Supervised Classification via Neural Networks for Replicated Point Patterns

Kateřina Pawlasová, Iva Karafiátová, and Jiří Dvořák

Abstract A spatial point pattern is a collection of points observed in a bounded region of $\mathbb{R}^d$, $d \geq 2$. Individual points represent, e.g., observed locations of cell nuclei in a tissue ($d = 2$) or centers of undesirable air bubbles in industrial materials ($d = 3$). The main goal of this paper is to show the possibility of solving the supervised classification task for point patterns via neural networks with general input space. To predict the class membership for a newly observed pattern, we compute an empirical estimate of a selected functional characteristic (e. g., the pair correlation function). Then, we consider this estimated function to be a functional variable that enters the input layer of the network. A short simulation example illustrates the performance of the proposed classifier in the situation where the observed patterns are generated from two models with different spatial interactions. In addition, the proposed classifier is compared with convolutional neural networks (with point patterns represented by binary images) and kernel regression. Kernel regression classifiers for point patterns have been studied in our previous work, and we consider them a benchmark in this setting.

Keywords: spatial point patterns, pair correlation function, supervised classification, neural networks, functional data

Kateřina Pawlasová (✉)
Charles University, Faculty of Mathematics and Physics, Ke Karlovu 3, 121 16 Praha 2, Czech Republic, e-mail: pawlasova@karlin.mff.cuni.cz

Iva Karafiátová
Charles University, Faculty of Mathematics and Physics, Ke Karlovu 3, 121 16 Praha 2, Czech Republic, e-mail: karafiatova@karlin.mff.cuni.cz

Jiří Dvořák
Charles University, Faculty of Mathematics and Physics, Ke Karlovu 3, 121 16 Praha 2, Czech Republic, e-mail: dvorak@karlin.mff.cuni.cz

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_32

1 Introduction

Spatial point processes have recently received increasing attention in a broad range of scientific disciplines, including biology, statistical physics, or material science [9]. They are used to model the locations of objects or events randomly occurring in $\mathbb{R}^d$, $d \geq 2$. We distinguish between the stochastic model (point process) and its realization observed in a bounded observation window (point pattern).

Typically, analyzing spatial point pattern data means working with just one pattern, which comes from a specific physical measurement. In this paper, we take another perspective: we suppose that a collection of patterns, which are independent realizations of some underlying stochastic models, is to be analyzed simultaneously. These independent realizations are then referred to as replicated point patterns. Recently, this type of data has become more frequent, encouraging the adaptation of methods such as supervised classification to the point pattern setting.

Since we are talking about supervised classification, our task is to predict the label variable (indicating class membership) for a newly observed point pattern, using the knowledge about a sample collection of patterns with known labels (training data). In the literature, this problem has been studied to a limited extent. Properties of a classifier constructed specifically for the situation where the observed patterns were generated by inhomogeneous Poisson point processes with different intensity functions are discussed in [5]. However, this method is based on the special properties of the Poisson point process, and its use is thus limited to a small class of models. On the other hand, no assumptions about the underlying stochastic models are made in [12], where the task for replicated point patterns is transformed, with the help of multidimensional scaling [16], to the classification task in $\mathbb{R}^2$. In [10, 11], the kernel regression classifier for functional data [4] is adapted for replicated point patterns. Instead of classifying the patterns themselves, a selected functional characteristic (e.g. the pair correlation function) is estimated for each pattern. These estimated values are considered functional observations, and the classification if performed in the context of functional data. The idea of linking point patterns to functional data also appears in [12] – the dissimilarity matrix needed for the multidimensional scaling is based on the same type of dissimilarity measure that is used for the kernel regression classifier in [10, 11]. Finally, [17] briefly discusses the model-based supervised classification. Unsupervised classification is explored in [2].

In this paper, our goal is to discuss the use of classifiers based on artificial neural networks in the context of replicated point patterns. We pay special attention to the procedure described in [14], where both functional and scalar observations enter the input layer. Hence, similarly as in [10, 11], each pattern can be represented by estimated values of a selected functional characteristic and the classification is performed in the context of functional data. The resulting decision about class membership is based on the spatial properties of the observed patterns that can be described by the selected characteristic. Therefore, with a thoughtfully chosen characteristic, this method has great potential within a wide range of possible classification scenarios. Moreover, it can be used without assuming stationarity of the underlying point processes, and it can be easily extended to more complicated settings (e.g., point patterns in non-Euclidean spaces or realizations of random sets).

We present a short simulation experiment that illustrates the behaviour of the neural network described in [14]. Binary classification is performed on realizations of two different point process models – the Thomas process (model for attractive interactions among pairs of points) and the Poisson point process (benchmark model for no interactions among points). This approach is then compared to the classification based on convolutional neural networks (CNNs) [8], where each pattern enters the network as a binary image. Finally, both methods based on artificial neural networks are compared to the kernel regression classifier studied in [10, 11] which can be considered a benchmark in the context of replicated point patterns.

This paper is organized as follows. Section 2 provides a brief theoretical background on spatial point processes and their functional characteristics, including the definition of the pair correlation function, which plays a crucial role in the sequel. Section 3 summarizes the methodology introduced in [14] about neural network models with general input space. Section 4 is devoted to a short simulation example.

2 Point Processes and Point Patterns

This section presents the necessary definitions from the point process theory. Our exposition closely follows the book [13]. For detailed explanation of the theoretical foundations, see, e.g., [7]. Throughout the paper, a simple point process X is defined as a random locally finite subset of $\mathbb{R}^d$, $d \geq 2$, where each point $x \in X$ corresponds to a specific object or event occurring at the location $x \in \mathbb{R}^d$. In applications, X can be used as a mathematical tool to model random locations of cell nuclei in a tissue (with $d = 2$) or centers of undesirable air bubbles in industrial materials ($d = 3$). We distinguish between the mathematical model X, which is called a point process, and its observed realization $\mathcal{X}$, which is called a point pattern. Examples of four different point patterns are given in Figure 1.

Before proper definition of the pair correlation function, a functional characteristic that plays a key role in the sequel, we need to define some moment properties of X. The *intensity function* $\lambda(\cdot)$ is a non-negative measurable function on $\mathbb{R}^d$ such that $\lambda(x)\,\mathrm{d}x$ corresponds to the probability of observing a point of X in a neighborhood of x with an infinitesimally small area $\mathrm{d}x$. If X is stationary (its distribution is translation invariant in $\mathbb{R}^d$), then $\lambda(\cdot) = \lambda$ is a constant function and the constant λ is called the *intensity* of X. In this case, λ is interpreted as the expected number of points of X that occur in a set with unit d-dimensional volume. Similarly, the *second-order product density* $\lambda^{(2)}(\cdot,\cdot)$ is a non-negative measurable function on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\lambda^{(2)}(x, y)\,\mathrm{d}x\,\mathrm{d}y$ corresponds to the probability of observing two points of X that occur jointly at the neighborhoods of x and y with infinitesimally small areas $\mathrm{d}x$ and $\mathrm{d}y$.

Assuming the existence of λ and $\lambda^{(2)}$, the *pair correlation function* $g(x, y)$ is defined as $\lambda^{(2)}(x, y)/(\lambda(x)\lambda(y))$, for $\lambda(x)\lambda(y) > 0$. If $\lambda(x) = 0$ or $\lambda(y) = 0$, we set

$g(x, y) = 0$. We write $g(x, y) = g(x - y)$ when g is translation invariant and $g(x, y) = g(\|x - y\|)$ when g is also isotropic (invariant under rotations around the origin). For the Poisson point process, a model for complete spatial randomness, $\lambda^{(2)}(x, y) = \lambda(x)\lambda(y)$ and $g \equiv 1$. Thus, $g(x, y)$ quantifies how likely it is to observe two points in X jointly occurring in infinitesimally small neighbourhoods of x and y, relative to the "no interactions" benchmark.

A large variety of characteristics (both functional and numerical) have been developed to capture various hypotheses about the stochastic models that generated the observed point patterns at hand. We have focused on the pair correlation function g mainly because of its widespread use in practical applications and ease of interpretation. Other popular characteristics are based on g, e.g., its cumulative counterpart, traditionally called the K-function. Others are based on inter-point distances, such as the nearest-neighbor distance distribution function G and the spherical contact distribution function F. A comprehensive summary of commonly used characteristics, including the list of possible empirical estimators, is presented in [9, 13]. Estimators of g, K, G, and F are implemented in the R package spatstat [3].

3 Neural Networks with General Input Space

This section prepares the theoretical background for the supervised classification of replicated point patterns via artificial neural networks. The recent approach of [14, 15] is the cornerstone of our proposed classifier, and hence we focus on its description in the following paragraphs. On the other hand, the approach based on CNNs is more established in the literature. We use it primarily for comparison and thus we refer the reader to [8] for a detailed description.

Following the setup in [14], let us assume that we want to build a neural network such that it takes $K \in \mathbb{N}$ functional variables and $J \in \mathbb{N}$ scalar variables as input. In detail, suppose that we have $f_k : \tau_k \longrightarrow \mathbb{R}$, $k = 1, 2, \ldots, K$ (τ_k are possibly different intervals in $\mathbb{R}$), and $z_j^{(1)} \in \mathbb{R}$, $j = 1, 2, \ldots, J$. Furthermore, suppose that the first layer of the network contains $n_1 \in \mathbb{N}$ neurons. We then want the i-th neuron of the first layer to transfer the value

$$z_i^{(2)} = g\left(\sum_{k=1}^{K} \int_{\tau_k} \beta_{ik}(t) f_k(t)\, \mathrm{d}t + \sum_{j=1}^{J} w_{ij}^{(1)} z_j^{(1)} + b_i^{(1)}\right), \quad i = 1, 2, \ldots, n_1,$$

where $b_i^{(1)} \in \mathbb{R}$ is the bias and $g : \mathbb{R} \longrightarrow \mathbb{R}$ is the activation function. Two types of weights appear in the formula: the functional weights $\{\beta_{ik} : \tau_k \longrightarrow \mathbb{R}\}$, and the scalar weights $\{w_{ij}^{(1)}, b_i^{(1)}\}$. The optimal value of all these weights should be found during the training of the network. To overcome the difficulty of finding the optimal weight functions β_{ik}, we can express β_{ik} as a linear combination of $\phi_1, \ldots, \phi_{m_k}$, where $\phi_1, \ldots, \phi_{m_k}$ are the basis functions (from the Fourier or B-spline basis) and m_k is chosen by the user. The sum $\sum_{k=1}^{K} \int_{\tau_k} \beta(t)_{ik} f_k(t)\, \mathrm{d}t$ can

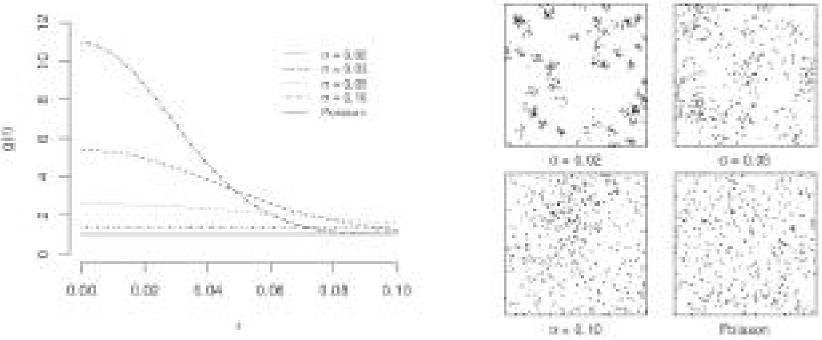

Fig. 1 Theoretical values of the pair correlation function g for the Poisson point process and the Thomas process with different values of the model parameter σ. For these models, g is translation invariant and isotropic. A single realization of the Poisson point process and the Thomas process with parameter σ set to 0.1, 0.05 and 0.02 respectively, is illustrated in the right part of the figure.

be expressed as $\sum_{k=1}^{K} \sum_{l=1}^{m_k} c_{ilk} \int_{\mathcal{T}_k} \phi_l(t) f_k(t)\,\mathrm{d}t$, where the integrals $\int_{\mathcal{T}_k} \phi_l(t) f_k(t)\,\mathrm{d}t$ can be calculated a priori and the coefficients of the linear combination of the basis functions $\{c_{ilk}\}$ act as scalar weights of the first layer and are learned by the network. The scalar values $z_i^{(2)}, i = 1, \ldots, n_1$, then propagate through the next fully connected layers as usual. An in-depth analysis of the computational point of view is provided in [14]. In the software `R`, neural networks with general input space are covered by the package `FuncNN` [15] built over the packages `keras` [6] and `tensorflow` [1]. The last two packages are used to handle CNNs.

