ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

Andrzej Dudek[*]

MULTIDIMENSIONAL SCALING FOR SYMBOLIC INTERVAL DATA

Abstract. The aim of multidimensional scaling is to represent dissimilarities among objects in high dimensional space as distances in low (usually 2- or 3-) dimensional space. Usually the input to multidimensional scaling procedure is a square, symmetric matrix indicating relationships (similarities or dissimilarities) among a set of items. There are many techniques of classical multidimensional scaling but all under assumption that each entry in relationship matrix is single numeric value.

Denoeux and Masson (2002) have proposed to extend multidimensional scaling onto symbolic interval data. The input to theirs INTERSCAL algorithm is interval dissimilarity table containing minimum and maximum distance between hyper-rectangles representing objects. The same approach is used in SYMSCAL and I-SCAL algorithms proposed by Groenen *et al.* (2005).

Article presents main algorithms of multi-dimensional scaling for symbolic data in form of intervals along with some examples on datasets taken from symbolic data repository (http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm).

Key words: Multidimensional scaling, visualization, symbolic data.

I. INTRODUCTION

Visualizing data in the form of illustrative diagrams and searching, in these diagrams, for structures, clusters, trends, dependencies etc. is one of the main aims of multivariate statistical analysis. In the case of symbolic data (e.g. data in form of: single quantitative value, categorical values, intervals, multi-valued variables, multi-valued variables with weights), some well-known methods are provided by suitable 'symbolic' adaptations of classical methods such as principal component analysis, factor analysis or multidimensional scaling (MDS) (Kruskal (1964)). The main difference between classical methods and "symbolic" methods is form of data they are dealing with. In case of symbolic data analysis the input contains intervals instead of single numerical values.

[*] Ph.D., Chair of Econometrics and Informatics, University of Economics, Wrocław.

This paper describes methods of multidimensional scaling of symbolic objects containing variables in form of intervals (symbolic interval data). The aim of multidimensional scaling of symbolic interval data is, like in classical, case: to represent dissimilarities among objects in n-dimensional as distances in reduced 2- or 3- dimensional space. But, while in classical MDS points of z n-dimensional space are transformed into points in p-dimensional space ($p = 2$ or 3), in case of multidimensional scaling of symbolic interval data hiper-cubes of higher dimensional space are translated into intervals in lower dimensional space.

First chapter describes the form of input and output data for algorithms of multidimensional scaling of symbolic interval data. Second presents three main methods of multidimensional scaling for symbolic interval data: Interscal, I-scal and Symscal, which is more detailed described in third chapter. Forth chapter presents example of usage of Symscal algorithm on symbolic data acquired from symbolic data repository. Finally some remarks and conclusions are given.

II. INPUT AND OUTPUT DATA

All methods of multidimensional scaling for symbolic interval data require matrix of minimal and maximal distances between objects in form similar to formula 1.

$$\begin{bmatrix} \underline{\delta_{11}},\overline{\delta_{11}} & \underline{\delta_{21}},\overline{\delta_{21}} & \cdots & \underline{\delta_{n1}},\overline{\delta_{n1}} \\ \underline{\delta_{21}},\overline{\delta_{21}} & \underline{\delta_{22}},\overline{\delta_{22}} & \cdots & \underline{\delta_{n2}},\overline{\delta_{n2}} \\ \cdots & \cdots & \ddots & \cdots \\ \underline{\delta_{n1}},\overline{\delta_{n1}} & \underline{\delta_{n2}},\overline{\delta_{n2}} & \cdots & \underline{\delta_{nn}},\overline{\delta_{nn}} \end{bmatrix} \tag{1}$$

where:

$\underline{\delta_{ij}}$ minimal distance between i-th and j-th symbolic object.

$\overline{\delta_{ij}}$ maximal distance between i-th and j-th symbolic object.

n number of symbolic objects.

