es $-\beta^{(r)'}x$. Los signos se suelen tomar de forma que el correspondiente coeficiente r-ésimo de correlación canónica, $\rho_{(r)}$, sea positivo.

g) Las correlaciones canónicas son invariantes ante cambios de origen y escala (transformaciones lineales) de las variables originales. Es decir, si transformamos las variables originales $y'=\{Y_1, Y_2, ..., Y_q\}$ y $x'=\{X_1, X_2, ..., X_p\}$ en $\tilde{y}'=(\tilde{Y}_1,\tilde{Y}_2,...,\tilde{Y}_q)$ y $\tilde{x}'=(\tilde{X}_1,\tilde{X}_2,...,\tilde{X}_p)$ mediante unas matrices regulares, *A* y *B,* y vectores escalares, *a* y *b*, de tal forma que:

$$\tilde{y} = A'y + a \quad \text{y} \quad \tilde{x} = B'x + b$$

entonces, las correlaciones canónicas entre $\tilde{y}$ y $\tilde{x}$ son las mismas que entre *y* y *x*. Y además, los vectores canónicos $\alpha^{(r)}$ y $\beta^{(r)}$ se transforman, respectivamente, en:

$$\tilde{\alpha}^{(r)} = A^{-1}\alpha^{(r)} \quad \text{y} \quad \tilde{\beta}^{(r)} = B^{-1}\beta^{(r)}$$

h) Las correlaciones canónicas no varían si se sustituyen las variables originales por un mismo número de combinaciones linealmente independientes de las mismas.

Contrastación del modelo y análisis de la dimensionalidad.

Enfoque muestral

En la práctica, los valores poblacionales serán desconocidos, debiéndose estimar a partir de las observaciones de las variables originales sobre *n* individuos.

Así, para centrar las variables, desconocemos los centroides poblacionales, por lo que restaremos los centroides de la muestra $\bar{x}, \bar{y}$.

Bajo la hipótesis de normalidad multivariante de la población, podemos estimar las correspondientes matrices de varianzas y covarianzas poblacionales $\Sigma_y, \Sigma_{yx}, \Sigma_x$ a partir de sus estimaciones de máxima verosimilitud, las matrices de varianzas y covarianzas muestrales S_y, S_{yx}, S_x. Y como las variables canónicas son funciones de aquellas matrices poblacionales, las mismas funciones de éstas matrices muestrales (estimadores de máxima

verosimilitud) proporcionarán estimaciones de máxima verosimilitud igualmente para aquéllas variables canónicas.

Así pues, podremos estimar las variables canónicas a partir de los autovalores y autovectores de las matrices muestrales:

$$S_y^{-1}S_{yx}S_x^{-1}S_{xy} \qquad \text{y} \qquad S_x^{-1}S_{xy}S_y^{-1}S_{yx}$$

siendo sus autovectores estimadores máximoverosímiles de los coeficientes $\alpha^{(r)}$ y $\beta^{(r)}$ recíprocamente, y sus autovalores, estimadores máximoverosímiles de los cuadrados de los correspondientes coeficientes de correlación canónica

Adecuación del Modelo (Test de independencia completa entre los grupos de variables)

Para poder aplicar el modelo del Análisis de Correlación Canónica, necesitamos que exista correlación entre los conjuntos de variables originales $\{Y_1, Y_2, ..., Y_q\}$ y $\{X_1, X_2, ..., X_p\}$. Por lo que un contraste para ver esto puede formularse como:

$$\begin{cases} H_0 : \Sigma_{yx} = 0 \quad (\text{matriz nula}) \\ H_1 : \Sigma_{yx} \neq 0 \end{cases}$$