4 Simulation Example

This section presents a simple simulation experiment in which we illustrate the performance of the classification rule based on the neural network with general input space. Binary classification is considered, where the group membership indicates whether a point pattern was generated by a stationary Poisson point process or a stationary Thomas process, the latter exhibiting attractive interactions among pairs of points [13]. The sample realizations can be seen in Figure 1.

We consider the Thomas process to be a model with one parameter σ. Small values of σ indicates strong, attractive short-range interactions between points, while larger values of σ result in looser clusters of points. Attractive interactions between the points of a Thomas process result in the values of the pair correlation function being greater than the constant 1, which corresponds to the Poisson case. The effect of σ on the shape of the theoretical pair correlation function of the Thomas process (which is translation invariant and isotropic) is illustrated in Figure 1.

Since the model parameter σ affects the strength and range of attractive interactions between points of the Thomas process, the complexity of the binary classification task described above increases with increasing values of σ [10, 11]. Therefore, this experiment focuses on the situation where σ is set to 0.1, and all realizations are observed on the unit square $[0, 1]^2$. We fix the intensity of the two models to 400 (in spatial statistics, patterns with several hundreds of points are standard nowadays). In this framework, we expect the classification task to be challenging enough to observe differences in the performance of the considered classifiers. On the other hand, it is still reasonable to distinguish (w.r.t. the chosen observation window) the realizations of the model with attractive interactions from the realizations corresponding to the complete spatial randomness.

Two different collections of labelled point patterns are considered as training sets. The first, referred to as *Training data 1*, is composed of 1 000 patterns per group. The second, called *Training data 2*, is then composed of 100 patterns per group. The test and validation sets have the same size and composition as the *Training data 2*. Table 1 presents the accuracy of three classification rules (described below) with respect to the test set. For the first two rules, the accuracy is in fact averaged over five runs corresponding to different settings of initial weights in the underlying neural network. Concerning the network architecture, we fix the ReLU function to be the activation function for all layers, except the output one. The output layer consists of one neuron with sigmoid activation function. The loss function is the binary cross-entropy. A detailed description of the individual layers is given below.

Rule 1 is based on the neural network with general input space. We set K and J from Sect. 3 to be 1 and 0, respectively, and $\tau_1 = (0, 0.25)$. The value 0.25 is related to the observation window of the point patterns at hand being $[0, 1]^2$. Then, f_1 is the vector of the estimated values of the pair correlation function g (estimated by the function `pcf.ppp` from the package `spatstat` [3] with default settings but the option `divisor` set to `d`), considered as a functional observation. Furthermore, we set $m_1 = 29$, and consider the Fourier basis. The data preparation (estimation of g, computation of integrals from Sect. 3) takes 740 s of elapsed time (w.r.t. the *Training data 1*, on a standard personal computer). To tune the hyperparameters of the final neural network (number of hidden layers, number of neurons per hidden layers, dropout, etc.), we performed a rough grid search (models with various combinations of the hyperparameters were trained on *Training data 1* and we used the loss function and the accuracy computed on the validation set to compare the performances). The resulting network consists of one hidden layer with 128 neurons followed by a dropout layer with a rate of 0.3. We use the Adam optimizer, and the learning rate is decaying exponentially, with initial value 0.001 and decay parameter 0.05. In total, the network has 3 969 trainable parameters. To train the network, we perform 50 epochs with an average elapsed time of 200 ms per epoch (w.r.t. *Training data 1*).

Rule 2 uses CNNs. Similarly to the previous case, our decision about the network architecture is based on a rough grid search. The final network has two convolutional layers, each of them with 8 filters, a squared kernel matrix with 36 (first layer) or 16 rows (second layer), and a following average pooling layer with the pool size fixed at 2 × 2. We add a dropout layer after the pooling, with a rate of 0.3 (after the first

Table 1 Accuracy for the three presented classification rules w.r.t. the testing set. For *Rule 1* and *Rule 2*, the accuracy is averaged over five runs corresponding to five different choices of initial weights in the underlying neural networks. In addition, the standard deviation computed from the five accuracy values is reported. Values close to 1 indicate a nearly perfect classification.

	Rule 1	*Rule 2*	*Rule 3*
Training data 1	**0.947** ±0.003	0.934 ±0.032	0.935
Training data 2	0.895 ±0.010	0.512 ±0.028	**0.925**

pooling) and 0.2 (after the second pooling). The batch size is set to 32. We use the Adam optimizer, and the learning rate is decaying exponentially, with initial value 0.001 and decay parameter 0.1. The total number of trainable parameters is equal to 32 785 and we perform 50 epochs with the average elapsed time per epoch (w.r.t. *Training data 1*) equal to 930 s. Data preparation (converting point patterns to binary images) takes less than 10 s of the elapsed time (w.r.t. *Training data 1*).

Rule 3 is the kernel regression classifier studied in [10, 11]. We use the Epanechnikov kernel together with an automatic procedure for the selection of the smoothing parameter. The underlying dissimilarity measure for point patterns is constructed as the integrated squared difference of the corresponding estimates of the pair correlation function g; for more details, see [10]. The elapsed time needed to compute the upper triangle of the dissimilarity matrix (containing dissimilarities between every pair of patterns from *Training data 1*) is equal to 390 s. To predict the class membership for the testing set (w.r.t. *Training data 1*), 206 s elapsed. During the classification procedure, no random initialization of any weights is needed. Thus, there is no reason to average the accuracy in Table 1 over multiple runs.

For *Training data 1*, Table 1 shows that the highest accuracy was achieved for the neural network with general input space. The standard deviation of the five different accuracy values is significantly higher for CNN which has almost ten times more trainable parameters than the network with general input space. For *Training data 2*, the kernel regression method achieved the highest accuracy. In this situation, the performance of the classifier is stable even in the case of small training data. For the first two rules, the neural network models chosen with the help of the grid search (where the networks were trained w.r.t. the bigger training set) are now trained w.r.t. the smaller training set. The resulting accuracy is still around 0.90 for the network with general input space, but it drops to 0.5 (random assignment of labels) for CNN. The size of *Training data 2* seems to be too small to successfully optimize the large amount of trainable parameters of the convolutional network.

To conclude, our simulation example suggests that the classifier based on CNN (using information about the precise configuration of points) is in the presented situation outperformed by the classifiers based on the estimated values of the pair correlation function (using information about the interactions between pairs of points). The high number of trainable parameters of the CNN makes its use rather demanding with respect to computational time. The approach based on neural networks with general input space proved to be competitive with or even outperform the current benchmark method (kernel regression classifier), especially for large datasets. Also, it has the lowest demands regarding computational time. In the case of a small dataset, the low number of hyperparameters speaks in favor of kernel regression. Finally, in the simple classification scenario that we have presented, the choice of the pair correlation function was adequate. In practical applications, a problem-specific characteristic should be constructed to achieve satisfactory performance.

Acknowledgements The work of Kateřina Pawlasová and Iva Karafiátová has been supported from the Grant schemes at Charles University, project no. CZ.02.2.69/0.0/0.0/19 073/0016935. The work of Jiří Dvořák has been supported by the Czech Grant Agency, project no. 19-04412S.

References

1. Allaire, J. J., Eddelbuettel, D., Golding, N., Tang, Y.: `tensorflow`: R Interface to TensorFlow (2016) Available at GitHub. `https://github.com/rstudio/tensorflow.Cited10Jan2022`
2. Ayala, G., Epifanio, I., Simo, A., Zapater, V.: Clustering of spatial point patterns. Comput. Stat. Data. Anal. **50**, 1016–1032 (2006)
3. Baddeley, A., Rubak, E., Turner, R.: Spatial Point Patterns: Methodology and Applications with R. Chapman & Hall/CRC Press, Boca Raton (2015)
4. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Theory and Practice. Springer-Verlag, New York (2006)
5. Cholaquidis, A., Forzani, L., Llop, P., Moreno, L.: On the classification problem for Poisson point processes. J. Multivar. Anal. **153**, 1–15 (2017)
6. Chollet, F., Allaire, J. J. and others: R Interface to Keras (2017) Available via GitHub. `https://github.com/rstudio/keras.Cited10Jan2022`
7. Daley, D., Vere-Jones, D.: An Introduction to the Theory of Point Processes. Vol II., 2nd edn. Springer-Verlag, New York (2008)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
9. Illian, J., Penttinen, A., Stoyan, H., Stoyan, D.: Statistical Analysis and Modelling of Spatial Point Patterns. Wiley, Chichester (2004)
10. Koňasová, K., Dvořák, J.: Techniques from functional data analysis adaptable for spatial point patterns (2021) In: Proceedings of the 22nd European Young Statisticians Meeting. `https://www.eysm2021.panteion.gr/publications.html.Cited10Jan2022`
11. Koňasová, K., Dvořák, J.: Supervised nonparametric classification in the context of replicated point patterns. *S*ubmitted (2021)
12. Mateu, J., Schoenberg, F. P., Diez, D. M., González, J. A., Lu, W.: On measures of dissimilarity between point patterns: classification based on prototypes and multidimensional scaling. Biom. J. **57**, 340–358 (2015)
13. Møller, J., Waagepetersen, R.: Statistical Inference and Simulation for Spatial Point Processes. Chapman & Hall/CRC, Boca Raton (2004)
14. Thind, B., Multani, K., Cao, J.: Deep Learning with Functional Inputs (2020) Available via arxiv. `https://arxiv.org/pdf/2006.09590.pdf.Cited10Jan2022`
15. Thind, B., Wu, S., Groenewald, R., Cao, J.: FuncNN: An R Package to Fit Deep Neural Networks Using Generalized Input Spaces (2020) Available via arxiv. `https://arxiv.org/pdf/2009.09111.pdf.Cited10Jan2022`
16. Torgerson, W.: Multidimensional Scaling: I. Theory and Method. Psychometrika. **17**, 401–419 (1952)
17. Vo, B. N., Dam, N., Phung, D., Tran, Q. N., Vo, B. T.: Model-based learning for point pattern data. Pattern Recognit. **84**, 136–151 (2018)

Parsimonious Mixtures of Seemingly Unrelated Contaminated Normal Regression Models

Gabriele Perrone and Gabriele Soffritti

Abstract In recent years, the research into linear multivariate regression based on finite mixture models has been intense. With such an approach, it is possible to perform regression analysis for a multivariate response by taking account of the possible presence of several unknown latent homogeneous groups, each of which is characterised by a different linear regression model. For a continuous multivariate response, mixtures of normal regression models are usually employed. However, in real data, it is not unusual to observe mildly atypical observations that can negatively affect the estimation of the regression parameters under a normal distribution in each mixture component. Furthermore, in some fields of research, a multivariate regression model with a different vector of covariates for each response should be specified, based on some prior information to be conveyed in the analysis. To take account of all these aspects, mixtures of contaminated seemingly unrelated normal regression models have been recently developed. A further extension of such an approach is presented here so as to ensure parsimony, which is obtained by imposing constraints on the group-covariance matrices of the responses. A description of the resulting parsimonious mixtures of seemingly unrelated contaminated regression models is provided together with the results of a numerical study based on the analysis of a real dataset, which illustrates their practical usefulness.

