A CLUSTERING MODEL FOR THREE-WAY ASYMMETRIC PROXIMITY DATA

Laura Bocci[1] and Donatella Vicari[2]

[1] Department of Social and Economic Sciences, Sapienza University of Rome
(e-mail: laura.bocci@uniroma1.it)

[2] Department of Statistical Sciences, Sapienza University of Rome
(e-mail: donatella.vicari@uniroma1.it)

ABSTRACT: This paper presents a model for clustering three-way asymmetric proximity data which represent flows or exchanges between objects observed at different occasions. In order to account for systematic differences between occasions, the asymmetric data are assumed to subsume two clustering structures common to all occasions: the first defines a standard partitioning of all objects which fits the average amount of the exchanges; the second one, which fits the imbalances, defines an "incomplete" partitioning of the objects, where some of them are allowed to remain unassigned. The model is fitted in a least-squares framework and an efficient Alternating Least Squares algorithm is given.

KEYWORDS: Asymmetric dissimilarities, three-way data, partition.

1 Introduction

In many real-world applications, information is measured or observed in the form of several pairwise asymmetric proximity (similarity or dissimilarity) matrices related to the same N objects observed at H occasions (i.e., times, subjects, scenarios). Such kind of data represent three-way two-mode asymmetric proximity data which, without loss of generality, can derive from mobility flows, brand-switching, import/export exchanges or other type of transactions or trade. For example, international student mobility between countries over several years gives rise to a three-way asymmetric proximity array where in each year-matrix the rows correspond to the origins and the columns to the destinations of the mobile students.

In the analysis of asymmetric data, the asymmetry has often been ignored by symmetrizing the proximities (i.e., averaging the two different values for any pair of objects). Nonetheless, if one hypothesizes that the asymmetries are meaningful and systematic across occasions, special models are needed (see Saito & Yadohisa, 2005, Bove et al., 2021, for extended reviews).

Clustering three-way asymmetric proximity data is a complex task since each proximity data matrix generally subsumes a (more or less) different clustering of objects due to the heterogeneity of the occasions and the asymmetry may incorporate some important information about clustering. Chaturvedi and Carroll (1994) generalized the INDCLUS model to asymmetric proximities by identifying two

different sets of (overlapping) clusters of the *N* objects (for rows and columns, respectively) common to all occasions, while the three-way heterogeneity is accounted for by occasion-specific weights for the clusters.

In order to extract as much information as possible from the three-way asymmetries taking into account the heterogeneity of the occasions, we present here a generalization to asymmetric three-way data of the model proposed by Vicari (2020) for clustering an asymmetric dissimilarity matrix. To account for the asymmetric structure of the data, the model relies on the decomposition of the asymmetric matrices into the sum of their symmetric and skew-symmetric components which are jointly modelled. The asymmetric dissimilarities are assumed to subsume two clustering structures common to all occasions: the first defines a standard partitioning of all objects which fits the symmetric component of the exchanges; the second one, which fits the imbalances, defines an *incomplete* partitioning of the objects, where some of them are allowed to remain unassigned. Objects within the same clusters in both clustering structures share the same behaviours in terms of exchanges directed to the other clusters and identify "origin" and "destination" clusters. Note that the partition to fit the imbalances is allowed to be incomplete to better identify the directions of the exchanges, so those objects not assigned to any cluster (incomplete partitioning) qualify objects with "small" asymmetries. Moreover, to account for the heterogeneity of the occasions, occasion-specific sets of weights are estimated which account for both the average amounts and the directions of the exchanges.

In Section 2, the model is formalized in a least-squares framework and an appropriate Alternating Least Squares algorithm is given.

2 The model

Let $\mathbf{X}$ be a three-way two-mode asymmetric array of size $(N \times N \times H)$, where the H frontal slices consist of square asymmetric matrices $\mathbf{X}_h$ $(h = 1, \dots, H)$ of pairwise dissimilarities between N objects observed in H occasions and where the generic element x_{ijh} is generally different from x_{jih}.

