ANALISIS ESTADISTICO MULTIVARIANTE Y REPRESENTACION CANONICA DE FUNCIONES ESTIMABLES

TESIS PARA OPTAR AL GRADO
DE DOCTOR EN CIENCIAS, SECCION
DE MATEMATICAS, PRESENTADA POR

CARLOS M. CUADRAS AVELLANA

UNIVERSIDAD DE BARCELONA

9.4. REPRESENTACION DE FUNCIONES ESTIMABLES POR ANALISIS DE COORDENADAS PRINCIPALES

Considerando un sistema de funciones estimables $\psi^*_1, \ldots, \psi^*_s$, como s individuos del espacio E^*, pueden representarse también utilizando el método general del análisis de coordenadas, introducido por GOWER (1966).

En esencia, consiste en formar una matriz $\alpha = (a_{ij})$, a partir de una matriz $A = (a_{ij})$ de asociaciones entre los individuos, utilizando la transformación:

$$(9.4.1) \qquad a_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a} \qquad a_i = \frac{1}{s} \sum_j a_{ij}$$

$$\bar{a} = \frac{1}{s} \sum_i \bar{a}_i$$

Las columnas de α suman todas cero.

Se define la distancia entre los individuos i,j:

$$D^2_{ij} = a_{ii} + a_{jj} - 2a_{ij}$$

Gower demuestra entonces que esta distancia queda inalterada con la nueva matriz α, que tiene los mismos vectores y valores propios que A, salvo al menos uno que es nulo. Los vectores propios de α dan las coordenadas de los individuos, referidos al individuo medio.

En nuestro caso, la distancia entre dos funciones estimables es (§ 8.1)

$$D^2(\psi^*_i, \psi^*_j) = (\psi^*_i - \psi^*_j)\, \Sigma^{-1} (\psi^*_i - \psi^*_j)$$

Si tomamos como coeficiente de asociación entre ψ^*_i y ψ^*_j

$$a_{ij} = (\psi^*_i - \bar{\psi}^*)\, \Sigma^{-1} (\psi^*_j - \bar{\psi}^*) \qquad \bar{\psi}^* = \frac{1}{s} \sum_i \psi^*_i$$

entonces

$$D^2(\psi_i^*, \psi_j^*) = \alpha_{ii} + \alpha_{jj} - 2\alpha_{ij}$$

Es fácil comprobar que $\sum_i \alpha_{ij} = 0$, por lo que obtenemos directamente la matriz α, sin necesidad de la transformación (9.4.1).

Las coordenadas principales de $\psi_1^*, \ldots, \psi_s^*$ referidas a $\bar{\psi}^*$, serán solución de

$$(9.4.2) \qquad (\alpha - \lambda_i I)\, V_i = 0$$

que vamos a relacionar con las coordenadas obtenidas en § 8.4.

Si prescindimos de la matriz de probabilidades M_ψ, las variables canónicas se obtendrán de

$$(B_\psi' \cdot B_\psi - \lambda_i \Sigma)\tilde{V}_i = 0.$$

y las coordenadas canónicas, referidas a $\bar{\psi}^*$, serán, según (8.4.2),

$$\mathbb{V}_c = \tilde{V}' B_\psi'$$

Ahora bien, $\alpha = B_\psi \Sigma^{-1} B_\psi'$, y por ser Σ^{-1} matriz simétrica definida positiva, admite una descomposición de la forma:

$$\Sigma^{-1} = U \cdot U'$$

de donde $\alpha = (B_\psi U)(B_\psi U)'$. Entonces, (9.4.2) es

$$((B_\psi U)(B_\psi U)' - \lambda_i I)V_i = 0$$

$$\Rightarrow \qquad ((B_\psi U)'(B_\psi U) - \lambda_i I)W_i = 0$$

siendo $W_i = (B_\psi U)' V_i$, y de aquí deducimos:

$$U'(B'_\psi \cdot B_\psi - \lambda_i U'^{-1} \bar{U}^{-1}) U \; W_i = 0$$

$$\Rightarrow \quad (B'_\psi B_\psi - \lambda_i \Sigma) U \; W_i = 0$$

Luego: $\tilde{V}_i = U \; W_i = U(B_\psi U)' V_i$

con lo cual:

$$B_\psi \tilde{V}_i = B_\psi U \; U' B'_\psi \; V_i = \alpha V_i = \lambda_i V_i$$

de donde, salvo un factor constante, los vectores V_i constituyen la matriz $\tilde{W}_c$ de coordenadas canónicas.

