UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

INTRODUCCIÓN AL ANÁLISIS DE ESCALAMIENTO MULTIDIMENSIONAL

T E S I S

QUE PARA OBTENER EL TÍTULO DE

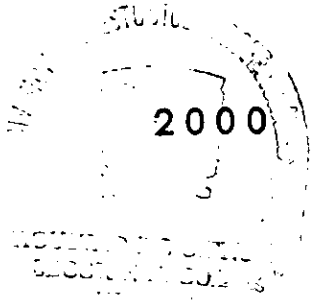
M A T E M A T I C A

P R E S E N T A :

MARÍA EUGENIA LEÓN CANO

DIRECTOR DE TESIS:

M. en A.P. MA. DEL PILAR ALONSO REYES



VNIVERIDAD NACIONAL
AVENMA DE
MEXICO

MAT. MARGARITA ELVIRA CHÁVEZ CANO
Jefa de la División de Estudios Profesionales
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:

Introducción al Análisis de Escalamiento Multidimensional

realizado por Ma. Eugenia León Cano

Con número de cuenta 8828602-4 , pasante de la carrera de Matemáticas

Dicho trabajo cuenta con nuestro voto aprobatorio.

A t e n t a m e n t e

Director de tesis Propietario	M. en A. P. Ma. del Pilar Alonso Reyes
Propietario	M. en C. José Antonio Flores Díaz
Propietario	Dr. Andrei Novikov
Suplente	Dra. Blanca Rosa Pérez Salvador
Suplente	Act. Ma. del Rosario Espinosa Tufiño

Consejo Departamental de Matemáticas
Dr. Héctor Méndez Lango

FACULTAD DE CIENCIAS
CONSEJO DEPARTAMENTAL
DE
MATEMATICAS

gracias . . .

. . . A Dios, por la luz de la existencia

. . . A mi Madre y a mi Padre, por la luz de la vida

. . . A Jorge, por la luz del amor

gracias también . . .

...A mis hermanas y a mis hermanos, por su incomparable ejemplo de valentia, entrega, y lucha.

El señor es mi pastor; nada me falta;
en verdes pastos él me hace reposar
y a donde brota agua fresca me conduce.

Fortalece mi alma,
por el camino del bueno me dirige
por amor de su Nombre.

Aunque pase por oscuras quebradas,
no temo ningún mal,
porque tú estas conmigo,
tú bastón y tu vara me protegen.

Me sirves a la mesa
frente a mis adversarios,
con aceite perfumas mi cabeza
y rellenas mi copa.

Me acompaña tu bondad y tu favor
mientras dura mi vida;
mi mansión será la casa del Señor
por largo, largo tiempo.

INTRODUCCIÓN AL

ANÁLISIS DE ESCALAMIENTO

MULTIDIMENSIONAL

Índice General

Capítulo 1

INTRODUCCIÓN

1.1 ESCALAMIENTO MULTIDIMENSIONAL

El análisis de escalamiento multidimensional es un conjunto de procedimientos desarrollados para investigar las relaciones y estructura existentes en una matriz de datos que representan ciertas medidas de disimilaridad entre objetos. El propósito es representar los objetos como puntos en un espacio de dimensión baja y las medidas de disimilaridad como las distancias entre ellos. Cuando el conjunto de datos es muy grande el valor práctico de este análisis es más aparente. La representación gráfica en dimensión dos o tres, que proporciona el análisis de escalamiento multidimensional permite que el investigador "vea" literalmente los datos y explore su estructura, lo que resultaría poco fructífero si sólo se observara el conjunto de números.

1.2 CUATRO PROPÓSITOS

El análisis de escalamiento multidimensional tiene tantos propósitos como aplicaciones, los cuales se pueden resumir en los siguientes cuatro:

1) Como un método exploratorio. Representa las medidas de disimilaridad como distancias en un espacio de dimensión baja para hacer los datos accesibles a una inspección y exploración visual. El análisis exploratorio se usa para estudiar datos teóricamente amorfos, es decir, aquellos que no están relacionados con una teoría explicita que prediga sus magnitudes o comportamientos.

2) Como una técnica para probar hipótesis estructurales. Permite verificar que los criterios con los que se diferencian conceptualmente los objetos reflejan las diferencias empíricas entre ellos. Cuando ya se tiene mayor información sobre el tema de estudio quedan atrás los métodos exploratorios.

3) Como una técnica para explorar estructuras psicológicas. Es un acercamiento analítico que permite descubrir las dimensiones que fundamentan los juicios de disimilaridad.

4) Como un modelo psicológico. Las matemáticas del análisis de escalamiento multidimensional pueden servir como modelos de juicios de disimilaridad. El enfoque más común es adoptar como hipótesis que las personas para emitir un juicio de disimilaridad entre un par de objetos, calcula un tipo particular de distancia en su propio espacio psicológico de estos objetos. Tal distancia es interpretada como una regla de composición psicológica.

El análisis de escalamiento multidimensional es útil en muchos campos de la investigación. La siguiente es sólo una lista ilustrativa de sus diversas aplicaciones.

a) Los científicos políticos usan frecuentemente el análisis de escalamiento multidimensional para entender porque los candidatos políticos son percibidos como similares o disimilares.

b) Los antropólogos han usado el análisis de escalamiento multidimensional para estudiar las diferencias culturales de varios grupos, basados en creencias, lenguaje e información sobre artefactos.

c) Los urbanistas han identificado las similaridades de varias ciudades, pueblos o regiones en términos de su posición en un espacio de dimensión reducida derivado de datos demográficos, fiscales y económicos.

d) Los psicólogos han utilizado estos procedimientos para entender las percepciones y las evaluaciones de colores, velocidades o factores de personalidad entre otras cosas.

e) Los mercadotecnistas utilizan el análisis de escalamiento multidimensional para investigar la manera en que los clientes evalúan los productos y entender la relación entre las características de los mismos.

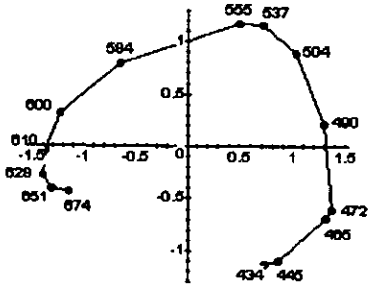
f) Los sociólogos utilizan este tipo de análisis para determinar la estructura de un grupo basado en las percepciones interpersonales y en las diferencias del comportamiento.

1.3 TRES EJEMPLOS

Ejemplo 1. Cuando una persona se enfrenta a la tarea de ordenar colores generalmente lo hace de manera que los colores no aparecen sobre una línea recta sino sobre una curva suave ordenados en forma creciente con respecto a su longitud de onda.

En un estudio que suele considerarse clásico, Shepard (1962) reanalizó los datos generados por 31 sujetos al evaluar unicamente la similaridad de cada uno de los 91 pares formados con 14 colores. Los colores varían en un rango de 434 a 674 nm. (nanómetros) de longitud de onda. Los puntajes de similaridad fueron transformados a un rango de 0 (no similares en absoluto) a 1 (idénticos).

Realizando un análisis de escalamiento multidimensional no métrico se puede obtener una solución en dos dimensiones. La representación gráfica de la configuración que se obtiene muestra una estructura que corresponde al concepto familiar del círculo de colores descrito arriba.

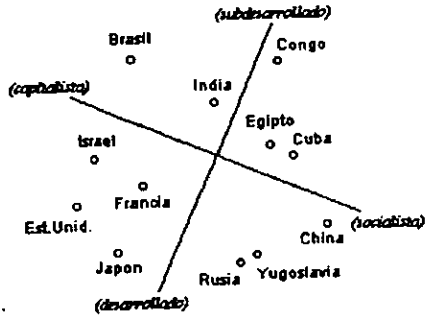


círculo de colores

En el capítulo final se analiza de manera más amplia este interesante ejemplo.

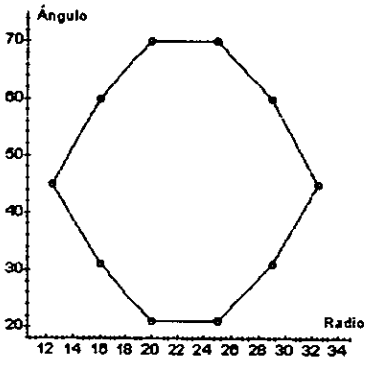
Ejemplo 2. Los datos de este ejemplo aparecieron originalmente en Wish (1971). Se trata de juicios de disimilaridad promediados de 18 estudiantes sobre 12 naciones. Los sujetos evaluaron la disimilaridad de cada par de naciones en un rango de 1 (muy similares) a 9 (muy diferentes), no se les dio indicación alguna acerca de las características sobre las que debían basar sus evaluaciones. Al aplicar un análisis de escalamiento multidimensional, con la suposición de medidas ordinales, se obtuvo un mapa en dos dimensiones.

Cada representacion gráfica obtenida debe ser interpretada por el investigador. Por ejemplo, se puede determinar, superponiendo dos líneas punteadas, que una dirección significativa es la afiliación política y otra el desarrollo económico de los países. Estas líneas (las cuales no son generadas como parte de la solución) representan los ejes del espacio derivado después de una rotación.



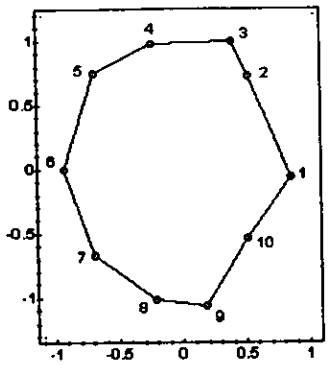
12 naciones

Ejemplo 3. En este ejemplo los datos son medidas de disimilaridad de 10 ruedas de un rayo. Es decir, 10 círculos cuyo diametro varía de 12.5 a 32.5 mm. con un radio dibujado a un ángulo que varía de 21 a 69°. Cada uno de los 45 pares fueron presentados a los sujetos para ser evaluados en una escala similaridad de 1 (mínima) a 7 (máxima). La hipótesis es que un sujeto llega a un juicio de similaridad calculando una distancia particular en su propio espacio psicológico. Este espacio debería corresponder esencialmente al espacio físico diseñado de donde fueron extraídos los objetos.



diseño original

La representación dada por el análisis de escalamiento multidimensional se calculó de manera que las "distancias ciudad" correspondieran al máximo con las medidas de similaridad. Broderson sostiene que los sujetos llegan a esos juicios comparando cada par de objetos dimensión por dimensión, sumando las diferencias percibidas y convirtiendo el resultado global a un formato de escala.



configuración obtenida

Los datos obtenidos con el análisis parecen sostener esta teoría. Las distancias ciudad y las medidas de similaridad entre cualesquiera dos puntos están altamente correlacionadas ($r = -0.92$) . Esta distancia geométrica bidimensional es de hecho un posible modelo de los juicios de similaridad para estos objetos.

1.4 RESUMEN HISTÓRICO

A continuación se presenta, a manera de resumen, una lista de grandes pasos dentro del desarrollo histórico del análisis de escalamiento multidimensional.

1) Análisis de escalamiento multidimensional métrico.

- Torgerson (1952), la primera propuesta. Ajuste del modelo euclideano.
- Attneave (1953) propuso los modelos no euclideanos.
- Messick y Abelson (1956) resuelven el problema de la constante aditiva.

2) Análisis de escalamiento multidimensional no métrico.

- Shepard (1962) propone un primer acercamiento intuitivo psicométrico.
- Kruskal (1964) proporciona el acercamiento matemático estadístico racional.
- Torgerson y Meuser (1962) combinan los trabajos de Torgerson y Kruskal.
- Guttman (1968) propone un acercamiento diferente muy importante.
- McGee (1968), Roskam (1968), de Leeuw (1973), Lingoes (1973), y Young (1972) contribuyen al desarrollo del análisis de escalamiento multidimensional no métrico.

3) Análisis de escalamiento multidimensional con diferencias individuales.

- Tucker y Messick (1963) los precursores con el análisis de puntos de vista.
- McGee (1968) propone el análisis de escalamiento no métrico replicado.
- Bloxom (1968) introduce el modelo ponderado.
- Carroll y Chang (1970) desarrollan el programa INDSCAL para aplicar el modelo ponderado

4) Consolidación del desarrollo.

- Takane, Young y de Leeuw (1977) combinan el análisis de escalamiento multidimensional métrico y no métrico, ponderado y no ponderado en el mismo algorítmo: ALSCAL, pero ajusta distancias cuadradas.
- Ramsay (1977) introduce pruebas de significancia basados en maxima verosimilitud del análisis de escalamiento multidimensional métrico ponderado y no ponderado, en el programa MultiScale.
- De Leeuw (1977) y De Leeuw y Heiser (1977) combinan el análisis de escalamiento multidimensional métrico y no métrico, ponderado y no ponderado, y ajusta distancias en un sólo algorítmo: SMACOF.

Capítulo 2

MEDIDAS DE PROXIMIDAD

2.1 MEDIDAS DE PROXIMIDAD

El **análisis de escalamiento multidimensional** se basa en la comparación de objetos. Con el término "objetos" se denomina a las cosas o eventos que se encuentran bajo interés y/o estudio (especies, zonas geográficas, estímulos, productos, etc.).

Las **medidas de proximidad** son un conjunto de números que indican el grado de semejanza (similaridad) o diferencia (disimilaridad) que guarda cada par de objetos, con relación a cierto número de características cualitativas y cuantitativas.

Tales medidas, reunidas en una matriz, son los elementos básicos para realizar un análisis de escalamiento multidimensional.

2.1.1 Similaridades

Sea A el conjunto de n objetos, indicado por $\{1, 2, ..., i, ..., j, ..., n\}$. Se llama **similaridad**, a la medida de proximidad que indica el grado de semejanza o similitud

entre el objeto i y el objeto j, y se denota por $s(i,j)$. Cuanto mayor es la semejanza entre los objetos, mayor es el valor de $s(i,j)$. Un valor pequeño de $s(i,j)$ indica poca similitud entre i y j.

Los coeficientes de similaridad tienen una larga historia, en la literatura temprana se les conocía como coeficientes de asociación.

La mayoría son considerados como simétricos, es decir, tales que $s(i,j) = s(j,i)$, y son acomodados de tal manera que sean positivos y tengan un límite superior igual a 1, aunque algunos cumplen la condición : $-1 \leq s(i,j) \leq 1$.

El coeficiente de correlación de Pearson es un índice de similaridad.

2.1.2 Disimilaridades

Sea A, nuevamente el conjunto de n objetos indicado por $\{1, 2, ..., i, ..., j, ..., n\}$. Se llama **disimilaridad** a la medida de proximidad que indica el grado de diferencia entre el objeto i y el objeto j, denotándolo por $\delta(i,j)$.

Cuanto mayor es la diferencia entre los objetos i y j, mayor es el valor de $\delta(i,j)$. Un valor pequeño de $\delta(i,j)$, indica poca diferencia entre los objetos.

Asociada a cada medida de similaridad limitada por cero y uno, existe una medida de disimilaridad definida por : $\delta(i,j) = 1 - s(i,j)$, la cual es simétrica y no negativa.

Un grado mayor de similaridad entre los objetos i y j incrementa el valor de $s(i,j)$ y disminuye el valor de $\delta(i,j)$.

La distancia euclideana es un ejemplo de disimilaridad.

2.1.3 Distancias

Sean r y s dos puntos (que pueden representar a dos objetos). Una función real denotada por $d(r,s)$ es una **función de distancia** si cumple con las propiedades:

I) $d(r,s) = d(s,r)$ (simetría)

II) $d(r,s) \geq 0$ (no negatividad)

III) $d(r,r) = 0$

Para algunas funciones de distancia se cumplen también las siguientes propiedades:

IV) $d(r,s) = 0$ si y sólo si $r = s$.

V) $d(r,s) \leq d(r,p) + d(p,s)$ (desigualdad del triángulo)

VI) $d(r,s) \leq \max\{d(r,p), d(s,p)\}$ (desigualdad ultramétrica)

VII) Existe un espacio euclideano R^m y dos puntos P_r , $P_s \in R^m$ de coordenadas $P_r = \{x_{r1}, x_{r2}, ..., x_{rm}\}$ y $P_s\{x_{s1}, x_{s2}, ..., x_{sm}\}$, tales que:

$d(r,s) = \left[\sum_{k=1}^{m} (x_{rk} - x_{sk})^2\right]^{1/2} = d_2(P_r, P_s)$, llamada distancia euclideana

Una función de distancia recibe diferentes denominaciones, según las propiedades que verifica:

Denominación	Propiedades
Disimilaridad	I II III
Distancia Métrica	I II III IV V
Distancia Ultramétrica	I II III VI
Distancia Euclideana	I II III IV VII

De la métrica de Minkowski : $d_p(r,s) = \left[\sum_{k=1}^{m} |x_{rk} - x_{sk}|^p\right]^{1/p}$ se obtienen las distancias más comunes.

Con $p = 1$ la distancia ciudad o de Manhattan:

$$d_1(r,s) = \sum_{k=1}^{m} |x_{rk} - x_{sk}|$$

Con $p = 2$ la distancia Euclideana:

$$d_2(r,s) = \left[\sum_{k=1}^{m} (x_{rk} - x_{sk})^2\right]^{1/2}$$

Una expresión límite da la Sup-norma:

$$d_\infty(r,s) = Sup_{1 \le k \le m} |x_{rk} - x_{sk}|$$

En general, estas distancias no son Euclideanas (propiedad VII) excepto para d_2, que justamente lleva este nombre.

Sin embargo $d_2(r,s)$ no es invariante bajo cambios de escala, es decir, cambios en la escala de medición pueden producir serios efectos sobre d_2.

El siguiente ejemplo lo muestra:

Se tiene el peso (medido en gramos) y la altura (medida en centímetros y en milímetros), de tres objetos como lo indica la tabla que se muestra a continuación:

Objeto	Peso (gm.)	Altura (cm.)	Altura (mm.)
(1) A	10	8	80
(2) B	20	3	30
(3) C	30	11	110

Calculando las distancias euclideanas para estos tres objetos, primero con la altura medida en centímetros y después en milímetros, se obtiene:

distancias	altura en cm.	altura en mm .
$d_2(1,2)$	11.18	50.99
$d_2(1,3)$	20.22	36.05
$d_2(2,3)$	12.80	80.62

Obsérvese que en el primer caso el objeto C estaría más cerca del objeto B que del objeto A, en cambio en el segundo caso, el objeto C estaría más cerca del objeto A que del objeto B, aún cuando en ambos casos la altura sea la misma.

Para remover la dependencia en las unidades de medida, cada variable se deberá dividir por su desviación estándar: $\sigma_k = \left[\frac{1}{n-1}\sum_{l=1}^{m}(x_{lk} - \overline{x}_{.k})^2\right]^{1/2}$ donde $\overline{x}_{.k} = \frac{1}{n}\sum_{l=1}^{n} x_{lk}$ para después calcular las distancias.

Aplicando esto a los datos del ejemplo se obtiene lo siguiente:

Objeto	Peso	Altura (cm.)	(mm)
(1) A	$\frac{10}{10} = 1$	$\frac{8}{4.04} = 1.98$	$\approx \frac{80}{40.4}$
(2) B	$\frac{20}{10} = 2$	$\frac{3}{4.04} = 0.74$	$= \frac{30}{40.4}$
(3) C	$\frac{30}{10} = 3$	$\frac{11}{4.04} = 2.72$	$= \frac{110}{40.4}$

Distancias
$d_2(1,2) = 1.59$
$d_2(2,3) = 1.29$
$d_2(1,3) = 1.01$

Puede observarse que los valores de la característica altura serán los mismos en cualquiera de los dos casos, en consecuencia las distancias serán únicas.

Otra función de distancia muy utilizada es la distancia de Mahalanobis:

$$D(x_r, x_s) = \left[(x_r - x_s)^t S^{-1} (x_r - x_s)\right]^{1/2}$$

donde $S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^t$ y $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Esta medida es invariante bajo transformaciones del tipo $y = Ax_i + b$ con A no singular. En particular es invariante bajo cambios de escala.

Aplicando esta función de distancia a los datos del ejemplo anterior se tendría:

$$x_1 = \begin{pmatrix} 10 \\ 8 \end{pmatrix} \quad x_2 = \begin{pmatrix} 20 \\ 3 \end{pmatrix} \quad x_3 = \begin{pmatrix} 30 \\ 11 \end{pmatrix} \quad \overline{x} = \begin{pmatrix} 10 \\ 7.33 \end{pmatrix}$$

$$S = \begin{pmatrix} 100 & 15 \\ 15 & 16.33 \end{pmatrix} \quad S^{-1} = \begin{pmatrix} 0.01\,1598 & -0.01\,0651 \\ -0.01\,0651 & 0.07\,1008 \end{pmatrix}$$

$$D(x_1, x_2) = 2.000025 \quad D(x_1, x_3) = 2.000038 \quad D(x_2, x_3) = 2.000038$$

2.2 MEDIDAS DE PROXIMIDAD DIRECTAS

Este término se aplica a las medidas de proximidad originadas al establecer directamente una relación binaria de similaridad o disimilaridad entre los objetos. Para esto se elige a N sujetos y a cada uno se le presentan los objetos por pares (en un conjunto de cartas, por ejemplo) y se le pide que emita su juicio de (dis)similaridad sobre cada par.

Los métodos más comúnes para obtener estos datos son los siguientes:

2.2.1 Evaluación categórica (puntaje)

Este método requiere que cada sujeto evalúe la similaridad o disimilaridad total de cada par de objetos, y asigne un puntaje de acuerdo al siguiente esquema:

Objeto i Objeto j

exactamente iguales	0	1	2	3	4	5	6	completamente diferentes
					×			

El valor que recibe el par de objetos (i, j) es el número de la categoría que fue marcada por el sujeto al evaluar este par.

Calculando la media aritmética de los valores asignados por cada uno de los sujetos al par (i, j) se obtiene una medida de disimilariadad $\delta(i, j)$.

2.2.2 Evaluación gráfica (puntaje gráfico)

Ésta es una ligera variación del método anterior. El sujeto debe marcar sobre una línea continua (de longitud 10 cm.) al evaluar cada par de objetos, de acuerdo al esquema siguiente:

Objeto i Objeto j

exactamente iguales _____/_____ completamente diferentes

A cada par se le asigna la longitud del segmento que queda a la izquierda de la marca. Se obtiene una medida de disimilaridad $\delta(i,j)$, calculando la media aritmética de los valores asignados por todos los sujetos, al par (i,j).

2.2.3 Evaluación condicional (objeto ancla)

Se selecciona un objeto-ancla y cada uno de los $n-1$ objetos restantes son comparados uno por uno con éste, de acuerdo a su similitud o diferencia. Se pueden utilizar los métodos antes descritos para obtener las disimilaridades.

Cada objeto sirve, en turno, de objeto-ancla. Esto da n conjuntos con $n-1$ medidas de proximidad cada uno.

Este método tiene la ventaja de que se necesita hacer menos comparaciones a un tiempo; en lugar de evaluar $\binom{n}{2} = \frac{n(n-1)}{2}$ pares de objetos, el método del objeto-ancla sólo necesita $n-1$. Sin embargo, las medidas así obtenidas son condicionales y menos comparables.

2.2.4 Clasificación de objetos

En este método se les pide a los N sujetos que clasifiquen los objetos en diferentes grupos, en función de su similaridad. El número de grupos puede ser determinado por el investigador o dejar a los sujetos decidirlo. Si N_{ij} es el número de sujetos que han incluido al objeto i y al j en el mismo grupo, entonces se define una medida de disimilaridad $\delta(i,j) = \frac{N-N_{ij}}{N}$.

Este coeficiente ha sido utilizado en lingüística para construir estructuras jerárquicas de vocablos.

2.2.5 Clasificación de pares

Al sujeto se le pide que clasifique los pares de objetos en diferentes grupos de tal manera que los pares formados por objetos muy similares deben quedar en el primer grupo y los pares formados por los objetos más disimilares deben quedar en el último. A cada grupo se le asigna un número $\{1, 2, ..., k\}$, comenzando por el primero y a cada par se le asigna el número del grupo al que pertenece.

