4/2
2ej

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

ESTUDIO COMPARATIVO ENTRE LOS ANALISIS DE CORRESPONDENCIAS SIMETRICO Y NO - SIMETRICO

TESIS PROFESIONAL

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A :

RICARDO MEDINA ALVAREZ

DIRECTOR: DR. RUBEN HERNANDEZ CID

MEXICO, D. F. 1991

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

CONTENIDO

1. INTRODUCCION

1.1 Relaciones entre variables cualitativas.

En los últimos años, el estudio de relaciones entre variables cualitativas se ha realizado, en una importante parte, mediante el Análisis de Correspondencias.

El caso principal de una matriz de datos conveniente para el Análisis de Correspondencias, es una tabla de contingencia, la cual expresa la asociación de las observaciones entre dos variables cualitativas.

H. O. Hartley (1935) publicó un artículo en el cual da una formulación algebraica de la asociación entre los renglones y las columnas de una tabla de contingencia, a este artículo se le atribuye el origen matemático del Análisis de Correspondencias. Después, Louis Guttman (1941) trató el caso general de más de dos variables cualitativas, esto da como resultado lo que ahora se conoce como Análisis de Correspondencias Múltiple. El francés Jean Paul Benzécri (1977),

trabajando en un contexto lingüístico, fue el creador de la forma geométrica del Análisis de Correspondencias, donde el término francés *correspondence* fue usado para denotar "sistemas de asociaciones" entre elementos de dos conjuntos de datos (renglones y columnas).

Todos los métodos desarrollados para este tipo de análisis están basados en la hipótesis implícita de considerar simétricas las relaciones entre las variables. Esta suposición aunada al uso e interpretación de la ji-cuadrada de Pearson, medida de asociación (simétrica) en cual se apoya el Análisis de Correspondencias, podrían ser inconvenientes en casos en que la relación entre las variables no sea simétrica.

El análisis de relaciones no-simétricas para variables cuantitativas ya ha sido tratado en diferentes contextos como en Regresión Multivariada y un primer acercamiento al tratamiento no-simétrico de variables cualitativas fue propuesto por los italianos Luigi D'Ambra y Natale Lauro (1982).

1.2 Acerca de los temas tratados.

Este trabajo presenta métodos para realizar los Análisis de Correspondencias Simétrico (clásico) y No-simétrico, usando la técnica numérica de Descomposición en Valores Singulares.

Como en este trabajo, se desarrolla el Análisis de Correspondencias basado en tablas de contingencia, en el Capítulo 2 se tratan algunos conceptos relacionados con éstas, los cuales serán utilizados en capítulos posteriores, así

como un ejemplo de una tabla de contingencia la cual ilustra cada uno de los conceptos.

Los objetivos del Capítulo 3 son: primero, tratar el problema de identificar subespacios óptimos, el cual se considera como el problema principal del Análisis de Correspondencias; segundo, mostrar la teoría de la técnica de Descomposición en Valores Singulares (DVS) y tercero, mostrar cómo la técnica de DVS resuelve el problema de identificar subespacios óptimos.

Estos dos primeros capítulos son considerados de apoyo, pero básicos, para los capítulos principales, 4 y 5; donde se desarrolla la teoría relacionada con cada análisis.

Un ejemplo que ilustra el uso de ambos análisis es tratado en el Capítulo 6.

En el Apéndice A se muestran algunos ejemplos de obtención de subespacios de menor dimensión (en dimensiones 2 y 3) y además se hace notar cuál es la importancia de asignar pesos a puntos y ejes. Este capítulo podría servir como introducción para el Capítulo 3 y es muy recomendable para quienes empiezan a trabajar con Análisis de Correspondencias, aunque se requieren conocimientos de Algebra Lineal y Estadística.

Dado que las técnicas que se usan en el Análisis de Correspondencias son un tanto laboriosas, el objetivo principal del Apéndice B, es el de explicar el funcionamiento del programa de cómputo ANACORR, el cual es una propuesta de un programa para poder obtener los resultados de ambos análisis.

2. TABLAS DE CONTINGENCIA

Las tablas de contingencia expresan la asociación de las observaciones entre variables cualitativas. Los siguientes conceptos son necesarios para desarrollar la teoría de los Análisis de Correspondencias. Las tablas de contingencias que se usarán en este trabajo, son las de dos variables cualitativas.

2.1 Matriz asociada a una tabla de contingencia

Sean ***I*** y ***J***, dos variables cualitativas con I y J categorías respectivamente. Una tabla de contingencia asociada a estas variables se puede representar por medio de la matriz N_{IxJ}, donde la celda n_{ij} representa el número de observaciones clasificadas en la categoría i y en la categoría j, de las variables ***I*** y ***J***.

Esta matriz debe cumplir

la celda $n_{ij} \geq 0$; para i = 1...I y j = 1...J y

$$\sum_{j=1}^{J} n_{ij} > 0 \text{ para } i = 1...I \quad y \quad \sum_{i=1}^{I} n_{ij} > 0 \text{ para } j = 1...J;$$

2.2 Matriz de correspondencias y vectores suma

Sea $n = \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij}$, la suma de todas las celdas de N.

La *matriz de correspondencias*, $P_{I \times J}$, se define como (1/n)N, por lo que el término general de P es $p_{ij} = n_{ij}/n$. Es claro que la suma de elementos de P es uno y se puede considerar como la matriz de densidad de probabilidad conjunta sobre las celdas de la matriz N.

El vector de las sumas de celdas por renglón de P, se denota como $r = [p_{1\cdot}, ..., p_{h\cdot}, ...p_{I\cdot}]^T$, y la h-ésima componente del vector es

$$p_{h\cdot} = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} n_{ij}/n = (n_{h\cdot})/n$$

donde $n_{h\cdot}$ es la suma de elementos del renglón de h de la matriz N.

La matriz diagonal, asociada con este vector se denota como

$Dr_{bd} = diag(p_{i\cdot})$

Análogamente, el vector de las sumas de celdas por columna de P se denota como $c = [p_{\cdot 1}, ..., p_{\cdot h}, ...p_{\cdot J}]^T$ y la h-ésima componente del vector es

$$p_{\cdot h} = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} n_{ij}/n = (n_{\cdot h})/n$$

donde $n_{\cdot h}$ es la suma de elementos de la columna h de la matriz N.

La matriz diagonal, asociada con este vector se denota como

$Dc_{J\times J} = diag(p_{\cdot j})$.

Los vectores **r** y **c** pueden ser considerados como densidades marginales sobre las celdas de la matriz N.

2.3 Matrices perfil de renglones y columnas

La matriz *perfil* de renglones de P se define como

$R_{I\times J} = Dr^{-1} P$, y representa la frecuencia relativa con respecto a la variable *I* en cada celda.

Notar que el termino general, r_{ij}, se escribe como

$r_{ij} = p_{ij} / p_{i\cdot} = (n_{ij}/n) / (n_{i\cdot}/n) = n_{ij}/n_{i\cdot}$;

esto es, los renglones de P (o también de N) divididos por su respectiva suma.

Análogamente, la matriz *perfil* de columnas de C se define como

$C_{I\times J} = P\, Dc^{-1}$, y representa a cada celda como la frecuencia relativa con respecto a la variable *J*, el término general; $c_{ij} = p_{ij}/p_{\cdot j} = n_{ij}/n_{\cdot j}$, esto es, los columnas de P (o de N) divididos por su respectiva suma.

Estas matrices juegan un papel muy importante en el Análisis de Correspondencias y en el siguiente ejemplo se tratará de mostrar su importancia.

2.4 Ejemplo Ilustrativo

Una muestra de 193 empleados de una empresa son clasificados con respecto a: su hábito de fumar en las categorías NADA, POCO (1 a 5 cigarros al día), REGULAR (6 a 18 cigarros) y DEMASIADO (más de 18 cigarros); y el puesto que ocupan dentro de la empresa con las categorías GERENTE, SUBGERENTE, JEFE, ANALISTA y SECRETARIA. La tabla de contingencia asociada a la clasificación, es la siguiente

	NADA	POCO	REGULAR	DEMASIADO	Total
GERENTE	4	2	3	2	11
SUBGERENTE	4	3	7	4	18
JEFE	25	10	12	4	51
ANALISTA	18	24	33	13	88
SECRETARIA	10	6	7	2	25
Total	61	45	62	25	193

Si las variables cualitativas *I* y *J* son el *puesto* y el *hábito de fumar*, respectivamente, entonces los números de categorías asociado con estas variables son I=5 y J=4.

La matriz de Correspondencias P_{5x4} está dada por

$$P = \begin{bmatrix} 0.021 & 0.010 & 0.015 & 0.010 \\ 0.021 & 0.015 & 0.036 & 0.021 \\ 0.130 & 0.052 & 0.062 & 0.021 \\ 0.093 & 0.124 & 0.171 & 0.067 \\ 0.052 & 0.036 & 0.036 & 0.010 \end{bmatrix}$$

Como a ésta se le puede considerar como la matriz de densidad de probabilidad conjunta estimada sobre las celdas de la tabla de contingencia, se pueden dar varias relaciones como

prob. (puesto=ANALISTA y hábito=REGULAR) = 0.171

Los vectores de sumas de celdas por renglón y por columna están dados por

$\mathbf{r} = [0.057, 0.093, 0.264, 0.456, 0.130]^T$

$\mathbf{c} = [0.316, 0.233, 0.321, 0.130]^T$

Como ya se mencionó, r y c pueden ser considerados como densidades marginales estimadas sobre las celdas de la matriz N y se pueden dar los siguientes ejemplos

prob. (puesto = SECRETARIA) = 0.130 y

prob. (hábito = NADA) = 0.316

La matriz perfil de renglones R_{5x4} está dada por

$$R = \begin{bmatrix} 0.364 & 0.182 & 0.273 & 0.182 \\ 0.222 & 0.167 & 0.389 & 0.222 \\ 0.490 & 0.196 & 0.235 & 0.078 \\ 0.205 & 0.273 & 0.375 & 0.148 \\ 0.400 & 0.240 & 0.280 & 0.080 \end{bmatrix}$$

Lo que se logra al construir esta matriz es representar al total de observaciones en cada celda como frecuencias relativas con respecto a la variable *puesto*, es decir, esta matriz expresa las probabilidades condicionales de la forma P(*hábito* dado *puesto*). Por ejemplo, la frecuencia con respecto al total de observaciones en la celda GERENTE-NADA es 0.021 al igual que en la celda JEFE-DEMASIADO, pero si se observan las mismas celdas en la matriz perfil R, los valores son 0.364 para la primera y 0.078 para la segunda, esto significa que mientras que en la categoría GERENTE, la categoría NADA juega un papel importante (de 11 gerentes, 4 no fuman); la categoría DEMASIADO no es tan relevante para la de JEFE (de 51 jefes, solo 4 fuman demasiado).

La matriz perfil de columnas $C_{5\times4}$ se interpreta de forma análoga y está dada por

$$C = \begin{bmatrix} 0.066 & 0.044 & 0.048 & 0.080 \\ 0.066 & 0.067 & 0.113 & 0.160 \\ 0.410 & 0.222 & 0.194 & 0.160 \\ 0.295 & 0.533 & 0.532 & 0.520 \\ 0.164 & 0.133 & 0.113 & 0.080 \end{bmatrix}$$

Esta matriz representa las frecuencias de observaciones en cada celda como frecuencias relativas con respecto a la variable *hábito de fumar*,esto es, son probabilidades condicionales de la forma P(*puesto* dado *hábito*). Por ejemplo, las celdas GERENTE-NADA y JEFE-DEMASIADO tienen una frecuencia con respecto al total de observaciones, como ya se ha visto, de 0.021, pero en la matriz perfil de columnas, estas celdas tienen valores de 0.066 y 0.161 respectivamente, lo que significa, que aunque el papel que juegan las categorías GERENTE y JEFE, en las categorías NADA y DEMASIADO respectivamente, no es tan fuerte, es un poco más importante el papel que juega la categoría JEFE en la categoría DEMASIADO.

Gracias a esta interpretación el Análisis de Correspondencias considera como puntos a las categorías de la variable *I* (*J*), en un espacio de dimensión J (I), donde los J (I) ejes están dados por las J (I) categorías de variable *J* (*I*) y la componente j (i) del punto i (j) es r_{ij} (c_{ij}); y trata de representarlos en un subespacio "conveniente" de menor dimensión, con el objeto de apreciar las *correspondencias* entre las categorías de ambas variables.

3. IDENTIFICACION DE SUBESPACIOS OPTIMOS

Como se ha mencionado anteriormente, el objetivo principal del Análisis de Correspondencias es representar "lo mejor posible" las categorías de las variables cualitativas como puntos en un subespacio "conveniente".

En este capítulo se explicará, en primer término, qué se quiere decir con "lo mejor posible" y "conveniente", y después se propondrá una técnica para resolver el problema.

3.1 Planteamiento general para identificar subespacios óptimos

Supóngase que se tiene una nube de puntos $U = \{u_1 \ldots u_I\}$ en un espacio F de dimensión m y se requiere representar los puntos de esta nube en un subespacio S de dimensión $k^* < m$, el objetivo es encontrar el subespacio S^* el cual esté "más cerca" de la nube de puntos.

¿Cómo definir distancia entre una nube de puntos a un subespacio dado?