Sometimes this matrix is not given directly but should be calculated from n symbolic objects containing intervals. In these case, according to Deneux and Masson (2000) minimal and maximal distances are computed due to formulas 2 and 3.

$$\overline{\delta_{ij}} = \frac{1}{2}\sqrt{\sum_{k=1}^{m}\left[\left(\overline{x}_{ik} - \underline{x}_{ik}\right) + \left(\overline{x}_{jk} - \underline{x}_{jk}\right) + 2\left|\frac{\overline{x}_{ik} - \underline{x}_{ik}}{2} - \frac{\overline{x}_{jk} - \underline{x}_{jk}}{2}\right|\right]^2} \tag{2}$$

$$\underline{\delta_{ij}} = \frac{1}{4}\sqrt{\sum_{k=1}^{m}\left[\left(\overline{x}_{ik} - \underline{x}_{ik}\right) + \left(\overline{x}_{jk} - \underline{x}_{jk}\right) + 2\left|\frac{\overline{x}_{ik} - \underline{x}_{ik}}{2} - \frac{\overline{x}_{jk} - \underline{x}_{jk}}{2}\right| - \right.}$$

$$\left.\overline{\left|\left(\overline{x}_{ik} - \underline{x}_{ik}\right) + \left(\overline{x}_{jk} - \underline{x}_{jk}\right) + 2\left|\frac{\overline{x}_{ik} - \underline{x}_{ik}}{2} - \frac{\overline{x}_{jk} - \underline{x}_{jk}}{2}\right|\right|}\right]^2 \tag{3}$$

where:

$\underline{\delta_{ij}}$ – minimal distance between i-th and j-th symbolic object,

$\overline{\delta_{ij}}$ – maximal distance between i-th and j-th symbolic object,

$(\underline{x_{ij}}, \overline{x_{ij}})$ – j-th variable of i-th object (the beginning and the end of interval)

m – number of symbolic variables describing each object.

The aim of multidimensional scaling of symbolic interval data is like in classical case: to represent dissimilarities among objects in n-dimensional as distances in 2- or 3- dimensional space thus output matrix also contains minimal and maximal distances between and can be written in form of 4.

$$\begin{bmatrix} \underline{d_{11}}, \overline{d_{11}} & \underline{d_{21}}, \overline{d_{21}} & \cdots & \underline{d_{n1}}, \overline{d_{n1}} \\ \underline{d_{21}}, \overline{d_{21}} & \underline{d_{22}}, \overline{d_{22}} & \cdots & \underline{d_{n2}}, \overline{d_{n2}} \\ \cdots & \cdots & \ddots & \cdots \\ \underline{d_{n1}}, \overline{d_{n1}} & \underline{d_{n2}}, \overline{d_{n2}} & \cdots & \underline{d_{nn}}, \overline{d_{nn}} \end{bmatrix} \tag{4}$$

where:

$\underline{d_{ij}}$ – minimal distance between i-th and j-th symbolic object in reduced space,

$\overline{d_{ij}}$ – maximal distance between i-th and j-th symbolic object in reduced space,

n – number of symbolic objects.

III. METHODS OF MULTIDIMENSIONAL SCALING FOR SYMBOLIC INTERVAL DATA

There are three main algorithms of multidimensional scaling for symbolic interval data, one non-iterative (Interscal) and two (I-scal and Symscal) iteratively searching for optimal value of transformation loss function.

Main steps of Interscal (Denoeux, Masson [2002]) method can be stated as:

- Calculation of modified $\tilde{\Delta}$ matrix containing $2n$ rows and $2n$ columns with minimal distance, maximal distance and average distance for every pair of objects.
- Calculation of B – matrix of scalar products of rows of $\tilde{\Delta}$,
- Calculation of eigenvectors v and eigenvalues λ of B,
- Calculation of $\underline{d_{ij}}$ $\overline{d_{ij}}$.

I-scal and Symscal algorithms starting from $\mathbf{X}$ (centers of intervals) and $\mathbf{R}$ (spread of intervals) matrices are iteratively searching for optimal values of I-Stress/Stress-Sym loss function .

IV. SYMSCAL

The idea of SymScal algorithm is majorization of loss function. In first steps matrices $\mathbf{X}$ (centers of intervals) and $\mathbf{R}$ (spread of intervals) are generated (usually in random way, but Groenen *et al.* (2006) suggests to use Interscal algorithm to find initial values of $\mathbf{X}$ and $\mathbf{R}$.

