An INDCLUS-type model for occasion-specific complementary partitions

*Un modello INDCLUS per partizioni complementari specifiche per occasione*

Laura Bocci and Donatella Vicari[[1]](#footnote-1)

**Abstract** This paper presents an INDCLUS-type model for partitioning the units in three-way proximity data taking into account the systematic differences among the occasions. Specifically, the proximity structure of each occasion is assumed to underlie two complementary partitions: the first, common to all occasions, defines a partitioning of a subset of units and the second, occasion-specific, defines a partitioning of the remaining units. The model is fitted in a least-squares framework and an efficient ALS algorithm is given.

**Abstract** *Si presenta un modello di tipo INDCLUS per partizionare le unità nel caso di dati di prossimità a tre vie tenendo conto delle differenze sistematiche esistenti tra le similarità a coppie rilevate in diverse occasioni. In particolare, si assume che la struttura di prossimità di ciascuna occasione si componga di due partizioni complementari: un sottogruppo di unità definisce dei gruppi comuni a tutte le occasioni, mentre le rimanenti unità sono allocate a dei gruppi specifici per ogni occasione. Il modello, formulato seguendo l’approccio dei minimi quadrati, è stimato utilizzando un algoritmo ALS.*

**Key words:** three-way proximity data, INDCLUS, clustering

1. Introduction

In many research domains, there is an increasing interest in how the perception or evaluation of several units (objects or stimuli) differs across several sources or occasions such as subjects, experimental conditions, situations, scenarios or times.

Such kind of information can be arranged in three-way data represented as a cube having multiple data matrices stacked as slices along the third dimension which represents the data sources. Our concern here is three-way two-mode data, so that the frontal slices consist of square symmetric matrices **S***h* () of pairwise proximities of a set of *N* units coming from *H* occasions.

Clustering three-way data is a complex task since each proximity data matrix **S***h* generally subsumes a (more or less) different clustering of units due to the occasion heterogeneity. Therefore, a three-way clustering should represent a consensus classification of the ones obtained from each occasion, but such a consensus is actually representative only when the classifications from each frontal slice of the three-way data do not contain systematic differences.

However, since this hypothesis is often not appropriate in real situations, several methodologies for clustering three-way proximity data have been proposed both in a least-squares (Carroll and Arabie (1983), Gordon and Vichi (1998), Vichi (1999), Vicari and Vichi (2009), Giordani and Kiers (2012), Bocci and Vicari (2017)) and in a maximum likelihood approach (Basford and McLachlan (1985), Hunt and Basford (1999), Bocci, Vicari and Vichi (2006)).

Carroll and Arabie (1983) introduced the INDCLUS (INdividual Differences CLUStering) model to extract (overlapping) clusters of *N* units from a set of proximity matrices provided in *H* occasions which are supposed to differently weigh each group of units. Therefore, all occasions in the INDCLUS model are assumed to employ the same clustering, but with different patterns of weights indicating the effect of each cluster on the proximities.

A fuzzy variant of the INDCLUS model, called FINDCLUS, has been proposed by Giordani and Kiers (2012), where a fuzzy membership matrix, still common to all occasions, is assigned to the classification of units.

Bocci and Vicari (2017) proposed a Generalized INDCLUS model, called GINDCLUS, for clustering both units and occasions incorporating (possible) information from external variables not directly involved in the collection process of the similarities.

In the context of clustering and multidimensional scaling for three-way data, the CLUSCALE (simultaneous CLUstering and SCAL[E]ing) model (Chaturvedi and Carroll, (2006)) combines additively INDCLUS and INDSCAL (Carrol and Chang, (1970)) searching for a common both (discrete) clustering representation and continuous spatial configuration of the units.

In order to extract more information from the three-way proximities accounting for and taking advantage of the heterogeneity of the occasions, we present an INDCLUS-type model where the pairwise proximities subsume two complementary partitions of the units: one is common to all occasions while the other is occasion-specific. Therefore, we assume that a subset of the *N* units defines a partition shared by all occasions, while the remaining units are differently clustered in each occasion. The model is formalized in a least-squares framework and an appropriate ALS algorithm is given.

1. The model

In order to present our INDCLUS-type model, let’s formally recall the INDCLUS model (Carroll and Arabie, (1983))

, *,* (1)

where ( for and ) is a *N* × *J* binary matrix defining the clustering of *N* units, is a non-negative diagonal weight matrix of order *J* for occasion *h*, is a real-valued additive constant for occasion *h*, **1***N* denotes the column vector with *N* ones and **E***h* is the square matrix of errors which represents the part of **S***h* not accounted for by the model.

On the other hand, when the occasions present systematic differences, it is reasonable to think that a subset of the *N* units shares the same partition and weights in all occasions but for each occasion there exists a different partitioning and weighing of the remaining units. This assumption specifies a new model

, *,* (2)

where ( for and ) is an *N* × *J* binary matrix defining the common partition of () units in *J* groups, ( for and ) is an *N* × *J* binary matrix defining the partition of units in *J* groups for occasion *h* (, is a non-negative diagonal weight matrix of order *J*, is a non-negative diagonal matrix of order *J* containing the *differential* weights for occasion *h* (

Therefore, on one hand, we suppose that there exists a subset of the *N* units whose pairwise proximities for each occasion subsume the same partition **P** and the same weights **W** for its clusters; on the other hand, the partition of the remaining units is supposed to be different for each occasion, so that *H* occasion-specific complementary partition **M***h* () can be recognised.