Keywords: contaminated normal distribution, ECM algorithm, mixture of regression models, model-based cluster analysis, seemingly unrelated regression

Gabriele Perrone (✉)
Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy, e-mail: `gabriele.perrone4@unibo.it`

Gabriele Soffritti
Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy, e-mail: `gabriele.soffritti@unibo.it`

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_33

1 Introduction

Seemingly unrelated (SU) regression equations are usually employed in a multivariate regression analysis whenever the dependence of a vector $\mathbf{Y} = (Y_1, \ldots, Y_M)'$ of M continuous variables on a vector $\mathbf{X} = (X_1, \ldots, X_P)'$ of P regressors has to be modelled by allowing the error terms in the different equations to be correlated and, thus, the regression parameters of the M equations have to be jointly estimated [14]. With such an approach, the researcher is also enabled to convey prior information on the phenomenon under study into the specification of the regression equations by defining a different vector of regressors for each dependent variable. This latter feature is particularly useful in any situation in which different regressors are expected to be relevant in the prediction of different responses, such as in [3, 6, 16]. This approach has been recently embedded into the framework of Gaussian mixture models, leading to multivariate SU normal regression mixtures [7]. In these models, the effect of the regressors on the dependent variables changes with some unknown latent sub-populations composing the population that has generated the sample of observations to be analysed. Thus, when the sample is characterised by unobserved heterogeneity, model-based cluster analysis is simultaneously carried out.

Another source of complexity which could affect the data and make the prediction of $\mathbf{Y}$ a difficult task to perform is represented by mildly atypical observations [13]. Robust methods of parameter estimation insensitive to the presence of such observations in a sample characterised by unobserved heterogeneity have been introduced in [9], where the conditional distribution $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ is modelled through a mixture of K multivariate contaminated normal models, where K is the number of the latent sub-populations. A limitation associated with these latter models is that the same vector of regressors has to be specified for the prediction of all the dependent variables. To overcome this limitation while preserving all the features mentioned above, a more flexible approach which employs mixtures of multivariate SU contaminated normal regression models has been recently introduced in [11]. These latter models are able to capture the linear effects of the regressors on the dependent variables from sample observations coming from heterogeneous populations. The researcher is also enabled to specify a different vector of regressors for each dependent variable. Finally, a robust estimation of the regression parameters and the detection of mild outliers in the data are ensured.

In the presence of many responses and many latent sub-populations, analyses based on these latter models can become unfeasible in practical applications because of a large number of model parameters. In order to keep this number as low as possible, an approach due to [4], based on the spectral decompositions of the K covariance matrices of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$, is exploited here so as to obtain fourteen different covariance structures. The resulting parsimonious mixtures of SU contaminated regression models are described in Section 2. The usefulness of these new models is illustrated through a study aiming at determining the effect of prices and promotional activities on sales of canned tuna in the US market. A summary of the obtained results is provided in Section 3.

2 Parsimonious SU Contaminated Normal Regression Mixtures

In a system of M SU regression equations for modelling the linear dependence of $\mathbf{Y}$ on $\mathbf{X}$, let $\mathbf{X}_m = (X_{m_1}, X_{m_2}, \ldots, X_{m_{P_m}})'$ be the P_m-dimensional sub-vector of $\mathbf{X}$ composed of the P_m regressors expected to be relevant for the explanation of Y_m, for $m = 1, \ldots, M$. Furthermore, let $\mathbf{X}_m^* = (1, \mathbf{X}_m')'$. The mixture of K SU normal regression models described in [7] can be defined as follows:

$$\mathbf{Y} = \begin{cases} \tilde{\mathbf{X}}^{*\prime}\boldsymbol{\beta}_1^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_M(\mathbf{0}_M, \boldsymbol{\Sigma}_1) \text{ with probability } \pi_1, \\ \cdots \\ \tilde{\mathbf{X}}^{*\prime}\boldsymbol{\beta}_K^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_M(\mathbf{0}_M, \boldsymbol{\Sigma}_K) \text{ with probability } \pi_K, \end{cases} \tag{1}$$

where π_k is the prior probability of the kth latent sub-population, with $\pi_k > 0$ for $k = 1, \ldots, K$; $\sum_{k=1}^{K} \pi_k = 1$; $\tilde{\mathbf{X}}^*$ is the following $(P^* + M) \times M$ partitioned matrix:

$$\tilde{\mathbf{X}}^* = \begin{bmatrix} \mathbf{X}_1^* & \mathbf{0}_{P_1+1} & \ldots & \mathbf{0}_{P_1+1} \\ \mathbf{0}_{P_2+1} & \mathbf{X}_2^* & \ldots & \mathbf{0}_{P_2+1} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{P_M+1} & \mathbf{0}_{P_M+1} & \ldots & \mathbf{X}_M^* \end{bmatrix},$$

with $\mathbf{0}_{P_m+1}$ denoting the $(P_m + 1)$-dimensional null vector; $P^* = \sum_{m=1}^{M} P_m$; $\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_{k1}^{*\prime}, \ldots, \boldsymbol{\beta}_{km}^{*\prime}, \ldots, \boldsymbol{\beta}_{kM}^{*\prime})'$ is the $(P^* + M)$-dimensional vector containing all the linear effects on the M responses in the kth latent sub-population, with $\boldsymbol{\beta}_{km}^* = (\beta_{0k,m}, \boldsymbol{\beta}_{km}')'$, for $m = 1, \ldots, M$; $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)'$ is the vector of the errors, which are supposed to be independent and identically distributed; $N_M(\mathbf{0}_M, \boldsymbol{\Sigma}_k)$ denotes the M-dimensional normal distribution with mean vector $\mathbf{0}_M$ and positive-definite covariance matrix $\boldsymbol{\Sigma}_k$. From now on, this mixture regression model is denoted as MSUN. When $\mathbf{X}_m = \mathbf{X}\ \forall m$ (the P regressors are employed in all the M equations), model (1) reduces to the mixtures of K normal (MN) regression models (see [8]).

When the data are contaminated by the presence of mild outliers, departures from the normal distribution could be observed within any of the K latent sub-populations. A model able to manage this situation has been recently introduced in [11]. It has been obtained from equation (1) by replacing the normal distribution with the contaminated normal distribution. Under this latter distribution, the probability density function (p.d.f.) of $\boldsymbol{\epsilon}$ within the kth sub-population is equal to $h(\boldsymbol{\epsilon}; \boldsymbol{\vartheta}_k) = \alpha_k \phi_M(\boldsymbol{\epsilon}; \mathbf{0}_M, \boldsymbol{\Sigma}_k) + (1 - \alpha_k)\phi_M(\boldsymbol{\epsilon}; \mathbf{0}_M, \eta_k\boldsymbol{\Sigma}_k)$, where $\phi_M(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p.d.f. of the distribution $N_M(\mathbf{0}_M, \boldsymbol{\Sigma}_k)$, $\alpha_k \in (0.5, 1)$ and $\eta_k > 1$ are the proportion of typical observations within the kth sub-population and a parameter that inflates the elements of $\boldsymbol{\Sigma}_k$, respectively, and $\boldsymbol{\vartheta}_k = (\alpha_k, \eta_k, \boldsymbol{\Sigma}_k)$. As a consequence, a mixture of K SU contaminated normal (MSUCN) regression models is given by:

$$\mathbf{Y} = \begin{cases} \tilde{\mathbf{X}}^{*\prime}\boldsymbol{\beta}_1^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim CN_M(\alpha_1, \eta_1, \mathbf{0}_M, \boldsymbol{\Sigma}_1) \text{ with probability } \pi_1, \\ \cdots \\ \tilde{\mathbf{X}}^{*\prime}\boldsymbol{\beta}_K^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim CN_M(\alpha_K, \eta_K, \mathbf{0}_M, \boldsymbol{\Sigma}_K) \text{ with probability } \pi_K, \end{cases} \tag{2}$$

where $CN_M(\alpha_k, \eta_k, \mathbf{0}_M, \mathbf{\Sigma}_k)$ denotes the M-dimensional contaminated normal distribution described by the p.d.f. $h(\boldsymbol{\epsilon}; \boldsymbol{\vartheta}_k)$. The parameter vector of model (2) is $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k, \ldots, \boldsymbol{\psi}_K)$, where $\boldsymbol{\psi}_k = (\pi_k, \boldsymbol{\theta}_k)$, $\boldsymbol{\theta}_k = (\boldsymbol{\beta}^*_k, \boldsymbol{\vartheta}_k)$. The number of free elements of $\boldsymbol{\psi}$ is $n_{\boldsymbol{\psi}} = 3K - 1 + K(P^* + M) + n_{\sigma}$, where n_{σ} denotes the total number of free variances and covariances, with $n_{\sigma} = K n_{\Sigma}$ and $n_{\Sigma} = \frac{M(M+1)}{2}$. When $\mathbf{X}_m = \mathbf{X}\, \forall m$, model (2) coincides with the mixture of K contaminated normal (MCN) regression models described in [9]. For $\alpha_k \rightarrow 1$ or $\eta_k \rightarrow 1\ \forall k$, model (2) reduces to model (1). Conditions ensuring identifiability of models (2) are provided in [11]. The ML estimation of $\boldsymbol{\psi}$ in equation (2) can be carried out by means of a sample $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I)\}$ of I independent observations drawn from model (2) and an expectation-conditional maximisation (ECM) algorithm [10]. Details about this algorithm, including strategies for the initialisation of $\boldsymbol{\psi}$ and convergence criteria, are illustrated in [11]. In practical applications, the value of K is generally unknown and has to be properly chosen. This task can be carried out by resorting to model selection criteria, such as the Bayesian information criterion [15]: $BIC = 2\ell(\hat{\boldsymbol{\psi}}) - n_{\boldsymbol{\psi}} \ln I$, where $\hat{\boldsymbol{\psi}}$ is the maximum likelihood estimator of $\boldsymbol{\psi}$. Another commonly used information criterion is the integrated completed likelihood [2], which admits two slightly different formulations: $ICL_1 = BIC + 2\sum_{i=1}^{I}\sum_{k=1}^{K} \text{MAP}(\hat{z}_{ik}) \ln \hat{z}_{ik}$ and $ICL_2 = BIC + 2\sum_{i=1}^{I}\sum_{k=1}^{K} \hat{z}_{ik} \ln \hat{z}_{ik}$, where $\hat{z}_{ik}$ is the estimated posterior probability that the ith sample observation come from the kth sub-population (for further details see [11]), $\text{MAP}(\hat{z}_{ik}) = 1$ if $\max_h\{\hat{z}_{ih}\}$ occurs when $h = k$ ($\text{MAP}(\hat{z}_{ik}) = 0$ otherwise). Whenever the specification of the subvectors $\mathbf{X}_m$, $m = 1, \ldots, M$, to be considered in the M equations of the multivariate regression model is questionable, such criteria can also be employed to perform subset selection.