Bajo la hipótesis de Normalidad Multivariante $N(\mathbf{0},\Sigma)$ para el conjunto de todas las variables observadas y, por tanto, de los vectores $y'=\{Y_1, Y_2, ..., Y_q\} \in N(\mathbf{0},\Sigma_y)$ y $x'=\{X_1, X_2, ..., X_p\} \in N(\mathbf{0},\Sigma_x)$, ocurrirá que:

$$\Lambda = \frac{|\Sigma|}{|\Sigma_y||\Sigma_x|} = \left|I - \Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\right| = \prod_{r=1}^{h}\left(1-\rho_{(r)}^2\right) \xrightarrow{Bajo\,H_0} \Lambda(p, n-1-q, q)$$

Por otro lado, en muestras de tamaño n, la razón de verosimilitudes es (Anderson, 1958):

$$\lambda^* = (\Lambda)^{n/2}$$

de donde el estadístico experimental del correspondiente contraste de razón de verosimilitudes será:

$$\lambda = -2\ln(\lambda^*) = -n\ln(\Lambda) = -n\ln\left(\prod_{r=1}^{h}\left(1-\rho_{(r)}^2\right)\right) = -n\left(\sum_{r=1}^{h}\ln\left(1-\rho_{(r)}^2\right)\right) \xrightarrow[n\to\infty]{Bajo\,H_0} \chi_{pq}^2$$

siendo h=Min(p,q).

Este contraste se mejora con la *aproximación de Bartlett*

$$-\left(s - \frac{n-t+1}{2}\right)\ln\left(\Lambda(n,s,t)\right) \xrightarrow{s\to\infty} \chi_{nt}^2$$

de forma que

$$\lambda = -2\ln(\lambda^*) = -n\ln(\Lambda) = -\frac{n}{\left(n-1-q-\frac{p-q+1}{2}\right)}\cdot\left(n-1-q-\frac{p-q+1}{2}\right)\ln(\Lambda(p,n-1-q,q))$$

donde, por la aproximación de Barttlet,

$$-\left(n-1-q-\frac{p-q+1}{2}\right)\ln(\Lambda(p,n-1-q,q)) \xrightarrow{n\to\infty} \chi^2_{pq}$$

de donde:

$$-\left(n-\frac{3+p+q}{2}\right)\ln(\Lambda) = -\left(n-\frac{3+p+q}{2}\right)\left(\sum_{r=1}^{h}\ln\left(1-\rho^2_{(r)}\right)\right) \xrightarrow[n\to\infty]{Bajo\,H_0} \chi^2_{pq}$$

Análisis de la Dimensionalidad

Sabemos que el número máximo inicial de parejas de variables canónicas es h=min(p,q). Sin embargo, podría ocurrir que fuesen realmente menos si algunas parejas presentaran un coeficiente de correlación canónica nulo; lo que ocurre cuando existen extrictamente menos de h autovalores no nulos en las matrices correspondientes a partir de las cuales determinamos las variables canónicas. En este caso, si existen exactamente $k< h$ parejas de variables canónicas asociadas a autovalores no nulos, entonces el espacio de relaciones entre los conjuntos de variables enfrentados es $k<h$, induciéndonos a pensar que, o bien existen variables colineales en cada grupo, o bien existen variables en sendos grupos totalmente independientes.

Para ver cuál es la dimensión del conjunto de relaciones canónicas de dependencia (parejas de variables canónicas asociadas a autovalores no nulos) podemos establecer el siguiente contraste de razón de verosimilitudes:

$$\begin{cases} H_0 : \rho^2_{(k+1)} = 0 \quad \left(\Rightarrow \rho^2_{(k+2)} = \ldots \rho^2_{(h)} = 0\right) \\ H_1 : \rho^2_{(k+1)} > 0 \end{cases} \quad , \mathrm{k} = 0,1,\ldots,\mathrm{h}-1$$

Bajo la hipótesis de Normalidad Multivariante N($\mathbf{0}$,Σ) para el conjunto de todas las variables observadas y, por tanto, de los vectores $y'=\{Y_1, Y_2, \ldots, Y_q\} \in$ N($\mathbf{0}$,Σ_y) y $x'=\{X_1, X_2, \ldots, X_p\} \in$ N($\mathbf{0}$,Σ_x), el estadístico experimental del correspondiente contraste de razón de verosimilitudes será:

$$\lambda = -\left(n-\frac{3+p+q}{2}\right)\left(\sum_{r=k+1}^{h}\ln\left(1-\rho^2_{(r)}\right)\right) \xrightarrow[n\to\infty]{Bajo\,H_0} \chi^2_{(p-k)(q-k)}$$

