Escalado Multidimensional

Objetivo

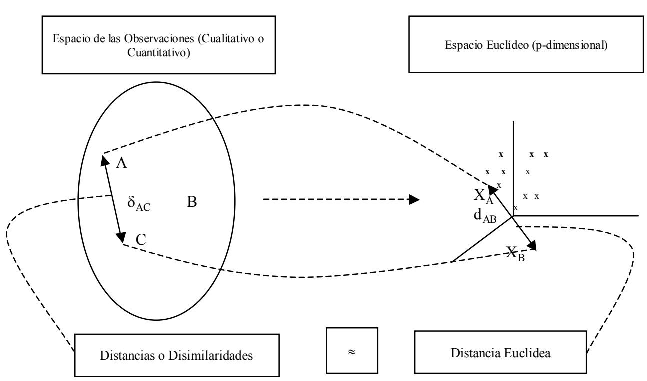
- Dados un conjunto de n objetos, para los que se conocen sus relaciones de proximidad mutuas (sus respectivas disimilaridades δ_{ij} i=1,...,n j=1,...,n),

  se trata de proyectarlos sobre un cierto espacio métrico de dimensión adecuada (p), asignándoles a cada objeto en este espacio unas coordenadas ($x_i=(x_{i1},...,x_{ip})$ i=1,...,n),

  de forma que una cierta distancia definida en este espacio entre los puntos proyectados (generalmente la distancia euclídea) d_{ij} i=1,...,n j=1,...,n, reproduzca lo más fielmente posible la relación original de proximidad entre los objetos.

Dpto. de Estadística Económica, Estructura Económica y O.E.I.
© Fco. Javier Callealta Barroso

Escalado Multidimensional



Escalado Multidimensional

Escalado Métrico vs No Métrico

- La evaluación de la fidelidad de la reproducción de la relación original de proximidad entre los objetos, induce dos metodologías diferentes:
- Escalado Multidimensional Métrico:
  - Emplea las propiedades métricas de las medidas de proximidad originales y las distancias reproducidas para evaluar la bondad de la solución.
  - Generalmente implica suponer una relación funcional entre la medida de proximidad original δ_{ij} y la distancia reproducida d_{ij}: por ejemplo, para un modelo de tipo lineal debe cumplirse que $|\delta_{ij}-(a+b\cdot d_{ij})|\leq\varepsilon\ \forall i,j$.
- Escalado Multidimensional No Métrico:
  - Emplea las propiedades de orden de las medidas de proximidad originales y las distancias reproducidas para evaluar la bondad de la solución: debe cumplirse que $\delta_{ij}\leq\delta_{kl} \Leftrightarrow d_{ij}\leq d_{kl}\quad \forall i,j,k,l$

Escalado Multidimensional Métrico

El Problema Clásico

- Conocida las distancias entre cada dos de n puntos, d_{ij} $i=1,...,n$ $j=1,...,n$, ¿podremos conocer sus coordenadas $x_i=(x_{i1},...,x_{ip})'$ en algún cierto espacio?
- Ejemplo:
  - Si conocemos el triángulo de distancias entre capitales de provincias españolas, ¿podremos identificas las coordenadas geográficas de las mismas de forma que nos permitan dibujar un mapa de posiciones?

Escalado Multidimensional Métrico.
El Problema Clásico: Ejemplo

Distancias d_{rs}	La Coruña	San Sebastián	Barcelona	Almería	Cádiz	Salamanca	Madrid
La Coruña	0						
San Sebastián	763	0					
Barcelona	1118	529	0				
Almería	1172	1032	809	0			
Cádiz	1072	1132	1284	484	0		
Salamanca	473	469	778	763	599	0	
Madrid	609	469	621	563	663	212	0

Escalado Multidimensional Métrico.
El Problema Clásico: La solución de Torgenson

- Torgenson en 1952 da la solución para cuando la distancia de partida es la distancia euclídea.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

  - Llamando $b_{ij}=x_i'x_j$, entonces $d_{ij}^2=b_{ii}+b_{jj}-2b_{ij}$
  - Este es un sistema indeterminado con n(n-1)/2 ecuaciones l.i. y n(n+1)/2 incógnitas, con $b_{ij}=b_{ji}$
  - Fijando como nuevo origen de coordenadas el centroide de los datos, entonces $\Sigma_i b_{ij}=\Sigma_j b_{ij}=0$
  - Así se convierte en sistema determinado.

