Escalamiento multidimensional

Dra. Rocío Maehara

16 de noviembre de 2019

Introducción

El análisis de escalamiento multidimensional (MDS) es una técnica de reducción de datos como otras que hemos visto anteriormente: análisis factorial o análisis de componentes principales, por ejemplo. El objetivo principal del MDS es representar N objetos en un espacio dimensional reducido (q dimensiones, siendo $q < N$) , de tal forma que la distorsión causada por la reducción de la dimensionalidad sea la menor posible , es decir, que las distancias entre los objetos representados en el espacio q dimensional sean lo más parecidas posible a las distancias en el espacio N dimensional.

Introducción

Dado que será difícil que las distancias coincidan, el objetivo del MDS es conseguir que ambas configuraciones dimensionales sean lo más parecidas posible, Para ello será necesario construir un indicador de esa proximidad que, como se detallará más adelante, denominaremos stress o sstress.

Valoración de la imagen de superficies comerciales

```
datos <- matrix(c(
0.0, 1.0, 2.1, 6.1, 5.2,
1.0, 0.0, 2.4, 6.9, 5.3,
2.1, 2.4, 0.0, 5.1, 4.1,
6.1, 6.9, 5.1, 0.0, 3.1,
5.2, 5.3, 4.1, 3.1,0.0), ncol=5, nrow=5, byrow=T,
dimnames=list(c("X1","X2","X3","X4","X5")))
datos
```

```
##    [,1] [,2] [,3] [,4] [,5]
## X1  0.0  1.0  2.1  6.1  5.2
## X2  1.0  0.0  2.4  6.9  5.3
## X3  2.1  2.4  0.0  5.1  4.1
## X4  6.1  6.9  5.1  0.0  3.1
## X5  5.2  5.3  4.1  3.1  0.0
```

Supongamos que hemos pedido a 100 consumidores que valoren la imagen que tienen de 5 superficies comerciales, atendiendo a la similitud con que las perciben. Para ello se utiliza una escala de 0 (idénticas) a 7 (totalmente diferentes). La siguiente matriz de disparidades originales o proximidades nos muestra las medias de las puntuaciones ofrecidas por los 100 consumidores.

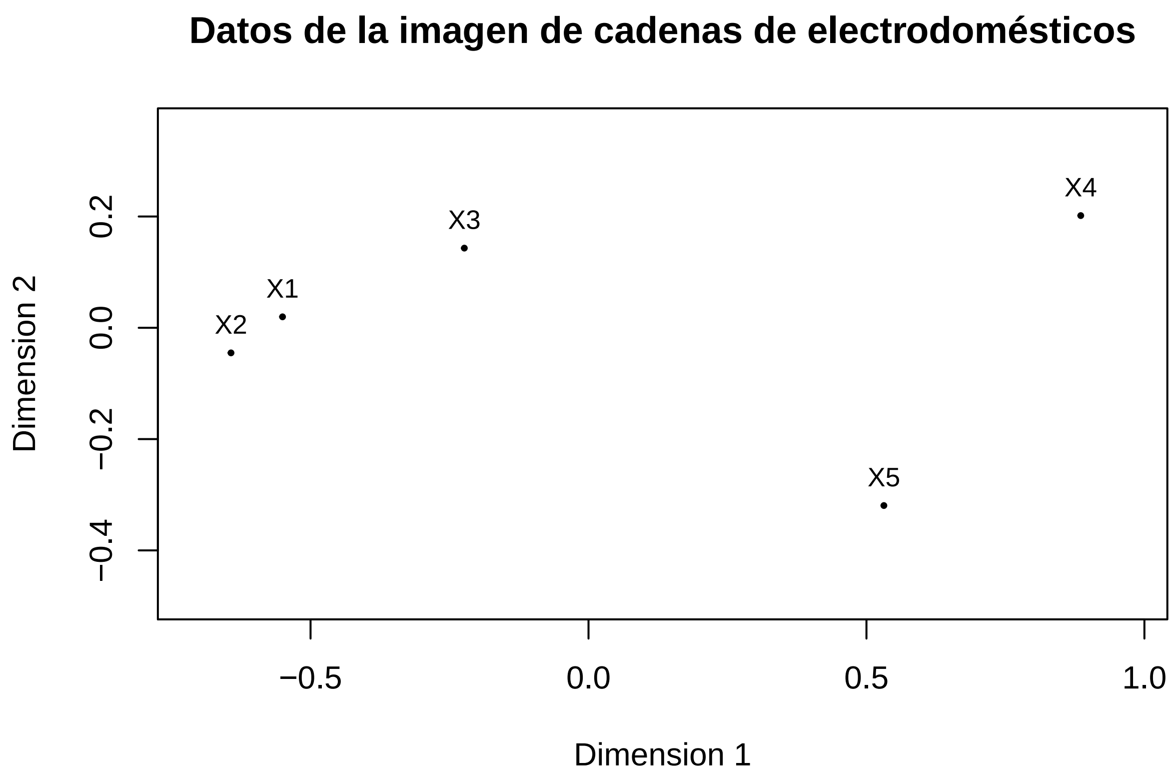
Valoración de la imagen de superficies comerciales


```
datos <- matrix(c(
0.0, 1.0, 2.1, 6.1, 5.2,
1.0, 0.0, 2.4, 6.9, 5.3,
2.1, 2.4, 0.0, 5.1, 4.1,
6.1, 6.9, 5.1, 0.0, 3.1,
5.2, 5.3, 4.1, 3.1,0.0), ncol=5, nrow=5, byrow=T,
dimnames=list(c("X1","X2","X3","X4","X5")))
datos
```

```
##     [,1] [,2] [,3] [,4] [,5]
## X1  0.0  1.0  2.1  6.1  5.2
## X2  1.0  0.0  2.4  6.9  5.3
## X3  2.1  2.4  0.0  5.1  4.1
## X4  6.1  6.9  5.1  0.0  3.1
## X5  5.2  5.3  4.1  3.1  0.0
```

Si creamos un mapa en dos dimensiones para ilustrar mejor la percepción de los consumidores, este mapa debería representar como puntos cercanos a las superficies X_1 y X_2 porque la disparidad entre ellas es pequeña (1.0), tal y como refleja la matriz. Asimismo, las superficies X_2 y X_4 deberían aparecer representadas muy distantes una de la otra, por cuanto su disparidad en la matriz es elevada (6.9).

Valoración de la imagen de superficies comerciales




```
library(smacof)
fit2 <- mds(delta=datos,ndim=2, type="interval")
plot(fit2, main = "Datos de la imagen de cadenas de electrodomésticos")
```

Se puede apreciar el mapa que se obtiene al representar las coordenadas bidimensionales resultantes de aplicar a la matriz anterior uno de los algoritmos que existen para efectuar un MDS, el implementado en `mds(smacof)`. Puede comprobarse como se constata la cercanía de X_1 y X_2 y la lejanía de X_2 y X_4 que esperábamos.

Desarrollo formal de la técnica MSD

Con el ejemplo de las superficies comerciales expuesto con anterioridad. Si partimos de N objetos (superficies comerciales), tendremos entonces $M = N(N^{-}1)/2$ disparidades originales entre pares de objetos (10 en nuestro ejemplo). Asumiendo que no haya empates (los distintos algoritmos resuelven los empates de distintos modos), las disparidades pueden escribirse en un orden estrictamente ascendente:

$$s_{i_1k_1} < s_{i_2k_2} < \ldots < s_{i_Mk_M}$$

donde $s_{i_1k_1}$ es la menor de las disparidades. El subíndice i_1k_1 indica el par de objetos que son más parecidos.

El algoritmo básico del MDS

En nuestro ejemplo esta ordenación sería la siguiente:

$$1.0 < 2.1 < 2.4 < 3.1 < 4.1 < 5.1 < 5.2 < 5.3 < 6.1 < 6.9$$

$$s_{12} < s_{13} < s_{23} < s_{45} < s_{35} < s_{34} < s_{15} < s_{25} < s_{14} < s_{24}$$

Nuestro objetivo es encontrar una nueva configuración q dimensional de los N objetos (2 dimensiones y 5 objetos en el ejemplo), de tal forma que las distancias calculadas entre ellos en ese espacio q dimensional mantengan la ordenación anterior. En el caso ideal de que se mantuviera el orden y las proporciones entre disparidades y distancias, el gráfico de dispersión entre ambas se representaría mediante una línea recta.

El algoritmo básico del MDS

```
# Solución bidimensional
fit2$conf
```

```
##            D1          D2
## X1 -0.5504010  0.01970998
## X2 -0.6430815 -0.04506942
## X3 -0.2233691  0.14315764
## X4  0.8854840  0.20160756
## X5  0.5313676 -0.31940576
```

En el MDS se van ensayando distintas configuraciones q dimensionales hasta que las distancias en ese espacio y las disparidades originales guarden una relación lo más próxima posible a esta recta ideal. El código muestra la solución bidimensional final resultante de la aplicación del MDS a los datos de nuestro ejemplo. En este cuadro aparecen las coordenadas de cada objeto (superficie comercial) en ese espacio bidimensional.

El algoritmo básico del MDS

```
#Distancias entre las configuraciones
print(fit2$confdist)
```

```
##           X1        X2        X3        X4
## X2 0.1130754
## X3 0.3495557 0.4599869
## X4 1.4473605 1.5483417 1.1103925
## X5 1.1336767 1.2060644 0.8852075 0.6299629
```

```
##            D1          D2
## X1 -0.5504010  0.01970998
## X2 -0.6430815 -0.04506942
## X3 -0.2233691  0.14315764
## X4  0.8854840  0.20160756
## X5  0.5313676 -0.31940576
```

A partir de esas coordenadas es sencillo derivar la matriz de distancia entre los distintos objetos. Así, tomando distancias euclídeas, la distancia entre, por ejemplo, X1 y X2 tomaría el valor:

$$d(X_1, X_2) = \sqrt{(-0.5504 - (-0.6431))^2 + (0.0197 - (-0.0451))^2} = 0.1131$$

Repitiendo los cálculos para todos los objetos (estímulos), obtendríamos la matriz de distancias **D** entre las configuraciones mostrada en el código.