Dada la distancia entre puntos, entonces, se puede definir la distancia entre un punto **u** y un subespacio S como la distancia más corta entre el punto **u** y todos los puntos contenidos en el subespacio, esto es

$d(\mathbf{u}, S) = \min (d(\mathbf{u},\mathbf{y}))$ para todo $\mathbf{y} \in S$.

Ahora, la distancia de la nube de puntos $U = \{u_1 \ldots u_I\}$ al subespacio S, se define como la suma de distancias al cuadrado (por conveniencias geométricas, como en Regresión y Análisis de Varianza), de los puntos u_i al subespacio S.

Si Φ es la función que describe la distancia entre la nube U y el subespacio S, Φ está dada con la siguiente relación

$$\Phi(S;u_1 \ldots u_I) = \sum_{i=1}^{I} w_i (u_i - \hat{u}_i)^T \, Dp \, (u_i - \hat{u}_i) = \sum_{i=1}^{I} w_i \, d(u_i,S)_{Dp}$$

donde;

$\hat{u}_1 \ldots \hat{u}_I \in S$ son los puntos más cercanos a los puntos $u_1 \ldots u_I \in P$ respectivamente,

w_i es el peso (masa) asignado al punto u_i.

Dp es la matriz diagonal asociada a los pesos de las dimensiones (métrica).

Esta función depende del subespacio S y el objetivo es encontrar el subespacio óptimo S^* que minimice la función Φ.

Si se elige al subespacio S de dimensión cero, éste sólo tendría al punto Φ y la función objetivo es

$$\Phi(\Phi;u_1 \ldots u_I) = \sum_{i=1}^{I} w_i (u_i - \Phi)^T \, Dp \, (u_i - \Phi)$$

El *centroide* de la nube I, es el punto que minimiza la función Φ.

Demostración.

El centroide está dado por

$$\bar{u} = \sum_{i=1}^{I} w_i u_i$$

su j-ésima coordenada es;

$$\bar{u}_j = \sum_{i=1}^{I} w_i u_{ij}$$

y se tiene que demostrar que ésta, es la j-ésima coordenada del vector Φ la cual minimiza la función Φ.

Por otra parte, la función Φ se puede reescribir como

$$\Phi(\Phi) = \sum_{i=1}^{I} w_i (q_1 (u_{i1}-s_1)^2 + \ldots + q_j (u_{ij}-s_j)^2 + \ldots + q_J (u_{iJ}-s_J)^2)$$

Entonces, la derivada de Φ con respecto con s_j es

$$D\Phi(s_j) = -2 \sum_{i=1}^{I} w_i q_j (u_{ij}-s_j) = -2q_j \sum_{i=1}^{I} w_i (u_{ij}-s_j)$$

Ahora, igualando la derivada a cero se tiene que resolver la siguiente ecuación

$$\sum_{i=1}^{I} w_i(u_{ij}-s_j) = 0 \quad \Rightarrow \quad \sum_{i=1}^{I} w_i u_{ij} = \sum_{i=1}^{I} s_j w_i = s_j \sum_{i=1}^{I} w_i \Rightarrow$$

$s_j = \sum_{i=1}^{I} w_i u_{ij}$, el cual es punto crítico de Φ.

Entonces, sólo resta verificar que este punto minimiza Φ y por medio de la segunda derivada se tiene que

$$D^2\Phi(s_j) = -2q_j \sum_{i=1}^{I} w_i(-1) = 2q_j > 0.$$

Por lo que el centroide minimiza Φ.■

Con base en este resultado se deduce que el subespacio óptimo S^*, debe contener al *centroide*.

Demostración. Sea S' un subespacio que no contiene al centroide $\bar{u}$, se demostrará que S' no es óptimo.

Sea $u_i' \in S'$ es el punto más cercano a u_i, entonces la distancia de la nube de puntos $I = \{u_1 \ldots u_I\}$, al subespacio S' está dada por la función Φ

$$\Phi(S';u_1 \ldots u_I) = \sum_{i=1}^{I} w_i(u_i - u_i')^T \, Dp \, (u_i - u_i').$$

Sin pérdida de generalidad se puede asumir que $\Sigma_i w_i = 1$. Sean $u' \in S'$ el punto que está más cercano a $\bar{u}$, $t = \bar{u} - u'$ y, finalmente, $\hat{u}_i = u_i' + t$. La función Φ puede escribirse como

$$\Phi(S';u_1 \ldots u_I) = \sum_{i=1}^{I} w_i (u_i - \Omega_i + \Omega_i - u_i')^T \; Dp \; (u_i - \Omega_i + \Omega_i - u_i') =$$

$$= \sum_{i=1}^{I} w_i (u_i - \Omega_i)^T \; Dp \; (u_i - \Omega_i) +$$

$$+ \sum_{i=1}^{I} w_i (\Omega_i - u_i')^T \; Dp \; (\Omega_i - u_i') +$$

$$+2 \sum_{i=1}^{I} w_i (u_i - \Omega_i)^T \; Dp \; (\Omega_i - u_i').$$

El segundo sumando de esta expresión es, simplemente, $t^T Dpt$ (cuadrado de la distancia entre $\bar{u}$ y u'); el tercero es 0, ya que $Dp \; (\Omega_i - u_i') = Dpt$ y

$$\sum_{i=1}^{I} w_i (u_i - \Omega_i) = \sum_{i=1}^{I} w_i u_i - \sum_{i=1}^{I} w_i \Omega_i = \bar{u} - \sum_{i=1}^{I} w_i (u_i' + t) = \bar{u} - (u' + t) = 0$$

Entonces la función Φ se puede escribir como

$$\Phi(S';u_1 \ldots u_J) = \sum_{i=1}^{I} w_i (u_i - \Omega_i)^T \; Dp \; (u_i - \Omega_i) + t^T Dpt.$$

Entonces el subespacio definido por todos los puntos de S' mas el vector t (los puntos Ω_i) es "mejor" subespacio que S'.

Por este resultado, los puntos Ω_i se pueden escribir como

$$\Omega_i = \bar{u} + \sum_{k=1}^{k^*} f_{ik} \; m_k$$

donde $m_1 \ldots m_{k^*}$ son los vectores base del subespacio S^*, y f_{ik} son las coordenadas con respecto a esta base. La función Φ se puede escribir, entonces, como

$$\Phi(S;u_1 \ldots u_I) = \sum_{i=1}^{I} w_i (u_i - \bar{u} - \sum_{k=1}^{k^*} f_{ik} m_k)^T D p (u_i - \bar{u} - \sum_{k=1}^{k^*} f_{ik} m_k)$$

$$= \sum_{i=1}^{I} w_i [(u_i - \bar{u}) - \sum_{k=1}^{k^*} f_{ik} m_k]^T D p [(u_i - \bar{u}) - \sum_{k=1}^{k^*} f_{ik} m_k]$$

Las variables de esta función objetivo son k^* ejes principales, $m_1 \ldots m_{k^*}$, implicando un total de Jk^* variables escalares.

3.2 Técnica para la solución de identificar subespacios óptimos

En esta sección se presentan algunos resultados relacionados a la técnica numérica de Descomposición en Valores Singulares y cómo puede ésta contribuir al Análisis de Correspondencias. Para una consulta más detallada se recomienda el texto de Micheal J. Greenacre (1984).

Teorema de DVS Ordinario. Cualquier matriz A_{IxJ} real, puede ser expresada como

$$A_{IxJ} = U_{IxK} D\alpha_{KxK} V^T_{KxJ}$$

donde;

$D\alpha$ es una matriz diagonal de números positivos $\alpha_1 \ldots \alpha_K$,

K es el rango de A ($K \leq \min(I,J)$),

$U^TU = V^TV = I$ (los vectores columna de U y V son ortonormales en el espacio euclidiano).

Una forma equivalente de expresar A es

$$A = \sum_{n=1}^{K} \alpha_n \, (u^n)\,(v^n)^T$$

donde $u^1 \ldots u^K$ y $v^1 \ldots v^K$ son las columnas de U y V.

Los valores $\alpha_1 \ldots \alpha_K$ son llamados valores singulares de A, los vectores $u^1 \ldots u^K$ son llamados vectores singulares izquierdos de A y los vectores $v^1 \ldots v^K$ son llamados vectores singulares derechos de A.

Los vectores singulares izquierdos, $u^1 \ldots u^K$, forman una base ortonormal para las columnas de A y las coordenadas, con respecto a esta base, están dadas por los renglones de $G = VD\alpha$. Esto es, si la *i*-ésima columna de A es $a^i = (a_{1i} \ldots a_{Ji})^T$ y el i-ésimo renglón de G es $g_i = (g_{i1} \ldots g_{iK})$, entonces

$$a^i = g_{i1}u^1 + g_{i2}u^2 + \ldots + g_{iK}u^K.$$

De forma análoga, los vectores singulares derechos, $v^1 \ldots v^K$, forman una base ortonormal para los renglones de A y las coordenadas, con respecto a esta base, están dadas por los renglones de $F = UD\alpha$. Esto es;

$$a_i = f_{i1}v^1 + f_{i2}v^2 + \ldots + f_{iK}v^K.$$

Un caso especial de DVS es la eigendescomposición de una matriz simétrica $B_{J\times J}$ de rango $K \leq J$, como sigue

$$B_{J\times J} = V_{J\times K} D\sigma_{K\times K} V^T_{K\times J}$$

En este caso los vectores singulares derechos e izquierdos son los mismos y conocidos como eigenvectores de B mientras que los valores singulares son conocidos como eigenvalores. La prueba de la existencia de la descomposición de la matriz A se facilita gracias a la existencia de la eigendescomposición de la

matriz cuadrada y simétrica B, tal que $B = A^TA = (UD\alpha V^T)^T (UD\alpha V^T) = VD\alpha(U^TU)D\alpha V^T = VD\alpha D\alpha V^T = VD\alpha^2V^T$.

Antes de mencionar cual es la importancia de DVS para este propósito, se generalizará el resultado con lo que se conoce como DVS Generalizado.

Teorema de Descomposición en Valores Singulares Generalizado.

Si Ω_{IxI} y ϕ_{JxJ} son matrices simétricas definidas positivas, entonces cualquier matriz A_{IxJ} de rango K puede ser expresada como

$$A_{IxJ} = N_{IxK}\, D\alpha_{KxK}\, M^T_{KxJ}$$

donde los vectores columna de N y M son ortonormales con respecto a Ω y ϕ, es decir

$N^T\Omega N = M^T\phi M = I$

La importancia de DVS para este propósito es que si se eligen los valores singulares de la forma $\alpha_1 \geq \alpha_2 \geq \ldots \alpha_K$ con sus respectivos vectores singulares y si se `definen` $\Omega = D_w$ (matriz de pesos asignadas a cada punto) y $\phi = D_p$ (matriz de pesos asignados a las dimensiones) entonces la matriz

$$A_{[k^*]} = N_{[k^*]}\, D\alpha_{[k^*]}\, M^T_{[k^*]} = \sum_{s=1}^{k^*} \alpha_s n_s m_s^T$$

minimiza la siguiente función

$$\| A - X \|^2_{Dp,Dw} = \sum_{i=1}^{I} w_i (a_i - x_i)^T D_p (a_i - x_i)$$

entre todas las matrices X de rango menor o igual a K.

Recordando la función a minimizar, para encontrar el subespacio S^*

$$\Phi(S;u_1 \ldots u_I) = \sum_{i=1}^{I} w_i [(u_i-\bar{u}) - \sum_{k=1}^{k^*} f_{ik} m_k]^T Dq [(u_i-\bar{u}) - \sum_{k=1}^{k^*} f_{ik} m_k]$$

Entonces, si cada punto de la nube U = $\{u_1 \ldots u_I\}$ es asociado con un renglón de una matriz U_{kJ} y, si A es definida como la matriz centrada de renglones de U, esto es, el i-ésimo renglón de A es $u_i-\bar{u}$, entonces DVS da la solución requerida, ya que los vectores $m^1 \ldots m^{k^*}$ (columnas de M) definen una base ortonormal (ejes principales) para el subespacio óptimo y las coordenadas de los vectores $u_i-\bar{u}$ con respecto a esta base están dadas en los renglones de la matriz F, donde $F_{[k^*]}=N_{[k^*]}D\alpha_{[k^*]}$.

El total de variación de la matriz A es cuantificado por la norma cuadrada $\|A\|^2_{Dp,Dw}$, la cual se puede expresar como el cuadrado de los valores singulares, esto da una idea de qué tan bien es representada la matriz sobre los ejes principales y se tiene que

$$\|A\|^2_{Dp,Dw} = \sum_{i=1}^{I} w_i a_i^T Dp a_i = \sum_{i=1}^{K} \alpha_i^2 \quad \text{(total de inercia)},$$

similarmente la variación de la aproximación $A_{[k^*]}$ es

$$\|A_{[k^*]}\|^2_{Dp,Dw} = \sum_{i=1}^{k^*} \alpha_i^2,$$

la variación no explicada es

$$\| A - A_{[k^*]} \|^2_{Dp,Dw} = \sum_{i=k^*+1}^{K} \alpha_i^2,$$

el cociente $\tau_{k^*} = \|A_{[k^*]}\|^2_{Dp,Dw} / \|A\|^2_{Dp,Dw}$ es usado para cuantificar la calidad de la matriz $A_{[k^*]}$ en la dimensión k^*.

