EKONOMETRIA ECONOMETRICS 3(41) • 2013

ISSN 1507-3866

Artur Zaborski, Marcin Pełka
Wroclaw University of Ecomomics

UNFOLDING ANALYSIS ADAPTATION FOR SYMBOLIC DATA – HYBRID AND SYMBOLIC-NUMERIC APPROACH

Abstract: The aim of this paper is to propose and present adaptations of unfolding analysis for symbolic data. In the article, the basic terms of unfolding analysis and symbolic data are presented. The paper presents two approaches – the internal hybrid approach and the external symbolic-numeric approach. In the empirical part, the external symbolic-numeric unfolding for LCD brands is presented. Symbolic multidimensional scaling R source codes were written by authors.

Keywords: symbolic data analysis, unfolding analysis, preference data.

1. Introduction

The classical unfolding analysis presents points which represent respondents and their preferences on one map. Thus dependencies between respondents and objects can be analyzed. Unfolding analysis, unlike classical multidimensional scaling, requires data in preference matrix or dissimilarity and preference matrix.

There are significant differences between symbolic data and well-known classical data. Objects can be described not only by single-valued variables but also by interval-valued variables, sets of categories, set of categories with associated weights and dependent variables. It is not possible to perform a classical multidimensional scaling with such representations. In order to perform unfolding analysis for symbolic objects an interval-valued dissimilarity or dissimilarity and preferences matrices should be used. Dissimilarity matrix is obtained by applying one of the symbolic dissimilarity measures.

The article presents an adaptation of the classical unfolding analysis for "classical" data to "symbolic" data case. The first part of the paper presents the basis of the unfolding analysis and concentrates on the differences between the internal and external unfolding. The second part presents the terms of symbolic data analysis like symbolic variable, symbolic objects. In the third part, adaptations of unfolding analysis for symbolic data are proposed with a special focus on the internal hybrid

approach and external symbolic-numeric approach. The empirical part demonstrates an example of the application on a real data set where the symbolic-numeric approach was applied.

2. Unfolding analysis for classical data

The unfolding analysis for classical data attempts to produce a configuration $\mathbf{Y}$ of points in r-dimensional space with each point $\mathbf{y}_k$ ($k = 1, \ldots, m$) representing one of m judges, together with another configuration $\mathbf{X}$ of points $\mathbf{x}_i$ ($i = 1, \ldots, n$) in the same space, which represent choice objects. Individuals are represented as "ideal" points in the multidimensional space, so that the distances between each ideal point and the object point correspond to the preference scores. The ideal point model is used to find a point in a stimulus space which is most like an attribute. If the attribute is a subject's preference for the stimuli, then this point is interpreted as a subject's ideal stimulus. This is the hypothetical stimulus which the subject would prefer the most if it existed.

There are two main approaches to the unfolding procedure:

– internal unfolding,
– external unfolding.

In internal unfolding both the object configuration and the ideal points are simultaneously derived only from the preference matrix. We can conceive the preference matrix as a submatrix of the dissimilarity matrix in which the dissimilarity between objects and between respondents are treated as missing values [Borg, Groenen 2005].

One of the most common methods for coordinates estimation in internal unfolding analysis is the Guttmann transform (see [Borg, Groenen 2005; Zaborski 2011]) realized by SMACOF (*Scaling by MAjorizing a COmplicated Function*) and by PREFSCAL (*PREFerence SCALing*). The possibility to avoid degeneracy solutions by the modification of the loss function is the main advantage of PREFSCAL. In the structure of the STRESS function the variation coefficient is used as a diagnostic for identifying solutions with constant interpoint distances (see: [Busing, Groenen, Heiser 2005]).

In external unfolding it is assumed that a similarity objects configuration is given. With preference data on these objects external unfolding puts the ideal point for each subject in the space, so that the closer this point lies to a point that represents an object, the more this object is preferred by an individual. The external analysis for the preference data is realized by PREFMAP (*PREFerence MAPping*), which consist of four preference-property models: vector model, simple unfolding model, weighted unfolding model and general unfolding model. Detailed algorithms for ideal points and vectors models in PREFMAP are presented by [Davison 1983].

