UN ENFOQUE UNIFICADO PARA TÉCNICAS DE REPRESENTACIÓN EUCLIDIANA

Elina Miret Barroso[1], Gladys Linares Fleites y María V. Mederos Brú
Facultad de Matemática y Computación, Universidad de La Habana

RESUMEN
En este trabajo se estudian algunos métodos del Escalamiento Multidimensional (EM) buscando nexos con las técnicas exploratorias que están incluidas en la teoría general de las Coordenadas Canónicas de R. Rao [1995] y se encuentra un enfoque unificado para todas las técnicas de representación euclidiana que facilita, desde el punto de vista metodológico, la enseñanza del EM como método exploratorio. A partir de una estrategia computacional programada en MATLAB, que incluye dos métodos de Escalamiento Multidimensional, se reconstruye el mapa de Cuba.

ABSTRACT
In this paper several methods of Multidimensional Scaling (MDS) are studied to find nexuses with the exploratory techniques included in the general theory of canonical coordinates given by R. Rao in 1995 and .an unified focus for all techniques with Euclidean representation approach to facilitate methodologically the MDS teaching as an exploratory method. Using a computational strategy programmed in MATLAB language, including two MDS methods, the Cuban map is reconstructed.

INTRODUCCIÓN

Algunos fenómenos de la realidad no pueden reflejarse directamente como observaciones por lo que es necesario observar algunas relaciones entre los objetos para poder estudiarlos. Los coeficientes de similitudes y disimilitudes así como las distancias entre los objetos son algunas medidas de semejanza para expresar estas relaciones en un fenómeno en estudio.

Los procedimientos estadísticos que trabajan con medidas de semejanzas en el Análisis de Datos se agrupan bajo el nombre de Escalamiento Multidimensional (EM), conocido en inglés con este mismo nombre: *Multidimensional Scaling* (MDS).

El problema del EM consiste en representar disimilitudes entre objetos o individuos como distancias entre puntos en un espacio de dimensión reducida. El punto de partida es un conjunto Ω de n objetos y una matriz $\Delta = (\delta_{ij})_n$ de disimilitudes entre los objetos. El resultado final es una configuración de puntos que se identifican con los n objetos en un espacio euclidiano de baja dimensión, de forma que las distancias d_{ij} entre los puntos representan lo "mejor posible" las disimilitudes δij iniciales.

Cuadras [1981, 1989] define una disimilitud δ sobre un conjunto Ω como una aplicación de Ω x Ω en R tal que a cada par (i, j) asigna el número real $\delta(i, j) = \delta_{ij}$ que satisface:

$$\delta_{ij} \geq 0$$

$$\delta_{ii} = 0$$

$$\delta_{ij} = \delta_{ji} \ , \ \forall i,j \ .$$

Un procedimiento sencillo para obtener similitudes puede ser el siguiente:

Cada individuo u objeto de una población está caracterizado por las presencias y ausencias de ciertas cualidades que le fueron medidas. Suponiendo que se midieron n cualidades, una información útil con respecto a los individuos u objetos que se analizan es el número de cualidades presentes en cada individuo u objeto con respecto a las n cualidades medidas a todos. Llamemos “a” al número de cualidades presentes comunes a dos individuos, “b” al número de cualidades ausentes comunes a los dos individuos, “c” al número de cualidades presentes en el primer individuo y ausentes en el segundo y “d” al número de cualidades presentes en el segundo individuo y ausentes en el primero. Entonces n = a + b + c + d.

E-mail: [1]elina@matcom.uh.cu

Existen diferentes expresiones para calcular similitudes a partir de a, b, c y d. Algunas de las citadas por Cuadras [1981, 1989], Cox & Cox [1994] y Borg & Groenen [1997] son el coeficiente de similitud de Rao y Russell [a/n] y el que propone Jaccard [a/(a + b + c)].

Existen coeficientes de similitud que cumplen con las propiedades de una disimilitud, entre estos, los de Sokal y Michener y de Rao y Russell, analizados por Zhang y Srihary [2003].

En el epígrafe 1 se esboza la teoría general de coordenadas canónicas de Rao [1995], se presentan los métodos del EM, en el epígrafe 2, luego en el tercer epígrafe se ofrece un enfoque unificado de la perdida de información para todos los métodos de representación euclidiana. En el epígrafe 4 se dan los gráficos de la aplicación de una estrategia combinada de dos métodos de EM para reconstruir el mapa de Cuba.

1. TEORÍA GENERAL DE LAS COORDENADAS CANÓNICAS (R. Rao, 1995)

Sea $X = (X_1:...:X_n)$ una matriz de datos.

X_i (Perfil de la población i-ésima) representa la medida de p variables a la i-ésima población o individuo.