Alternativamente, el test de Bartlett-Lawley contrasta las hipótesis:

$$\begin{cases} H_0 : \rho^2_{(k)} = 0 \quad \left(\Rightarrow \rho^2_{(k+1)} = ...\rho^2_{(h)} = 0\right) \\ H_1 : \rho^2_{(k)} > 0 \end{cases} \text{, k} = 1,...,\text{h}$$

Bajo la hipótesis de Normalidad Multivariante N($\mathbf{0}$,Σ) para el conjunto de todas las variables observadas y, por tanto, de los vectores $y'=\{Y_1, Y_2, ..., Y_q\} \in$ N($\mathbf{0}$,Σ_y) y $x'=\{ X_1, X_2, ..., X_p\} \in$ N($\mathbf{0}$,Σ_x), el estadístico experimental de Bartlett-Lawley será:

$$L_k = -\left(n - k - \frac{3+p+q}{2} + \sum_{r=1}^{k} \rho^{-2}_{(r)}\right)\left(\sum_{r=k+1}^{h} \ln\left(1-\rho^2_{(r)}\right)\right) \xrightarrow[n\to\infty]{Bajo\, H_0} \chi^2_{(p-k)(q-k)}$$

Relación con otras técnicas de análisis multivariante.

Regresión simple

El Análisis de Correlación Canónica, cuando cada grupo de variables se compone de una única variable, se reduce al análisis de regresión lineal simple.

En este caso, $p=q=1$ y las submatrices Σ_y, Σ_{yx}, Σ_x son respectivamente los escalares varianza de Y, Covarianza de Y con X y varianza de X; por lo que el coeficiente de correlación canónica es:

$$\rho^2 = \Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy} = \frac{\sigma^2_{xy}}{\sigma^2_x \cdot \sigma^2_y}$$

o sea, el cuadrado del propio coeficiente de correlación lineal de las variables X y Y.

Regresión Múltiple

Cuando un grupo de variables se compone de una única variable y el otro tiene varias (por ejemplo, una única Y y varias X's), el Análisis de Correlación Canónica se reduce al análisis de regresión lineal múltiple.

En este caso, $q=1$, $p=p$ y las submatrices Σ_y, Σ_{yx}, Σ_x son respectivamente el escalar varianza de Y, el vector de dimensión p de las covarianzas de la variable Y con las X's y la matriz de dimensión $p \cdot p$ de varianzas y covarianzas de las X's; por lo que el coeficiente de correlación canónica es:

$$\rho^2 = \Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy} = \frac{\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}}{\sigma^2_y} = \frac{\sigma^2_{\hat{y}}}{\sigma^2_y}$$

o sea, el coeficiente de determinación múltiple o cuadrado del coeficiente de correlación lineal múltiple de las variables Y con respecto de las X 's.

Análisis Discriminante

Desembocamos en este caso cuando definimos las *q* variables *Y*'s como indicadoras de pertenencia a cada uno de los *q+1* grupos de la variable categórica grupo del Análisis Discriminante, de la forma usual (*método indicador*)

$$Y_i = \begin{cases} 1 \text{ si el caso pertenece al grupo i - ésimo,} & \text{i} = 1,...,\text{q} \\ 0 \text{ si el caso pertenece a otro grupo distinto} \end{cases}$$

En este caso, el cuadrado del coeficiente de correlación canónica es:

$$\rho_r^2 = \Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy} = \frac{\lambda_r}{1+\lambda_r}$$

siendo λ_r el autovalor asociado a la función discriminante r-ésima

Análisis de Tablas de Contingencia

Desembocamos en este análisis cuando definimos las *q* variables *Y*'s como indicadoras de la observación de cada una de las *q* modalidades de uno de los atributos de la tabla (por ejemplo, atributo-columna) y las *p* variables *X*'s como indicadoras de ocurrencia de cada una de las *p* modalidades del otro atributo de la tabla (por ejemplo, atributo-fila).

