# ESCALAMIENTO MULTIDIMENSIONAL

Práctica 3: Técnicas de Investigación

10 noviembre 2009

# Escalamiento Multidimensional (E.M.)

- Las técnicas de E.M son una generalización de la idea de C.P. cuando en lugar de disponer una matriz de observaciones por variables se dispone de una matriz de disimilaridades, llamada **Δ**, entre los *n* elementos del conjunto.

# Ejemplos

(a) Similitudes o distancias entre *n* productos fabricados en una empresa

(b) Distancias percibidas entre *n* candidatos políticos

Estas distancias pueden haberse obtenido a partir de ciertas variables o bien estimándolas directamente, por ejemplo, mediante cuestionarios.

# Objetivo

El objetivo será representar la matriz de disimilaridades $\Delta$ mediante un conjunto de variables ortogonales {z1,..,zp} a las que llamaremos coordenadas principales pertenecientes a una matriz **Z** y cuya matriz de distancias euclideas al cuadrado **D** reproduzca exacta o de forma aproximada a $\Delta$.

Al igual que C.P. el objetivo del E.M. será describir e interpretar los datos

# ¿Qué utilidad tiene el E.M?

- Imaginemos que existen muchos datos, en este caso la matriz de disimilaridades $\Delta$ seria muy grande y la representación por unas pocas variables de los elementos nos permitiría entender su estructura.

- Además, si podemos interpretar las variables aumentará nuestro conocimiento del problema, al entender cómo se han generado los datos.

# Escalado Multidimensional (E.M.)

*¿Es siempre posible encontrar estas z1,..,zp variables?*

En general no es posible encontrar siempre *p* variables que reproduzcan **exactamente** las distancias iniciales, sin embargo es frecuente encontrar variable que reproduzcan **aproximadamente** las distancias originales.

# Ejemplo:


- Realizamos una encuesta para determinar qué similiritudes encuentran los consumidores entre *n* productos de refrescos, y que la información se resume en una matriz cuadrada de similitudes entre los productos.

  Supongamos que utilizando el E.M. descubrimos que estas similitudes pueden generarse mediante dos variables. Entonces, seria razonable suponer que los consumidores han estimado la similitud entre los productos utilizando estas dos variables.

# Escalado métrico

| X | -> | Y | -> | Q | -> | D |
|---|---|---|---|---|---|---|
| Matriz Observa. | | Matriz Columnas centrada | | Matriz productos escalares | | Matriz distancias euclideas |

El escalado métrico es el proceso contrario

| Δ | -> | Q | -> | Z | D | = | Δ |
|---|---|---|---|---|---|---|---|
| Matriz desimilarid. | | Matriz productos escalares | | Matriz coordenadas principales | Matriz distancias euclideas | | Matriz desimilarid |

# Caso 1: Si Δ es una matriz D

| D* | -> | Q | -> | Z | D = | D* |
|---|---|---|---|---|---|---|
| Matriz distancias euclideas original | | Matriz productos escalares | | Matriz coordenadas principales | Matriz distancias euclideas | Matriz destancias euclideas original |

Q = -1/2 P D* P siendo P = I - 1/n **1 1'**

Z = V Λ ^ ½ donde Λ ^ ½ = diag ($\lambda_1$ ^ ½,....., $\lambda_p$ ^ ½)

Comparar la matriz D con la D* que es la original

# Caso 2: Si Δ es una matriz de disimilaridades

- Partimos de una matriz simétrica y con ceros en la diagonal.

Origen:

a) Estimando las distancias entre los elementos.
b) Ordenando las distancias por pares, en este caso, se pueden transformar para conseguir la matriz de disimiliridades.

Hacer el mismo proceso para encontrar Z pero partiendo de Δ

# Número de dimensiones

- Podríamos considerar solo r<p coordenadas principales donde se pueda resumir de una forma adecuada nuestra información. En este caso, seleccionaríamos las r coordenadas principales asociadas a los mayores valores propios.

# Escalado no métrico

- La diferencia entre el escalado metrico y no metrico consiste simplemente en la manera de encontrar la matriz de coordenadas Z.