Cada X_i puede representarse como un punto del espacio vectorial Rp con el producto interno $<x,y> = x^tMy$, $\forall x,y \in R^p$, siendo M una matriz definida positiva y la norma asociada: $\| x \| = <x,x>^{1/2} = (x^tMx)^{1/2}$, $x \in Rp$.

Sea $W = (w_i)_n$, donde cada w_i es un peso asociado a la población o individuo i-ésimo.

El espacio (Rp, M, W) se llama MW-espacio o espacio básico métrico.

Problema a resolver:

Encontrar una matriz de tamaño kxn $Y_{(k)} = (Y_1:...:Y_n)$ con $k < p$ cuyas columnas Y_i representen a los correspondientes perfiles iniciales X_i en un espacio euclidiano k-dimensional R^k con el producto interno usual y las distancias $\hat{d}_{ij} = d(Y_i, Y_j)$ sean tales que "preserven en lo posible" las correspondientes relaciones de distancia iniciales $d_{ij} = d(X_i, X_j)$ entre los perfiles del espacio básico métrico.

Para este propósito se necesita una función de pérdida que mida la información que se pierde al reducir la dimensión.

Rao define cierta transformación de X en el espacio básico inicial que contiene la información correspondiente a las distancias entre las poblaciones o individuos en estudio en dicho espacio, llamada matriz de configuración inicial, con el propósito de construir su función de pérdida.

La matriz de configuración inicial en el espacio básico R^p viene dada por la expresión: $C_i = (X - \xi 1^t)^t M(X - \xi 1^t)$, siendo ξ cierto vector de referencia y 1 un vector formado por unos.

Análogamente, la matriz de configuración final en el espacio Rk ($k < p$) tiene la expresión: $C_f = Y^tY$.

Entonces, la búsqueda de Y en R^k se plantea en términos de las matrices de configuración de la siguiente forma:

Encontrar $Y(k) = (Y_1:...:Y_n)$ en R^k que resuelva el problema de optimización:

$$\min_{Y_{(k)} \in R^k} \| C_i - C_f \|$$

Teorema 1: (Rao [1995])

Sea la s.v.d. de la matriz: $M^{1/2}(X - \xi 1^t)W^{1/2} = \sigma_1 U_1 V_1^t + ... + \sigma_p U_p V_p^t$, con valores singulares $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p$, donde $M^{1/2}$ y $W^{1/2}$ son las raíces cuadradas simétricas de M y W respectivamente. Entonces, la solución óptima al problema de optimización: $\min_{Y_{(k)} \in R^k} \| C_i - C_f \|$, es:

$$Y = \begin{pmatrix} \sigma_1 V_1^t W^{-1/2} \\ \vdots \\ \sigma_k V_k^t W^{-1/2} \end{pmatrix}$$

A las componentes del i-ésimo n-vector $\sigma_i W^{-1/2} V_i$ se les llama *coordenadas canónicas* en la i-ésima dimensión para los diferentes individuos o poblaciones.

La elección del espacio básico y su correspondiente función de distancias d, así como de la función de pérdida, dependen del problema concreto que se investiga y deben construirse atendiendo a consideraciones prácticas.

2. TIPOS DE ESCALAMIENTO MULTIDIMENSIONAL

En la literatura, existen dos criterios bien diferenciados en cuanto a la clasificación de los métodos del EM:

En uno, se considera que el escalamiento es métrico cuando las disimilitudes permanecen fijas en la función objetivo y no métrico cuando las mismas son variables (Trosset [1993]).

El otro criterio, más difundido que el primero, es el relativo al tipo de transformación que se haga a las disimilitudes en la función objetivo Borg & Groenen [1997]. Se tiene escalamiento métrico cuando se trabaja con una función paramétrica de las disimilitudes y escalamiento no métrico cuando se emplea una función monótona. En este trabajo se sigue la segunda clasificación.

Escalamiento Métrico

El caso más simple de los métodos de Escalamiento Multidimensional es el Escalamiento Clásico (EC) o Análisis de Coordenadas Principales. **Escalamiento Clásico**.

El caso más simple de los métodos de EM es el Escalamiento Clásico (EC) también conocido por Análisis de Coordenadas Principales (Ver Mardia **et al**. [1979]). Se hace EC Euclídeo si la matriz de disimilitudes inicial Δ es de distancias y EC No Euclídeo si no lo es. Miret [2005] incluye el EC Euclídeo en la teoría de Rao y propone la inclusión del EC No Euclídeo utilizando elementos de la teoría de espacios vectoriales con métrica indefinida.

