PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Artur Zaborski
Uniwersytet Ekonomiczny we Wrocławiu

WYKORZYSTANIE METODY MAJORYZACJI FUNKCJI DOPASOWANIA W MODELACH RÓŻNIC INDYWIDUALNYCH

Streszczenie: Majoryzacja jest metodą o charakterze iteracyjnym aproksymującą minimalne wartości funkcji STRESS. Celem artykułu jest prezentacja metodologii skalowania różnic indywidualnych za pomocą metody majoryzacji. Podejście to nosi nazwę SMACOF i jest realizowane w środowisku R. Na zakończenie zaprezentowano przykład, w którym wykorzystano funkcję `smacofIndDiff` pakietu `smacof`.

1. Wstęp

Skalowanie wielowymiarowe jest zbiorem technik mających na celu przedstawienie (zazwyczaj w przestrzeni dwuwymiarowej) relacji zachodzących między obiektami traktowanymi jako punkty w przestrzeni wielowymiarowej. Na podstawie macierzy niepodobieństw między obiektami i oraz j ($i, j = 1, 2, \ldots, n$) w przestrzeni wielowymiarowej $\mathbf{\Delta} = \left[\delta_{ij} \right]_{n \times n}$ dokonuje się, przy wykorzystaniu odpowiednich procedur, takiego rozmieszczenia na mapie percepcyjnej punktów $\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n \right]^T$, aby dopasowanie konfiguracji odległości w przestrzeni wielowymiarowej i dwuwymiarowej było możliwie najlepsze. Miarą dopasowania obydwu konfiguracji punktów jest funkcja STRESS (*STandardized REsidual Sum of Squares*). W najprostszej odmianie funkcja ta przyjmuje postać:

$$S = \sum_{i,j} \left(f(\delta_{ij}) - d_{ij}(\mathbf{X}) \right)^2 , \tag{1}$$

gdzie: $d_{ij}(\mathbf{X})$ – odległość między punktami $\mathbf{x}_i$ oraz $\mathbf{x}_j$,
$f(\delta_{ij})$ – funkcja regresji między d_{ij} a δ_{ij}.

Podstawowym zadaniem skalowania wielowymiarowego jest więc znalezienie takiej konfiguracji punktów w przestrzeni o zredukowanej liczbie wymiarów, która minimalizuje wartość funkcji dopasowania. Jedną z metod minimalizacji funkcji dopasowania jest metoda majoryzacji.

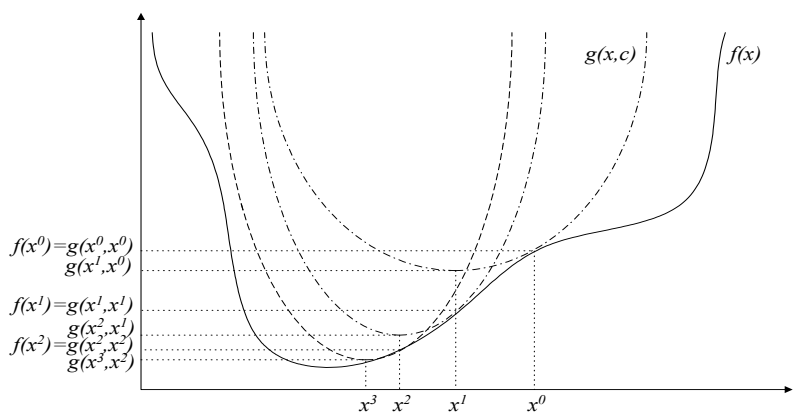
Artykuł jest prezentacją możliwości zastosowania metody majoryzacji w modelach różnic indywidualnych, tzn. gdy w badaniu zależności między obiektami wykorzystywanych jest wiele macierzy niepodobieństw.