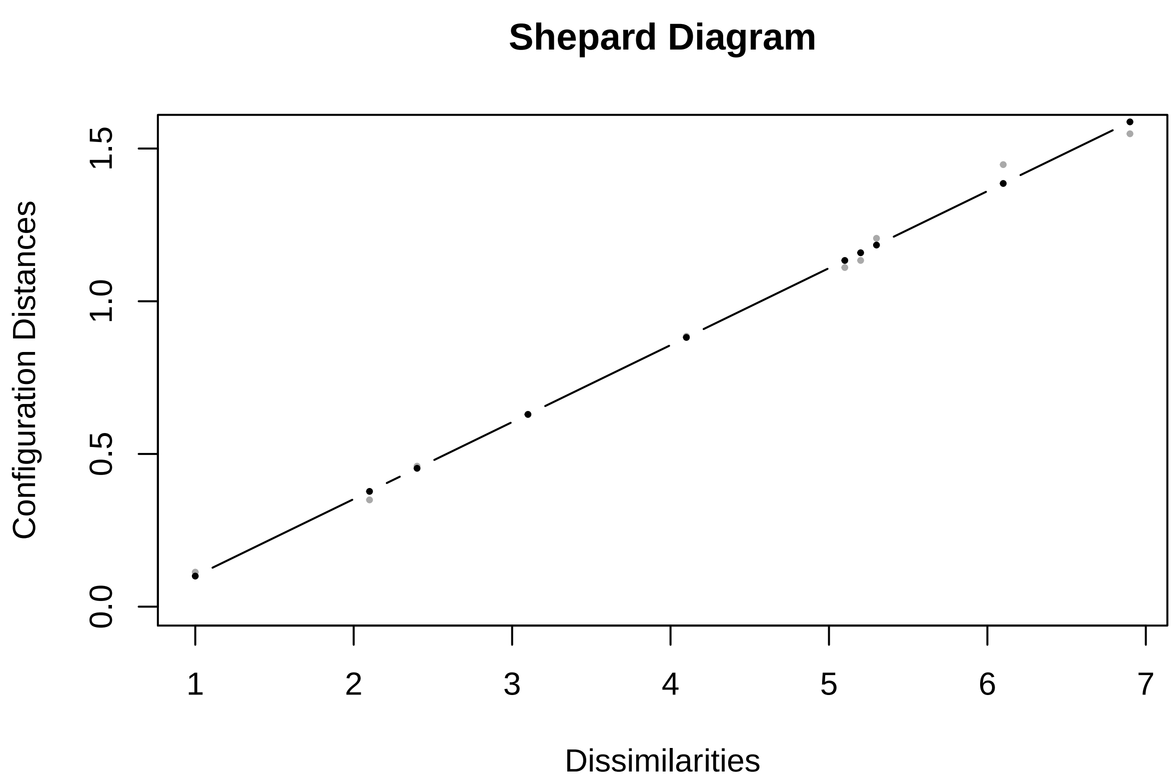
El algoritmo básico del MDS

```
# Matriz de disparidades
print(fit2$dhat)
```

```
##           X1        X2        X3        X4
## X2 0.1001109
## X3 0.3773818 0.4530011
## X4 1.3856397 1.5872913 1.1335752
## X5 1.1587817 1.1839881 0.8815108 0.6294463
```

En el algoritmo del MDS se obtiene, como hemos señalado, una transformación monótona de la matriz de distancias originales en el espacio N dimensional y es respecto a esa transformación (matriz de disparidades Δ) obtenida en cada iteración con la que se va comparando la matriz **D**. El código muestra la solución final bidimensional alcanzada en la última iteración proporciona la matriz de disparidades Δ.

El algoritmo básico del MDS




```
plot(fit2,plot.type="Shepard",plot.dim=c(1,2),sphere=TRUE,bubscale=0.1,col=1,
label.conf=list(label=TRUE,pos=3,col=1,cex=0.8),
shepard.x=NULL,identify=FALSE,
type="p",pch=20,asp=1,col.hist=NULL)
```

Así pues, la matriz de disparidades Δ es simplemente una transformación monótona de la matriz de distancias originales **S**. Esto se puede comprobar representando, simplemente, en un gráfico de dispersión las distancias que aparecen en ambas. Este gráfico es conocido como diagrama de Shepard, donde los puntos sobre la diagonal muestran la transformación monótona y los puntos grises muestran las discrepancias que se producen, que serán un indicador de la bondad del ajuste del modelo que veremos posteriormente.

El algoritmo básico del MDS

El algoritmo usado en la técnica MDS ensaya distintas configuraciones bidimensionales hasta dar con aquella que reduce en mayor grado las diferencias entre las matrices de distancias **D** y disparidades $\boldsymbol{\Delta}$. Para ello necesitamos una función objetivo que se minimizará en cada iteración. Kruskal (1964a) propuso la siguiente función, que denominó stress:

$$Stress = \sqrt{\frac{\sum_{i \neq j}(d_{ij} - \delta_{ij})^2}{\sum_{i \neq j} d_{ij}^2}}$$

donde d_{ij} son los elementos de la matriz de distancias resultante de la solución q dimensional en la interación que se esté realizando y δ_{ij} son los elementos de la matriz de disparidades que, recordemos, no son sino una transformación monótona de los elementos de la matriz de disparidades originales entre los distintos objetos (estímulos).

El algoritmo básico del MDS

$Stress$	Bondad de ajuste
$0.15 \leq Stress$	Pobre
$0.10 \leq Stress < 0.15$	Regular
$0.05 \leq Stress < 0.10$	Bueno
$0.01 \leq Stress < 0.05$	Excelente
$0.00 \leq Stress < 0.01$	Perfecto

Table: Escala para el índice *Stress*

En síntesis, el stress no es sino un indicador de cuánto difieren en promedio la matriz con las distancias de la solución dimensional reducida respecto a la matriz con las disparidades originales. El cuadrado del numerador pretende, únicamente, que no se compensen diferencias positivas con negativas. El valor del stress deberá ser tan pequeño como sea posible y, en todo caso, reducirse en cada iteración. De no ser así, el algoritmo se detendrá.

El algoritmo básico del MDS

Una segunda medida de las discrepancias entre las matrices de disparidades y distancias y, por ello, de calidad de la representación lograda por el MDS, es el estadístico denominado s-stress que fue propuesto por Takane et al. (1977) autores del algoritmo ALSCAL y que, por ello, es la función que se minimiza en ese algoritmo:

$$S - stress = \sqrt{\frac{\sum_{i \neq j}(d_{ij} - \delta_{ij})^2}{\sum_{i \neq j} d_{ij}^4}}$$

El valor del s-stress está siempre comprendido entre 0 y 1 y cualquier valor inferior a 0.1 indica que la solución obtenida es una buena representación de los objetos de la solución N dimensional inicial.

El algoritmo básico del MDS

Estímulos	d	δ	$(d-\delta)^2$	d^2	δ^2	d^4	$(d^2-\delta^2)^2$
1-2	0.113	0.100	0.000	0.013	0.010	0.000	0.000
1-3	0.350	0.377	0.001	0.122	0.142	0.015	0.000
1-4	1.448	1.386	0.004	2.097	1.920	4.395	0.031
1-5	1.134	1.159	0.001	1.286	1.343	1.654	0.003
2-3	0.460	0.453	0.000	0.212	0.205	0.045	0.000
2-4	1.549	1.587	0.001	2.399	2.519	5.757	0.014
2-5	1.207	1.184	0.001	1.456	1.402	2.119	0.003
3-4	1.111	1.134	0.001	1.234	1.285	1.523	0.003
3-5	0.886	0.882	0.000	0.785	0.777	0.615	0.000
4-5	0.630	0.629	0.000	0.397	0.396	0.158	0.000
Suma	8.887	8.891	0.008	10.000	10.000	16.281	0.055

Matriz de disparidades Δ

```
print(fit2$dhat)
```

```
##           X1        X2        X3        X4
## X2 0.1001109
## X3 0.3773818 0.4530011
## X4 1.3856397 1.5872913 1.1335752
## X5 1.1587817 1.1839881 0.8815108 0.6294463
```

Distancias entre configuraciones $\mathbf{D}$

```
print(fit2$confdist)
```

```
##           X1        X2        X3        X4
## X2 0.1130754
## X3 0.3495557 0.4599869
## X4 1.4473605 1.5483417 1.1103925
## X5 1.1336767 1.2060644 0.8852075 0.6299629
```

El algoritmo básico del MDS

Stress

```
print(fit2$stress)
```

```
## [1] 0.02826073
```

Stress por punto

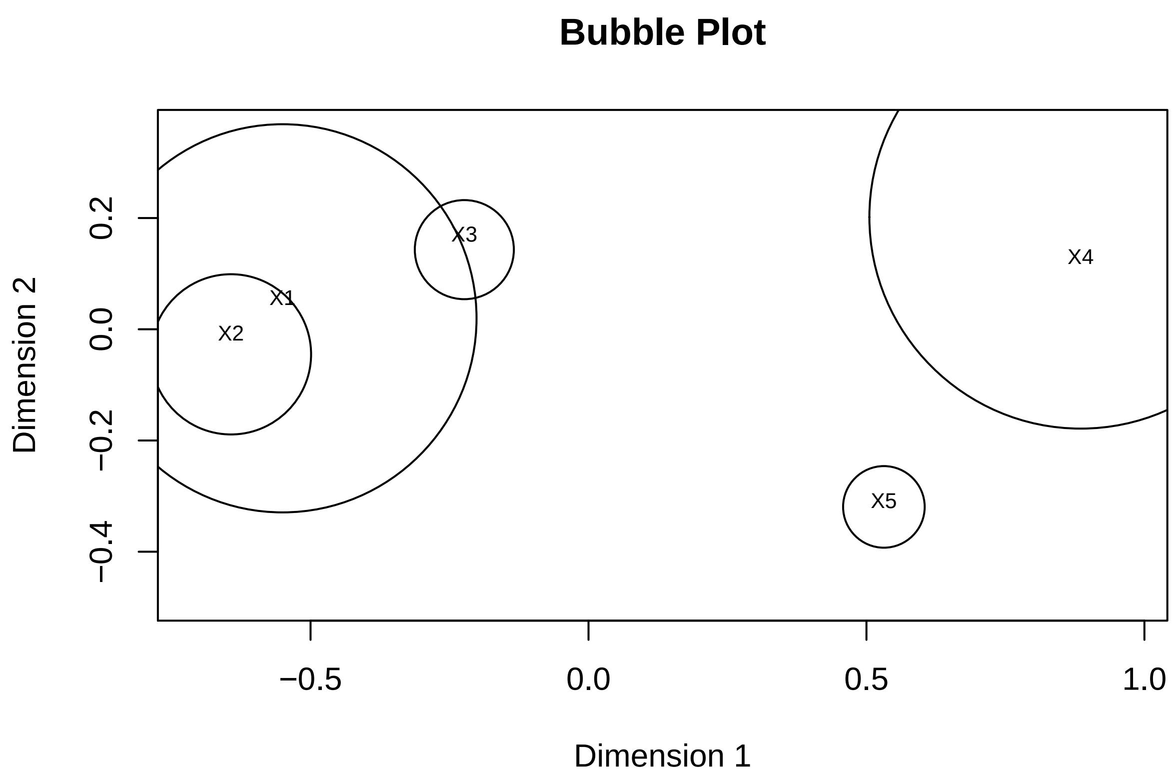
```
print(fit2$spp)
```

```
##        X1        X2        X3        X4        X5
## 33.694099 13.906328  8.603052 36.712519  7.084002
```