El análisis de coordenadas principales de funciones estimables coincide con el análisis canónico, si en éste prescindimos de las probabilidades atribuidas a cada una de las funciones.

10 PROGRAMACION Y METODOS NUMERICOS

La representación canónica efectiva de uno o varios sistemas de funciones estimables está ligada a un complicado proceso de cálculo numérico, debidamente estudiado en este último capítulo. Su resolución se basa en las propiedades de la inversa generalizada de una matriz (PRINGLE y RAYNER,1971), y en la descomposición en valores singulares de una matriz utilizando el algoritmo numérico de GOLUB y REINSCH (1970).

Después del estudio numérico, se incluye una breve reseña de las características del programa CANG (análisis canónico generalizado), que hemos preparado y verificado en el Laboratorio de Cálculo de la Facultad de Ciencias.

Otras referencias: ANDERSON(1963), GOLUB(1969), GRAYBILL(1969).

10.1. PROBLEMAS NUMERICOS QUE PLANTEA EL ANALISIS CANONICO GENERALIZADO

El análisis multivariante y representación canónica de un sistema de funciones estimables, exige superar los siguientes pasos: resolución de las ecuaciones normales, obtención de la expresión Δ-óptima, separametrización, comparación estadística de funciones estimables. Todos estos pasos están asociados a procesos bastante complejos de álgebra lineal, y su programación en un lenguaje de alto nivel, solo puede hacerse después de un detallado análisis numérico.

Exponemos, a continuación, cada uno de los problemas que se plantean, pasando, más adelante, a su resolución.

10.1.1. Resolución de las ecuaciones normales y obtención de la expresión Δ-óptima.

Para poder calcular la estimación de la matriz Σ, (§5.2), necesitamos obtener una solución LS de los parámetros β_{Y_i}, es decir, resolver las ecuaciones normales:

$$X'\Delta X\ \beta_{Y_i} = X'\Delta\ \bar{Y}_i \qquad i=1,\dots,p \tag{10.1.1.a}$$

Un problema parecido, es la obtención dela expresión Δ-óptima de la función paramétrica estimable $\psi^* = P'\beta^*$ que nos exige hallar

$$\hat{D} = \Delta X\ D_1 \quad \text{siendo} \quad P' = X'\Delta X\ D_1 \tag{10.1.1.b}$$

Resolveremos ambos problemas utilizando las propiedades de la inversa generalizada de una matriz.

Definición 10.1.1.- Sea A una matriz de orden (m,n). Llamaremos inversa generalizada de A, a la matriz A^{-}, de orden (n,m), tal que:

$$1) \quad A.A^{-}.A = A$$
$$2) \quad A^{-}.A.A^{-} = A^{-}$$
$$3) \quad (A.A^{-})' = A.A^{-}$$
$$4) \quad (A^{-}.A)' = A^{-}.A$$

Se demuestra que la matriz A^{-} existe, es única, y que dado un sistema de ecuaciones lineales compatible, de la forma:

$$A.X = h$$

una solución es:

$$X = A^{-}.h$$

(véase PRINGLE Y RAYNER, 1971)

El sistema (10.1.1.a) es compatible, porque el subespacio generado por las columnas de $(X' \Delta X)$, coincide con el generado por las columnas de X' (lema 4.5.1). El sistema 10.1.1.b), es también compatible, porque ψ^* es función estimable.

Si hallamos una inversa generalizada de $(X' \Delta X)$, la solución de ambos sistemas será:

$$\hat{\beta}_{Y_i} = (X' \Delta X)^{-} X' \Delta \bar{Y}_i \qquad i=1,\ldots,p.$$

$$\hat{D} = X(X' \Delta X)^{-} P'$$

10.1.2. Transformación de los parámetros bajo una hipótesis nula.

Si se verifica una hipótesis nula sobre los parámetros ,

$$H.\beta = 0 \qquad H = (t,m) \qquad \text{rang } H = r'$$

para obtener los nuevos parámetros, necesitaremos hallar una matriz C, de orden $(m,m-r')$, tal que:

$$H.B = 0$$

La nueva matriz del diseño factorial, en función de los nuevos parámetros $(\theta_1,\ldots,\theta_{m-r'})$, será:

$$\tilde{X} = X.B$$

La obtención de B nos exige resolver un sistema de ecuaciones homogéneas.