2. Idea majoryzacji funkcji

Majoryzacja jest metodą wyznaczenia minimum funkcji $f(x)$ o skomplikowanej postaci analitycznej przez zastępowanie jej w kolejnych cyklach iteracyjnych przez pewną funkcję pomocniczą $g(x,c)$, gdzie c jest ustaloną stałą. Funkcję $g(x,c)$ nazywamy funkcją majoryzującą, jeżeli:

- minimalizacja funkcji $g(x,c)$ jest znacznie łatwiejsza niż minimalizacja funkcji $f(x)$ (np. $g(x,c)$ jest funkcją kwadratową zmiennej x);
- funkcja $f(x)$ zawsze przyjmuje wartości nie większe niż funkcja $g(x,c)$, czyli $\forall x\, f(x) \le g(x,c)$;
- wykres funkcji pomocniczej jest styczny do wykresu funkcji podstawowej w tzw. punkcie podparcia c, czyli $f(c) = g(c,c)$.

Zasadę metody majoryzacji obrazuje rys. 1.



Rys. 1. Trzy cykle iteracyjne metody majoryzacji dla funkcji jednej zmiennej

Źródło: opracowano na podstawie pracy [Groenen 1993, s. 5].

Pierwszy cykl iteracyjny rozpoczyna się od wyznaczenia funkcji pomocniczej $g(x,x^0)$, której wykres jest styczny do $f(x)$ w punkcie podparcia. Następnie wyznacza się punkt x^1 będący minimum funkcji $g(x,x^0)$. Jeżeli $\left|f(x^0) - f(x^1)\right| < \varepsilon$, gdzie ε jest ustaloną niewielką stałą nieujemną, zadanie zostaje zakończone. Jeżeli nie – przechodzi się do kolejnego cyklu, w którym punktem podparcia jest punkt x^1.

3. Charakterystyka modelu SMACOF

Wykorzystanie algorytmu majoryzacji do minimalizacji funkcji dopasowania w skalowaniu wielowymiarowym zaprezentowali De Leeuw i Heiser [1977] na przykładzie modelu SMACOF (*Scaling by MAjorizing a COmplicated Function*). W modelu tym funkcja dopasowania przyjmuje postać:

$$SS = \sum_{i<j} w_{ij} \left(\delta_{ij} - d_{ij}(\mathbf{X}) \right)^2 , \tag{2}$$

gdzie $w_{ij} = 1$, jeżeli δ_{ij} jest dane (w przeciwnym przypadku $w_{ij} = 0$).

Funkcją majoryzującą funkcję (2) jest (zob. [De Leeuw i Heiser 1977; Groenen 1993, s. 190; De Leeuw, Mair 2008]):

$$\tau(\mathbf{X}, \mathbf{Y}) = 1 + r\mathbf{X}^T\mathbf{V}\mathbf{X} - 2tr\mathbf{X}^T\mathbf{B}(\mathbf{Y})\mathbf{Y} , \tag{3}$$

gdzie: $\mathbf{V}$ – macierz o elementach $v_{ij} = \begin{cases} \sum\limits_{\substack{j=1 \\ i \neq j}}^{n} w_{ij} & \text{dla } i = j \\ -w_{ij} & \text{dla } i \neq j \end{cases}$,

$\mathbf{B(Y)}$ – macierz o elementach $b_{ij} = \begin{cases} \dfrac{-w_{ij}\delta_{ij}}{d_{ij}(\mathbf{Y})} & \text{dla } i \neq j \text{ i } d_{ij}(\mathbf{Y}) \neq 0 \\ 0 & \text{dla } i \neq j \text{ i } d_{ij}(\mathbf{Y}) = 0 \\ -\sum\limits_{\substack{j=1 \\ i \neq j}}^{n} b_{ij} & \text{dla } i = j \end{cases}$,

$\mathbf{Y}$ – konfiguracja punktów wyznaczonych w u-1 cyklu iteracyjnym.

Minimum funkcji (3) otrzymamy, przyrównując pochodne $\tau(\mathbf{X}, \mathbf{Y})$ do zera, tzn.

$$\nabla\tau(\mathbf{X}, \mathbf{Y}) = 2\mathbf{V}\mathbf{X} - 2\mathbf{B}(\mathbf{Y})\mathbf{Y} = \mathbf{0} , \tag{4}$$

gdzie $\nabla\tau(\mathbf{X}, \mathbf{Y})$ jest gradientem $\tau(\mathbf{X}, \mathbf{Y})$.