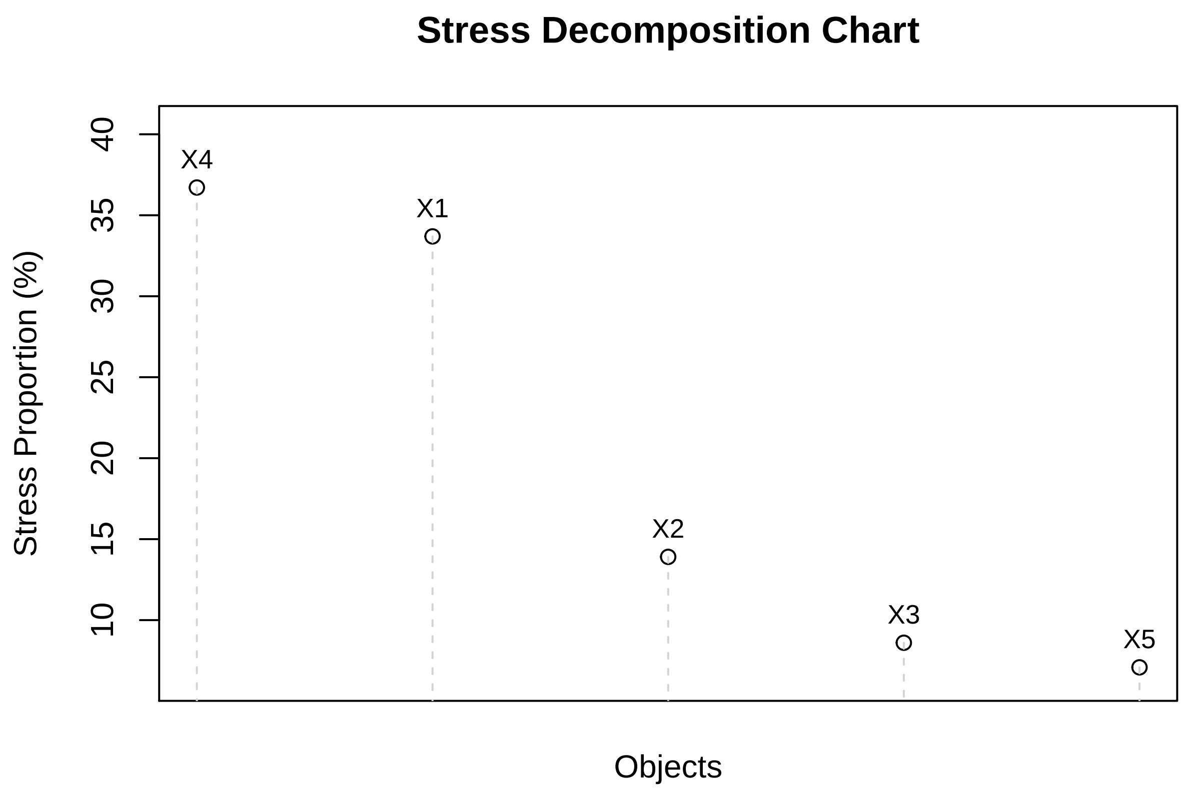
La salida de smacof proporciona el valor final del stress. Dado que este paquete no optimiza no calcula el s-stress. Se comprueba que el stress 0.0282 obtenido alcanza el valor de "excelente". El paquete smacof permite también analizar la contribución de cada punto representado al stress, es decir, del total del desajuste, qué parte es debida a una mala representación de un punto determinado. En el caso del ejemplo, vemos que el punto peor representado sería el X_4 que supone el 36,7% del total del stress.

El algoritmo básico del MDS

```
plot(fit2, plot.type = "bubbleplot")
```




```
plot(fit2, plot.type = "stressplot")
```



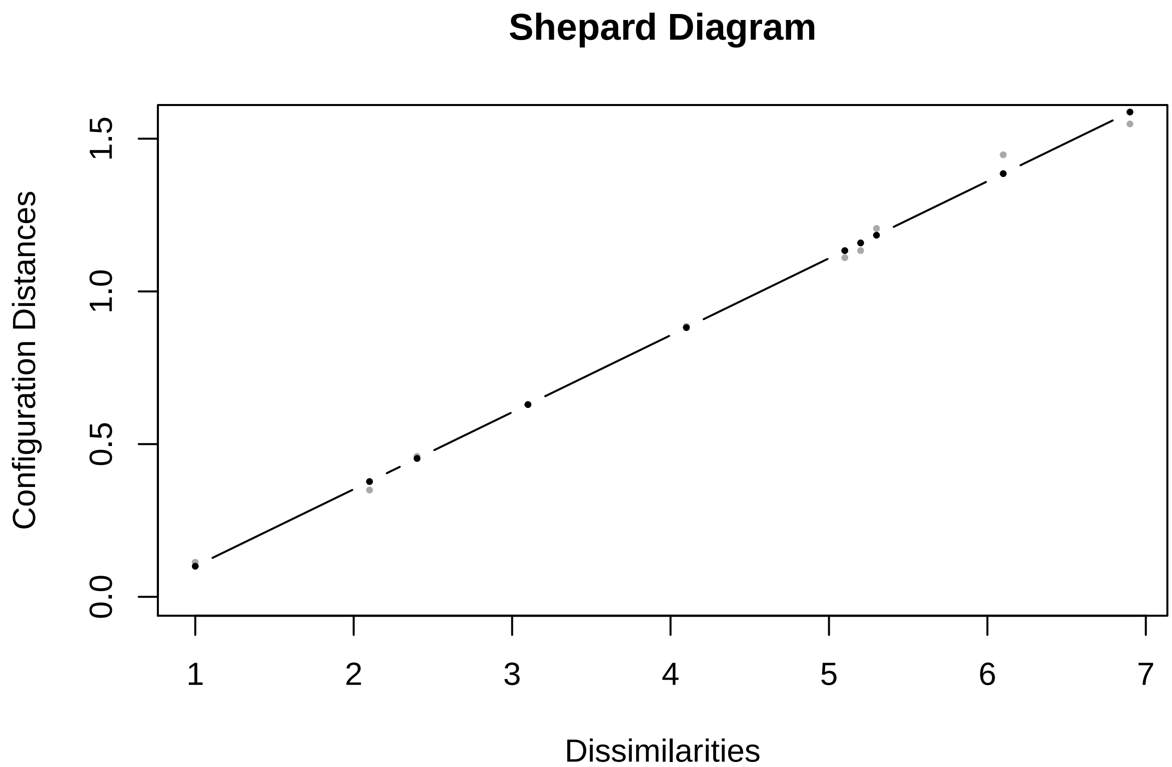
El algoritmo básico del MDS

```
print(1-fit2$rss)
```

```
## [1] 0.9920133
```

Como ya se indicó, la solución dimensional reducida es una buena representación de la solución N dimensional si la ordenación de las distancias entre los objetos de la primera mantiene la ordenación de las disparidades originales de la segunda. En el caso ideal, el gráfico de dispersión que representara distancias y disparidades debería ser una línea recta.

El algoritmo básico del MDS



Del análisis del diagrama de Shepard se desprende que la ordenación lograda con las distancias coincide de manera prácticamente perfecta con las disparidades. Ahora cabe preguntarse ¿en qué medida se aleja la relación que se ha representado de la relación ideal? Si efectuáramos una regresión simple tomando las disparidades como variables independientes y las distancias como dependientes, el coeficiente de determinación de esta regresión sería un buen indicador de lo cercano al ideal de la solución obtenida. Pues bien, ese coeficiente de determinación es el indicador RSQ que como vimos tiene un valor cercano a la unidad (0.9920).

Recogida de datos para un escalamiento

El input básico del MDS es, como ya hemos señalado, la similaridad entre cada par de los N objetos que se están analizando. A esta medida, se la suele denominar también proximidad, tal como se ha apuntado. Estas medidas pueden obtenerse de muy diversas formas. Las dos más habituales son: - Pedir a los individuos que emitan un juicio de similaridad entre cada par de estímulos. - Que puntúen en qué grado un atributo determinado está presente en el estímulo. Al primer tipo de medidas se las conoce como similaridades directas y al segundo como similaridades derivadas.

Tipos de escalamiento multidimensional

El escalamientos multidimensional no es una técnica, sino un conjunto de ellas. Los elementos que permiten diferenciarlas son:

- El número de matrices de proximidades.
- La forma de las mismas (cuadradas o rectangulares).
- Si el algoritmo contempla o no ponderaciones.

Escalamiento multidimensional clásico

Esta formado por una única matriz de proximidades y esta es cuadrada. Existen dos tipos de esclamiento multidimensional clásico en función de cómo sean las medidas de similaridad.

- Métrico: Asume que la medidas de similaridad son de intervalo o de razón. Este seria el caso, por ejemplo, de una matriz donde los estímulos son ciudades, y la medida de proximidad, la distancia en kilémetros.
- No métrico: Donde el nivel de medida de la variables es ordinal.

Escalamiento multidimensional clásico

En la función `mds{smacof}` que es la que estima este tipo de escalamiento, la opción por uno u otro instrumento de medida se elige con el modificador `type=c("ratio","interval","ordinal","mspline")`.

Desarrollo educativo de distintas zonas del mundo

Etiqueta	Descripción	I1	I2	I3	I4	I5	I6	I7
Z1	América del Norte	213	68,1	97,0	84,0	5,5	23	148,0
Z2	Asia/Oceanía	520	53,3	107,8	45,3	4,0	21	108,2
Z3	Europa	250	77,3	111,4	47,8	5,4	24	82,2
Z4	Países en transición	114	54,0	86,9	34,2	5,2	27	6,1
Z5	África subsahariana	12	9,2	24,3	3,5	5,6	10	1,5
Z6	Estados árabes	37	15,4	53,7	12,5	5,2	20	2,7
Z7	América Latina y Caribe	80	51,1	56,6	17,3	4,5	22	10,7
Z8	Asia Oriental y Oceanía	57	28,9	61,5	8,9	3,0	14	7,3
Z9	China	43	28,7	66,6	5,3	2,3	13	5,5
Z10	Asia Meridional	27	9,8	44,5	6,5	4,3	9	1,9
Z11	India	32	4,8	48,7	6,4	3,5	9	1,9
Z12	Países menos adelantados	7	10,8	18,4	3,2	2,5	8	0,4

Supongamos que un investigador en economía de la educación desea saber cuál es la posición relativa de distintas zonas del mundo respecto al nivel de desarrollo educativo. En la tabla se recogen los principales indicadores que este investigador maneja para esta finalidad. Es obvio que podría recurrirse a un análisis de conglomerados para conseguir un número de grupos territoriales homogéneos, pero estos grupos estarán próximos o alejados entre sí. Estas distancias, para el investigador, son relevantes porque le muestran las diferencias regionales y el camino que queda por recorrer.