Para obtener DVS generalizado es recomendable partir del DVS ordinario como sigue

Sea la matriz A a la cual se desea aplicar DVS generalizado, entonces

$A = N D\alpha M^T$ donde $N^T Dw N = M^T Dq M = I$ y se desea conocer N, $D\alpha$, M.

Se definen una matriz auxiliar B, tal que

$$B = Dw^{\frac{1}{2}} A Dq^{\frac{1}{2}}$$

Ahora se aplica DVS ordinario a B, esto es

$B = U D\phi V^T$ donde $U^T U = V^T V = I$ y se encuentran U, $D\phi$ y V.

Entonces, $N = Dw^{-\frac{1}{2}} U$, $M = Dq^{-\frac{1}{2}} V$ y $D\phi = D\alpha$, ya que;

$$B = Dw^{\frac{1}{2}} A Dq^{\frac{1}{2}} = Dw^{\frac{1}{2}} (N D\alpha M^T) Dq^{\frac{1}{2}} = (Dw^{\frac{1}{2}} N) D\alpha (M^T Dq^{\frac{1}{2}})$$

Se puede poner $U = Dw^{\frac{1}{2}} N$ y $V = Dq^{\frac{1}{2}} M$, ya que

$$U^T U = (N^T Dw^{\frac{1}{2}}) (Dw^{\frac{1}{2}} N) = N^T Dw N = I$$ y

$V^T V = (M^T Dq^{\frac{1}{2}}) (Dq^{\frac{1}{2}} M) = M^T Dq M = I$, cumple con las restricciones de ortonormalización, entonces

$N = Dw^{-\frac{1}{2}} U$, $M = Dq^{-\frac{1}{2}} V$ y $D\phi = D\alpha$.

4. ANALISIS DE CORRESPONDENCIAS SIMETRICO

El Análisis de Correspondencias puede ser presentado como una técnica, para representar los renglones y columnas de una matriz de datos, como puntos en un espacio de menor dimensión, el cual tiene ciertas características.

4.1 Análisis en R^J

Considérese la nube de puntos N_i (I puntos en un espacio de dimensión J) cuyas coordenadas, con respecto a la base estándar, están dadas por los renglones de la matriz perfil de renglones R. Es decir, el i-ésimo punto de la nube es

$r_i = [r_{i1}, r_{i2}, ...,r_{iJ}]^T =$

$[n_{i1}/n_{i+},...,n_{iJ}/n_{i+}]$, esto es, el renglón i de la matriz asociada a la tabla de contingencia, en proporción con el total del renglón.

Ahora, antes de aplicar DVS para encontrar el subespacio de menor dimensión más cercano a N_I, se deben definir los pesos asociados a cada punto y los pesos asociados a las dimensiones, es decir, la matriz de pesos y la métrica.

Como matriz de pesos a puntos se elige la matriz Dr, de suma de elementos por renglón.

Como métrica se elige la matriz Dc^{-1}, inversa de la matriz de suma de elementos por columna (métrica ji-cuadrada).

Notar que, considerando la tabla de contingencia N, la proporción del renglón i (asociado con el punto i) con respecto al total, está dado por

$(n_{i1} + n_{i2} + ...+ n_{iJ})/ n = n_{i1}/n + n_{i2}/n + ...+ n_{iJ}/n = p_{i1} + p_{i2} + ...+ p_{iJ} = p_{i\cdot}$

que es, justamente, el elemento i de la diagonal de la matriz Dr (matriz asignada a los pesos de puntos).

Ahora, el peso de la columna j (asociado con el eje j) está dado por

$(n_{1j} + n_{2j} + ...+ n_{Ij})/ n = n_{1j}/n + n_{2j}/n + ...+ n_{Ij}/n = p_{1j} + p_{2j} + ...+ p_{Ij} = p_{\cdot j}$

que es el elemento j de la diagonal de la matriz Dc (inversa de la matriz asociada a la métrica).

Una vez definida la matriz de pesos a puntos se puede ver que el centroide de la nube N_I es el vector **c**, de suma de columnas de P, de la siguiente forma

$$\text{centroide de } N_I = \sum_{i=1}^{I} (p_{i\cdot}) r_i = \sum_{i=1}^{I} (p_{i\cdot}) [r_{i1},\ldots,r_{iJ}]^T =$$

$$\sum_{i=1}^{I} (p_{i\cdot}) [p_{i1}/p_{i\cdot},\ldots,p_{iJ}/p_{i\cdot}]^T = \sum_{i=1}^{I} [p_{i1},\ldots,p_{iJ}]^T =$$

$$[\sum_{i=1}^{I} p_{i1},\ldots,\sum_{i=1}^{I} p_{iJ}]^T = [p_{\cdot 1},\ldots,p_{\cdot J}]^T = \mathbf{c}$$

Entonces la matriz centrada de R, R^{-}, es aquella cuyo i-ésimo renglón es $\mathbf{r}_i - \mathbf{c}$ ($R^{-} = R - \mathbf{1}\mathbf{c}^T$, donde $\mathbf{1}$ es un vector de unos de dimensión I).

El subespacio de dimensión k^*, el cual está más cerca a la nube de puntos N_i está definido por los primeros k^* vectores singulares derechos de R^{-}, es decir, los vectores singulares derechos definen los ejes principales del nube N_i.

Para conocer los ejes principales y las coordenadas de la nube con respecto a estos ejes; se tiene que obtener DVS generalizado de la matriz R^{-}, donde los vectores singulares izquierdos y derechos están ortonormalizados con respecto a Dr y Dc^{-1} respectivamente.

Esto es, R^{-} se puede descomponer como (DVS generalizado)

$R^{-} = L\, D\mu\, M^T$ donde $L^T Dr\, L = M^T Dc^{-1} M = I$

entonces, las columnas de M (vectores singulares derechos) definen los ejes principales y los renglones de $L\, D\mu = R^{-} Dc^{-1} M$ definen las coordenadas de la nube N_i con respecto a estos ejes.

Los ejes principales del mejor subespacio de dimensión k^* son las primeras k^* columnas de la matriz M.

4.2 Análisis en R^J (Análisis dual)

En la sección anterior, se ha tratado el problema de encontrar el mejor subespacio de menor dimensión para la nube de puntos N_i, asociada a la matriz perfil de renglones R (a la nube N_i suele llamársele nube de renglones). Con un tratamiento similar y ***simétrico*** se trabaja con la nube de columnas N^J (J puntos en un espacio de dimensión I) cuyas coordenadas, con respecto a la base

estandard, están dadas por las columnas de la matriz perfil de columnas C (renglones de C^T). Es decir el j-ésimo punto de la nube es

$c^j = [c_{1j}, c_{2j}, ..., c_{Ij}]^T =$

$[n_{1j}/n_{\cdot j},...,n_{Ij}/n_{\cdot j}]^T$, es decir, la columna j de la matriz asociada a la tabla de contingencia, en proporción con en el total de esa columna.

Como matriz de pesos a los puntos se elige la matriz Dc, de suma de columnas.

Notar que, el peso de la columna j (asociada con el punto j) con respecto al total, está dado por

$(n_{1j} + n_{2j} + ...+ n_{Ij})/ n = n_{1j}/n + n_{2j}/n + ...+ n_{Ij}/n = p_{1j} + p_{2j} + ...+ p_{Ij} = p_{\cdot j}$

que es, justamente, el elemento j de la diagonal de la matriz Dc (matriz asignada a los pesos de puntos).

Como, el peso del renglón i (asociado con el eje i), considerando a N, está dado por

$(n_{i1} + n_{i2} + ...+ n_{iJ})/ n = n_{i1}/n + n_{i2}/n + ...+ n_{iJ}/n = p_{i1} + p_{i2} + ...+ p_{iJ} = p_{i\cdot}$

que es el elemento i de la diagonal de la matriz Dr, entonces, como métrica (peso a las dimensiones) se elige a la matriz Dr^{-1} (métrica ji-cuadrada).

El centroide de la nube N^J es el vector r, de suma de renglones de P;

$$\text{centroide de } N^J = \sum_{j=1}^{J} (p_{\cdot j})\, r_j = \sum_{j=1}^{J} (p_{\cdot j})[c_{1j}\ c_{2j}\ \cdots\ c_{Ij}]^T =$$

$$\sum_{j=1}^{J} (p_{\cdot j})[p_{1j}/p_{\cdot j},\ldots,p_{Ij}/p_{\cdot j}]^T = \sum_{j=1}^{J} [p_{1j},\ldots,p_{Ij}]^T =$$

$$[\sum_{j=1}^{J} p_{1j},\ldots,\sum_{j=1}^{J} p_{Ij}]^T = [p_{1\cdot},\ldots,p_{I\cdot}]^T = r$$

La matriz centrada de C^T, C^{T_r}, es aquella cuyo j-ésimo renglón es $c_j - r$ ($C^{T_r} = C^T - r1^T$, 1 es un vector de unos de dimensión J).

Para encontrar los k^* ejes principales, los cuales definen a la nube de puntos N^J, se tiene que aplicar la técnica de DVS generalizado a la matriz C^{T_r}, donde los vectores singulares izquierdos y derechos son ortonormales con respecto a Dc y Dr^{-1}, en la cual los vectores singulares derechos definen a los ejes principales de N^J.

Es decir, C^{T_r} se puede descomponer como

$C^{T_r} = U D_\mu V^T$ donde $U^T Dc\, U = V^T Dr^{-1} V = I$

entonces, las columnas de V (vectores singulares derechos) definen los ejes principales y los renglones de $UD_\mu = C^{T_r} Dr^{-1} V$ definen las coordenadas de la nube N^J con respecto a estos ejes.

Los ejes principales del mejor subespacio de dimensión k^* son las primeras k^* columnas de la matriz V.

4.3 Simetría

Observar la siguiente comparación entre los dos Análisis (R^J y R^I):

ANALISIS EN R^J, NUBE N_I (I PUNTOS DE DIMENSION J):

Puntos :
I renglones, de la matriz perfil de renglones R.

Matriz de pesos asociados a los puntos :
la matriz diagonal Dr.

Métrica :
la matriz diagonal Dc^{-1}.

Centroide :
vector de suma de columnas **c**.

ANALISIS EN R^I NUBE N^J (J PUNTOS DE DIMENSION I):

Puntos :
J columnas, de la matriz perfil de columnas, C.

Matriz de pesos asociados a los puntos :
la matriz diagonal Dc.

Métrica :
la matriz diagonal Dr^{-1}.

Centroide :
vector de suma de columnas **r**.

Se puede apreciar que existe una "gran" simetría, impuesta por las definiciones de métricas y matriz de pesos, entre los análisis en R^J y R^I, además el trato para las variables es exactamente el mismo, esto es, los resultados obtenidos, al efectuar el análisis en R^I (R^J) sobre una matriz N, asociada a una tabla de contingencia, son exactamente los mismos; a los resultados después de efectuar el análisis en R^J (R^I) donde se ha trabajado con una matriz transpuesta a N.

Existen algunos casos en los cuales esta "suposición simétrica" resulta inconveniente, como cuando se requiera trabajar con variables cualitativas no-simétricas, es decir, que una preceda a otra o explique a otra, que tengan una dependencia de estructura. Algunos casos podrían ser: droga vs. reacción, diagnosis vs. medicamento, nivel socio-económico vs. voto. Es por este tipo de casos que en el capítulo siguiente se propone un análisis, desde el punto de vista

de Análisis de Correspondencias, para estudiar este tipo de variables, al cual se le ha denominado Análisis de Correspondencias No-simétrico.

4.4 Resultados y comentarios

1.- La variación global de cada nube de puntos es cuantificada por su *total de inercia*, esto es, la suma de distancias al cuadrado a sus respectivos centroides (considerando los pesos a los puntos y la métrica).

Total de inercia de la nube N_I.

$$in(N_I) = \sum_{i=1}^{I} (p_{i\cdot})\ (r_i - c)^T\ Dc^{-1}\ (r_i - c)$$

Total de inercia de la nube N^J.

$$in(N^J) = \sum_{j=1}^{J} (p_{\cdot j})\ (c^j - r)^T\ Dr^{-1}\ (c^j - r)$$

2.- El total de inercia en ambas nubes es igual al coeficiente de contingencia de significancia al cuadrado calculado sobre N, esto es, la estadística ji-cuadrada para "independencia" dividida por el total n.