As the number of free parameters $n_{\boldsymbol{\psi}}$ incresases quadratically with M, analyses based on model (2) can become unfeasible in real applications. A way to manage this problem can be based on the introduction of suitable constraints on the elements of $\mathbf{\Sigma}_k$, $k = 1, \ldots, K$, based on the following eigen-decomposition [4]: $\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k$, where $\lambda_k = |\mathbf{\Sigma}_k|^{1/M}$, $\mathbf{A}_k$ is a diagonal matrix with entries (sorted in decreasing order) proportional to the eigenvalues of $\mathbf{\Sigma}_k$ (with the constraint $|\mathbf{A}_k| = 1$) and $\mathbf{D}_k$ is a $M \times M$ orthogonal matrix of the eigenvectors of $\mathbf{\Sigma}_k$ (ordered according to the eigenvalues). This decomposition allows to obtain variances and covariances in $\mathbf{\Sigma}_k$ from λ_k, $\mathbf{A}_k$ and $\mathbf{D}_k$. From a geometrical point of view, λ_k determines the volume, $\mathbf{A}_k$ the shape and $\mathbf{D}_k$ the orientation of the kth cluster of sample observations detected by the fitted model. By constraining λ_k, $\mathbf{A}_k$ and $\mathbf{D}_k$ to be equal or variable across the K clusters, a class of fourteen mixtures of K SUCN regression models is obtained (see Table 1). With variable volumes, shapes and orientations (VVV in Table 1), the resulting model coincides with (2). When $K > 1$, the other covariance structures allow to obtain thirteen different parsimonious mixtures of K SUCN regression models (i.e.: with a reduced n_{σ}). When $K = 1$, the possible covariance structures for $\mathbf{\Sigma}_1$ are: diagonal with different entries, diagonal with the same entries and fully unconstrained. The ML estimation of $\boldsymbol{\psi}$ under model (2) with any of these parameterisations can be carried out through an ECM algorithm in which the CM-step update for $\mathbf{\Sigma}_k$ can be computed either in closed form or using iterative procedures, depending on the parameterisation to be employed (see [4]).

Table 1 Features of the parameterisations for the covariance matrices $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$ $(K > 1)$.

Acronym	Covariance structure	Volume	Shape	Orientation	CM step	n_σ
EEE	$\lambda\mathbf{DAD}'$	Equal	Equal	Equal	Closed	n_Σ
VVV	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k'$	Variable	Variable	Variable	Closed	Kn_Σ
EII	$\lambda\mathbf{I}$	Equal	Spherical	–	Closed	1
VII	$\lambda_k\mathbf{I}$	Variable	Spherical	–	Closed	K
EEI	$\lambda\mathbf{A}$	Equal	Equal	Axis-aligned	Closed	M
VEI	$\lambda_k\mathbf{A}$	Variable	Equal	Axis-aligned	Iterative	$M + K - 1$
EVI	$\lambda\mathbf{A}_k$	Equal	Variable	Axis-aligned	Closed	$MK - (K - 1)$
VVI	$\lambda_k\mathbf{A}_k$	Variable	Variable	Axis-aligned	Closed	MK
EEV	$\lambda\mathbf{D}_k\mathbf{AD}_k'$	Equal	Equal	Variable	Iterative	$Kn_\Sigma - (K - 1)M$
VEV	$\lambda_k\mathbf{D}_k\mathbf{AD}_k'$	Variable	Equal	Variable	Iterative	$Kn_\Sigma - (K - 1)(M - 1)$
EVE	$\lambda\mathbf{DA}_k\mathbf{D}'$	Equal	Variable	Equal	Iterative	$n_\Sigma - (K - 1)(M - 1)$
VVE	$\lambda_k\mathbf{DA}_k\mathbf{D}'$	Variable	Variable	Equal	Iterative	$n_\Sigma - (K - 1)M$
VEE	$\lambda_k\mathbf{DAD}'$	Variable	Equal	Equal	Iterative	$n_\Sigma - (K - 1)$
EVV	$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k'$	Equal	Variable	Variable	Iterative	$Kn_\Sigma - (K - 1)$

3 Analysis of U.S. Canned Tuna Sales

The models illustrated in Section 2 have been fitted to a dataset [5] containing the volume of sales (Move), a measures of the display activity (Nsale) and the log price (Lprice) for seven of the top 10 U.S. brands in the canned tuna product category in the $I = 338$ weeks between September 1989 and May 1997. The goal of the analysis is to study the dependence of canned tuna sales on prices and promotional activites for two products: Star Kist 6 oz. (SK) and Bumble Bee Solid 6.12 oz. (BBS). To this end, the following vectors have been considered: $\mathbf{Y}' = (Y_1 =$ Lmove SK, $Y_2 =$ Lmove BBS), $\mathbf{X}' = (X_1 =$ Nsale SK, $X_2 =$ Lprice SK, $X_3 =$ Nsale BBS, $X_4 =$ Lprice BBS), where Lmove denotes the logarithm of Move. The analysis has been carried out using all the parameterisations of the MSUN, MN, MCSUN and MCN models for each $K \in \{1, 2, 3, 4, 5, 6\}$. Furthermore, MSUN and MCSUN models have been fitted by considering all possible subvectors of $\mathbf{X}$ as vectors $\mathbf{X}_m$, $m = 1, 2$, for each K. In this way, best subset selections for Lmove SK and Lmove BBS have been included in the analysis both with and without contamination. The overall number of fitted models is 37376, including the fully unconstrained models (i.e., with the VVV parameterisation) previously employed in [11] to perform the same analysis.

Table 2 reports some information about the nine models which best fit the analysed dataset according to the three model selection criteria over the six examined values of K within each model class. An analysis based on a single linear regression model $(K = 1)$, both with and without contamination, appears to be inadequate according to all criteria. All the examined criteria indicate that the overall best model for studying the effect of prices and promotional activities on sales of SK and BBS tuna is a parsimonious mixture of two SU contaminated Gaussian linear regression models with the EVE parameterisation for the covariance matrices in which the log unit sales of SK tuna are regressed on the log prices and the promotional activites of the same brand, while the regressors selected for the BBS log unit sales are the log prices of both brands and the promotional activites of BBS. Thus, the analysis suggests that two sources of complexity affect the analysed dataset: unobserved heterogeneity over time ($K = 2$ clusters of weeks have been detected) and the presence of mildly atypical observations. Since the two estimated proportions of typical observations are quite similar (see the values of $\hat{\alpha}_k$ in Table 3), contamination seems to characterise the two clusters of weeks detected by the model almost in the same way. As far as the strength of the contaminating effects on the conditional variances and covariances of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ is concerned, it appears to be stronger in the first cluster, where the estimated inflation parameter is larger ($\hat{\eta}_1 = 15.70$). By focusing the attention on the other estimates, it appears that also some of the estimated regression coefficients, variances and covariances are affected by heterogeneity over time. Sales of SK tuna results to be negatively affected by prices and positively affected by promotional activites of the same brand within both clusters detected by the model, but with effects which are sligthly stronger in the first cluster of weeks. A similar behavior is detected for the estimated regression equation for `Lmove BBS`, which also highlights that `Lmove BBS` are positively affected by the log prices of SK tuna, especially in the first cluster of weeks. Furthermore, typical weeks in the first cluster show values of `Lmove SK` which are more homogeneous than those of `Lmove BBC`; the opposite holds true for the typical weeks belonging to the second cluster. Also the correlation between log sales of SK and BBS products results to be affected by heterogeneity over time: while in the largest cluster of weeks this correlation has been estimated to be slightly positive (0.200), the first cluster is characterised by a mild estimated negative correlation (−0.151). An interesting feature of this latter cluster is that 17 out of the 20 weeks which have been assigned to this cluster are consecutive from week no. 58 to week no. 74, which correspond to the period from mid-October 1990 to mid-February 1991 characterised by a worldwide boycott campaign encouraging consumers not to buy Bumble Bee tuna because Bumble Bee was found to be buying yellow-fin tuna caught by dolphin-unsafe techniques [1]. Such events could represent one of the sources of the unobserved heterogeneity detected by the model. According to the overall best model, some weeks have beed detected to be mild outliers. In the first cluster, this has happened for week no. 60 (immediately after Halloween 1990) and week no. 73 (two weeks immediately before Presidents day 1999). The analysis of the estimated sample residuals $\mathbf{y}_i - \hat{\boldsymbol{\mu}}_1(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_1^*)$ for the 20 weeks belonging to the first cluster (see the scatterplot on the left side of Figure 1) clearly show that weeks 60 and 73 noticeably deviates from the other weeks. Among the 318 weeks of the second cluster, 32 have resulted to be mild outliers, most of which are associated with holidays and special events that took place between September 1989 and mid-October 1990 or between mid-February and May 1997 (see the scatterplot on the right side of Figure 1). These results are almost equal to those obtained using the best overall fully unconstrained fitted model in the analysis presented in [11]. However, the `EVE` parameterisation for the MSUCN model has allowed to obtain a better trade-off among the fit, the model complexity and the uncertainty of the estimated partition of the weeks; furthermore, it has led to a slightly lower number of mild outliers in the second cluster of weeks.

Table 2 Maximised log-likelihood $\ell(\hat{\psi})$ and values of BIC, ICL_1 and ICL_2 for nine models selected from the classes MSUCN, MCN, MSUN and MN in the analysis of tuna sales.

Model class	K	Acronym	$\mathbf{X}_1$	$\mathbf{X}_2$	$\ell(\hat{\psi})$	n_ψ	BIC	ICL_1	ICL_2
MSUCN	2	EVE	X_1, X_2	X_2, X_3, X_4	−242.9	23	−619.8	−625.7	−635.8
MCN	2	EVI	$\mathbf{X}$	$\mathbf{X}$	−239.6	28	−642.2	−648.9	−663.2
MCN	2	EEV	$\mathbf{X}$	$\mathbf{X}$	−240.8	29	−650.6	−650.8	−652.0
MCN	3	EVI	X_1, X_2, X_4	X_1, X_2, X_4	−214.2	36	−638.0	−703.1	−788.6
MSUN	2	VEV	X_1, X_2	X_3, X_4	−279.3	18	−663.4	−673.1	−692.1
MSUN	3	EEV	X_2, X_3	X_2, X_3, X_4	−259.8	28	−682.7	−684.7	−688.0
MSUN	5	VVV	X_2, X_3	X_1, X_4	−167.4	49	−620.0	−701.1	−780.3
MN	3	EEV	X_2, X_3, X_4	X_2, X_3, X_4	−258.7	31	−697.9	−699.6	−702.1
MN	4	VVE	X_2, X_4	X_2, X_4	−216.6	36	−642.9	−725.3	−832.9

Table 3 Parameter estimates of the overall best model for the analysis of tuna sales.

$\hat{\psi}$	$k = 1$	$k = 2$
$\hat{\pi}_k$	0.062	0.938
$\hat{\alpha}_k$	0.810	0.844
$\hat{\eta}_k$	15.70	6.94
$\hat{\beta}'^{*}_{k1}$	$(8.87, 0.56, -4.70)$	$(8.64, 0.27, -3.09)$
$\hat{\beta}'^{*}_{k2}$	$(15.04, 3.92, 2.83, -17.76)$	$(9.98, 0.25, 0.12, -3.83)$
$\hat{\boldsymbol{\Sigma}}_k$	$\begin{pmatrix} 0.034 & -0.009 \\ -0.009 & 0.105 \end{pmatrix}$	$\begin{pmatrix} 0.121 & 0.012 \\ 0.012 & 0.030 \end{pmatrix}$

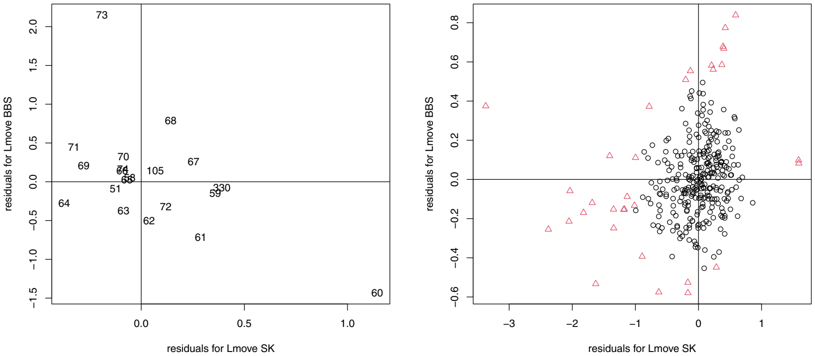

Fig. 1 Scatterplots of the estimated residuals for the weeks assigned to the first (left) and second (right) clusters detected by the overall best model. Points of the first scatterplot are labelled with the number of the corresponding weeks. Black circle and red triangle in the second scatterplot correspond to typical and outlying weeks, respectively.