$$Y_i = \begin{cases} 1 \text{ si el caso presenta el atributo columna i - ésimo} \\ 0 \text{ si el caso pertenece a otro grupo distinto} \end{cases}$$

$$X_i = \begin{cases} 1 \text{ si el caso presenta el atributo fila i - ésimo} \\ 0 \text{ si el caso pertenece a otro grupo distinto} \end{cases}$$

En este caso, la asociación entre los atributos (variables cualitativas) puede estudiarse a partir de *h-1*=min(*p,q*)-1 relaciones canónicas; lo que recuerda por su analogía al análisis factorial de correspondencias. En el caso de que la primera pareja de variables canónicas presente correlación canónica nula, ello indicaría la no asociación o independencia de los dos atributos enfrentados en la tabla.

Análisis Canónico Asimétrico. Redundancias y sus aplicaciones.

El planteamiento del Análisis Canónico realizado hasta ahora es simétrico en el sentido de que si se cambian las *X*'s por las *Y*'s los resultados sería los mismos. Y provee las parejas de variables canónicas más correlacionadas que pueden encontrarse.

Sin embargo, aún estando las variables canónicas, $Y^{(k)}$ y $X^{(k)}$, máximamente correlacionadas podría ocurrir que $Y^{(k)}$ (combinación lineal de las variables $Y_1, Y_2, ..., Y_q$ y en cierta forma predictoras de las *X*'s) presentara bajas correlaciones con cada una de las

variables $X_1, X_2, ..., X_p$; y viceversa, que $X^{(k)}$ (combinación lineal de las variables $X_1, X_2, ..., X_p$ y en cierta forma predictoras de las Y's) presentara bajas correlaciones con cada una de las variables $Y_1, Y_2, ..., Y_q$.

Por ello, cuando se dispone de un esquema causal claro, de forma que uno de los conjuntos de variables (por ejemplo las X's) se consideran variables exógenas o explicativas de las variables del otro grupo que se consideran variables endógenas o explicadas, más que buscar altas correlaciones entre las variables canónicas, se pretenderá encontrar altas correlaciones entre las variables $X^{(k)}$ (funciones de las exógenas) y el conjunto de variables endógenas $Y_1, Y_2, ..., Y_q$ que se pretenden explicar; o recíprocamente con el otro conjunto de variables.

Este enfoque da lugar al llamado Análisis Canónico Asimétrico. Y para abordar su planteamiento, plantearemos antes cómo medir la correlación de esas variables $X^{(k)}$ con el conjunto de variables endógenas $Y_1, Y_2, ..., Y_q$, para lo que introduciremos primero el concepto de coeficiente de redundancia.

Coeficiente de Redundancia

Se define *Coeficiente de Redundancia* entre $X^{(k)} = \beta^{(k)'} x = \beta_1^{(k)} X_1 + \beta_2^{(k)} X_2 + ... + \beta_P^{(k)} X_p$ y el conjunto de variables endógenas $Y_1, Y_2, ..., Y_q$, como el promedio de los cuadrados de las correlaciones entre la variable $X^{(k)}$ y cada una de las $Y_1, Y_2, ..., Y_q$.