$in(N_I) = in(N^J) = X^2/n$.

donde $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - e_{ij})^2/e_{ij}$ y $e_{ij} = n_{i\cdot}.n_{\cdot j}/n$

Demostración.

$$in(N_I) = \sum_{i=1}^{I} (p_{i\cdot})\ (r_i - c)^T\ Dc^{-1}\ (r_i - c)$$

$$= \sum_{i=1}^{I} (p_{i\cdot}) \sum_{j=1}^{J} (1/p_{\cdot j}) \ (r_{ij} - p_{\cdot j})^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{i\cdot}/p_{\cdot j}) \ (p_{ij}/p_{i\cdot} - p_{\cdot j})^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{i\cdot}/p_{\cdot j}) \ (p_{ij} - p_{i\cdot}p_{\cdot j})^2 \ / \ p_{i\cdot}^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - p_{i\cdot}p_{\cdot j})^2 \ / \ p_{i\cdot}p_{\cdot j}$$

Ahora el mismo procedimiento con el total de inercia de la nube N^J;

$$in(N^J) = \sum_{j=1}^{J} (p_{\cdot j}) \ (c_j - r)^T \ Dr^{-1} \ (c_j - r)$$

$$= \sum_{j=1}^{J} (p_{\cdot j}) \sum_{i=1}^{I} (1/p_{i\cdot}) \ (c_{ij} - p_{i\cdot})^2$$

$$= \sum_{i=1}^{J} \sum_{j=1}^{I} (p_{\cdot j}/p_{i\cdot}) \ (p_{ij}/p_{\cdot j} - p_{i\cdot})^2$$

$$= \sum_{i=1}^{J} \sum_{j=1}^{I} (p_{\cdot j}/p_{i\cdot}) \ (p_{ij} - p_{i\cdot}p_{\cdot j})^2 \ / \ p_{\cdot j}^2$$

$$= \sum_{i=1}^{J} \sum_{j=1}^{I} (p_{ij} - p_{i\cdot}p_{\cdot j})^2 \ / \ p_{i\cdot}p_{\cdot j}$$

Con esto se ha demostrado que $in(N_I) = in(N^J)$, ahora se demostrará que $in(N^J) = X^2/n$, para concluir la prueba.

$$in(N^J) = \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - p_{i\cdot}p_{\cdot j})^2 \ / \ p_{i\cdot}p_{\cdot j}$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij}/n - n_{i.}n_{.j}/n^2)^2 \; / \; (n_{i.}n_{.j}/n^2)$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij}/n - e_{ij}/n)^2 \; / \; (e_{ij}/n)$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} ((n_{ij} - e_{ij})^2 \; /n^2) / \; (e_{ij}/n)$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - e_{ij})^2 / \; (ne_{ij})$$

$$= 1/n \sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - e_{ij})^2 / \; (e_{ij}) = X^2/n. \; \blacksquare$$

3.- Los respectivos subespacios de dimensión k^* de las nubes N_I y N^J, están definidos por los k^* vectores singulares derechos e izquierdos respectivamente de $P - rc^T$, con las métricas Dr^{-1} y Dc^{-1}, correspondientes a los k^* valores singulares más grandes. En otras palabras, los vectores singulares derechos e izquierdos definen los ejes principales de la nube de N_I y de la nube N^J respectivamente.

Esto es, la matriz $P - rc^T$ se puede descomponer como

$P - rc^T = A \, D_\mu \, B^T$ donde $A^T Dr^{-1} A = B^T Dc^{-1} B = I$

$\mu_1 \geq ... \geq \mu_k > 0$. Entonces las columnas de A y B definen los ejes principales de las nubes N_I y N^J respectivamente.

Demostración. Recordar que, DVS de $R^* = R - 1c^T = Dr^{-1}P - 1c^T$;

$Dr^{-1}P - 1c^T = L \, D_\mu M^T$ donde $L^T Dr \, L = M^T Dc^{-1} M = I$

entonces, las columnas de M definen los ejes principales y los renglones de $L \, D_\mu$ las coordenadas principales.

Si se multiplica por Dr por la izquierda se obtiene

$P - rc^T = (DrL)\, D\mu\, M^T$ donde $(DrL)^T\, Dr^{-1}\, (DrL) = M^T Dc^{-1} M = I$

Notar que $(DrL)^T\, Dr^{-1}\, (DrL) = (L^T Dr)\, Dr^{-1}\, (DrL) = L^T\, Dr\, L$.

Esto muestra que las columnas de M son iguales a las columnas de B.

Para mostrar que los ejes principales de la nube N^J se sigue un razonamiento análogo.∎

Comentario. Este resultado dice que no es necesario aplicar dos veces DVS, a N_I y N^J, para encontrar los ejes principales asociados a cada nube. Pero no dice nada, directamente, acerca de las coordenadas de las nubes con respecto a lo ejes (coordenadas principales). En los siguientes resultado se verá como a partir de este, se encontrarán las coordenadas principales de las nubes.

4.- Las respectivas coordenadas de las nubes N_I y N^J con respecto a sus ejes principales (coordenadas principales) están relacionadas con los ejes principales de la otra nube. Esto es,

Coordenadas principales de la nube N_I:

$F = R\, Dc^{-1}\, B$, entonces $F = Dr^{-1}\, A\, D\mu$.

Coordenadas principales de la nube N^J:

$G = C^T\, Dr^{-1}\, A$, entonces $G = Dc^{-1}\, B\, D\mu$.

Demostración. Se puede escribir F como

$F = Dr^{-1}(P - rc^T)Dc^{-1} B$ y si al DVS, $P - rc^T = A D\mu B^T$ se multiplica por $Dc^{-1} B$ por la derecha se obtiene

$(P - rc^T)Dc^{-1} B = A D\mu B^T Dc^{-1} B = A D\mu i$.

entonces $F = Dr^{-1} A D\mu$.

La demostración para G es análoga.■

Comentario. Las matrices F y G definen las coordenadas de las nubes N_I y N^J con respecto a todos los ejes principales . Las coordenadas de los puntos con respecto a un subespacio óptimo de dimensión k^* están contenidas en los renglones de las primeras k^* columnas de F y G.

5.- Los conjuntos de coordenadas de F y G están relacionadas por medio de las siguientes fórmulas.

Transición de F a G:

$G = C^{T-} F D\mu^{-1}$.

Transición de G a F:

$F = R^- G D\mu^{-1}$.

Demostración.Se tiene que $F = Dr^{-1}AD\mu$ entonces $FD\mu^{-1} = Dr^{-1}A$ y como,$G = C^{T-} Dr^{-1} A$ entonces $G = C^{T-} F D\mu^{-1}$.

De igual forma, $G = Dc^{-1} B D\mu$ entonces $GD\mu^{-1} = Dc^{-1}B$ y como, $F = R^- Dc^{-1} B$ entonces $F = R^- G D\mu^{-1}$.■

Comentario. A partir de este resultado se deduce que no es necesario calcular las dos matrices de coordenadas principales, sino que a partir de una se puede calcular la otra.

5. ANALISIS DE CORRESPONDENCIAS NO-SIMETRICO

La suposición de simetría impuesta sobre las variables, en el análisis anterior, puede ser inconveniente en casos en los que alguna de las variables tenga antecedentes lógicos de la otra, es decir, cuando haya una dependencia de estructura entre las variables (variables no-simétricas). Algunos casos que podrían presentar estas características podrían ser estudios con variables tales como: droga-reacción o sexo-deporte practicado. En estos casos resulta inconveniente dar un tratamiento simétrico a estas variables.

Los italianos Luigi D'Ambra y Natale Lauro (1982) desarrollan el Análisis en Componentes Principales sobre Subespacios Referencia, para el uso de variables no-simétricas. A partir de esta técnica, obtienen el Análisis de Correspondencias No-simétrico, para variables cualitativas no-simétricas.

En este trabajo es presentado el Análisis de Correspondencias No-simétrico usando la herramienta de Descomposición en Valores Singulares

El Análisis de Correspondencias No-simétrico hace un cierto tipo de discriminación para cada una de estas variables y su objetivo es evaluar la influencia de las J categorías, de la variable cualitativa *J*, a la cual se le dará el tratamiento de variable explicativa; sobre la distribución de la variable cualitativa *I*, a la cual se le dará el tratamiento de variable respuesta. En otras palabras, que tanta influencia hay de *J* sobre *I*.

Para tratar de establecer la dependencia de estructura entre la variable respuesta *I* y la variable explicativa *J*, se realizan dos análisis, en el espacio de columnas y en el espacio de renglones, obteniendo así los subespacios duales para cada nube de puntos.

5.1 Análisis en R^I

Considérese una nube de puntos N^J (J puntos en un espacio de dimensión I) en la cual, se puede relacionar los J puntos con la variable explicativa *J* y las I dimensiones con la variable respuesta *I*.

De la misma forma que en el caso simétrico, las coordenadas de N^J, con respecto a la base estandard, están dadas por las columnas de la matriz perfil de columnas C. Entonces el j-ésimo punto de la nube es

$$c^j = [c_{1j}, c_{2j}, \ldots, c_{Ij}]^T = [n_{1j}/n_{.j}, \ldots, n_{Ij}/n_{.j}]^T$$

Es decir, la columna j de la matriz asociada a la tabla de contingencia, en proporción con el total de esa columna. Esta nube se eligió así ya que está relacionada con la variable explicativa.

La matriz de pesos a los puntos seguirá siendo Dc.

En el caso simétrico se eligió como métrica a la matriz D_r^{-1}, pero en el caso no-simétrico podría ser inconveniente, ya que cuando la frecuencia de la categoría i, de la variable respuesta *I*, es baja, entonces, el coeficiente $1/p_{i.}$ causa que esta categoría tenga un papel importante. Es por esto que la métrica a elegir es la matriz identidad (I_{DJ}), es decir, el peso a la dimensión i es 1, para que el papel que tenga cada categoría sea el mismo.

Notar que el centroide es el mismo, es decir, **r** (suma de renglones de P), ya que la nube N^J es la misma y los pesos a los puntos también y por lo tanto la matriz centrada C^T sigue siendo igual.

Ahora, aplicando DVS:

$C^T = U D_\mu V^T$ donde $U^T Dc\, U = V^T V = I$

Las columnas de V definen los ejes principales y los renglones de $UD_\mu = C^T V$ son las coordenadas (coordenadas principales) con respecto a estos ejes.

5.2 Análisis en R^J

El Análisis en R^J, en el caso no-simétrico, es muy diferente al del caso simétrico, ya que se tiene que elegir una nube N_I (I puntos de dimensión J) más representativa considerando que hay un tratamiento diferente para cada una de

las variables *J* e *I*. Observar que la variable *I* está relacionada con los puntos y que la variable *J* con las dimensiones.

Entonces, para cada punto de la nube N_i, se consideran las J distribuciones condicionales $p_{ij}/p_{\cdot j}$, con referencia al vector de suma de renglones **r**.

Sea la matriz T_{IxJ} con estas características, entonces el término general t_{ij} está dado por

$t_{ij} = [p_{ij}/p_{\cdot j} - p_{i\cdot}]$

Es decir, las coordenadas del i-ésimo punto, con respecto a la base estandard, están dadas por el i-ésimo renglón de la matriz T

$t_i = [p_{i1}/p_{\cdot 1} - p_{i\cdot}, \ldots, p_{iJ}/p_{\cdot J} - p_{i\cdot}] =$

$[(p_{i1} - p_{i\cdot}p_{\cdot 1})/p_{\cdot 1}, \ldots, (p_{iJ} - p_{i\cdot}p_{\cdot J})/p_{\cdot J}] =$

$[(n_{i1} - n_{i\cdot}n_{\cdot 1}/n)/n_{\cdot 1}, \ldots, (n_{iJ} - n_{i\cdot}n_{\cdot J}/n)/n_{\cdot J}]$

Como matriz de pesos a los puntos, relacionada con la variable *I*, se elige a la matriz identidad, ya que se quiere dar el mismo peso a cada punto. Como matriz de pesos a las dimensiones (métrica) se elige la matriz Dc, ya que el peso a cada dimensión está en proporción del total de cada una de las columnas de P.

Es fácil ver que el centroide es el vector **0**;

centroide de N_i =

$$\sum_{i=1}^{I} (1)[p_{i1}/p_{\cdot 1} - p_{i\cdot}, \ldots, p_{iJ}/p_{\cdot J} - p_{i\cdot}] / n =$$

$$[\sum_{i=1}^{I} p_{i1}/p_{\cdot 1} - \sum_{i=1}^{I} p_{i\cdot}, \ldots, \sum_{i=1}^{I} p_{iJ}/p_{\cdot J} - \sum_{i=1}^{I} p_{i\cdot}] / n =$$

$$[p_{\cdot 1}/p_{\cdot 1} - 1, \ldots, p_{\cdot J}/p_{\cdot J} - 1] / n = 0.$$

Ahora para encontrar los ejes principales del mejor subespacio de dimensión k*, se tiene que aplicar DVS de la siguiente manera:

$T = M D_\mu L^T$ donde $M^T M = L^T Dc L = I$

Las columnas de N definen los ejes principales y los renglones de $M D_\mu = T Dc L$, son las coordenadas con respecto a estos ejes.

5.3 Indice τ de Goodman-Kruskal

El análisis clásico de tablas de contingencia sugiere medidas de asociación no simétricas como el índice τ de descomposición de predictibilidad debido a Goodman-Kruskal, el cual permite dar una interpretación probabilística.

El Análisis de Correspondencias No-simétrico se basa en el índice descomposición de predictibilidad de Goodman-Kruskal (1954), ya que ofrece algunas ventajas más, que si se usara el índice ϕ^2 de Pearson ($\phi^2 = X^2/n$).

El índice τ está dado por

$\tau = (\Sigma(p^2_{ij}/p_{\cdot j}) - \Sigma p^2_{i\cdot})/(1 - \Sigma p^2_{i\cdot})$.