10.1.3. Comparación de funciones estimables y obtención de los ejes canónicos.

La comparación estadística de funciones estimables se decide mediante un estadístico que es función de las p raices de (§ 5.6),

$$\det(R_1 - \lambda R_0) = 0$$

La obtención de los ejes canónicos exige el cálculo de los vectores y valores propios de $\hat{\Sigma}_\psi$ respecto a $\hat{\Sigma}$ (§ 8.4),

$$\hat{\Sigma}_\psi V_i = \lambda_i \hat{\Sigma} V_i$$

En uno y otro caso necesitaremos un algoritmo que nos proporcione los vectores y valores propios de una matriz simétrica semidefinida o definida positiva, respecto a una matriz simétrica definida positiva.

10.2. DESCOMPOSICION EN VALORES SINGULARES DE UNA MATRIZ

Sea A una matriz de orden (m,n) y rango r, siendo $m \geq n$. Existe (v. PRINGLE Y RAYNER, 1971, p. 3) una única descomposición de A en la forma

$$A = U \Sigma V' \qquad (10.2.1)$$

siendo:

1) U matriz de orden (m,n), verificando $U'.U = I$. Sus n vectores columna son ortonormales y asociados con los vectores propios no nulos de $A.A'$.

2) V matriz ortogonal de orden (n,n), cuyos vectores columna son los vectores propios de $A'A$.

3) $\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$ es una matriz diagonal tal que:

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0 \qquad r=\text{rang}\,(A)$$

Se verifica: $\sigma_i = \lambda_i^{-1/2} \qquad i=1,\ldots,n.$

siendo $\{\lambda_i\}$ los valores propios de $A'A$.

A la descomposición (10.2.1) se la llama descomposición en valores singulares de A. GOLUB Y REINSCH (1970) han desarrollado un algoritmo numérico para efectuar una descomposición de una matriz A cualquiera, que consiste en reducirla por transformaciones de Householder a la forma bidiagonal, y hallar entonces la descomposición en valores singulares de la matriz bidiagonal. Este trabajo, que incluye además un programa en lenguaje ALGOL, une a una depurada precisión, una muy notable rapidez.

Aplicaremos la descomposición en valores singulares de una matriz, para hallar la inversa generalizada, re-

solver ecuaciones homogéneas, y hallar los vectores y valores propios de una matriz simétrica.

10.2.1. Cálculo de la inversa generalizada

Sea A una matriz de orden (m,n), $m \geq n$, y sea

$$A = U \Sigma V' \qquad \text{rang } (A) = r$$

la descomposición en valores singulares de A. La inversa generalizada de A es

$$A^{-} = V \cdot \Sigma^{-} \cdot U'$$

siendo $\Sigma^{-} = \text{diag } (1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0)$.

En efecto:

1) $A\,A^{-}\,A = U \Sigma V' V \Sigma^{-} U' U \Sigma V' = U \Sigma V' = A$
2) $A^{-}\,A\,A^{-} = V \Sigma^{-} U' U \Sigma V' V \Sigma^{-} U' = V \Sigma^{-} U' = A^{-}$
3) $(A\,A^{-})' = (U \Sigma V' V \Sigma^{-} U')' = U \Sigma V' V \Sigma^{-} U' = A \cdot A^{-}$
4) $(A^{-} \cdot A)' = (V \Sigma^{-} U' U \Sigma^{-} U)' = A^{-} \cdot A$

Es, pues, muy fácil calcular la inversa generalizada de una matriz, a partir de su descomposición en valores singulares. Podemos aplicarlo para hallar $(X' \Delta X)^{-}$, y resolver las ecuaciones normales, y obtener la expresión Δ-óptima de una función estimable.

10.2.2. Resolución de un sistema de ecuaciones homogéneas.

Sea H una matriz de orden (t,m) y rang $(H) = r$.

Propongámonos hallar una matriz C de orden $(m,m-r)$ tal que:

$$H \cdot C = 0$$

Si es $t \geq m$, sea

$$H = U \Sigma V'$$

la descomposición en valores singulares de M.