Se define una medida de disimilaridad $\delta(i,j)$ como la media aritmética de los valores asignados al par de objetos i y j, por todos los sujetos.

2.2.6 Ordenamiento de Pares

A cada sujeto se le pide que ordene los pares de objetos de acuerdo a su similaridad o disimilaridad, de tal manera que queden acomodados en una pila, donde el par de objetos más similares quede en la base y el de los más disimilares en el tope de la

pila. A cada par se le asigna un número del conjunto $\left\{1, 2, ..., \frac{n(n-1)}{2}\right\}$ comenzando por la base de la pila.

Se define un medida de disimilaridad $\delta(i, j)$ como la media aritmética de los valores asignados al par de objetos i y j , por todos los sujetos.

2.2.7 Observaciones

Claramente la facilidad o dificultad de obtener las medidas de proximidad directas está en gran medida determinada por el número de objetos. Un número grande de objetos implica una gran cantidad de comparaciones. Con $n = 7$ se tienen 21 pares; con $n = 20$ se tienen 190; en general con n objetos se tienen $\binom{n}{2} = \frac{n(n-1)}{2}$ pares, un número que crece demasiado rápido con n.

Si el número de objetos es muy grande, la tarea de evaluar todos los pares es muy demandante para los sujetos. Una alternativa es emplear un diseño de datos incompletos, en el cual los pares de objetos se dividen en subconjuntos no ajenos y cada individuo evalúa uno. Cada par de objetos es evaluado por igual número de sujetos pero no por todos.

2.3 MEDIDAS DE PROXIMIDAD DERIVADAS

En la práctica, generalmente las medidas de proximidad no se obtienen de juicios directos de similaridad, sino más bien son índices derivados de otra información sobre los objetos, dando lugar a las medidas de proximidad derivadas.

2.3.1 Variables Cuantitativas

Supóngase que se tiene para cada objeto i un vector de valores $(x_{i1}, x_{i2}, ..., x_{ip})$ sobre p variables o características de tipo cuantitativo.

Coeficientes de Correlación

Las medidas de proximidad derivadas más comunes son los coeficientes de correlación. Por ejemplo el coeficiente de correlación de Pearson:

$$s(i,j) = \frac{\sum_{k=1}^{p}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\left[\sum_{k=1}^{p}(x_{ik}-\bar{x}_i)^2 \sum_{k=1}^{p}(xj-\bar{x}j)^2\right]^{\frac{1}{2}}}$$

sería una medida de similaridad para los objetos i y j, la cual cumple la condición $-1 \leq s(i,j) \leq 1$.

Además de no estar entre 0 y 1, tiene la desventaja de que, por ejemplo, si $s(i,j) = 1$, esto no significa que $i = j$, sólo significa que los elementos de x_i están relacionados de manera lineal con los de x_j.

Los coeficientes de correlación miden diferentes comportamientos, el de Pearson mide el grado de relación lineal. Si se sustituyen los datos originales por rangos, se obtiene el coeficiente lineal por rangos ρ de Spearman, que mide la relación monotónica de x_i y x_j.

Otra opción es el coeficiente μ_2 de Guttman, que es usado en combinación con escalamiento multidimensional ordinal.

Distancias

Otro tipo de medidas de proximidad derivadas son las de tipo distancia, es decir, se puede definir una medida de disimilaridad como $\delta(i,j) = d(i,j)$.

Se tienen entonces las siguientes medidas de disimilaridad:

$$\delta_1(i,j) = d_1(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

$$\delta_2(i,j) = d_2(i,j) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right]^{1/2}$$

$$\delta_\infty(i,j) = d_\infty(i,j) = Sup_{1\leq k\leq p} |x_{ik} - x_{jk}|$$

$$\delta_D(i,j) = D(x_i, x_j) = [(x_i - x_j)^t S^{-1}(x_i - x_j)]^{1/2}$$

Los inconvenientes y la alternativa para $\delta_2(i,j)$ así definida, se discuten en la parte dedicada a distancias de la primera sección de este capítulo.

Algunas versiones de d_1 han sido propuestas como medidas de proximidad, por ejemplo:

$\delta(i,j) = \frac{1}{p}\sum_{k=1}^{p} \frac{|x_{ik}-x_{jk}|}{R_k}$ donde R_k es el rango de la variable k (Gower 1971 a)

$\delta(i,j) = \frac{\sum_{k=1}^{p}|x_{ik}-x_{jk}|}{\sum_{k=1}^{p} x_{ik} + \sum_{k=1}^{p} x_{jk}}$ llamada de Bray-Curtis

Esta última medida se utiliza ocasionalmente en Ecología, aunque no necesariamente es una métrica.

$\delta(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik}-x_{jk}|}{x_{ik}-x_{jk}}$ llamada Métrica de Camberra, la cual es insensible a sesgos y datos extremos.

2.3.2 Variables dicotómicas

Supóngase que cada objeto i tiene asociado un vector $(x_{i1}, x_{i2}, ..., x_{ip})$ de valores sobre p variables de tipo binario o dicotómico.

Se puede además suponer, sin pérdida de generalidad, que se trata de la presencia (denotada por +) o ausencia (denotada por -) de cierta característica.

Para cada par de objetos se puede formar la tabla de contingencia:

objeto j

objeto i		(+)	(-)	total
	(+)	a	b	$a+b$
	(-)	c	d	$c+d$
	total	$a+c$	$b+d$	$a+b+c+d$

Donde, por ejemplo b es el número de características presentes en el objeto i , pero ausentes en el objeto j.

Se han propuesto numerosas medidas de similaridad $s(i,j)$ basadas en estos números, la mayoría satisfacen la condición $0 \leq s(i,j) \leq 1$.

Las más populares son:

Jaccard (1928) : $s(i,j) = \frac{a}{a+b+c}$

Czekanowski (1913) : $s(i,j) = \frac{2a}{2a+b+c}$

proporción simple : $s(i,j) = \frac{a+d}{a+b+c+d}$

Cada uno de estos índices pone el énfasis en un aspecto concreto de la relación. La elección de uno depende mucho de los pesos relativos que se quiera dar a las relaciones positivas a y negativas d . Cuadras (1990) sugiere que no deben depender de d, pues añadiendo carácteres arbitrarios no comunes, podrían hacerse falsamente similares objetos que no lo son.

Muchos autores consideran como similaridades, únicamente a este tipo de medidas.

2.4 CONCLUSIONES

La gran variedad de medidas de proximidad hace parecer difícil la elección de una. No hay una regla general para definir una que sirva para todos los casos de análisis de escalamiento multidimensional, depende de muchos factores, tales como la naturaleza de los objetos, las características de las variables sobre las que se miden los objetos, la finalidad del análisis y la aplicabilidad del método de obtención, entre otras.

Capítulo 3

SOLUCIÓN CLÁSICA

3.1 TEOREMA FUNDAMENTAL

Sea A el conjunto de n objetos que se indicará abreviadamente por $\{1, 2, ..., i, ..., j, ...n\}$. Supóngase conocida una medida de disimilaridad o distancia $d(i, j) = d_{ij}$ para los objetos del conjunto A. Se llama *matriz de distancia* a la matriz D cuyos elementos son las medidas de disimilaridad entre los objetos: $D = [d_{ij}]$.

Una matriz de distancia es euclideana si la medida de distancia que la define lo es. Es decir si existe una configuración de puntos $x_1, x_2, ...x_n$ en algún espacio euclideano R^m, de tal manera que las distancias entre dichos puntos están dadas por los elementos de D.

El siguiente teorema habilita para saber cuando una matriz es euclideana y en dado caso describe un método para construir la configuración de puntos correspondiente. A esta construcción se le conoce como la **solución clásica** del análisis de escalamiento multidimensional.

TEOREMA 1. *Sea* $D = [d_{ij}]$ *una matriz de distancia y defínanse las matrices:* $A = [a_{ij}]$ *con* $a_{ij} = -\frac{1}{2}d_{ij}^2$ y $B = HAH$ *donde* $H = I_n - n^{-1}1_n 1_n^t$ *con* $1_n = (1, 1, ..., 1)^t$, *de modo que* $b_{ij} = a_{ij} - \bar{a}_{.j} - \bar{a}_{i.} + \bar{a}_{..}$. *Entonces* D *es euclideana si y sólo si* B *es semidefinida positiva.*

En particular el teorema establece lo siquiente :

a) Si D es la matriz de distancias euclideanas entre los puntos de una configuración $Z = (z_1, z_2, ..., z_n)^t$ entonces $b_{ij} = (z_i - \bar{z})^t(z_j - \bar{z})$, lo que equivale a decir que $B = (HZ)(HZ)^t$ y B es semidefinida positiva.

B se puede interpretar como la matriz de producto interno centrado para la configuración Z.

b) Inversamente, si B es semidefinida positiva de rango p entonces una configuración correspondiente a D se puede construir como sigue:

Sean $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ los eigenvalores positivos de B con sus correpondientes eigenvectores $X = (x_{(1)}, x_{(2)}, ..., x_{(p)})$ normalizados de modo que $x_{(k)}^t x_{(k)} = \lambda_k$ para $k = 1, ..., p$.

Entonces los puntos x_i con coordenadas $(x_{i1}, x_{i2}, ..., x_{ip})^t$ $i = 1, .., n$ definen una configuración donde las distancias entre ellos estan dadas por los elementos de D . Además esta configuración tiene centro de gravedad $\bar{x} = 0$ y B representa su matriz de producto interno.

Demostración.

a) Utilizando las definiciones de b_{ij} y a_{ij} se probará que $b_{ij} = (z_i - \bar{z})^t(z_j - \bar{z})$. Para la configuración Z se tiene que

$$(z_i - z_j)^t(z_i - z_j) = d_{ij}^2 = -2a_{ij}$$

entonces $-2a_{ij} = z_i^t z_i - z_j^t z_i - z_i^t z_j + z_j^t z_j$

$$= z_i^t z_i - 2z_i^t z_j + z_j^t z_j$$

Sumando sobre $i = 1, ..., n$ y dividiendo por $\frac{1}{n}$ se obtiene

$$-2\bar{a}_{.j} = \frac{1}{n}\sum_{i=1}^{n} z_i^t z_i - \frac{2}{n}\sum_{i=1}^{n} z_i^t z_j + \frac{1}{n}\sum_{i=1}^{n} z_j^t z_j$$
$$= \frac{1}{n}\sum_{i=1}^{n} z_i^t z_i - 2\overline{z}^t z_j + z_j^t z_j$$

Sumando sobre $j = 1, ..., n$ y dividiendo por $\frac{1}{n}$ se obtiene

$$-2\bar{a}_{i.} = \frac{1}{n}\sum_{j=1}^{n} z_i^t z_i - \frac{2}{n}\sum_{j=1}^{n} z_i^t z_j + \frac{1}{n}\sum_{j=1}^{n} z_j^t z_j$$
$$= z_i^t z_i - 2z_i^t\overline{z} + \frac{1}{n}\sum_{j=1}^{n} z_j^t z_j$$

Sumando sobre $i, j = 1, ..., n$ y dividiendo por $\frac{1}{n^2}$ se obtiene

$$-2\bar{a}_{..} = \frac{1}{n^2}\sum_{i,j=1}^{n} z_i^t z_i - \frac{2}{n^2}\sum_{i,j=1}^{n} z_i^t z_j + \frac{1}{n^2}\sum_{i,j=1}^{n} z_i^t z_j$$
$$= \frac{1}{n}\sum_{i=1}^{n} z_i^t z_i - 2\overline{z}^t\overline{z} + \frac{1}{n}\sum_{j=1}^{n} z_j^t z_j$$
$$= \frac{2}{n}\sum_{i=1}^{n} z_i^t z_i - 2\overline{z}^t\overline{z}$$

Sustituyendo en b_{ij}

$$b_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}$$
$$= -\tfrac{1}{2}z_i^t z_i + \tfrac{1}{2}z_j^t z_i + \tfrac{1}{2}z_i^t z_j - \tfrac{1}{2}z_j^t z_j - \left(-\tfrac{1}{2}z_i^t z_i + z_i^t\overline{z} - \tfrac{1}{2n}\sum_{j=1}^{n} z_j^t z_j\right)$$
$$- \left(-\tfrac{1}{2n}\sum_{i=1}^{n} z_i^t z_i + \overline{z}^t z_j - \tfrac{1}{2}z_j^t z_j\right) + \left(-\tfrac{1}{n}\sum_{i=1}^{n} z_i^t z_i + \overline{z}^t\overline{z}\right)$$

$$= -\tfrac{1}{2}z_i^t z_i + \tfrac{1}{2}z_j^t z_i + \tfrac{1}{2}z_i^t z_j - \tfrac{1}{2}z_j^t z_j + \tfrac{1}{2}z_i^t z_i - z_i^t\overline{z} + \tfrac{1}{2n}\sum_{j=1}^{n} z_j^t z_j$$
$$+\tfrac{1}{2n}\sum_{i=1}^{n} z_i^t z_i - \overline{z}^t z_j + \tfrac{1}{2}z_j^t z_j - \tfrac{1}{n}\sum_{i=1}^{n} z_i^t z_i + \overline{z}^t\overline{z}$$

$$= z_i^t z_j - z_i^t\overline{z} - \overline{z}^t z_j + \overline{z}^t\overline{z}$$

$$= (z_i - \overline{z})^t(z_j - \overline{z})$$

De esto se obtiene $B = \overline{X}\ \overline{X}^t \geq 0$ donde $\overline{X}^t = (z_1 - \bar{z}, z_2 - \bar{z}, ..., z_n - \bar{z})$

Así la parte a) queda demostrada.

b) Supóngase que $B \geq 0$ es de rango p y sea $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$, entonces por el Teorema espectral de descomposición existe una matriz ortogonal $V = (v_{(1)}, v_{(2)}, ..., v_{(p)})$ tal que $B = V\Lambda V^t$

$$= (V\Lambda^{\frac{1}{2}})(V\Lambda^{\frac{1}{2}})^t = XX^t$$

De aqui se obtiene que $b_{ij} = x_i^t x_j$ y B representa la matriz de producto interno de esta configuración .

En lo siguiente se prueba que los elementos de D determinan las distancias entre los puntos de la configuración.

$$\begin{aligned} d_2^2(x_i, x_j) &= (x_i - x_j)^t(x_i - x_j) = x_i^t x_i - x_i^t x_j - x_j^t x_i + x_j^t x_j \\ &= x_i^t x_i - 2x_i^t x_j + x_j^t x_j \\ &= b_{ii} - 2b_{ij} + b_{jj} \\ &= a_{ii} - \bar{a}_{i\cdot} - \bar{a}_{\cdot i} + \bar{a}_{\cdot\cdot} - 2a_{ij} + 2\bar{a}_{i\cdot} + 2\bar{a}_{\cdot j} - 2\bar{a}_{\cdot\cdot} \\ &\qquad + a_{jj} - \bar{a}_{j\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot} \end{aligned}$$

como $a_{ij} = a_{ji}$, entonces $\bar{a}_{i\cdot} = \bar{a}_{j\cdot}$ y se tiene que:

$$\begin{aligned} d_2^2(x_i, x_j) &= a_{ii} - 2a_{ij} + a_{jj} \\ &= -\tfrac{1}{2}d_{ii}^2 - 2(-\tfrac{1}{2}d_{ij}^2) + (-\tfrac{1}{2}d_{jj}^2) \\ &= d_{ij}^2 \end{aligned}$$

Finalmente nótese que:

$$\begin{aligned} H1_n &= (I_n - n^{-1}1_n 1_n^t)1_n = 1_n - n^{-1}1_n 1_n^t 1_n \\ &= 1_n - n^{-1}1_n = 1_n - 1_n = 0_n \end{aligned}$$

Por lo que $B1_n = HAH1_n = 0_n = 0(1_n)$ y 1_n es un eigenvector de B correspondiente al eigenvalor $\lambda = 0$. Esto implica que 1_n es ortogonal a las columnas de X: $x_{(i)}^t 1_n = 0$ para $i = 1, ..., n$.

De esto se deduce que

$$n\overline{x} = \sum_{i=1}^{n} x_i = X^t 1_n = (x_{(1)}^t 1_n, ..., x_{(n)}^t 1_n)^t = 0_n$$

Asi que el centro de gravedad de esta configuración está en el origen.

Y con esto termina la demostración del teorema 1.

3.1.1 Observaciones

1.- La matriz X puede ser visualizada como sigue, en términos de los eigenvectores de B y los puntos correspondientes.

	λ_1	λ_2	...	λ_p	*Notación Vectorial*
P_1	x_{11}	x_{12}		x_{1p}	x_1^t
P_2	x_{21}	x_{22}	...	x_{2p}	x_2^t
$\vdots$	$\vdots$			$\vdots$	$\vdots$
P_n	x_{n1}	x_{n2}		x_{np}	x_n^t
Notación Vectorial	$x_{(1)}$	$x_{(2)}$	...	$x_{(p)}$	

Centro de gravedad:

$$\overline{x}_i = \frac{1}{p}\sum_{h=1}^{p} x_h = 0 \quad \text{para} \quad i = 1, ..., n \quad \text{y} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0_n$$

En resumen, el $i - ésimo$ renglón de X contiene las coordenadas del $i - ésimo$ punto de la configuración, mientras que la $j - ésima$ columna de X contiene el eigenvector correspondiente al eigenvalor λ_j.

2.- Geométricamente, si B es la matriz de producto interno centrado para la configuración Z, entonces $b_{ii}^{1/2}$ es igual a la distancia entre z_i y $\overline{z}$, y $b_{ij}/(b_{ii}b_{jj})^{1/2}$ es igual al coseno del ángulo subtendido desde $\overline{z}$ entre z_i y z_j.

3.- El vector 1_n es eigenvector de B sea D euclideana o no.

4.- Nótese que la solución que se obtiene no es única, un cambio del origen y una rotación o una reflexión no cambiaría las distancias entre los puntos, por ejemplo supóngase que L es una matriz ortogonal de $(p \times p)$, entonces: Lx_i también proporciona una solución.

$$
\begin{aligned}
d_2^2(Lx_i, Lx_j) &= (Lx_i - Lx_j)^t(Lx_i - Lx_j) \\
&= (x_i^tL^t - x_j^tL^t)(Lx_i - Lx_j) \\
&= x_i^tL^tLx_i - x_j^tL^tLx_i - x_i^tL^tLx_j + x_j^tL^tLx_j \\
&= x_i^tx_i - x_j^tx_i - x_i^tx_j + x_j^tx_j \\
&= d_2^2(x_i, x_j)
\end{aligned}
$$

Sin embargo, el problema del cambio de origen se soluciona al hacer que el centro de gravedad sea igual a cero; y la posibilidad de rotar o reflejar la solución es favorable para la interpretación de la misma.

5) Este resultado fue probado primero por Schoenberg(1935) y, Young y Householder (1938). Su utilización como base del análisis de escalamiento multidimensional se debe a Torgerson (1958).

3.1.2 Similaridades

En algunas situaciones no se comienza con una matriz de distancia sino con una de similaridad $C = (c_{ij})$. Para utilizar los procedimientos del escalamiento multidimensional clásico se deben transformar las similaridades en distancias, la llamada transformación estándar es muy útil y está dada por:

$$d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{\frac{1}{2}}$$

Si la matriz de similaridad es semidefinida positiva entonces la matriz de distancia que se obtiene resulta ser euclideana, como lo establece el siguiente teorema.

TEOREMA 2. *Si* $C \geq 0$ *entonces* $D = (d_{ij})$ *definida por*

$$d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{\frac{1}{2}}$$

es euclideana con matriz de producto interno centrado $B = HCH$.

Demostración.

Primero obsérvese que: como $C \geq 0$ entonces $x^tCx \geq 0$ para todo $x \in R^n$. En particular si x es un vector con un 1 en el lugar $i - ésimo$, un -1 en el lugar $j - ésimo$ y ceros en los demás lugares, se tiene:

$$d_{ij}^2 = c_{ii} - 2c_{ij} + c_{jj} = x^tCx \geq 0$$

Con lo que se prueba que d_{ij} está bien definida.

Ahora, sean las matrices A y B como en el teorema 1. Dado que HCH también es semidefinida positiva, es suficiente probar que $B = HCH$ para deducir que D es euclideana con matriz de producto interno centrado HCH.

Los elementos de $B = HAH$ se pueden escribir, utilizando su definición y la de d_{ij}, como sigue:

$$b_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$$

$$= -\tfrac{1}{2}d_{ij}^2 - \tfrac{1}{n}\sum_{h=1}^{n}\left(-\tfrac{1}{2}d_{ih}^2\right) - \tfrac{1}{n}\sum_{h=1}^{n}\left(-\tfrac{1}{2}d_{hj}^2\right) + \tfrac{1}{n^2}\sum_{h,g=1}^{n}\left(-\tfrac{1}{2}d_{hg}^2\right)$$

de donde se tiene que

$$-2b_{ij} = d_{ij}^2 - \tfrac{1}{n}\sum_{h=1}^{n} d_{ih}^2 - \tfrac{1}{n}\sum_{h=1}^{n} d_{hj}^2 + \tfrac{1}{n^2}\sum_{h,g=1}^{n} d_{hg}^2$$

$$= c_{ii} - 2c_{ij} + c_{jj} - \tfrac{1}{n}\sum_{h=1}^{n}(c_{hh} - 2c_{hj} + c_{jj}) - \tfrac{1}{n}\sum_{h=1}^{n}(c_{ii} - 2c_{ih} + c_{hh})$$
$$+\tfrac{1}{n^2}\sum_{h,g=1}^{n}(c_{hh} - 2c_{hg} + c_{gg})$$

$$= c_{ii} - 2c_{ij} + c_{jj} - \tfrac{1}{n}\sum_{h=1}^{n} c_{hh} + 2\bar{c}_{\cdot j} - c_{jj} - c_{ii} + 2\bar{c}_{i\cdot} - \tfrac{1}{n}\sum_{h=1}^{n} c_{hh}$$
$$+\tfrac{1}{n}\sum_{h=1}^{n} c_{hh} - \tfrac{2}{n}\bar{c}_{\cdot\cdot} + \tfrac{1}{n}\sum_{g=1}^{n} c_{gg}$$
$$= -2c_{ij} + 2\bar{c}_{\cdot j} + 2\bar{c}_{i\cdot} - 2\bar{c}_{\cdot\cdot}$$

Finalmente se obtiene que

$$b_{ij} = c_{ij} - \bar{c}_{\cdot j} - \bar{c}_{i\cdot} + \bar{c}_{\cdot\cdot} \qquad \text{y} \qquad B = HCH \geq 0\,.$$

Queda entonces demostrado el teorema 2.

3.2 SOLUCIÓN CLÁSICA EN K DIMENSIONES

Algunas veces la matriz de distancia D no es euclideana y entonces algunos de los eigenvalores de B son negativos. En este caso no se puede escribir V como $V^{1/2}V^{1/2}$ y el teorema fundamental no se cumple.

Por otro lado, aún cuando la matriz D sea euclideana, la dimensión del espacio en el que puede ser representada, generalmente resulta demasiado grande para ser de interés práctico.

En ambos casos, se puede construir una configuración en un espacio R^k cuyas coordenadas están determinadas k eigenvectores de B, correspondientes a los primeros k eigenvalores de B, suponiendo que éstos son positivos y comparativamente grandes y que los restantes $p-k$ eigenvalores están cercanos a cero (positivos o negativos).

Cuando se obtiene una configuración de esta manera, a los puntos también suele llamarseles las *coordenadas principales de X en k dimensiones.* Y, a esta configuración se le concoce como la **solución clásica en k dimensiones** del análisis de escalamiento multidimensional.