Si se comparan ambos índices se puede observar que:

-τ contrariamente a ϕ^2, tiene definido un limite superior;

-τ contrariamente a ϕ^2, no varía si los elementos de la tabla de contingencia son multiplicados por una constante;

-τ, es menos sensible cuando la distribución marginal de la variable respuesta es muy asimétrica;

-τ contrariamente a ϕ^2, crece cuando la varianza de la variable respuesta decrece;

-τ y ϕ^2, decrecen cuando la varianza de la variable explicativa decrece.

6. APLICACION

En este capítulo, se mostrará un ejemplo de ambos tipos de análisis usando el programa ANACORR (Apéndice B). Este, es un ejemplo médico donde se intenta mostrar algunas de las ventajas y desventajas aunadas al uso de los dos análisis, aunque cada uno de ellos esté basado en supuestos diferentes.

En la primera parte del capítulo, se definirán las variables cualitativas a usar, así como las categorías de cada una de ellas, y a partir de éstas se construirá la tabla de contingencia. En la segunda y tercera partes, se le aplicarán a este caso los Análisis de Correspondencias Simétrico y No-simétrico, respectivamente, mostrando en ambas partes los resultados obtenidos en ANACORR. En la última parte, se tratan algunos comentarios de los resultados obtenidos.

6.1 Datos médicos

Se considerará un caso de estudio, donde los datos conciernen a la frecuencia, con la cual seis medicamentos fueron prescritos para siete enfermedades.

Las variables cualitativas, sus categorías y las claves de éstas, se describen en los siguientes cuadros.

MEDICAMENTO	
1. Penicilina	PENI
2. Tifomicina	TIFO
3. Tetraciclina	TETR
4. Eritromicina	ERIT
5. Tiofenicina	TIOF
6. Gentamicina	GENT

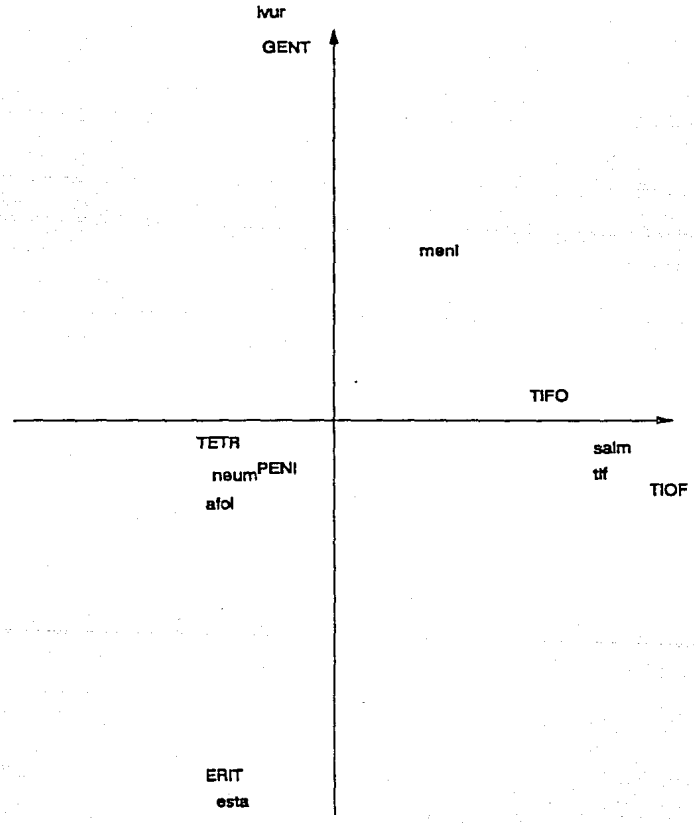
ENFERMEDAD	
1. Tifoidea	tif
2. Salmonelosis	salm
3. Afección Otorrino-laringea	afol
4. Neumocistosis	neum
5. Meningitis	meni
6. Infección vías urinarias	ivur
7. Estafilococcia	esta

69 pacientes que presentaron tener síntomas de alguna de estas *enfermedades* se les preguntó que *medicamento* se les prescribió, a partir de esta información (Hospital La Raza) se obtiene la siguiente tabla de contingencia.

MEDICAMENTOS	ENFERMEDADES							Tot ren
	tif	salm	afol	neum	meni	ivur	esta	
PENI	0	0	8	7	2	4	3	24
TIFO	4	2	0	0	2	0	0	8
TETR	0	0	5	5	0	2	1	13
ERIT	0	0	3	2	0	0	3	8
TIOF	2	1	0	0	0	0	0	3
GENT	0	0	3	3	1	6	0	13
Tot col	6	3	19	17	5	12	7	69

Este, parece ser un caso en donde una variable cualitativa (ENFERMEDAD) precede a otra (MEDICAMENTO) y por lo tanto parece ser más conveniente realizar Análisis de Correspondencias No-simétrico.

6.2 Aplicación simétrica


ivur
GENT
meni
TIFO
TETR
neumPENI
afol
salm
tif
TIOF
ERIT
esta

Al realizar Análisis de Correspondencias Simétrico se obtuvieron los siguientes valores singulares

$\alpha_1^2 = 0.8783$

$\alpha_2^2 = 0.2037$

$\alpha_3^2 = 0.0492$

$\alpha_4^2 = 0.0286$

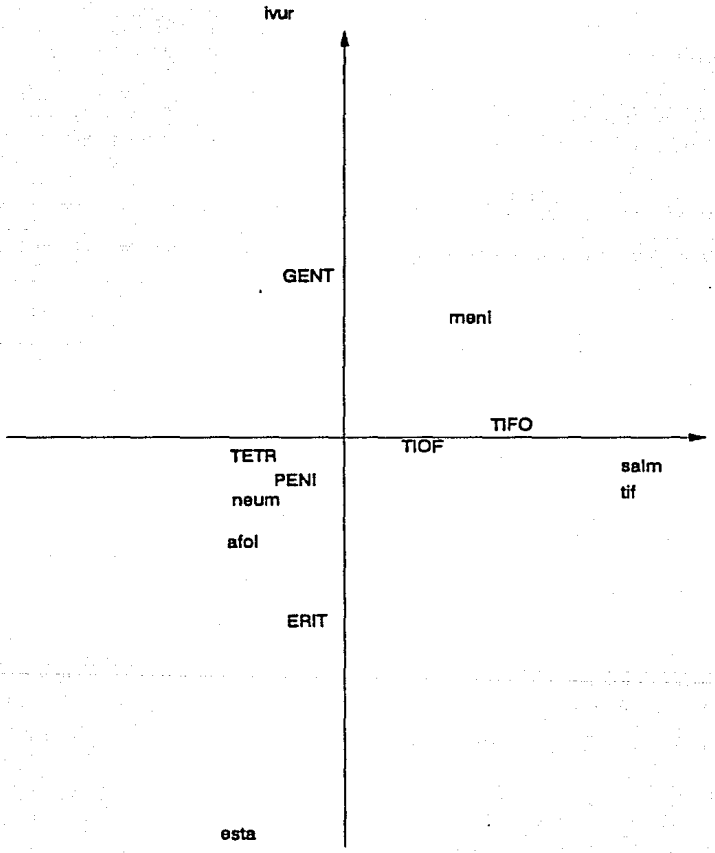
$\alpha_5^2 = 0.00004$

Por lo tanto, el total de inercia (ji-cuadrada de Pearson) obtenido en este caso es **1.159.**

El "mejor" desplegado (utilizando los ejes de mayor inercia)de puntos simétrico está dado en la gráfica anterior.

6.3 Aplicación no-simétrica

Considerando a la variable MEDICAMENTO como respuesta y como explicativa a la variable ENFERMEDAD, se efectúa Análisis de Correspondencias No-simétrico, con el objetivo de poder prescribir alguno de los medicamentos (categorías de la variable MEDICAMENTO) cuando un individuo presente tener síntomas de alguna de las enfermedades (categorías de la variable ENFERMEDAD). Se obtiene para valor de $\tau = 0.180$ (Indice Goodman-Kruskal) y el desplegado, no-simétrico, obtenido es el siguiente


ivur
GENT
meni
TIFO
TIOF
TETR
PENI
neum
afol
salm
tif
ERIT
esta

6.4 Comentarios de las aplicaciones

Si se observan ambas gráficas se puede notar que en la gráfica simétrica, la cercanía entre algunos puntos es más marcada que en la gráfica no-simétrica, es decir, las "parejas" entre categorías de ambas variables son más notorias. Por ejemplo, las categorías *ERIT-esta* y *GENT-ivur* aparecen claramente como "parejas" en la gráfica simétrica, no así en la no-simétrica.

Con referencia en la tabla de contingencia es evidente que, por ejemplo, *ERIT* es intermedia con respecto a las enfermedades para las cuales el medicamento es usado (*afol* 3, *neum* 2, *esta* 3), y como el Análisis de Correspondencias No-simétrico asigna un mismo peso a las categorías *afol*, *neum* y *esta* (y a las otras categorías de la variable ENFERMEDAD), en la gráfica se aprecia que *ERIT* está intermedia a estas categorías. A diferencia el Análisis de Correspondencias Simétrico considera los pesos de las categorías y como el menor de estos pesos es el de la categoría *esta*, en la gráfica simétrica forman una notoria "pareja". En otras palabras, la interpretación que se le podría dar en el caso no-simétrico es: de 69 personas a 8 se les prescribió eritromicina de las cuales 3 presentaron síntomas de afección otorrino-laringea; 2, de neumocistosis y 3, de estafilococcia, entonces eritromicina podría estar explicado por estas tres enfermedades. En el caso simétrico la interpretación podría ser: a 8 se les prescribió eritromicina; de las cuales 3 tenían afección otorrino-laringea (de 12 totales), 2 neumocistosis (de 17) y 3 estafilococcia (de 7); y a 7 se les encontraron síntomas de estafilococcia; de las cuales a 3 les recetaron penicilina (de 24), a 1 tetraciclina (de 13) y a 3 eritromicina (de 8); entonces se puede apreciar,

apreciar, considerando los totales marginales, que hay más dependencia entre eritromicina y estafilococcia que, por ejemplo, entre eritromicina y neumocistosis, o entre penicilina y estafilococcia.

Comentarios generales

En este trabajo, se han desarrollado los Análisis de Correspondencias Simétrico (clásico) y No-simétrico, basados en la poderosa técnica numérica de Descomposición en Valores Singulares (DVS).

El Análisis de Correspondencias Simétrico, parte de dos variables cualitativas no independientes (sentido de probabilidad) y trata de encontrar la dependencia entre estas variables; no así el Análisis de Correspondencias No-simétrico, el cual parte de dos variables no-simétricas (variable explicativa y otra de respuesta) y trata de mostrar la dependencia de estructura entre ambas variables.

Considerando lo anterior puede observarse que los análisis están basados sobre hipótesis diferentes, es decir se puede elegir uno u otro dependiendo de lo que se quiera investigar. En este trabajo se ha querido mostrar la diferencia de los métodos para desarrollar cada análisis, estos métodos, basados en DVS, están apoyados teóricamente en Análisis en Componentes Principales, para el caso simétrico, y en Análisis en Componentes Principales sobre un Subespacio Referencia (N. Lauro, D'Ambra), para el caso no-simétrico.

Por último, dado que realizar los análisis manualmente resultaría casi imposible, muy probablemente erróneo (considerando errores de cálculo) y no se conoce algún programa que realice Análisis de Correspondencias No-simétrico, se

ha propuesto el programa en computadora ANACORR el cual realiza ambos análisis, en una forma muy sencilla.

APENDICE A
EJEMPLOS DE OBTENCION DE SUBESPACIOS Y ASIGNACION DE PESOS

A.1 Ejemplos de obtención de subespacios de dimensión menor para puntos en R^2 y R^3

Considérese que se está interesado en las medidas de altura y de peso de personas. Así, los vectores de datos, de orden 2, para tres personas (personas A, B, C) son:

$$x_A = [175\ 70]^T \quad x_B = [170\ 68]^T \quad x_C = [150\ 60]^T.$$

Estos vectores pueden ser vistos como puntos en el plano, (espacio de dimensión 2) los valores de cada vector son las coordenadas del punto.

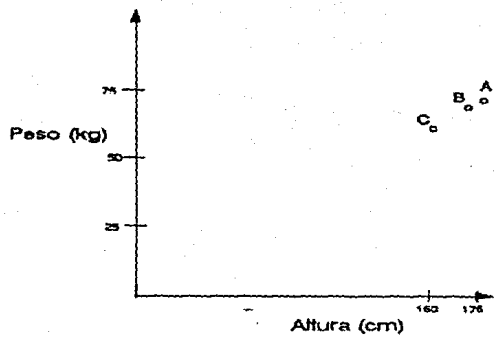


Fig 1.1

Se puede observar que los tres puntos; x_A, x_B y x_C; están en una línea recta que, además, pasa por el origen. Esto significa que los tres vectores de datos pueden ser expresados como múltiplos de un solo vector, por ejemplo el vector $\mathbf{b} = [10\ 4]^T$:

$$x_A = 17.5b \quad x_B = 17b \quad x_C = 15b$$

Así estos tres puntos se representan en un subespacio de dimensión 1, definido por el vector base **b** y con coordenadas, con respecto a **b**, de 17.5, 17 y 15 respectivamente.