Sea $U = (u_1,\ldots,u_m)$ y $V = (v_1,\ldots,v_m)$

Puesto que:

$$M\, v_i = U \Sigma V' v_i = \sigma_i u_i \qquad i=1,\ldots,m$$

deducimos que:

$$M\, v_i = 0 \qquad i=r+1,\ldots,m$$

y la matriz $C = (v_{r+1},\ldots,v_m)$ verifica $M.C = 0$.

Si es $t < m$, busquemos los m vectores propios de la matriz simétrica $M'M$. Habrá m-r vectores propios de valor propio nulo

$$M'M\, v_i = 0 \qquad i=r\ 1,\ldots,m.$$

Puesto que los ortocomplementarios de M y de $M'M$ coinciden (lema 4.5.1), deducimos $M\, v_i = 0$ $(i=r+1,\ldots,m)$ Luego, la matriz C formada por estos m-r vectores verifica $M.C = 0$.

Con esto resolvemos el problema de la reparametrización.

A continuación tratamos el problema de hallar los vectores y valores propios de una matriz simétrica mediante la descomposición en valores singulares.

10.2.3. Diagonalización de una matriz simétrica.

Sea A una matriz simétrica de orden (n,n). Es bien conocido que existe una transformación ortogonal U tal que:

(10.2.3.a) $$A = U \Sigma U'$$

siendo Σ una matriz diagonal con los valores propios de A. Pero (10.2.3.a) es una descomposición de A en valores singulares. Luego, la descomposición (10.2.1) de A, nos da directamente los vectores y valores propios de A, siendo U = V.

10.3. VECTORES Y VALORES PROPIOS DE UNA MATRIZ A RESPECTO A UNA MATRIZ B

Sea A una matriz simétrica, y B una matriz simétrica definida positiva. Si V es vector propio de A, respecto a B,

(10.3.1) $$(A-\lambda B)V = 0$$

entonces V es vector propio de $B^{-1}A$, pues:

$$B^{-1} A \; V = \lambda B^{-1} B \; V = \lambda \; V$$

y recíprocamente.

La obtención de los vectores y valores propios de A respecto a B, se reduce a diagonalizar la matriz no simétrica $B^{-1}A$. Para hallar los valores propios de $B^{-1}A$ utilizaremos el algoritmo de COOLEY Y LOHNES (1971).

Sea U la matriz con los vectores propios de B, y D la matriz diagonal con los valores propios correspondientes. Podemos obtener ambas matrices de la descomposición en valores singulares de B.

Entonces:

$$B = U\ D\ U' \Rightarrow B^{-1} = U\ D^{-1}\ U' = (U\ D^{-1/2})(D^{1/2}U')$$

$$\Rightarrow B^{-1/2} = U\ D^{-1/2}$$

Sustituyendo en (10.3.1):

$$(A - \lambda U\ D^{-1/2} D^{1/2} U')\ V = 0$$

se deduce, después de algunas transformaciones,

$$(B^{-1/2\,\prime} A\ B^{-1/2} - I)(B^{-1/2\,\prime} V) = 0$$

Hallando, pues, los valores propios $\tilde{V}$ de $B^{-1/2\,\prime} A\ B^{-1/2}$, los valores propios de A, respecto a B, serán:

$$V = B^{-1/2}\ \tilde{V}$$

La diagonalización de $B^{-1}A$ ha quedado resuelta con la diagonalización de dos matrices simétricas.

Los vectores propios V de A respecto a B, no quedan normalizados.

En el caso de las variables canónicas, éstas deben normalizarse de modo que

$$V'\ \Sigma\ V = \text{var}\ (V) = 1$$

10.4. UN ALGORITMO PARA EL CALCULO EFECTIVO DE $\hat{\Sigma}$

La estimación $\hat{\Sigma}$ de la matriz Σ (§ 5.2), exige el cálculo de

$$R_o(i,j) = (mY_i - X_a \beta_{Y_i})'(mY_j - X_a \beta_{Y_j}) =$$

$$= mY_i' \, mY_j - (X_a \beta_i)'(X_a \beta_{Y_j})$$

Para muestras grandes, el valor de

$$mY_i' \, mY_j = \sum_{t=1}^{k} \sum_{h=1}^{n_t} y_{ith} \, y_{jth}$$

puede llegar a ser muy elevado, provocando errores de OVERFLOW (imposibilidad de guardar un número en coma flotante que supere la capacidad de almacenamiento de una "palabra" de ordenador).