- Partiendo de Δ el escalado no metrico consiste en encontrar unas coordenadas cuyas distancias euclideas al cuadrado mantengan el orden (no magnitud) de las disimilaridades.

# En general

- Partimos de una matriz de disimilaridades Δ
- Usando Δ, escribimos el orden de las distancias de menor a mayor.

- Se calculan las coordenadas iniciales,  Zo

- Calcular su matriz de distancias, Do
- Usando Do, escribir el orden de las distancias de menor a mayor

El orden se tendría que conservar

# Medidas de ajuste

- STRESS

  Adecuado si S $\in$ ( 0.05,  0.1] , si es más pequeño mejor

- S-STRESS

0<S-S<1, más cercanos a 0 indican un buen ajuste

- RSQ.  Coeficiente de correlación al cuadrado entre las distancias y las disparidades.

  Adecuado si  RSQ >= 0.6

  Si estas medidas no son satisfactorias se minimiza alguna de ellas, por ejemplo, el STRESS

# Generar los resultados en PASW

- Ir a la pagina web de la asignatura
- Copiar los 2 ficheros excel de la practica 3: “datos excel” y “datos extra”
- Desde PASW abrir el fichero “datos excel”

TEXTO

# Para 3 dimensiones

Iteration history for the 3 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

| Iteration | S-stress | Improvement |
|---|---|---|
| 1 | ,05468 | |
| 2 | ,04939 | ,00528 |
| 3 | ,04886 | ,00053 |
| 4 | ,04875 | ,00011 |
| 5 | ,04872 | ,00003 |

Iterations stopped because
S-stress improvement is less than ,000100

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.

For matrix

Stress = ,04308 RSQ = ,98803

Configuration derived in 3 dimensions

S-STRESS
0<S-S<1
cercano a 0
buen ajuste

STRESS
adecuado si
S Є ( 0.05, 0.1)

RSQ > 0.6
buen ajuste

# Para 3 dimensiones

Stimulus Coordinates

| Stimulus Number | Stimulus Name | Dimension 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | a | 1,9254 | ,1530 | -,5043 |
| 2 | b | 1,2856 | ,6664 | -,8746 |
| 3 | c | 1,1216 | -1,5687 | ,7720 |
| 4 | d | 1,2045 | ,6513 | ,0706 |
| 5 | e | ,7153 | -1,1192 | ,1023 |
| 6 | f | ,2176 | ,6396 | -,3651 |
| 7 | g | -,1672 | -,2656 | -,2030 |
| 8 | h | -,4350 | -,6408 | ,7000 |
| 9 | i | -1,6229 | -,7990 | -,5625 |
| 10 | j | -,3170 | ,6836 | ,8166 |
| 11 | k | -2,8051 | -,4572 | -,6735 |
| 12 | l | -1,1226 | 2,0566 | ,7215 |

Matriz de coordenadas principales **Z**

# Para 2 dimensiones

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

| Iteration | S-stress | Improvement |
|---|---|---|
| 1 | ,11539 | |
| 2 | ,10401 | ,01138 |
| 3 | ,10327 | ,00074 |
| 4 | ,10322 | ,00005 |

Iterations stopped because
S-stress improvement is less than ,000100

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.