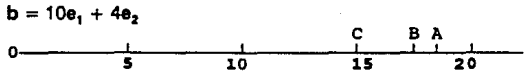


FIG 1.3

Esta nueva dimensión, la cual es una combinación de las dimensiones originales de altura y de peso, puede ser interpretada como una dimensión de "tamaño" y la fig 1.3 muestra dónde se encuentra cada persona sobre esta dimensión, así la persona B es más pequeña que A y C es mucho más pequeña que B.

Es claro que hay una infinidad de caminos para la elección de un vector base de la dimensión de "tamaño". Por ejemplo, $\mathbf{c} = [5\ 2]^T$ es otro vector base y las coordenadas de los tres puntos con respecto a **c** son 35, 34 y 30 respectivamente.

Similarmente la base estandard en el espacio de 2 dimensiones es solo una, de un infinito de posibles bases.

Considérese otro ejemplo en el espacio de dos dimensiones, donde los tres vectores

$$\mathbf{y}_1 = [1500\ 2000]^T \quad \mathbf{y}_2 = [2000\ 1000]^T \quad \mathbf{y}_3 = [2200\ 600]^T$$

representan las exportaciones e importaciones de cierto producto en tres años consecutivos.

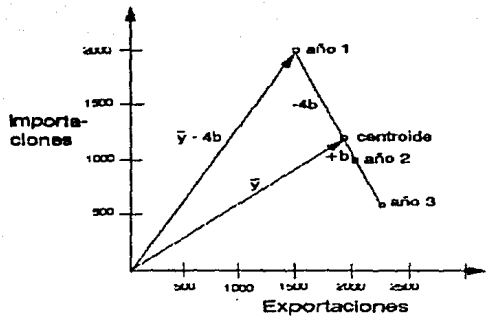


Fig 1.4

Estos tres puntos están sobre una linea recta, pero, ésta no pasa por el origen y los vectores no pueden ser expresados, directamente, como múltiplo de un vector base.

Entonces se usará el hecho de que cualquier vector puede ser expresado como la suma de un vector fijo desde el origen mas un múltiplo de un vector a lo largo de la recta.

El vector fijo a elegir es el centroide (vector de medias)

$$\bar{\mathbf{y}} = [(1500+2000+2200)/3,\ (2000+1000+600)/3]^T =$$

$$[1900\ \ 1200]^T$$

El vector **b** a lo largo de la recta puede ser elegido como $\mathbf{b} = [100\ -200]^T$ y así los tres puntos y_1, y_2 e y_3 pueden expresarse como

$$y_1 = \bar{y} - 4b \qquad y_2 = \bar{y} + b \qquad y_3 = \bar{y} + 3b$$

Alternativamente, se expresan los vectores como desviaciones desde su centroide

$$z_1 = y_1 - \bar{y} \qquad z_2 = y_2 - \bar{y} \qquad z_3 = y_3 - \bar{y}$$

Este "centrado" da como resultado que el origen, sea el centroide de los vectores z_1, z_2 y z_3; y éstos sean múltiplos del vector base **b**.

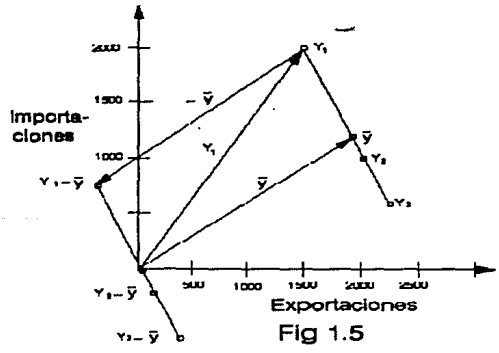


Fig 1.5

La representación en un espacio de una dimensión, de los tres puntos está dada en la fig. 1.6., y las coordenadas de los puntos centrados son -4, 1, 3 respectivamente.

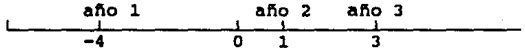


FIG 1.6

Esta dimensión podría interpretarse como una medida de "crecimiento" de la manufactura local del producto, con movimiento a la derecha indicando un "crecimiento" positivo. El vector base **b** ha sido etiquetado como una unidad de "crecimiento" positivo: un incremento en exportaciones de 100 acoplado de un decremento en importaciones de 200. Para una interpretación más real de la situación se necesita información adicional; como, por ejemplo, el consumo local del producto año con año, ya que, el decrecimiento en importaciones se debe a un bajo consumo.

Supóngase que se tiene un consumo local por año, de 2500, 2700, 3200 respectivamente, los vectores nuevos son de dimensión 3:

$\mathbf{y}_1 = [1500\ 2000\ 2500]^T$ $\mathbf{y}_2 = [2000\ 1000\ 2700]^T$ $\mathbf{y}_3 = [2200\ 600\ 3200]^T$

El centroide de estos puntos con la nueva dimensión es

$\bar{\mathbf{y}} = [1900\ 1200\ 2800]$

Como $\mathbf{b} = [100\ -200]^T = 100\mathbf{e}_1 - 200\mathbf{e}_2$ donde $\mathbf{e}_1$ y $\mathbf{e}_2$ son los vectores unitarios de los ejes de exportaciones e importaciones entonces,

$$\mathbf{y}_1 = [1500\ 2000\ 2500]^T = \bar{\mathbf{y}} - 4\mathbf{b} - 300\mathbf{e}_3$$

donde $\mathbf{e}_3$ es el vector unitario sobre el nuevo eje de consumo. Representando los puntos centrados en un espacio de dos dimensiones con ejes **b** y $\mathbf{e}_3$ da como resultado la fig 1.8. y las coordenadas de los puntos con respecto a estos ejes son [-4 -300], [1 -100] y [3 400].

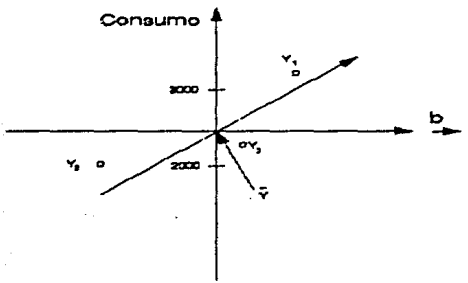


Fig 1.8

Estos puntos están muy próximos a una recta, definida por un vector **c**. Si se obtienen las desviaciones de los tres puntos con respecto a este vector, se podrían reducir los puntos originales en un espacio de dimensión 3 a un espacio de dimensión 1, desplegado a lo largo de un solo eje definido por el vector **c**. Este

vector es ahora la combinación del eje de consumo y el eje definido por **b**, en otras palabras, combinación de las tres dimensiones originales.

Este ejemplo, en tres dimensiones, es un simple ejemplo de reducción de dimensionalidad de un conjunto, *o nube*, de puntos en un espacio multidimensional. En aplicaciones actuales, usualmente se trabaja con puntos de en espacios de dimensión mucho más alta.

La pregunta obligada es ¿como encontrar un subespacio de menor dimensión el cual esté "más cerca" a la nube de puntos?. Lo cual es tratada en el Capítulo 2.

A.2 Asignación de pesos a ejes y puntos

Como se sabe, se puede definir distancia entre dos puntos a partir de la definición de producto escalar de vectores.

El producto escalar entre los vectores $\mathbf{a} = [a_1\ a_2 ... a_J]^T$ y $\mathbf{b} = [b_1\ b_2 ... b_J]^T$, se denota como

$$\langle \mathbf{a},\ \mathbf{b} \rangle = \sum_{j=1}^{J} a_j b_j = \mathbf{a}^T\mathbf{b}$$

Entonces la distancia entre **a** y **b** es

$$d(\mathbf{a},\mathbf{b}) = \left(\sum_{j=1}^{J} (a_j - b_j)^2 \right)^{\frac{1}{2}} = \langle \mathbf{a}-\mathbf{b},\ \mathbf{a}-\mathbf{b} \rangle^{\frac{1}{2}}$$

Estos resultados dependen, absolutamente, de un sistema perpendicular de coordenadas, ya que la base estandard de vectores $\mathbf{e}_1, \mathbf{e}_2 ..., \mathbf{e}_J$ son *ortogonales* ($\langle \mathbf{e}_j, \mathbf{e}_i \rangle = 0$ si $j \neq i$), y además estos vectores tienen longitud de una unidad,

normalizados, entonces la base estandard, elemental para estos resultados es una base *ortonormal*.

En la definición anterior de distancia se presupone que las dimensiones asociadas a cada dimensión son de la misma naturaleza. Se trata de longitudes medidas de acuerdo a una misma cantidad.

En casi todas las situaciones que se podrían presentar, a cada eje (dimensión) se le asocia una característica que implícitamente tiene unidad diferente.

Recordando el ejemplo de las medidas de altura y peso sobre diferentes personas, es claro que se está trabando con unidades completamente diferentes (cm, para altura y kg, para peso). La altura siempre tendrá un valor mucho más alto al del peso, así la medida de altura contribuirá más, relativamente, al resultado de la distancia; pero si la unidad de altura es m, entonces, el peso es el que contribuirá más en el resultado de la distancia.

Claramente, no es conveniente que las distancias dependan directamente de la elección de la escala de las medidas.

Es necesario, entonces, "ponderar" cada eje con respecto a las diferencias de las variables y así trabajar en un Espacio Euclidiano con Peso. Comúnmente, se dividen las medidas por sus respectivas desviaciones estándares. Por ejemplo, suponer que la desviación estandard de la altura y peso de una muestra de gente es 30 cm y 10 kg respectivamente. Sean **x** e **y** los vectores que representan las medidas de dos personas, los vectores estandarizados son

$\mathbf{x_s} = [x_1/30\ x_2/10]^T$ y $\mathbf{y_s} = [y_1/30\ \ y_2/10]^T$ y su producto escalar $\langle \mathbf{x_s}, \mathbf{y_s} \rangle = x_1y_1/30^2 + x_2y_2/10^2$.

Este producto escalar puede ser escrito como

$$\langle \mathbf{x_s}, \mathbf{y_s} \rangle = x_1y_1/30^2 + x_2y_2/10^2 = \mathbf{x}^T D_s^{-2} \mathbf{y}^T$$

donde

$$D_s^{-2} = \begin{bmatrix} 1/30^2 & 0 \\ 0 & 1/10^2 \end{bmatrix}$$

es la matriz diagonal de las inversas de la varianzas, asociada con el Espacio Euclidiano Pesado.

Si se está trabajando con puntos en un espacio de dimensión J, la matriz asociada (comunmente llamada métrica), será de JxJ, simétrica, y definida positiva.

Las métricas más utilizadas son las diagonales, ya que son de fácil interpretación.

En general, se define el producto escalar como

$\mathbf{x}^T$ Dp $\mathbf{y} = \Sigma_j p_j x_j y_j$, donde $p_1 \ldots p_J$ son números reales positivos elementos de la diagonal de Dp (Dp = diag (p_j)) que significan; el peso asociado a cada dimensión del espacio.

Entonces la distancia al cuadrado se expresa como

$d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T$ Dp $(\mathbf{x} - \mathbf{y}) = \Sigma_j p_j(x_j - y_j)^2$.

Una base, $\mathbf{b_1} \ldots \mathbf{b_j}$, ortonormal para este espacio (ortonormal con respecto a la métrica Dp) debe satisfacer

$\mathbf{b}_i^T$ Dp $\mathbf{b}_j = 0$, si $i \neq j$ (ortogonal) y

$\mathbf{b}_i^T$ Dp $\mathbf{b}_i = 1$ (normal).

Lo anterior referente a pesos asignados a ejes, y tratando lo referente a pesos asignados a puntos, se pretende que los métodos usados puedan adaptarse a situaciones prácticas, entonces se considera, que las observaciones no necesariamente tengan la misma importancia. En la estadística clásica cada observación es, generalmente, considerada como aportadora de la misma cantidad de información. En la realidad, se puede privilegiar a algunas observaciones más que otras en función de su importancia con respecto a cierto criterio. De tal forma, a cada observación se le asociará un *peso* w_i tal que; $w_i > 0$. Es claro, que si todas las observaciones se consideren con la misma importancia, entonces $w_i = 1/$(total de obs.).

Generalmente estos pesos de puntos serán almacenados en una matriz diagonal (matriz de pesos asignados a I puntos) $Dw_{I \times I}$.

Así, el centroide del conjunto de puntos, $I = \{x_1\ x_2\ \ldots\ x_I\}$, con pesos $w_1\ w_2 \ldots w_I$ se define como

$$\bar{x} = \sum_{i=1}^{I} w_i x_i / \sum_{i=1}^{I} w_i .$$

Normalmente, se supondrá que $\sum_i w_i = 1$, a menos que se especifique lo contrario, en este caso la expresión para el centroide es

$$\bar{x} = \sum_{i=1}^{I} w_i x_i$$

APENDICE B
PROPUESTA DE UN PROGRAMA EN COMPUTADORA

Para realizar los Análisis de Correspondencias Simétrico y No-simétrico, es decir, obtener resultados y un desplegado de puntos, se propone el programa ANACORR, el cual realiza los análisis, en un ambiente de microcomputadora amigable y de fácil operación. ANACORR está programado en el lenguaje de Turbo Pascal (Versión 5.0), apoyado en el manejador de bases de datos Btrieve (Versión 5.1) y basado en la teoría presentada en los capítulos anteriores.