El procedimiento del EC para encontrar la solución $Y_{(k)}$ consiste en:

1. A partir de $\Delta = (\delta_{ij})_n$, construir $A = (a_{ij})_n$, siendo $a_{ij} = (-1/2)\delta_{ij}^2$.

2. Obtener B de A. ($B = HAH$, donde $H = I_n - (1/n)1_n 1_n^t$, siendo 1_n el vector columna formado por n unos.

3. Extraer los k valores propios (si se hace EC Euclídeo) o singulares (si se hace EC No Euclídeo) estrictamente positivos y más grandes $\lambda_1,...,\lambda_k$ de la descomposición espectral de $B = V\Sigma V^t$ y sus correspondientes vectores propios normalizados. Si se denota por $V(k) = (V_1,...,V_k)$ a la matriz de las k primeras columnas de V y $\Sigma_{(k)}^{1/2} = \text{diag}(\lambda_1^{1/2},..., \lambda_1^{1/2})$.

Entonces: $Y_{(k)} = B_{(k)}^{1/2} = \Sigma_{(k)}^{1/2} V_{(k)}^t$.

Para averiguar si $\Delta = (\delta_{ij})_n$ es euclídea basta analizar si la matriz $B = (b_{ij})_n$ del algoritmo anterior, es semi-definida positiva de rango $p \geq k$ (donde k es la dimensión del subespacio de Rp en el cual se quiere representar a los individuos u objetos).

La construcción del vector $Y_{(k)}$ con las nuevas coordenadas se puede plantear en términos de la búsqueda de la solución de un problema de optimización donde la función objetivo depende de la diferencia de las d_{ij} iniciales transformadas (B) y las $d_{ij(k)}$ finales transformadas ($B_{(k)}$). El óptimo de esta función coincide con la solución algebraica ofrecida en el algoritmo anterior (Borg & Groenen [1997]). Carroll & Chang, en 1972, dieron a conocer esta función objetivo que recibe el nombre de STRAIN. Luego, el problema de búsqueda

del mejor subespacio de representación puede expresarse en términos del problema de optimización del STRAIN puede expresarse de la forma siguiente:

$$\min_{Y_{(k)} \in R^k} \left\| B - B(k) \right\|^2 \qquad (1)$$

donde B y $B^{(k)}$ se calculan según el algoritmo antes descrito.

Tarasaga & Trosset [1998] presentan el problema de optimización del EC como sigue:

Sean los siguientes conjuntos en el espacio $S_n(R)$ de las matrices simétricas reales de orden n:

$\Delta_n(p)$: el conjunto de todas las matrices de disimilitudes de orden n.

$\Lambda_n(p)$: el conjunto de todas las matrices de disimilitudes que son euclidianas o de distancias, es decir, aquellas para las cuales existen $X_1,..,X_n \in R^p$ tales que:

$$d_{ij} = \left\| X_i - X_j \right\|_F^2$$

$\Omega_n(p)$: el conjunto de todas las matrices semi-definidas positivas de rango $k \leq p$.

Es obvio que, $\Lambda_n(p) \subset \Delta_n(p)$, pues toda matriz de distancias es de disimilitudes.

Entre los conjuntos $\Lambda_n(p)$ y $\Omega_n(p)$ se tiene la siguiente equivalencia que establece una generalización del criterio anterior de Mardia [1979] y que determina cuando una matriz de disimilitudes es euclídea o de distancias.

Teorema 2.

$D = (d_{ij})_n \in \Lambda_n(p)$ si y sólo si existen $B \in \Omega n(p)$ y una función lineal $\kappa:\Omega_n(p) \rightarrow \Lambda_n(p)$ tales que: $\kappa(B) = D$.

La demostración de este resultado aparece en el trabajo de Tarazaga & Trosset [1998].

Si $X = (X_1,...,X_n)$ es la matriz de configuración inicial, entonces $B = X^tX$, donde $D = (d_{ij})_n$ cumple que $d_{ij} = \left\| X_i - X_j \right\|_F^2$.

El problema de optimización a resolver por el EC Euclídeo tiene la forma:

$$\left\| \Delta * \Delta - D * D \right\|_F^2 \qquad (2)$$

donde:

$||.||_F$ denota la norma de Frobenius o norma inducida por el producto escalar matricial $<A, B>_F = tr(B^tA)$,

$D*D$ y $\Delta*\Delta$ son las matrices de orden n formadas por las distancias al cuadrado y las disimilitudes al cuadrado, respectivamente; es decir, $D*D = (d_{ij}^2)_n$ y $\Delta*\Delta = (d_{ij}^2)_n$ y la operación $*$ se llama producto de Hadamard de la matriz consigo misma,

τ: $R^{nxn} \rightarrow R^{nxn}$ es el operador de doble centrado, definido para cada matriz $A = (a_{ij})_n$ como $\tau(A) = B$, es decir a cada a_{ij} de A le hace corresponder $b_{ij} = - ½ (a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..})$, siendo los a_{ij} resultado del producto de Hadamard de una matriz P consigo misma, es decir, en A, cada $a_{ij} = p_{ij}^2$. Si P = D, entonces $\tau(D*D) = B$.