Równanie (4) jest prawdziwe, jeżeli $\mathbf{V}\mathbf{X} = \mathbf{B}(\mathbf{Y})\mathbf{Y}$. Ponieważ wyznacznik macierzy $\mathbf{V}$ jest równy zeru, do wyznaczenia macierzy $\mathbf{X}$ wykorzystuje się macierz odwrotną Moora-Penrose'a (zob. [Groenen 1993, s. 11]): $\mathbf{V}^+ = (\mathbf{V} + \mathbf{1}\mathbf{1}^T)^{-1} - n^{-1}\mathbf{1}\mathbf{1}^T$, gdzie $\mathbf{1}$ – wektor kolumnowy złożony z jedynek. Ostatecznie rozwiązaniem równania (4) jest macierz:

$$\mathbf{X}^u = \begin{cases} \mathbf{V}^+\mathbf{B}(\mathbf{Y})\mathbf{Y} \\ n^{-1}\mathbf{B}(\mathbf{Y})\mathbf{Y} \quad \text{gdy} \underset{\mathrm{i}\neq\mathrm{j}}{\forall} \; w_{ij} = 1 \end{cases}, \tag{5}$$

nazywana transformacją Guttmana.

SMACOF, który wykorzystuje metodę majoryzacji STRESS-u, jest modelem skalowania wielowymiarowego o charakterze iteracyjnym, przybliżającym w kolejnych cyklach obliczeniowych optymalne rozwiązanie. Algorytm tego modelu jest następujący:

1. Wyznaczenie konfiguracji początkowej punktów w przestrzeni *r*-wymiarowej reprezentujących obiekty $\mathbf{Y} = \mathbf{X}^0$.
2. Obliczenie wartości funkcji dopasowania STRESS dla konfiguracji początkowej $SS(\mathbf{X}^0)$.
3. Za pomocą transformacji Guttmana wyznaczenie $\mathbf{X}^u$, gdzie u – numer iteracji.
4. Obliczenie wartości funkcji dopasowania $SS(\mathbf{X}^u)$.
5. Jeżeli $|SS(\mathbf{X}^u) - SS(\mathbf{X}^{u-1})| < \varepsilon$ (ε – nieujemna stała) lub u jest równe maksymalnej, z góry ustalonej, liczbie iteracji – następuje koniec procesu obliczeniowego. W przeciwnym przypadku dokonuje się podstawienia $\mathbf{Y} = \mathbf{X}^u$ i przechodzi do kroku 3.

Podstawową zaletą procedury jest to, że w kolejnych cyklach iteracyjnych gwarantuje ciąg nierosnących wartości funkcji dopasowania.

4. Modele różnic indywidualnych

Przy badaniu relacji zachodzących między obiektami często korzysta się z opinii wielu respondentów. Ponieważ podstawą przeprowadzenia skalowania wielowymiarowego jest macierz $\mathbf{\Delta} = \left[\delta_{ij}\right]_{n\times n}$, której elementy δ_{ij} przedstawiają niepodobieństwa między obiektami i oraz j ($i, j = 1, 2, \ldots, n$, gdzie n jest liczbą obiektów), mamy do czynienia z sytuacją, kiedy danych jest m macierzy $\mathbf{\Delta}_k = \left[\delta_{ij,k}\right]$ ($k = 1, 2, \ldots, m$), przy czym każda macierz $\mathbf{\Delta}_k$ przedstawia różnice między obiektami postrzegane przez *k*-tego respondenta. Na podstawie m macierzy $\mathbf{\Delta}_k$ określa się:

– wspólną dla wszystkich respondentów konfigurację punktów $\mathrm{X} = \left[x_{ia}\right]$, gdzie $a = 1, 2, ..., r$, reprezentującą badane obiekty w *r*-wymiarowej przestrzeni, która nazywana jest grupową przestrzenią zmiennych (*group stimulus space*),
– przestrzeń wag. Punkty przestrzeni wag przedstawiają respondentów, a współrzędne tych punktów są wagami przypisywanymi poszczególnym wymiarom przez daną osobę,

– indywidualne konfiguracje punktów $\mathbf{X}_k$, które otrzymuje się, mnożąc współrzędne punktów z grupowej przestrzeni zmiennych przez odpowiednie wagi. Elementy macierzy $\mathbf{X}_k$ wynoszą:

$$\mathbf{X}_k = \mathbf{X}\mathbf{W}_k ,$$

gdzie $\mathbf{W}_k$ jest diagonalną macierzą wag o wymiarach $r \times r$ *k*-tego respondenta.