Desarrollo educativo de distintas zonas del mundo

Etiqueta	Descripción	I1	I2	I3	I4	I5	I6	I7
Z1	América del Norte	213	68,1	97,0	84,0	5,5	23	148,0
Z2	Asia/Oceanía	520	53,3	107,8	45,3	4,0	21	108,2
Z3	Europa	250	77,3	111,4	47,8	5,4	24	82,2
Z4	Países en transición	114	54,0	86,9	34,2	5,2	27	6,1
Z5	África subsahariana	12	9,2	24,3	3,5	5,6	10	1,5
Z6	Estados árabes	37	15,4	53,7	12,5	5,2	20	2,7
Z7	América Latina y Caribe	80	51,1	56,6	17,3	4,5	22	10,7
Z8	Asia Oriental y Oceanía	57	28,9	61,5	8,9	3,0	14	7,3
Z9	China	43	28,7	66,6	5,3	2,3	13	5,5
Z10	Asia Meridional	27	9,8	44,5	6,5	4,3	9	1,9
Z11	India	32	4,8	48,7	6,4	3,5	9	1,9
Z12	Países menos adelantados	7	10,8	18,4	3,2	2,5	8	0,4

Los indicadores se corresponden con la siguiente leyenda:

- I1. Tirada de diarios. Número de ejemplares por mil habitantes.
- I2. Tasa bruta de escolarización en la enseñanza preprimaria.
- I3. Tasa bruta de escolarización en la enseñanza secundaria.
- I4. Tasa bruta de escolarización en la enseñanza superior.
- I5. Gasto público en educación (% del PIB).
- I6. Número de docentes de todos los niveles educativos (miles) por cada mil habitantes de 15 a 64 años.
- I7. Consumo de papel de imprenta (kg por habitante).

Desarrollo educativo de distintas zonas del mundo

Etiqueta	Descripción	I1	I2	I3	I4	I5	I6	I7
Z1	América del Norte	213	68,1	97,0	84,0	5,5	23	148,0
Z2	Asia/Oceanía	520	53,3	107,8	45,3	4,0	21	108,2
Z3	Europa	250	77,3	111,4	47,8	5,4	24	82,2
Z4	Países en transición	114	54,0	86,9	34,2	5,2	27	6,1
Z5	África subsahariana	12	9,2	24,3	3,5	5,6	10	1,5
Z6	Estados árabes	37	15,4	53,7	12,5	5,2	20	2,7
Z7	América Latina y Caribe	80	51,1	56,6	17,3	4,5	22	10,7
Z8	Asia Oriental y Oceanía	57	28,9	61,5	8,9	3,0	14	7,3
Z9	China	43	28,7	66,6	5,3	2,3	13	5,5
Z10	Asia Meridional	27	9,8	44,5	6,5	4,3	9	1,9
Z11	India	32	4,8	48,7	6,4	3,5	9	1,9
Z12	Países menos adelantados	7	10,8	18,4	3,2	2,5	8	0,4

Las zonas geográficas que requieren alguna aclaración respecto a los países que las componen serán las siguientes:

- Asia / Oceanía. Exclusivamente Australia, Israel, Japón y Nueva Zelanda.
- Europa: no comprende la Europa del Este, que se incluye en países en transición.
- Países en transición: Albania, Armenia, Azerbaiyán, Eslovaquia, Federación Rusa, República Checa, Bulgaria, Hungría, Yugoslavia, entre otros.
- África subsahariana: Angola, Benín, Botsuana, Burkina Faso, Gambia, Guinea, Kenia, Sudáfrica, Sudán, entre otros.
- Estados árabes: Arabia Saudí, Argelia, Bahréin, Yibuti, Egipto, Emiratos Árabes, Irak, Jordania, Kuwait, Líbano, Marruecos, Siria, entre otros.
- Asia Oriental y Oceanía: Brunéi, Camboya, Fiyi, Filipinas, Hong Kong, Indonesia, Macao, Malasia, Corea del Norte, Corea del Sur, Singapur, Tailandia, entre otros. Por su importancia relativa, China se considera aparte.
- Asia Meridional: Afganistán, Bangladés, Bután, Irán, Maldivas, Nepal, Pakistán y Sri Lanka. Por su importancia relativa, India se considera aparte.

Desarrollo educativo de distintas zonas del mundo

Etiqueta	Descripción	I1	I2	I3	I4	I5	I6	I7
Z1	América del Norte	213	68,1	97,0	84,0	5,5	23	148,0
Z2	Asia/Oceanía	520	53,3	107,8	45,3	4,0	21	108,2
Z3	Europa	250	77,3	111,4	47,8	5,4	24	82,2
Z4	Países en transición	114	54,0	86,9	34,2	5,2	27	6,1
Z5	África subsahariana	12	9,2	24,3	3,5	5,6	10	1,5
Z6	Estados árabes	37	15,4	53,7	12,5	5,2	20	2,7
Z7	América Latina y Caribe	80	51,1	56,6	17,3	4,5	22	10,7
Z8	Asia Oriental y Oceanía	57	28,9	61,5	8,9	3,0	14	7,3
Z9	China	43	28,7	66,6	5,3	2,3	13	5,5
Z10	Asia Meridional	27	9,8	44,5	6,5	4,3	9	1,9
Z11	India	32	4,8	48,7	6,4	3,5	9	1,9
Z12	Países menos adelantados	7	10,8	18,4	3,2	2,5	8	0,4

- Países menos adelantados. Se consideran, de todos los anteriores, aquellos países con un atraso mayor. Este punto servirá en el MDS de referencia de nivel. Incluye Afganistán, Angola, Bangladés, Benín, Bután, Etiopía, Gambia, Guinea, Haití, Nepal, Níger, Congo, Tanzania, Yemen, Zambia, entre otros.

Desarrollo educativo de distintas zonas del mundo

```
datos <- read.table("educa.txt", header=F)
rownames(datos)<- c("Z1", "Z2", "Z3","Z4", "Z5", "Z6", "Z7", "Z8", "Z9", "Z10","Z11","Z12")
colnames(datos) <- c("I1","I2","I3","I4","I5","I6","I7")
# Normalizamos los indicadores
datos_norm <- scale(datos)
# Calculamos la matriz de distancias
datosf <-  dist(datos_norm,method = "euclidian",diag=T,upper=T)
m <- as.matrix(datosf)
datos <- as.dist(m)
```

Como puede comprobarse, los datos de la tabla anterior no son una medida de proximidad entre los estímulos (las 12 zonas). Para obtener esta matriz calcularíamos las distancias euclídeas entre los estímulos teniendo en cuenta las variables que los caracterizan (I1 a I7). Lógicamente las variables son previamente estandarizadas, como se observa en la sintaxis.

Desarrollo educativo de distintas zonas del mundo

```
##
## Call:
## mds(delta = datos, ndim = 2, type = "ratio")
##
## Model: Symmetric SMACOF
## Number of objects: 12
## Stress-1 value: 0.071
## Number of iterations: 64
```

```
## [1] 0.07110678
```

```
library(smacof)
fit <- mds(delta=datos,ndim=2,type="ratio")
```

```
fit
# Stress
print(fit$stress)
```

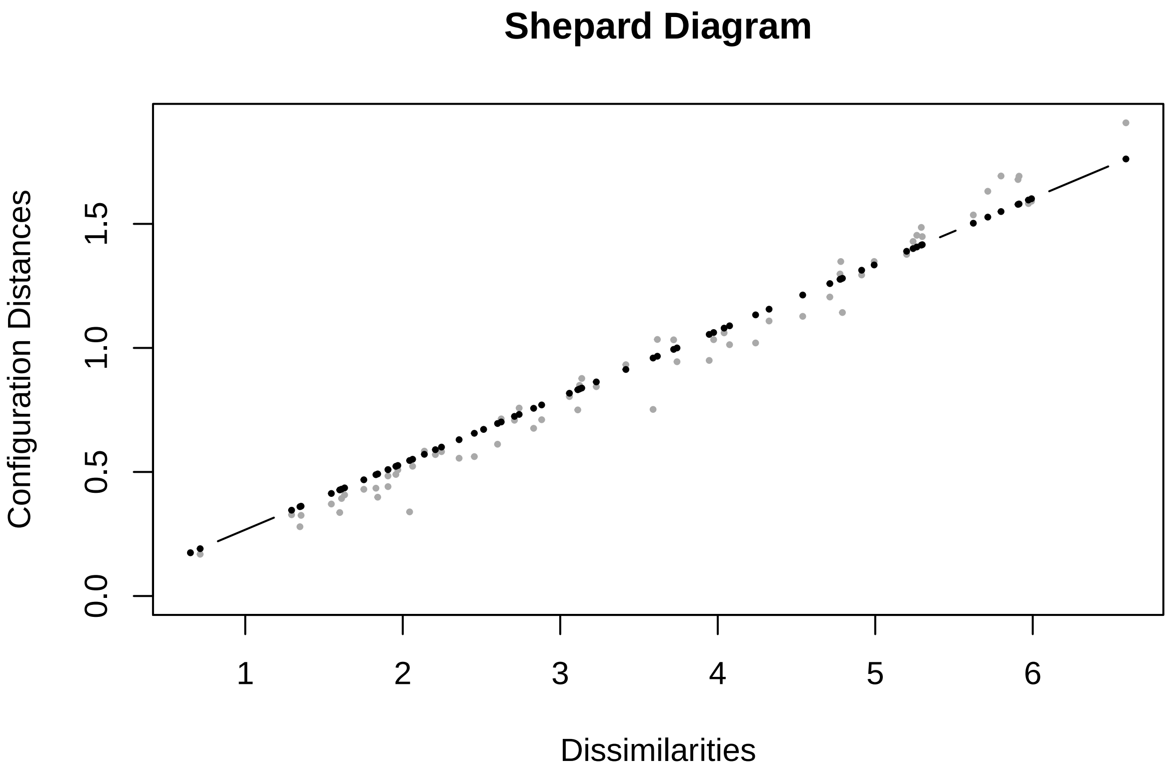
Puede verse que el algoritmo convergió tras 64 iteraciones ofreciendo un stress de 0.071, cifra que puede ser considerada entre "bueno".