For matrix

Stress = ,09021 RSQ = ,96057

Configuration derived in 2 dimensions

S-STRESS

0<S-S<1

mas alejado de 0

STRESS

regular pues esta

mas cercano a 0.1

RSQ > 0.6

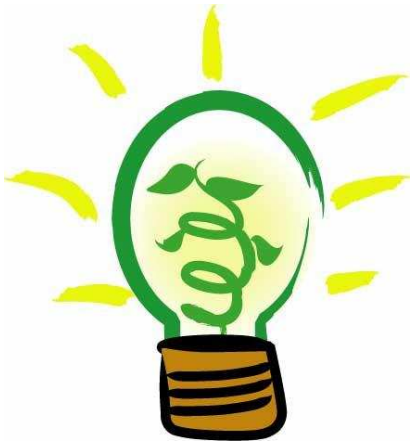
es mejor en tres dimensiones

# Para 2 dimensiones

Stimulus Coordinates

| Stimulus Number | Stimulus Name | Dimension 1 | Dimension 2 |
|---|---|---|---|
| 1 | a | 1,6659 | ,1048 |
| 2 | b | 1,1649 | ,6051 |
| 3 | c | 1,0246 | -1,3705 |
| 4 | d | ,9988 | ,5091 |
| 5 | e | ,5991 | -,8875 |
| 6 | f | ,1549 | ,4757 |
| 7 | g | -,1334 | -,1923 |
| 8 | h | -,3730 | -,5519 |
| 9 | i | -1,3836 | -,7048 |
| 10 | j | -,2646 | ,6438 |
| 11 | k | -2,4688 | -,4368 |
| 12 | l | -,9849 | 1,8053 |

Matriz de coordenadas principales **Z**



Los resultados son mejores si consideramos 3 dimensiones

Vamos a repetir el análisis agregando algunos resultados adicionales

Alscal Procedure Options

Data Options-

| | |
|---|---|
| Number of Rows (Observations/Matrix). | 12 |
| Number of Columns (Variables) . . . | 12 |
| Number of Matrices . . . . . | 1 |
| Measurement Level . . . . . | Interval |
| Data Matrix Shape . . . . . | Symmetric |
| Type . . . . . | Dissimilarity |
| Approach to Ties . . . . . | Leave Tied |
| Conditionality . . . . . | Matrix |
| Data Cutoff at . . . . . | ,000000 |

Model Options-

| | |
|---|---|
| Model . . . . . | Euclid |
| Maximum Dimensionality . . . . | 3 |
| Minimum Dimensionality . . . . | 3 |
| Negative Weights . . . . . | Not Permitted |

Output Options-

| | |
|---|---|
| Job Option Header . . . . . | Printed |
| Data Matrices . . . . . | Printed |
| Configurations and Transformations . | Plotted |
| Output Dataset . . . . . | Not Created |
| Initial Stimulus Coordinates . . . | Computed |

# Matriz de disimilaridades original Δ

Raw (unscaled) Data for Subject 1

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ,000 | | | | | | | | | | | |
| 2 | 1,000 | ,000 | | | | | | | | | | |
| 3 | 4,000 | 5,000 | ,000 | | | | | | | | | |
| 4 | 2,000 | 1,000 | 4,000 | ,000 | | | | | | | | |
| 5 | 3,000 | 4,000 | 1,000 | 3,000 | ,000 | | | | | | | |
| 6 | 3,000 | 2,000 | 5,000 | 1,000 | 3,000 | ,000 | | | | | | |
| 7 | 4,000 | 3,000 | 4,000 | 3,000 | 2,000 | 2,000 | ,000 | | | | | |
| 8 | 5,000 | 5,000 | 3,000 | 4,000 | 2,000 | 3,000 | 1,000 | ,000 | | | | |
| 9 | 7,000 | 6,000 | 6,000 | 6,000 | 4,000 | 4,000 | 3,000 | 3,000 | ,000 | | | |
| 10 | 5,000 | 4,000 | 5,000 | 3,000 | 4,000 | 2,000 | 2,000 | 2,000 | 4,000 | ,000 | | |
| 11 | 9,000 | 8,000 | 8,000 | 8,000 | 7,000 | 6,000 | 5,000 | 5,000 | 2,000 | 6,000 | ,000 | |
| 12 | 7,000 | 6,000 | 8,000 | 5,000 | 7,000 | 4,000 | 5,000 | 5,000 | 6,000 | 3,000 | 6,000 | ,000 |

# Para 3 dimensiones

Iteration history for the 3 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

| Iteration | S-stress | Improvement |
|---|---|---|
| 1 | ,05468 | |
| 2 | ,04939 | ,00528 |
| 3 | ,04886 | ,00053 |
| 4 | ,04875 | ,00011 |
| 5 | ,04872 | ,00003 |

Iterations stopped because
S-stress improvement is less than ,000100

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.