4 Conclusions

The parsimonious mixtures of seemingly unrelated linear regression models for contaminated data introduced here can account for heterogeneous regression data both in the presence of mild outliers and multivariate correlated dependent variables, each of which is regressed on a different vector of covariates. Models from this class allow for simultaneous robust clustering and detection of mild outliers in multivariate regression analysis. They encompass several other types of Gaussian mixture-based linear regression models previously proposed in the literature, such as the ones illustrated in [7, 8, 9], providing a robust and flexible tool for modelling data in practical applications where different regressors are considered to be relevant for the prediction of different dependent variables. Previous research (see [9, 11]) demonstrated that BIC and ICL could be effectively employed to select a proper value for K in the presence of mildly contaminated data. Thanks to an imposition of an eigen-decomposed structure on the K variance-covariance matrices of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$, the presented models are characterised by a reduced number of variance-covariance parameters to be included in the analysis, thus improving flexibility, usefulness and effectiveness of an approach to multivariate linear regression analysis based on finite Gaussian mixture models in real data applications.

References

1. Baird, I. G., Quastel, N.: Dolphin-safe tuna from California to Thailand: localisms in environmental certification of global commodity networks. Ann. Assoc. Am. Geogr. **101**, 337–355 (2011)
2. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 719–725 (2000)
3. Cadavez, V. A. P., Hennningsen, A.: The use of seemingly unrelated regression (SUR) to predict the carcass composition of lambs. Meat. Sci. **92**, 548–553 (2012)
4. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**, 781–793 (1995)
5. Chevalier, J. A., Kashyap, A. K., Rossi, P. E.: Why don't prices rise during periods of peak demand? Evidence from scanner data. Am. Econ. Rev. **93**, 15–37 (2003)
6. Disegna, M., Osti, L.: Tourists' expenditure behaviour: the influence of satisfaction and the dependence of spending categories. Tour. Econ. **22**, 5–30 (2016)
7. Galimberti, G., Soffritti, G.: Seemingly unrelated clusterwise linear regression. Adv. Data Anal. Classif. **14**, 235–260 (2020)
8. Jones, P. N., McLachlan, G. J.: Fitting finite mixture models in a regression context. Aust. New Zeal. J. Stat. **34**, 233–240 (1992)
9. Mazza, A., Punzo, A.: Mixtures of multivariate contaminated normal regression models. Stat. Pap. **169**, 787–822 (2020)
10. Meng, X. L., Rubin, D. B.: Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. **80**, 267–278 (1993)
11. Perrone, G., Soffritti, G.: Seemingly unrelated clusterwise linear regression for contaminated data. Under review (2021)
12. R Core Team R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2022) `http://www.R-project.org`
13. Ritter, G.: Robust cluster analysis and variable selection. Chapman & Hall, Boca Raton (2015)
14. Srivastava, V. K., Giles, D. E. A.: Seemingly unrelated regression equations models. Marcel Dekker, New York (1987)
15. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
16. White, E. N., Hewings, G. J. D.: Space-time employment modelling: some results using seemingly unrelated regression estimators. J. Reg. Sci. **22**, 283–302 (1982)

Penalized Model-based Functional Clustering: a Regularization Approach via Shrinkage Methods

Nicola Pronello, Rosaria Ignaccolo, Luigi Ippoliti, and Sara Fontanella

Abstract With the advance of modern technology, and with data being recorded continuously, functional data analysis has gained a lot of popularity in recent years. Working in a mixture model-based framework, we develop a flexible functional clustering technique achieving dimensionality reduction schemes through a L_1 penalization. The proposed procedure results in an integrated modelling approach where shrinkage techniques are applied to enable sparse solutions in both the means and the covariance matrices of the mixture components, while preserving the underlying clustering structure. This leads to an entirely data-driven methodology suitable for simultaneous dimensionality reduction and clustering. Preliminary experimental results, both from simulation and real data, show that the proposed methodology is worth considering within the framework of functional clustering.

Keywords: functional data analysis, L_1-penalty, silhouette width, graphical LASSO, mixture model

Nicola Pronello (✉)
Department of Neurosciences, Imaging and Clinical Sciences, University of Chieti-Pescara, Chieti, Italy, e-mail: nicola.pronello@unich.it

Rosaria Ignaccolo
Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Torino, Italy, e-mail: rosaria.ignaccolo@unito.it

Luigi Ippoliti
Department of Economics, University of Chieti-Pescara, Pescara, Italy,
e-mail: luigi.ippoliti@unich.it

Sara Fontanella
National Heart and Lung Institute, Imperial College London, London, United Kingdom,
e-mail: s.fontanella@imperial.ac.uk

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_34

1 Introduction

In recent decades, technological innovations have produced data that are increasingly complex, high dimensional, and structured. A large amount of these data can be characterized as functions defined on some continuous domain and their statistical analysis has attracted the interest of many researchers. This surge of interests is explained by the ubiquitous examples of functional data that can be found in different application fields (see for example [2], and references therein for specific examples). With functions as the basic units of observation, the analysis of functional data poses significant theoretical and practical challenges to statisticians. Despite these difficulties, methodology for clustering functional data has advanced rapidly during the past years; recent surveys of functional data clustering are presented in [7] and [2]. Popular approaches have extended classical clustering concepts for vector-valued multivariate data to functional data.

In this paper, we consider a finite mixture as a flexible model for clustering. In particular, applying a functional model-based clustering algorithm with an L_1-penalty function on a set of projection coefficients, we extend the results of [8] and [9] for vector-valued multivariate data to a functional data framework. This approach appears particularly appealing in all cases in which the functions are spatially heterogeneous, meaning that some parts of the function can be smoother than in other parts, or that there may be distant parts of the function that are correlated with each other. Furthermore, the introduction of a shrinkage penalty allows to look for directions in the feature space (that is now the space of expansion/projection coefficients) that are the most useful in separating the underlying groups without first applying dimensionality reduction techniques.

In Section 2 we present at first the methodology along with some details on model estimation (subsection 2.2). Secondly, in Section 3, we perform a validation study with simulated and real data for which the classes are known a-priori.

2 Shrinkage Method for Model-based Clustering for Functional Data

Here we consider the problem of clustering a set of n observed curves into K homogeneous groups (or clusters). To this end, we propose a flexible model based on a finite mixture of Gaussian distributions, with a L_1 penalized likelihood, which we name *Penalized model-based Functional Clustering* (PFC-L_1).

2.1 Model Definition

We consider a set of n observed curves, $x_1, \dots, x_n$, that are independent realizations of a continuous stochastic process $X = \{X(t)\}_{t\in[0,T]}$ taking values in $L_2[0,T]$. In practice, such curves/trajectories are available only at a discrete set of the domain points $\{t_{is} : i = 1, \ldots, n,\ \ s = 1, \ldots, m_i\}$ and the n curves need to be reconstructed. To this goal, it is common to assume that the curves belong to a finite dimensional space spanned by a basis of functions, so that given a basis of functions $\mathbf{\Phi} = \{\psi_1, ..., \psi_p\}$ each curve $x_i(t)$ admits the following decomposition:

$$x_i(t) = \sum_{j=1}^{p} \beta_{j,i}\psi_j(t), \qquad i = 1, \ldots, n; \tag{2.1}$$

that is the stochastic process X admits a corresponding truncated basis expansion

$$X(t) = \sum_{j=1}^{p} \beta_j(X)\psi_j(t),$$

where $\boldsymbol{\beta} = \{\beta_1(X), \ldots, \beta_p(X)\}$ is a random vector in $\mathbb{R}^p$. By considering observations with a sampling error, such that

$$x_i^{obs}(t) = x_i(t) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{2.2}$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, the realizations of the random coefficients $\beta_{j,i}$ for $j = 1, \ldots, p$ describing each curve can be obtained via least squares as $\hat{\boldsymbol{\beta}}_i = (\mathbf{\Theta}_i^{'}\mathbf{\Theta}_i)^{-1}\mathbf{\Theta}_i^{'}\mathbf{X}_i^{obs}$ where $\mathbf{\Theta}_i = (\psi_j(t_{is})), 1 \leq j \leq p, 1 \leq s \leq m_i$ contains the basis functions evaluated at the fixed domain points and $\mathbf{X}_i^{obs} = (x_i^{obs}(t_{i1}), \ldots, x_i^{obs}(t_{im_i}))'$ is the vector of observed values of the i-th curve.

With the goal of dividing into K homogeneous groups the observed curves $x_1, \ldots, x_n$, let us assume that it exists an unobservable grouping variable $\mathbf{Z} = (Z_1, ..., Z_K) \in [0, 1]^K$ indicating the cluster membership: $z_{i,k} = 1$ if x_i belongs to cluster k, 0 otherwise (and $z_{i,k}$ is indeed what we want to predict for each curve).

In adopting a model-based clustering approach, we denote with π_k the (a-priori) probabilities of belonging to a group:

$$\pi_k = \mathbb{P}(Z_k = 1), \qquad k = 1, \ldots, K,$$

such that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ for each k, and we assume that, conditionally on Z, the random vector $\boldsymbol{\beta}$ follows a multivariate Gaussian distribution, that is for each cluster

$$\boldsymbol{\beta}|(Z_k = 1) = \boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$$

where $\boldsymbol{\mu}_k = (\mu_{1,k}, \ldots, \mu_{p,k})^T$ and $\mathbf{\Sigma}_k$ are respectively the mean vector and the covariance matrix of the k-th group. Then the marginal distribution of $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$ can be written as a finite mixture with mixing proportions π_k as

$$p(\boldsymbol{\beta}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{\beta}_k; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k),$$

where f is the multivariate Gaussian density function. The log-likelihood function can then be written as

$$l(\boldsymbol{\theta};\boldsymbol{\beta}) = \sum_{i=1}^{n} log \sum_{k=1}^{K} \pi_k f(\boldsymbol{\beta}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

where $\boldsymbol{\theta} = \{\pi_1,\ldots,\pi_K;\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K;\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K\}$ is the vector of parameters to be estimated and $\boldsymbol{\beta}_i = (\beta_{1,i},\ldots,\beta_{p,i})^T$ is the vector of projection coefficients of the i-th curve.

In this modeling framework, we consider a very general situation without introducing any kind of constraints neither for cluster means nor for covariance matrices, that can be different in each cluster. This flexibility, however, leads to overparameterization and, as an alternative to any kind of constraints, we consider a penalty that allows regularized parameters' estimation.

To define a suitable penalty term, we follow the penalized approach introduced by Zhou et al. [8] in the high-dimensional setting, and so we consider a penalty composed by two terms: the first one on the mean vector of each cluster $\boldsymbol{\mu}_k$, and the second one on the inverse of the covariance matrix in each group $\mathbf{W}_k = \boldsymbol{\Sigma}_k^{-1}$, otherwise said "precision" matrix, with elements $W_{k;j,l}$. The proposed penalized log-likelihood function, given the projection coefficients $\boldsymbol{\beta}_i$, is

$$l_P(\boldsymbol{\theta};\boldsymbol{\beta}) = \sum_{i=1}^{n} log \sum_{k=1}^{K} \pi_k f(\boldsymbol{\beta}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) - \lambda_1 \sum_{k=1}^{K} ||\boldsymbol{\mu}_k||_1 - \lambda_2 \sum_{k=1}^{K} \sum_{j,l}^{p} |W_{k;j,l}|,$$

where $||\boldsymbol{\mu}_k||_1 = \sum_{j=1}^{p} |\mu_{k,j}|$, $\lambda_1 > 0$ and $\lambda_2 > 0$ are penalty parameters to be suitably chosen.