Desarrollo educativo de distintas zonas del mundo

```
##
## Call:
## lm(formula = dist ~ dism)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09232 -0.03865 -0.01007  0.03535  0.17497
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09872    0.01546   6.385  2.2e-08 ***
## dism         0.91158    0.01550  58.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05644 on 64 degrees of freedom
## Multiple R-squared:  0.9818, Adjusted R-squared:  0.9815
## F-statistic:  3459 on 1 and 64 DF,  p-value: < 2.2e-16
```

```
dist <- cbind(c(fit$dhat))
dism <- cbind(c(fit$confdist))
summary(lm(dist~dism))
```

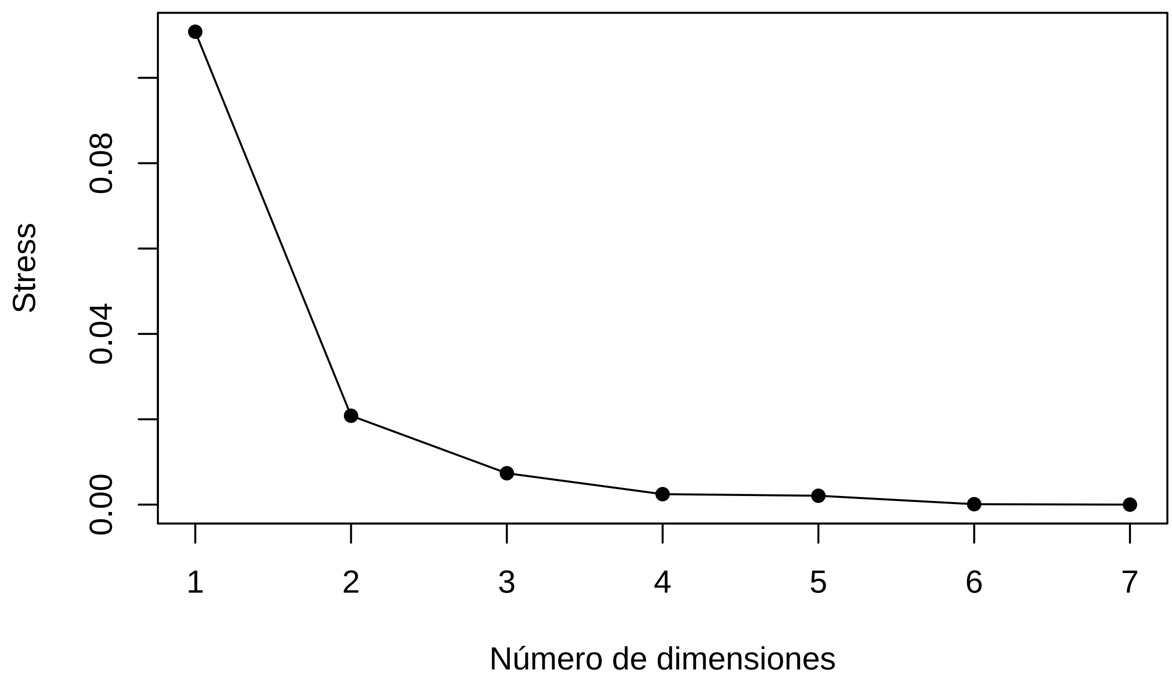
Asimismo el coeficiente de determinación (RSQ) está muy cercano a 1 (0.9818), demostrando así que el ajuste entre disparidades y distancias es casi perfecto.




```
plot(fit,plot.type="Shepard",plot.dim=c(1,2),sphere=TRUE,bubscale=0.1,col=1,
     label.conf=list(label=TRUE,pos=3,col=1,cex=0.8),
     shepard.x=NULL,identify=FALSE,
     type="p",pch=20,asp=1,col.hist=NULL)
```

El ajuste casi perfecto entre disparidades y distancias se constata también gráficamente usando el gráfico de Shepard.

Desarrollo educativo de distintas zonas del mundo



Para corroborar que esta solución bidimensional es razonable. De no ser así, si fuera necesario recurrir a cuatro o cinco dimensiones, el MDS no ofrecería demasiada ayuda respecto al análisis de conglomerados, puesto que el análisis gráfico de la solución sería complicado. Para corroborar ensayamos distintas soluciones dimensionales (digamos de 1 a 7) y las representamos en un gráfico las dimensiones en abscisas y el stress en ordenadas. Si la solución bidimensional es válida, el stress debe caer rápidamente hasta alcanzar esa dimensión, disminuyendo mucho menos en dimensiones adicionales. La figura obtenida confirma que la solución bidimensional es la adecuada.

Desarrollo educativo de distintas zonas del mundo

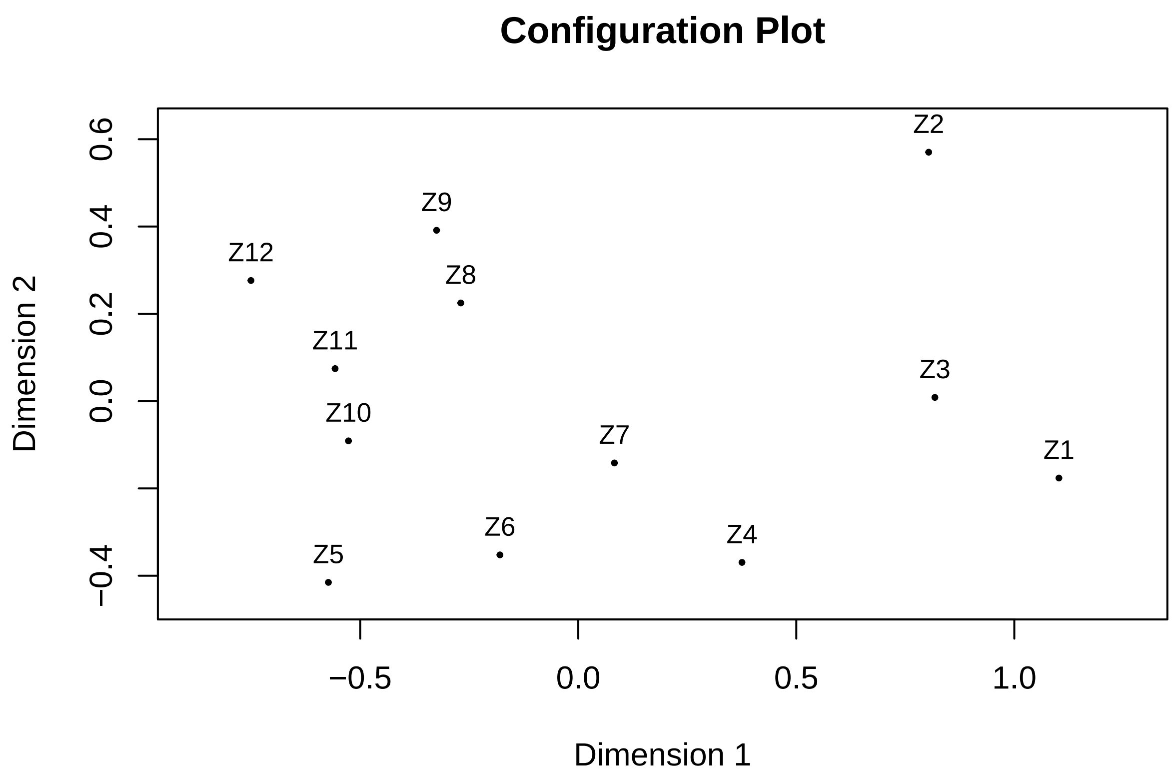
```
svec <- NULL
for (i in 1:7) {
  svec[i] <- mds(delta=datos, ndim = i, type = "ordinal")$stress
}

plot(seq(1:7),svec,type="overplotted", pch=16, ylab="Stress",xlab="Número de dimensiones")
```

Desarrollo educativo de distintas zonas del mundo

Este código de R permite graficar la representación bidimensional de las regiones usando el MSD.

```
plot(fit,plot.type="confplot",plot.dim=c(1,2),sphere=TRUE,bubscale=0.1,col=1,
     label.conf=list(label=TRUE,pos=3,col=1,cex=0.8),
     shepard.x=NULL,identify=FALSE,
     type="p",pch=20,asp=1,col.hist=NULL)
```

Desarrollo educativo de distintas zonas del mundo



La figura nos muestra que existen, al menos, tres grupos de regiones con un desarrollo educativo muy diferenciado. Un grupo lo formarían América del Norte (Z1), Asia / Oceanía (Z2) y Europa (Z3). Este grupo vendría caracterizado por contener a los países con mayor desarrollo de sus sistemas educativos. Téngase en cuenta que Z2 incluye solo a aquellos países de Asia y Oceanía más avanzados, como son Japón, Nueva Zelanda o Australia. Interpretamos que estos países son los más alejados no solo por el sentido común, difícil de aplicar en otro tipo de análisis, sino porque son los más alejados de Z12, que, recordemos, incluía como referencia a aquellos países menos adelantados.