Escalado Multidimensional Métrico.
El Problema Clásico: La solución de Torgenson

- Debe resolverse: $d_{ij}^2=b_{ii}+b_{jj}-2b_{ij}$ con $\Sigma_i b_{ij}=\Sigma_j b_{ij}=0$

$$\left.\begin{array}{l} d_{i\cdot}^2 = nb_{ii}+T, i=1,2,\dots,n \\ d_{\cdot j}^2 = nb_{jj}+T, j=1,2,\dots,n \\ d_{\cdot\cdot}^2 = nT+nT=2nT \end{array}\right\} \Rightarrow \begin{cases} T=\dfrac{d_{\cdot\cdot}^2}{2n} \\ b_{ii}=\dfrac{d_{i\cdot}^2-T}{n}=\dfrac{d_{\cdot i}^2-T}{n}; i=1,2,\dots,n \\ b_{ij}=\dfrac{b_{ii}+b_{jj}-d_{ij}^2}{2}; i\neq j; \begin{cases} i=1,2,\dots,n \\ j=1,2,\dots,n \end{cases} \end{cases}$$

- Así se obtiene la matriz $B=((b_{ij}))=XX'$

Dpto. de Estadística Económica, Estructura Económica y O.E.I.

Escalado Multidimensional Métrico.
El Problema Clásico: La solución de Torgenson

- Si (λ_i, e_i) son los autovalores y los respectivos autovectores ortonormalizados de B, entonces una solución al problema es:

$$X=(\sqrt{\lambda_1}\cdot e_1, \sqrt{\lambda_2}\cdot e_2, \dots, \sqrt{\lambda_n}\cdot e_n)$$

- Las variables X_i están incorrelacionadas
- Sus varianzas valen: $Var(X_i)=\lambda_i$
- Pero la solución no es única: Cualquier rotación ortogonal de la misma, provee una nueva solución:
  - Si X es solución y R ortogonal $\Rightarrow XR'$ es solución

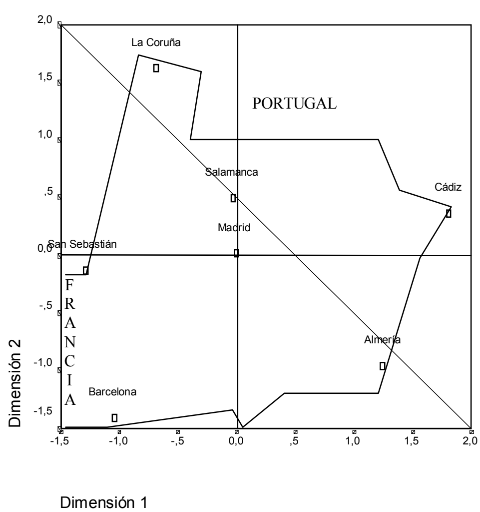
Escalado Multidimensional Métrico.
El Problema Clásico: Ejemplo

Distancias d_{rs}	La Coruña	San Sebastián	Barcelona	Almería	Cádiz	Salamanc	Madrid
La Coruña	0						
San Sebastián	763	0					
Barcelona	1118	529	0				
Almería	1172	1032	809	0			
Cádiz	1072	1132	1284	484	0		
Salamanca	473	469	778	763	599	0	
Madrid	609	469	621	563	663	212	0

Coordenadas de las Capitales

	Dimensión 1	Dimensión 2
La Coruña	-0,6884	1,6225
San Sebastián	-1,2945	-0,1263
Barcelona	-1,0401	-1,4116
Almería	1,2471	-0,9566
Cádiz	1,8101	0,3605
Salamanca	-0,0286	0,4996
Madrid	-0,0058	0,0118

Escalado Multidimensional Métrico.
El Problema Clásico: Ejemplo



Escalado Multidimensional No Métrico.

- Conocida las disimilaridades entre cada dos de n puntos, δ_{ij} i=1,...,n j=1,...,n, ¿podremos asignar a cada punto unas coordenadas $x_i=(x_{i1},...,x_{ip})'$ en algún cierto espacio, cuyas distancias reproduzcan las relaciones de orden de las disimilaridades entre cada dos individuo?

- Ejemplo:
  - Si construimos un triángulo de distancias aproximadas según nuestro común conocimiento (disimilaridades) entre capitales de provincias españolas, ¿podríamos obtener unas coordenadas de las mismas de forma que nos permitan dibujar un mapa aproximado de sus posiciones?