The penalty term on the cluster mean vectors allow for component selection in the functional data framework (whereas it would be variable selection in the multivariate case), considering that when the j-th component in the basis expansion is not useful in separating groups it has a common mean across groups, that is $\mu_{1,j} = \ldots = \mu_{K,j} = 0$. Then to realize component selection the considered term is $\sum_{k=1}^{K} ||\boldsymbol{\mu}_k||_1$.

The second part of the penalty, namely $\sum_{k=1}^{K} \sum_{j,l}^{p} |W_{k;j,l}|$, imposes a shrinkage on the elements of the precision matrices, thus avoiding possible singularity problems and facilitating the estimation of large and sparse covariance matrices.

2.2 Model Estimation via E-M Algorithm

Since the membership of each observation to a cluster is unobservable, data related to the grouping variable $\mathbf{Z}$ is inevitably missing and the maximum penalized log-likelihood estimator can be obtained by means of the E-M algorithm [4], that iterates over two steps: expectation (E) of the complete data (penalized) log-likelihood by considering the unknown parameters equal to those obtained at the previous iteration (with initialization values), and maximization (M) of a lower bound of the obtained expected value with respect to the unknown parameters.

In particular, at the d-th iteration, given a current estimate $\boldsymbol{\theta}^{(d)}$, the lower bound after the E-step assumes the following form:

$$Q_P(\boldsymbol{\theta};\boldsymbol{\theta}^{(d)})=\sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{k,i}^{(d)}\left[\log\pi_k+\log f(\boldsymbol{\beta}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right]-\lambda_1\sum_{k=1}^{K}||\boldsymbol{\mu}_k||_1-\lambda_2\sum_{k=1}^{K}\sum_{j,l}^{p}|W_{k;j,l}|,$$

where $\tau_{k,i} = \mathbb{P}(Z_k = 1|X = x_i)$ is the posterior probability of observation i to belong to group k. The M-step maximizes the function Q_P in order to update the estimate of $\boldsymbol{\theta}$.

As suggested by [9], it is possible to maximize each of the K term using a “graphical lasso” (GLASSO) algorithm (first proposed by [5]), thanks to the close connection between fitting Gaussian mixture models and Gaussian graphical models. Indeed, in GLASSO the objective function looks like $\log\det(\mathbf{W}) - \mathrm{tr}(\mathbf{SW}) - \lambda\sum_{j,l}^{p}|W_{j,l}|$ so that the algorithm implemented in the R package “`glasso`” can be used with $\mathbf{W} = \mathbf{W}_k$, $S = \tilde{\mathbf{S}}_k$ and $\lambda = \frac{2\lambda_2}{\sum_{i=1}^{n}\tau_{k,i}^{(d)}}$ for each k to obtain the elements $\widehat{W}_{k;j,l}^{(d+1)}$ of the precision matrices.

2.3 Model Selection via Silhouette Profile

A fundamental, and probably unsolved, problem in cluster analysis is determining the “true” number of groups in a dataset. To this purpose, for simplicity, here we approach the problem choosing the number of groups as cluster validation problem and use the *average silhouette width* index as a model selection heuristic. The silhouette value for curve i is given by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance of curve i to all other curves h assigned to the same cluster (if i is the only observation in its cluster, then $s(i) = 0$), and $b(i)$ is the minimum average distance of curve i to observations h which are assigned to a different cluster. This definition ensures that $s(i)$ takes values in $[-1, 1]$, where values close to one indicate “better”clustering solutions. Conditional on K and a pair of values (λ_1, λ_2), we thus assess the overall cluster solution using the total average of silhouette values

$$S(K, \lambda_1, \lambda_2) = \frac{1}{n}\sum_{i=1}^{n} s(i).$$

In particular, by doing a grid search for the triple $(K, \lambda_1, \lambda_2)$, the best cluster solution is obtained by looking for the largest value of the *average silhouette width* (*ASW*) index. Note that, to evaluate $s(i), i = 1, \ldots, n$, and then the objective function $S(K, \lambda_1, \lambda_2)$, we need to compute a distance between pairs of curves X_i and X_h. One possibility is to compute the euclidean distance

$$d_E^2(i,h) = \int \|X_i(t) - X_h(t)\|^2 dt.$$

3 Experimental Results

3.1 Simulation

We present here a simulated scenario in order to investigate the effectiveness of the L_1 regularization in removing noise while preserving dominant local features, accommodating for spatial heterogeneity of the curves.

The statistical analysis is illustrated for data simulated by means of a finite mixture of multivariate Gaussian distributions. In particular, based on equation (2.1) and (2.2), the curves are simulated using a combination of $p = 25$ Fourier basis functions defined over a one-dimensional regular grid with 100 observations. We consider a mixture of four ($K = 4$) multivariate Gaussian distributions with isotropic covariance matrices, i.e.

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}_k; \mathcal{I}_k) \text{ where } \epsilon_i \sim \mathcal{N}(0; 0.5), \quad k = 1, \ldots, 4.$$

With the exclusion of 3 entries per group, the means $\boldsymbol{\mu}_k$ are all zero mean vectors. Under this scenario, the simulated curves (25 per group) and the non-zero group expansion coefficients are represented in Figure 1. For this simple simulation setting, estimation results suggest that, using euclidean distance to computed the *ASW*, the grid search procedure is always able to correctly select the cluster-relevant basis functions. This is confirmed by Figure 2 which shows both the distribution (over 100 replications) of the selected basis functions and the data projected on these bases that clearly highlight the identification of 4 clusters. Under this scenario, the quality of the estimated clusters thus appears very good as the analysis of the misclassification rate suggests an 100% of accuracy in all the replicated datasets.

Similar results hold for more complex simulation designs, where we consider different structure of the covariance matrices in the data generating process.

3.2 Performance on Real Data Sets

We evaluate the PFC-L_1 model on a well-known benchmark data set, namely the electrocardiogram (ECG) data set (data can be found at the UCR Time Series Classification Archive [3]).

The ECG data set comprises a set of 200 electrocardiograms from 2 groups of patients, myocardial infarction and healthy, sampled at 96 time instants in time.

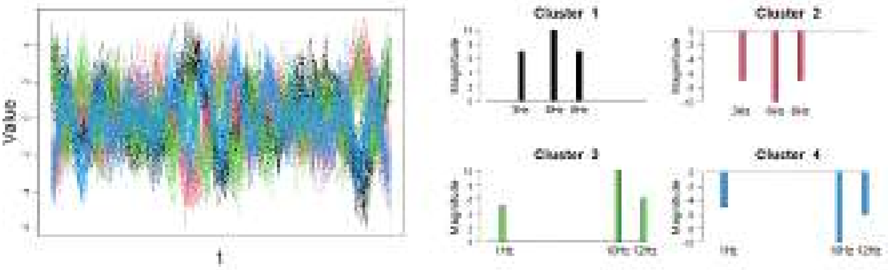

Fig. 1 Left: 25 simulated curves for each group. Right: Vector of expansion coefficients for each group, with only three non-zero coefficients corresponding to basis functions with specific periodicities (Hertz values).

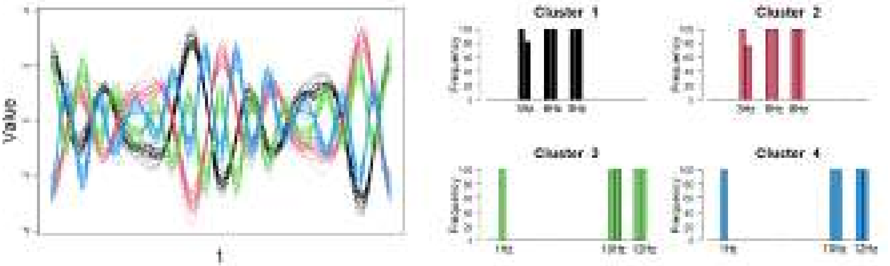

Fig. 2 Left: Data projected on cluster specific functional subspace generated by the selected basis functions. Right: Distribution (over 100 replications) of the selected basis functions shown for pairs of sine and cosine basis functions, according to the Hertz values.

This data set were previously used to compare the performance of several functional clustering models in [1]. The results in Table 5 of [1] show that the FunFEM models, compared to other state of the art methodologies, achieved the best performances in terms of accuracy. Hence, here, we limit the comparison to the results obtained with the PFC-L_1 and the FunFEM models. Although FunFEM models relay on a mixture of Gaussian distributions describing the likelihood of the data similarly to our proposal, they differ on facing the intrinsic high dimension of the problem by estimating a latent discriminant subspace in parallel with the steps of an EM algorithm.

For all the data, we reconstruct the functional form from the sampled curves choosing arbitrarily 20 cubic spline basis of functions. We tested the PFC-L_1 models considering five different values for the number of clusters, $K = \{2, 3, 4, 5, 6\}$, and six values for $\lambda_1 = \{0.5, 1, 5, 10, 15, 20\}$.

Considering that the GLASSO penalty parameter λ depends linearly from λ_2, the choice of λ_2 has to provide suitable values for λ. A practical approach is to choose values avoiding convergence problems with GLASSO. Here λ_2 was set to $\{5, 7.5, 10, 12, 15, 20\}$ for the ECG data. Both PFC-L_1 and FunFEM algorithms were initialized using a K-means procedure.

The clustering accuracies, computed with respect to the known labels, are 69% for FunFEM DFM$_{[\alpha_{kj}\beta_k]}$ (choosing among 12 different model parameterizations with BIC index), and 75% for PFC-L$_1$ [$\lambda_1 = 0.5$, $\lambda_2 = 5$] (values of tuning parameters chose by ASW index) . Thus PFC-L_1 achieves good performance, with an increase in the accuracy about 9%.

4 Discussion

In this paper we tried to investigate the potential of shrinkage methods for clustering functional data. Our numerical examples show the advantages of performing clustering with features selection, such as uncover interesting structures underlying the data while preserving good clustering accuracy. To the best of our knowledge, this is the first proposal that considers a penalty for both means and covariances of mixture components in functional model-based clustering. In the model selection section we defined an heuristic criterion to choose among different model parameterizations based on average silhouette index. It may be interesting to evaluate different distances (i.e. not euclidean) to compute this index in future research. Moreover, we will consider more complex simulation designs to investigate the robustness of the proposal and extend the comparison with the state of the art methodologies on more benchmark datasets.

References

1. Bouveyron, C., Come, E., Jacques, E.: The discriminative functional mixture model for a comparative analysis of bike sharing systems. Ann. Appl. Stat. **9**, 1726–1760 (2015)
2. Chamroukhi, F., Nguyen, H.: Model-based clustering and classification of functional data. Wiley Interdiscip. Rev.: Data Min. and Knowl. Discov. **9**, e1298, 1–36 (2019)
3. Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., Hexagon-ML: The UCR Time Series Classification Archive (October 2018) `https://www.cs.ucr.edu/$\sim$eamonn/time\_series\_data\_2018/`
4. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.). **39**, 1–38 (1977)
5. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostat. **9**, 432–41 (2008)
6. Friedman, J., Hastie, T., Tibshirani, R.: glasso: Graphical Lasso: Estimation of Gaussian Graphical Models, R package version 1.11 (2019). `https://CRAN.R-project.org/package=glasso`
7. Jacques, J., Preda, C.: Functional data clustering: A survey. Adv. Data Anal. Classif. **8**, 231–255 (2013)
8. Pan, W., Shen, X.: Penalized model-based clustering with application to variable selection. J. Mach. Learn. Res. **8**, 1145–1164 (2007)
9. Zhou, H., Pan, W., Shen, X.: Penalized model-based clustering with unconstrained covariance matrices. Electron. J. Stat. **3**, 1473–1496 (2009)

Emotion Classification Based on Single Electrode Brain Data: Applications for Assistive Technology

Duarte Rodrigues, Luis Paulo Reis, and Brígida Mónica Faria

Abstract This research case focused on the development of an emotion classification system aimed to be integrated in projects committed to improve assistive technologies. An experimental protocol was designed to acquire an electroencephalogram (EEG) signal that translated a certain emotional state. To trigger this stimulus, a set of clips were retrieved from an extensive database of pre-labeled videos. Then, the signals were properly processed, in order to extract valuable features and patterns to train the machine and deep learning models.There were suggested 3 hypotheses for classification: recognition of 6 core emotions; distinguishing between 2 different emotions and recognising if the individual was being directly stimulated or merely processing the emotion. Results showed that the first classification task was a challenging one, because of sample size limitation. Nevertheless, good results were achieved in the second and third case scenarios (70% and 97% accuracy scores, respectively) through the application of a recurrent neural network.