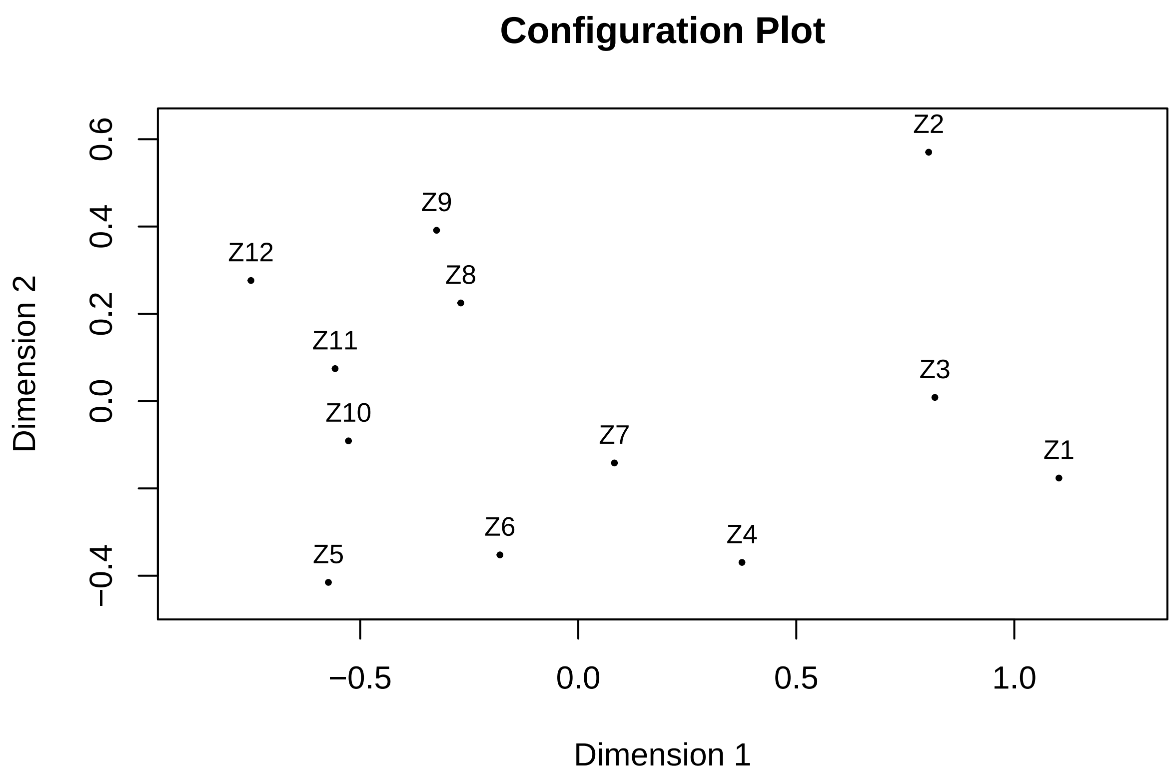
Desarrollo educativo de distintas zonas del mundo

Etiqueta	Descripción	I1	I2	I3	I4	I5	I6	I7
Z1	América del Norte	213	68,1	97,0	84,0	5,5	23	148,0
Z2	Asia/Oceanía	520	53,3	107,8	45,3	4,0	21	108,2
Z3	Europa	250	77,3	111,4	47,8	5,4	24	82,2
Z4	Países en transición	114	54,0	86,9	34,2	5,2	27	6,1
Z5	África subsahariana	12	9,2	24,3	3,5	5,6	10	1,5
Z6	Estados árabes	37	15,4	53,7	12,5	5,2	20	2,7
Z7	América Latina y Caribe	80	51,1	56,6	17,3	4,5	22	10,7
Z8	Asia Oriental y Oceanía	57	28,9	61,5	8,9	3,0	14	7,3
Z9	China	43	28,7	66,6	5,3	2,3	13	5,5
Z10	Asia Meridional	27	9,8	44,5	6,5	4,3	9	1,9
Z11	India	32	4,8	48,7	6,4	3,5	9	1,9
Z12	Países menos adelantados	7	10,8	18,4	3,2	2,5	8	0,4

Las zonas geográficas que requieren alguna aclaración respecto a los países que las componen serán las siguientes:

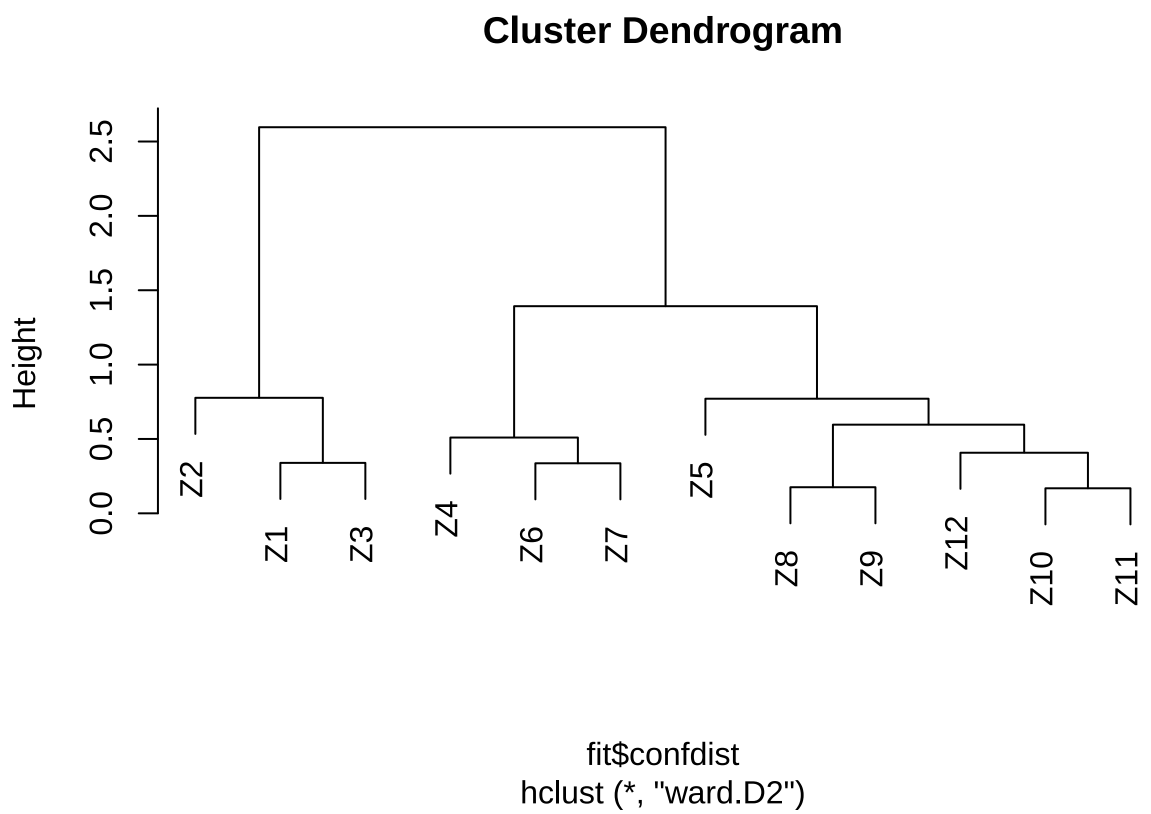
- Asia / Oceanía. Exclusivamente Australia, Israel, Japón y Nueva Zelanda.
- Europa: no comprende la Europa del Este, que se incluye en países en transición.
- Países en transición: Albania, Armenia, Azerbaiyán, Eslovaquia, Federación Rusa, República Checa, Bulgaria, Hungría, Yugoslavia, entre otros.
- África subsahariana: Angola, Benín, Botsuana, Burkina Faso, Gambia, Guinea, Kenia, Sudáfrica, Sudán, entre otros.
- Estados árabes: Arabia Saudí, Argelia, Bahréin, Yibuti, Egipto, Emiratos Árabes, Irak, Jordania, Kuwait, Líbano, Marruecos, Siria, entre otros.
- Asia Oriental y Oceanía: Brunéi, Camboya, Fiyi, Filipinas, Hong Kong, Indonesia, Macao, Malasia, Corea del Norte, Corea del Sur, Singapur, Tailandia, entre otros. Por su importancia relativa, China se considera aparte.
- Asia Meridional: Afganistán, Bangladés, Bután, Irán, Maldivas, Nepal, Pakistán y Sri Lanka. Por su importancia relativa, India se considera aparte.

Desarrollo educativo de distintas zonas del mundo



Un grupo intermedio vendría formado por los países euroasiáticos en transición (Z4), América Latina y el Caribe (Z7) y los Estados árabes (Z6), que tendrían un desarrollo educativo intermedio entre los más avanzados y los de menor desarrollo. Estos últimos serían los recogidos en el mapa más cerca de Z12, es decir, Africa subsahariana, Asia Oriental y Oceanía, Asia Meridional e India.

Desarrollo educativo de distintas zonas del mundo



Por qué considerar tres agregados y no más (o menos) en la interpretación. Para ayudar en este tipo de análisis de los gráficos de MDS, Kruskal y Wish (1978) recomiendan complementar esta técnica con otras ya expuestas, como el análisis de conglomerados. Atendiendo a ese dendograma, se han formado los grupos que, como se puede comprobar, sirvió de base para la interpretación que expusimos con anterioridad.

Ejemplo sobre psicología clínica

El conjunto de datos proviene del área de psicología clínica. McNally et al. (2015) recopilaron datos sobre los síntomas del TEPT (trastorno de estrés postraumático) informados por los sobrevivientes del terremoto de Wenchuan en China utilizando la lista de control civil (TEP-C; Weathers et al., 1993). En total, hay 17 ítems de síntomas de trastorno de estrés postraumático escalados en una escala de calificación de 5 puntos (1 … "para nada"; 5 … "extremadamente"). Estamos interesados en representar asociaciones entre el TEPT. los síntomas.

Ejemplo sobre psicología clínica

```
library("MPsychoR")
data(Wenchuan)
head(Wenchuan)
```

```
##   intrusion dreams flash upset physior avoidth avoidact amnesia lossint
## 1         2      2     2     2       3       2        3       2       3
## 2         2      2     2     3       3       3        3       2       3
## 3         2      4     4     4       3       3        3       5       4
## 4         2      1     2     2       1       1        2       2       2
## 5         2      2     2     2       2       2        2       2       3
```

```
## 6         4      3     2     2       2       2        3       3       2
##   distant numb future sleep anger concen hyper startle
## 1       2    2      1     3     4      3     4       2
## 2       3    2      2     3     3      2     3       3
## 3       3    2      3     4     4      4     3       4
## 4       1    1      2     2     1      2     3       3
## 5       2    2      2     3     2      3     2       3
## 6       2    2      3     2     3      2     3       2
```

