Three-way Spectral Clustering

Cinzia Di Nuzzo and Salvatore Ingrassia

Abstract In this paper, we present a spectral clustering approach for clustering *three-way data*. Three-way data concern data characterized by three modes: n units, p variables, and t different occasions. In other words, three-way data contain a $t \times p$ observed matrix for each statistical observation. The units generated by simultaneous observation of variables in different contexts are usually structured as three-way data, so each unit is basically represented as a matrix. In order to cluster the n units in K groups, the spectral clustering application to three-way data can be a powerful tool for unsupervised classification. Here, one example on real three-way data have been presented showing that spectral clustering method is a competitive method to cluster this type of data.

Keywords: spectral clustering, kernel function, three-way data

1 Introduction

Spectral clustering methods are based on the graph theory, where the units are represented by the vertices of an undirected graph and the edges are weighted by the pairwise similarities coming from a suitable kernel function, so the clustering problem is reformulated as a graph partition problem, see e.g. [16, 6]. The spectral clustering algorithm is a very powerful method for finding non-convex clusters of data, moreover, it is a handy approach for handling high-dimensional data since it works on a transformation of the raw data having a smaller dimension than the space of the original data.

Cinzia Di Nuzzo (✉)
Department of Statistics, University of Roma La Sapienza, Piazzale Aldo Moro, 5, 00185 Roma, Italy, e-mail: `cinzia.dinuzzo@uniroma1.it`

Salvatore Ingrassia
Department of Economics and Business, University of Catania, Piazza Università, 2, 95131 Catania, Italy, e-mail: `s.ingrassia@unict.it`

Three-way data derives from the observation of various attributes measured on a set of units in different situations; some examples are longitudinal data on multiple response variables and multivariate spatial data. Three-way data can also derive from temporal measurements of a feature vector, thus having the dataset composed of three modes: n units (matrices), p variables (columns), and t times (rows). Clustering of three-way data has attracted a growing interest in literature, see e.g. [14], [1]; model-based clustering of three-way data has been introduced by [15] in the framework of matrix-variate normal mixtures; recent papers include [9] handle on parsimonious models for modeling matrix data; [11] introduce two matrix-variate distributions, both the elliptical heavy-tailed generalization of the matrix-variate normal distribution; [12] deal with three-way data clustering using matrix-variate cluster-weighted models (MV-CWM); and, [13] consider an application to educational data via mixtures of parsimonious matrix-normal distribution.

In this paper, we present a spectral clustering approach for clustering *three-way data* and a suitable kernel function between matrices is introduced. As a matter of fact, the data matrices represent the vertices of the graph, consequently, the edges must be weighted by a single value.

The rest of the paper is organized as follows: in Section 2 the spectral clustering method is summarized; in Section 3 a method to select the parameters in the spectral clustering algorithm is described; in Section 4 the three-way spectral clustering with a new kernel function are introduced; in Section 5 an application based on real three-way data is presented. Finally, in Section 5 we provide concluding remarks.

2 Spectral Clustering

Spectral clustering algorithm for two-way data has been described in [8, 16, 6]. Here, we summarize the main step of this algorithm.

Let $V = \{x_1, x_2, \ldots, x_n\}$ be a set of points in $X \subseteq \mathbb{R}^p$. In order to group the data V in K cluster, the first step concerns the definition of a symmetric and continuous function $\kappa : X \times X \rightarrow [0, \infty)$ called the *kernel function*. Afterwards, a *similarity matrix* $W = (w_{ij})$ can be assigned by setting $w_{ij} = \kappa(x_i, x_j) \geq 0$, for $x_i, x_j \in X$. and finally the *normalized graph Laplacian* matrix $L_{\text{sym}} \in \mathbb{R}^{n \times n}$ is introduced

$$L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}, \tag{1}$$

where $D = \text{diag}(d_1, d_2, \ldots, d_n)$ is the *degree matrix* and d_i is the *degree* of the vertex x_i defined as $d_i = \sum_{j \neq i} w_{ij}$ and I denotes the $n \times n$ identity matrix. The Laplacian matrix L_{sym} is positive semi-definite with n non-negative eigenvalues. For a fixed $K \ll n$, let $\{\gamma_1, \ldots, \gamma_K\}$ be the eigenvectors corresponding to the smallest K eigenvalues of L_{sym}. Then, the *normalized Laplacian embedding in the K principal subspace* is defined as the map $\Phi_{\Gamma} : \{x_1, \ldots, x_n\} \rightarrow \mathbb{R}^K$ given by