3. Symbolic data

Bock and Diday have defined five different symbolic variables types [Bock, Diday (eds.) 2000]:
– single quantitative value,
– categorical value,
– quantitative variable of interval type,
– set of values or categories (multivalued variable),
– set of values or categories with weights (multivalued variable with weights),
– modal interval-valued variable proposed in [Billard, Diday (eds.) 2006].

More details on symbolic data, symbolic variables and differences between them and classical data can be found for example in: Noirhomme-Fraiture, Brito [2011], Bock, Diday [2000], Billard, Diday [2006], Dudek [2013], Diday, Noirhomme-Fraiture [2008].

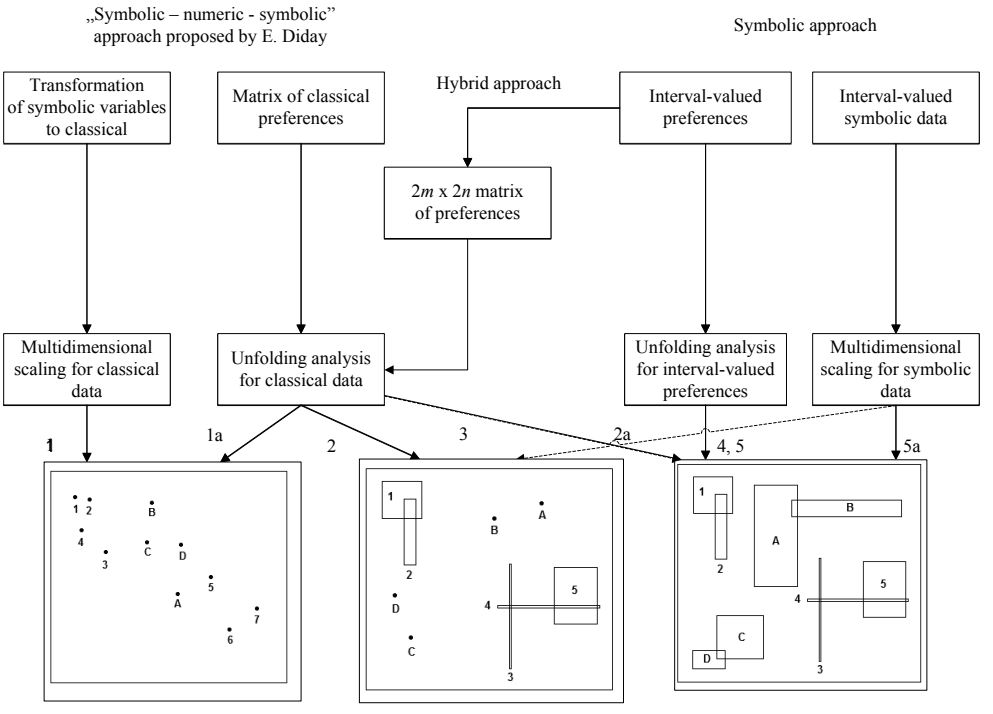
Symbolic data allows to take into account variability of the data but it requires special methods to deal with this kind of data. There are many methods of symbolic data analysis (see for example: [Dudek 2013; Bock, Diday (eds.) 2000; Billard, Diday (eds.) 2006]). However, there is a need to develop new methods for this kind of data.

4. Unfolding analysis for symbolic data

Unfolding analysis for symbolic data will allow to analyze consumer preferences towards products, companies, etc. that are represented by symbolic objects. Symbolic multidimensional unfolding allows also to take into account variability and uncertainty of the data.

The classical approach of unfolding for symbolic objects is based on the transformation of symbolic variables into classical ones. Symbolic objects can be presented as points, however some information about original data structure is lost due to the transformation. In contrast, there is no loss of information in the methods based on symbolic dissimilarity measures, but then symbolic objects are also treated as points, whereas symbolic objects should not be treat as points due to the fact that they are not points in multidimensional space. Therefore, symbolic methods based on interval-valued dissimilarities should be applied.