5. SMACOF dla modeli różnic indywidualnych

W modelu różnic indywidualnych na podstawie macierzy $\mathbf{\Delta}_k = \left[\delta_{ij,k} \right]$ poszukuje się takich konfiguracji $\mathbf{X}_k$ dla poszczególnych respondentów, aby wartość funkcji dopasowania STRESS:

$$\sigma^2(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m) = \sum_{k=1}^{m} \sum_{i<j} \left(\delta_{ijk} - d_{ij}(\mathbf{X}_k) \right)^2 \tag{6}$$

osiągała wartość najmniejszą, przy założeniu, że $\mathbf{X}_k = \mathbf{X}\mathbf{W}_k$.

Proces minimalizacji wartości STRESS-u przebiega przez dwa etapy (zob. [Borg, Groenen 2005, s. 475-476]), na które składa się:

1) wyznaczenie macierzy konfiguracji indywidualnych $\mathbf{X}^* = \left[\mathbf{X}_1 \quad \mathbf{X}_2 \quad \cdots \quad \mathbf{X}_m \right]^T$ na podstawie macierzy niepodobieństw $\mathbf{\Delta}_k$ $\left(k = 1, 2, \ldots, m \right)$,

2) wyznaczenie konfiguracji $\mathbf{X}$ oraz macierzy wag $\mathbf{W}_k$ minimalizujących wartość wyrażenia:

$$\sum_{k=1}^{m} tr\ n(\mathbf{X}\mathbf{W}_k - \mathbf{X}_k)^T (\mathbf{X}\mathbf{W}_k - \mathbf{X}_k) . \tag{7}$$

Macierze $\mathbf{X}_k$ $\left(k = 1, 2, \ldots, m \right)$ szacowane są za pomocą transformacji Guttmana. Po *u*-tym cyklu iteracyjnym $\mathbf{X}_k$ przyjmuje wartość (por. [Borg, Groenen 2005, s. 475-476]):

$$\mathbf{X}_k^u = \mathbf{V}_k^+ \mathbf{B}(\mathbf{Y}_k)\mathbf{Y}_k . \tag{8}$$

Dla ustalonych $\mathbf{X}_k$ minimalizacja wyrażenia (7) ze względu na $\mathbf{X}$ oraz wagi $\mathbf{W}_k$ jest równoznaczna z minimalizacją

$$\sum_{a=1}^{p} tr \left(\mathbf{x_a} \mathbf{w_a^T} - \mathbf{X_a} \right)^T \left(\mathbf{x_a} \mathbf{w_a^T} - \mathbf{X_a} \right), \tag{9}$$

gdzie: $\mathbf{x}_a$ – a-ta kolumna macierzy $\mathbf{X}$,

$\mathbf{w}_a$ – wektor wag w_{aak} a-tego wymiaru dla k-tego respondenta o wymiarach $m \times 1$,

$\mathbf{X}_a$ – macierz o wymiarach $n \times m$, w której elementy poszczególnych kolumn przedstawiają a-te współrzędne obiektów k-tego respondenta.

W tym celu, oddzielnie dla każdego wymiaru a, w kolejnych cyklach iteracyjnych wyrażenie (9) jest minimalizowane ze względu na $\mathbf{x}_a$, przy założeniu, że $\mathbf{w}_a$ jest ustalone, a następnie ze względu na $\mathbf{w}_a$, przy założeniu, że wyznaczone wcześniej $\mathbf{x}_a$ jest dane.

6. SMACOF dla modeli różnic indywidualnych w programie R

Skalowanie różnic indywidualnych wykorzystujące metodę majoryzacji funkcji dopasowania jest możliwe w programie R za pomocą funkcji smacofIndDiff pakietu smacof. Składnię funkcji oraz opis jej podstawowych argumentów prezentuje tab. 1.