$$\Phi_{\Gamma}(x_i) = (\gamma_{1i}, \ldots, \gamma_{Ki}), \quad i = 1, \ldots, n,$$

where $\gamma_{1i}, \ldots, \gamma_{Ki}$ are the i-th components of $\gamma_1, \ldots, \gamma_K$, respectively. In other words, the function $\Phi_{\Gamma}(\cdot)$ maps the data from the input space $\mathcal{X}$ to a feature space defined by the K principal subspace of L_{sym}. Afterwards, let $Y = (y'_1, \ldots, y'_n)$ be the $n \times K$ matrix given by the embedded data in the feature space, where $y_i = \Phi_{\Gamma}(x_i)$ for $i = 1, \ldots, n$. Finally, the embedded data Y are clustered according to some clustering procedure; usually, the k-means algorithm is taken into account in literature. However, to this end Gaussian mixtures have been proposed because they yield elliptical cluster shapes, i.e. more flexible cluster shapes with respect to the k-means, see [2]. Finally, we point out that the performances of other mixture models based on non-Gaussian component densities have been analyzed, but Gaussian mixture models can be considered as a good trade-off between model simplicity and effectiveness, see [3] for details.

3 A Graphical Approach for Parameter Selection

According to spectral clustering algorithm introduced in Section 2, the spectral approach requires to set: i) the number of clusters K, ii) the kernel function κ (with the corresponding parameter). In order to select these quantities, in the following we summarize the method proposed in [4].

To begin with, we point out that the choice of the kernel function affects the entire data structure in the graph, and consequently, the structure of the Laplacian matrix and its eigenvectors. An optimal kernel function should lead to a similarity matrix W having (as much as possible) diagonal blocks: in this case, we get well-separated groups and we are also able to understand the number of groups in that data set by counting the number of blocks. For the sake of simplicity, we consider here the self-tuning kernel introduced by [17]

$$\kappa(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\epsilon_i \epsilon_j}\right) \tag{2}$$

with $\epsilon_i = \|x_i - x_h\|$, where x_h is the h-th neighbor of point x_i (similarly for ϵ_j). This function allow to get a similarity matrix that does not depend on any parameter so that the algorithm of spectral clustering will be based on the pairwise proximity between units. On the contrary, we need to select the h-th neighbor of the unit in (2).

The main novelty of the joint-graphical approach concerns the analysis of some graphic features of the Laplacian matrix including the shape of the embedded space. Indeed, the embedded data provide useful information for the clustering, in particular the main results in [10] and [5] allow to deduce that if the embedded data assume a cones structure, then the number of clusters is equal to the number of the cones/spikes in the feature space; furthermore, a clearer clustering structure emerges when the spikes are narrower and well separated.

The idea behind the graphical approach is to select the number K of groups and the parameter h in the kernel function from a joint analysis of three main characteristics: the plot of the Laplacian matrix; the maxima values of the eigengaps between two

consecutive eigenvalues; the scatter plot of the mapped data in the feature space and in particular the number of spikes counted in the embedded data space.

We remark that we cannot analyze all possible values of $h \in \{1, 2, \ldots, n-1\}$ and hence we choose a suitable subset $\mathcal{H} \subset \{1, 2, \ldots, n-1\}$, in particular we choose $\mathcal{H} = \{1\%, 2\%, 5\%, 10\%, 15\%, 20\%\} \times n \subset \{1, 2, \ldots, n-1\}$, and select $h \in \mathcal{H}$, see the following procedure for details.

Parameter selection (K and h)

Input: data set V, kernel function κ, $\mathcal{H}$.

1. For each h in $\mathcal{H}$, compute the matrix M_s and analyze the block structure in the greyscale plot of M_s.
2. For each h in $\mathcal{H}$, plot the embedded data in the feature space and analyze the shape of the cone structure.
3. If the number of blocks in Step 1 is equal to the number of spikes in Step 2, then set K equal to the number of blocks. Go to Step 5.
4. Otherwise, analyze the eigengap plot.

   a. If this plot shows a unique maximum eigengap for each $h \in \mathcal{H}$, then set K according to this maximum. Go to Step 5.
   b. If this plot shows multiple maxima for different $h \in \mathcal{H}$, select the number of clusters K not to be smaller than the number of tight spikes in the corresponding plot of the embedded data.