Escalado Multidimensional No Métrico

Algoritmo Básico

1.- Asignación de coordenadas $x_i=(x_{i1},...,x_{ip})'$
- Si es la primera vez, por algún método (p.e.: azar)
- Si no, usando las informaciones de 4 y 5 $\Rightarrow$ $\Delta x_{ij} = -\alpha \dfrac{dS}{dx_{ij}}$

2.- Cálculo de todas las distancias euclídeas $((d_{ij}))$

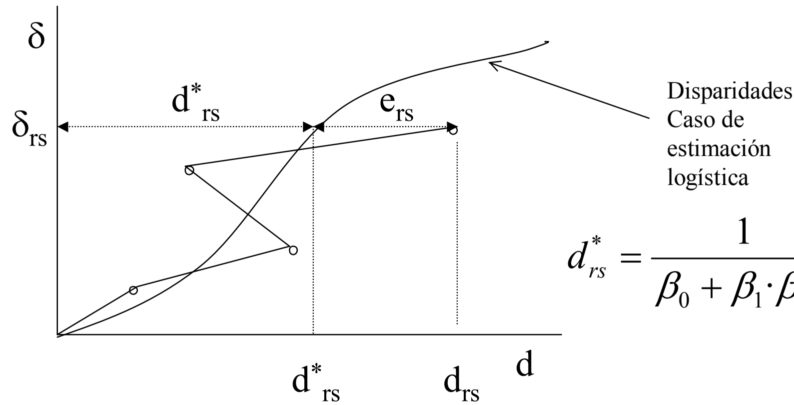
3.- Ver su adecuación con las disimilaridades iniciales $((\delta_{ij}))$:
- debe cumplirse que $\delta_{ij}\leq\delta_{kl} \Leftrightarrow d_{ij}\leq d_{kl} \quad \forall i,j,k,l$
- Si se adecua (con error tolerado), terminar. Si no, seguir

4.- A partir de las $((d_{ij}))$, obtener disparidades $((d^*_{ij}))$ tales que $\delta_{ij}\leq\delta_{kl} \Leftrightarrow d^*_{ij}\leq d^*_{kl}$ $\forall i,j,k,l$ y medida de Stress

5.- Determinar cómo varía el Stress para variaciones de cada una de las coordenadas x_{ij} , $\dfrac{dS}{dx_{ij}}$, e ir al paso 1

Escalado Multidimensional No Métrico.

Diagrama de Shepard.



$$d^*_{rs} = \frac{1}{\beta_0 + \beta_1 \cdot \beta_2^{\delta_{rs}}}$$

Escalado Multidimensional No Métrico.

Medidas de Stress

- Coef. de Determinación $R^2 = 1 - \frac{\sum e_{rs}^2}{\sum (d_{rs} - \bar{d}_{rs})^2}; \quad 0 \le R^2 \le 1$
- Stress 2 $S_2 = \left(\frac{\sum_r \sum_s e_{rs}^2}{\sum_r \sum_s (d_{rs} - \bar{d}_{rs})^2} \right)^{1/2} = \left(\frac{\sum_r \sum_s (d_{rs} - d^*_{rs})^2}{\sum_r \sum_s (d_{rs} - \bar{d}_{rs})^2} \right)^{1/2} = \sqrt{1 - R^2}; \quad 0 \le S_2 \le 1$
- Stress 1 (Kruskal) $S_1 = \left(\frac{\sum_r \sum_s e_{rs}^2}{\sum_r \sum_s d_{rs}^2} \right)^{1/2} = \left(\frac{\sum_r \sum_s (d_{rs} - d^*_{rs})^2}{\sum_r \sum_s d_{rs}^2} \right)^{1/2}; \quad 0 \le S_1 \le \sqrt{1 - \frac{2}{n}}$
  - Kruskal (1964) caracteriza los valores de S_1 como:
    0-perfect; 0,025-excel;0,05-bueno; 0,1-aceptable; 0,2-pobre
- S-Stress(1) de Young-Takane-de Leeuw $S - Stress = \left(\frac{\sum_r \sum_s \left(d_{rs}^2 - d^{*\,2}_{rs} \right)^2}{\sum_r \sum_s d^{*\,4}_{rs}} \right)^{1/2}$

Escalado Multidimensional
Dimensionalidad

- En el MDS Métrico, la dimensión del espacio la marca el número de autovalores no nulos
- En el MDS No Métrico, necesitamos estudiar cuál es la dimensión del espacio mas adecuada

- Opciones:
  - Kruskal sugiere realizar el análisis con varias dimensiones y dibujar el Stress conseguido en cada caso, para quedarnos con aquella dimensión a partir de la cual no se reduce significativamente éste

Escalado Multidimensional
El Modelo de diferencias individuales

- Desarrollado por Carroll, J.D y Chang, J. en 1970.
- Las disimilaridades entre unos mismos individuos se presentan en diversas tablas de estructura similar (p.e. distintos años)
- La distancia ajustada es:

$$d_{ij}^t = \sqrt{\sum_{k=1}^{p} w_{kt}(x_{ik} - x_{jk})^2}$$

- Como resultado, se conocen:
  - coordenadas de cada objeto en el espacio común
  - pesos de paso del espacio común a los individuales
  - coordenadas de cada objeto en cada espacio individual