Figure 1 presents some propositions of possible unfolding methods for symbolic data. These methods, similarly to the methods in unfolding analysis for classical data, present two groups of approaches – internal and external. This paper discusses only two approaches: the internal hybrid approach and the external symbolic-numeric approach.



1 and 1a – external approach (after transformation of symbolic variables); 2 and 2a – external approach (multidimensional scaling for symbolic data and classical unfolding analysis); 3 – internal hybrid unfolding (which uses $\hat{P}$ matrix); 4 – internal symbolic unfolding analysis; 5 and 5a – external symbolic unfolding analysis (which uses symbolic multidimensional scaling and symbolic unfolding).

Figure 1. Unfolding methods for symbolic data

Source: authors' elaboration based on [Groenen et al. 2005, 2006; Lattin, Carroll, Green 2003].

4.1. Internal hybrid unfolding

The algorithm of internal hybrid method has the following stages:

1. Collection of preferences from *n* sources (judges, experts, etc.),
2. Construction of interval-valued preference matrix **P** from *n* sources (see: [Lechevallier (ed.) 2001]).[1]
3. Construction of $\hat{\mathbf{P}}$ matrix defined as follows:

[1] Besides the methods presented in [Lechevallier (ed.) 2001], the authors propose to apply mean minus standard deviation as the beginning of the interval and mean plus standard deviation as the end of the preference interval.

$$\hat{\mathbf{P}} = \begin{bmatrix} \underline{p}_{11} & \dfrac{\underline{p}_{11} + \overline{p}_{11}}{2} & \cdots & \dfrac{\underline{p}_{n1} + \overline{p}_{n1}}{2} \\ \dfrac{\underline{p}_{11} + \overline{p}_{11}}{2} & \overline{p}_{11} & \cdots & \overline{p}_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\underline{p}_{m1} + \overline{p}_{m1}}{2} & \overline{p}_{m1} & \cdots & \overline{p}_{mn} \end{bmatrix},$$

where: $\underline{p}_{ik}$ – lower bound of preference interval; $\overline{p}_{ik}$ – upper bound of preference interval.

4. Application of unfolding analysis for classical data, where $\hat{\mathbf{P}}$ matrix is the input.

5. Attainment of $2m$ points for rows (y_{ha}) and $2n$ points for columns (x_{la}), where: $h = 1,...,2m$ – number of rows, $l = 1,..., 2n$ – number of columns.

6. Interval-valued coordinates computation of rectangles as follows: $\underline{y}_{ka} = \min_{h\in(2k-1,\ 2k)} \{y_{ha}\}$; $\overline{y}_{ka} = \max_{h\in(2k-1,\ 2k)} \{y_{ha}\}$; $\underline{x}_{ia} = \min_{l\in(2i-1,\ 2i)} \{x_{la}\}$; $\overline{x}_{ia} = \max_{l\in(2i-1,\ 2i)} \{x_{la}\}$, where: $k = 1,..., m$ – number of respondent, $i = 1,..., n$ – number of object, $a = 1,...,$ r – space dimensions.

4.2. External symbolic-numeric unfolding

The algorithm of external symbolic-numeric unfolding has the following stages:

1. Attainment of interval-valued variables for symbolic objects or collection of m judgments, opinions, etc. on objects.

2. Performance of the multidimensional scaling for symbolic data with one of the methods: Interscal, SymScal, I-Scal (for details see: [Groenen et al. 2005; Lattin, Carroll, Green 2003]).

3. Attainment of rectangles for the preference map.

4. Construction of matrix **R** of rectangles centers for unfolding analysis.

5. Collection of m preferences for n objects.

6. Mapping points representing respondents through unfolding analysis for classical data (e.g. PREFMAP) on the configuration of rectangles centers (elements of matrix **R**).