Keywords: emotions, brain-computer interface, EEG, supervised learning, machine and deep learning

Duarte Rodrigues
Faculty of Engineering of University of Porto (FEUP), Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal, e-mail: `up201705420@fe.up.pt`

Luis Paulo Reis
Faculty of Engineering of University of Porto (FEUP) and Artificial Intelligence and Computer Science Laboratory (LIACC), Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal,
e-mail: `lpreis@fe.up.pt`

Brígida Mónica Faria (✉)
School of Health, Polytechnic of Porto (ESS-P.PORTO) and Artificial Intelligence and Computer Science (LIACC), Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal,
e-mail: `monica.faria@ess.ipp.pt`

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_35

1 Introduction

Emotions are a part of our lives, as humans we know how to identify the tiniest of microexpressions to unveil what someone is feeling, but also how to use them to express our hearts. From the youngest of ages we see and interact with others and build a database of patterns of, for example, what joy is and how different it is from fear or sadness. Computers, on the other hand, do not have any idea of what an emotion is or how to recognize it. Or do they?

The Artificial Intelligence and Computer Science Laboratory (LIACC) established 2 projects where emotion recognition can be of the utmost importance. The first project, the "IntellWheels 2.0" [1], intends to develop an interactive and intelligent electric wheelchair. This innovative equipment will have a diverse set of features, such as an adaptive control system (through eye gaze, a brain-computer interface, hand orientation, among others) and a personalized multi-modal interface which will allow communication to multiple devices both from the patients and the caregivers. In this case, having information about the mood of the patient is very beneficial, because the interface can give updates to the nursing staff of the emotional condition of the patient. The second project, the "Sleep at the Wheel" [2], focuses on the research of an interface that can sense and predict a driver's drowsiness state, being able to detect if he fell asleep while driving and, consequently, support an alarm system to provide safer routing and driving. Here the state of mind of the driver is a very important aspect, as different emotions, like anger or fear, can provoke dangerous situations or unpredictable scenarios, making the driver less attentive to his surroundings.

In this work, emotions will be sensed through a brain-computer interface (BCI). These are commercial devices that allow to acquire a surface electroencephalogram (EEG). This signal is used to measure the electrical activity of the brain, that fluctuates according to the firing of the neurons in the brain, being quantified in micro-volts. In this research, the BCI used was the "NeuroSky MindWave2" which possesses one single electrode on the forehead, from which it collects a signal from the activity of the frontal lobe. This brain area is responsible for the higher executive functions, including emotional regulation, planning, reasoning and problem solving [3].

The study of emotion recognition started with psychologist Paul Ekman that defined, based on a cross cultural study, six core emotions - Fear, Anger, Happiness, Sadness, Surprise and Disgust [4]. Later, psychologist Robert Plutchik established a model called "Wheel of Emotions", a diagram where every emotion can be derived from the core 6.

It is also important to have a way to measure what someone is feeling or what emotion they are experiencing. An easy way to do this is through the "Discrete Emotion Questionnaire", a psychological validated questionnaire to verify the intensity of a certain emotion. This assessment presents the 6 core emotions to the subjects asking them to rate the intensity they felt, from 1 to 7 [5].

As a first approach in this area, the current work aims to be able to identify the core emotions using EEG signals collected with the BCI.

2 Experimental Methodology

In order to correctly identify the core emotions, the first step is to trigger them in an efficient way for the brain data collected to be as informative as possible.To do so, the emotions were prompted via a set of video clips, that lasted 5-7 seconds. These videos were selected from a certified database, where the videos were labeled according to the intensity and kind of emotion it caused in the subjects [6]. For each of the 6 core emotions, the 4 videos classified with the biggest intensity were selected to be presented to the participants of this research work.

For each of the 24 video clips (4 videos per each of the 6 emotions), 3 EEG samples are collected. The first is before the display of the video, where a fixation cross is presented, in order to collect the idle/blank state of the user, where he is asked to relax. The second sample is the EEG during the video (active visual stimulus); and the third sample is after the video finishes where the volunteer is processing the emotion triggered (higher level thinking), while getting back to the initial relaxed state, where the fixation cross is presented again. To confirm that the volunteers experience the same emotion defined in the pre-determined label, they are a prompted to answer the “Discrete Emotion Questionnaire”, after the 3 EEG samples are collected.

Regarding the physiological signal processing, this step is important because the raw EEG signal that comes directly from the BCI has a low signal-to-noise ratio, as well as many surrounding artifacts that contaminate the readings, especially eye blinks and facial movements triggered by the various emotions. These interfering signals caused by the latter, denominated electromyograms (EMG), are characterized by high frequencies (50-150 Hz) that make the underlying signal very noisy. Every time a person blinks, the EEG signal shows a very high peak with a very low frequency (<1Hz). To remove these muscle artifacts, a 5th order utterworth bandpass filter (this type of filter was chosen because it has the flattest frequency response, which leads to less signal distortion) with cut-off frequencies in 1 Hz and 50 Hz [7].The attenuation of very low frequencies is important to remove the eye blinks artifacts. Considering the top cut-off frequency, it is very convenient to use 50 Hz since it mitigates the effects of the power line noise and the EMG artifacts. Like this, no important brain data is lost. At this step, the EEG was segmented in the brain waves of interest, i.e., the alpha and beta brain waves. The best way to perform this is to apply bandpass filters (same filter type as before) in the corresponding bandwidths, 8-13Hz and 13-32 Hz, to have alpha and beta bands, respectively.

The EEG signals, at this stage possess the "emotional data" exposed allowing to extract the features. To do so, multiple mathematical equations were applied to obtain relevant information from the signals. Feature extraction methods depend on the domain, as will be seen ahead [8]. Most strategies to extract features from the EEG are formulas applied in the time domain, such as, the common statistical equations, the Hjorth statistical parameters, the mean and zero crossings (number of times the signal crosses these 2 thresholds) [8]. Besides these, there were applied more advanced feature extraction methods, based on fractal dimensions and entropy analysis (methods to assess the complexity, or irregularity, of a time-series) [9].

Regarding frequency domain approaches, these features can only be calculated in the filtered EEG and not in the brain waves, as their spectrum is very narrow. In terms of the pure frequency band, the only feature computed was the Power Spectral Density (PSD), based on the Welch method. These domains can be combined creating the time-frequency domain, leading to more sophisticated methods, like the Hilbert – Huang Transform, where the original signal is decomposed in intrinsic mode functions (IMF) [10].

The resulting number of features is too high to compute machine learning models, because the correlation between most of the features is very low, which means that between different classes the information is virtually the same. This would introduce uncertainty in the weights for each class in the models, thus the number of features needs to be reduced. To do this the "Min Redundancy Max Relevance" (MRMR) method was applied, with the objective of finding the optimal number of features to have a higher inter-class variability, in order to find distinct patterns between emotions [11]. The features were used raw, normalized or standardized to train the models.

In this study, all the models implemented are based on supervised learning and fully depend on the data that is inputted. Concerning emotion classification there is not a specific machine learning approach that is optimal, thus 9 different types of models were implemented to verify which has the best performance. These models are designed to be able to adapt to various kinds of input data, through the definition of hyper-parameters. Hence, to tune them to the best possible configuration, it was performed a GridSearchCV. This method exhaustively searches over a given list of possible parameters applying cross validation between them. In the end, the model with the best performance is chosen to be trained with the resulting feature matrix.

A deep learning model was also implemented, based on recurrent neural network (RNN), a very common architecture in classification problems using EEG. A particularity of this network is that it has a GRU, i.e., a layer that helps to mitigate the problem of vanishing gradients (common issue on artificial neural networks), giving long term memory to the model [12].

3 Evaluation and Discussion of Results

In this experiment, 12 subjects volunteered to participate. Each EEG recording is labeled according to the emotion registered in the original database, as well as if it was before video, during or after the video. The answers of the “Discrete Emotion Questionnaire” were used to validate if the emotion triggered by the video was as expected and, if so, the data was used. With this dataset structure, 3 hypotheses were tested and their results are discussed ahead.

An important aspect to have in consideration is that the EEG collected while the subject is relaxing, i.e., while the fixation cross presented before the video, does not have relevant cognitive information regarding emotions. Therefore, these segments were not considered to train any of the models.

3.1 Core Emotions Classification

This first hypothesis describes the main goal of the project where a model was developed to classify 6 emotions.

First, the feature extraction was computed. At this step, the optimal number of features to get selected was tested, iterating from 5 to 50, 5 at a time. The best number found was 30, which gave the best accuracies, with a balanced computation time and power. This value was chosen for the 3 feature matrixes (raw, normalized and standardized). The dataset was then divided into training and testing with an 80% ratio and fully independent of one another. Each model was then trained and assessed, by computing the accuracy in the test dataset. Table 1 presents the results for each model.

Table 1 Results of the 6 Core Emotions Classification.

Classification Models	Raw Features	Normalized Features	Standardized features
	Accuracy (%)		
Gaussian Naïve Bayes Classifier	12.07	12.93	10.34
Support Vector Classifier	12.07	12.93	16.38
Decision Tree Classifier	18.96	18.10	18.10
Random Forest Classifier	24.13	18.10	20.69
K Nearest Neighbors	21.55	18.96	16.38
Logistic Regression	**25.00**	14.66	18.10
Linear Discriminant Analysis	24.13	14.65	18.96
Linear Support Vector Classifier	18.10	13.79	19.82
Multi-Layer Perceptron	20.69	13.79	12.93
Recurrent Neutral Network	13.79	20.69	23.27

When comparing the various models, the average accuracy is around 16-18%, logically due to the number of classes in the problem (100%/6 = 16,6%). Despite this, the best result reached was 25% accuracy, with the features in their raw state, since the magnitude information was not lost, so patterns in different emotions could be more easily identified due to the high discrepancy in the values. These results are not discouraging since the main objective of the study is very ambitious, as we are trying to create a model to define universally what an emotion is. There is no work more subjective or abstract, and the only way to achieve this universal standardization would be with a sample population as wide and diverse as possible with different beliefs, nationalities, age groups, etc. Although this is an initial study, it shows that it is possible to register and identify differences in the electrical changes of the prefrontal cortex and, with that information, categorize what someone is feeling.

3.2 One vs One – Dual Emotion Classification

As the results in the previous hypothesis could not precisely identify an emotion when compared to the other 5, the problem was narrowed down and a new hypothesis was tested, to continue the proposed research. In this experiment, the model was trained to discern between only 2 emotions, decided *a priori*. For demonstration purposes, a concrete example can be seen in Table 2 where it compares "fear" vs "surprise".

Table 2 Results of "Fear vs Surprise" Classification.