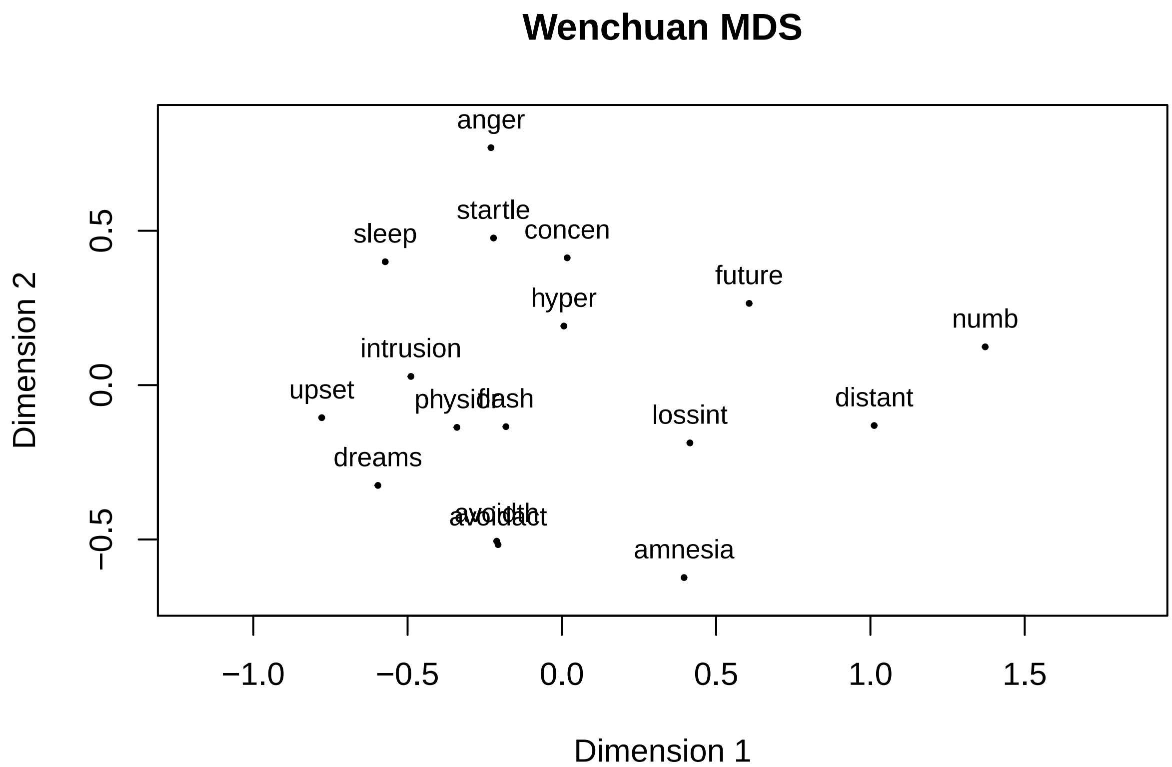
Ejemplo sobre psicología clínica

```
library("smacof")
Wdelta <- dist(t(Wenchuan)) ## Euclidean distances
fit.wenchuan1 <- mds(Wdelta, type = "ordinal") ## MDS fit
fit.wenchuan1
```

```
##
## Call:
## mds(delta = Wdelta, type = "ordinal")
##
## Model: Symmetric SMACOF
## Number of objects: 17
## Stress-1 value: 0.133
## Number of iterations: 35
```

Primero, calculamos las similitudes derivadas utilizando la distancia euclidiana, lo que resulta en una matriz de disimilitud simétrica de 17×17. Por un momento, consideremos un nivel de escala ordinal para las diferencias de entrada y ajustemos un MDS ordinal bidimensional. El ajuste de MDS da como resultado un valor de stress de 0.133. El valor del stress en este ejemplo sugiere un ajuste regular del modelo.

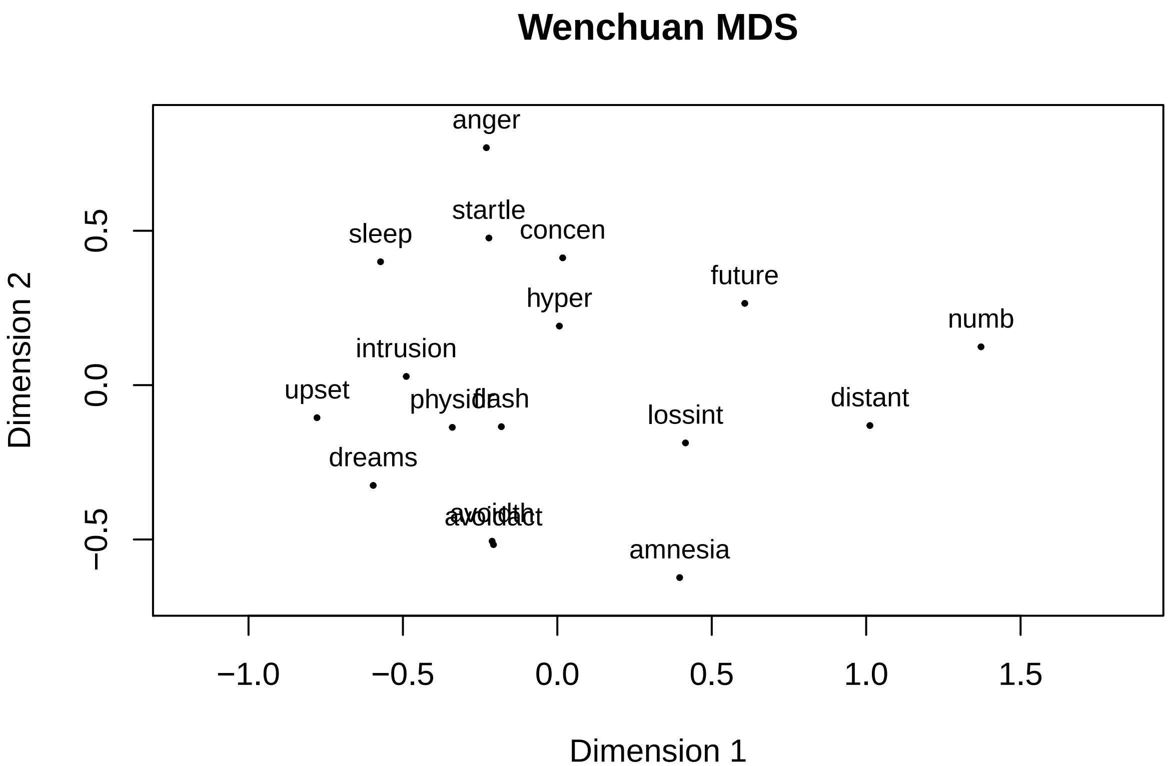
Ejemplo sobre psicología clínica




```
plot(fit.wenchuan1, main = "Wenchuan MDS")
```

El gráfico muestra cómo los síntomas de TEPT están relacionados entre sí. Por ejemplo, vemos que “evitar pensar o hablar sobre una experiencia estresante del pasado o evitar tener sentimientos relacionados con ella” (avoidth) y “evitar actividades o situaciones porque le recordaron una experiencia estresante del pasado”(avoidact) están virtualmente en el mismo lugar.

Ejemplo sobre psicología clínica



También vemos que "sentirse muy molesto cuando algo te recordó una experiencia estresante del pasado?" (upset) y "sentirse emocionalmente adormecido o ser incapaz de tener sentimientos amorosos para las personas cercanas a usted" (numb) son los puntos más lejanos en la primera dimensión, mientras que "problemas para recordar partes importantes de una experiencia estresante del pasado" (amnesia) y "sentirse irritable o tener arrebatos de ira"(anger) son los puntos extremos en la segunda dimensión. Estos pares de síntomas no están relacionados entre sí.

Ejemplo sobre psicología clínica

Un error común es interpretar el stress de manera demasiado mecánica al confiar solo en la tabla presentada anteriormente. Esto es problemático debido a que la magnitud del stress depende de la cantidad de objetos n (cuanto más grande es n, mayor es el stress). En las aplicaciones MDS modernas, n puede ser bastante grande. En lugar de utilizar estas reglas generales, podemos considerar enfoques de simulación. Un enfoque antiguo (Spence y Ogilvie, 1973) es simular diferencias aleatorias para n y p fijas y ajustar el modelo MDS correspondiente en estas matrices. Esto conduce a `random stress norms`.

Ejemplo sobre psicología clínica

```
set.seed(123)
rsvec <- randomstress(n = attr(Wdelta, "Size"), ndim = 2,
                      nrep = 500, type = "ordinal")
mean(rsvec)
mean(rsvec) - 2*sd(rsvec)
```

```
## [1] 0.2770934
```

```
## [1] 0.2548536
```

El llamado `randomstress` da 500 valores de stress aleatorios. Estos estándares de stress representan un punto de referencia de "mal ajuste". Nuestra solución debe estar claramente por debajo del stress aleatorio promedio. A menudo, en la literatura, los valores de stress observados se consideran "significativos" si son más pequeños que el límite de estrés aleatorio inferior $2 \times$ sd, que es claramente el caso en nuestro ejemplo (el estrés fue 0.133).

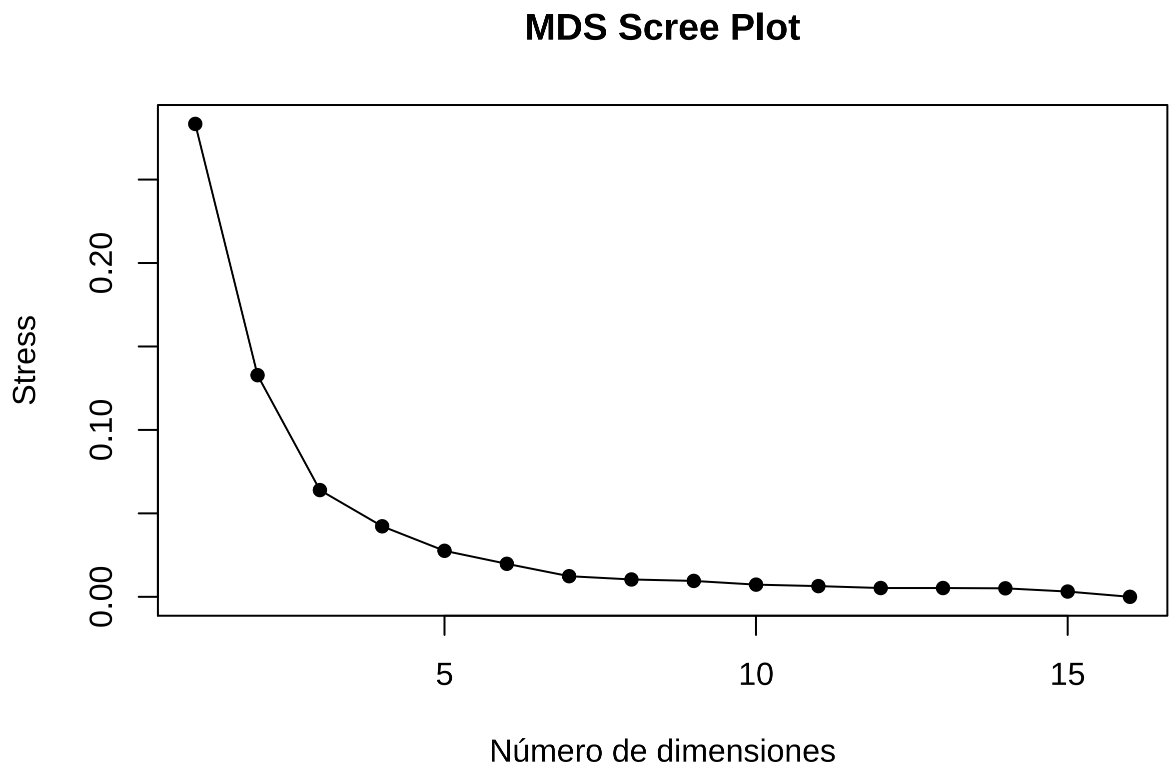
Ejemplo sobre psicología clínica

```
set.seed(123)
permmds <- permtest(fit.wenchuan1, data = Wenchuan,
                    method.dat = "euclidean", nrep = 500,
                    verbose = FALSE)
permmds
```

```
##
## Call: permtest.smacof(object = fit.wenchuan1, data = Wenchuan, method.dat = "euclidean",
##     nrep = 500, verbose = FALSE)
##
## SMACOF Permutation Test
## Number of objects: 17
## Number of replications (permutations): 500
##
## Observed stress value: 0.133
## p-value: <0.001
```

Un enfoque más preciso es utilizar una prueba de permutación como se describe en Mair et al. (2016). Para las diferencias similares, vuelve a muestrear los datos originales, y para cada matriz de desigualdad resultante, se lleva a cabo un ajuste MDS. Esto proporciona una distribución del estrés bajo el H_0: "el stress / configuración se obtiene de una permutación aleatoria de diferencias". Como $p < 0.05$, rechazamos el H_0 y concluimos que el stress / configuración se obtiene de algo distinto a una permutación aleatoria de diferencias.