The model proposed here aims at clustering the N objects by decomposing the observed asymmetries into symmetric and skew-symmetric effects, modelled as functions of two nested partitions of the objects which subsume clustering structures common to all occasions. Specifically, all occasions are supposed to share the same partition of the *N* objects into J disjoint clusters $\{C_1, \dots, C_J\}$ uniquely identified by an (*N*×*J*) binary membership matrix $\mathbf{U} = [u_{ij}]$ $(u_{ij} = \{0,1\}$ for $i = 1, \dots, N$ and $j = 1, \dots, J$ and $\sum_{j=1}^{J} u_{ij} = 1$ for $i = 1, \dots, N)$, where $u_{ij} = 1$ if object i belongs to cluster C_j and $u_{ij} = 0$ otherwise. Since any object is required to be assigned to some cluster C_j, such a partition is referred to as *complete partition*. Furthermore, a second partition of the *N* objects into *J* clusters $\{G_1, \dots, G_J\}$ common to all occasions is identified by an (*N*×*J*) binary membership matrix $\mathbf{V} = [v_{ij}]$ $(v = \{0,1\}$ for $i = 1, \dots, N$ and $j = 1, \dots, J)$, where any object i is allowed either not to be assigned to any cluster or to

belong to cluster G_j if it belongs to cluster C_j in the complete partition, i.e. $v_{ij} \leq u_{ij}$ ($i = 1, \dots, N$ and $j = 1, \dots, J$). The partition identified by $\mathbf{V}$ is referred to as an *incomplete partition* because a number N_0 ($N_0 \leq N$) out of N objects are allowed to remain unassigned to any cluster. Moreover, the *complete* and the *incomplete* partitions are *common* to all occasions and linked each other, the latter being constrained to be nested into the former one ($G_j \subseteq C_j$ for $j = 1, \dots, J$).

Hereafter, $\mathbf{I}_N$ denotes the identity matrix of size N, $\mathbf{1}_{AB}$ and $\mathbf{1}_A$ denote the matrix of size ($A \times B$) of all ones and the column vector with A ones, respectively.

Let us recall that any square matrix $\mathbf{X}_h$ ($h = 1, \dots, H$) can be uniquely decomposed into a sum of a symmetric matrix $\mathbf{S}_h$ and a skew-symmetric matrix $\mathbf{K}_h$, which are orthogonal to each other (i.e., $trace(\mathbf{S}_h\mathbf{K}_h) = 0$), as

$$\mathbf{X}_h = \mathbf{S}_h + \mathbf{K}_h = \tfrac{1}{2}(\mathbf{X}_h + \mathbf{X}_h') + \tfrac{1}{2}(\mathbf{X}_h - \mathbf{X}_h') \,, \qquad (h = 1, \dots, H). \tag{1}$$

Both components in $\mathbf{X}_h$ can be modeled by defining two clustering structures depending on matrices $\mathbf{U}$ and $\mathbf{V}$, respectively, as introduced in Vicari (2020) for a two-way asymmetric dissimilarity matrix.

Specifically, the symmetric component $\mathbf{S}_h$ and the skew-symmetric component $\mathbf{K}_h$ for occasion h are modeled by the two clustering structures introduced in Vicari (2014, 2018) and depend on the *common complete* membership matrix $\mathbf{U}$ and the *common incomplete* membership matrix $\mathbf{V}$, respectively, as

$$\mathbf{S}_h = \mathbf{U}\mathbf{C}_h\big(\mathbf{1}_{NJ} - \mathbf{U}\big)' + \big(\mathbf{1}_{NJ} - \mathbf{U}\big)\mathbf{C}_h\mathbf{U}' + \mathbf{E}_{hS} \,, \qquad (h = 1, \dots, H), \tag{2}$$

$$\mathbf{K}_h = \mathbf{V}\mathbf{D}_h\big(\mathbf{1}_{NJ} - \mathbf{V}\big)' - \big(\mathbf{1}_{NJ} - \mathbf{V}\big)\mathbf{D}_h\mathbf{V}' + \mathbf{E}_{hK} \,, \qquad (h = 1, \dots, H), \tag{3}$$

where $\mathbf{C}_h$ and $\mathbf{D}_h$ are ($J \times J$) occasion-specific diagonal weight matrices associated with the clusters of the complete and incomplete partition, respectively, and the error terms $\mathbf{E}_{hS}$ and $\mathbf{E}_{hK}$ represent the parts of $\mathbf{S}_h$ and $\mathbf{K}_h$ not accounted for by the model, respectively. For identifiability reasons, any matrix $\mathbf{V}\mathbf{D}_h$ is constrained to sum to zero: $\mathbf{1}_N'(\mathbf{V}\mathbf{D}_h)\mathbf{1}_J = 0$ ($h = 1, \dots, H$).

Models (2) and (3) can be combined in (1) to specify the model accounting for the asymmetric dissimilarities between clusters

$$\begin{aligned}\mathbf{X}_h = {} & \Big[\mathbf{U}\mathbf{C}_h\big(\mathbf{1}_{NJ} - \mathbf{U}\big)' + \big(\mathbf{1}_{NJ} - \mathbf{U}\big)\mathbf{C}_h\mathbf{U}'\Big] + \Big[\mathbf{V}\mathbf{D}_h\big(\mathbf{1}_{NJ} - \mathbf{V}\big)' - \big(\mathbf{1}_{NJ} - \mathbf{V}\big)\mathbf{D}_h\mathbf{V}'\Big] \\ & + b_h(\mathbf{1}_{NN} - \mathbf{I}_N) + \mathbf{E}_h \,, \quad (h = 1, \dots, H),\end{aligned} \tag{4}$$

where b_h is the additive constant term and the general error term $\mathbf{E}_h$ represents the part of $\mathbf{X}_h$ not accounted for by the model.