Pero podemos solventar este inconveniente, introduciendo una matriz auxiliar de productos cruzados

$$s_{ij} = \sum_{t=1}^{k} \sum_{h=1}^{n_t} (y_{ith} - \bar{y}_{it.})(y_{jth} - \bar{y}_{jt.})$$

en la que cada factor $(y_{ith} - \bar{y}_{it.})$ queda notablemente reducido en magnitud.

Finalmente tendremos:

$$R_o(i,j) = s_{ij} + \sum_{t=1}^{k} n_t \left[\bar{y}_{it.} \, \bar{y}_{jt.} - \right.$$

$$\left. - \left(\sum_{h=1}^{m} x_{th} \, \beta_{ih} \right) \left(\sum_{h=1}^{m} x_{th} \, \beta_{jh} \right) \right]$$

que se calcula sin dificultad.

10.5 EL PROGRAMA CANG

El programa CANG realiza un análisis canónico completo a uno o varios sistemas de funciones estimables.

Está escrito en FORTRAN IV, y preparado para el sistema IBM 360/30 con 96K de memoria.

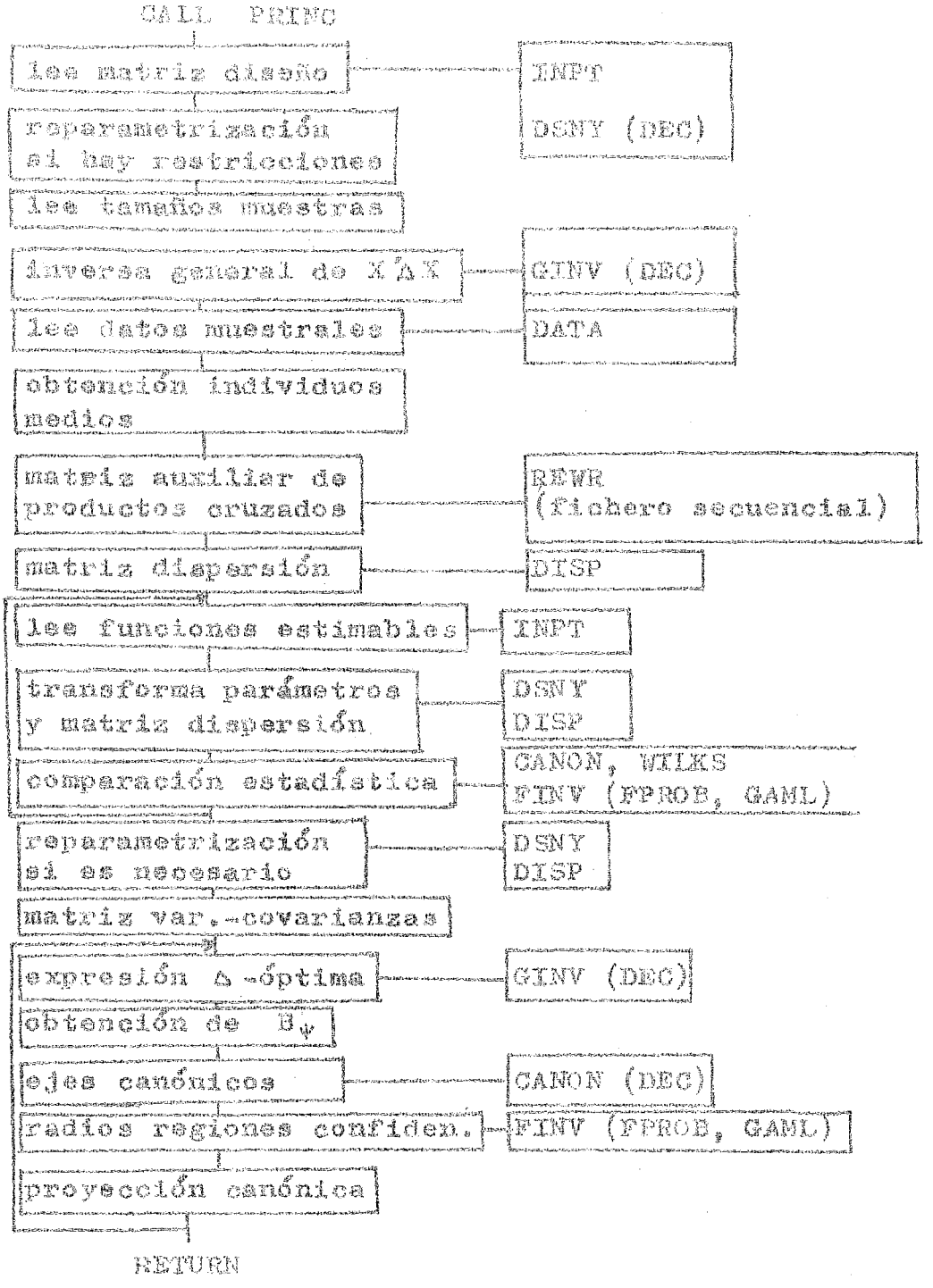
A partir de una matriz de diseño y de las observaciones de cada variable sobre los individuos de la muestra, calcula la matriz de varianzas-covarianzas, y, para cada sistema, efectua el test de comparación de funciones estimables. Si es signicativo, obtiene la expresión Δ-óptima y la proyección canónica de las funciones. Si no lo es, efectua previamente una reparametrización y transforma la matriz del diseño.