3.2.1 Propiedades óptimas

Dada una matriz de distancia $D = [d_{ij}]$ el objetivo del análisis de escalamiento multidimensional clásico es encontrar una configuración $\widehat{X}$ en un espacio euclideano de dimensión pequeña, de manera que las distancias entre los puntos estén dadas por: $\hat{d}_{ij} = (\hat{x}_i - \hat{x}_j)^t(\hat{x}_i - \hat{x}_j) = d(i,j)$

El sombrero ($\widehat{\ }$) se utiliza en esta parte para indicar las distancias $\widehat{D}$ de la configuración $\widehat{X}$ que ajustan a las verdaderas distancias D. Sea $\hat{B} = \widehat{X}\widehat{X}^t$ la matriz de producto interno centrada ajustada.

Sea X la configuración en R^p y $L = (L_1, L_2)_{pxp}$ una matriz ortogonal. con

L_1 de orden $p \times k$. Entonces XL representa una proyección de la configuración X en el subespacio de R^p expandido por las columnas de L_1 . Se puede pensar en $\widehat{X} = XL$ como la configuración ajustada de dimensión k.

Como L es ortogonal, las distancias entre los renglones de X son las mismas que entre los de XL:

$$d_{ij}^2 = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2 = \sum_{h=1}^{p} (x_i^t 1_h - x_j^t 1_h)^2$$

donde 1_h es un vector con un 1 en el lugar $h - ésimo$ y ceros en los demás lugares.

Si se denotan las distancias entre los renglones de XL_1 por $\widehat{D}$ entonces

$$\hat{d}_{ij}^2 = \sum_{h=1}^{k} (x_i^t 1_h - x_j^t 1_h)^2$$

De donde se observa que $\hat{d}_{ij} \leq d_{ij}$, lo que significa que proyectando una configuración se reducen las distancias entre los puntos.

Una medida de discrepancia entre la configuración original X y la configuración proyectada $\widehat{X}$ está dada por:

$$\Phi = \sum_{i,j=1}^{n} (d_{ij}^2 - \hat{d}_{ij}^2)$$

Entonces, la solución clásica del análisis de escalamiento multidimensional en k dimensiones tiene las siguientes propiedades óptimas.

TEOREMA 3. *Sea D una matriz de distancia euclideana correspondiente a la configuración X en R^p y sea k un número fijo $(1 \leq k < p)$. Entonces de entre todas las proyecciones XL de X en subespacios de R^p de dimensión k la cantidad Φ se minimiza cuando X se proyecta sobre sus coordenadas principales en k dimensiones.*

Demostración.

Usando la definición de d_{ij}^2 y de $\hat{d}_{ij}^2$, se puede escribir Φ como sigue

$$\begin{aligned}\Phi &= \sum_{i,j=1}^{n}\sum_{h=k+1}^{p}(x_i^t 1_{(h)} - x_j^t 1_{(h)})^2 \\ &= tr\ L_2^t\left(\sum_{i,j=1}^{n}(x_i - x_j)(x_i - x_j)^t\right)L_2 \\ &= 2n^2\ tr\ L_2^t S L_2\end{aligned}$$

Esto último porque:

$$\begin{aligned}\sum_{i,j=1}^{n}(x_i - x_j)(x_i - x_j)^t &= \sum_{i,j=1}^{n}(x_i x_i^t - x_i x_j^t - x_j x_i^t + x_j x_j^t) \\ &= \sum_{i,j=1}^{n} x_i x_i^t - \sum_{i,j=1}^{n} x_i x_j^t - \sum_{i,j=1}^{n} x_j x_i^t + \sum_{i,j=1}^{n} x_j x_j^t \\ &= n\sum_{i=1}^{n} x_i x_i^t - \sum_{j=1}^{n} n\overline{x}x_j^t - \sum_{i=1}^{n} n\overline{x}x_i^t + n\sum_{j=1}^{n} x_j x_j^t \\ &= 2n\sum_{i=1}^{n} x_i x_i^t - 2n\sum_{i=1}^{n}\overline{x}x_i^t \\ &= 2n\sum_{i=1}^{n}(x_i x_i^t - \overline{x}x_i^t - x_i\overline{x}^t + x_i\overline{x}^t) \\ &= 2n\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^t \\ &= 2n(nS)\end{aligned}$$

Sean $\lambda_1, ..., \lambda_p$ los eigenvalores de nS con sus correspondientes eigenvectores estándarizados $V = (v_{(1)}, ..., v_{(p)})$, entonces Φ se puede escribir

$$\Phi = 2n\ tr\ F_2^t \Lambda F_2 \quad \text{con} \quad F_2 = V^t L_2 \quad \text{matriz columna ortonormal}$$

Utilizando que Φ se minimiza cuando $F_2 = (0, I_{p-k})^t$, esto es, cuando $L_2 = (v_{(k+1)}, ..., v_{(p)})$. Se tiene entonces que las columnas de L_1 expanden el espacio de los primeros k eigenvectores de nS; y XL_1 representa las coordenadas principales de X en k dimensiones.

Nótese que para esta proyección de coordenadas principales se tiene:

$$\Phi = 2n(\lambda_{(k+1)}, ..., \lambda_{(p)})$$

Finalizando la demostración.

Cuando D no necesariamente es euclideana es más conveniente trabajar con la matriz $B = HAH$. Si $\widehat{X}$ es una configuración ajustada con matriz de producto interno centrado $\widehat{B}$, entonces la medida de discrepancia entre B y $\widehat{B}$ está dada (Mardia 1978) por:

$$\Psi = \sum_{i,j=1}^{n} (b_{ij} - \hat{b}_{ij})^2 = tr\ (B - \widehat{B})^2$$

Para esta medida también se puede probar que la solución clásica del escalamiento multidimensional es óptima.

TEOREMA 4. *Si D es una matriz de distancia (no necesariamente euclideana), entonces para un número k fijo Ψ se minimiza sobre todas las configuraciones $\widehat{X}$ en k dimensiones cuando se obtiene de la solución clásica del análisis de escalamiento multidimensional.*

Demostración.

Sean $\lambda_1 \geq ... \geq \lambda_n$ los eigenvalores de B, algunos de los cuales pueden ser negativos, con sus correspondientes eigenvectores estándarizados $V = (v_{(1)}, ..., v_{(n)})$. Por simplicidad supóngase que $\lambda_k > 0$.

Si $\widehat{\lambda}_1 \geq ... \geq \widehat{\lambda}_n \geq 0$ son los eigenvalores de $\widehat{B}$. Por el teorema espectral de descomposición se puede escribir:

$$V^t \widehat{B} V = G \widehat{\Lambda} G^t \qquad \text{con } G \text{ ortogonal}$$

entonces:

$$\Psi = tr\ (B - \widehat{B})^2 = tr\ V^t (B - \widehat{B}) V V^t (B - \widehat{B}) V$$

$$= tr\ (V^t B V - V^t \widehat{B} V)(V^t B V - V^t \widehat{B} V)$$

$$= tr\ (\Lambda - G\hat{\Lambda}G^t)(\Lambda - G\hat{\Lambda}G^t)$$

Obsérvese que para $\hat{\Lambda}$ fija Ψ se minimiza para $G = I$ de tal manera que

$$\Psi = \sum_{h=1}^{n} (\lambda_h - \hat{\lambda}_h)^2$$

Como $\widehat{X} \subset R^k$, $B = H\widehat{X}\widehat{X}^tH$ tendrá al menos k eigenvalores diferentes que sean positivos, y se puede ver que Ψ se minimiza si:

$$\hat{\lambda}_h = \begin{cases} \lambda_h & h = 1, ..., k \\ 0 & h = k+1, ..., n \end{cases}$$

De donde $\widehat{B} = V_1 \Lambda V_1^t$ con $V_1 = (v_{(1)}, ..., v_{(k)})$ y $\Lambda = diag(\lambda_1, ..., \lambda_k)$. Entonces $\widehat{X}$ puede ser tomada igual que $V_1\Lambda_1^{1/2}$, la solución clásica en k dimensiones del escalamiento multidimensional.

Nótese que el valor mínimo de Ψ está dado por:

$$\Psi = \lambda_{k+1}^2 + ... + \lambda_p^2$$

Termina con esto la demostración del teorema 4.

Los dos teoremas anteriores sugieren una posible *medida de ajuste* para la *proporción explicada de una matriz de distancia* por la solución clásica en k dimensiones.

Supóngase que $\lambda_k > 0$, entonces quedan definidas dos medidas como:

$$\alpha_{1k} = \left(\frac{\sum_{h=1}^{k} \lambda_h}{\sum_{h=1}^{n} /\lambda_h/} \right) \times 100\ \%$$

$$\alpha_{2k} = \left(\frac{\sum_{h=1}^{k} \lambda_h}{\sum_{h=1}^{n} \lambda_h^2} \right) \times 100\ \%$$

Se necesita valor absoluto o elevar al cuadrado porque los eigenvalores más pequeños pueden ser negativos.

Si se desea una solución en cierto número de dimensiones p^* $(p^* \leq k)$, entonces se puede examinar el conjunto de eigenvectores asociados con los p^* eigenvalores más grandes. La medida de ajuste de la representación obtenida estaría dada de manera similar con los índices arriba citados, sustituyendo el valor de k por el de p^*.

Capítulo 4

CLASIFICACIÓN

Existen diversos tipos de análisis de escalamiento multidimensional. Una primera clasificación distingue entre el análisis *métrico* y el *no métrico*. La diferencia proviene de la naturaleza de los datos de entrada, es decir, las medidas de proximidad. El modelo métrico asume que se dispone de variables cuantitativas medidas en escalas intervalar o de razón, mientras que el no métrico supone que los datos son cualitativos, principalmente ordinales.

Se puede distinguir también el análisis no ponderado del *ponderado*, el cual incorpora en el modelo la información debida a las diferencias individuales de los sujetos. Este modelo supone que existe un espacio de cierta dimensión común a todos los sujetos en el cual se pueden representar los n objetos , pero que las distancias entre los puntos en dicho espacio difieren de un sujeto a otro de acuerdo a la importancia o peso que cada uno le da a las dimensiones.

4.1 MODELO MÉTRICO

Este tipo de modelos supone que los datos están medidos en una escala intervalar o de razón. De aqui que existe una forma funcional exacta (lineal, cuadrática, etc.) que relaciona las medidas de proximidad con las distancias. El análisis de escalamiento métrico supone directamente que existe una relación lineal.

Nótese que, por ejemplo, si la matriz de proximidad tiene elementos que son realmente distancias entre los objetos con escala de razón, entonces el uso de un procedimiento métrico produce una solución en la que las distancias entre los puntos de la configuración obtenida están en la misma razón que las distancias originales.

El primer método aplicable del **análisis de escalamiento multidimensional métrico** se debe a Torgerson (1952), quien trabajó sobre las investigaciones de Young y Householder (1938,1941). Este método produce la solución clásica que se analizó en el capítulo anterior, aunque las hipótesis de Torgerson eran más restrictivas que las presentadas.

La suposición fundamental de Torgerson es que las medidas de disimilaridad son en realidad distancias euclideanas, es decir que:

$$\delta_{ij} = d_{ij} = d_2(i,j) = \left[\sum_{h=1}^{n}(x_{ih} - x_{jh})^2\right]^{1/2}$$

Para obtener la solución clásica del análisis de escalamiento multidimensional escencialmente lo que se hace es una reconstrucción algebraica a partir de la matriz de proximidad que representa exacta o aproximadamente distancias euclideanas, para encontrar la configuración de puntos en el espacio euclideano de dimensión dada, que le corresponde.

4.1.1 Un algoritmo práctico

La mecánica del análisis métrico se puede resumir con el siguiente algoritmo, suponiendo que se tiene una matriz $D = [d_{ij}]$ que representa exacta o aproximadamente distancias euclideanas entre los objetos.

a) Se construye la matriz $A = [a_{ij}] = \left[-\frac{1}{2}d_{ij}^2\right]$

b) Se construye la matriz $B = HAH$ donde $H = I_n - \frac{1}{n}1_n 1_n^t$ y $1_n = (1, ..., 1)$ de modo que los elementos de B quedan definidos por:

$$b_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$$

c) Se obienen los k eigenvalores positivos más grandes de B

$$\lambda_1, \lambda_2, ..., \lambda_k \qquad (k \leq n)$$

con sus correspondientes eigenvectores normalizados

$$X = \left[x_{(1)}, x_{(2)}, ..., x_{(k)}\right] \quad \text{donde} \quad x_{(i)}^t x_{(i)} = 1$$

d) Se obtiene la configuración buscada, dada por los renglones de la matriz X:

$$x_1^t, x_2^t, ..., x_k^t$$

Los posibles problemas y las propiedades óptimas de la solución clásica ya se han discutido en el capítulo anterior.

4.1.2 Otros modelos métricos

El algoritmo de Torgerson tiene hipótesis muy restrictivas. Un modelo ligeramente menos restrictivo es el que supone que $\delta_{ij} = d_{ij} + c$, donde c es una constante aditiva.

Una forma de analizar estos datos podría ser estimar primero el valor de c,

sustraerlo de cada disimilaridad δ_{ij} y entonces aplicar a los nuevos datos $(\delta_{ij} - c)$ el análisis clásico del escalamiento multidimensional.

Al problema de estimar el valor de c se le conoce como el *problema de la constante aditiva*. Algunos autores han propuesto diversas formas de resolver este problema, dando lugar a diferentes modelos métricos

Una opción para estimar el valor de c es como sigue:

$$\hat{c} = (-1)\max_{(i,j,h)}(\delta_{hj} - \delta_{hi} - \delta_{ij})$$

La constante aditiva estimada $\hat{c}$ es el valor más pequeño que se puede sustraer de la medida de proximidad δ_{ij} que garantiza que los datos transformados satisfacen la desigualdad del triángulo. Esto es, si se definen las nuevas medidas de proximidad γ_{ij} por: $\gamma_{ij} = 0$ si $i = j$ y $\gamma_{ij} = \delta_{ij} - \hat{c}$ si $i \neq j$, entonces estas medidas satisfacen la desigualdad del triángulo.

4.2 MODELO NO MÉTRICO

El fundamento teórico del **análisis de escalamiento multidimensional no métrico** lo proporcionó Kruskal (1964) y ha constituido la base de casi todo el trabajo posterior en esta área. Fue precisamente Kruskal quién le dió el nombre a este análisis. La hipótesis fundamental de este modelo es que las medidas de proximidad (básicamente disimilaridades) unicamente están relacionadas con las distancias entre los puntos mediante una función monótona (ordenada), a diferencia del modelo métrico que supone una forma funcional exacta. Esto es congruente con la suposición de que los datos están medidos en una escala ordinal ya que sólo se utiliza su relación de

orden. El algoritmo consiste en construir una configuración de puntos que minimice cierta medida de bondad de ajuste sobre dicha hipótesis fundamental.

Como antecedente del trabajo de Kruskal está el desarrollado por Shepard (1962), quien propuso un método para estimar las coordenadas de los puntos suponiendo que las medidas de disimilaridad y las distancias están relacionadas mediante una función monótona. Específicamente su algoritmo asumía que:

$$\delta_{ij} = f(d_{ij}) = f\left([\sum(x_{ih} - x_{jh})^2]^{1/2}\right)$$

con f tal que $d_{ij} < d_{rs} \Rightarrow f(\, d_{ij}) \leq f(\, d_{rs})$. Ejemplos de este tipo de funciones son las lineales, la exponencial y la logarítmo.

A diferencia del algoritmo de Shepard, el de Kruskal permite que las distancias sean euclideanas o no, siempre que pertenezcan a la métrica de Minkowski.

4.2.1 Medidas de bondad de ajuste

Un factor importante del análisis de escalamiento multidimensional no métrico es la medida de bondad de ajuste que indica el grado en que las distancias entre los puntos que representan a los objetos reproducen el orden de las disimilaridades.

La mayoría de los algoritmos de escalamiento no métrico, incluyendo el de Kruskal, calcula tres conjuntos de parámetros sobre los que se basa la medida de bondad de ajuste. El primero y más importante contiene las coordenadas x_{ik} de los puntos de la configuración. El segundo contiene las distancias entre los puntos, calculadas a partir de las coordenadas como sigue:

$$d_{ij} = [\sum (x_{ik} - x_{jk})^p]^{1/p}$$

El tercer conjunto contiene ciertas variables "dummy" que la mayoría de los autores llama disparidades. Las disparidades $\hat{d}_{ij}$ son valores calculados de manera que estén tan próximos a las distancias d_{ij} como sea posible y relacionadas con las disimilaridades originales de manera monótona, es decir tales que:

$$\delta_{ij} < \delta_{rs} \;\Rightarrow\; \hat{d}_{ij} \leq \hat{d}_{rs}$$

Este método no trabaja directamente con las disimilaridades δ_{ij} originales, sólo utiliza su relación de orden a través de las disparidades $\hat{d}_{ij}$.

Hay dos formas de plantear la condición de monotonicidad en caso de empate, es decir, en caso de que $\delta_{ij} = \delta_{rs}$. La primera forma es llamada comúnmente *planteamiento primario* y consiste en no restringir la relación de orden de las disparidades, en este caso puede ocurrir que: $\hat{d}_{ij} < \hat{d}_{rs}$, $\hat{d}_{ij} = \hat{d}_{rs}$, o $\hat{d}_{ij} > \hat{d}_{rs}$. La segunda forma es llamada *planteamiento secundario* y consiste en restringir a las disparidades a ser iguales en caso de que sus proximidades correspondientes lo sean, es decir, que se cumpla la condición adicional: $\delta_{ij} = \delta_{rs} \;\Rightarrow\; \hat{d}_{ij} = \hat{d}_{rs}$.

Stress

Supóngase conocida una configuración $X = (x_1, x_2, ..., x_n)$ en R^k con $x_i = (x_{i1}, x_{i2}, ..., x_{ik})^t$ para $i = 1, ...n$. La medida de bondad de ajuste propuesta por Kruskal es llamada *Stress*, aunque algunas veces se le llama Stress fórmula uno para diferenciarla de una variación de la misma llamada Stress fórmula dos. Estas médidas son:

$$S_1 = \left[\frac{\sum(\hat{d}_{ij} - d_{ij})^2}{\sum d_{ij}^2}\right]^{1/2} \quad \text{y}$$

$$S_2 = \left[\frac{\sum(\hat{d}_{ij} - d_{ij})^2}{\sum(d_{ij}^2 - \bar{d}_{..})^2}\right]^{1/2} \quad \text{donde } \bar{d}_{..} = \frac{1}{n^2}\sum_{i,j} d_{ij}$$

S_1 y S_2 difieren únicamente en la constante de normalización que aparece en el denominador en ambas fórmulas. S_1 utiliza la suma de cuadrados de las distancias. S_2 utiliza una cantidad proporcional a la varianza de las distancias, la suma de cuadrados de las desviaciones con respecto a la media.

El stress mide cuanto se desvían las disparidades de las distancias. Es invariante bajo cambios de escala (alargamientos o contracciones uniformes) y bajo movimientos rígidos (rotaciones, traslaciones y reflexiones), debido a que está normalizada.

S-Stress

Los algoritmos de Young emplean otra medida de bondad de ajuste, llamada *S-Stress* para diferenciarla del Stress de Kruskal. Igual que el Stress, el S-Stress tiene dos variantes llamadas fórmula uno y dos. Tales medidas son:

$$SS_1 = \frac{[\sum(\hat{d}_{ij}^2 - d_{ij}^2)^2]^{1/2}}{\sum(d_{ij}^2)^2} \quad \text{y}$$

$$SS_2 = \left[\frac{\sum(\hat{d}_{ij}^2 - d_{ij}^2)^2}{\sum(d_{ij}^2 - \bar{d}_{..})^2}\right]^{1/2} \quad \text{donde } \bar{d}_{..} = \frac{1}{n^2}\sum_{i,j} d_{ij}^2$$

La diferencia con el Stress de Kruskal estriba en que el S-Stress utiliza los cuadrados de las disparidades y de las distancias.

Coeficiente de alienación

Guttman (1968) propuso una tercera medida de ajuste para el análisis de escalamiento no métrico. Primero define un coeficiente de monotonicidad:

$$\mu = \frac{\sum(\hat{d}_{ij} - d_{ij})}{(\sum \hat{d}_{ij}^2)(\sum d_{ij}^2)}$$

Escencialmente μ mide el grado de asociación ordinal entre los medidas originales

y las distancias, siendo también una medida de ajuste. El *coeficiente de alienación* queda definido por:

$$\kappa = (1 - \mu^2)^{1/2}$$

Al igual que el Stress y el S-Stress, un valor grande de κ indica un mal ajuste, un valor pequeño indica un buen ajuste. El programa *SSA (Smallest Space Analysis)* produce coordenadas que maximizan el coeficiente de monotonicidad y minimizan el coeficiente de alienación.

4.2.2 Algoritmo

Una vez elegida la medida de bondad de ajuste S, se define una solución en k dimensiones como la configuración que minimiza su valor. El problema de encontrar esta configuración óptima se puede abordar de tres formas:

a) Intuitivamente se describe un método de aproximación. Comenzando con una configuración inicial, se mueven un poco los puntos para mejorar el ajuste, repitiendo el procedimiento hasta que la configuración no pueda ser mejorada.

b) Formalmente significa acercar los puntos x_i y x_j si $\hat{d}_{ij} < d_{ij}$ y alejarlos en caso contrario, de forma que $\hat{d}_{ij}$ se parece cada vez más a d_{ij}.

c) A nivel teórico, se trata del problema de minimizar una función de varias variables. De hecho $S = S(x_1, x_2, ..., x_n)$ es una función de nk variables, cada x_i tiene k coordenadas. El problema se resuelve con un método iterativo de análisis numérico llamado *método del gradiente* o de *"steepest descent"*. A nivel práctico se han desarrollado algoritmos de computadora que realizan este método exitósamente.

Aunque hay una gran variación entre ellos, a continuación se presenta uno que da la idea general de como funcionan, suponiendo sin pérdida de generalidad que la medida de ajuste es el Stress fórmula uno.

La mayoría de los algoritmos consisten en cuatro fases. En la primera se obtiene una configuración inicial. En la segunda se estándarizan las coordenadas de los puntos y las distancias entre ellos. La tercera, llamada fase no métrica, tiene como mayor propósito calcular las disparidades. La cuarta, llamada fase métrica, calcula las nuevas coordenadas de los puntos.

Después de calcular la configuración inicial, cada iteración consiste en una estandarización, una fase no métrica y una fase métrica. Al finalizar se calcula el Stress. Las iteraciones continúan hasta que el cambio en su valor de una iteración a la siguiente sea menor que algún valor establecido, por ejemplo 0.001.

Configuración inicial

Cualquier algoritmo del análisis de escalamiento no métrico comienza con la obtención de una configuración inicial, un paso realizado una sola vez. Los mejores métodos utilizan de alguna manera la solución clásica. En general, las primeras k coordenadas principales constituyen una solución en k dimensiones. Sean $x_1^o, x_2^o, ...x_n^o$ los puntos de la configuración inicial, las distancias entre ellos se calculan de manera usual como:

$$d_{ij}^{(o)} = \left[\sum_h \left(x_{ih}^o - x_{jh}^o\right)^2\right]^{1/2}$$

Estandarización

Después de calcular la configuración inicial comienza la primera iteración. Al comienzo de cada iteración se estándarizan las coordenadas de los puntos y las distancias. La experiencia indica que con esto se reduce la probabilidad de encontrar una solución degenerada

Si se utiliza el Stress fórmula uno, es conveniente estándarizar las distancias de manera que la suma de cuadrados $\sum_{i,j} d_{ij}^2$ sea igual a 1. Asi, S_1 se reduce a

$$S_1 = \left[\sum_{i,j} \left(\hat{d}_{ij} - d_{ij}\right)^2\right]^{1/2}$$

De aqui que minimizar S_1 es equivalente a minimizar $S = S_1^2 = \sum_{i,j} \left(\hat{d}_{ij} - d_{ij}\right)^2$

La estandarización puede realizarse, por ejemplo, multiplicando cada distancia por una constante adecuada. Para expresar las coordenadas en la misma escala que las distancias, deberán multiplicarse por la misma constante.