7. Presentation of rectangles (representing columns) and points (representing rows) on one perceptual map.

The main difference between internal hybrid unfolding for symbolic data and external symbolic-numeric unfolding is that in the first approach (internal hybrid unfolding) the rows and columns of the preference matrix (with interval-valued preferences) are assumed to be symbolic objects. On the resulting perceptual map we

shall have rectangles representing columns and rows. In the case of the external symbolic-numeric approach, we use interval-valued data for symbolic multidimensional scaling and numeric preferences. On the resulting perceptual map we have rectangles representing symbolic rows and points representing columns. So internal hybrid unfolding can be applied when we deal only with interval-valued preferences (without any prior knowledge about the objects). External symbolic-numeric unfolding can be applied when we deal with interval-valued data (or dissimilarities) for symbolic objects and numerical preferences.

5. Empirical example

Stage 1. Data were collected from 100 PC experts/dealers about 8 LCD display brands (Samsung, LG, Maxdata, Philips, BenQ, NEC, Neovo, Hyundai). They were asked to compare brands from 0 (identical brands offer) to 100 (different brands product range). The experts also compared the variety of LCD models in each brand from 0 (all models are identical) to 100 (all models differ from each other). The aim of this part was to perform the multidimensional scaling of LCD brands using the Interscal procedure. The results are presented in Figure 2.

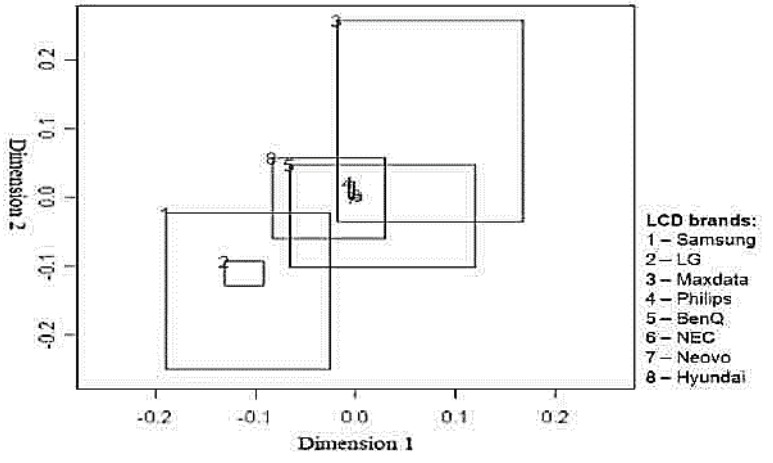


Figure 2. Results of multidimensional scaling for symbolic data

Source: authors' computations in R software.

Stage 2. Preference data were collected from 28 PC experts/dealers (different than in stage one). The experts were asked to rank 8 LCD display brands (the same brands as in stage one) according to their preferences on a 8-point scale: 1 – the highest preferred brand, 8 – the least preferred brand. The aim was to combine the results of multidimensional scaling from stage two with the experts' preferences.

The symbolic-numeric approach was applied and points representing rows were computed with the application of PREFSCAL. The results of the symbolic-numeric external unfolding are presented in Figure 3.

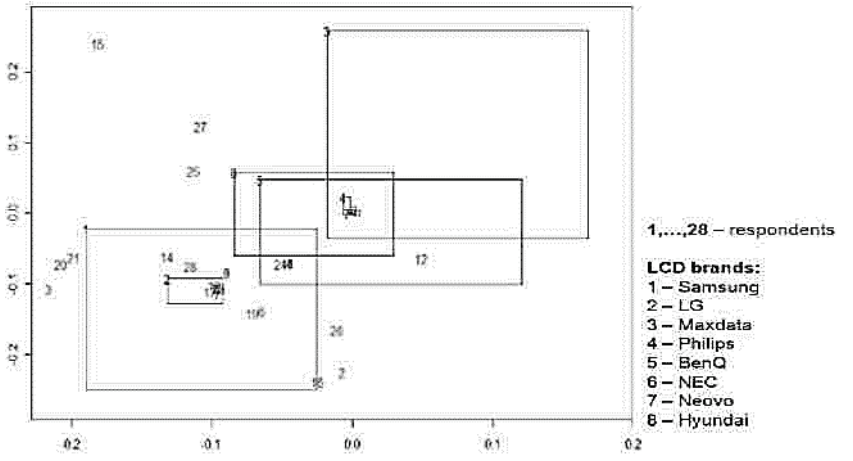


Figure 3. Results of unfolding analysis for symbolic data

Source: authors' computations in R software.