ANACORR está basado en un ambiente de menús en los cuales, para ejecutar alguna opción se puede lograr de dos maneras:

a) presionar las teclas de flechas, en la dirección de la opción deseada hasta llegar a ésta y después presionar la tecla <ENTER>,

b) Presionar la tecla de la letra de la opción que tiene el color diferente.

B.1 Instalación

Es recomendable que el programa ANACORR se pueda ejecutar en un disco duro, es por esto que cuenta con una herramienta para instalarlo del disco flexible al disco duro, para lo cual se deben de seguir los siguientes pasos:

1) colocarse en el lugar, del disco duro, donde será instalado el programa. Por ejemplo si se quiere trabajar en el directorio PRUEBA, se debe estar en C:\PRUEBA>,

2) cambiarse al disco flexible (generalmente A: o B:) y elegir el subdirectorio ANACORR (A:>CD ANACORR),

3) teclear el comando, A:\ANACORR\INSTALA C:

La herramienta INSTALA copiará todo lo necesario para trabajar con ANACORR sobre el directorio C:\PRUEBA.

B.2 ANACORR

Para entrar a ANACORR se deben seguir dos pasos:

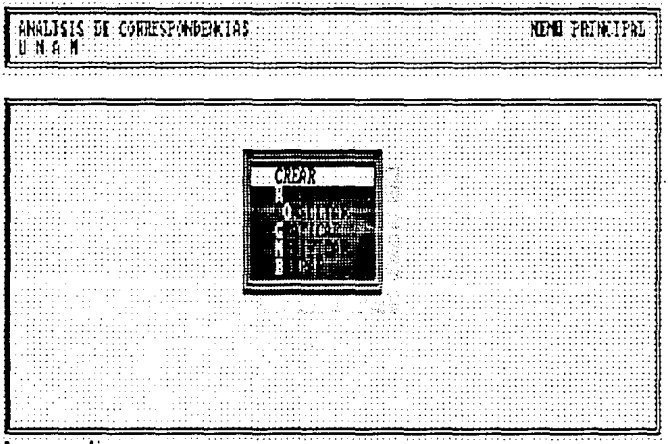
1) Cambiarse al subdirectorio ANACORR

CD ANACORR

2) Teclear ANACORR

\ANACORR>ANACORR

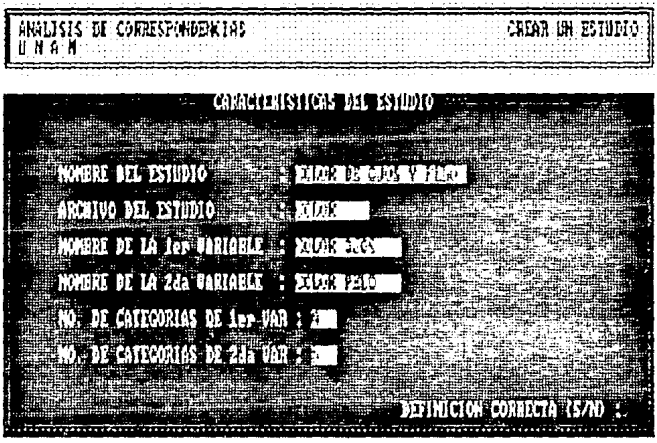
Una vez realizado esto, ya se está dentro del programa y éste desplegará la siguiente pantalla, correspondiente al MENU PRINCIPAL.

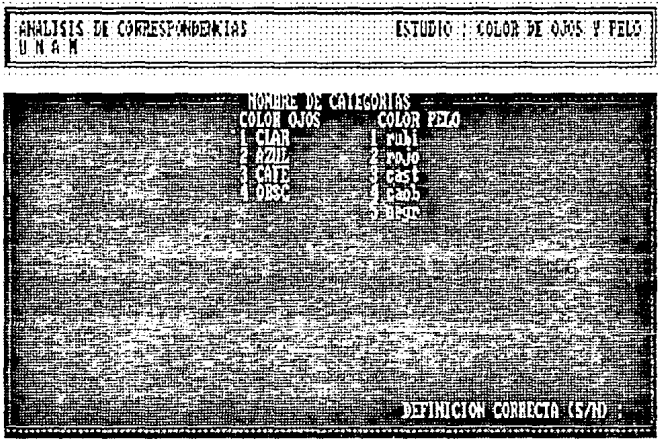


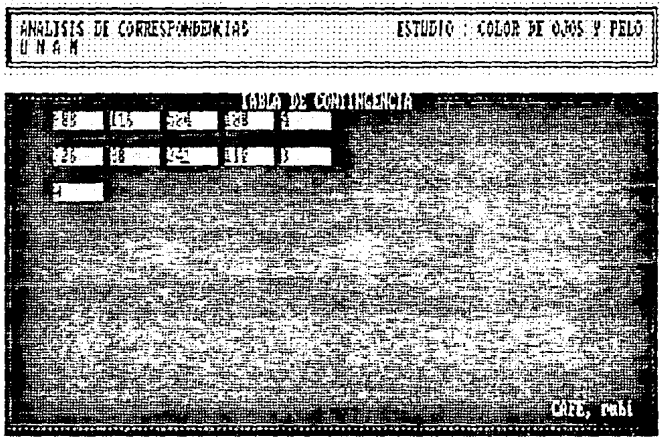
Crear estudio

A continuación se explicará a detalle como pueden ser utilizadas la opciones de este menú.

CREAR. Esta opción permite alimentarle al programa la información necesaria de un estudio determinado, para después realizar, alguno de los análisis. La opción está formada por tres pantallas en las cuales se puede apreciar la información requerida por ANACORR.


ANALISIS DE CORRESPONDENCIAS
U N A M
CREAR UN ESTUDIO
CARACTERISTICAS DEL ESTUDIO
NOMBRE DEL ESTUDIO
ARCHIVO DEL ESTUDIO
NOMBRE DE LA 1er VARIABLE
NOMBRE DE LA 2da VARIABLE
NO. DE CATEGORIAS DE 1er VAR
NO. DE CATEGORIAS DE 2da VAR
DEFINICION CORRECTA (S/N)


ANALISIS DE CORRESPONDENCIAS
U N A M
ESTUDIO : COLOR DE OJOS Y PELO
NOMBRE DE CATEGORIAS
COLOR OJOS
1 CLAR
2 AZUL
3 CAFE
4 OBSC
COLOR PELO
1 rubi
2 rojo
3 CAS1
4 CAS2
5 NEG
DEFINICION CORRECTA (S/N)



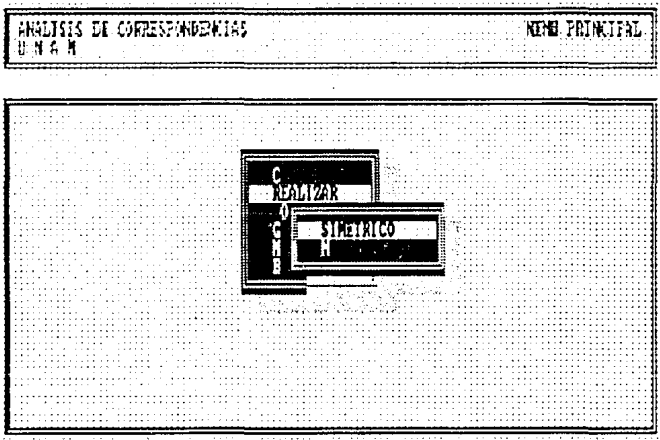
En la primer pantalla se requieren datos generales del estudio, tales como: nombre del estudio, nombre del archivo en el cual se almacenará la información, el programa no permitirá que el nombre de este archivo ya se le haya asignado a otro; nombre y número de categorías de las variables, para el caso no-simétrico se considerará la primer variable como respuesta y la segunda como explicativa. La información que se requiere en la segunda pantalla es, básicamente, el nombre o una clave para identificar las categorías de cada variable. La última pantalla consta de la tabla de contingencia.

Al final de cada pantalla, se pregunta si la información que se ha tecleado es correcta, si ésta no lo es se puede volver a teclear desde el principio de la

pantalla, además con la tecla <ESC> en cualquier momento se puede cancelar la creación del estudio.

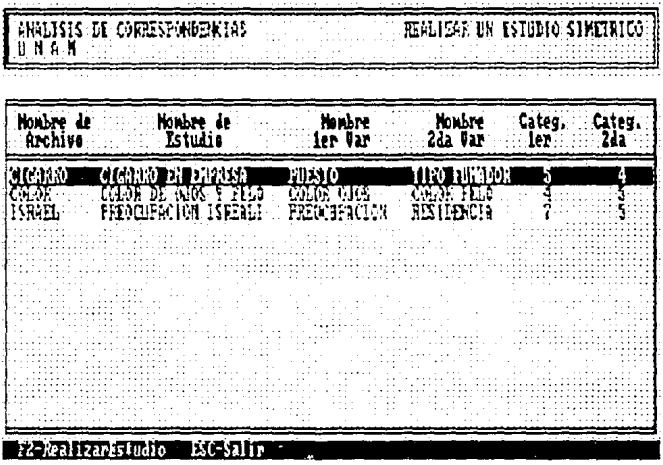
REALIZAR. Una vez creado un estudio, el siguiente paso sería aplicarle alguno de los análisis, éste es el objetivo de esta opción.

Al ejecutar la opción de REALIZAR, se permitirá elegir el tipo de análisis (SIMETRICO o NO-SIMETRICO).



Realizar Análisis de Correspondencias Simétrico

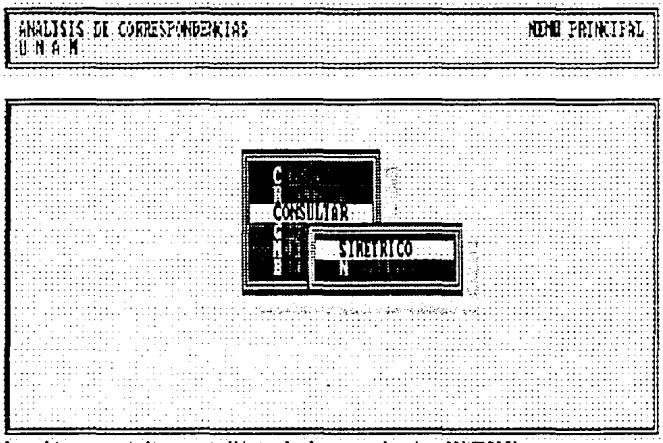
Ya elegido el tipo de análisis se mostrará la lista de todos los estudios creados a los cuales se les puede realizar este tipo de análisis.



Cuando se haya seleccionado el estudio, para llevar a cabo el análisis solo se tiene que presionar la tecla <F2>.

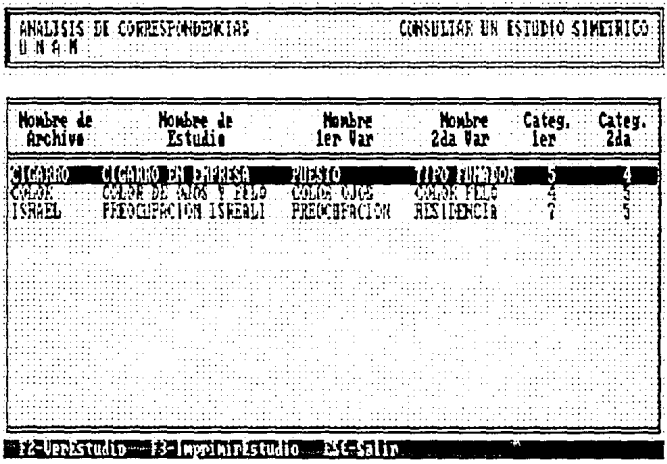
CONSULTAR. Permite consultar los resultados de un estudio al cual se le aplicó el análisis requerido.

Al ejecutar esta opción se tendrá que elegir el tipo de análisis que se desea consultar.



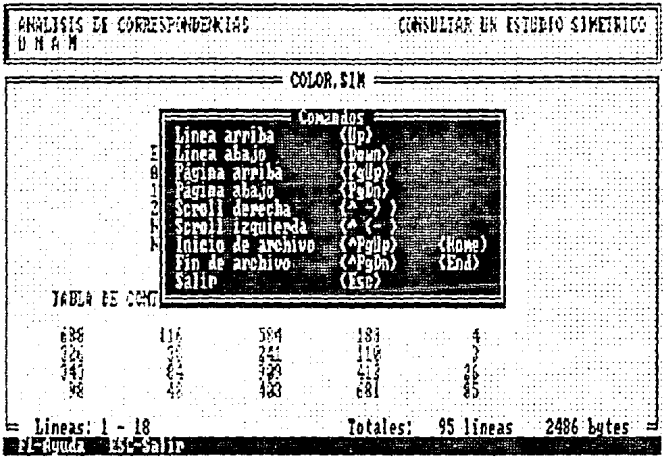
Consultar un estudio con Análisis de Correspondencias SIMETRICO

Cuando se elija el análisis requerido se mostrarán todos los estudios a los cuales se les ha realizado éste análisis.



