`interscal`, `symscal`, `iscal` (`symbolicDA`)

Multidimensional scaling for symbolic interval-valued data

Symbolic multidimensional scaling aims to present relations between objects treated as hypercubes in multidimensional space. To allow interpretation and graphical representation of the results usually two-dimensional space is used.

Most of symbolic multidimensional scaling methods require interval dissimilarity matrix as input. This matrix can be obtained from n judges, opinions or from a dissimilarity measure for interval-valued variables that produces interval-valued dissimilarities (see: Lechevallier 2001).

Fig. 1 presents main two main approaches in symbolic multidimensional scaling, classical approach – also known as ”symbolique-numerique-symbolique” proposed by E. Diday in 1978 – and symbolic multidimensional scaling based on interval-valued distances. The fig. 1 presents also main methods for each approach.

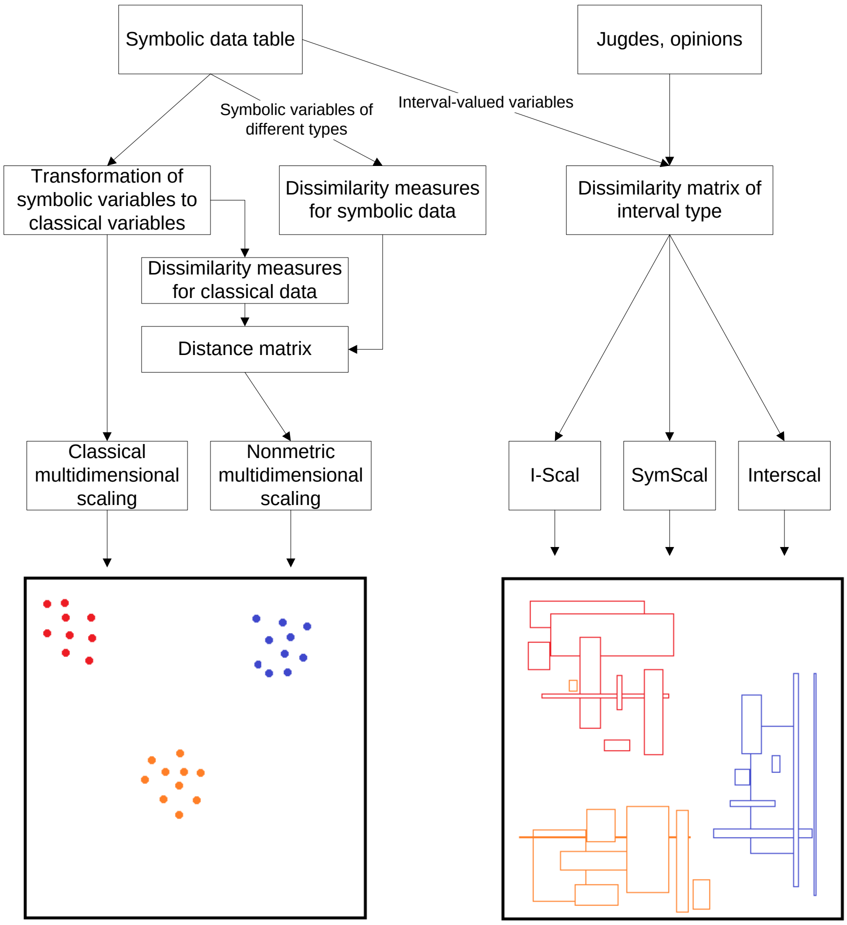

Fig. 1. Different approaches in multidimensional scaling of interval-valued symbolic data

Source: Pełka 2010.

The classical approach is based on transformation of symbolic variables to classical variables. It allows to present symbolic objects as points, but transformation causes some information loss about original data structure.

Methods based on symbolic dissimilarity measures don't cause loss of information, but the result symbolic objects are treated as points. Symbolic objects shouldn't be treated as points due to the fact that they are not points in multidimensional space. That's why symbolic multidimensional scaling methods based on interval-valued dissimilarities should be applied.