For matrix

Stress = ,04308 RSQ = ,98803

Configuration derived in 3 dimensions

S-STRESS
0<S-S<1
cercano a 0
buen ajuste

STRESS
adecuado si
S Є ( 0.05, 0.1)

RSQ > 0.6
buen ajuste

# Para 3 dimensiones

Stimulus Coordinates

| Stimulus Number | Stimulus Name | Dimension 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | a | 1,9254 | ,1530 | -,5043 |
| 2 | b | 1,2856 | ,6664 | -,8746 |
| 3 | c | 1,1216 | -1,5687 | ,7720 |
| 4 | d | 1,2045 | ,6513 | ,0706 |
| 5 | e | ,7153 | -1,1192 | ,1023 |
| 6 | f | ,2176 | ,6396 | -,3651 |
| 7 | g | -,1672 | -,2656 | -,2030 |
| 8 | h | -,4350 | -,6408 | ,7000 |
| 9 | i | -1,6229 | -,7990 | -,5625 |
| 10 | j | -,3170 | ,6836 | ,8166 |
| 11 | k | -2,8051 | -,4572 | -,6735 |
| 12 | l | -1,1226 | 2,0566 | ,7215 |

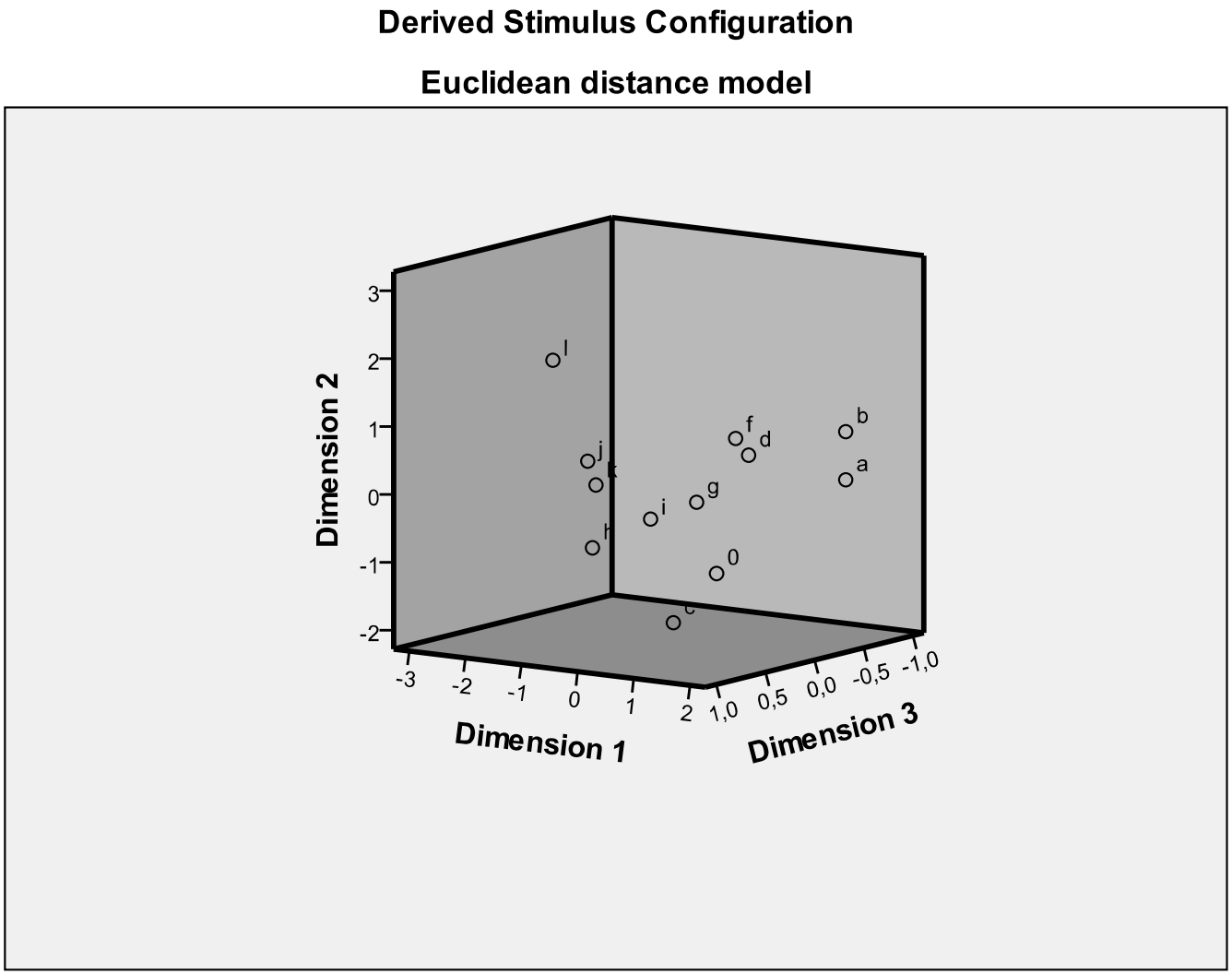
Matriz de coordenadas principales **Z**

# Matriz de disparidades

Optimally scaled data (disparities) for subject 1