5. Select $h \in \mathcal{H}$ such that the clearest orthogonal data structure emerges from the plot of the embedded data.
6. Stop.

Output: K, h.

4 Three-way Spectral Clustering

In this section, we propose a spectral approach for clustering three-way data. Three-way data consists of a data set referring to the same sets of units and variables, observed in different situations, i.e., a set of multivariate matrices, that can be organized in three modes: n units, p variables, and t situations. Therefore, given n matrices that represent the vertices of the graph, each matrix is composed by p columns that represent our variables and t rows that represent the time or another feature. So we have a tensor of dimension $n \times t \times p$, thus the dataset is a tensor $\{X\}_{isk}$ for $i = 1, \ldots, n$, $s = 1, \ldots, t$, $k = 1, \ldots, p$.

We define a distance function δ_M between two matrices $A, B \in \mathbb{R}^{p \times t}$ such that $\delta_M : R^{t \times p} \times R^{t \times p} \to [0, +\infty)$ is defined as

$$\delta_M(A, B) := \|A - B\|_F = \sqrt{\sum_{s=1}^{t} \sum_{k=1}^{p} |a_{sk} - b_{sk}|^2} \tag{3}$$

where $\| \cdot \|_F$ is Frobenius norm[1]. Thus the distance between two units in the matrix data X is equal to

$$\delta_M(X_{i_1 sk}, X_{i_2 sk}) = \sqrt{\sum_{s=1}^{t} \sum_{k=1}^{p} |X_{i_1 sk} - X_{i_2 sk}|^2}, \qquad \text{for } i_1, i_2 = 1, \ldots, n. \tag{4}$$

For simplicity, in the following, we denote $\delta_M(X_{i_1 sk}, X_{i_2 sk})$ by $\delta_M(i_1, i_2)$. Moreover, we define the three-way self-tuning kernel function as

$$\kappa_S : X \times X \to [0, +\infty), \qquad \kappa_S(i_1, i_2) = \exp\left(-\frac{\delta_M(i_1, i_2)}{\epsilon_{i_1} \epsilon_{i_2}}\right) \tag{5}$$

where ϵ_{i_1} and ϵ_{i_2} need to be selected like in the kernel defined in (2).

Afterwards, we compute the similarity matrix W given by $w_{i_1 i_2} = \kappa(i_1, i_2)$, so that we can apply the spectral clustering algorithm.

Finally, we point out that, differently from approaches based on mixtures of matrix-variate data, the number of variables of the data set is not a critical issue because the spectral clustering algorithm is based on distance measures.

5 A Real Data Application

We apply the three-way spectral clustering to the analysis of the Insurance data set, available in the `splm` R package. This dataset was initially introduced by [7] and has recently been analyzed by [12]. The goal is to study the consumption of non-life insurance during the years 1998-2002 in the 103 Italian provinces, so $t = 5$ and $n = 103$. As regards the number of variables, we consider all the variables contained in the data set, so $p = 11$. Thus, we have 103 matrices of dimensions 5×11.

The 103 Italian provinces are divided into north-west (24 provinces), north-east (22 provinces), center (21 provinces), south (23 provinces), and islands (13 provinces).

As regard the choice of K and h, we consider the graphical approach introduced in Section 3. In Figure 1 the geometric features of spectral clustering are plotted as h varies. From the number of blocks of the Laplacian matrix (Figure 1-a)), the first maximum eigengap (Figure 1-b)) and the number of spikes in the feature space (Figure 1-c)), we deduce that the number of clusters is $K = 2$. For the selection of

---

[1] In general, given a matrix $A \in \mathbb{R}^{n \times m}$, with $A = (a_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The Frobenius norm is defined by

$$\|A\|_F := \sqrt{\sum_{j=1}^{m} \sum_{i=1}^{n} |a_{ij}|^2}.$$

1%
h = 1

2%
h = 2

5%
h = 5

10%
h = 10

15%
h = 15

20%
h = 21

a)   b)   c)
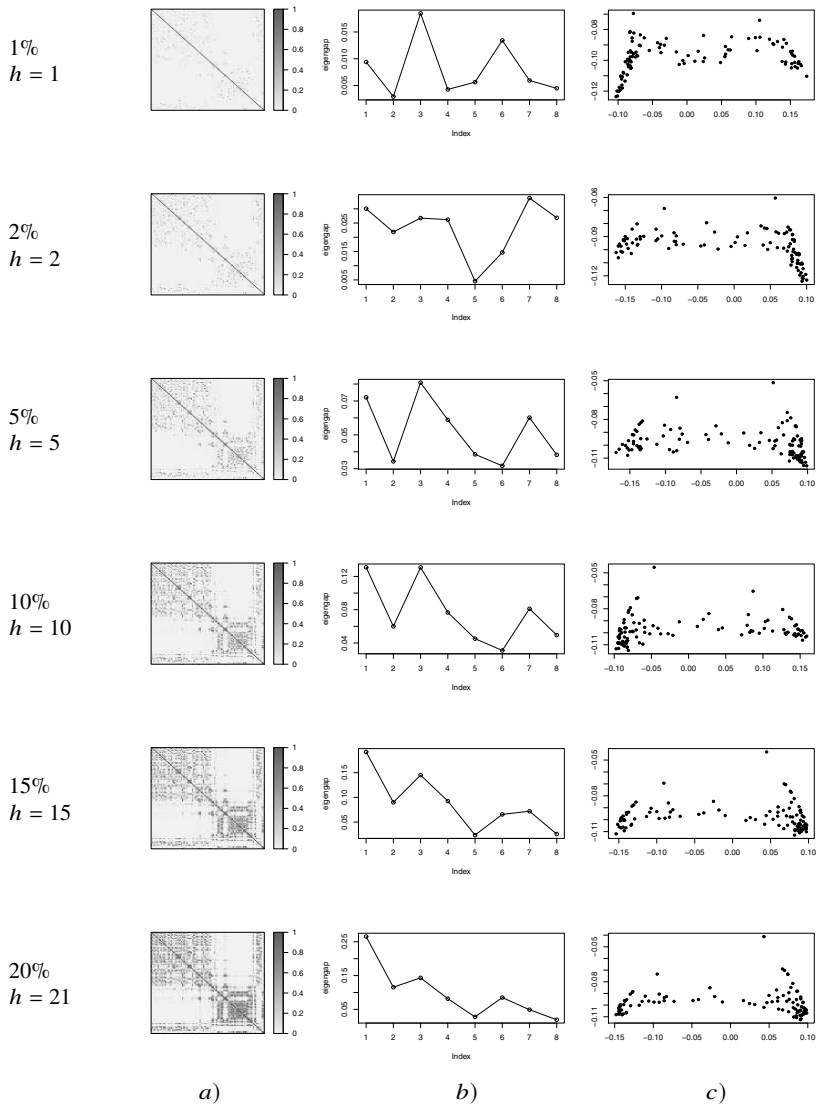
Fig. 1 *Insurance data.* Spectral clustering features: *a*) plot of Laplacian matrix in greyscale; *b*) plot of the first eight eigengap values; *c*) scatterplot of the embedded data along with directions (γ_1, γ_2).

Table 1 *Insurance data.* Table of spectral clustering result.

	NORTHWEST (24 provinces)
Cluster 1	NORTH EAST (22 provinces)
	CENTRE (15 provinces)
	CENTRE (6 provinces)
Cluster 2	SOUTH (23 provinces)
	ISLANDS (13 provinces)

h we choose indifferently $h = 15$ and $h = 21$ because in these cases the maximum eigengap highlights the maximum values corresponding to $K = 2$. In Table 1 the clustering results are presented. This table shows that only 6 center provinces are classified together with the southern provinces. But to be sure that these provinces are neighboring the south provinces, let us analyze spectral clustering results on the map of Italy. Figure 2-a) illustrates the partition deriving from spectral clustering in the political map of Italy, where Italian regions are described by the yellow lines, while the provinces are by the black lines. The result shows a clear separation between center-north Italy and south-insular Italy, in fact, the center-north has a level of insurance penetration close to the European averages, while the South is less developed economically. However, the Massa-Carrara province should belong to the centre-north group. Moreover, we remark that the Rome province, being the capital of Italy, has one socio-economic development comparable to that of north Italy justifying belonging to the centre-north group.