Algorithm of the Interscal method (Denoeux, Masson 2000, Lechevallier 2001):

1. Obtain interval-valued dissimilarities either by using interval-valued variables or judgments, opinions of n respondents, experts, etc.
2. Construct $\mathbf{\Delta}$ matrix of interval-valued dissimilarities, where $\overline{\delta}_{ij}$ is the upper bound of dissimilarity between i-th and *j*-th object, $\underline{\delta}_{ij}$ is the lower bound of dissimilarity between i-th and *j*-th object.
3. Construct $\tilde{\mathbf{\Delta}}$ matrix defined as follows:

$$\tilde{\mathbf{\Delta}} = \begin{bmatrix} 0 & 0 & \underline{\delta}_{12} & \frac{\overline{\delta}_{12}+\underline{\delta}_{12}}{2} & \cdots & \underline{\delta}_{1n} & \frac{\overline{\delta}_{1n}+\underline{\delta}_{1n}}{2} \\ 0 & 0 & \frac{\overline{\delta}_{12}+\underline{\delta}_{12}}{2} & \overline{\delta}_{12} & \cdots & \frac{\overline{\delta}_{1n}+\underline{\delta}_{1n}}{2} & \overline{\delta}_{1n} \\ \underline{\delta}_{21} & \frac{\overline{\delta}_{21}+\underline{\delta}_{21}}{2} & 0 & 0 & \cdots & \underline{\delta}_{2n} & \frac{\overline{\delta}_{2n}+\underline{\delta}_{2n}}{2} \\ \frac{\overline{\delta}_{21}+\underline{\delta}_{21}}{2} & \overline{\delta}_{21} & 0 & 0 & \cdots & \frac{\overline{\delta}_{2n}+\underline{\delta}_{2n}}{2} & \overline{\delta}_{2n} \\ \underline{\delta}_{31} & \frac{\overline{\delta}_{31}+\underline{\delta}_{31}}{2} & \underline{\delta}_{32} & \frac{\overline{\delta}_{32}+\underline{\delta}_{32}}{2} & \cdots & \underline{\delta}_{3n} & \frac{\overline{\delta}_{3n}+\underline{\delta}_{3n}}{2} \\ \frac{\overline{\delta}_{31}+\underline{\delta}_{31}}{2} & \overline{\delta}_{31} & \frac{\overline{\delta}_{32}+\underline{\delta}_{32}}{2} & \overline{\delta}_{32} & \cdots & \frac{\overline{\delta}_{3n}+\underline{\delta}_{3n}}{2} & \overline{\delta}_{3n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \underline{\delta}_{n1} & \frac{\overline{\delta}_{n1}+\underline{\delta}_{n1}}{2} & \underline{\delta}_{n2} & \frac{\overline{\delta}_{n2}+\underline{\delta}_{n2}}{2} & \cdots & 0 & 0 \\ \frac{\overline{\delta}_{n1}+\underline{\delta}_{n1}}{2} & \overline{\delta}_{n1} & \frac{\overline{\delta}_{n2}+\underline{\delta}_{n2}}{2} & \overline{\delta}_{n2} & \cdots & 0 & 0 \end{bmatrix}$$

4. Find the matrix $\mathbf{B} = -\frac{1}{2}\mathbf{J}\tilde{\mathbf{\Delta}}^{(2)}\mathbf{J}$ with $\mathbf{J}$ the centering matrix.
5. Find eigenvalues $\mathbf{\Phi}^2$ and eigenvectors $\mathbf{P}$ of matrix $\mathbf{B}$.
6. Compute $2n$ points in *S*-dimensions using the formula: $y_{is} = p_{is}\phi_{ss}$ for $i = 1,2,\ldots,2n$ and $s = 1,2,\ldots,S$.

7. When applying I-STRESS loss function, that can be interpreted in the same way as well-known STRESS, construct the center coordinates $\mathbf{X}$ and spreads of rectangle $\mathbf{R}$ for each object i and each dimension s as follows:

$$x_{is} = \frac{\left(y_{2i,s} + y_{2i+1,s}\right)}{2},$$

$$r_{is} = \frac{\left|y_{2i,s} - y_{2i+1,s}\right|}{2}$$

8. Compute I-STRESS value.

Algorithms of I-Scal and SymScal methods differ only in type of loss function applied. I-Scal uses normalized I-STRESS, SymScal uses unnormalized STRESS-Sym. The both algorithms are the same in other parts.