La expresión (2) es otra forma de expresar la función STRAIN de Carroll & Chang.

Escalamiento Mínimo Cuadrático

Los métodos de Escalamiento Mínimo Cuadrático construyen ciertas funciones f(δij) de las disimilitudes y como criterio de ajuste emplean una transformación L de la suma de cuadrados de los errores $e_{ij} = f(\delta_{ij}) - g(d_{ij}(X))$ dada, de forma general, por:

$$L(e_{ij}) = \sum_{i,j} w_{ij}\left[f(\delta_{ij}) - g(d_{ij}(X))\right]^2 \qquad (3)$$

Emplean procesos de optimización numérica que requieren de una matriz de configuración inicial X^0. Usualmente toman la solución del EC como dicho punto de partida.

Si en la expresión (3) se sustituye la función g por la identidad y todos los pesos por 1, la función resultante se llama raw-Stress y se denota por:

$$\sigma_r = \sigma_r(X) = \sum_{(i,j)} (d_{ij}(X) - f(\delta_{ij}))^2$$

Borg & Groenen [1997] plantean que la función raw-Stress no es muy informativa en la práctica debido a su variabilidad por cambios de escala en las disimilitudes. Para evitar esta dependencia de escala, sugieren normalizar el raw-Stress como sigue:

$$\sigma_1^2 = \sigma_1^2(X) = \sum_{(i,j)} \left[f(\delta_{ij}) - d_{ij}(X)\right]^2 / \sum_{i,j} d_{ij}^2(X) = \sigma_r(X) / d_{ij}^2(X)$$

Como σ_1^2 casi siempre toma valores muy pequeños, suele trabajarse con su raíz cuadrada. Entonces, σ_1 se identifica con la función Stress-1 de Kruskal, que queda explícitamente:

$$\text{Stress-1} = = \left[\sum_{(i,j)} \left[f(\delta_{iij}) - d_{ij}(X)\right]^2 / \sum_{i,j} d_{ij}^2(X)\right]^{1/2}$$

Si en la fórmula de σ_r se considera que $f(\delta_{ij}) = \delta_{ij}$ (es decir, la función identidad) y se tienen en cuenta ciertos pesos w_{ij}, entonces se obtiene una de las funciones de ajuste más utilizadas en el EM Mínimo Cuadrático: el STRESS, cuya expresión es:

$$\text{STRESS} = 2\sum_{i<j} w_{ij}\left[d_{ij}(X) - \delta_{ij}\right]^2$$

Sustituyendo la expresión de la distancia euclidiana para el caso p-dimensional en la expresión anterior, queda:

$$\text{STRESS} = 2\sum_{i<j} w_{ij}\left\{\left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2 - \delta_{ij}^2\right]^{1/2}\right\}^2 \qquad (4)$$

La fórmula anterior es la empleada en la implementación numérica de los métodos que utilizan esta función.

Otra función frecuentemente utilizada en el EM Mínimo Cuadrático es el SSTRESS, que es resultado de emplear en (3) las funciones $f(\delta_{ij}) = d_{ij}^2$ y $g(d_{ij}(X)) = d_{ij}^2(X)$. Se expresa como sigue:

$$\text{SSTRESS} = 2\sum_{i<j} w_{ij}\left[d_{ij}^2(X) - \delta_{ij}^2\right]^2$$

Análogamente a como se obtuvo (4), se tiene la expresión empleada en la implementación numérica de los métodos que utilizan esta función, dada por:

$$\text{SSTRESS} = 2\sum_{i<j} w_{ij}\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2 - \delta_{ij}^2\right]^2 \qquad (5)$$

Método Lineal o Escalamiento de Intervalo

Entre los métodos de Escalamiento Métrico Mínimo Cuadrático, caracterizados por hacer transformaciones paramétricas a las disimilitudes, se encuentra el EM Lineal (en inglés, Interval MDS, Borg & Groenen [1997]) que trabaja en la función (3) con $f(\delta_{ij}) = a + b\delta_{ij}$, donde a y b son los estimadores mínimo cuadráticos ordinarios de la regresión. Un caso particular del EM Lineal es el EM de Razón, método en el que la función de las disimilitudes es de la forma $f(\delta_{ij}) = b\delta_{ij}$. El caso más simple se tiene cuando $f(\delta_{ij}) = \delta_{ij}$ y se conoce como EM Absoluto.