Fase no métrica

La fase no métrica utiliza las disimilaridades y las distancias estándarizadas de la iteración previa (o en su caso de la configuración inicial) para calcular las disparidades. Después de ordenar las disimilaridades en forma ascendente se requiere una serie de pasos para obtener las disparidades de cada iteración . Obsérvese que durante este proceso no se ajustan las coordenadas de los puntos o las distancias, únicamente las disparidades cambian en esta fase.

Antes de describir propiamente los pasos de esta fase, se establece la notación necesaria.

Sea $\widehat{d}_{ij}^{(c+1)}$ la disparidad para el par de objetos (i, j), calculado en la $(c+1)-ésima$ iteración. Sean $x_1^c, x_2^c, ..., x_n^c$ las coordenadas obtenidas en la iteración c $(c = 0, .., C)$. Si $c = 0$, se trata de las coordenadas de la configuración inicial. Finalmente, sea $d_{ij}^{(c)}$ la distancia en la iteración c, calculada como:

$$d_{ij}^{(c)} = \left[\sum_h \left(x_{ih}^c - x_{jh}^c\right)^2\right]^{1/2}$$

En la fase no métrica más común (Kruskal, 1964) las disparidades se calculan de modo que constituyan una transformación monótona de las disimilaridades originales. Obsérvese que el algoritmo no aplica simplemente una función monótona bien conocida (como un logaritmo o una potencia) a los datos. En lugar de eso, la fase no métrica consiste en una serie de pasos en los cuales cada disparidad $\widehat{d}_{ij}^{(c+1)}$ se iguala a la correspondiente distancia $d_{ij}^{(c)}$ o al promedio de varias de ellas. Las distancias de la iteración c son las disparidades iniciales de la iteración $(c+1)$.

El primer paso de la fase no métrica es arreglar las disimilaridades y sus correspondientes disparidades en forma ascendente. Surge una pequeña complicación cuando hay empates, es decir, cuando $\delta_{ij} = \delta_{rs}$ para algún par de puntos, pero es fácil de resolver. Si $\delta_{ij} = \delta_{rs}$ y $\widehat{d}_{ij}^{(c)} = \widehat{d}_{rs}^{(c)}$ entonces no hay problema en cuál precede. Sin embargo, si $\delta_{ij} = \delta_{rs}$ y $\widehat{d}_{ij}^{(c)} \neq \widehat{d}_{rs}^{(c)}$, entonces la disparidad asociada con la distancia $d^{(c)}$ más pequeña precede. En el primer paso de esta fase, las disparidades son iguales a las distancias obtenidas en la iteración anterior, en los pasos siguientes las disparidades son las calculadas en el paso previo.

Cada paso después del primero comienza con la división del conjunto de disparidades en bloques con disparidades iguales, si no las hay entonces cada una constituye su propio bloque.

El resto de cada paso consiste en comparar bloques adyacentes. Sea m $(m = 1, ..., M)$ el subíndice que distingue a los bloques, desde el primero que contiene a las disparidades más pequeñas $(m = 1)$, hasta el que contiene a las más grandes $(m = M)$. Comenzando con $m = 1$, los elementos del $m - ésimo$ bloque se comparan con los elementos del $(m + 1) - ésimo$. Si los elementos del $m - ésimo$ bloque son menores que los del $(m + 1) - ésimo$, entonces simplemente se prosigue con la comparación de los siguientes dos bloques. Si por el contrario, los elementos del $m - ésimo$ bloque son mayores que los del $(m + 1) - ésimo$, entonces se calcula la media aritmética de los elementos de ambos bloques y todos ellos se sustituyen por este valor. Al hacer esta sustitución los elementos del $m - ésimo$ y $(m + 1) - ésimo$ bloques son iguales por lo tanto se pueden reagrupar formando el nuevo $m - ésimo$ bloque. Se prosigue comparando este nuevo $m - ésimo$ bloque con el siguiente. El paso termina una vez que todos los bloques adyacentes hayan sido comparados. El resultado un nuevo conjunto de disparidades. Si no se reagruparon elementos en el último paso entonces termina la fase métrica de la iteración.

Sin embargo, si algunos elementos fueron reagrupados, entonces da incio un nuevo paso. Las disparidades obtenidas en el último paso de la fase son las disparidades $\widehat{d}_{ij}^{(c+1)}$ de la iteración $(c + 1)$.

Cuando termina esta fase las disparidades satisfacen la condición de monotonicidad. La fase no métrica es una aplicación de la regresión monotónica de Kruskal (1964) de las disimilaridades contra las distancias. Las disparidades de la última iteración son las disparidades finales $\widehat{d}_{ij}$ del análisis.

La fase no métrica propuesta por Guttman (1968) e incorporada en el programa

Smallest Space Analysis (*SSA*) es una alternativa de la fase no métrica de Kruskal (1964) ya descrita, y ha jugado un papel muy importante en la literatura del análisis de escalamiento multidimensional no métrico. Igual que la de Kruskal, la de Guttman utiliza las disimilaridades originales y las distancias calculadas en la iteración anterior para calcular las disparidades de cada iteración, pero ejecuta los cálculos de manera diferente.

En la fase no métrica de Guttman cada disparidad de la iteracion $(c+1)$ es igualada a una de las distancias de la iteración c. Específicamente, si el par de objetos (i,j) corresponde a la $r-ésima$ disimilaridad más pequeña δ_{ij}, entonces la disparidad correspondiente $\hat{d}_{ij}^{(c+1)}$ es igual a la $r-ésima$ distancia más pequeña (que no necesariamente es $d_{ij}^{(c)}$).

Fase métrica

Esta fase utiliza las disparidades $\hat{d}_{ij}^{(c+1)}$ de la fase no métrica recién concluida, las coordenadas $x_1^c, x_2^c, ..., x_n^c$ y las distancias $d_{ij}^{(c)}$ de la iteración previa para calcular el gradiente negativo de S_1 y las nuevas coordenadas $x_1^{(c+1)}, x_2^{(c+1)}, ..., x_n^{(c+1)}$, para las cuales se obtienen las distancias $d_{ij}^{(c+1)}$. Las disparidades permanecen sin cambio en esta fase.

Considerando que se llevó a cabo la estándarización se calcula el gradiente negativo de $S = \sum_{i,j} \left(\hat{d}_{ij} - d_{ij}\right)^2$. Es decir, se obtiene $\left(-\frac{\partial S}{\partial x_{11}}, ..., -\frac{\partial S}{\partial x_{ih}}, ..., -\frac{\partial S}{\partial x_{np}}\right)$ donde

$\frac{\partial S}{\partial x_{ih}} = \frac{\partial}{\partial x_{ih}} \left(\sum_{i,j} \left(\hat{d}_{ij} - d_{ij}\right)^2 \right) = -2 \sum_j \frac{\left(\hat{d}_{ij} - d_{ij}\right)}{d_{ij}} \left(x_{ih} - x_{jh}\right)$

Existen diversas formas de definir las nuevas coordenadas. Kruskal (1964) las

define como sigue: $x_{ih}^{(c+1)} = x_{ih}^{c} - 2\alpha \sum_{j} \frac{\left(d_{ij}^{(c)} - \hat{d}_{ij}^{(c+1)}\right)}{d_{ij}^{(c)}} \left(x_{ih}^{c} - x_{jh}^{c}\right)$

donde α es es la magnitud del movimiento en dirección del gradiente. El valor de α se desconoce y debe ser estimado en cada iteración.

Lingoes y Roskman (1973) proponen lo siguiente:

$$x_{ih}^{(c+1)} = x_{ih}^{c} - \frac{1}{n} \sum_{j} \left(1 - \frac{\hat{d}_{ij}^{(c+1)}}{d_{ij}^{(c)}}\right) \left(x_{ih}^{c} - x_{jh}^{c}\right)$$

Para evitar la división por cero el radio $\frac{\hat{d}_{ij}^{(c+1)}}{d_{ij}^{(c)}}$ se iguala a cero cuando $d_{ij}^{(c)} = 0$.

4.2.3 Problemas de cálculo

Cualquier algoritmo del análisis de escalamiento multidimensional no métrico comienza con el cálculo de una configuración inicial y la va modificando a través de varias iteraciones. En cada una de ellas las coordenadas de la configuración se ajustan de tal manera que se reduzca el valor de la medida de bondad de ajuste. Si todo va bien, las iteraciones continúan hasta que se obtiene una solución óptima.

Desafortunadamente, hay algunas cosas que pueden ir mal en el proceso iterativo. Estos problemas potenciales son llamados *problema del mínimo local, problema de la solución degenerada* y *problema de convergencia.* A continuación se analiza brevemente en que consiste cada uno.

Mínimo local

Una configuración para la cual la medida de ajuste sea mínimo es por definición un *mínimo local*, sin embargo puede ser que éste no sea un *mínimo global.* El método de "steepest descent" genera soluciones que proporcionan mínimos locales.El problema

del mínimo local es el de asegurar que la solución encontrada corresponde efectivamente al mínimo global.

Una primera forma de solucionar el problema es utilizando diferentes configuraciones iniciales para el algoritmo. En principio cada configuración inicial puede derivar en un mínimo local diferente. Considerando el valor más pequeño de éstos como el mínimo global, se tendría como solución la configuración que le corresponde.

Otra forma, y al parecer la mejor, de asegurar que la solución obtenida corresponda al mínimo global y no a un mínimo local es utilizar una buena configuración inicial. Algunas variantes de la solución clásica proveen los mejores métodos para obtener tal configuración. Aunque varían, los diversos programas estadísticos comienzan aplicando alguna transformación monótona a las disimilaridades para después aplicarles el análisis clásico y obtener la solución en k dimensiones.

La configuración obtenida por alguna variante de la solución clásica es llamada *configuración inicial racional*. Los algoritmos que utilizan configuraciones iniciales racionales resultan ser menos propensos al problema del mínimo local, que aquellos que utilizan cualquier otra configuración inicial no racional.

Solución degenerada

Una solución se dice degenerada cuando el número de puntos distintos en la configuración es pequeño comparado con el número de objetos a representar. La solución es degenerada porque algunos objetos se colapsan en un sólo punto en la gráfica que los representa.

Se deben siempre revisar las soluciones por degeneración, particularmente aquellas

en una o dos dimensiones. Shepard (1974) establece que las soluciones con un valor de stress muy cercano o igual a cero son generalmente soluciones degeneradas, por lo cual se debe hacer un esfuerzo extra al revisar alguna solución con un valor de ajuste sospechosamente bajo.

Frecuentemente una solución degenerada significa que la solución correcta debe buscarse en una dimensionalidad mayor, por lo que debe ser desechada.

Convergencia

De los tres problemas de cálculo, el de la falta de convergencia es el más fácil de solucionar, pero probablemente el más frecuente.

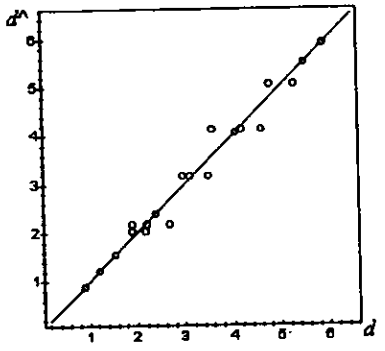
Debido a que se debe establecer el número máximo de iteraciones que se pueden llevar a cabo, en algunos casos no hay convergencia, lo que significa que el número de iteraciones requeridas para alcanzar la solución excede al establecido . La mayoría de los programas muestra un mensaje cuando esto ha ocurrido. El problema puede resolverse incrementando el número de iteraciones.

4.2.4 Diagramas de Shepard

Existen tres gráficas ampliamente discutidas en la literatura del análisis de escalamiento multidimensional que proporcionan aspectos importantes del modelo no métrico.

La primera es llamada por Guttman (1968) *diagrama de imagen*. Las distancias d_{ij} se toman como las abscisas y las disparidades $\hat{d}_{ij}$ como las ordenadas de los puntos en un plano cartesiano. Hay un punto en la gráfica por cada par de objetos, (aunque puede haber menos debido a que los valores de d_{ij} y $\hat{d}_{ij}$ de algún par de

objetos sean exactamente iguales a los de algún otro par). La siguiente figura muestra un ejemplo de este tipo de gráficas.

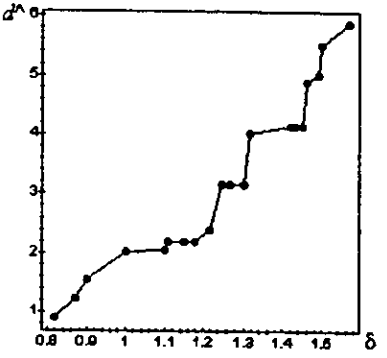


d vs. $\hat{d}$

Si los datos satisfacen perfectamente el modelo, en cuyo caso la medida de ajuste será igual a cero, los puntos caerán sobre la recta identidad (la recta con pendiente igual a 1, que pasa por el origen). Mientras más puntos se desvíen de esta línea, peor será el ajuste del modelo a los datos. Los puntos que se desvían más marcadamente, los llamados *"outliers"*, son los que menos satisfacen el modelo no métrico. Una cuidadosa inspección de ellos ayuda a descubrir desviaciones del modelo teóricamente importantes.

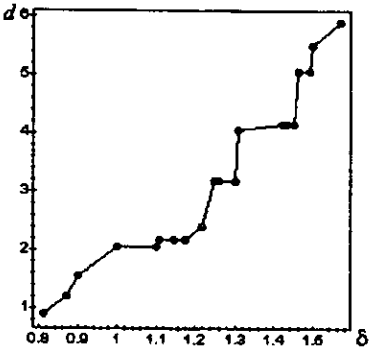
Los puntos de la segunda gráfica tienen como abscisas los valores de las disimilaridades originales δ_{ij} y como ordenadas los de las disparidades $\hat{d}_{ij}$. Nuevamente hay un punto por cada par de objetos (aunque puede haber menos debido a que dos pares de objetos tengan valores exactamente iguales). Esta gráfica es una representación gráfica de la función monótona f que relaciona las disimilaridades con las distancias.

La siguiente figura muestra un ejemplo.



δ vs. $\hat{d}$

La tercera gráfica también provee una representación gráfica de la función f. Se construye estableciendo las disimilaridades δ_{ij} como las abscisas de los puntos y las distancias d_{ij} como las ordenadas. Hay un punto en la gráfica por cada par de objetos, y puede haber menos.



δ vs. d

A estas dos últimas gráficas (δ_{ij} vs. $\hat{d}_{ij}$ y δ_{ij} vs. d_{ij}) se les llama, más generalmente, *diagramas de Shepard.*

4.2.5 Observaciones

En general, el análisis de escalamiento multidimensional no métrico, donde sólo se utiliza la relación de orden, provee espacios de dimensión baja mejor ajustados, que su contraparte métrico. Sin embargo, las soluciones obtenidas utilizando ambos modelos pueden ser muy similares. De hecho, la elección del modelo métrico o no métrico, es un tema de controversia tanto a nivel teórico como aplicado.

4.3 MODELO PONDERADO

El análisis métrico y el no métrico podrían llamarse modelos no ponderados (sin pesos) porque trabajan con una sola matriz, tal vez obtenida al promediar los valores de las matrices de varios de sujetos. A este tipo de análisis también suele llamársele de dos vías. En contraste los modelos ponderados trabajan con dos o más matrices, separando la información en común de todos los sujetos, de la información particular de cada uno de ellos. Es decir, el modelo ponderado incluye espacios adicionales que reflejan la manera en que cada sujeto percibe a los objetos. Por esta razón también se le conoce como análisis de tres vías. En particular, el **análisis de escalamiento multidimensional ponderado** incorpora explícitamente las diferencias individuales en el modelo métrico básico.

Con este modelo puede analizarse cómo varía la estructura debido a las diferencias de los sujetos, pero también cómo varía en un mismo sujeto a través del tiempo o debido a diferentes condiciones experimentales. En el primer caso se tendría una matriz por cada individuo, en el segundo una matriz por cada situación de observación.

El primer intento de incorporar las diferencias individuales a los métodos del análisis de escalamiento multidimensional se debe a Tucker y Messick (1963), quienes introdujeron el llamado *"análisis de puntos de vista"*, sin embargo este método apenas es más potente que realizar un análisis a cada uno de los sujetos.

Posteriormente, varios autores propusieron algunos métodos para resolver este problema. En la literatura del escalamiento multidimensional hay dos que han sobresalido. El primero, y al parecer el más exitóso, propuesto por Carroll y Chang (1970), incorpora en el modelo métrico euclideano los pesos que cada sujeto le da a las dimensiones. A este método se le conoce como *INDSCAL* (*INDividual diferences SCALing*) porque se le identifica con el programa de computadora del mismo nombre desarrollado por Carroll y Chang para llevarlo a cabo. En este trabajo se le llamará simplemente modelo euclideano ponderado.

El segundo método, desarrollado por Tucker (1972), es llamado modelo de *tres modas* (*three-mode model*). Este modelo supone que los sujetos difieren en la importancia que le dan a las dimensiones representadas en sus juicios, pero también en el grado de interacción entre ellas.

4.3.1 Modelo euclideano

Para llevar a cabo la discusión sobre este modelo es necesario establecer primero la notación adecuada. Sea $X = [x_{ik}]$ la matriz de coordenadas de la configuración en el espacio común a todos los sujetos, llamada la *matriz del grupo*. Este modelo supone que existe una matriz idiosincrática de coordenadas $X_s = [x_{iks}]$ para cada sujeto, donde x_{iks} es la coordenada del objeto i sobre la dimensión k para el sujeto s. Los elementos de la matriz de coordenadas del sujeto s están relacionados con los de la matriz del grupo mediante la ecuación: $x_{iks} = x_{ik}w_{ks}$ donde w_{ks} es un peso desconocido que el sujeto s atribuye a la dimensión k.

Si W_s es una matriz diagonal cuyos elementos son los pesos w_{ks}, entonces la relación queda establecida de forma matricial como:

$$X_s = XW_s$$

De acuerdo al modelo euclideano los juicios de disimilaridad de los sujetos se pueden expresar como una función de distancia euclideana de las coordenadas, esto es: $\delta_{ijs} = \left[\sum_k (x_{iks} - x_{jks})^2\right]^{1/2}$ donde δ_{ijs} es la disimilaridad del par de objetos (i, j) de acuerdo al juicio del sujeto s. Sustituyendo los valores de las coordenadas por su equivalente en coordenadas del grupo, se obtiene:

$$\delta_{ijs} = \left[\sum_k w_{ks}^2 (x_{ik} - x_{jk})^2\right]^{1/2}$$

Esto expresa la hipótesis fundamental del **modelo euclideano ponderado.**

Los pesos w_{ks} son llamados a veces pesos de importancia. Si w_{ks} crece, las diferencias entre los objetos sobre la dimensión k tienen mayor influencia en el juicio de disimilaridad entre los objetos i y j.

Algoritmo

Existen diversos algoritmos para ajustar el modelo euclideano ponderado (Bloxom, 1978; Carroll y Chang, 1970; De Leeuw y Pruzansky, 1978; Lingoes y Borg, 1978), a continuación se presenta una versión del algoritmo de Carroll y Chang con el propósito de dar una idea general de cómo funcionan tales algoritmos.

El primer paso es convertir la matriz de disimilaridad de cada sujeto en una de producto interno, mediante una transformación como la utilizada en la solución clásica: $b_{ijs} = -\frac{1}{2}(\delta^2_{ijs} - \delta^2_{i.s} - \delta^2_{.js} + \delta^2_{..s})$. Sea $B_s = [b_{ijs}]$ la matriz de producto interno para el sujeto s.

Para evitar que la matriz de algún sujeto tenga una influencia indebida en la solución final, cada B_s $(s = 1, ..., S)$ puede ser estándarizada de manera que las varianzas de todas ellas sean iguales.

El segundo paso del algoritmo es obtener una configuración inicial. Una forma de hacerlo (Schönemann, 1972) consiste en calcular la matriz promedio de producto interno: $\overline{B} = \frac{1}{S}\sum_{h=1}^{S} B_h$

Y, utilizando un análisis de componentes principales para $\overline{B}$, obtener una matriz $\widetilde{X}$ que cumpla la condición: $\widetilde{X}\widetilde{X}^t = \overline{B}$. Las primeras k columnas de $\widetilde{X}$ forman una excelente configuración inicial para la solución en k dimensiones.

El resto del algoritmo se compone de iteraciones. Cada una de ellas contiene dos fases, una para estimar los pesos w_{ik} y la otra para calcular las coordenadas x_{ik} de los puntos.

En la primera fase de cada iteración se construyen dos matrices A y C. La

matriz A tiene S renglones, uno por cada sujeto, y n^2 columnas, una por cada par de objetos. Los primeros n elementos del renglón s de A son las entradas del renglón 1 de B_s. Los segundos n elementos son las entradas del renglón 2, y asi sucesivamente. Básicamente, el renglón s de la matriz A es un vector que contiene todas las entradas en la matriz de producto interno del sujeto s.

La matriz C tiene K renglones, uno por cada dimensión, y n^2 columnas, una por cada posible par de objetos. El elemento en el renglón correspondiente a la dimensión k y en la columna que corresponde al par de objetos (i,j), se define como: $c_{k(i,j)} = \widehat{x}_{ik}\widehat{x}_{jk}$ donde $\widehat{x}_{ik}$ y $\widehat{x}_{jk}$ son las coordenadas calculadas en la iteración anterior (o provienen de la configuración inicial si se trata de la primera iteración).

Una vez que se han construido las matrices A y C, se obtiene la matriz de pesos $\widehat{W}^2$ mediante la ecuación:

$$\widehat{W}^2 = (CC^t)^{-1} CA^t$$

Los elementos de $\widehat{W}^2$, o más precisamente, sus raíces cuadradas w_{ks}, son los pesos estimados de la iteración.

Para comenzar la segunda fase de la iteración, se deben reconstruir las matrices A y C. La matriz A contiene nuevamente productos escalares. En esta fase A tiene n renglones, uno por cada objeto, y nS columnas, una por cada posible par (s,i) de sujetos y objetos. El elemento en el renglón i y en la columna correspondiente al sujeto s y al objeto j es b_{ijs}, el producto interno para los objetos i y j de la matriz de producto interno B_s del sujeto s.

La matriz C se reconstruye de manera que tenga K renglones, uno por cada

dimensión, y nS columnas, una por cada posible par de sujetos y objetos. El elemento en el renglón k y en la columna correspondiente al sujeto s y al objeto j se define como: $c_{k(s,j)} = \hat{w}^2_{ks}\hat{x}_{jk}$

Aqui $\hat{x}_{jk}$ y $\hat{w}^2_{ks}$ son los valores estimados en la iteración previa (o en la configuración inicial) y en la fase anterior de la misma iteración, respectivamente.

Una vez que se han reconstruido las matrices A y C, se obtiene la matriz $\widehat{X}$ cuyos elementos son las nuevas coordenadas, calculada mediante la ecuación:

$$\widehat{X} = AC^t(CC^t)^{-1}$$

Después de calcular las nuevas coordenadas termina la segunda fase y con ella la iteración.

El algoritmo descrito está diseñado para minimizar la suma de las discrepancias al cuadrado entre los productos escalares estimados y los predichos por el modelo:

$$F = \sum_{(i,j,s)} (b_{ijs} - \hat{b}_{ijs})^2$$

donde el producto escalar $\hat{b}_{ijs}$ se define como: $\hat{b}_{ijs} = \sum_k \hat{x}_{ik}\hat{x}_{jk}\hat{w}^2_{ks}$

Al final de cada iteración se calculan los valores de $\hat{b}_{ijs}$ y de F. Las iteraciones continuarán hasta que la reducción en el valor de F de una iteración a la siguiente sea menor que algún valor pequeño establecido, por ejemplo 0.001. Después de terminado el procedimiento, se pueden estándarizar las columnas de $\widehat{X}$ de manera que la varianza de los valores sobre cada dimensión sea igual a 1, entonces la matriz de pesos $\widehat{W}^2$ debe ser estimada una vez más, en una repetición extra de la fase uno.