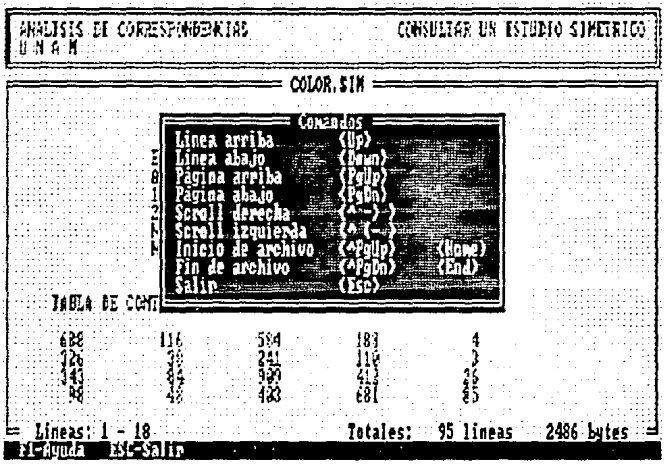
Para consultar el estudio en la pantalla se tiene que seleccionar éste y después presionar la tecla <F2>, para imprimir los resultados, solo se tiene que presionar la tecla <F3>.

Los comandos que se pueden realizar para consultar los resultados en la pantalla, pueden observarse al presionar la tecla de ayuda <F1>.



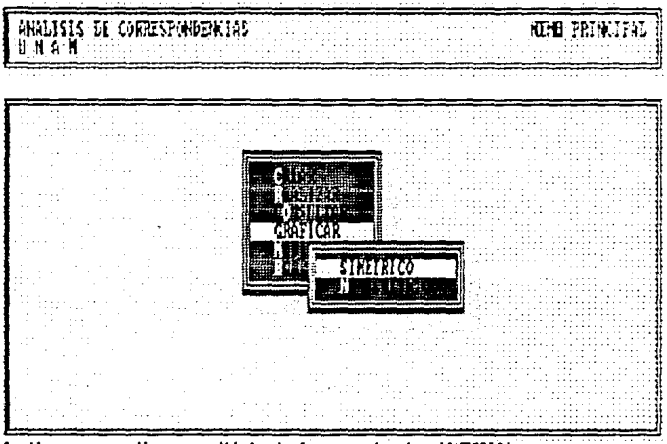
GRAFICAR. Con esta opción se puede visualizar la representación en el plano de las categorías, con la posibilidad de elegir los ejes principales por medio de los cuales se desea hacer dicha representación.

Al igual que en la opción de CONSULTAR, después de haber elegido el tipo de análisis requerido se mostrarán todos los estudios a los cuales se les ha realizado este análisis. Esto lo muestran las siguientes pantallas.

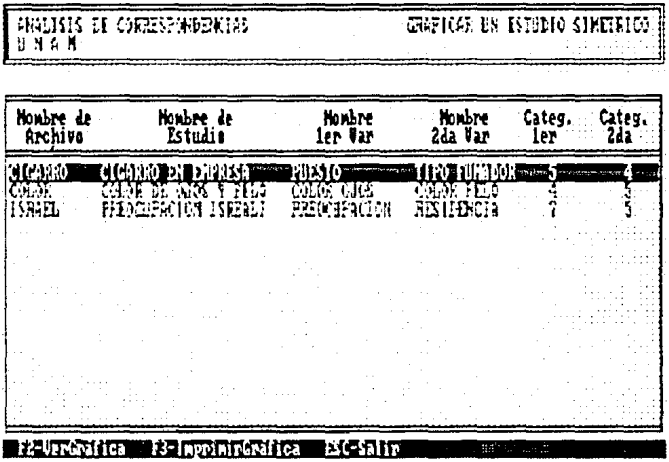


GRAFICAR. Con esta opción se puede visualizar la representación en el plano de las categorías, con la posibilidad de elegir los ejes principales por medio de los cuales se desea hacer dicha representación.

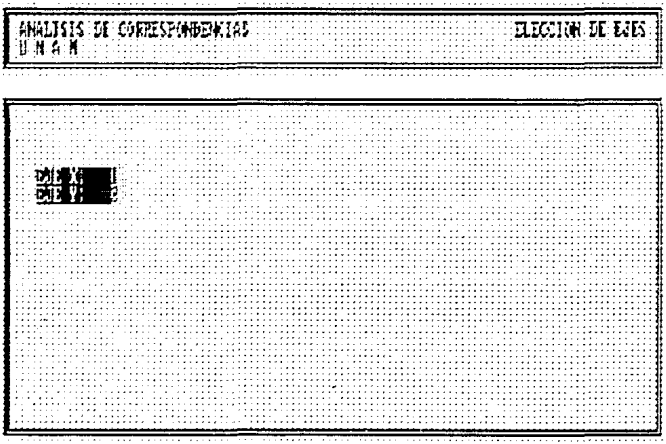
Al igual que en la opción de CONSULTAR, después de haber elegido el tipo de análisis requerido se mostrarán todos los estudios a los cuales se les ha realizado este análisis. Esto lo muestran las siguientes pantallas.



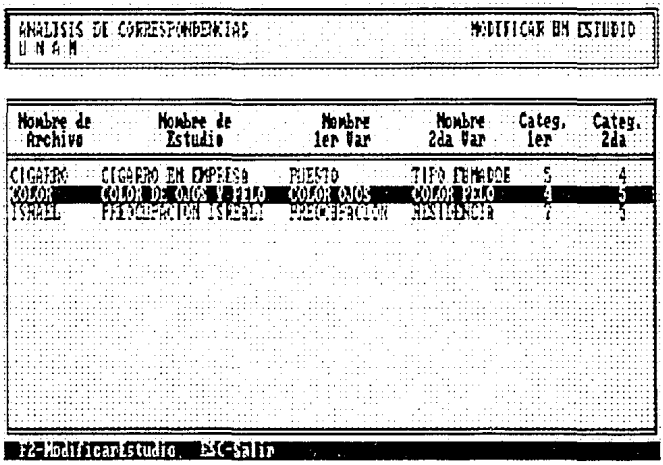
Graficar un estudio con análisis de Correspondencias SIMETRICO



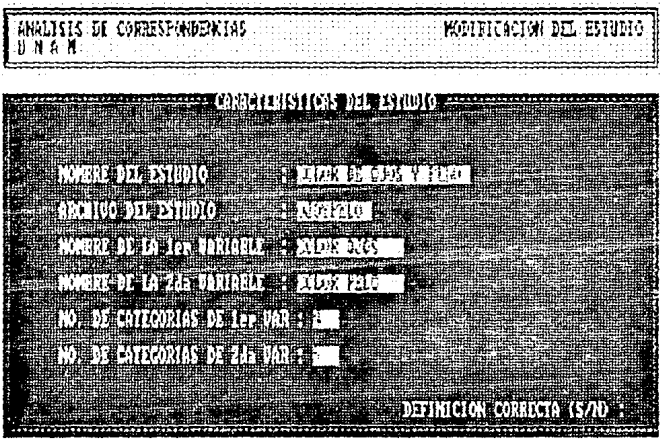
Como se puede observar en esta pantalla la representación se puede ver en la pantalla (tecla <F2>) o se puede enviar a la impresora(tecla <F3>). Cuando ya se ha elegido el estudio a graficar y en caso que el número de ejes principales sea mayor que dos, se tendrá que elegir en cual de los ejes se realizará la representación (recordar que la mejor representación se obtiene con los dos primeros ejes).

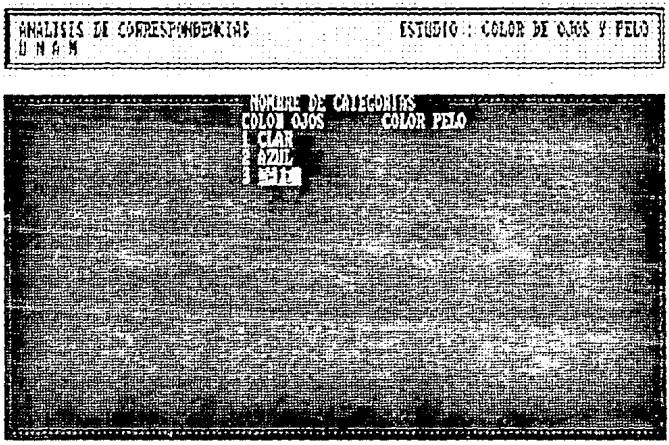


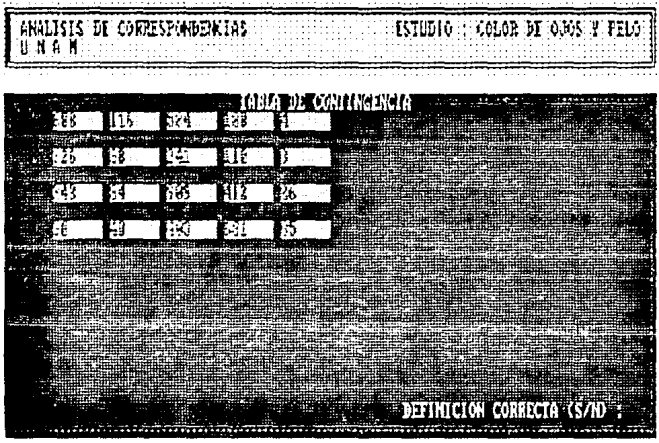
MODIFICAR. Cuando a un estudio creado es necesario hacerle algún cambio se tiene que recurrir a la opción de MODIFICAR, la cual, cuando se haya seleccionado, mostrará todos los estudios creados con la posibilidad de hacerle algún cambio presionando la tecla <F2>.



La opción de MODIFICAR consta de las mismas pantallas que la opción de CREAR con la diferencia de que éstas tienen la información alimentada cuando se creó el estudio.

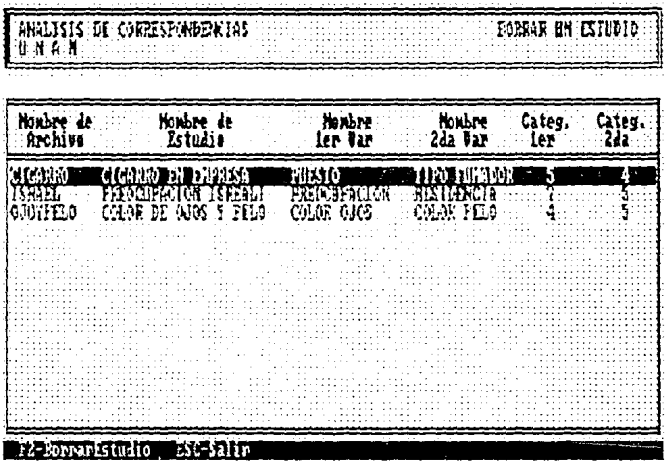

ANALISIS DE CORRESPONDENCIAS
U N A M
MODIFICACION DEL ESTUDIO
CARACTERISTICAS DEL ESTUDIO
NOMBRE DEL ESTUDIO : Libro de Ojos y Filas
ARCHIVO DEL ESTUDIO : Acerolo
NOMBRE DE LA 1er VARIABLE : Ojos
NOMBRE DE LA 2da VARIABLE : Filas
NO. DE CATEGORIAS DE 1er VAR : 3
NO. DE CATEGORIAS DE 2da VAR : 3
DEFINICION CORRECTA (S/N) :


ANALISIS DE CORRESPONDENCIAS
U N A M
ESTUDIO : COLOR DE OJOS Y PELO
NOMBRE DE CATEGORIAS
COLOR OJOS
COLOR PELO
1 CLAR
2 AZUL
3 CAFE



Es importante mencionar, que cuando se haya realizado algún análisis a un estudio que se modifica éste se considerará como un estudio al cual no se le ha realizado ningún tipo de análisis.

BORRAR. Esta opción se usa cuando ya no se desea considerar algún estudio creado y se decide eliminarlo.


ANALISIS DE CORRESPONDENCIAS
U N A M
BORRAR UN ESTUDIO
Nombre de Archivo
Nombre de Estudio
Nombre 1er Var
Nombre 2da Var
Categ. 1er
Categ. 2da
CICARRO
CICARRO EN EMPRESA
PUESTO
TIPO FUNCION
3
4
ISRAEL
PERCEPCION ISRAELI
PERCEPCION
RESILIENCIA
?
5
OJOSPELO
COLOR DE OJOS Y PELO
COLOR OJOS
COLOR PELO
4
5
F2-Borrar Estudio
ESC-Salir

Referencias

D'Ambra Luigi, Lauro Natale (1982). Principal components analysis onto reference subespaces.

D'Ambra Luigi, Lauro Natale (1983). Non-symmetrical correspondence analysis.

D'Ambra Luigi, Lauro Natale (1984). Non-symmetrical analysis of three-way contingency tables.

Greenacre Micheal J. (1984). "Theory and applications of correspondence analysis". Academic Press, London.

Hernández Cid Rubén (1988). Notas del curso de Análisis de datos.

Lang Serge (1976). "Algebra lineal". Fondo Educativo Interamericano, México.

Leach Chris (1982). "Fundamentos de estadística. Enfoque no paramétrico para ciencias sociales". LIMUSA, México.

Mendenhall William, Scheaffer Richard L., Wackerly Dennis D. (1986). "Estadística matemática con aplicaciones". Grupo Editorial Iberoamerica, México.

Paquetes usados para el desarrollo del programa ANACORR:

Btrieve (Ver 5.1).

Turbo Pascal (Ver 5.0).

Turbo Pascal, Graphix Toolbox (Ver 4.0).

Turbo Pascal, Numerical Toolbox (Ver 4.0).

Turbo Professional (Ver 4.0).

Xtrieve Plus (Ver 3.0).