In second and next steps, *STRESS-Sym* measure is calculated due to formula 5

$$\text{STRESS-Sym}\left(\mathbf{X},\mathbf{R}\right)=\sum_{i<j}^{n}\omega_{ij}\left[\overline{\delta}_{ij}-\overline{d}_{ij}(\mathbf{X},\mathbf{R})\right]^2+\sum_{i<j}^{n}\omega_{ij}\left[\underline{\delta}_{ij}-\underline{d}_{ij}(\mathbf{X},\mathbf{R})\right]^2 \quad (5)$$

where:

$\mathbf{X},\mathbf{R}$ – centers of intervals and spread of intervals,

ω_{ij} – weights (usually equal)

$\underline{\delta_{ij}}$ – minimal distance between i-th and j-th symbolic object,

$\overline{\delta_{ij}}$ – maximal distance between i-th and j-th symbolic object,

$\underline{d_{ij}}$, $\overline{d_{ij}}$ – minimal and maximal distance between i-th and j-th symbolic object in reduced space, calculated from $\mathbf{X}$ and $\mathbf{R}$ according to 6 and 7,

$$\overline{d}_{ij}(\mathbf{X},\mathbf{R}) = \sqrt{\sum_{s=1}^{p}\left[\left|x_{is} - x_{js}\right| + \left(r_{is} + r_{js}\right)\right]^2}, \tag{6}$$

$$\underline{d}_{ij}(\mathbf{X},\mathbf{R}) = \sqrt{\sum_{s=1}^{p}\max\left[0, \left|x_{is} - x_{js}\right| + \left(r_{is} + r_{js}\right)\right]^2}, \tag{7}$$

p dimensionality of reduced space.

Method of majorization of *STRESS-Sym* is described in details in Groenen et al. (2006). The main weakness of this method is fact that *STRESS-Sym* is not a normalized measure. Thus one can observe the improvement of loss function during iteration process but cannot compare quality of transformation to other datasets.

V. EXAMPLE OF USE OF MULTIDIMENSIONAL SCALING FOR SYMBOLIC INTERVAL DATA

As an illustration of usage of multidimensional scaling for symbolic interval data the wine.xml set taken from symbolic data repository has been used. These set contained 21 symbolic objects described by 23 symbolic interval variables. Figure 1 shows original

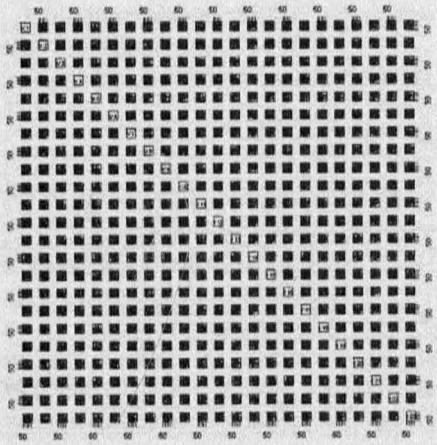


Fig. 1. wine.xml dataset in original space

As one can see there is no clear structure of data. In fact there are to many dimensions on the scatterplot to observe anything.

Figure 2 shows data in first 6 dimensions but even with this limitation no structure can be observed.

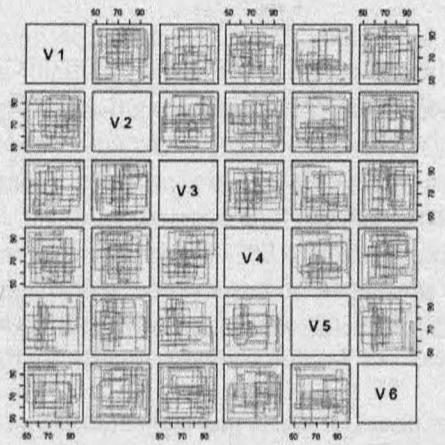
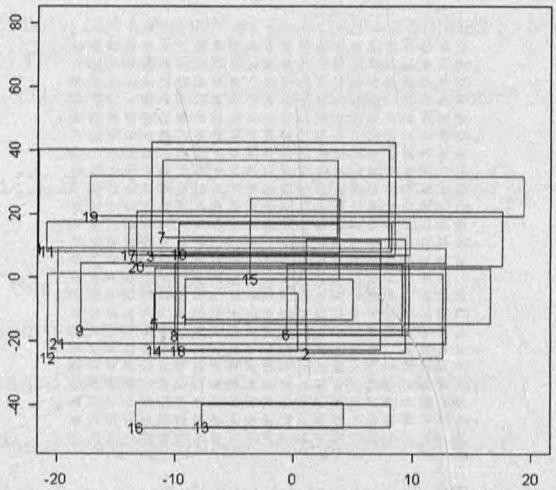


