

A Method of Asymmetric Multidimensional Scaling with External Information

*Un Metodo di Scaling Multidimensionale Asimmetrico
con Informazione Esterna*

Giuseppe Bove

Università di “Roma Tre”, Via del Castro Pretorio 20, Roma, e-mail:bove@uniroma3.it

Roberto Rocci

Università di “Tor Vergata”, Via Columbia 2, Roma, e-mail:roberto.rocci@uniroma2.it

Riassunto: In questa comunicazione viene proposto un nuovo metodo di multidimensional scaling, con informazione esterna, per la rappresentazione grafica di matrici di prossimità quadrate non necessariamente simmetriche.

Keywords: Data Representation, Graphical Methods, Asymmetric MDS

1. Introduction

Multidimensional scaling (MDS) is a set of techniques for the analysis of square data matrices whose rows and columns correspond to the same set of "objects". Examples of this type of matrices are: proximities (e.g. similarity ratings), preferences (e.g. socio-matrices), flow data (e.g. import-export, brand switching). Scalar product and Euclidean distance models allow their representation in low-dimensional spaces when the relationships between rows and columns are symmetric. These models have to be suitably modified by increasing the number of parameters, and complicating the graphical representation, when not random asymmetry is present in the data (see e.g. Zielman and Heiser, 1996). Some of these methods are based on the unique decomposition of the data matrix in its symmetric and skew-symmetric components. Following this approach, Rocci and Bove (2002) proposed a general model, named GEG, for a joint representation of the two components, allowing for rotations that make easier the graphical interpretation in more than two dimensions. In this paper the model is extended for taking into account external information. This allows to analyze the contribution of variables suggested from theoretical knowledge to the explanation of the relationships in the data. In the next sections the main aspects of the extension are reported along with an application to Morse code confusion data (Rothkopf, 1957) illustrating the proposal.

2. Generalized Escoufier and Grolud model (GEG) with external information

Each square data matrix $\mathbf{X} = [x_{ij}]$, whose rows and columns correspond to the same set of n objects, is uniquely decomposable into the symmetric and the skew-symmetric

components: $\mathbf{X} = \mathbf{S} + \mathbf{K}$, where $\mathbf{S}$ is the symmetric matrix of averages $s_{ij} = \frac{1}{2}(x_{ij} + x_{ji})$ and $\mathbf{K}$ is the skew-symmetric matrix of differences $k_{ij} = \frac{1}{2}(x_{ij} - x_{ji})$. In order to analyse $\mathbf{S}$ and $\mathbf{K}$ jointly Escoufier and Grolud (1980) proposed the following EG model:

$$x_{ij} = \sum_{m=1}^M (a_{im}a_{jm} + b_{im}b_{jm} + a_{im}b_{jm} - b_{im}a_{jm}) + \varepsilon_{ij}$$

where ε_{ij} is a residual term. When $M = 1$ the model allows us to represent objects in a plane (*bimension*) having coordinates (a_{i1}, b_{i1}) ; for a pair of points in this plane, the scalar product describes the symmetric component, while twice the area of the triangle having the two points and the origin as vertices describes the absolute value of the skew-symmetric component, whose algebraic sign is associated with the orientation of the plane (positive counter-clockwise, negative clockwise). The previous EG model can be considered a constrained version of the GEG model proposed by Rocci and Bove (2002)

$$x_{ij} = \sum_{m=1}^M [\gamma(a_{im}a_{jm} + b_{im}b_{jm}) + \delta(a_{im}b_{jm} - b_{im}a_{jm})] + \varepsilon_{ij}$$

where $\gamma = 0$ if the data matrix is skew (*Gower decomposition*) and $\gamma = 1$ otherwise, while $\delta \geq 0$. When $\gamma = 1$ and $M=1$ the GEG model is equivalent to Chino's GIPSCAL, while giving different weights to the skew-symmetric component we obtain Generalized GIPSCAL. GEG can be rotated simplifying the interpretation of bimensions. In order to improve data interpretation (e.g. data theory compatible MDS), we can incorporate in the analysis external information represented by a matrix of *external variables* $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p]$. We usually seek for an MDS solution whose dimensions are linear combination of the columns of $\mathbf{E}$. By rewriting GEG in vector notation