Tabela 1. Opis funkcji `smacofIndDiff` w programie R

`smacofIndDiff(delta, ndim = 2, weightmat = NULL, init = NULL, metric = TRUE, constraint = NULL, verbose = FALSE, itmax = 1000, eps = 1e-6)`	
`delta`	lista macierzy niepodobieństw lub lista odległości między obiektami wyznaczone za pomocą funkcji `dist`
`ndim`	wymiar przestrzeni, w której prezentowane są wyniki skalowania
`weightmat`	macierz wag (opcjonalnie)
`init`	konfiguracja początkowa (opcjonalnie)
`metric`	FALSE oznacza niemetryczne skalowanie wielowymiarowe
`constrain`	NULL oznacza, że wyznaczana jest tylko konfiguracja wspólna; `diagonal` – wyznaczana jest również diagonalna macierz wag; `idioscal` – dopuszcza się dodatkowo rotację osi dla konfiguracji indywidualnych
`verbose`	TRUE oznacza, że podawane są wartości STRESS w kolejnych cyklach iteracyjnych
`itmax`	maksymalna liczba iteracji
`eps`	kryterium zbieżności

Źródło: opracowanie własne z wykorzystaniem dokumentacji programu R.

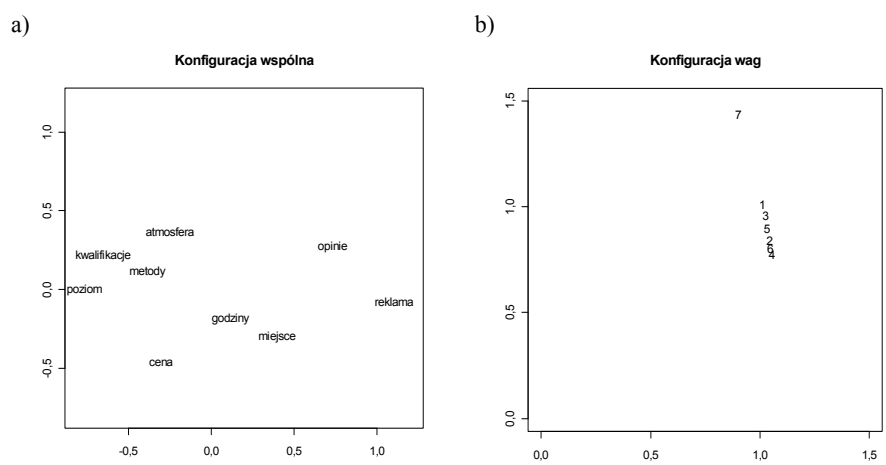
Przykład

Respondentom 7 względnie jednorodnych klas słuchaczy szkół językowych w Jeleniej Górze, wyodrębnionych za pomocą drzewa klasyfikacyjnego (zob. [Kurzydłowski, Zaborski 2005]), przedstawiono 9 czynników mających wpływ na wybór szkoły, z prośbą o określenie, zgodnie z własnymi preferencjami, istotności

tych czynników przez przyporządkowanie im liczb od 1 do 7. Liczba jeden oznaczała czynnik bardzo istotny, liczba siedem zaś – czynnik nieistotny przy wyborze szkoły. Na podstawie otrzymanych macierzy preferencji obliczono dla każdej klasy odległości między czynnikami za pomocą miary GDM dla zmiennych mierzonych na skali porządkowej, a następnie przeprowadzono skalowanie wielowymiarowe z wykorzystaniem funkcji `smacofIndDiff` z następującą składnią poleceń:

```
> library(smacof)
> options(OutDec=",")
> klasy <- list(k1, k2, k3, k4, k5, k6, k7)
>    klas_diag    <-    smacofIndDiff(klasy,    ndim=2,
metric=FALSE,
  constraint="diagonal")
> summary(klas_diag)
> wykr_klas <- klas_diag$gspace
> plot(wykr_klas, type="n", main="Konfiguracja wspól-
na",
  xlim=c(-0.8,   1.2),   ylim=c(-0.8,   1.2),   xlab="",
ylab="")
> text(wykr_klas, labels=rownames(wykr_klas))
> print(klas_diag$cweights)
> plot(klas_diag$cweights, type="n", main="Konfiguracja
wag",
  xlim=c(0, 1.5), ylim=c(0, 1.5), xlab="", ylab="").
```

W wyniku zastosowanej procedury otrzymano konfigurację wspólną czynników i konfigurację wag (rys. 2).



