Direct Individual Differences Scaling for Evaluation of Research Quality

DINDSCAL per la Valutazione della qualità della Ricerca

Gallo M., Trendafilov N., and Simonacci V.

Abstract *The eValuation of Research Quality (VQR) is one of the most important assessment processes achieved by the National Agency for the Evaluation of Universities and Research Institutes. Its main task is to provide information on the status of the Italian research system by assessing the performance of universities in various scientific areas. The basic evaluation criteria were defined by panels of experts according to the specific characteristics of each subject area and through a synthetic statement on the products submitted by researchers. With the aim of studying this phenomenon in depth, DINDSCAL (Direct Individual Differences Scaling) model is proposed for a compositional analysis of VQR dataset.*

Abstract *La Valutazione della Qualità della Ricerca (VQR) è uno dei processi di valutazione più importanti realizzati dall'Agenzia Nazionale di Valutazione del Sistema Universitario. Il suo compito principale è fornire informazioni sullo stato del sistema di ricerca italiano valutando le prestazioni delle universitın varie aree scientifiche. I criteri di valutazione di base sono stati definiti da gruppi di esperti in base alle caratteristiche specifiche di ciascuna area tematica e attraverso una dichiarazione sintetica sui prodotti presentati dai ricercatori. Con l'obiettivo di studiare approfonditamente tale fenomeno, il modello DINDSCAL è proposto per l'analisi composizionale dei dati VQR.*

Key words: compositional data, log-ratios, DINDSCAL, Stiefel mainifold, VQR data.

Michele Gallo
Department of Human and Social Sciences, University of Naples "L'Orientale", Italy.
e-mail: mgallo@unior.it

Nickolay Trendafilov
School of Mathematics and Statistics, Open University, UK.
e-mail: nickolay.trendafilov@open.ac.uk

Violetta Simonacci
Department of Human and Social Sciences, University of Naples "L'Orientale", Italy.
e-mail: mgallo@unior.it

1 Introduction

The National Agency for the Evaluation of the University and Research Systems (ANVUR), in the framework of an evaluation project (VQR), collected research outputs from 96 Italian universities, including 18 research Institutes. Sixteen panel of experts in evaluation (GEV), one for each scientific area of research, classified the products in specific merit classes. According to Ministerial Decree no. 458 dated 27 June 2015, common guidelines were defined for all GEVs. A new approach based on DINDSCAL (Direct Individual Differences Scaling [2]) model and compositional analysis is proposed in this work to extract information regarding the criteria used by the GEV, if the dimension of structure and the geographic location influence the quality of research outputs. In literature the INDSCAL (Individual Differences Scaling) model is used to study the individual differences in three-way data by doubly centered set of matrices of squared dissimilarity measures between a range of stimuli [1]. A direct approach as DINDSCAL is here preferred, in order to directly analyse simultaneous slices of squared dissimilarity matrices organized as compositional data.
The new approach is shortly described in Sect. 2. Sect. 3 summarizes the analysis of the VQR data.

2 Theory

2.1 Compositional data

Let $\underline{V}$ $(n \times p \times m)$ be a three-way array where each row v_{ik} $(i = 1, \cdots, n, k = 1, \cdots, m)$ is a compositional vector of p parts observed after the kth treatment (or occasion). From a geometrical point of view the sample space for all vectors v_{ik} is the simplex. There is a rich literature for CoDa proprieties and how is possible to handle them in simplex space (for a detailed review and references see [3]).
Here the CoDa are transformed in centred log-ratio (clr) in order to move from simplex to real space [4]. Let $\underline{L}$ $(n \times p \times m)$ be a three-way array with the CoDa in logarithm scale $[l_{ijk} = log(v_{ijk})]$. The clr-coordinates for each frontal slice of $\underline{X}$ are defined $X_k = L_k J_p$, where L_k is a $n \times p$ matrix and J_p is a $p \times p$ centring matrix, $J_p = I_p - \frac{1}{p} E_p$ with I is an identity matrix and E is a matrix of ones. Thus, when the columns of X_k are centred, $X_k = J_n L_k J_p$, the metric multidimensional scaling (MDS) for each X_k is given by the following identity:

$$-\frac{1}{2} J_n (D_k \odot D_k) J_n = X_k X_k^\top, \tag{1}$$

where '$\odot$' denotes the usual elementwize (Hadamard) matrix product and D_k is a $n \times n$ symmetric matrix containing zero on main diagonal and the dissimilarity measure between the n compositions collected at the kth occasion.