Problemas de cálculo

La mayoría de los algoritmos utilizados para ajustar el modelo euclideano ponderado son iterativos y, por tanto, están sujetos a los problemas de mínimo local, de la solución degenerada y de convergencia discutidos en la sección análoga del modelo no métrico. Las soluciones a estos problemas se obtienen en la misma forma que para dicho modelo.

4.3.2 Modelo de tres modas

Carroll y Chang (1972), Harshman (1972) y Tucker (1972) propusieron un modelo ponderado más general que el recien discutido. El modelo que a continuación se presenta sigue la formulación de Tucker, es por eso que se utiliza su término **modelo de tres modas.**

Cómo en el modelo euclideano ponderado, se supone que existe una matriz X de coordenadas del grupo. Las disimilaridades se pueden expresar como una función de distancia euclideana de ellas. Cada sujeto tiene una matriz idiosincrática de coordenadas X_s, pero la función que relaciona X con X_s es más compleja en el modelo de tres modas. Mientras que en el modelo euclideano ponderado la relación se determina por $X_s = XW_s$, en el modelo tres modas se determina mediante $X_s = XW_sT_s$

Aqui, W_s es una matriz diagonal cuyos elementos son los pesos w_{ks}, como antes, y T_s es una matriz transformación. En otras palabras las coordenadas del espacio de cada sujeto se pueden obtener combinando una ponderación y una rotación. El modelo euclideano ponderado resulta ser un caso particular del de tres modas, en el

cual $T_s = I$ para todos los sujetos.

Considérese la matriz $R_s = T_s T_s^t$ con elementos $r_{kk's}$. El modelo de tres modas puede escribirse en términos de los parámetros w_{ks} y $r_{kk's}$ de los sujetos, como sigue:

$$\delta_{ijs} = \left[\sum_k w_{ks}^2 (x_{ik} - x_{jk})^2 + \sum_{(k,k')} w_{ks} w_{k's} r_{kk's} r'_{kk's} (x_{ik} - x_{jk})(x_{ik'} - x_{jk'}) \right]^{1/2}$$

Si $r_{kk's} = 0$ para todos los pares de dimensiones (k, k'), entonces la segunda suma en esta expresión desaparece y la ecuación se reduce a la del modelo euclideano ponderado. La suma del lado derecho sobre los pares de dimensiones (k, k') es la suma de las interacciones entre la diferencia de coordenadas sobre la dimensión k y sobre la dimensión k'. El parámetro $r_{kk's}$ caracteriza la interacción entre la diferencia de las coordenadas sobre las dimensiones k y k' para el sujeto s. El signo de $r_{kk's}$ indica la dirección de la interacción y el valor absoluto indica su magnitud.

La matriz de pesos de los sujetos W_s, y las matrices transformacion T_s no representan la descripción más concisa de las características de los sujetos. Tal descripción estaría dada por $2S$ matrices, una matriz transformación y una de pesos por cada uno de los S sujetos. De aqui que, Tucker (1972) desarrolló un método para estimar un conjunto de coordenadas en cierto espacio, llamado el espacio de los sujetos, para representarlos; tal conjunto esta contenido en una matriz $\hat{Z}$ $(S \times M)$, con elementos $\hat{z}_{sm}$. Las coordenadas en $\hat{Z}$ proporcionan una descripción dimensional cuantitativa de los sujetos, análoga a la que da X de los objetos.

$\hat{Z}$ tiene un renglón por cada sujeto y una columna por cada dimensión en el espacio de los sujetos. La dimensión M del espacio de los sujetos no necesariamente es igual

a K, la dimensión del espacio del grupo (donde están las coordenadas de los objetos). Esto significa que se deben tomar dos decisiones sobre dimensionalidad, una para el espacio de los objetos y otra para el de los sujetos.

4.3.3 Observaciones

De los modelos del análisis de escalamiento multidimensional el de tres modas es el más rico, e incluye el euclideano ponderado como caso particular. Este modelo proporciona una solución que describe las diferencias entre los sujetos, no sólo en términos de pesos para cada sujeto, sino también en términos de matrices transformación. Con el incremento en la riqueza del modelo de tres modas aumenta la complejidad. Cuando las configuraciones de los objetos y de los sujetos contienen varias dimensiones, la rotación y la interpretación de la solución suele ser un problema desafiante para el investigador.

Capítulo 5

DIMENSIONALIDAD, ROTACIÓN E INTERPRETACIÓN

Tres temas importantes en las aplicaciones reales del análisis de escalamiento multidimensional son (1) el número de dimensiones a conservar, (2) la rotación de la solución obtenida y (3) la interpretación del espacio obtenido y la correspondiente configuración.

5.1 DIMENSIONALIDAD

Los análisis presentados suponen que el número de dimensiones de la configuración buscada es conocido y está fijo. En la práctica, sin embargo, debe determinarse como parte del mismo análisis. La forma más viable es obtener soluciones en diferentes

dimensiones y elegir entre ellas con base en algunos criterios como: ajuste de los datos, interpretabilidad y reproducibilidad. Davison (1983) sugiere que si k^* es el valor a priori que el investigador considera apropiado, entonces se deben obtener soluciones en cada dimensión de $k^* - 3$ a $k^* + 3$. Si $k^* \leq 3$, haciendo $k^* - 3 \leq 0$, entonces de dimensión 1 a $k^* + 3$.

5.1.1 Ajuste de los datos

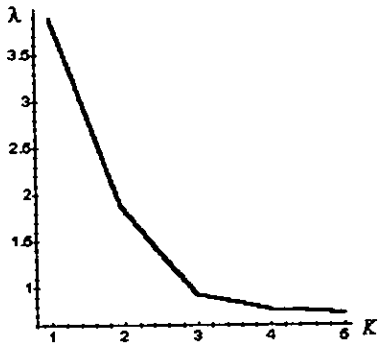
Modelo métrico

La solución clásica del análisis de escalamiento multidimensional es uno de los métodos en los cuales la medida de ajuste no juega un rol muy importante en la decisión sobre la dimensionalidad adecuada. Hay, sin embargo, una serie de eigenvalores que si juegan un rol importante en dicha decisión. Cada eigenvalor está asociado a una dimensión de la solución. Para este propósito el valor asociado con cada dimensión es la suma de cuadrados de los valores sobre esa dimensión, es decir: $\lambda_k = \sum_{i=1}^{n} x_{ik}^2$.

Una gráfica muy útil para determinar la dimensionalidad adecuada es llamada *Scatter diagram*, la cual consiste en establecer sobre un plano cartesiano los puntos cuyas abscisas son los valores de las dimensiones y cuyas ordenadas son los correspondientes eigenvalores.

Si los datos satisfacen completamente el modelo, entonces la gráfica mostrará un levantamiento exactamente sobre la dimensión $K + 1$. En otras palabras, aparecerá un codo en la gráfica una unidad adelante de K, el valor correcto o apropiado de la dimensionalidad. En algunos casos, donde los datos contienen gran cantidad de

errores de medición o muestreo no se satisface completamente el modelo, y es difícil de observar un codo en la gráfica correspondiente. Entonces no es suficiente este tipo de gráfica para determinar la dimensionalidad, hay que considerar los otros criterios. A continuación se muestra una gráfica ejemplo.



K vs. λ

Modelo no métrico

La decisión sobre el número de dimensiones de la solución puede hacerse simple al elegir la dimensionalidad que proporcione el valor de Stress más pequeño. Sin embargo, antes de usar ciegamente el stress como un indicador de la "verdadera" o "correcta" dimensión, se deben comprender ciertos aspectos de esta medida.

1. La forma exacta de la medida de Stress puede variar con el programa de computadora empleado para el análisis de escalamiento multidimensional. Aunque las diversas medidas son similares, un valor numérico dado puede reflejar un ajuste bueno o excelente para una pero pobre o pésimo para otra.

2. La terminación prematura del proceso de iteración antes de la convergencia o la terminación en un mínimo local, puede producir valores del Stress más altos que el verdadero valor mínimo para esa dimensión.

3. Los valores del Stress muy cercanos a cero (por ejemplo 0.01 o menos) pueden significar una solución total o parcialmente degenerada. Esta situación se da cuando la posición de los objetos en el espacio ajustado sugieren grupos o conglomerados naturales, es decir, los objetos se encuentran agrupados en un número reducido de regiones, y no se cumple la condición de monotonicidad.

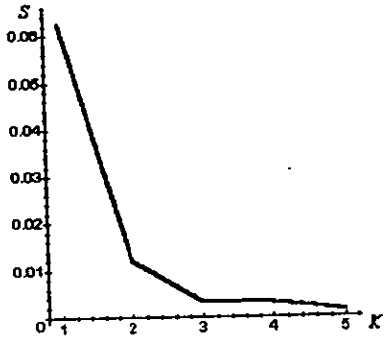
4. La interpretación del valor del Stress es sensible al número de objetos y de dimensiones.

Por estas razones, se deben tomar algunas medidas de precaución. Los problemas de convergencia, del mínimo local y de la solución degenerada se discutieron en el capítulo anterior. Los diagramas de Shepard, que se presentan más adelante, son muy útiles para detectar algunas otras anomalías.

No obstante las cuestiones presentadas, el valor del Stress es un buen elemento para determinar la dimensionalidad adecuada de la solución. Para este propósito se construye una gráfica con los valores del stress sobre el eje de las ordenadas y las dimensiones correspondientes sobre el eje de las abscisas en un plano cartesiano. Al igual que la gráfica de eigenvalores contra dimensiones del modelo métrico, esta gráfica de valores del stress contra dimensiones es llamada *"scatter diagram"*.

Asumiendo que no existe un error de muestreo o de medición demasiado grande, la gráfica debe mostrar un codo sobre el valor apropiado K de la dimensionalidad.

Obsérvese que a diferencia de la gráfica del modelo métrico el codo debe aparecer directamente sobre el valor K y no sobre el valor $K+1$. Nuevamente, si la gráfica no muestra claramente el valor de la dimensioanalidad apropiada, o aún en ese caso, no deben olvidarse los otros criterios de elección. La siguiente es una gráfica de ejemplo.



K vs. *Stress*

Modelo ponderado

Al igual que en los otros modelos es útil construir una gráfica para discernir sobre el número de dimensiones a conservar.

En el modelo euclideano ponderado la gráfica se construye con los valores de las dimensiones como las abscisas de los puntos y los valores de la medida de ajuste como las ordenadas. La gráfica debe mostrar un codo visible sobre el valor K de la dimensionalidad más apropiada.

Cabe mencionar que, si se utiliza el programa *ALSCAL (individuAL diferences SCALing,* Young y Lewyckyj, 1979) para realizar el análisis de escalamiento pondera-

do, la medida de ajuste preferible es una variación del S-Stres fórmula uno, adaptado a la situación en la que hay varias matrices de disimilaridad, una por cada sujeto. Tal medida es: $SS_1 = \frac{1}{S}\sum_s \left[\frac{\sum(\hat{d}^2_{ijs}-d^2_{ijs})^2}{\sum(d^2_{ijs})^2}\right]^{1/2}$

donde $\hat{d}_{ijs}$ es una disparidad calculada para el par de objetos (i,j) y el sujeto s. Se deben considerar aqui, los puntos que se discuten en la sección anterior acerca del Stress

Si se utiliza el programa *INDSCAL* (Carroll y Chang,1970) o el *SINDSCAL* (Pruzansky,1975) la medida de ajuste es la correlación entre los productos escalares b_{ijs} y los productos escalares estimados $\hat{b}_{ijs}$. Para el algoritmo presentado es:

$$r = \frac{\sum_{i,j,s}(b_{ijs}-b_{...})(\hat{b}_{ijs}-\hat{b}_{...})}{\left[\sum_{i,j,s}(b_{ijs}-b_{...})^2 \sum_{i,j,s}(\hat{b}_{ijs}-\hat{b}_{...})^2\right]^{1/2}} \quad \text{con} \quad b_{...} = \frac{1}{n^2 S}\sum_{i,j,s} b_{ijs} \quad \text{y} \quad \hat{b}_{...} = \frac{1}{n^2 S}\sum_{i,j,s}\hat{b}_{ijs}.$$

El valor de r crece cuando el ajuste del modelo a los datos mejora.

En el modelo de tres modas, como en la solución clásica los eigenvalores juegan un papel importante en la decisión sobre dimensionalidad. Más precisamente, hay dos conjuntos de eigenvalores, uno para la configuración de los objetos y otro para la de los sujetos. En ambos casos se construye una gráfica cuyos puntos tienen por abscisas los valores de las dimensiones y por ordenadas los eigenvalores. Si aparece un codo sobre la dimension $K+1$ en la gráfica correspondiente a X, esto sugiere que K es un número razonable de dimensiones a retener en el espacio de los objetos. Similarmente si aparece un codo sobre la dimension $M+1$ en la gráfica que corresponde a $\hat{Z}$, entonces M es un valor razonable para la dimensionalidad del espacio de los sujetos.

5.1.2 Interpretabilidad

La interpretabilidad como criterio para la elección de la dimensionalidad adecuada requiere algunos juicios subjetivos. Establece que se debe conservar el espacio en el cual aparezcan todas las características importantes de los objetos.

Esto significa que una solución en una dimensión alta es preferible si provee una mejor interpretación, es decir, si en ella aparecen importantes características de los objetos que no aparecen en una solución de menor dimensión. Inversamente, se prefiere una solución en dimensión pequeña si en la de dimensión mayor no aparecen características significativas.

Puesto en términos simples, las dimensiones que no se pueden interpretar, posiblemente no existen.

5.1.3 Reproducibilidad

La reproducibilidad puede ser utilizada como criterio para la decisión sobre la dimensionalidad, solamente cuando hay dos o más muestras de la población de objetos. La idea básica consiste en obtener una solución para cada muestra y si hay k dimensiones que aparecen consistentemente en todas ellas, entonces la solución final debe contener k dimensiones.

5.2 ROTACIÓN

Si en el análisis se utiliza alguna métrica no euclideana entonces, no hay problema de rotación. Si X es una configuración solución entonces, en general, $X^* = XT$ con T una matriz ortogonal, no es una solución. De hecho, $X^* = XT$ será una solución sólo si T es diagonal con elementos iguales a 1 o -1.

El problema de rotación surge cuando se utiliza una métrica euclideana, como es el caso de la solución clásica, pues como ya se demostró anteriormente si X es una solución, para cualquier matriz ortogonal T, $X^* = XT$ es también una solución.

Sin embargo, si la solución contiene sólo dos dimensiones, el problema de rotación no es grave; las características importantes deben ser fácilmente identificables independientemente de la rotación. El problema se torna serio cuando hay más de dos dimensiones. Las coordenadas pueden no corresponder con las características significativas de los objetos y sería difícil interpretar las dimensiones.

En la decisión sobre rotar una solución para hacerla más interpretable, hay tres opciones básicas. Si la solución no rotada es interpretable, entonces se conserva de esta manera. Si la solución no es fácilmente interpretable, entonces se puede realizar una rotación objetiva o una rotación a mano.

Una rotación objetiva es un algoritmo matemático para encontrar una solución más interpretable. Tales rotaciones fueron diseñadas primero para utilizarse en el análisis de factores y se usan ocasionalmente en el análisis de escalamiento multidimensional. En algunas aplicaciones una rotación objetiva como Varimax (Kaiser, 1958) o Equimax (Saunders,1960) proveen una solución altamente interpretable, sin

embargo, no se debe suponer ciegamente que tales algoritmos de rotación objetiva producirán automáticamente la solución más interpretable.

Una rotación a mano es la realizada por el investigador y se basa en su inspección visual de la solución no rotada. En la práctica, algunas veces, se puede ver, literalmente, cual rotación de la solución podría hacerla más interpretable.

Para los modelos ponderados, la situación es diferente. En el caso del modelo euclideano ponderado, si X es una solución, en general, $X^* = XT$ con T una matriz ortogonal, no es una solución alternativa. Debido a que este modelo permite que los sujetos alarguen o cantraigan los ejes diferentemente, mientras que sólo los alargamientos uniformes preservan las distancias, no se puede rotar los ejes sin alterar de manera sensible el ajuste de los datos. Afortunadamente la solución obtenida es altamente interpretable por lo que el problema de rotación desaparece.

En el caso del modelo de tres modas, tanto el espacio de los objetos como el de los sujetos pueden ser sometidos a una rotación a mano o una rotación objetiva para obtener una solución más interpretable. Si X es una matriz solución que representa las coordenadas de los objetos, entonces $X^* = XT$ es otra solución, con T una matriz ortogonal. Similarmente, si la matriz Z es la solución para el espacio de los sujetos, entonces $Z^* = ZT$ es otra solución, con T una matriz ortogonal. En general, la matriz transformación que proporciona la solución rotada más interpretable para el espacio de los objetos no es la misma que para el espacio de los sujetos.

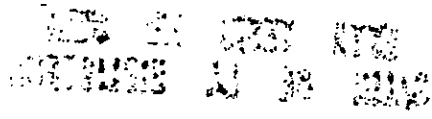
Mientras que el problema de rotación desaparece con el modelo euclideano ponderado, con el modelo de tres modas potencialmente se duplica.

5.3 INTERPRETACIÓN

Una vez que se ha decidido el valor de la dimensionalidad adecuada, la configuración de puntos que representan a los objetos debe ser interpretada. Se puede simplemente permitir que la posición de los puntos en el espacio determine la interpretación, lo que se llama método subjetivo, o seguir un método más objetivo como ajustar un vector característico o realizar un análisis de correlación canónica.

5.3.1 Método subjetivo

El método subjetivo para interpretar la configuración recae únicamente en la posición de los puntos en el espacio. La frase "características significativas" que se emplea en la sección anterior, se refiere básicamente a ordenamientos o agrupamientos de los puntos. Un agrupamiento sustantivamente significativo es un conjunto de puntos que aparecen juntos en una región del espacio multidimensional y cuyos objetos representados poseen atributos en común. Un ordenamiento es un arreglo ordenado de los puntos con respecto a cierta característica de los objetos que representan. El primer paso para identificar un ordenamiento es descubrir los puntos que se encuentran en posiciones extremas y observar las características que poseen los objetos que representan. De estas características, elegir la que mejor explica la posición relativa de los puntos en el espacio.



5.3.2 Ajuste de un vector característico

Este tipo de método objetivo para interpretar la configuración obtenida se basa en el siguiente razonamiento. Supóngase que se conoce una variable que mide cierta característica de los objetos que se sospecha tiene una relación sistemática con la posición de los puntos en el espacio. En este caso resulta conveniente utilizar dicha variable para tratar de explicar la configuración.

Escencialmente lo que se hace es buscar una dirección a través del espacio, la cual corresponda al incremento en la variable que mide la característica elegida. Geométricamente, significa insertar una línea, que se denotará por L, con la propiedad de que la proyección de cada punto sobre esta línea corresponda tanto como sea posible al grado en que el objeto representado posee la característica en cuestión. A la línea L se le llama *vector característico.*

Si la variable está fuertemente relacionada con la configuración, entonces los valores reales de la característica que poseen los objetos corresponderán considerablemente con las proyecciones de los puntos en el espacio y habrá una correlación alta entre ellos. Si por el contrario, la variable no está muy relacionada con la configuración entonces la correlación será muy baja.

El procedimiento para encontrar el vector característico hace uso del análisis de regresión múltiple. Sean a_i el valor específico de la característica que posee el objeto i, con $i = 1, ..., n$, y sean $x_{i1}, x_{i2}, ..., x_{ik}$ las coordenadas del punto que lo representa, entonces la ecuación de regresión múltiple ordinaria es:

$$a_i \approx \hat{a}_i = b_o + b_1 x_{i1} + \ldots + b_k x_{ik}$$

A los valores $b_1, b_2, ..., b_k$ se les llama coeficientes de regresión y al término b_o el intercepto. El valor de $\hat{a}_i$ es el mejor estimador para la proyección del objeto con coordenadas x_{ir}, $r = 1, ..., k$, sobre el vector característico.

Se calcula el coeficiente de correlación múltiple entre las proyecciones y los valores de la característica, y si su valor es bajo entonces se puede concluir con seguridad que los sujetos no utilizaron la variable en cuestión cuando hicieron sus juicios de similaridad.

Para graficar el vector característico, primero se calculan los valores estimados de los coeficientes de regresión, denotándolos por $\beta_1, \beta_2, ..., \beta_k$. En seguida, se determina en el espacio de la configuración el punto $\beta^* = (\beta_1, \beta_2, ..., \beta_k)$. Finalmente, suponiendo que la media de las coordenadas en cada dimensión es cero, se dibuja una línea que parte del origen y pasa por el punto β^*. Es costumbre que la longitud de la línea sea proporcional al cuadrado del coeficiénte de regresión (aunque en un sentido estricto la longitud es arbitraria), y poner una flecha en el punto final del vector.

Los coeficientes $\beta_1, \beta_2, ..., \beta_k$ son los cosenos directores del vector característico. Después de una transformación inversa del coseno se convierten en los ángulos entre el vector característico L y los ejes coordenados.

Estos ángulos se pueden utilizar para hacer una rotación de la solución, de manera que el vector característico quede como una de las dimensiones del espacio.

5.3.3 Análisis de correlación canónica

En la sección previa se consideró ajustar un vector característico en el espacio de la configuración. La discusión se limita al caso en que una sola variable puede ser utilizada para interpretar la solución. En la mayoría de las aplicaciones es razonable esperar que se pueda obtener cierto número de variables para hacerlo.

Cuando se tienen varias características para interpretar la solución, se puede obtener un vector característico para cada una, pero de esta manera se ignoran las interrelaciones entre ellas. En tal caso, se requiere un procedimiento que permita relacionar simultáneamente varias variables con los puntos de la configuración, tal procedimiento se llama *análisis de correlación canónica*.

Dados dos conjuntos de variables X y Y, el análisis de correlación canónica busca una asociación lineal entre los dos conjuntos. El proceso consiste en determinar dos combinaciones lineales, una para X y otra para Y de tal manera que la correlación producto-momento entre las dos combinaciones sea tan grande como sea posible.

En el contexto del análisis de escalamiento multidimensional, el conjunto Y consta de los valores de cada una de las variables para cada objeto y el conjunto X contiene las coordenadas de los puntos en cada dimensión. La correlación es entre la suma ponderada de los valores de las variables y las proyecciones de los puntos sobre el vector característico canónico.

Para el modelo ponderado se debeb interpretar la configuración en el espacio del grupo y los espacios de los sujetos.

Capítulo 6

RELACIÓN CON OTRAS TÉCNICAS

El análisis de escalamiento multidimensional está relacionado con otras técnicas del análisis multivariado, especialmente con aquellas empleadas en la reducción de dimensionalidad.

6.1 ANÁLISIS DE COMPONENTES PRINCIPALES

Dado un conjunto de variables $X_1, ..., X_p$ con valores para n individuos (que pueden ser objetos), el análisis de componentes principales lo transforma, mediante combinaciones lineales, en un conjunto más pequeño de índices $Y_1, ..., Y_m$ que describen la mayor parte posible de la varianza del conjunto original.

El proceso del análisis de componentes principales se resume con la siguiente definición. Sean $\lambda_1 > \lambda_2 > ... > \lambda_p$ los eigenvalores de la matriz de covarianzas S de $X = (X_1, ..., X_p)$. El $j - ésimo$ componente principal Y_j de X se define como la combinación lineal:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + ... + a_{pj}X_p \ \lambda_j$$

donde los coeficientes $a_{1j}, a_{2j}, ..., a_{pj}$ son los elementos del eigenvector a_j correspondiente al $j - ésimo$ eigenvalor más grande λ_j de la matriz de covarianza S, normalizado de manera que $a_j^t a_j = 1$. Si $\lambda_j \neq \lambda_h$ entonces los coeficientes del $j - ésimo$ y del $h - ésimo$ componentes principales son ortogonales. La varianza del $j - ésimo$ componente principal es λ_j, y la varianza total del sistema está dado por: $tr(S) = \lambda_1 + ... + \lambda_p$

Para cada individuo i existe un conjunto de índices (centrados) sobre los componentes principales dados por:

$$y_{i1} = a_1^t(x_i - \overline{x}), \ y_{i2} = a_{i2}^t(x_i - \overline{x}), \ . \ . \ . \ , y_{im} = a_m^t(x_i - \overline{x})$$

donde x_i es el $i - ésimo$ vector de observaciones sobre las p variables y $\overline{x}$ es la media sobre todos estos vectores.