Para simplificar, supongamos que todas las variables se encuentran tipificadas. En este caso, las correlaciones entre cada dos variables coinciden con sus covarianzas, por lo que el vector columna q-dimensional de estas correlaciones sería:

$$Corr(y, X^{(k)}) = Cov(y, X^{(k)}) = E\left[y\left(\beta^{(k)'} x\right)' \right] = E[yx']\beta^{(k)} = \Sigma_{yx}\beta^{(k)}$$

de donde el promedio de sus cuadrados (suma de sus cuadrados dividido por el número de variables endógenas) será:

$$CR(y \mid X^{(k)}) = \frac{1}{q}\left(\Sigma_{yx}\beta^{(k)}\right)'\left(\Sigma_{yx}\beta^{(k)}\right) = \frac{1}{q}\beta^{(k)'}\Sigma_{xy}\Sigma_{yx}\beta^{(k)}$$

Y si tenemos h=Min(p,q) combinaciones lineales, la medida de la redundancia global que recibe el nombre de *Redundancia Total* se define como la suma de las redundancias anteriores para las h combinaciones lineales:

$$CR(y \mid x) = \sum_{k=1}^{h} CR(Y \mid X^{(k)}) = \frac{1}{q}\sum_{k=1}^{h} \beta^{(k)'}\Sigma_{xy}\Sigma_{yx}\beta^{(k)}$$

Estas definiciones se aplican de forma general sobre k combinaciones lineales cualesquiera, no necesariamente las variables canónicas. Cuando se aplican sobre las variables canónicas obtenidas por el procedimiento anteriormente expuesto y las variables están tipificadas,

puede demostrarse que:

$$CR(y \mid x) = \frac{Traza(\mathrm{P}_{yx}\mathrm{P}_{xx}^{-1}\mathrm{P}_{xy})}{Traza(\mathrm{P}_{yy})} = \frac{1}{q}\sum_{k=1}^{q} R^2_{Y_k;X_1,X_2,\ldots,X_p}$$

siendo $R^2_{Y_k;X_1,X_2,\ldots,X_p}$ el coeficiente de determinación múltiple (cuadrado del coeficiente de correlación múltiple) entre la variable endógena Y_j y el conjunto de variables exógenas X_1, X_2, ..., X_p.

Análisis Canónico Asimétrico

Supongamos que uno de los conjuntos de variables (por ejemplo las *X*'s) se consideran variables exógenas o explicativas de las variables del otro grupo (en este caso, las *Y*'s) que se consideran variables endógenas o explicadas. Y supongamos, por comodidad, que todas las variables *X*'s e *Y*'s se encuentran tipificadas.

Bajo el enfoque del Análisis Canónico Asimétrico, desarrollado por Stewart y Love (1968) y Gudmundsson (1977), se trataría pues de encontrar combinaciones lineales tipificadas de las variables exógenas

$$X^{(k)} = \beta^{(k)'}x = \beta_1^{(k)}X_1 + \beta_2^{(k)}X_2 + \ldots + \beta_P^{(k)}X_p$$

de forma que presentaran las más altas correlaciones con las variables del conjunto de variables endógenas Y_1, Y_2, ..., Y_q que se pretenden explicar.

En otras palabras, para encontrar la primera variable canónica (en este enfoque asimétrico) se pretende

$$* \text{ Encontrar } \beta^{(1)} \text{ tal que : } \quad Var(X^{(1)}) = 1 \quad y \quad CR(y \mid X^{(1)}) \text{ sea máximo}$$

o lo que es equivalente,

$$\text{Maximizar } \beta^{(1)'}\Sigma_{xy}\Sigma_{yx}\beta^{(1)} \quad \text{sujeto a la restricción : } \beta^{(1)'}\Sigma_{xx}\beta^{(1)} = 1$$

lo que conduce a la ecuación:

$$\Sigma_{xy}\Sigma_{yx}\beta^{(1)} = \lambda\Sigma_{xx}\beta^{(1)} \quad \text{o equivalentemente} \quad \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yx}\beta^{(1)} = \lambda\beta^{(1)}$$

de donde se deduce que los coeficientes buscados serían los autovectores de la matriz $H = \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yx}$ asociado a su autovalor λ.

Análogamente, encontrar la segunda y sucesivas variables canónicas (en este enfoque asimétrico), imponiendo además la condición de que las nuevas variables canónicas sean ortogonales a las anteriores, conduce a ir extrayendo los sucesivos autovectores y autovalores de la anterior matriz *H*.