As well as recalled, the measures in D_k are Aitchison distances. Thus, they have all properties necessary for a meaningful interpretation of compositional results. Moreover, the identity (1) shows that $-\frac{1}{2}J_n(D_k \odot D_k)J_n$ is positive semi-definite.
With the aim to a simultaneous metric of m symmetric slices D_k, the INDSCAL model decomposes the each slice as

$$-\frac{1}{2}J_n(D_k \odot D_k)J_n = Q\Lambda_k Q^\top + \Delta_k, \tag{2}$$

where Q is $n \times r$ assumed of full column rank, Λ_k diagonal matrix and Δ_k is $n \times n$ matrix, containing the errors of the model fit. In other words all slices share a common loading matrix Q and differ each other only by the (non-negative) diagonal elements of Λ_k called idiosyncratic saliences.
Unfortunately, in (2) the parameter set of all $n \times r$ matrices Q with full column rank is a non-compact Stiefel manifold. To solve this drawback an approach called direct INDSCAL was proposed by [2].

2.2 DINDSCAL problem

Following the approach proposed by [5], it is easy to show that the squared Aitchison distances are given by the following identity:

$$D_k \odot D_k = (I_n \odot X_k X_k^\top)E_n + E_n(I_n \odot X_k X_k^\top) - 2X_k X_k^\top \quad k = 1, ..., m. \tag{3}$$

Thus, the DINDSCAL fitting problem for CoDa is concerned with the following equality constrained optimization problem:

$$\min_{Q,\Lambda_k} \sum_{k=1}^{m} \| D_k \odot D_k - (I_n \odot Q\Lambda_k^2 Q^\top)E_n - E_n(I_n \odot Q\Lambda_k^2 Q^\top) + 2Q\Lambda_k^2 Q^\top \| , \tag{4}$$

subject to $(Q, \Lambda_1, \Lambda_2, \ldots, \Lambda_m) \in \mathscr{O}_0(n,r) \times \mathscr{D}(r)^m$, where $\mathscr{D}(r)^m = \underbrace{\mathscr{D}(r) \times \ldots \times \mathscr{D}(r)}_{m}$, and $\mathscr{D}(r)$ denotes the set of all $r \times r$ diagonal matrices. $\mathscr{O}_0(n,r)$ denotes the set of all $n \times r$ orthonormal matrices with zero column-sums, i.e.: $\mathrm{O}_0(n,r) := \{Q \in \Re^{n \times r} |\ Q^\top Q = I_r \text{ and } E_n Q = 0_n\}$. It is easy to observe that the additional constraint of clr-transformation zero column- rows sums, i.e. $E_n X_k = 0_n$ and $X_k E_p = 0_p$ does not introduce new constrains. Thus the problem of minima defined in (4) can be solved by a non-linear conjugate gradient algorithm, which leads to globally convergent algorithms.
For a formal introduction to the method and its numerical integration see [6], while all information about the gradient dynamical system for the DINDSCAL problem is given in [2].

3 Case study

The research outputs submitted from each university was calculated considering the number of university staff members. Each product was classified in a specific merit classes, that is, Excellent, Good, Fair, Acceptable, and Limited. Products classified as not eligible are assigned to a specific merit class. According to the kind of research outputs (articles, monographs, book chapters, etc.) bibliometric algorithm or peer-review methodology were used to evaluate the 117,079 research outputs submitted in the 16 MIUR scientific areas.
The data are preprocessed according to the procedure described in Sect. 2.1, DINDSCAL is able to find some important differences between two groups of scientific areas and between the scientific structures with different size. In short, there is a tangible contraposition between large-size scientific structures, especially those localized in the Centre of Italy, and the others. Moreover, the medium size structures localized in the Centre and in the South of Italy are characterized only by specific scientific area.

References

[1] Carroll, J.D., Chang J.J.: Analysis of individual differences in multidimensional scaling via an $n-$way generalization of "Eckart-Young" decomposition. *Psychometrika,* 35, 283–319 (1970).
[2] Trendafilov, N.T.: Dindscal: direct INDSCAL. *Statistics and Computing*, 22(2), 445-454 (2012).
[3] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: *Modeling and analysis of compositional data.* John Wiley & Sons (2015).
[4] Gallo, M.: Tucker3 model for compositional data. *Communications in Statistics-Theory and Methods*, pp 4441-4453 (2015).
[5] Browne, M.W.: The Young-Householder algorithm and the least squares multidimensional scaling of squared distances, *Journal of Classification*, 4, 175–219 (1987).
[6] Trendafilov, N.T.: The dynamical system approach to multivariate data analysis, a review, *Journal of Computational and Graphical Statistics*, 15, 628–650 (2006).