Within this framework, the *i*-th unit () at occasion *h* () is assigned to one of the *J* groups in either **P** or (): if for some then for all , while if for all then for some .

Therefore, conditionally on occasion *h*, the binary membership matrices **P** and **M***h* () specify pairs of *incomplete and complementary* partitions which together define the occasion-specific clustering structure of the *N* units in *J* groups. Note that () is a membership matrix defining a *complete* occasion-specific partition where each group is the union of the corresponding groups from **P** and **M***h* ().

The *J* groups identified in **P** can be considered as *core* clusters in the sense that they represent the part common to all occasions, while the *complementary* partitions identified in **M***h* () capture the heterogeneity of the occasions.

Moreover, the *H* clustering structures () are weighted by two patterns of weights: **W** which is common to all occasions and **R***h* which is occasion-specific and expresses how the *complementary* partition **M***h* is *differently* weighted with respect to the *core* partition **P** in occasion *h*.

1. The algorithm

In model (2), the incomplete membership matrices **P** and **M***h* (), the weight matrices **W** and **R***h* () and constants *ch* () can be estimated by solving the following least-squares fitting problem:

 (3)

subject to

 (; ) and (), (3a)

 (; ; ) and

 (; ), (3b)

if then (), (3c)

if then (), (3d)

 (), (3e)

 (; ). (3f)

For the sake of clarity, the set of constraints (3a)-(3d) specify that in the partitions defined by **P** and **M***h* () each units is assigned to only one clusters in either **P** or **M***h* (), while constraints (3e)-(3f) specify the non-negativity of the weights **W** and **R***h* (), respectively.

Problem (3) can be solved using an appropriate coordinate descent algorithm also known as Alternating Least-Squares (ALS) algorithm, which alternates the estimation of a set of parameters when all the other are fixed.

The algorithm proposed here estimates in turn:

1. the *core* membership matrix **P** by sequentially solving assignment problems for the different rows of **P**, where each unit can possibly remain not assigned;
2. the occasion-specific membership matrix **M***h* () by sequentially solving *h* () assignment problems for the rows of **M***h* corresponding to the *NI* units not assigned in **P**;
3. the common weight matrix **W** by solving a regression problem using nonnegative least squares (Lawson and Hanson, 1974);
4. the occasion-specific weight matrix **R***h* () by solving *h* () regression problems using nonnegative least squares (Lawson and Hanson, 1974);
5. the additive constant *ch* () by successive residualizations of the three-way data matrix.

The five main steps are alternated and iterated until convergence and the best solution over different random starting classification matrices is retained to prevent from local minima.

Results from applications to real and artificial data will be presented to show the performance of the algorithm and the capability of the model to capture the heterogeneity of the occasions.

References

1. Basford, K.E., McLachlan, G.J.: The Mixture Method of Clustering Applied to Three-way Data. J. Classif. **2**, 109--125 (1985).
2. Bocci, L., Vicari, D., Vichi, M.: A mixture model for the classification of three-way proximity data. Comput. Stat. Data An. **50**, 1625--1654 (2006).
3. Bocci, L., Vicari, D: GINDCLUS: Generalized INDCLUS with external information. Psychometrika **82**, 355--381 (2017).
4. Carroll, J.D., Arabie, P.: INDCLUS: an Individual Differences Generalization of ADCLUS model and the MAPCLUS algorithm. Psychometrika **48**, 157--169 (1983).
5. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-generalization of the Eckart-Young decomposition. Psychometrika **35***,* 283--319 (1970).
6. Chaturvedi, A., Carroll, J.D.: CLUSCALE (CLUstering and multidimensional SCAL[E]ing): a three-way hybrid model incorporating clustering and multidimensional scaling structure. J. Classif. **23**, 269--299 (2006).
7. Giordani, P., Kiers, H.A.L.: FINDCLUS: Fuzzy INdividual Differences CLUStering. J. Classif. **29**, 170--198 (2012).
8. Gordon, A.D., Vichi, M.: Partitions of Partitions. J. Classif. **15**, 265--285 (1998).
9. Hunt, L.A., Basford, K.E.: Fitting a Mixture Model to Three-mode Three-way Data with Categorical and Continuous Variables. J. Classif. **16**, 283--296 (1999).
10. Lawson, C.L., Hanson, R.J.: Solving least squares problems. Prentice-Hall, Englewood Cliffs, NJ (1974).
11. Vichi, M.: One Mode Classification of a Three-way Data Set. J. Classif. **16**, 27--44 (1999).
12. Vicari, D., Vichi, M.: Structural Classification Analysis of Three-Way Dissimilarity Data. J. Classif. **26**, 121--154 (2009).
1. Laura Bocci, Department of Communication and Social Sciences, Sapienza University of Rome; email: laura.bocci@uniroma1.it

 Donatella Vicari, Department of Statistical Sciences, Sapienza University of Rome; email: donatella.vicari@uniroma1.it [↑](#footnote-ref-1)