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ,000 | | | | | | | | | | | |
| 2 | ,748 | ,000 | | | | | | | | | | |
| 3 | 2,243 | 2,742 | ,000 | | | | | | | | | |
| 4 | 1,247 | ,748 | 2,243 | ,000 | | | | | | | | |
| 5 | 1,745 | 2,243 | ,748 | 1,745 | ,000 | | | | | | | |
| 6 | 1,745 | 1,247 | 2,742 | ,748 | 1,745 | ,000 | | | | | | |
| 7 | 2,243 | 1,745 | 2,243 | 1,745 | 1,247 | 1,247 | ,000 | | | | | |
| 8 | 2,742 | 2,742 | 1,745 | 2,243 | 1,247 | 1,745 | ,748 | ,000 | | | | |
| 9 | 3,738 | 3,240 | 3,240 | 3,240 | 2,243 | 2,243 | 1,745 | 1,745 | ,000 | | | |
| 10 | 2,742 | 2,243 | 2,742 | 1,745 | 2,243 | 1,247 | 1,247 | 1,247 | 2,243 | ,000 | | |
| 11 | 4,735 | 4,237 | 4,237 | 4,237 | 3,738 | 3,240 | 2,742 | 2,742 | 1,247 | 3,240 | ,000 | |
| 12 | 3,738 | 3,240 | 4,237 | 2,742 | 3,738 | 2,243 | 2,742 | 2,742 | 3,240 | 1,745 | 3,240 | ,000 |