$$\mathbf{X} \approx \sum_{m=1}^M \gamma(\mathbf{a}_m \mathbf{a}_m' + \mathbf{b}_m \mathbf{b}_m') + \delta(\mathbf{a}_m \mathbf{b}_m' - \mathbf{b}_m \mathbf{a}_m')$$

and by imposing $\mathbf{a}_m = \mathbf{E} \mathbf{c}_m$ and $\mathbf{b}_m = \mathbf{E} \mathbf{d}_m$, we obtain

$$\mathbf{X} \approx \gamma \mathbf{E} \mathbf{H}^{(s)} \mathbf{E}' + \delta \mathbf{E} \mathbf{H}^{(k)} \mathbf{E}'$$

where: $\mathbf{H}^{(s)} = \sum_{m=1}^M (\mathbf{c}_m \mathbf{c}_m' + \mathbf{d}_m \mathbf{d}_m')$ is symmetric, and $\mathbf{H}^{(k)} = \sum_{m=1}^M (\mathbf{c}_m \mathbf{d}_m' - \mathbf{d}_m \mathbf{c}_m')$ is skew. In vector notation

$$\mathbf{X} \approx \gamma \sum_{j \geq i} h_{ij}^{(s)} \mathbf{e}_i \mathbf{e}_j' + \delta \sum_{j > i} h_{ij}^{(k)} (\mathbf{e}_i \mathbf{e}_j' - \mathbf{e}_j \mathbf{e}_i'), \quad i, j = 1, 2, \dots, p.$$

The previous equation shows that the constrained representation of symmetry and skew-symmetry is a linear combination of the symmetry and skew-symmetry of the different planes determined by each pair of external variables.

3. Application to Morse code confusion data

Rothkopf (1957) studied confusion rates on 36 Morse code signals (26 for the alphabet, 10 for the digits 0, ..., 9). A Morse code signal is a sequence of up to five tones represented by dots · (a short beep of 0.05 sec.) and dashes – (a long beep of 0.15 sec.) separated by a silence of 0.05 sec. (e.g. the signal for A is short-silence-long, or, symbolically, · –). 598 subjects, who did not know Morse code, were asked to judge whether two signals presented acoustically one after another (separated by a quiet period of 1.4 sec.) were the same or not. The entries in the proximity matrix (36×36) are the percentages (*rates*) of roughly 150 subjects who responded “same” for the two signals they heard. Each signal pair was presented in two orders, e.g. X following H (confusion rate 33%) and also H following X (confusion rate 6%). Symmetric MDS was performed on the symmetric part of the data by several authors and the obtained configuration could be partitioned by different criteria (e.g. number of elements in the signal, duration, type of composition, etc.). Even if the symmetric part of the Morse code data is dominant (96% of the total SSQ without the diagonal) many authors argued that asymmetry may still reveal interesting not random relations. Unconstrained GEG was applied in one bimension ($M=1$) with different values for the parameters γ and δ (i.e. Gower decomposition, EG and GIPSCAL models). For each model, constrained solutions were obtained by using three external variables: number of elements, number of short/long beep inversion, duration. The reduction in the fit of the constrained solutions showed that the three external variables can not provide a complete explanation of symmetry and skew-symmetry in the data, and more external variables have to be considered. Nevertheless it is interesting to compare, for example, the unconstrained and constrained solutions obtained for the GEG-Gipscale model in Figure 1 and Figure 2. The effect of the linear constraints can be detected in Figure 2, where points are aligned when they have similar duration (e.g. the line of symbols 5 H S I E). The distances along the line reflect the differences in the number of elements. The skew-symmetry in this bimension is in fact strongly correlated with skew-symmetry of the plane corresponding to the external variables: duration and number of elements. In particular, the skew-symmetry between X and H is represented by the area of the triangle determined by the two points with the origin.

4. Conclusions

The proposed extension of GEG allows to analyse the influence of external variables on the relationships (symmetry and skew-symmetry) contained in a square matrix. Future research could consider other type of constraints (e.g. non linear transformations of the external variables).