Existe una dualidad importante entre el análisis de componentes principales y el análisis de escalamiento multidimensional métrico, el siguiente teorema lo establece.

TEOREMA 5. *Considérese la matriz de datos $X_{n \times p}$ y sea D la matriz de distancias entre los renglones de X. Si D es euclideana entonces la solución clásica del escalamiento multidimensional en k dimensiones está dada por los índices para los n individuos sobre los primeros k componentes principales de X.*

Demostración.

Sean $a_{(i)}$ el $i-$ *ésimo* vector de coeficientes del análisis de componentes principales normalizado de manera que $a_{(i)}^t a_{(i)} = 1$ y $A = (a_{(1)}, ..., a_{(p)})$; sea, además $V = (v_{(1)}, ..., v_{(p)})$ la matriz de eigenvectores de $B = (HX)^t HX$, normalizados tal que $v_{(i)}^t v_{(i)} = \lambda_i$ y H definida como en la solución clásica. Por el teorema espectral de descomposición se puede escribir $nS = X^t HX = A\Lambda A^t$, donde $\Lambda = diag(\lambda_1, ..., \lambda_p)$ Por el teorema del valor singular de descomposición se pueden elegir los signos de $a_{(i)}$ y de $v_{(i)}$ para escribir HX en términos de estos eigenvectores como: $HX = VA^t$.

Los índices para los n individuos sobre el $i-$ *ésimo* componente principal están dados por los n elementos de $HXa_{(i)}$. Esto es, si $A_k = (a_{(1)}, ..., a_{(k)})$, los índices estan dados por $HXA_k = VA^tA_k = V(I_k, 0)^t = V_k = (v_{(1)}, ..., v_{(k)})$, puesto que las columnas de A son ortogonales. Con lo que queda probado el teorema.

Como las columnas de A_k son ortogonales, entonces $A_k^t A_k = I_k$, y se puede ver que $V_k = XA_k$ representa una proyección de X en un subespacio de R^p de dimensión k, la cual es óptima de entre todas las posibles proyecciones en subespacios de dimensión k por ser la más cercana a la configuración original en dimensión p, como lo establece el teorema 3 del capítulo 3.

Este resultado es dual de aquel que establece que la suma de las varianzas de los primeros k componentes principales es mayor que la suma de varianzas de cualquier otra combinación lineal no correlacionada de columnas de X.

6.2 ANÁLISIS DE CONGLOMERADOS

El análisis de conglomerados fue diseñado para resolver el siguiente problema: Dados n objetos con valores sobre p variables, enconcontrar un esquema de agrupación de los objetos en clases o grupos, de manera que los objetos similares esten en el mismo grupo o clase.

De los diferentes algoritmos propuestos para resolver el problema destacan dos, frecuentemente más utilizados. Primero, el método *jerárquico*, el cuál produce un dendograma. Este método comienza calculando las distancias de cada objeto con cada uno de los restantes. Los grupos se van formando por un proceso de aglomeración o de división. Con aglomeración todos los objetos comienzan siendo su propio grupo. Los conglomerados surgen gradualmente por una serie de fusiones hasta que finalmente todos los objetos pertenecen al mismo.

Con división todos los objetos comienzan en un solo grupo, éste se va dividiendo en subgrupos cada vez más finos hasta que cada objeto sea su propio grupo.

El segundo método involucra particiones, permitiendo a los objetos moverse de un grupo a otro en diferentes estados del análisis. Para comenzar, se eligen, de forma más o menos arbitraria los grupos de objetos y se calculan los centros de cada uno. Un objeto se mueve a un nuevo grupo si está más cerca de su centro que del centro del grupo al que actualmente pertenece. Se calculan los nuevos centros y el proceso se repite. Los grupos muy cercanos se unen , los grupos muy dispersos se dividen, etc. El proceso continúa iterativamente hasta que se alcanza alguna estabilidad con cierto número de grupos predeterminado.

El análisis de conglomerados jerárquico es el que está más relacionado con el análisis de escalamiento multidimensional; los dos métodos son usados frecuentemente para investigar la estructura de los objetos. Hay por lo menos tres paralelismos entre estos métodos:

a) Con ambos métodos se pueden analizar matrices de proximidades;

b) Ambos métodos se construyen sobre modelos de distancia;

c) Se puede representar la solución en términos de dimensiones coordenadas, aunque la de conglomerados rara vez se representa asi.

No obstante estos tres paralelismos, el análisis de conglomerados y el análisis de escalamiento multidimensional difieren fundamentalmente. Primero, la relación entre las proximidades δ_{ij} y las distancias d_{ij} en el análisis de conglomerados no puede ser expresada por una función lineal o una función monótona como en el análisis de escalamiento. Segundo, las distancias del análisis de conglomerados no son espaciales como en el análisis de escalamiento, además, deben satisfacer una desigualdad más fuerte, la ultramétrica. Tercero, las dimensiones coordenadas en el análisis de escalamiento multidimensional son variables contínuas mientras que en el análisis de conglomerados, si se utiliza esta representación, las variables son discretas.

Como el análisis de conglomerados jerárquico y el análisis de escalamiento emplean diferentes representaciones de la estructura, son vistos frecuentemente como métodos complementarios para sobresaltar las diferentes características de los objetos.

6.3 ANÁLISIS FACTORIAL

El objetivo del análisis factorial es simplificar las diversas relaciones que existen dentro de un conjunto de variables observadas, calculando dimensiones o factores que ayuden a analizar la estructura de los datos. La idea básica es un tanto similar a la del análisis de componentes principales: describir un conjunto de p variables (con valores para n individuos) en términos de un número más pequeño de índices o factores.

Bajo el modelo del análisis factorial cada variable se representa como una función lineal de un número pequeño de factores comunes inobservables y un factor específico único. Los factores comunes generan las covarianzas entre las variables observadas, mientras que los términos específicos contribuyen solamente a la varianza de su variable particular. Los coeficientes de los factores comunes no están restringidos a ser ortogonales, y de hecho su matriz es única sólo con una postmultiplicación por una matriz ortogonal.

El modelo matemático del análisis factorial está dado como sigue:

$$X_i = \lambda_{i1} Y_1 + ... + \lambda_{im} Y_m + e_i$$

donde Y_j es el $j-ésimo$ factor común, λ_{ij} es un parámetro que refleja la importancia del $j-ésimo$ factor en composición con la $i-ésima$ variable y e_i es el $i-ésimo$ término específico.

Existen diversos métodos para calcular los factores, dos son los más utilizados. El primero y más viejo es el de factores principales; este método frecuentemente se confunde con el análisis de componentes principales. El segundo es el método de máxima verosimilitud y es el único que proporciona bases estadísticas para comprobar

la adecuación del modelo analítico factorial. El método de factores principales extrae factores de manera que cada uno contiene la mayor cantidad posible de varianza del conjunto de variables originales. La idea básica del método de máxima verosimilitud es suponer que se conoce la forma general de la distribución poblacional de donde fueron extraidos los valores de las variables, pero que no se conocen los parámetros. Se obtienen ciertos valores de los parámetros que hacen que los valores observados de las variables tengan la mayor verosimilitud conjunta, y se toman como los estimadores de máxima verosimilitud de los parámetros de la población.

El análisis de escalamiento multidimensional se relaciona más con el análisis factorial que con el análisis de conglomerados. Los datos básicos en muchas de las aplicaciones del análisis factorial son medidas de proximidad entre pares de objetos. En la práctica los objetos son frecuentemente pruebas psicológicas o sobre productos y la medida de proximidad es generalmente el coeficiente de correlación. Como en el análisis de escalamiento el análisis factorial produce una representacion dimensional cuantitativa de una estructura entre los objetos.

Con estos paralelismos ente análisis factorial y de escalamiento no debe sorprender que se utilicen algunas veces sobre los mismos temas de investigación. Ambos han sido utilizados para estudiar las dimensiones de las percepciones interpersonales, la estructura de habilidades humanas, y la organización de ambientes urbanos, por citar algunos ejemplos.

Cuando el análisis factorial y el análisis de escalamiento se utilizan para estudiar el mismo tema, hay tres razones por las que las conclusiones pueden diferir, dos de las cuales no tienen nada que ver con el método. Primero, los estudios pueden emplear

diferentes medidas de similaridad para los objetos. El análisis factorial prefiere el coeficiente de correlación, el análisis de escalamiento ha empleado, además, otras medidas de proximidad.

Segundo, el análisis factorial y el de escalamiento aplicados al mismo tema frecuentemente emplean muy diferentes procedimiento experimentales. Esto resulta ser una fuente grande de variaciones en los resultados.

Los métodos analíticos mismos representan un tercera fuente de diferencia. Aunque, ambos análisis producen una representanción de la estructura de los objetos en términos de coordenadas espaciales (llamadas "factor loadings" en análisis factorial y "scale values" en el análisis de escalamiento), el modelo factorial y el de análisis de escalamiento tienen hipótesis diferentes acerca de la relación entre las coordenadas y las proximidades observadas.

En análisis factorial la correlación r_{ij} (o alguna otra medida de proximidad) se presume relacionada con las coordenadas x_{ij} y x_{jk} por una función de la forma:

$$r_{ij} = \sum_k x_{ik} x_{jk}$$

En el análisis de escalamiento métrico, por otro lado, las disimilaridades se presumen relacionadas con las coordenadas por una función de la forma:

$$\delta_{ij} = [\sum_k (x_{ik} - x_{jk})^2]^{1/2}$$

Tales diferencias pueden provocar diferentes resultados.

6.4 ANÁLISIS DE CORRESPONDENCIAS

El análisis de correspondencias es una técnica de interdependencia desarrollada recientemente que facilita la reducción de dimensionalidad y provee una representación gráfica. Su aplicación más directa es representar la correspondencia de categorias de variables, particularmente medidas en términos nominales, la cual es base del desarrollo de representaciones gráficas de percepciones .

En su forma más básica el análisis de correspondencias emplea una tabla de contingencia, la cual es la tabulación cruzada de dos variables categóricas. Transforma los datos no métricos en forma métrica y efectúa una reducción similar al análisis factorial.

Aunque en menor grado también tiene paralelismos con el análisis de escalamiento multidimensional, por ejemplo, utiliza medidas de proximidad que indican el grado de asociaciòn entre categorías renglón o columna, como elementos básicos. Y efectúa una representación gráfica de las categorías en un espacio multidimensional. El análisis de correspondencias proporciona una representación multivariada de interdependencia para datos no métricos que no es posible con otros métodos.

Capítulo 7

APLICACIONES

En este capítulo se realiza el análisis de escalamiento multidimensional de tres conjuntos de datos. Para el primero se retoma el ejemplo 1 del capítulo de introducción sobre el círculo de colores. El segundo consta de una matriz de disimilaridad sobre 8 naciones y se aplica el modelo métrico. El tercero contiene cuatro matrices de disimilaridad sobre las mismas 8 naciones, utilizándose el modelo ponderado.

El análisis se llevará a cabo con el programa *INDSCAL* disponible dentro del paquete estadístico *SPSS (Statistical Package for the Social Sciences)* en la opción llamada *Multidimensional Scaling*

7.1 Círculo de colores

Cuando a una persona se le pide que ordene objetos de diferente color, casi seguramente llegará al orden: rojo, naranja, amarillo, verde, azúl, azúl-violeta, que corresponde al orden de la longitud de onda electromagnética de estos colores. Para el color

rojo-violeta, la persona no está segura si ponerlo en el extremo rojo o en el extremo violeta de la escala (o en ambos). Este problema se resuelve ordenando los colores en forma de herradura o círculo. Existe la hipótesis de que la mayoría de las personas enfrentadas a la tarea de ordenar colores llegará tarde o temprano a esta solución, y parece haber componentes perceptuales e intelectuales envueltos en ello. Se puede argumentar que el círculo de colores es una implicación de la forma en que se percibe la similaridad entre éstos.

Ekman (1954) utilizó 14 colores que diferian en su longitud de onda (434 a 674) pero no en su brillo o saturación. Cada uno de los 91 pares posibles (de colores diferentes) fue proyectado en una pantalla, y 31 sujetos juzgaron su similaridad dando un puntaje de 0 (no similares) a 4 (idénticos). Los puntajes fueron promediados sobre todos los sujetos y finalmente divididos entre cuatro para dar una escala de 0 a 1. Esto dio por resultado la matriz D, que se muestra más adelante.

Estas medidas pueden interpretarse como correlaciones, y se les puede aplicar un análisis de componentes principales, obteniendose cinco factores que comprenden los colores: 434-445, 465-490, 504-555, 584-600 y 610-674, los cuales corresponden a cinco diferentes grupos de puntos del espectro electromagnético, dando lugar a las categorías subjetivas: azúl-púrpura, azúl, verde, amarillo y rojo respectivamente.

Para realizar el análisis de escalamiento se utiliza el modelo no métrico. Es decir se busca una configuración de 14 puntos en la cual la relación de orden de las distancias entre los puntos corresponda (inversamente) con la relación de orden de los datos. Se obtienen soluciones en dimensión 1, 2 y 3, y se elige entre ellas la que representa mejor los datos, con base en los criterios discutidos en los capítulos tres y cuatro.

$D =$

nm.	434	445	465	472	490	504	537	555	584	600	610	628	651	674
434	1.0	0.86	0.42	0.42	0.18	0.06	0.07	0.04	0.02	0.07	0.09	0.12	0.13	0.16
445	0.86	1.0	0.50	0.44	0.22	0.09	0.07	0.07	0.02	0.04	0.07	0.11	0.13	0.14
465	0.42	0.50	1.0	0.81	0.47	0.17	0.10	0.08	0.02	0.01	0.02	0.01	0.05	0.03
472	0.42	0.44	0.81	1.0	0.54	0.25	0.10	0.09	0.02	0.01	0.00	0.01	0.02	0.04
490	0.18	0.22	0.47	0.54	1.0	0.61	0.31	0.26	0.07	0.02	0.02	0.01	0.02	0.00
504	0.06	0.09	0.17	0.25	0.61	1.0	0.62	0.45	0.14	0.08	0.02	0.02	0.02	0.01
537	0.07	0.07	0.10	0.10	0.31	0.62	1.0	0.22	0.22	0.14	0.05	0.02	0.02	0.00
555	0.04	0.07	0.08	0.09	0.26	0.45	0.73	1.0	0.33	0.19	0.04	0.03	0.02	0.02
584	0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	1.0	0.58	0.37	0.27	0.20	0.23
600	0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	1.0	0.74	0.50	0.41	0.28
610	0.09	0.07	0.02	0.00	0.02	0.02	0.05	0.04	0.37	0.74	1.0	0.76	0.62	0.55
628	0.12	0.11	0.01	0.01	0.01	0.02	0.02	0.03	0.27	0.50	0.76	1.0	0.85	0.68
651	0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.20	0.41	0.62	0.85	1.0	0.76
674	0.16	0.14	0.03	0.04	0.00	0.01	0.00	0.02	0.23	0.28	0.55	0.68	0.76	1.0

A continuación se presentan los resultados obtenidos con el paquete estadístico SPSS más o menos con el mismo formato pero en idioma español.

Resumen de datos del Procedimiento

Casos					
Validos		Faltantes		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
14	100.0%	0	.0%	14	100.0%

Utilizando Distancia Euclideana

Alscal Opciones del Procedimiento

Opciones de los Datos -
Número de Renglones (Observaciones/Matriz). 14
Número de Columnas (Variables). 14
Número de Matrices. 1
Nivel de Medición . . . . . Ordinal
Forma de la Matriz . . . . . Simétrica
Tipo . . . . . Disimilaridad
Tratamiento de Empates . . . . . Primario
Condicionalidad . . . . . Matriz
Corte de los Datos 0.000000

Opciones Del Modelo -
Modelo . . . . . Euclideano
Máxima Dimensionalidad 1
Mínima Dimensionalidad 1
Pesos Negativos . . . . . No Permitidos

Opciones de Salida -
Encabezado . . . . . Impreso
Matrices de datos . . . . . Impresas
Configuraciones y Transformaciones . Graficadas
Archivo de salida. . . . . No creado
Coordenadas Iniciales . . . . . Calculadas

Opciones del Algoritmo -
Máximo de Iteraciones 30
Criterio de Convergencia 0.00100
S-stress Mínimo 0.00500
Datos Faltantes Estimados por . . "Ulbounds"
"Tiestore" 91

Datos originales(distancias) para el Sujeto 1

	1	2	3	4	5	6	7
1	0.000						
2	0.226	0.000					
3	1.047	0.966	0.000				
4	1.105	1.041	0.298	0.000			
5	1.496	1.450	1.022	0.923	0.000		
6	1.781	1.759	1.561	1.487	0.807	0.000	
7	1.860	1.856	1.773	1.745	1.255	0.704	0.000
8	1.854	1.849	1.786	1.765	1.339	0.907	0.440
9	1.855	1.889	1.937	1.945	1.758	1.595	1.448
10	1.940	1.991	2.103	2.122	2.005	1.895	1.797
11	2.004	2.058	2.217	2.244	2.178	2.130	2.082
12	1.991	2.042	2.236	2.263	2.219	2.187	2.161
13	1.922	1.970	2.162	2.194	2.163	2.147	2.133
14	1.810	1.862	2.070	2.095	2.082	2.071	2.061

	8	9	10	11	12	13	14
8	0.000						
9	1.303	0.000					
10	1.691	0.790	0.000				
11	2.010	1.215	0.632	0.000			
12	2.096	1.420	1.001	0.506	0.000		
13	2.075	1.472	1.126	0.698	0.294	0.000	
14	1.997	1.411	1.193	0.823	0.558	0.412	0.000

Desarrollo de las iteraciones para la solucion en dimensión 1
Se utiliza el S-stress formula 1 de Young.

Iteración	S-stress	Mejoramiento
1	0.20632	
2	0.17814	0.02818
3	0.16790	0.01025
4	0.16127	0.00662
5	0.15661	0.00466
6	0.15315	0.00346
7	0.15072	0.00243
8	0.14926	0.00145
9	0.14842	0.00084

Las iteraciones se detuvieron porque el cambio en el S-stress es menor que 0.001000. El Stress y la correlación cuadrática (RSQ) de las distancias: los valoresde RSQ dan la proporción de varianza de los datos transformados (disparidades), la cual es contada por sus correspondientes distancias.
Los valores son del Stress formula 1 de Kruskal.
Para la matriz

Stress = 0.17592 RSQ = 0.90233

Configuración derivada in 1 dimensión

Núm. del Objeto	Nombre del Objeto	Dimensión 1
1	434	-0.2193
2	445	-0.3734
3	465	-1.2069
4	472	-1.3025
5	490	-1.1107
6	504	-0.8826
7	537	-0.7096
8	555	-0.5368
9	584	0.2863
10	600	0.7712
11	610	1.4097
12	628	1.5517
13	651	1.3684
14	674	0.9545

Datos transformados (disparidades) para el sujeto 1

	1	2	3	4	5	6	7
1	0.000						
2	0.149	0.000					
3	0.792	0.662	0.000				
4	0.792	0.792	0.149	0.000			
5	0.861	0.861	0.662	0.401	0.000		
6	0.861	0.861	0.861	0.861	0.401	0.000	
7	0.861	0.861	0.861	0.861	0.792	0.401	0.000
8	0.861	0.861	0.861	0.861	0.792	0.401	0.272
9	0.861	1.049	1.455	1.513	0.861	0.861	0.861
10	1.455	1.513	2.102	2.102	1.843	1.455	0.861
11	1.634	1.843	2.644	2.735	2.503	2.188	2.092
12	1.634	1.843	2.735	2.854	2.662	2.503	2.261
13	1.455	1.513	2.503	2.644	2.503	2.251	2.188
14	0.861	1.049	1.973	2.102	2.092	1.973	1.843

	8	9	10	11	12	13	14
8	0.000						
9	0.792	0.000					
10	0.861	0.401	0.000				
11	1.843	0.792	0.401	0.000			
12	2.102	0.861	0.662	0.272	0.000		
13	1.973	0.861	0.792	0.401	0.149	0.000	
14	1.634	0.792	0.792	0.401	0.401	0.272	0.000

Configuración Derivada en 1 dimensión

Modelo Euclideano

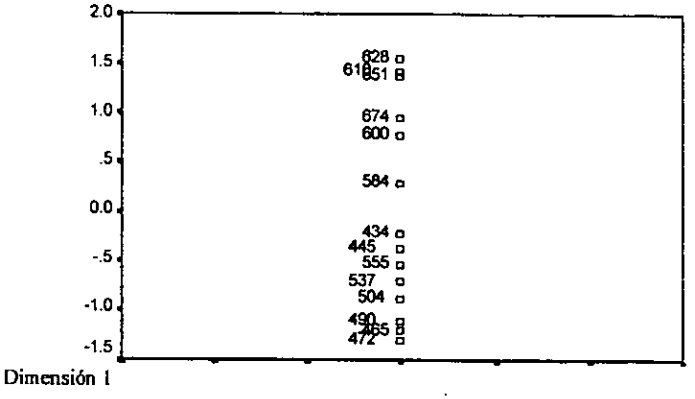


Diagrama de ajuste lineal

Modelo Euclideano

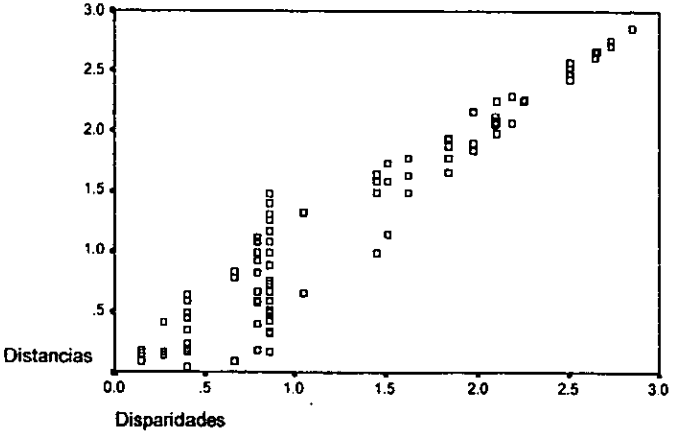


Diagrama de Shepard a

Modelo Euclideano

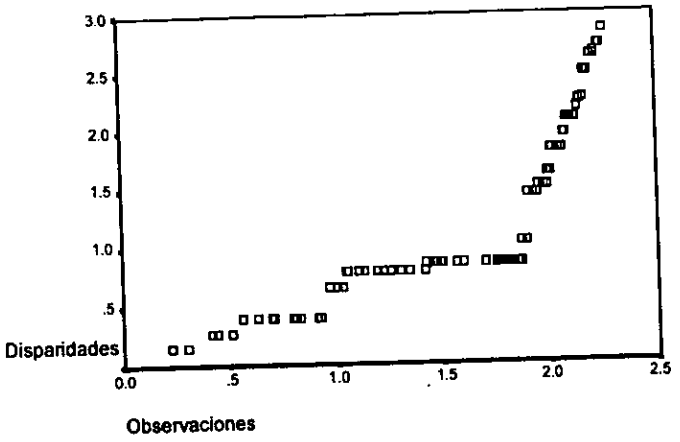
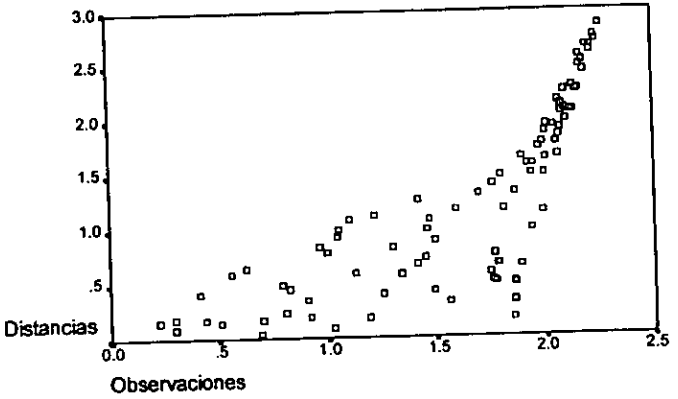