Borg & Groenen [1997] analizan otros métodos métricos que emplean la función (3) como son el EM Logarítmico, donde $f(\delta_{ij}) = a + b \log(\delta_{ij})$ y el Exponencial, donde $f(\delta_{ij}) = a + b \exp(\delta_{ij})$. Un caso particular del primero es el *MULTISCALE* debido a Ramsay, en 1977 (Ver Cox & Cox [1994]), que trabaja con $f(\delta_{ij}) = \log(\delta_{ij})$.

En Química Molecular se emplea el EM Polinómico o Spline, que generaliza al EM Lineal, donde $f(\delta_{ij}) = a_0 + a_1\delta_{ij} + ... + a_r\delta_{ij}^6$, particularmente se usa la función $f(\delta_{ij}) = \delta_{ij}^6$.

El Método Lineal (*Interval MDS*, Borg & Groenen, [1997]) construye $f(\delta_{ij}) = a + b\delta_{ij}$ donde a y b son los estimadores mínimo cuadráticos ordinarios de la regresión. Nótese que en el caso del Método Lineal la función (3) tiene la expresión particular:

$$S = \sum_{i\neq j}\left[(\alpha + \beta\,\delta_{ij}) - d_{ij}(X)\right]^2 / \sum_{i\neq j} d_{ij}^2(X)$$

Escalamiento no métrico

Los métodos *no métricos* construyen ciertos valores $d_{ij}^* = f(\delta_{ij})$ preservando el orden inicial. También se conocen en la literatura con el nombre de EM Ordinal (del inglés, Ordinal MDS, Borg & Groenen [1997]). Dos técnicas muy conocidas son los métodos de Kruskal y Guttman.

Método de Kruskal

En el método de Kruskal, para construir las disparidades se construye una configuración inicial y se calcula la matriz de distancias derivadas de dicha configuración. A continuación, se ordena monótonamente la matriz de disimilitudes inicial $\Delta = (\delta_{ij})_n$ en forma de vector y a éste se asocia el vector de las distancias derivadas correspondientes. Al analizar la relación de orden en el vector de las distancias derivadas, se construye un nuevo vector, el de las disparidades $d_{ij}^* = f(\delta_{ij})$, que consiste en que, si las distancias conservan el orden de las disimilitudes, en las posiciones correspondientes a esas distancias aparecerán las mismas y en caso contrario, se sustituyen ambos valores de las distancias derivadas por su valor promedio. Seguidamente se procede a optimizar la función Stress-1. Aunque, se suele llamar método de Kruskal a esta metodología aún cuando se utilice cualquiera de las variantes de Stress, muchos autores al aplicar el método emplean como función de pérdida el Stress-1 dado por:

$$\text{Stress-1} = L(d_{ij}^*, d_{ij}) = \left\{\sum_{i<j}\left[d_{ij}^* - d_{ij}(X)\right]^2 / \sum_{i<j} d_{ij}^2(X)\right\}^{1/2}$$

Método de Guttman

En el método de Guttman para encontrar las $d_{ij}^* = f(\delta_{ij})$, llamadas imágenes de rango, se procede análogamente al de Kruskal. Estas transformaciones d_{ij}^* se construyen ordenando las disimilitudes en un vector con valores crecientes y formando posteriormente, el vector de distancias derivadas correspondiente, las imágenes de rango d_{ij}^* resultan las distancias ordenadas en forma creciente. También Guttman construyó su propia medida de ajuste, el coeficiente de alienación:

$$G = (1 - u^2)^{1/2} \text{ siendo } u = \sum_{i<j} \left[d_{ij}^* - d_{ij}(X) \right]^2 / \left[\sum_{i<j} d_{ij}^{*2}(X) \right] \left[\sum_{i<j} d_{ij}^2(X) \right]$$

3. UN ENFOQUE UNIFICADO DE LA PÉRDIDA DE INFORMACIÓN

Empleando la norma de Frobenius, pueden expresarse el STRESS y el SSTRESS de la siguiente forma:

$$\text{STRESS} = 2\sum_{i<j} w_{ij} \left[d_{ij}(X) - \delta_{ij}(X) \right]^2 = \| D - \Delta \|_F^2$$

$$\text{STRESS} = 2\sum_{i<j} w_{ij} \left[\left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2} - (\delta_{ij}^2)^{1/2} \right]^2 = \| D * D - \Delta * \Delta \|_F^2$$