Algorithm of I-Scal method, SymScal respectively, (see: Groenen et. al. 2006; Groenen et. al. 2005):

1. Obtain interval-valued dissimilarities either by using interval-valued variables or judgments, opinions of n respondents, experts, etc.
2. Set matrix $\mathbf{X}_0$ to initial matrix for coordinate centers of rectangles (I-Scal random start point) or obtain matrix $\mathbf{X}_0$ by applying Interscal (I-Scal rational start point).
3. Set matrix $\mathbf{R}_0$ to initial matrix of nonnegative values for rectangles widths (I-Scal random start point) or obtain matrix $\mathbf{R}_0$ from Interscal (I-Scal rational start point).
4. Set maximum iteration number T and the convergence criterion ε to a small positive value – e.g. 10^{-6}.
5. Set iteration counter $k = 1$ and $\mathbf{X}_{-1} = \mathbf{X}_0$ and $\mathbf{R}_{-1} = \mathbf{R}_0$.
6. While $\text{I-STRESS}_{k-1} - \text{I-STRESS}_k \le \varepsilon$ and $k \le t$:
   a) $k = k + 1$,
   b) set $\mathbf{Y}_k = \mathbf{X}_{k-1}$ and $\mathbf{Q}_k = \mathbf{R}_{k-1}$.

   For every dimension

   c) compute $A_s^{(1)}$ and $B_s^{(1)}$ (see: Groenen et. al. 2006 for details).
   d) compute and update matrix $\mathbf{X}$ of coordinate centers for rectangles.
   e) compute $A_s^{(2)}$ and $b_s^{(2)}$ (see: Groenen et. al. 2006 for details).
   f) compute and update matrix of nonnegative values of rectangle width $\mathbf{R}$.
   g) Set $\mathbf{X}_k = \mathbf{X}$ and $\mathbf{R}_k = \mathbf{R}$.

The I-STRESS loss function, that takes values between 0 and 1, is defined as follows (see: Groenen et. al. 2006):

$$\text{I-STRESS} = \frac{\sigma_I^2(\mathbf{X},\mathbf{R})}{\sum_{i<j}^{n} w_{ij}\left[\delta_{ij}^{(U)}\right]^2 + \sum_{i<j}^{n} w_{ij}\left[\delta_{ij}^{(L)}\right]^2},$$

where: $\sigma_I^2(\mathbf{X},\mathbf{R}) = \sum_{i<j}^{n} w_{ij}\left[\delta_{ij}^{(U)} - d_{ij}^{(U)}(\mathbf{X},\mathbf{R})\right]^2 + \sum_{i<j}^{n} w_{ij}\left[\delta_{ij}^{(L)} - d_{ij}^{(L)}(\mathbf{X},\mathbf{R})\right]^2,$

$\mathbf{X},\mathbf{R}$ – matrices of rectangle centers $(\mathbf{X})$ and rectangles span $(\mathbf{R})$,

w_{ij} – weights,

$\delta_{ij}^{(U)}$ and $\delta_{ij}^{(L)}$ – upper and lower distances between *i*-th and *j*-th hiperrectangle,

$d_{ij}^{(U)}$ and $d_{ij}^{(L)}$ – upper and lower distances between rectangles.

References:

1. Billard L., Diday E. (Eds.) (2006), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons Ltd, Chichester.
2. Bock H.-H., Diday E. (Eds.) (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*. Springer Verlag, Berlin-Heidelberg.
3. Denoeux T., Masson M. (2000), *Multidimensional scaling of interval-valued dissimilarity data*. "Pattern Recognition Letters", vol. 21, issue 1:83–92.
4. Groenen P.J.F., Winsberg S., Rodrguez O., Diday E. (2005), *Multidimensional Scaling of Interval Dissimilarities*. Econometric Report, 2005-15, Erasmus University, Rotterdam.
5. Groenen P.J.F., Winsberg S., Rodriguez O., Diday E. (2006), *I-Scal: multidimensional scaling of interval dissimilarities*. "Computational Statistics and Data Analysis", vol. 51:360-378.
6. Lechevallier Y. (Ed.) (2001), *Scientific report for unsupervised classification, validation and cluster representation*. Analysis System of Symbolic Official Data – Project number IST-2000-25161, project report.
7. Pełka M. (2010), *Symbolic multidimensional scaling versus noisy variables and outliers* [in:] H. Locarek-Junge, C. Weihs (Eds.) Classification as a tool for research, Springer-Verlag, Berlin-Heidelberg, s. 341-350.
8. Sammon J.W. (1969), *A nonlinear mapping for data structure analysis*. "IEEE Transactions on Computers", vol. C-18, no. 5:401–409.