Rys. 2. Konfiguracje punktów reprezentujących: a) czynniki, b) poszczególne klasy respondentów

Źródło: opracowanie własne z wykorzystaniem funkcji `smacofIndDiff`.

```
Nonmetric stress: 0,08120585
Number of iterations: 247
Group Stimulus Space (Joint Configurations):
   D1      D2
1 -0,7640  0,0070
2 -0,6522  0,2186
3 -0,3857  0,1217
4 -0,2474  0,3732
5 -0,3004 -0,4576
6  0,3950 -0,2886
7  0,1126 -0,1784
8  0,7360  0,2758
9  1,1061 -0,0716
klas_diag$cweights
       [,1]      [,2]
[1,] 1,0111692 1,0179589
[2,] 1,0450749 0,8471151
[3,] 1,0241923 0,9650909
[4,] 1,0536151 0,7826529
[5,] 1,0328875 0,9047070
[6,] 1,0485146 0,8095302
   [7,] 0,8995695 1,4435935.
```

Dwuwymiarowa prezentacja konfiguracji wspólnej pozwala na stwierdzenie, że na decyzję o wyborze szkoły mają wpływ dwie grupy czynników odpowiadające wymiarom mapy percepcyjnej. Są to: jakość kształcenia (poziom kształcenia, kwalifikacje kadry i stosowane metody nauczania) oraz dostępność szkoły (cena zajęć, lokalizacja, godziny odbywania zajęć i atmosfera panująca w szkole). Zróżnicowanie preferencji jest przede wszystkim związane z przypisywaniem różnego znaczenia wymiarowi drugiemu (zob. rys. 2b). Najmniej istotny jest on dla słuchaczy, którzy przy wyborze szkoły kierowali się stosowanymi metodami nauczania (klasa IV), największy zaś dla tych, którzy dokonali wyboru szkoły pod wpływem reklamy.

7. Podsumowanie

Modele różnic indywidualnych są ważnym narzędziem geometrycznej prezentacji danych w badaniach zależności między obiektami opierających się na ocenach wielu respondentów. Dzięki nim można otrzymać zarówno grupową konfigurację, jak i konfiguracje indywidualne oraz przestrzeń wag, co umożliwia badanie relacji między obiektami w ramach różnych (indywidualnych) kryteriów ocen.

W artykule zaprezentowano procedurę modelu SMACOF, który w swojej konstrukcji wykorzystuje metodę majoryzacji funkcji dopasowania. Do podstawowych

walorów tego podejścia należy zaliczyć to, że gwarantuje ciąg nierosnących wartości funkcji dopasowania w kolejnych cyklach iteracyjnych, dopuszcza zerowe odległości między obiektami oraz jest powszechnie dostępny dzięki istniejącemu oprogramowaniu w środowisku R.

Literatura

Borg I., Groenen P., *Modern Multidimensional Scaling. Theory and Applications. Second Edition*, Springer-Verlag, New York 2005.

De Leeuw J., Heiser W.J., *Convergence of Correction-Matrix Algorithms for Multidimensional Scaling*, [w:] Recent Developments in Statistics, J.R. Barra, F. Brodeau, G. Romier, B. van Cutsem (red.), Amsterdam: North-Holland 1977, s. 133-145.

De Leeuw J., Mair P., *Multidimensional Scaling Using Majorization: SMACOF in R*, Department of Statistics, UCLA, Department of Statistics Papers, Paper 2008010903, http://repositories.cdlib.org/uclastat/papers/2008010903.

Groenen P.J.F., *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions*, Leiden: DSWO Press, Leiden University, 1993.

Kurzydłowski A., Zaborski A., *Ocena atrakcyjności szkół językowych z wykorzystaniem wybranych metod wielowymiarowej analizy statystycznej*, [w:] *Klasyfikacja danych – teoria i zastosowania*, Taksonomia 12, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław 2005, s. 453-461.

THE APPLICATION OF MAJORIZATION ALGORITHM FOR BADNESS-OF-FIT MEASURE IN INDIVIDUAL DIFFERENCES MODELS

Summary: Iterative majorization is a method of trying to get increasingly better estimates of STRESS function. The aim of the paper is to present the methodology of individual differences models scaling by means of the majorization algorithm. This strategy is called SMACOF and it is implemented in an R environment. Finally, an example is presented in which the smacofIndDiff function of smacof package is used.