Diagrama de Shepard b

Modelo Euclideano



Desarrollo de las iteraciones para la solución en dos dimensiones
(calculando distancias cuadradas)
Se utiliza el S-stress formula 1 de Young

Iteración	S-stress	Mejoramiento
1	0.03678	
2	0.02059	0.01619
3	0.01719	0.00340
4	0.01570	0.00149
5	0.01466	0.00104
6	0.01399	0.00068

Las iteraciones se detuvieron porque el cambio en el S-Stress es menor que 0.001000. El Stress y la correlación cuadratica(RSQ) de las distancias: el valor de RSQ mide la varianza de los datos transformados (disparidades), la cual está contenida en sus correspondientes distancias.
Los valores son del Stress formula 1 de Kruskal.
Para la matriz

Stress = 0.02071 RSQ = 0.99780

Configuración derivada en 2 dimensiones

Num. del Objeto	Nombre del Objeto	Dimensión 1	Dimensión 2
1	434	0.7410	-1.1226
2	445	0.8584	-1.0982
3	465	1.3206	-0.6865
4	472	1.3680	-0.6036
5	490	1.2947	0.2061
6	504	1.0308	0.8889
7	537	0.7392	1.2682
8	555	0.5695	1.2543
9	584	-0.6509	0.7476
10	600	-1.2267	0.3144
11	610	-1.5680	-0.0770
12	628	-1.6204	-0.3075
13	651	-1.5154	-0.4142
14	674	-1.3406	-0.3698

Datos transformados (disparidades) para el sujeto 1

	1	2	3	4	5	6	7
1	0.000						
2	0.120	0.000					
3	0.773	0.723	0.000				
4	0.773	0.773	0.126	0.000			
5	1.485	1.404	0.773	0.723	0.000		
6	2.036	2.007	1.602	1.485	0.620	0.000	
7	2.360	2.360	2.036	2.001	1.216	0.478	0.000
8	2.360	2.360	2.081	2.022	1.298	0.620	0.175
9	2.360	2.360	2.435	2.435	2.007	1.688	1.466
10	2.435	2.508	2.737	2.752	2.526	2.360	2.185
11	2.526	2.632	2.952	2.983	2.877	2.772	2.676
12	2.508	2.602	2.965	3.003	2.960	2.899	2.849
13	2.365	2.470	2.849	2.899	2.877	2.849	2.813
14	2.214	2.360	2.676	2.704	2.697	2.676	2.647

	8	9	10	11	12	13	14
8	0.000						
9	1.298	0.000					
10	2.001	0.620	0.000				
11	2.526	1.216	0.439	0.000			
12	2.704	1.404	0.736	0.236	0.000		
13	2.676	1.466	0.773	0.439	0.126	0.00	
14	2.508	1.313	0.773	0.620	0.287	0.175	0.000

Configuración Derivada en 2 dimensiones

Modelo Euclideano

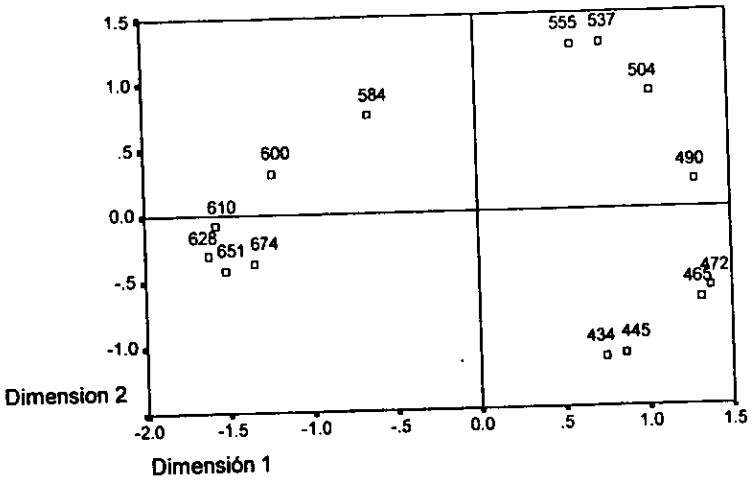


Diagrama de ajuste lineal

Modelo Euclideano

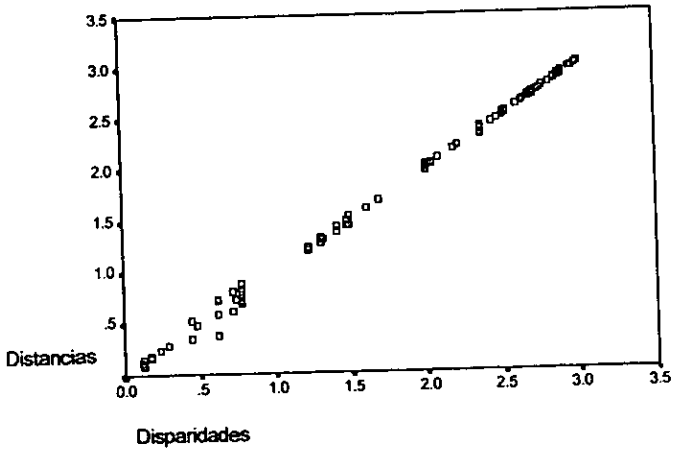


Diagrama de Shepard a

Modelo Euclideano

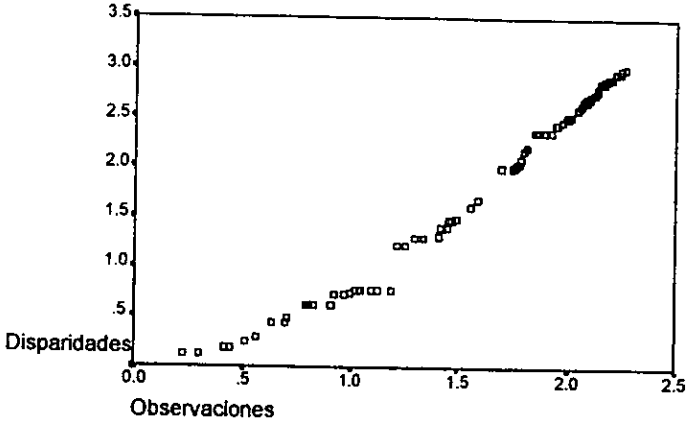
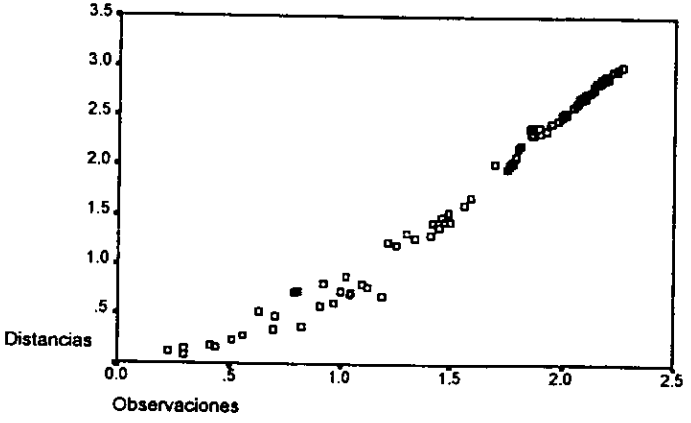


Diagrama de Shepard b

Modelo Euclideano



>Precaución # 14654
>El número total de parámetros a estimar (el número de coordenadas más el número de pesos (si los hay)) es muy grande comparado con el número de dato en la matriz. Los resultados pueden no ser confiables dado que no hay suficientes datos para estimar de manera precisa los valores de los parámetros. Debería reducirse el número de parámetros (i.e. requerir menos dimensiones) o incrementar el número de observaciones.
El numero de parámetros es 42. El numero de datos es 91

Desarrollo de las iteraciones para la solución en dimensión 3
(calculando distancias cuadradas)
Se utiliza el S-stress formula 1 de Young.

Iteración	S-stress	Mejoramiento
1	0.03143	
2	0.02062	0.01082
3	0.01651	0.00411
4	0.01394	0.00257
5	0.01223	0.00170
6	0.01099	0.00125
7	0.01000	0.00099

Las iteraciones se detuvieron porque el cambio en el S-stress es menor que 0.001000. El Stress y la correlación cuadrada(RSQ) de las distancias: el valor de RSQ es la proporción de la varianza de los datos transformados (disparidades), la cual es contada por sus correspondientes distancias.
Los valores son del Stress formula 1 de Kruskal.

Para la matriz
Stress = 0.01429 RSQ = 0.99882

Configuración derivada en 3 dimensiones

Núm. del Objeto	Nombre del Objeto	Dimension 1	2	3
1	434	0.9642	-1.3426	0.4387
2	445	1.0514	-1.3310	0.3821
3	465	1.5367	-0.8767	-0.3718
4	472	1.5985	-0.7933	-0.3687
5	490	1.5099	0.2821	-0.4859
6	504	1.1975	1.1364	-0.4040
7	537	0.8742	1.5421	0.1553
8	555	0.7049	1.4684	0.4903
9	584	-0.7775	0.8663	0.5906
10	600	-1.4774	0.4143	0.3532
11	610	-1.8651	-0.1558	0.0930
12	628	-1.9069	-0.3560	-0.2210
13	651	-1.8067	-0.4151	-0.3156
14	674	-1.6035	-0.4393	-0.3362

Datos transformados (disparidades) para el sujeto 1

	1	2	3	4	5	6	7
1	0.000						
2	0.105	0.000					
3	1.112	1.055	0.000				
4	1.128	1.112	0.129	0.000			
5	1.959	1.863	1.112	1.055	0.000		
6	2.626	2.583	2.042	1.959	0.800	0.000	
7	2.884	2.884	2.583	2.501	1.562	0.763	0.000
8	2.823	2.823	2.626	2.583	1.734	1.055	0.365
9	2.823	2.884	3.031	3.053	2.583	2.228	1.863
10	3.031	3.077	3.358	3.365	3.095	2.884	2.626
11	3.095	3.165	3.493	3.544	3.443	3.365	3.221
12	3.095	3.165	3.493	3.544	3.493	3.443	3.382
13	3.018	3.077	3.382	3.443	3.393	3.382	3.365
14	2.823	2.884	3.191	3.222	3.221	3.215	3.191

	8	9	10	11	12	13	14
8	0.000						
9	1.603	0.000					
10	2.427	0.800	0.000				
11	3.095	1.562	0.625	0.000			
12	3.264	1.852	1.055	0.365	0.000		
13	3.221	1.877	1.128	0.625	0.129	0.000	
14	3.095	1.802	1.128	0.800	0.365	0.206	0.000

Configuración derivada en 3 dimensiones

Modelo Euclideano

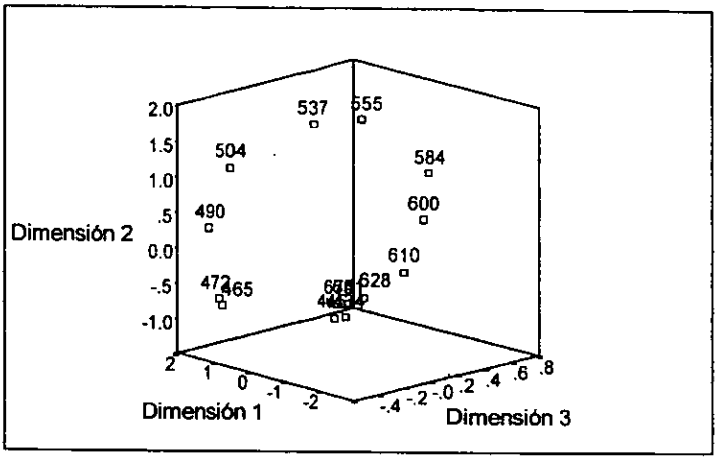


Diagrama de ajuste lineal

Modelo Euclideano

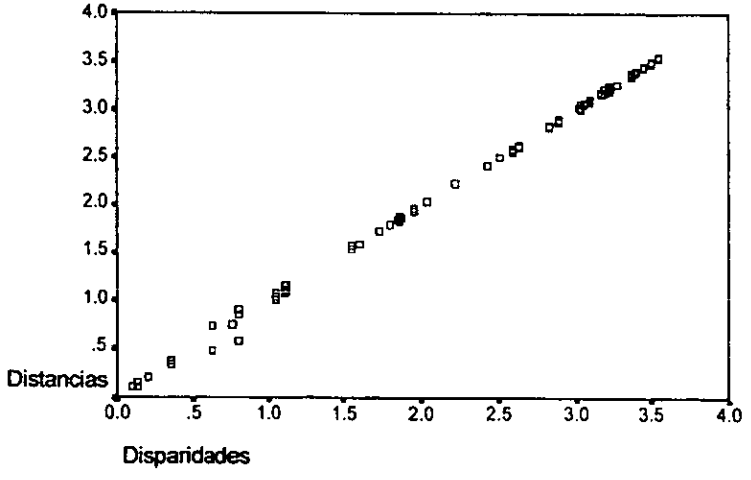


Diagrama de Shepard a

Modelo Euclideano

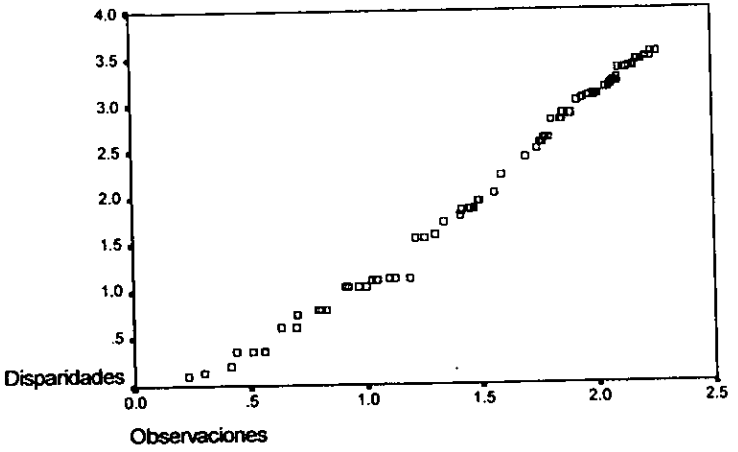
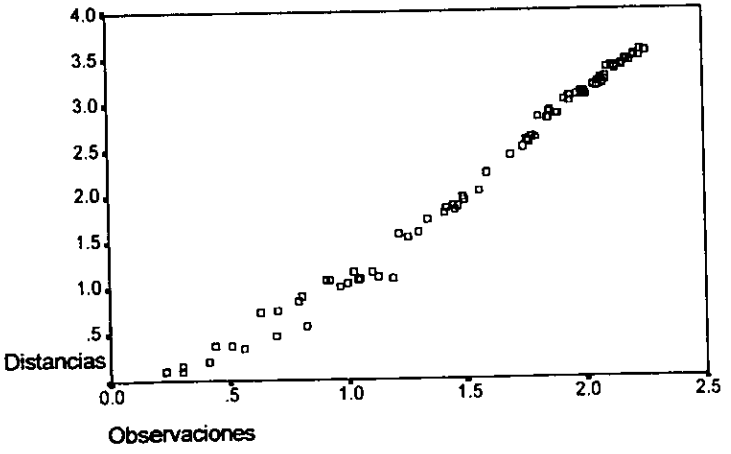


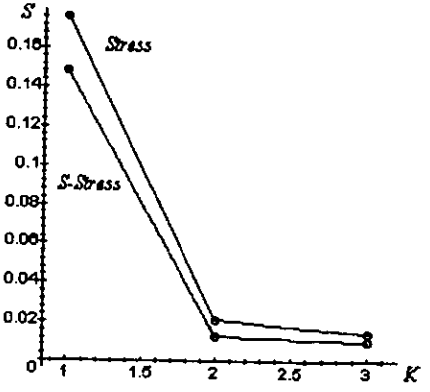
Diagrama de Shepard b

Modelo Euclideano



Para llevar a cabo la discusión sobre las tres soluciones, y decidir cuál representa mejor los datos; en lo sucesivo, a la solución de dimensión uno se le llamara *solución I*; a la de dimensión dos, *solución II*; y *solución III*, a la de dimensión tres. Lo primero que se puede analizar son los diagramas de ajuste lineal (pags. 98, 101 y 104). Obsérvese que para la solución I, su diagrama muestra muchos puntos que se desvían de la recta identidad lo que significa que existe un mal ajuste, siendo mejor el que proporcionan las soluciones II y III; además, nótese que los diagramas de éstas presentan poca diferencia entre si.

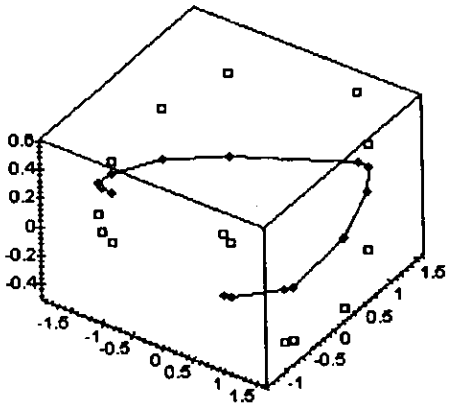
Medidas en términos del Stress y S-Stress, la solución I tiene los valores más altos, aumentando una dimensión se obtiene una reducción bastante considerable y pasando a dimensión 3 la diferencia no es tan significativa, pues para la solución II los valores son ya muy cercanos a cero. La gráfica muestra un codo bastante visible sobre la dimensión 2.



K vs. Stress y S-Stress

Analizando los diagramas de Shepard, también se llega a la conclusión de que las soluciones II y III son mejores, pues existen demasiados puntos en el diagrama b de la solución I alejados de lo que sería la mejor curva de monotonicidad (diagrama a).

Por otro lado, obsérvese que la proyección sobre el plano X-Y de los puntos de la solución III corresponden de manera sensible a los de la solución II. Y no parece haber una interpretación sustantiva para la tercera dimensión en la solución III. Además de que aumentando una dimensión los cambios no son sensiblemente notorios, no existe una teoría que fundamente la solución en dimensión 3.



Proyección de la solución III en el plano XY

Con estos criterios y consideraciones se concluye que la solución II es la representación de los datos más apropiada, correspondiendo de manera precisa con el familiar círculo de colores, aún cuando a los sujetos no se les pidió ordenarlos, sólo evaluar su similaridad.

7.2 Ocho naciones

Supóngase que a un sujeto se le pide que emita su juicio de disimilaridad sobre 8 naciones sin darle indicación alguna sobre los lineamientos que deba seguir para evaluar estos objetos (los datos fueron tomados de Davison, 1983). De estos juicios se genera la siguiente matriz:

$D =$	Ang	Arg	Aus	Chi	Cub	Jap	E.U.	Zim
Angola	0.00	1.41	1.00	1.00	1.41	1.41	1.73	0.71
Argentina	1.41	0.00	1.00	1.73	1.41	1.41	1.00	1.41
Australia	1.00	1.00	0.00	1.41	1.73	1.00	1.41	1.00
China	1.00	1.73	1.41	0.00	1.00	1.00	1.41	1.00
Cuba	1.41	1.41	1.73	1.00	0.00	1.41	1.00	1.41
Japón	1.41	1.41	1.00	1.00	1.41	0.00	1.00	1.41
Est.Unid.	1.73	1.00	1.41	1.41	1.00	1.00	0.00	1.73
Zimbawe	0.71	1.41	1.00	1.00	1.41	1.41	1.73	0.00

Considerando estas medidas como distancias euclideanas se aplica el modelo métrico y, utilizando el programa *Maple V Release 4* para realizar los cálculos, se obtiene la matriz de producto interno centrado B y sus eigenvalores con sus correspondientes eigenvectores. Estos resultados se muestran a continuación.

B =

0.6936328128 -0.2076671878 0.1628703126 0.1628703126 -0.2120421876
-0.3322234376 -0.7090234376 0.4415828126

-0.2082921877 0.8785078128 0.2549953127 -0.7414546877 -0.1149171876
-0.2400984376 0.3795515626 -0.0208292187

0.1622453127 0.2549953128 0.6314828127 -0.3625671877 -0.7408296877
0.1304390625 -0.2380109377 0.1622453125

0.1622453128 -0.7414546878 -0.3625671877 0.6314828127 0.2556203126
0.1304390626 -0.2380109376 0.1622453126

-0.2082921877 -0.1155421877 -0.7414546877 0.2549953126 0.8791328127
-0.2400984375 0.3795515626 -0.2082921875

-0.3328484378 -0.2400984378 0.1304390627 0.1304390627 -0.2394734376
0.6293953126 0.2549953126 -0.3328484376

-0.7096484378 0.3795515628 -0.2380109377 -0.2380109377 0.3801765626
0.2549953126 0.8805953126 -0.7096484376

0.4409578128 -0.2082921878 0.1622453127 0.1622453127 -0.2076671876
-0.3328484376 -0.7096484376 0.6930078126

Eigenvalores de B:

[2.525276998 1.205404820 1.990500003 0.2520500010 -0.06256638902
-0.99 e-10 0.008972070961 -0.002400000128]

Eigenvectores de B:

[0.6058562529 0.2296964724 -0.7677207368 0.4316787311 -0.5172437759
0.01417372281 -0.46580781e-9 -0.8143649716e-7]

[-0.7819214747 -0.05827782406 -0.3 e-9 0.5693027760 -0.5172437708
-0.3527316908 -0.001435998137 -0.00496138898]

[-0.3491743539 0.5108815667 0.32 e-8 -0.2936974758 -0.5172437978
0.0175352834 -0.5007180056 -0.2475194203]

[0.2333299732 -0.2883930378 -0.12 e-8 -0.5262683194 -0.5172437694
-0.2003806451 -0.4992820090 0.2524805787]

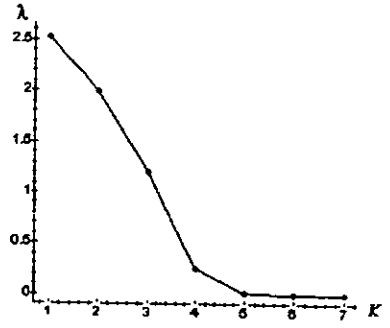
[0.2333299736 -0.2883930376 -0.11 e-8 -0.5262683198 -0.5172437943
-0.2003806592 0.5007180039 -0.2475194211]

[-0.3491743539 0.5108815671 0.33 e-8 -0.2936974755 -0.5172437723
0.1753528491 0.4992820068 0.2524805801]

[-0.1973342590 -0.8487715651 -0.23 e-8 0.2026038040 -0.5172437838
0.3780467548 -0.51547803 e-9 -0.736406339 e-7]

[0.6050882392 0.2323758621 0.7677207371 0.4363462826 -0.5172437786
0.01056679709 0.001435997359 -0.004961382982]

Observando la gráfica de las dimensiones contra los eigenvalores se puede notar un codo sobre la dimensión 4, por lo que se considera como la solución más apropiada la de dimensión 3.

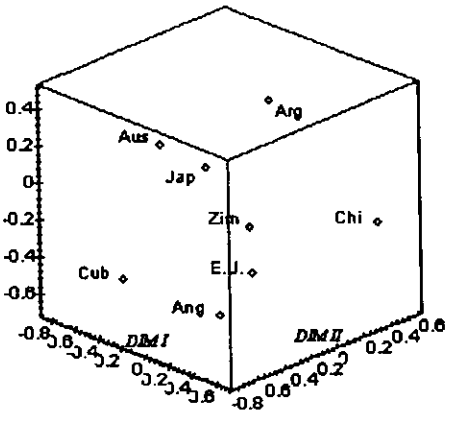


K vs. λ

Configuración derivada en dimensión tres

	Dimensión I	Dimensión II	Dimensión III
Angola	0.6058562529	-0.7819214747	-0.349174353
Argentina	0.2296964724	-0.5827782406e-1	0.510881566
Australia	-0.7677207368	-0.3 e-9	0.32e-8
China	0.4316787311	0.5693027760	-0.2936974758
Cuba	-0.5172437759	-0.5172437708	-0.5172437978
Japón	0.0141737228	-0.3527316908	0.175352834
Est. Unid.	-0.46580781e-9	-0.1435998137e-2	-0.500718005
Zimbawe	-0.8143649716e-7	-0.4961388983e-2	-0.247519420

La proporción explicada de la varianza de la matriz original queda especificada por los valores de $\alpha_{1,3} = 94.6\%$ y $\alpha_{2,3} = 48.2\%$. Sin embargo, la interpretación de la configuración obtenida no es sencilla si no se posee un conocimiento amplio sobre estas naciones y sus características, lo cual no forma parte del propósito de este trabajo.