Fig. 2. wine.xml dataset in original space (first 6 dimensions)

Symscal procedure has been to those data. All calculations have been made in R statistical environment with use of package SymbolicDA written by author. Effects of multidimensional scaling presents figure 6. The *STRESS-Sym* loss function has changed from 476620 to 816,35 during 8 iterations steps.



1 Ausone; 2 Cheval Blanc; 3 Cos d'Estournel; 4 Ducru-Beaucaillou; 5 Haut-Brion; 6 Lafite-Rothschild; 7 Lafleur; 8 Latour; 9 Leoville Las Cases; 10 L'Evangile; 11 Lynch-Bages; 12 Margaux; 13 Mission Haut-Brion; 14 Montrose; 15 Mouton-Rothschild; 16 Petit Village; 17 Petrus; 18 Pichon C.de Lalande; 19 Pichon Longueville; 20 Sassicaia; 21 Trotanoy;

Fig. 3. wine.xml dataset in 2-dimensional space

VI. FINAL REMARKS

Methods of multidimensional scaling can be adapted to symbolic interval data. There are three main methods of multidimensional scaling for symbolic interval data: Interscal, Symscal and I-Scal. The main weakness of those method is lack of an objective measure of quality of transformation.

An open issue is also adaptation or development of new method of multidimensional scaling of other types of symbolic data type (nominal and multinominal, categorical data, distributions).

REFERENCES

Billard L., Diday E. (2006), *Symbolic data analysis. Conceptual statistics and data mining*, Wiley, Chichester.

Bock H.-H., Diday E. (eds.), (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*, Springer Verlag, Berlin.

Denoeux T., Masson M. (2000), Multidimensional scaling of interval-valued dissimilarity data, *Pattern Recognition Letters*, vol. 21, issue 1, 83–92.

Groenen P. J. F., Winsberg S., Rodríguez O., Diday E. (2005), SymScal: Symbolic Multidimensional Scaling of Interval Dissimilarities, *Econometric Report EI 2005-15*, Erasmus University, Rotterdam.

Groenen P. J. F., Winsberg S., Rodríguez O., Diday E. (2006), *I-Scal:* Multidimensional scaling of interval dissimilarities, *Computational Statistics & Data Analysis* vol. 51, issue 1, 360-378.

Kruskal, J. B. (1964a), Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27.

Kruskal, J. B. (1964b),s Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.

Andrzej Dudek

SKALOWANIE WIELOWYMIAROWE DLA DANYCH SYMBOLICZNYCH PRZEDZIAŁOWYCH

Podstawowym celem skalowania wielowymiarowego jest przedstawienie relacji między obiektami w przestrzeni wielowymiarowej jako odległości w przestrzeni 2- lub 3- wymiarowej. Dane wejściowe do procedur skalowania wielowymiarowego to zazwyczaj symetryczna macierz kwadratowa wskazująca na relacje (podobieństwa lub niepodobieństwa) pomiędzy obiektami pewnego zbioru. Istnieje wiele technik klasycznego skalowania wielowymiarowego, jednak wszystkie z nich wymagają aby w poszczególnych komórkach tej macierzy znajdowały się pojedyncze wartości liczbowe.

Denoeux and Masson (2002) zaproponowali rozszerzenie klasycznego skalowania wielowymiarowego na dane symboliczne w postaci przedziałów liczbowych. Danymi

wejściowymi do opracowanego przez nich algorytmu INTERSCAL jest tabela zawierająca minimalne i maksymalne odległości pomiędzy hiperprostopadłościanami reprezentującymi obiekty. Takie same podejście występuje w algorytmach SYMSCAL i I-SCAL zaproponowanych przez Groenena i in. (2005).

W artykule przedstawiony zostały najważniejsze algorytmy skalowania wielowymiarowego dla danych symbolicznych w postaci przedziałów liczbowych oraz przykłady ich zastosowania dla danych symbolicznych pochodzących z repozytorium http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm.