SDS
Statistica e
Data Science

Palermo, 11-12 April 2024

Proceedings of the Statistics and Data Science 2024 Conference

New perspectives on Statistics and Data Science

Edited by

Antonella Plaia – Leonardo Egidi
Antonino Abbruzzo

Proceedings of the SDS 2024 Conference
Palermo, 11-12 April 2024
Edited by: Antonella Plaia – Leonardo Egidi - Antonino Abbruzzo
-
Palermo: Università degli Studi di Palermo.

ISBN Ebook (Pdf)
978-88-5509-645-4

Contents

A Simultaneous Spectral Clustering for Three-Way Data

Cinzia Di Nuzzo and Salvatore Ingrassia

Abstract We introduce a novel approach to spectral clustering for three-way data, which integrates simultaneous dimensionality reduction and clustering. While conventional spectral clustering methods focus on two-way data and treat dimensionality reduction and clustering separately, our proposed method extends to handle three-way data, capturing temporal repetition and multivariate interactions. This is the first method, which tackles this challenge purely through statistical techniques.

Key words: Spectral Clustering, Dimensionality reduction, Three-way data

1 Introduction and Background Theory

The Spectral Clustering algorithm is a clustering technique that leverages the spectral structure of data to partition them into homogeneous groups, see [11], [10]. Unlike traditional approaches such as K-Means, Spectral Clustering does not require assuming a specific cluster shape and can handle non-linearly separable data. The Spectral Clustering procedure can be divided into several phases. Initially, a similarity matrix capturing the relationships between data instances is constructed. Subsequently, the Laplacian matrix of the similarity matrix is computed, and the eigenvectors corresponding to the smallest eigenvalues of the Laplacian matrix are determined. These eigenvectors are used to represent the data in a reduced-dimensional space, where the cluster structure is more evident. Finally, a clustering algorithm (such as K-Means, but also mixture models have been taken into account [2]) is applied to this reduced representation to assign instances to clusters. Spectral Clus-

Cinzia Di Nuzzo
Department of Economics and Business, University of Catania, Corso Italia, 55, Catania, 95129, Italy, e-mail: cinzia.dinuzzo@unict.it

Salvatore Ingrassia
Department of Economics and Business, University of Catania, Corso Italia, 55, Catania, 95129, Italy, e-mail: salvatore.ingrassia@unict.it

tering is particularly useful when data exhibit complex or non-linear structures, and when assuming a specific cluster shape is inappropriate. It is effective even when clusters have irregular shapes or when there are clusters of vastly different sizes. Moreover, Spectral Clustering can efficiently handle high-dimensional data and can be adapted to address clustering problems on graphs or non-Euclidean/categorical data [5].

The simultaneous integration of the spectral clustering method as a dimensionality reduction and clustering algorithm is crucial for maximizing efficiency and coherence in data representation. A simultaneous approach allows capturing the intrinsic relationships among the data, reducing computational complexity, and enhancing precision in creating homogeneous groups. A first method to combine the dimensionality reduction and the clustering step in this context has been proposed by [9]. Moreover, extending the spectral clustering method simultaneously for three-way data is essential for addressing situations where the temporal repetition of data is relevant and interactions between three variables can influence results over time. An example can be data concerning user behaviour and interactions on an online social media platform over a certain time interval. Some developments of spectral clustering for three-way data have already been proposed in [4], [3]. By using the spectral clustering method simultaneously for dimensionality reduction and clustering on these three-way data, complex patterns in user behaviour over time can be identified. For example, discovering groups of users with similar characteristics interacting similarly on the platform during specific periods of the day or in response to certain types of content. This would enable better customization of the user experience, adapting the type of content shown or the timing of notifications based on the different user segments identified.

This paper proposes a novel approach to spectral clustering that integrates simultaneous dimensionality reduction and clustering for three-way data. The basic idea is to extend the results presented in [9]. By simultaneously conducting dimensionality reduction and clustering, our approach captures intricate patterns in data representation, thereby enhancing computational efficiency and coherence. We demonstrate the utility of our method through first empirical evaluations on synthetic and real-world datasets. The results showcase the effectiveness of our proposed approach in uncovering complex patterns and facilitating meaningful insights in diverse application domains. Overall, our contribution advances the field of spectral clustering by addressing the challenges posed by three-way data structures and providing a versatile framework for clustering in various contexts.

The rest of the work is structured as follows: Section 2 introduces the foundational concept of the model, while Section 3 provides an illustration using one artificial data and one real-world data.

2 The Model

A novel spectral clustering method for three-way data is presented. Spectral clustering is a two-step sequential procedure involving the reduction of data dimensionality through Laplacian embedding, followed by the application of a clustering algorithm to partition the data into K clusters. Although spectral clustering methods have conventionally been applied to two-way data, we recently proposed a spectral clustering approach for three-way data in [3] and [4]. The objective is to partition a three-way dataset comprising N units, M variables, and H occasions into K clusters. The innovation of the proposed method lies in modelling a set of H Laplacian matrices $\mathbf{L}_h$ (for $h = 1,\dots,H$) of size $N \times N$ as follows:

$$\mathbf{L}_h = \mathbf{A}\mathbf{C}_h\mathbf{A}' + \mathbf{E}_h, \tag{1}$$

such that $\mathbf{A}\mathbf{A}' = \mathbf{I}$, for $h = 1,\dots,H$; where $\mathbf{L}_h$ denotes the Laplacian matrix, $\mathbf{A}$ represents the $N \times K$ matrix for Laplacian embedding, $\mathbf{C}_h$ is a $K \times K$ diagonal matrix with non-negative elements, and $\mathbf{E}_h$ is the $N \times N$ error term. We adopt a least squares approach to estimate the part of dimensionality reduction of the model, based on the following minimisation under orthogonality constraints

$$\underset{\mathbf{A},\mathbf{C}_h}{\arg\min} \sum_{h=1}^{H} \left|\left|\mathbf{L}_h - \mathbf{A}\mathbf{C}_h\mathbf{A}'\right|\right|^2. \tag{2}$$

According to [1], because the direct minimisation of (2) seems difficult; a Candecomp method can be developed to minimize (2), therefore the minimisation problem can be solved by

$$\underset{\mathbf{A},\mathbf{B},\mathbf{C}_h}{\arg\min} \sum_{h=1}^{H} \left|\left|\mathbf{L}_h - \mathbf{A}\mathbf{C}_h\mathbf{B}'\right|\right|^2, \tag{3}$$

where $\mathbf{B}$ is an $N \times K$ matrix. In this context, a notable contribution arises from the findings presented in [6] and the associated algorithm known as Indort. Particularly, [6] demonstrates that if the matrices $\mathbf{L}_h$ are positive definite and the elements of $\mathbf{C}_h$ are non-negative, then the resulting matrices $\mathbf{A}$ and $\mathbf{B}$ in (3) obtained from the Indort algorithm are identical, thus (3) solves the least squares problem of (2). These results facilitate the derivation of a unique configuration of Laplacian embedding shared among all H Laplacian matrices $\mathbf{L}_h$, rendering the matrix $\mathbf{A}$ representative of the entire dataset. Within this framework, to conduct dimensionality reduction, an algorithm for clustering is also modelled concurrently, employing K-means and essentially extending the model outlined in [9] to the three-way case.

Finally, it should be noted that the computational burden of the algorithm proposed by [9] is of the order $O(N^2K)$; in the model proposed here, an additional step is merely appended wherein individual occasions sums are computed within the specific SVD calculation in the estimation process of matrices $\mathbf{A}$ and $\mathbf{B}$. Another difference from the method proposed in [9] is the estimation of matrix $\mathbf{C}_h$, which is introduced in this model to represent the weights of the occasion h in each cluster. It is also important to note that the latter part of the method, concerning the clustering process, remains unaltered compared to the method already proposed in [9], as in the model we are proposing here, we opt to reduce the original data dimensionality of size $N \times M \times H$ to a common space of dimension $N \times K$ shared across all occasions. Hence, in terms of computational burden and utility of the method in the dimensionality reduction process, this model facilitates resolving situations with initially high dimensions, which are subsequently condensed into a much smaller common space containing the structure of clusters from the original data.

3 Some results on synthetic and real-word datasets

Here, we provide some examples illustrating the effectiveness of the method. The first example concerns an artificial dataset originating from $K = 3$ normal distributions each comprising $N = 20$ units, with means $\mu_1^{(1)} = (1.5, 1, 2)$, $\mu_2^{(1)} = (0, 0, 0)$, and $\mu_3^{(1)} = (-1, -2.5, -2)$, respectively, and, covariance matrices equal to $\Sigma^{(1)} = \mathbf{I}$. We repeat the random generation of such data $H = 3$ times, specifically, we generate randomly the second occasion from 3 normal distribution with $\mu_1^{(2)} = (2.5, 1.5, 2.5)$, $\mu_2^{(2)} = (0.2, 0.3, 0.4)$, and $\mu_3^{(2)} = (-2.5, -2.4, -2)$ and $\Sigma^{(2)} = 2\mathbf{I}$. Finally we generate the third occasion from normal distribution with $\mu_1^{(3)} = (6, 5, 5)$, $\mu_2^{(3)} = (5, 4, 3.7)$, and $\mu_3^{(3)} = (3.7, 2.8, 1.8)$ and $\Sigma^{(3)} = 0.5\mathbf{I}$, respectively.

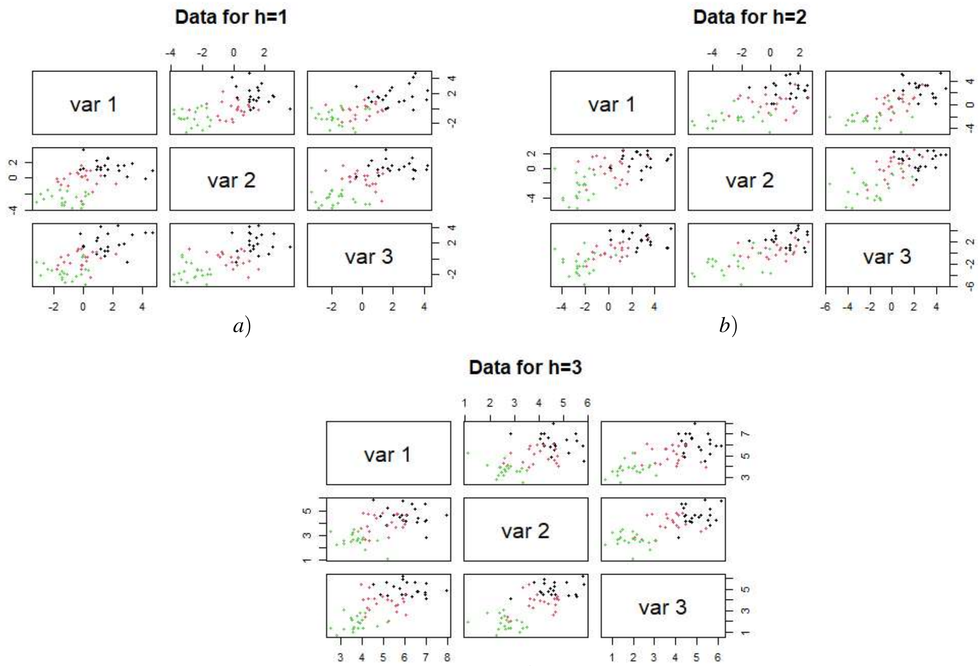

Fig. 1 *Synthetic data. a)* $\mathbf{X}_1$. *b)* $\mathbf{X}_2$. *c)*$\mathbf{X}_3$.

Denoting the dataset of $N = 60$ units and $M = 3$ variables across $H = 3$ occasions as $\mathbf{X}$, the original data are depicted in Figure 1, it should be noted that the other settings exhibit similar patterns, differing primarily in the magnitude of the data. Upon constructing the original dataset $\mathbf{X}$, three Laplacian matrices, namely $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_3$, are generated such that $\mathbf{L}_h = \mathbf{D}_h^{1/2}\mathbf{W}_h\mathbf{D}_h^{-1/2}$, for $h = 1,2,3$, where $\mathbf{W}_h$ represents the similarity matrix constructed using a kernel function (refer to [12]):

$$w_{ij}^{(h)} = \exp\left(-\frac{\|\mathbf{x}_i^{(h)} - \mathbf{x}_j^{(h)}\|^2}{\varepsilon_i^{(h)}\varepsilon_j^{(h)}}\right), \qquad \text{for} \quad h = 1,\dots,3 \tag{4}$$

with $\varepsilon_i^{(h)} = \|\mathbf{x}_i^{(h)} - \mathbf{x}_m^{(h)}\|$, where $\mathbf{x}_m^{(h)}$ is the m-neighbor of the point $\mathbf{x}_i^{(h)}$, for $i, j = 1,\dots,N$, and $m = 2$. Additionally, the matrix $\mathbf{D}_h$ signifies the degree matrix associated with $\mathbf{W}_h$ for $h = 1,\dots,3$. Employing the proposed algorithm outlined in (3) to process the set of matrices $\mathbf{L}_h$, with the number of groups predetermined, the resulting reduced-dimensional space, as determined by the clustering outcome, is depicted in Figure 2 (ARI=1).

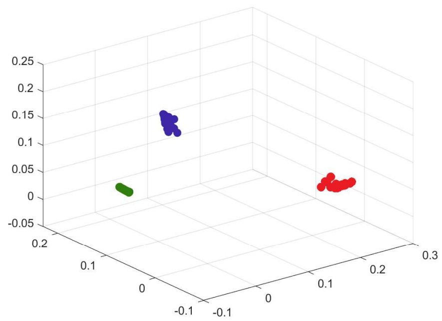

Fig. 2 *Synthetic data.* Clustering result on the embedded space.

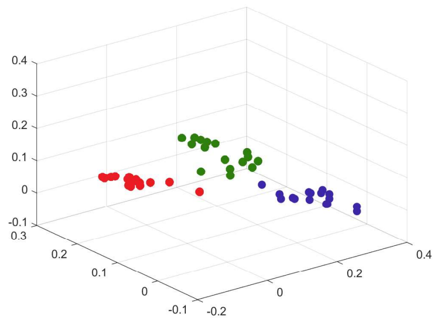

Fig. 3 *Blue Crabs data.* Clustering result on the embedded space.

The second example pertains to a commonly used dataset called Blue Crabs data. Blue crabs data can be accessed at https://three-mode.leidenuniv.nl/ (see [7] and [8]). The dataset comprises gill, hepatopancreas, and muscle tissue samples extracted from blue crabs, categorized into three groups based on their origin and health condition: Albermarle Sound (Healthy), Pamlico River (Healthy), and Pamlico River (Diseased). The dataset consists of $N = 48$ units (tissue samples) by $M = 3$ variables (tissue types) by $H = 25$ occasions (trace elements). Our model was applied to this dataset, and the result in the embedded space is depicted in Figure 3. In this case as well, our method succeeded in achieving correct classification (ARI=1).

Acknowledgement

This study was funded by the GRINS project PE00000018-CUP E63C22002120006 (*Di Nuzzo Cinzia*), and by PNRR MUR project PE0000013-FAIR (*Salvatore Ingrassia*).

References

1. Carroll, J.D., and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition, *Psychometrika*, **35**, 283-319.
2. Di Nuzzo, C., Ingrassia, S. (2022). A mixture model approach to spectral clustering and application to textual data, *Statistical Methods & Applications*. **31**, 1071-–1097.
3. Di Nuzzo, C., Ingrassia, S., Vicari, D. (2022). An INDSCAL-Type Approach for Three-Way Spectral Clustering. In "García-Escudero L.A., Gordaliza A., Mayo-Iscar A., Lubiano Gomez M.A., Gil M.A, Grzegorzewski P., Hryniewicz O. (Eds.) *Building Bridges between Soft and Statistical Methodologies for Data Science", Advances in Intelligent Systems and Computing*, Springer, **1433**, 128–135.
4. Di Nuzzo C., Ingrassia S. (2023). Three-way Spectral Clustering. In "Brito P., Dias J.G., Lausen B., Montanari A., Nugent R. (Eds.) *Classification and Data Science in the Digital Age", Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, 111–118.
5. Di Nuzzo, C. (2024). Advancing Spectral Clustering for Categorical and Mixed-Type Data: Insights and Applications. *Mathematics*. **12**(4): 508.
6. Dosse, M.B., ten Berge, J.M. & Tendeiro, J.N. (2011). Some New Results on Orthogonally Constrained Candecomp. *Journal of Classification*, **28**(2), 144–155.
7. Gemperline, P. J., Miller, K. H., West, T. L., Weinstein, J. E., Hamilton, J. C., & Bray, J. T. (1992). Principal component analysis, trace elements, and blue crab shell disease. *Analytical Chemistry*, **64**, 523–531.
8. Kroonenberg, P. M., Basford, K. E., & Gemperline, P. J. (2004). Grouping three-mode data with mixture methods: the case of the diseased blue crabs. *Journal of Chemometrics*, **18**, 508–518.
9. Labiod, L., Nadif, M. (2021). Efficient regularized spectral data embedding. *Advances Data Analysis Classification*. **15**, 99–119.
10. Ng, A.Y.,Jordan, M., Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In "Dietterich T., Becker S. and Ghahramani Z. (Eds.) *Advances in neural information processing systems"*, MIT Press, **14**.
11. von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing* **17**, 4, 395–416.
12. Zelnik-Manor, L., Perona, P. (2004). Self-tuning spectral clustering. In "L. Saul and Y. Weiss and L. Bottou (Eds.) *Advances in Neural Information Processing Systems*", MIT Press, **17**.