Configuración obtenida en tres dimensiones

7.3 Ocho naciones y cuatro sujetos

Supóngase que la disimilaridad de ocho naciones es evaluada por cuatro sujetos generando las matrices que se muestran más adelante.

El análisis de estos datos se lleva a cabo en dos pasos, primero se utiliza el modelo ponderado considerando las cuatro matrices, después se obtiene el promedio de ellas y se le aplica el modelo no métrico. Al final se comparan las soluciones obtenidas.

Datos originales (disimilaridades) para el sujeto 1

	Ang 1	Arg 2	Aus 3	Chi 4	Cub 5	Jap 6	E.U. 7	Zim 8
1	0.00							
2	1.41	0.00						
3	1.00	1.00	0.00					
4	1.00	1.73	1.41	0.00				
5	1.41	1.41	1.73	1.00	0.00			
6	1.41	1.41	1.00	1.00	1.41	0.00		
7	1.73	1.00	1.41	1.41	1.00	1.00	0.00	
8	0.71	1.41	1.00	1.00	1.41	1.41	1.73	0.00

Datos originales (disimilaridades) para el sujeto 2

	Ang 1	Arg 2	Aus 3	Chi 4	Cub 5	Jap 6	E.U. 7	Zim 8
1	0.00							
2	1.00	0.00						
3	2.00	2.00	0.00					
4	3.00	3.00	2.00	0.00				
5	1.00	1.00	3.00	1.00	0.00			
6	2.00	2.00	1.00	2.00	1.00	0.00		
7	3.00	3.00	2.00	3.00	2.00	2.00	0.00	
8	1.00	1.00	3.00	1.00	3.00	3.00	1.00	0.00

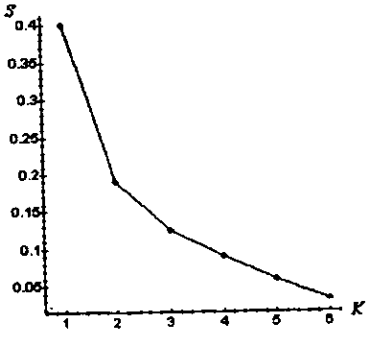
Datos originales (disimilaridades) para el sujeto 3

	Ang 1	Arg 2	Aus 3	Chi 4	Cub 5	Jap 6	E.U. 7	Zim 8
1	0.00							
2	1.41	0.00						
3	1.41	1.00	0.00					
4	1.00	1.41	1.41	0.00				
5	1.00	1.41	1.41	1.00	0.00			
6	1.41	1.00	1.00	1.41	1.41	0.00		
7	1.41	1.00	1.00	1.41	1.41	1.00	0.00	
8	1.00	1.41	1.41	1.00	1.00	1.41	1.41	0.00

Datos originales (disimilaridades) para el sujeto 4

	Ang 1	Arg 2	Aus 3	Chi 4	Cub 5	Jap 6	E.U. 7	Zim 8
1	0.00							
2	1.00	0.00						
3	0.00	1.00	0.00					
4	1.00	0.00	1.00	0.00				
5	1.41	1.00	1.41	1.00	0.00			
6	1.00	1.41	1.00	0.00	1.00	0.00		
7	1.41	1.00	1.41	1.00	0.00	1.00	0.00	
8	0.00	1.00	0.00	1.00	1.41	1.00	1.41	0.00

Los resultados del análisis ponderado fueron obtenidos por Davison (1983) con el programa *SINDSCAL* (Pruzansky,1975), unicamente se presenta la solución en dimensión 3, aunque en la gráfica de las dimensiones contra los valores del Stress aparece un codo muy visible sobre la dimensión 2 y otro menos notorio sobre la 3.



K vs. Stress

Configuración derivada en dimensión 3

	Nombre del Objeto	Dimensión 1	2	3
1	Angola	0.33	0.37	-0.32
2	Argentina	-0.35	-0.13	-0.56
3	Australia	-0.38	0.43	0.03
4	China	0.40	0.00	0.45
5	Cuba	0.34	-0.53	0.06
6	Japón	-0.33	0.02	0.54
7	Est. Unid.	-0.36	-0.51	0.08
8	Zimbawe	0.34	0.35	-0.28

Matriz de pesos para los sujetos

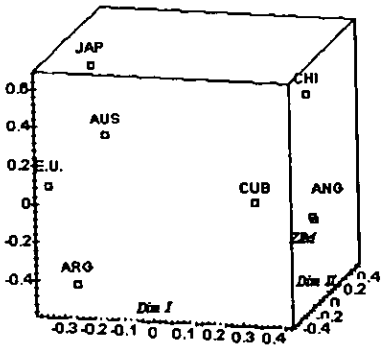
	Sujeto 1	2	3	4	Dimensión
	0.57	0.34	1.00	0.05	Marxista-Capitalista
W =	0.66	0.24	0.00	0.78	Este-Oeste
	0.33	0.46	0.00	0.35	Norte-Sur

Una posible interpretación de la configuración derivada se obtiene al observar que a lo largo de la dimensión 1, aparecen los paises de régimen marxista: Angola, China, Cuba y Zimbawe en la parte positiva, y los de régimen capitalista: Argentina, Australia, Japón y Estados Unidos en la negativa. Esta podría llamarse la dimensión Marxista-Capitalista (*vista uno*).

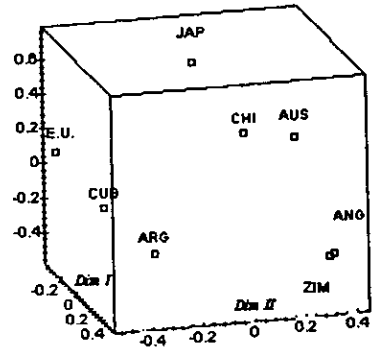
Los países en el hemisferio Este del planeta: Angola, Australia, China y Japón aparecen en la parte positiva o a la mitad de la dimensión 2. Argentina, Cuba y los Estados unidos, los países del Oeste, aparecen en la parte negativa. Esta podría llamarse la dimensión Este-Oeste (*vista dos*).

Excepto por la posición de Australia la dimensión 3 podría ser la dimensión Norte-Sur. Tres de las naciones del hemisferio Sur: Angola, Argentina y Zimbawe tienen coordenada negativa. Japón, China y los Estados Unidos, los cuales están en el

hemisferio Norte tienen coordenada positiva (aunque para el último es demasiado pequeña). Cuba, que es el país más cercano al ecuador, tiene la coordenada menor sobre esta dimensión, siendo consistente con el interpretación Norte-Sur (vista uno o dos).



vista uno



vista dos

Analizando la matriz de pesos W (pag. 113) se tiene que para el primer sujeto las dimensiones Marxista-Capitalista y la de Este-Oeste son las más importantes.

Los pesos que el segundo sujeto asigna a las tres dimensiones muestran un nivel de diferencia pequeño. Los juicios del tercer sujeto parace que reflejan únicamente la dimensión Marxista-Capitalista. Para el sujeto cuatro la dimensión Este-Oeste es la más importante, seguida de la Norte-Sur, siendo poca la importancia dada a la dimensión 1.

Para la matriz de disimilaridad promedio P que se muestra a continuación, se obtienen los resultados siguientes con el programa *INDSCAL* del paquete *SPSS*.

$P =$	Ang	Arg	Aus	Chi	Cub	Jap	E.U.	Zim
Angola	0.000	1.205	1.1025	1.500	1.205	1.455	1.8877	0.6775
Argentina	1.205	0.000	1.250	1.535	1.205	1.455	1.500	1.205
Australia	1.1025	1.250	0.000	1.455	1.8877	1.000	1.455	1.3525
China	1.500	1.535	1.455	0.000	1.000	1.1025	1.705	1.000
Cuba	1.205	1.205	1.8877	1.000	0.000	1.205	1.1025	1.705
Japón	1.455	1.455	1.000	1.1025	1.205	0.000	1.250	1.705
Est.Unid.	1.8877	1.500	1.455	1.705	1.1025	1.250	0.000	1.3875
Zimbawe	0.6775	1.205	1.3525	1.000	1.705	1.705	1.3875	0.000

Resumen del Procedimiento

Casos					
Validos		Faltantes		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
8	100.0%	0	0.0%	8	100.0%

a. Distancia Euclideana

Alscal Opciones del Procedimiento

```
  Opciones de los Datos-
Numero de Renglones (Observaciones/Matriz).    8
Numero de Columnas (Variables)  . . . . .      8
Numero de Matrices  . . . . . .  . . . .       1
Nivel de Medición   . . . . . .  .  Ordinal
Forma de la Matriz  . . . . . .  .  Simétrica
Tipo  . . . . . .  . . . . . .  .  Disimilaridad
Tratamiento de Empates  . . . . .  .  Primario
Conditionalidad  . . . . . .  . . .  Matriz
Corte de los datos  . . . . . .  .  0.000000

  Optiones del Modelo-
Modelo  . . . . . .  . . . . .  Euclideano
Máxima Dimensionalidad  . . . . . .  . . .  3
Minima Dimensionalidad  . . . . . .  . . .  1
Pesos Negativos  . . . . .  No Permitidos

  Opciones de Salida-
Encabezado . . . . . .  . . . . .  Impreso
Matrices de Datos  . . . . . .  .  Impresas
Configuraciones y Transformaciones .  Impresas
Archivo de Salida  . . . . . .  .  No Creado
Coordenadas Iniciales  . . . . . .  Calculadas

  Opciones del Algoritmo-
Máximo de Iteraciones  . . . . . .  .  30
Criterio de Convergencia  . . . . . .  0.00100
Minimo S-stress  . . . . . .  . . .  0.00500
Datos faltantes estimados por  . .  "Ulbounds"
"Tiestore" . . . . . .  . . . . .  28
```

Datos originales (distancias) para el Sujeto 1

	1	2	3	4	5	6	7	8
1	0.000							
2	1.832	0.000						
3	1.937	1.959	0.000					
4	2.244	2.257	2.336	0.000				
5	2.346	2.017	2.763	1.804	0.000			
6	2.435	2.204	1.710	1.842	1.962	0.000		
7	2.817	2.263	2.378	2.479	1.942	1.999	0.000	
8	1.339	1.950	2.141	1.938	2.590	2.624	2.545	0.000

A pesar de que en la gráfica de las dimensiones contra los valores del Stres y del S-Stres puede observarse un codo sobre la dimensión 2 y uno más pequeño sobre la dimensión 3, se considera únicamente la solución en dimensión 3 para poder hacer la comparación con los resultados de los dos apartados anteriores.

>Precaución # 14654
El número total de parámetros a estimar (el número de coordenadas más el número de pesos si los hay) es muy grande comparado con el número de datos en la matriz. Los resultados pueden no ser confiables debido a que no hay suficientes datos para estimar de manera precisa los valores de los parámetros. Debería reducirse el número de parámetros (i.e. requerir menos dimensiones) o incrementar el número de observaciones.
El número de parámetros es 24. El número de datos es 28.

Desarrollo de las Iteraciónes para la solución en dimensión 3
(calculando distancias cuadradas)
Se utiliza el S-stress formula 1 de Young.

Iteración	S-stress	Mejoramiento
1	0.10865	
2	0.08920	0.01945
3	0.08378	0.00542
4	0.07993	0.00385
5	0.07658	0.00335
6	0.07351	0.00307
7	0.07063	0.00288
8	0.06779	0.00284
9	0.06495	0.00284
10	0.06210	0.00285
11	0.05928	0.00282
12	0.05710	0.00218
13	0.05564	0.00146
14	0.05481	0.00084

Las iteraciones se detuvieron debido a que el cambio en el S-stres es menor que 0.00100. El Stress y la correlación cuadrada (RSQ) de las distancias. Los valores de RSQ dan la proporción de la varianz de los datos transformados (disparidades), la cual es contada en s correspondientes distancias.

Los valores son del Stress formula 1 de Kruskal.
Para la matriz
Stress = 0.03732 RSQ = 0.98412

Configuración derivada en 3 dimensiones

Número del Objeto	Nombre del Objeto	Dimensión 1	2	3
1	ANGOLA	1.7462	0.5505	0.1028
2	ARGENTINA	-0.0627	-0.0932	0.7503
3	AUSTRALIA	0.9800	-1.2890	-0.6634
4	CHINA	0.1512	0.8333	-1.1254
5	CUBA	-1.6682	1.0502	-0.0348
6	JAPON	-1.2170	-0.8783	-0.8024
7	EST.UNID.	-1.6583	-0.5912	1.3324
8	ZIMBAWE	1.7290	0.4177	0.4405

Datos transformados (disparidades) para el sujeto 1

	1	2	3	4	5	6	7	8
1	0.000							
2	2.107	0.000						
3	2.107	2.107	0.000					
4	2.107	2.107	2.325	0.000				
5	3.404	2.107	3.589	2.107	0.000			
6	3.404	2.107	2.107	2.107	2.107	0.000		
7	3.796	2.107	3.404	3.404	2.107	2.107	0.000	
8	0.363	2.107	2.107	2.107	3.529	3.529	3.529	0.00

Configuración Derivada en 3 dimensiones

Modelo Euclideano

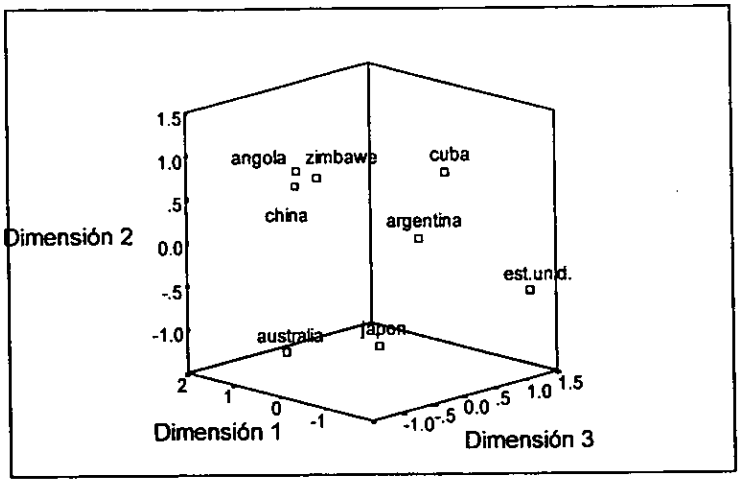


Diagrama de ajuste lineal

Modelo Euclideano

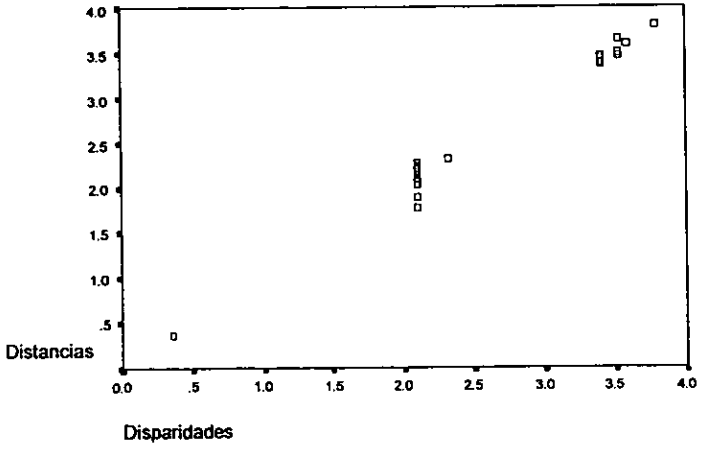


Diagrama de Shepard a

Modelo Euclideano

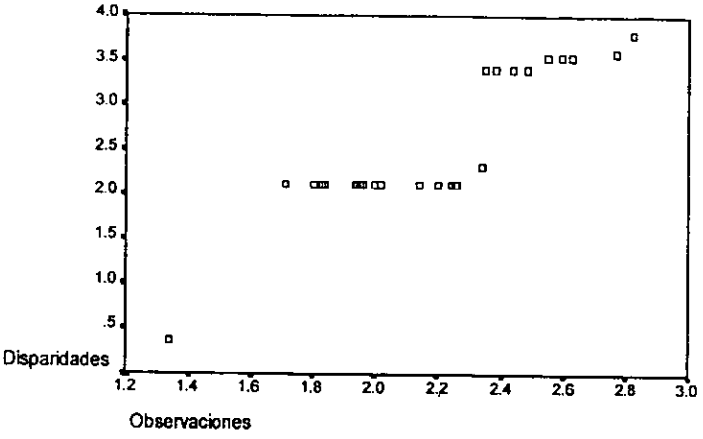
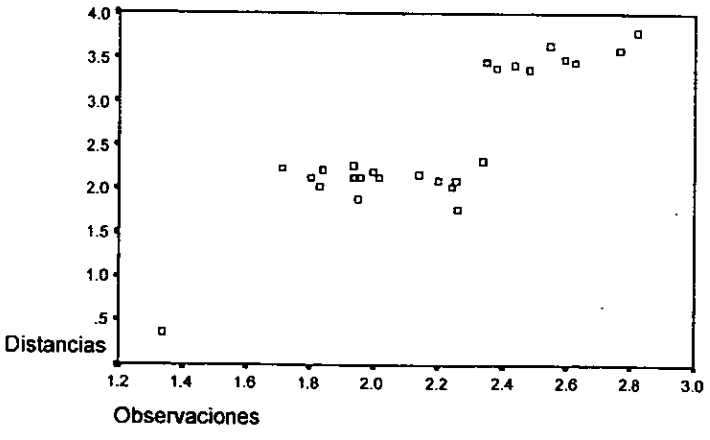


Diagrama de Shepard b

Modelo Euclideano



Analizando la configuración obtenida se puede observar que a lo largo de la dimensión 2 los países están ordenados de forma que los de régimen Marxista: Angola, China, Cuba y Zimbawe quedan en la parte positiva y los capitalistas: Argentina, Australia, Japón y los Estados Unidos en la negativa, correspondiendo con la posición a lo largo de la dimensión 1 de la solución obtenida con el modelo ponderado.

Asimismo, sobre la dimensión 1 tres de los países del hemisferio Este del planeta: Angola, Australia y China (el otro es Japón) aparecen del lado positivo, mientras que los del hemisferio Oeste: Argentina, Cuba y los Estados Unidos del lado negativo. Esta dimensión corresponde de manera sensible a la dimensión 2 de la solución del modelo ponderado.

Finalmente, para la dimensión 3 (a excepción de Australia y los Estados Unidos) la coordenada de los países del hemisferio Sur del planeta: Angola, Argentina y Zimbawe es positiva y para los del Norte: China Cuba y Japón, es negativa. El país más cercano al Ecuador: Cuba, tiene la coordenada más proxima a cero sobre esta dimension. La misma interpretación se obtiene de la dimensión 3 en la solución del modelo ponderado, pero con las coordenadas de signo contrario.

Puede verse claramente que para este caso las soluciones obtenidas con el modelo ponderado y las que se obtienen con el no métrico y la matriz promedio son muy similares. Sin embargo, al promediar las matrices no se dispone de la información sobre la diferencia en la percepción de los sujetos, que proporciona la matriz de pesos.

Referencias

Attneave, F. (1950) *Dimensions of similarity.* American journal of Psychology 53, p. 516-556.

Bloxom, B. (1968) *Individual Differences in Multidimensional Scaling.* Research Bulletin, p. 68-95. Princeton N. J. Educational Teaching Science.

Bloxom, B. (!978). *Constrained Multidimensional Scaling in N spaces.* Psychometrika 43, p. 397-408

Carroll, J. D. y Chang, J. J. (1970) *Analysis of individual differences in Multidimensional Scaling via an N-way generalization of "Ekcart-Young" decomposition.* Psychometrika 35, p. 283-319

Carroll, J. D. y Chang, J. J. (1972) *IDISCAL (Individual Differences In Orientation Scaling) A generalization of INDSCAL Allowing idiosyncratic reference systems as well as an analytic aproximation to INDSCAL.* Paper presented to the Psychometric Society. Princeton, N. J., March.

De Leeuw, J. (1977) *Correctness of Kruskal's algorithms for monotone regresion with ties.* Psychometrika 42, p. 141-144.

De Leeuw, J. y Pruzansky, S.(1978) *A new computational method to fit the wheighted Euclidean distance model.* Psychometrika 43, p. 479-490.

Guttman, l (1968) *A general nonmetric technique for finding the smallest coordinate space for a configuration of points.* Psychometrika 33, p. 469-504.

Kaiser, H. F. (1958) *The varimax criterion for analytic rotation in factor analysis.* Psychometrika 23, p. 187-200.

Kruskal, J. B. (1964) *Multidimensional Scaling by optimizing goodness of fit to a nonmetric hypothesis.* Psychometrika29, p. 1-28.

Lingoes, J. C. (1973) *The Guttman-Lingoes nonmetric program series.* Ann Arbor, M. I. Mathesis Press.

Lingoes, J. C. y Borg, I. (1978) *A direct approach to individual differences scaling using increasingly complex transformation.* Psychometrika 43, p. 491-519.

Mardia, K. V. (1978) *Some properties of classical multidimensional scaling.* Comm. Satist-Teor. Math A 7, p. 1233-1241.

Ramsay, J. O. (1978) *MULTISCALE. Four programs for multidimensional scaling by the method of maximum likelihood.* Chicago: International Education Services.

Roskman, E. E. (1969) *Data theory and algorithms for nonmetric scaling (I y II).* Unpublished mannuscrippt. Catholic University Nijmegan. The Netherlands.

Saunder, D. R. (1960) *A computer program to find de best-fitting orthogonal factor for a given hypothesis.* Psychometrika 25, p. 207-210.

Schönemann , P. H. (1972) *An algebraic solution for a class of subjetive metrics models.* Psychometrika 37, p. 441-451.

Shepard,R. N. (1962) *The analysis of proximities: multidimensional scaling with an unknown distance function.* Psychometrika 35, p. 245-255.

Takane, V., Young, F. W. y De Leeuw, I. (1977) *Nonmetric Individual differences in multidimensional scaling: An alternating least squares method with optimal scaling features.* Psychometrika 25, p. 207-210.

Torgerson, W: S. (1952) *Multidimensional Scaling: I. Theory and Method.* Psychometrika 17, p. 401-419.

Torgerson, W. S. (1958) *Theory and methods of scaling.* New York Wiley.

T ucker, L. R. (1970) *Relations between multidimensional scaling on three-mode factor analysis.* Psychometrika 37, p. 3-28.

Young, F. W. (1972) *Multidimensional Scaling: Theory and applications in the behavioral sciencies.* Vol. I. New York Seminar Press.

Young, F. W. y Lewyckyj (1979) *ALSCAL 4 User'Guide (2a. Ed.)* Chapel Hill NC: Data analysis an Theory Associates.

Young, F. W. y Householder, J (1938) *Discussion of a set of points in terms of their mutual distances.* Psychometrika 3, p. 19-22

Young, F. W. y Householder, J (1941) *A note on multidimensional psychophysics.* Psychometrika 6, p. 331-333.

Bibliografía

Borg, Ingwer y Groeneng, Patrick. *Modern Multidimensional Scaling. Theory and Applications.* 1997. Springer Series in Statistics Editorial.

Davison, Mark L. *Multidimensional Scaling.* 1983. John Wiley and Sons Inc.

Dillon, William R. y Goldstein, Matthew. *Multivariate Analysis. Methods and Applications.* 1987. Wiley.

Mardia, K. V. *Multivariate Analysis.* 1982. Academic Press Inc.