# Modelo de distancia euclidea



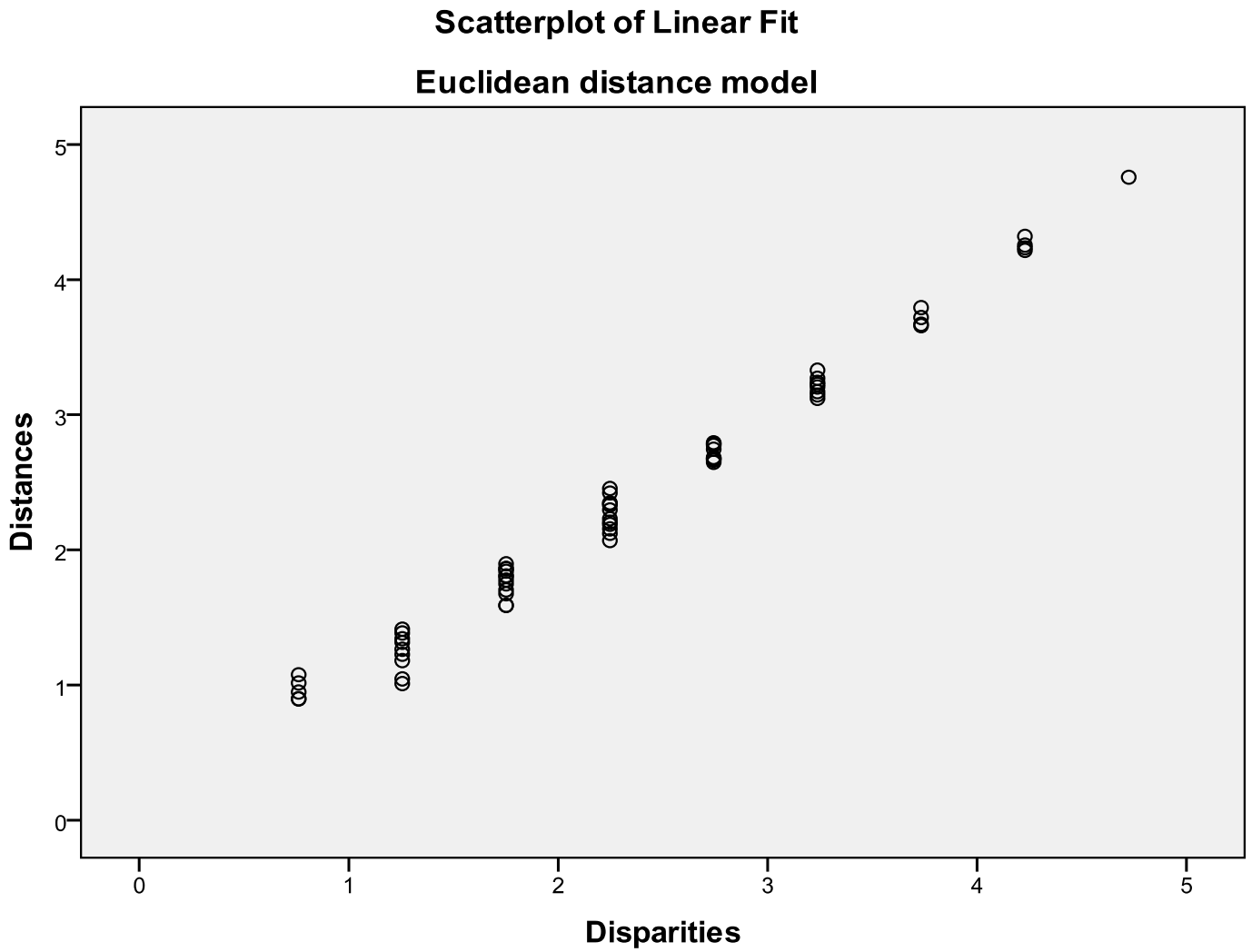



Gráfico de disparidades frente a distancias.

Si son iguales es porque las distancias tienen el orden de las disimilaridades originales

Abrir el fichero “DatosExtraPractica3EMD.xls”

Crear las variables Z1, Z2 y Z3, son variables numéricas con 4 decimales

Analyze - > Dimension Reduction ->  Factor

# Correlaciones: Carácter. coches *vs*. Coordenadas Z

**Correlation Matrix**

| | | Precio | Potencia | Tamaño | Consumo | Z1 | Z2 | Z3 |
|---|---|---|---|---|---|---|---|---|
| Correlation | Precio | 1,000 | ,997 | -,064 | ,241 | **-,999** | ,001 | -,031 |
| | Potencia | ,997 | 1,000 | -,042 | ,233 | **-,997** | ,025 | -,027 |
| | Tamaño | -,064 | -,042 | 1,000 | -,059 | ,053 | **,949** | ,135 |
| | Consumo | ,241 | ,233 | -,059 | 1,000 | -,247 | ,012 | **-,938** |
| | Z1 | -,999 | -,997 | ,053 | -,247 | 1,000 | -,007 | ,033 |
| | Z2 | ,001 | ,025 | ,949 | ,012 | -,007 | 1,000 | ,034 |
| | Z3 | -,031 | -,027 | ,135 | -,938 | ,033 | ,034 | 1,000 |

Las correlaciones nos sirven para interpretar el sentido de las coordenadas principales Z

- El **Precio** y **potencia** están correlacionados negativamente con Z1.

| (-) | (+) |
| --- | --- |
| Caros y mucha potencia | bajo precio y potencia |

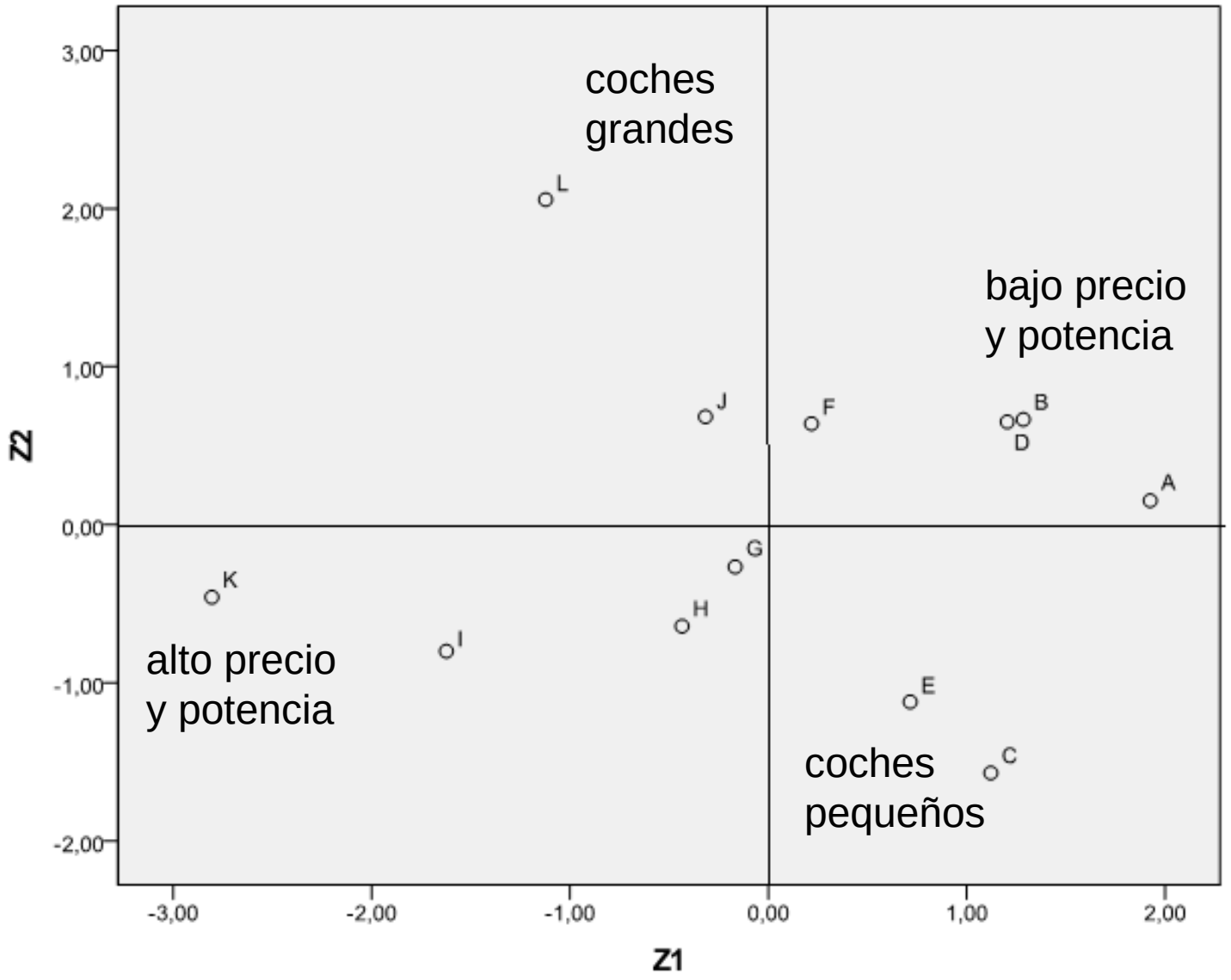
- El **tamaño** de coche esta relacionado con Z2