Samsung and LG (1 and 2 in Figure 3) are the most preferred brands by all PC experts. Maxdata (3 in Figure 3) is the least preferred brand by all PC experts. Some experts prefer some other brands, such as Philips, BenQ, NEC, Neovo, Hyundai (4, 5, 6, 7, 8 in Figure 3).

6. Conclusions

Four different unfolding analysis methods for symbolic data where proposed on the basis of the literature review: symbolic-classical-symbolic approach, hybrid approach, mixed: symbolic-classical approach and symbolic approach based on I-STRESS. Two of them were described in a detailed way – internal hybrid unfolding based on interval-valued preferences and external symbolic-numeric unfolding based on numeric preferences and interval-valued data. In the empirical part external symbolic-numeric unfolding for symbolic data was applied to analyze LCD brands.

The location of points in Figure 3 representing respondents indicates that the most preferred brands for most respondents are Samsung and LG. In this particular example, the fact that the objects perceived as similar might be differently preferred may pose a problem. Also objects which are treated as different might be preferred in the same way.

The comparison of research results obtained by unfolding analysis for the internal hybrid approach and symbolic-numeric approach might be interesting. However, it

was impossible in this case, since the preference data used in the research were not symbolic.

An open issue for further research is the development of unfolding analysis for symbolic data based on I-STRESS, satisfying the monotone transformation.

Literature

Billard L., Diday E. (eds.), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester 2006.

Bock H.H., Diday E. (eds.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg 2000.

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

Busing F.M.T.A., Groenen P.J.K., Heiser W.J., *Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation*, "Psychometrika" 2005, no. 1.

Davison M.L., *Multidimensional Scaling*, John Wiley and Sons, New York 1983.

Diday E., Noirhomme-Fraiture M., *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester 2008.

Dudek A., *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wydawnictwo UE, Wrocław 2013.

Groenen P.J.F., Winsberg S., Rodriguez O., Diday E., *I-Scal: Multidimensional scaling of interval dissimilarities*, "Computational Statistics and Data Analysis" 2006, vol. 51.

Groenen P.J.F., Winsberg S., Rodriguez O., Diday E., *Multidimensional Scaling of Interval Dissimilarities*, Econometric Report, 2005-15, Erasmus University, Rotterdam 2005.

Lattin J., Carroll J.D., Green P.E., *Analyzing Multivariate Data*, Thomson Learning, Toronto 2003.

Lechevallier Y. (ed.), *Scientific report for unsupervised classification, validation and cluster representation,* Analysis System of Symbolic Official Data − Project number IST-2000-25161, Project Report, 2001.

Noirhomme-Fraiture M., Brito P., *Far beyond the classical data models: Symbolic data analysis*, "Statistical Analysis and Data Mining" 2011, vol. 4, issue 2, pp. 157-170.

Zaborski A., *Zastosowanie algorytmu SMACOF do badań opartych na prostokątnej macierzy preferencji*, [w:] Taksonomia 18, *Klasyfikacja i analiza danych – teoria i zastosowania,* K. Jajuga, M. Walesiak (red.), Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, Wrocław 2011, pp. 262-271.

ADAPTACJA ANALIZY *UNFOLDING* DLA DANYCH SYMBOLICZNYCH – PODEJŚCIE HYBRYDOWE I SYMBOLICZNO-NUMERYCZNE

Streszczenie: Celem artykułu jest zaprezentowanie propozycji adaptacji analizy *unfolding* dla danych symbolicznych. W tekście scharakteryzowano podstawowe pojęcia związane z analizą *unfolding* i danymi symbolicznymi. Przedstawione zostały dwa podejścia – wewnętrzne – hybrydowe, i zewnętrzne − symboliczno-numeryczne. W części empirycznej przedstawiono wyniki zastosowania podejścia zewnętrznego − symboliczno-numerycznego − na przykładzie danych dotyczących monitorów LCD. Na użytek obliczeń przygotowano kody źródłowe programu R.

Słowa kluczowe: analiza danych symbolicznych, *unfolding*, dane preferencji.