Igualmente, el Raw-Stress, el Raw-Stress Normalizado, el Stress-1 y en general, la función $L(e_{ij})$ pueden expresarse en términos de la norma de Frobenius, pues:

$$\sigma_r = \sigma_r(X) = \sum_{(i,j)} w_{ij} \left(d_{ij}(X) - \delta_{ij}(X) \right)^2 = \| D - f(\Delta) \|_F^2 \qquad \text{(Raw-Stress)}$$

$$\sigma_1^2 = \frac{\sigma_r}{\sum_{(i,j)} d_{ij}^2(X)} = \frac{\sum_{(i,j)} \left(d_{ij}(X) - f(\delta_{ij}) \right)^2}{\sum_{(i,j)} d_{ij}^2(X)} = \frac{\| D - f(\Delta) \|_F^2}{\| D \|^2} \qquad \text{(Raw-Stress Normalizado)}$$

$$\sigma_1 = \sqrt{\frac{\sum_{(i,j)} \left(d_{ij}(X) - f(\delta_{ij}) \right)^2}{\sum_{(i,j)} d_{ij}^2(X)}} = \frac{\| D - f(\Delta) \|_F}{\| D \|_F} \qquad \text{(Stress-1)}$$

Y en general,

$$L(e_{ij}) = \sum_{i>j} w_{ij} \left(f(\delta_{ij}) - g(d_{ij}(X)) \right)^2 = \| f(\Delta) - g(D) \|_F^2$$

(Función de ajuste del Escalamiento Mínimo Cuadrático).

Por lo que todas las funciones de ajuste del Escalamiento Métrico Mínimo Cuadrático que en su mayoría son empleadas también en el Escalamiento No Métrico Mínimo Cuadrático pueden expresarse en términos de la norma de la diferencia entre una transformación de las disimilitudes [$f(\Delta)$] y una transformación de las distancias derivadas del EC [$g(D)$].

Para la función de ajuste del método de Guttman: $G = \sqrt{(1-u^2)}$ siendo $u = \dfrac{\sum_{i<j}(d_{ij} - d_{ij}^*)}{\left(\sum_{i<j} d_{ij}^2\right)\sum_{i<j}(d_{ij}^*)}$, se tiene también una expresión en términos de la de la diferencia de una función g(D) y una función f(Δ):

El numerador de u no es más que el Raw-Stress (σ_r) tomando g(D) = D* = $(d_{ij}^*)_{n\times 1}$, siendo las $d_{ij}^* = \hat{\delta}_{ij}$ las imágenes de rango que construye el método.

Luego: $1 - u = \dfrac{\sigma_r}{\|\Delta\|_F^2} = \dfrac{\|D - D^*\|_F^2}{\|\Delta\|_F^2}$.

Por lo tanto: $G = \sqrt{1-u} = \sqrt{\dfrac{\|D - D^*\|_F^2}{\|\Delta\|_F^2}} = \dfrac{\|D - D^*\|_F}{\|\Delta\|_F}$.

Otra forma de plantear el problema a resolver por Rao en el epígrafe 1 consiste en:

Buscar una matriz T de tamaño kxp que transforme las coordenadas dadas por las columnas de la matriz X inicial en las coordenadas Y(k) finales con respecto a un espacio de R^k con el producto escalar inducido por la matriz $TM^{-1}T^t$ y de modo que haga mínima la diferencia $D^{(2)} - D_{(k)}^2$, donde: $D^{(2)} = (d_{ij}^2)_n$ y $D^{(2)} = D_{(k)}^{(2)} = (d_{ij(k)}^2)_n$, siendo $d_{ij}^2 = (X_i - X_j)^t M(X_i - X_j)$ la distancia cuadrada entre X_i y X_j en el espacio básico inicial y $d_{ij(k)}^2 = (X_i - X_j)^t T^t)TM^{-1}T^t)^{-1}T(X_i - X_j)$ la distancia cuadrada entre X_i y X_j en el espacio R^k de dimensión reducida (k < p).

El problema anterior lleva a la misma solución del que resuelve el teorema de Rao y se cumple que:

$$\min_{T_{k\times n}} \left\| W^{1/2}(D^2 - D_{(k)}^2)W^{12} \right\| = \min_{Y_{(k)}\in R^k} \left\| W^{1/2}(C_i - C_f)W^{12} \right\| = \min_{Y_{(k)}\in R^k} \left[\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j (d_{ij}^2 - d_{ij(k)}^2) \right]$$

Según lo anterior,

La función de pérdida de Rao puede expresarse como la norma de la diferencia ponderada entre los cuadrados de las distancias iniciales y finales, es decir:

$$\min_{T_{k\times n}} \left\| W^{1/2}(D^2 - D_{(k)}^2)W^{12} \right\|$$

Esto permite buscar analogías con las expresiones de las funciones de pérdida de los métodos del EM.

Nótese que $D^{(2)}$ es la transformación g de D llamada producto de Hadamard, o sea, aquella que transforma a toda matriz en la de sus coeficientes al cuadrado, $g(D) = D^*D = (d_{ij}^2)_n$.

D coincide con Δ ya que toda matriz de distancias es de disimilitudes.

f es la composición de g con la aplicación proyección sobre todos los subespacios R^k del espacio euclidiano R^p.

La norma empleada por Rao es la de Frobenius, es decir, la norma matricial inducida por el producto escalar:$\langle A, B\rangle = trB^tA$

Como se ha visto, la función de pérdida que sirve para abarcar todas las posibles expresiones empleadas en los métodos del EM tiene en la función de pérdida de Rao, un caso particular.

Por tanto, el problema que plantean resolver las técnicas del EM puede considerarse como el problema general de una teoría que involucra a todos los métodos de representación de datos pues en todos se trata de minimizar una función de la forma: $\| f(\Delta) - g(D) \|_F^p$, sujeto a restricciones que define cada método y asumiendo que el subíndice F en la norma es relativo a la norma de Frobenius.

Si se trata de técnicas cuyo óptimo se obtiene por métodos algebraicos, adopta la forma:

$\min_{Y_{(k)} \in R^k} \| f(\Delta) - g(D) \|_F^p = \min_{Y_{(k)} \in R^k} \| f(D) - f(D_{(k)}) \|_F^p$, donde f es el producto de Hadamard y la solución se construye algebraicamente según se vio en el teorema anterior.

Si, por el contrario, se trata con técnicas cuya solución se obtiene por métodos numéricos (procedimientos del EM), deben diferenciarse dos casos:

(I) Técnicas métricas

En la función $\| f(\Delta) - g(D) \|_F^p$, f es una función paramétrica de Δ.

Si se escribe D en función de la matriz de configuración X, se minimiza la función de ajuste con respecto a sus componentes x_{ij}.

El problema dependiente sólo de X es irrestricto.

(II) Técnicas no métricas

En la función $\| f(\Delta) - g(D) \|_F^p$, f es una función monótona de Δ.

Se minimiza la función de ajuste con respecto a los coeficientes de Δ y a las componentes de la matriz de configuración $X = (x_{ij})$.

Luego, la función de pérdida $\| f(\Delta) - g(D) \|_F^p$, que sirve para abarcar a todas las funciones de ajuste del EM, tiene en la función de pérdida de Rao, un caso particular. Por tanto, del estudio de los métodos del EM como técnicas de representación de datos se han encontrado comunalidades con los restantes métodos incluidos en la teoría de Rao y aún cuando sus procedimientos de trabajo para obtener la solución son diferentes, todos tienen un mismo propósito: representar datos en un espacio euclidiano de baja dimensión que haga mínima la función $\| f(\Delta) - g(D) \|_F^p$. Esta invariante de todos los métodos analizados permite enfocarlos unificadamente y contribuye desde el punto de vista metodológico a conocer las bondades del EM en su función exploratoria. En la literatura actual, son muy variadas las esferas de aplicación del EM que con mucha frecuencia rebasan los límites del Análisis Exploratorio de Datos Multivariados y dada la naturaleza numérica de sus algoritmos, no se suelen relacionar con las otras técnicas bien conocidas de representación euclidiana como las incluidas por Rao en su formulación del epígrafe 1.

4. RECONSTRUCCIÓN DEL MAPA DE LAS CAPITALES DE PROVINCIA DE CUBA

Una de las aplicaciones más conocidas del EM es la reconstrucción de mapas de ciudades, conocidas sus distancias por carretera. (Mardia **et al**. [1979], Borg & Groenen [1997])

Empleando dos de los programas confeccionados por Borrego para preparar la ponencia presentada en el CLAPEM (Borrego & Miret [2001]), se aplicó EC a la matriz de distancias de las capitales de 15 provincias de Cuba (14 provincias y el municipio especial Isla de la Juventud). En el gráfico obtenido, como puede apreciarse (Figura 1), el ajuste no es bueno con respecto a la distribución de puntos esperada, aunque cabe destacar que, la matriz de distancias por carretera (y por mar, en el caso del municipio especial Isla de la Juventud) entre las 15 capitales de las provincias de nuestro país, es en realidad una matriz de disimilitudes no euclídea porque las distancias por carretera no se deben a mediciones en línea recta.

Considerando la configuración final debida al EC se aplicó el método métrico EM Absoluto empleando en la optimización del STRESS el método de Newton globalizado (Kearsley, Tapia & Trosset [1998]), que varía del de Newton en que realiza una búsqueda por regiones de confianza (trust region) en lugar de la búsqueda lineal usual.

Los resultados finales obtenidos son muy parecidos a los relativos a la ubicación real de las capitales de provincias de Cuba en un mapa plano. El valor del STRESS en esta estrategia es considerablemente pequeño lo que confirma la pérdida de información mínima. (Figura 2)

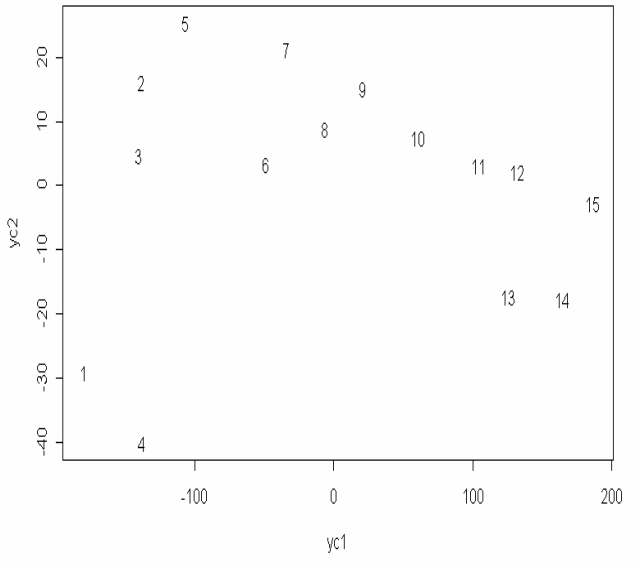

Figura 1. Solución Clásico.
Construcción del mapa de Cuba.

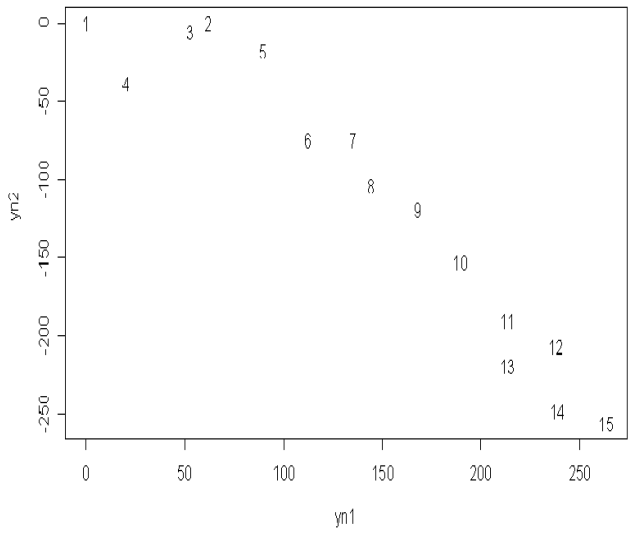

Figura 2. Solución Clásico-EM Absoluto empleando Newton Globalizado.
Construcción del mapa de Cuba.

REFERENCIAS

BORREGO, J.A. & MIRET, E. [2001]. "Algunos algoritmos de optimización en el Escalamiento Multidimensional". Ponencia presentada en CLAPEM'2001. Cuba. Noviembre.

BORG, I. and P. GROENEN [1997]: **Modern multidimensional scaling**. Springer-Verlag New York, Inc.

COX, T.F. and M.A.A. COX [1994]: **Multidimensional Scaling**. CHAPMAN & HALL. London.

CUADRAS, C.M. [1981]: **Métodos de Análisis Multivariante**. EUNIBAR, Barcelona.

CUADRAS, C.M. [1989]: "Distancias estadísticas". **Estadística Española** 30(119), 295-378.

KEARSLEY, A.; R.A. TAPIA and N.W. TROSSET [1998]: "The solution of the metric SSTRESS and STRESS problems in Multidimensional Scaling using Newton's method". **Computational Statistic** 13, 369-396.

MARDIA, K.V.; J.T. KENT and J.M. BIBBY [1979]: Multivariate Analysis. Academic Press, Inc., London.

MATLAB (1994): "The Matrix Laboratory". **The Math. Works**, Inc. Version 4.2c.

MIRET, E. [2005]. "Un enfoque unificado para técnicas de representación euclidiana". Tesis Presentada para optar por el grado de doctor. Universidad de La Habana, Cuba.

RAO, C. R. [1995]: "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance". **Qüestiió**, Barcelona. 19(1,2,3). 23-63.

TARAZAGA, P. & TROSSET, M. W. [1998]. “An approximate Solution to the Metric SSTRESS Problem in Multidimensional Scaling”. http:\\www.researchindex.com

TROSSET, M.W. [1993]. “Numerical Algorithms for Multidimensional Scaling” En: Information and Classification. R. Klar & O. Opitz (Eds). Springer, Heidelgerg, pp 81-92.

ZHANG, B. & SRIHARY, S. N. [2003]. “Properties of Binary Dissimilarity Measures.” Cedar. Publications. http://www.cedar.buffalo.edu/papers/pubs2000.html.