Furthermore, in Figure 2-b) we also represented the partition produced by MN-CWM proposed in [12], we note that the two clustering results are very similar to each other and differ only for one province of central Italy (precisely for the province of Terni). It should also be emphasized that the dataset analyzed by [12] is different from the one analyzed here, since, to avoid excessive parameterization of the models, the authors select only $p = 5$ variables in the data set.

a)   *b*)
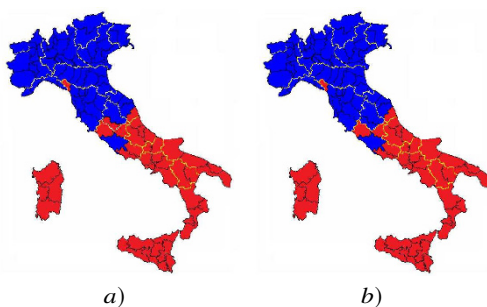
Fig. 2 *Insurance data. a*) Three-way spectral clustering; *b*) Method proposed by [12].

6 Conclusion

In this paper, a spectral approach to cluster three-way data has been proposed. So the data are organized in a tensor and the vertices in the graph are represented by the matrices of dimension $t \times p$. In order to weigh the matrices in the graph, a kernel function based on the Frobenius norm between the matrix difference has been introduced. The performance of the spectral clustering algorithm has been shown in one real three-way data set. Our method is competitive with respect to other clustering methods proposed in the literature to perform matrix-data clustering. Finally, in order to provide suggestions for future research, other kernel functions can be introduced considering different distances with respect to the Frobenius norm.

References

1. Bocci, L., Vicari, D.: ROOTCLUS: Searching for "ROOT CLUSters" in Three-Way Proximity Data. Psychometrika. **84**, 941–985 (2019)
2. Di Nuzzo, C., Ingrassia, S.: A mixture model approach to spectral clustering and application to textual data. Stat. Meth. Appl. Forthcoming (2022)
3. Di Nuzzo, C.: Model selection and mixture approaches in the spectral clustering algorithm. Ph.D. thesis, Economics, Management and Statistics, University of Messina (2021)
4. Di Nuzzo, C., Ingrassia, S.: A joint graphical approach for model selection in the spectral clustering algorithm. Tech. Rep. (2022)
5. Garcia Trillos, N., Hoffman, F., Hosseini, B.: Geometric structure of graph Laplacian embeddings. arXiv preprint arXiv:1901.10651. (2019)
6. Meila, M.: Spectral clustering. In Hennig, C., Meila, M., Murtagh, F., Rocci, R. (eds.). Handbook of Cluster Analysis. Chapman and Hall/CRC (2015)
7. Millo, G., Carmeci, G.: Non-life insurance consumption in Italy: A sub-regional panel data analysis. J. Geogr. Syst. **12**, 1–26 (2011)
8. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. Adv. Neural Inf. Process. Syst. **14** (2002)
9. Sarkar, S., Zhu, X., Melnykov, V., Ingrassia, S.: On parsimonious models for modeling matrix data. Comput. Stat. Data Anal. **142**, 106822 (2020)
10. Schiebinger, G., Wainwright, M. J., Yu, B.: The geometry of kernelized spectral clustering. Ann. Stat. **43**(2), 819–846 (2015a)
11. Tomarchio, S. D., Punzo, A., Bagnato, L.: Two new matrix-variate distributions with application in model-based clustering. Comput. Stat. Data Anal. **152**, 107050 (2020)
12. Tomarchio, S. D., McNicholas, P., Punzo, A.: Matrix normal cluster-weighted models. J. Classif. **38**, 556-575 (2021)
13. Tomarchio, S. D., Ingrassia, S., Melnykov, V.: Modeling students' career indicators via mixtures of parsimonious matrix-normal distributions. Aust. New Zeal. J. Stat. Forthcoming (2022)
14. Vichi, M., Rocci, R., Kiers, H. A. L.: Simultaneous component and clustering models for three-way data: Within and between approaches. J. Classif. **24**, 71–98 (2007)
15. Viroli, C.: Finite mixtures of matrix normal distributions for classifying three-way data. Stat. Comput. **21**, 511–522 (2011)
16. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
17. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. Adv. Neural Inf. Process. Syst. **17** (2004)