El progrma principal (FORTMAIN) se limita a leer los parámetros iniciales y llamar, de forma distinta según que el número de variables supere o no al de parámetros, al subprograma PRINC, que realiza los cálculos y llama a los demás programas. Se utilizan 21 subprogramas, de los cuales, 7 son del S.S.P. (1), y 2 son del Multivariance Prog.(2). La ejecución precisa de 5 ficheros secuenciales.

Las subrutinas especialmente programadas para CANG son:

- GINV (halla la inversa generalizada de una matriz)
- DSNY (reparametriza la matriz del diseño a partir de unas restricciones a los parámetros)
- DISP (matriz de dispersión de las v.a.)
- CANON (valores propios de una matriz A respecto a otra matriz B)
- INPT (lee matriz del diseño y funciones paramétricas)
- DATA (lee datos muestra que guarda en fichero sec.)
- WILKS (razón Λ y valor F correspondiente)
- FINV (dada la probabilidad, halla el valor F)
- REWR (lee y escribe matrices en ficheros secuenciales)
- DEC (valores singulares de una matriz; adaptada del programa en Algol de GOLUB y REINSCH, 1970).

ORGANIGRAMA POR MODULOS DE LA SUB. PRINC



(1) IBM. Scientific Subroutine Package, 360-A-CM-03X, ver.III. Programmer's Manual H2o-0205-3 .

(2) MULTIVARIANCE Program, versión IV, por Jeremy D.FINN, Comp.Center, State Univ. of N.York, Buffalos New York, U.S.A.

ANEXO : UN EJEMPLO DE APLICACION

Fuera de texto, incluimos un resumen del análisis canónico generalizado que forma parte de los cálculos estadísticos realizados sobre los datos del trabajo " Influencia de los factores de adaptación en la relación inteligencia-rendimiento escolar ", preparado por D.Javier Rudi Uriz, del Departamento de Pedagogía (Pedagogía Tecnológica) de la Universidad de Barcelona.

Se midieron cuantitativamente cuatro factores del test de Bell (familiar, salud, social, emotivo) a cada individuo de una muestra de 400 alumnos,de un mismo curso de enseñanza media. Los alumnos fueron clasificados en 12 grupos, de acuerdo con un diseño experimental de dos factores con interacción: factor rendimiento (3 niveles R_1, R_2, R_3, según la nota media en ciencias y letras), factor capacidad intelectual (4 niveles C_1, C_2, C_3, C_4, según la puntuación en el test D-48 de dominó). Se supone que el valor medio de las variables que miden la adaptación se ajusta al diseño paramétrico

$$E(X_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad i=1,2,3;\ j=1,2,3,4\ ,$$

siendo: α_i el efecto de rendimiento, β_j el efecto de capacidad y γ_{ij} la interacción rendimiento-capacidad.

Utilizando el programa CANG, se realizó un análisis canónico a los sistemas de funciones estimables:

a) Nivelesde rendimiento,

$\mu + \alpha_1$, $\mu + \alpha_2$, $\mu + \alpha_3$

b) Niveles de capacidad intelectual,

$\mu + \beta_1$, $\mu + \beta_2$, $\mu + \beta_3$, $\mu + \beta_4$

c) Efectos de interacción,

$\mu + \gamma_{ij}$ $i=1,2,3$; $j=1,2,3,4$.

La representación canónica de este último sistema, con un coeficiente de confianza del 94%, la damos a continuación.