| (-) | (+) |
| --- | --- |
| pequeños | grandes |

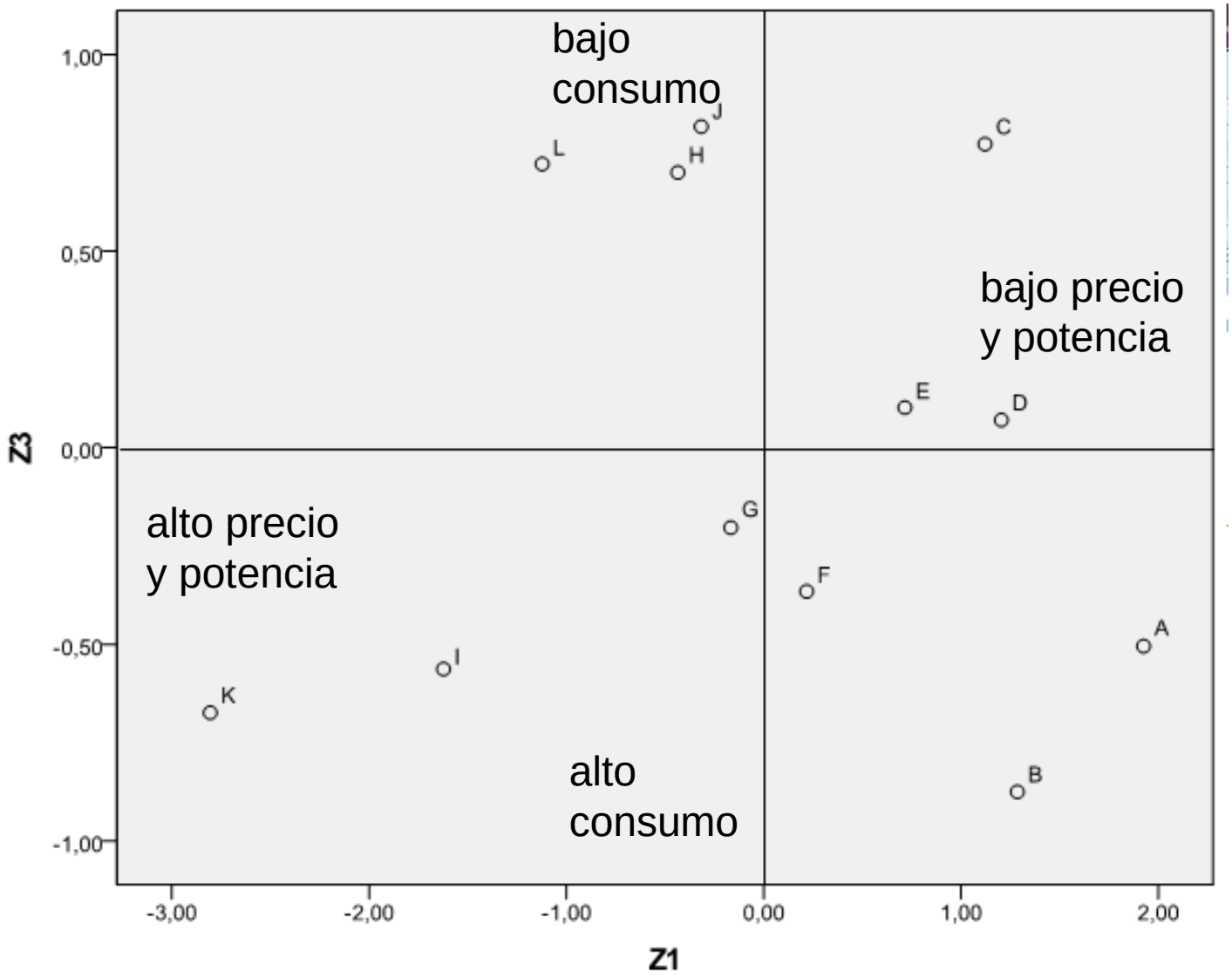
- El **consumo** esta relacionado con Z3

| (-) | (+) |
| --- | --- |
| alto consumo | bajo consumo |

# Z2 *vs.* Z1


coches
grandes
bajo precio
y potencia
alto precio
y potencia
coches
pequeños
L
J
F
B
D
A
G
K
H
I
E
C
3,00
2,00
1,00
0,00
-1,00
-2,00
-3,00
-2,00
-1,00
0,00
1,00
2,00
Z2
Z1

# Z3 *vs.* Z1


bajo consumo
bajo precio y potencia
alto precio y potencia
alto consumo
1,00
0,50
0,00
-0,50
-1,00
Z3
-3,00
-2,00
-1,00
0,00
1,00
2,00
Z1
L
H
J
C
E
D
G
F
A
I
K
B