Figure 1: *GEG-Gipscal representation of Morse code data*

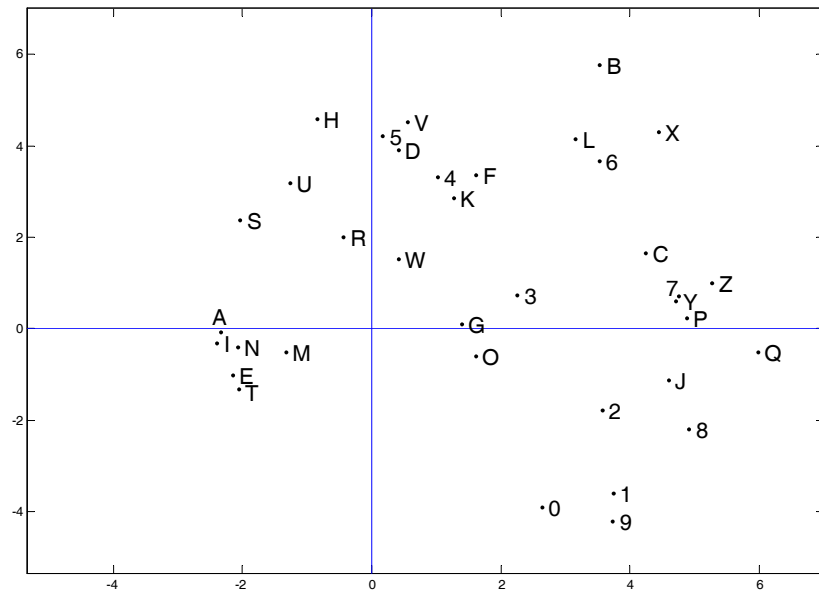
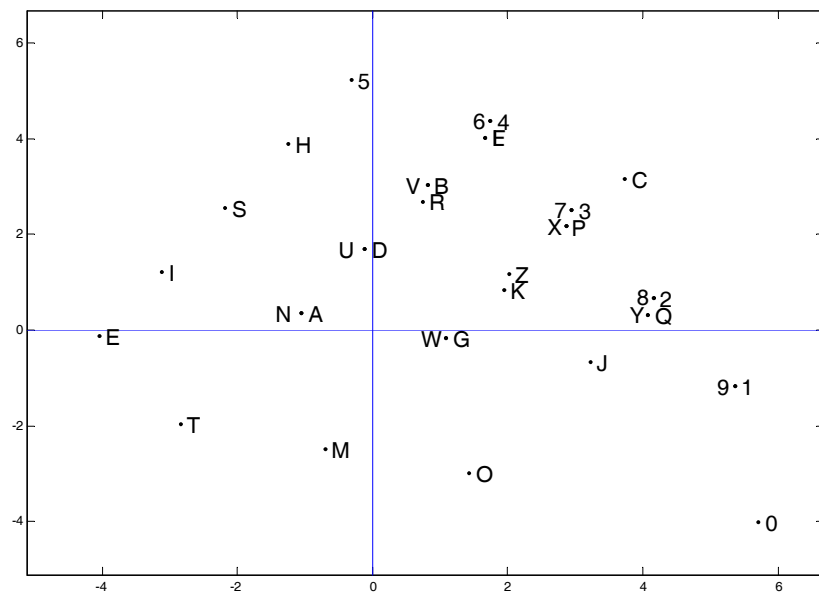


Figure 2: *Constrained GEG-Gipscal representation of Morse code data*



References

- Escoufier Y., Grorud A. (1980) Analyse factorielle des matrices carrees non symetriques, *Data Anal. & Informatics*, Diday E. et al. (eds.), N.-Holland, 263-276.
- Rocci R., Bove G. (2002) Rotation techniques in asymmetric multidimensional scaling, *Journal of Computational and Graphical Statistics*, 11, 405-419.
- Rothkopf E.Z. (1957) A measure of stimulus similarity and errors in some paired associated learning, *Journal of Experimental Psychology*, 53, 94-101.
- Zielman B., Heiser W. J. (1996) Models for asymmetric proximities, *British Journal of Mathematical and Statistical Psychology*, 49, 127-146.