It is worth noting that all occasions are assumed here to share the same clustering structure but with different patterns of weights which account for the heterogeneity of the occasions. In fact, the occasion-specific diagonal entries of $\mathbf{C}_h$ and $\mathbf{D}_h$ provide quantifications of the exchanges between clusters in terms of amounts and directions and allow to measure at what extent the exchanges vary across occasions.

In model (4), the *complete* and the *incomplete* membership matrices **U** and **V**, the weight matrices $\mathbf{C}_h$ and $\mathbf{D}_h$ $(h = 1, \dots, H)$ and the constants b_h $(h = 1, \dots, H)$ can be estimated by solving the following least-squares fitting problem:

$$\min F(\mathbf{U}, \mathbf{V}, \mathbf{C}_h, \mathbf{D}_h, b_h) = \sum_{h=1}^{H} \left\| \mathbf{X}_h - \left[\mathbf{U}\mathbf{C}_h \left(\mathbf{1}_{NJ} - \mathbf{U} \right)' + \left(\mathbf{1}_{NJ} - \mathbf{U} \right) \mathbf{C}_h \mathbf{U}' \right] - \left[\mathbf{V}\mathbf{D}_h \left(\mathbf{1}_{NJ} - \mathbf{V} \right)' - \left(\mathbf{1}_{NJ} - \mathbf{V} \right) \mathbf{D}_h \mathbf{V}' \right] - b_h (\mathbf{1}_{NN} - \mathbf{I}_N) \right\|^2 \quad (5)$$

subject to

$$u_{ij} = \{0,1\} \quad (i = 1, \dots, N; j = 1, \dots, J) \text{ and } \sum_{j=1}^{J} p_{ij} = 1 \quad (i = 1, \dots, N), \quad (5a)$$

$$v_{ij} = \{0,1\} \quad (i = 1, \dots, N; j = 1, \dots, J) \text{ and } v_{ij} \leq u_{ij} \quad (i = 1, \dots, N), \quad (5b)$$

$$\mathbf{1}_N' (\mathbf{V}\mathbf{D}_h) \mathbf{1}_J = 0 \quad (h = 1, \dots, H). \quad (5c)$$

Problem (5) can be solved by using an Alternating Least-Squares algorithm which alternates the estimation of a set of parameters when all the others are kept fixed. The algorithm proposed here estimates in turn: a) the *complete* and *incomplete* membership matrices **U** and **V** by sequentially solving joint assignment problems for the different rows of **U** and **V**: given any row *i*, setting $u_{ij} = 1$ implies that either $v_{ij} = 0$ or $v_{ij} = u_{ij}$ for $j = 1, \dots, J$; b) the occasion-specific weight matrices $\mathbf{C}_h$ and $\mathbf{D}_h$ $(h = 1, \dots, H)$ by solving regression problems; c) the additive constant b_h $(h = 1, \dots, H)$ by successive residualizations of the three-way data matrix. The main steps are alternated and iterated until convergence and the best solution over different random starts is retained to prevent from local minima.

Results from applications to real data will be presented to show the performance of the algorithm and the capability of the model to identify common clusters of objects which best account for their pairwise dissimilarities.

References

BOVE, G., OKADA, A., & VICARI, D. 2021. *Methods for the Analysis of Asymmetric Proximity Data*. Springer Nature Singapore.

CHATURVEDI, A., & CARROLL, J.D. 1994. An alternating combinatorial optimization approach to fitting the INDCLUS and Generalized INDCLUS models. *Journal of Classification*, **11**, 155–170.

SAITO, T., & YADOHISA, H. 2005. *Data analysis of asymmetric structures. Advanced approaches in computational statistics*. New York: Marcel Dekker.

VICARI, D. 2014. Classification of asymmetric proximity data. *Journal of Classification*, **31**(3), 386–420.

VICARI, D. 2018. CLUXEXT: CLUstering model for Skew-symmetric data including EXTernal information. *Adv Data Anal Classif*, **12**, 43–64.

VICARI, D. 2020. Modeling Asymmetric Exchanges Between Clusters In: T. Imaizumi, A. Nakayama and S. Yokoyama (Eds), Advanced Studies in Behaviormetrics and Data Science. Springer Nature Singapore, 297-313.