Ejemplo sobre psicología clínica




```
n <- attr(Wdelta, "Size")
svec <- NULL
for (i in 1:(n-1)) {
  svec[i] <- mds(Wdelta, ndim = i, type = "ordinal")$stress
}

plot(seq(1:(n-1)),svec,type="overplotted", pch=16, ylab="Stress",
xlab="Número de dimensiones", main="MDS Scree Plot")
```

Basándonos únicamente en el scree plot, probablemente elegiríamos una solución 3D, pero el stress de la solución 2D tampoco es demasiado malo, como lo juzgan las `random stress norms`, los resultados de las pruebas de permutación y la tabla de referencia de los valores de stress (considerados aquí como n = 17 es bastante pequeño).

Ejemplo sobre psicología clínica

Se sabe que la función de objetivo de estrés es irregular. Por lo tanto, puede suceder fácilmente que terminemos en un mínimo local (es decir, no obtenemos la mejor solución posible). Donde el algoritmo termina al final depende de donde comienza. De forma predeterminada, las funciones en el paquete `smacof` utilizan una solución de escala clásica (Torgerson, 1952) como configuración inicial. Esta no es necesariamente la mejor opción. El problema mínimo local que incluye una búsqueda sistemática de soluciones iniciales se describe en detalle en Borg y Mair (2017). Aquí utilizamos una estrategia ad hoc simple al probar diferentes inicios aleatorios y verificar si la mejor solución de inicio aleatorio conduce a un valor de estrés más bajo que la configuración predeterminada.

Ejemplo sobre psicología clínica

```
## [1] 0.1327943
```

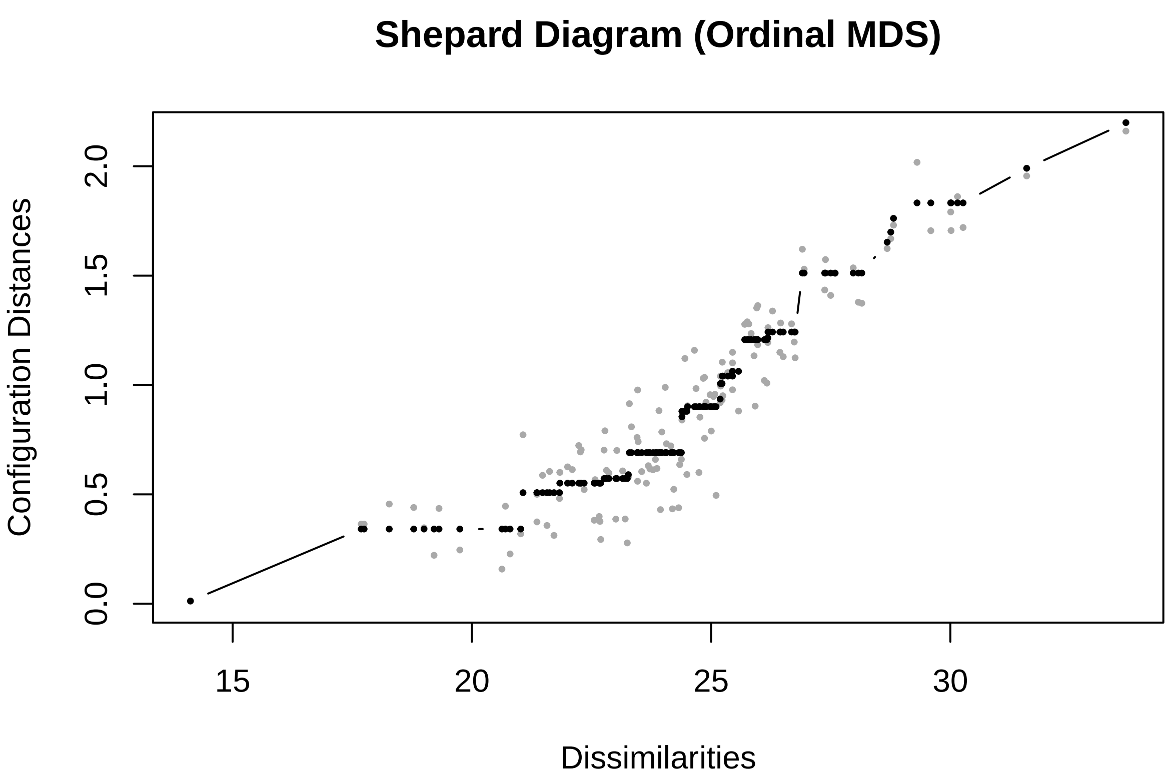
```
## [1] 0.1328058
```

A continuación examinamos 100 inicios aleatorios y extraemos los valores de estrés. Vemos que un inicio aleatorio en particular proporcionó un stress ligeramente mejor que nuestra solución original. Dado que la diferencia de estrés es tan mínima, podemos elegir cualquiera de las dos soluciones.

Ejemplo sobre psicología clínica

```
set.seed(123)
fit.wenchuan <- NULL ## 100 random starts
for(i in 1:100) {
  fit.wenchuan[[i]] <- mds(Wdelta, type = "ordinal",
                           init = "random")
}
## extract the best solution
ind <- which.min(sapply(fit.wenchuan,
                        function(x) x$stress))
fit.wenchuan2 <- fit.wenchuan[[ind]]
fit.wenchuan2$stress ## lowest stress (random start)

fit.wenchuan1$stress ## stress (classical scaling start)
```

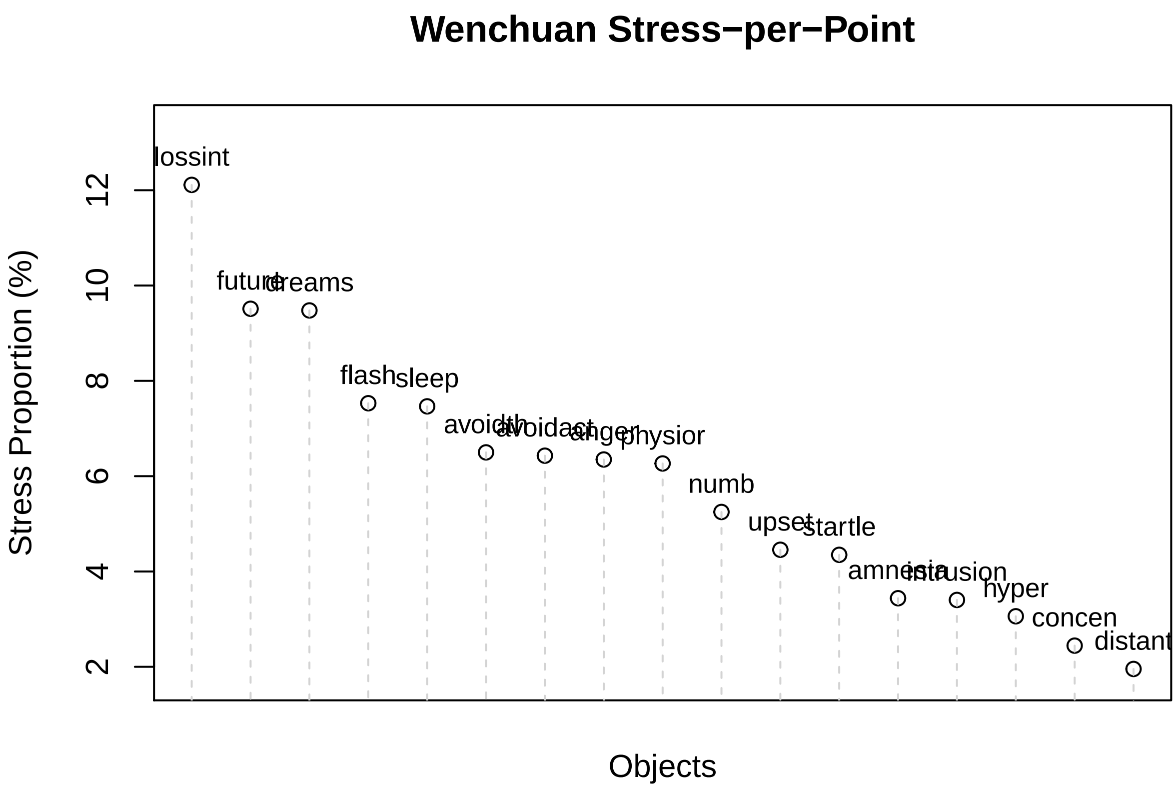
Ejemplo sobre psicología clínica



```
plot(fit.wenchuan2, plot.type = "Shepard",
     main = "Shepard Diagram (Ordinal MDS)")
```

Vemos varios puntos grises alejados de la recta, lo cual explica porque el valor de stress no está en el rango de bueno a excelente.

Ejemplo sobre psicología clínica




```
plot(fit.wenchuan2, plot.type = "stressplot",
     main = "Wenchuan Stress-per-Point")
```

La contribución de cada objeto al stress total se puede calcular fácilmente. Los valores resultantes, que normalmente se informan como porcentajes, se denominan stress por punto (SPP). Podemos pensar en puntos con una alta contribución de SPP de manera similar a los valores atípicos influyentes en la regresión. En nuestro ejemplo, vemos que la "pérdida de interés en actividades que solía disfrutar" (lossint) proporciona una contribución de estrés bastante alta del 12.11%.