Classification Models	Raw Features	Normalized Features	Standardized features
	Accuracy (%)		
Gaussian Naïve Bayes Classifier	48.27	55.17	53.44
Support Vector Classifier	51.72	51.72	53.44
Decision Tree Classifier	56.89	50.00	44.83
Random Forest Classifier	48.27	50.00	60.34
K Nearest Neighbors	46.55	44.82	50.00
Logistic Regression	50.00	53.45	53.45
Linear Discriminant Analysis	50.00	48.28	53.44
Linear Support Vector Classifier	50.00	51.72	55.17
Multi-Layer Perceptron	50.00	50.00	58.62
Recurrent Neutral Network	**69.23**	51.23	56.21

In this case, most of the machine learning algorithms have accuracies in the order of the 50-53%. This results are not ideal, as they are no better than a random choice between the two classes, however this can be justified by the low population sample, which is not high enough to bring to the surface concrete patterns on the features. Regarding the deep learning approach, the RNN has an advantage in this case, giving a final accuracy of 69%. This result shows that this model is reliable, and in the majority of the cases the 2 emotions can be distinguished. In this particular case, the facial expressions and their muscle activity, can induce big artifacts in the EEG. Someone who feels surprised has the tendency to raise their eyebrows and open the mouth. These movements can lead to a difference in the EEG and, consequently, in the patterns of the features, making the distinction between surprise and fear more noticeable. The same thinking applies to other emotions that trigger facial movement, like laugh, frowning, among others.

3.3 Stimulus vs No Stimulus Classification

Besides the good results presented in the last premise, one last hypothesis was assessed, regarding the difference between experiencing the emotion while watching the video (direct stimulus), and after, when the fixation cross is presented, while the volunteer is simply thinking and cognitively processing the emotion.

Table 3 summarizes the results of the various models.

Table 3 Results of Stimulus vs No Stimulus classification.

Classification Models	Raw Features	Normalized Features	Standardized features
	Accuracy (%)		
Gaussian Naïve Bayes Classifier	61.20	58.62	85.34
Support Vector Classifier	58.62	58.62	91.37
Decision Tree Classifier	39.65	58.62	89.65
Random Forest Classifier	39.65	58.62	91.37
K Nearest Neighbors	37.93	58.62	89.65
Logistic Regression	34.48	58.62	87.06
Linear Discriminant Analysis	29.31	37.06	80.17
Linear Support Vector Classifier	34.48	58.62	87.06
Multi-Layer Perceptron	31.03	58.62	88.79
Recurrent Neutral Network	**96.55**	61.20	88.79

As it can be seen, for this experiment, most models did fairly well using the standardized feature, being all accuracies higher than 80%. However, when testing the deep learning approach, this architecture revealed to fit almost perfectly to the testing data, with an accuracy higher than 96%. This hypothesis is the proof of concept that the characteristics of the signal collected during the stimulus itself are very different from the ones from a signal obtained when the person is simply thinking and cognitively processing the emotion (this change would be obvious if the EEG was collected from the occipital lobe, which is responsible for the visual perception, but is remarkable when spotted in the prefrontal cortex).

4 Conclusions

In conclusion, as a first approach, the results achieved are very satisfactory and reveal a high potential to be greatly efficient in the proposed applications both in "IntellWheels2.0" and "Sleep at the Wheel projects". Nevertheless by collecting more data the models will get more generalized resulting in more realistic patterns and, consequently, increasing the prediction's accuracies.

Comparing to the literature, using simple visual stimuli to distinguish six emotions, in a relaxed state, is a novel tactic. Most studies, complement the stimulus with forced facial expression, introducing different characteristics to the signal, leading to better results. Other studies use BCIs with more electrodes (channels), covering a wider cranial surface and, consequently, getting more EEG and information, which leads to more robust results.

As future work, the preprocessing of the data could be polished, improving the removal of artifacts and enhancing the underlying information of the EEG's. To obtain better results, it could also be used a transfer learning approach, by pre-training the models with another emotion related EEG databases.

Acknowledgements This work was financially supported by Base Funding - UIDB/00027/2020 of the Artificial Intelligence and Computer Science Laboratory – LIACC - funded by national funds through the FCT/MCTES (PIDDAC), Sono ao Volante 2.0 - Information system for predicting sleeping while driving and detecting disorders or chronic sleep deprivation (NORTE-01-0247-FEDER-039720), and Intellwheels 2.0 - IntellWheels2.0 – Intelligent Wheelchair with Flexible Multimodal Interface and Realistic Simulator (POCI-01-0247-FEDER-39898), supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement.

References

1. IntellWheels2.0 – Intelligent Wheelchair with Flexible Multimodal Interface and Realistic Simulator. Optimizer, Lda, FEUP, UA, Rehapoint, GroundControl. Available at `http://www.intellwheels.com/en/client/skins/geral.php?id=25` Cited 24 May 2021
2. Sono ao Volante 2.0 - Information system for predicting sleeping while driving and detecting disorders or chronic sleep deprivation. Optimizer, Lda, FEUP, IS, IPCA. Available at `http://sonoaovolante.com/en/client/skins/geral.php?id=25` Cited 24 May 2021
3. Lobes of the Brain. UQ-Queensland Brain Institute (2018). Available at `https://qbi.uq.edu.au/brain/brain-anatomy/lobes-brain.Cited26May2021`
4. Eckman, P.: Facial Expressions of Emotion: New Findings, New Questions In: Psychological Science, 34-38. Sage Journals (1992)
5. Harmon-Jones, C., Bastian, B., Harmon-Jones, E.: The Discrete Emotions Questionnaire: A New Tool for Measuring State Self-Reported Emotions. In: PLoS One **11**(8), e0159915 (2016) doi: 10.1371/journal.pone.0159915.
6. Cowen, A., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients In: Proceedings of the National Academy of Sciences of the United States of America **14**(38), E7900-E7909 (2017) doi: 10.1073/pnas.1702247114.
7. López-Gil, J.-M., Virgili-Gomá, J., Gil, R., Guilera, T., Batalla, I., Soler-González, J., García, R.: Method for Improving EEG Based Emotion Recognition by Combining It with Synchronized Biometric and Eye Tracking Technologies in a Non-invasive and Low Cost Way. In: Frontiers in Computational Neuroscience **10**, 85 (2016) doi: 10.3389/fncom.2016.00119
8. Jenke, R., Peer, A., Buss, M.: Feature Extraction and Selection for Emotion Recognition from EEG. In: IEEE Transactions on Affective Computing, **5**(3), 327-339, (2014) doi: 10.1109/TAFFC.2014.2339834
9. Richman, J. S., Moorman, J. R.: Physiological time-series analysis approximate entropy and sample entropy. In: American Journal of Physiology-Heart and Circulatory Physiology (2000) doi: 10.1152/ajpheart.2000.278.6.H2039
10. Junsheng, C., Dejie, Y., Yu, Y.: Research on the intrinsic mode function (IMF) criterion in EMD method In: Mechanical Systems and Signal Processing, **20**(4), 817-824. (2006) doi: 10.1016/j.ymssp.2005.09.011
11. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. In: Bioinformatics Computation Biol., **3**(2) 185–205, (2003) doi: 10.1142/S0219720005001004
12. Zain, M. A.: Predicting Emotions Using EEG Data with Recurrent Neural Networks. Geek Culture (2021) Available at `https://medium.com/geekculture/predicting-emotions-using-eeg-data-with-recurrent-neural-networks-8acf384896f5` Cited 19 May 2021

The Death Process in Italy Before and During the Covid-19 Pandemic: a Functional Compositional Approach

Riccardo Scimone, Alessandra Menafoglio, Laura M. Sangalli, and Piercesare Secchi

Abstract In this talk, based on [1], we propose a spatio-temporal analysis of daily death counts in Italy, collected by ISTAT (Italian Statistical Institute), in Italian provinces and municipalities. While in [1] the focus was on the elderly class (70+ years old), we here focus on the middle class (50-69 years old), carrying out analogous analyses and comparative observations. We analyse historical provincial data starting from 2011 up to 2020, year in which the impacts of the Covid-19 pandemic on the overall death process are assessed and analysed. The cornerstone of our analysis pipeline is a novel functional compositional representation for the death counts during each calendar year: specifically, we work with mortality densities over the calendar year, embedding them in the Bayes space B^2 of probability density functions. This Hilbert space embedding allows for the formulation of functional linear models, which are used to split each yearly realization of the mortality density process in a predictable and an unpredictable component, based on the mortality in previous years. The unpredictable components of the mortality density are then spatially analysed in the framework of Object Oriented Spatial Statistics. Via spatial downscaling of the results obtained at the provincial level, we obtain smooth predictions at the fine scale of Italian municipalities; this also enable us to perform

Riccardo Scimone (✉)
MOX, Dipartimento di Matematica, Politecnico di Milano and Center for Analysis, Decision and Society, Human Technopole, Milano, Italy, e-mail: riccardo.scimone@polimi.it

Alessandra Menafoglio
MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy,
e-mail: alessandra.menafoglio@polimi.it

Laura M. Sangalli
MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy,
e-mail: laura.sangalli@polimi.it

Piercesare Secchi
MOX, Dipartimento di Matematica, Politecnico di Milano and Center for Analysis, Decision and Society, Human Technopole, Milano, Italy, e-mail: piercesare.secchi@polimi.it

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_36

anomaly detection, identifying municipalities which behave unusually with respect to the surroundings.

Keywords: COVID-19, O2S2, functional data analysis, spatial downscaling

1 Introduction and Data Presentation

At the dawn of the third year of global pandemic, we can affirm that no aspect of people's everyday life has been left untouched by the consequences of Covid-19. The virus, in addition to exacting an heavy death toll, has caused great upheavals in global economy, education systems, technological development and in countless other aspects of human life. Given this global reaching, we deem appropriate to analyse death counts from all causes, and not just those directly attributed to Covid-19, as a proxy of how Italian administrative units, be they municipalities or provinces, have been affected by the pandemic. This choice is driven by the following considerations:

- Death counts from all causes are, on many levels, high quality data: they have a very fine spatial and temporal granularity, being collected daily in each Italian municipality, they are finely stratified in many age classes, and they are not affected by errors due to incorrect attribution of the cause of death, as may happen, for example, in deciding whether or not a given death is due to Covid-19;
- They incorporate any possible shock, be it direct or indirect, which the natural death process underwent: less deaths from road accidents due to restrictive policies, more deaths from other pathologies which are left untreated because of the unnatural stress on the welfare systems, and so on;
- They are made freely available by ISTAT[1], with a substantial amounts of historical data; in particular, in the following analysis we consider data starting from the beginning of 2011 up to the end of 2020.

The purpose of the analysis of such data is twofold: (1) to study the correlation structure of the death process in Italy before and during the pandemic, assessing possible perturbations caused by its outbreak, and (2) to assess local anomalies at the municipality level (i.e., identifying municipalities which behave unusually with respect to the surrounding). This talk will entirely be devoted to presenting data and results concerning people aged between 50 and 69 years. The elderly class was the focus of [1], while analyses focusing on younger age classes can be freely examined at `https://github.com/RiccardoScimone/Mortality-densities-italy-analysis.git`.

Daily death counts for the 107 Italian provinces, in the time interval spanning from 2017 to 2020, are shown in Fig. 1: for each province, we draw death counts along the year in light blue. The black solid line is the weighted mean number of deaths, where each province has a weight proportional to its population. We also

[1] `https://www.istat.it/it/archivio/240401`

highlight four provinces with colours: Rome, Milan, Naples, and Bergamo. By a visual inspection, it is easy to see that, during the years 2017, 2018 and 2019, the mortality in this age class has an almost uniform behaviour, with only a very slight increase in deaths during winter, for some Provinces. Conversely, 2020 presents an abnormal behaviour in many provinces, due to the pandemic outbreak: look for example at the double peak for Milan, hit by both pandemic waves, or the single, dramatically sharp peak of Bergamo, which reached, during the first wave, higher death counts than the ones associated to provinces which are several times bigger, as Rome or Naples. By comparison with the plots in [1], on can see how all these peaks are less sharper with respect to the elderly class: this is perfectly reasonable, since people aged more than 70 years are much more susceptible to death by Covid-19.

Daily death counts, Italian provinces, 50-69 years

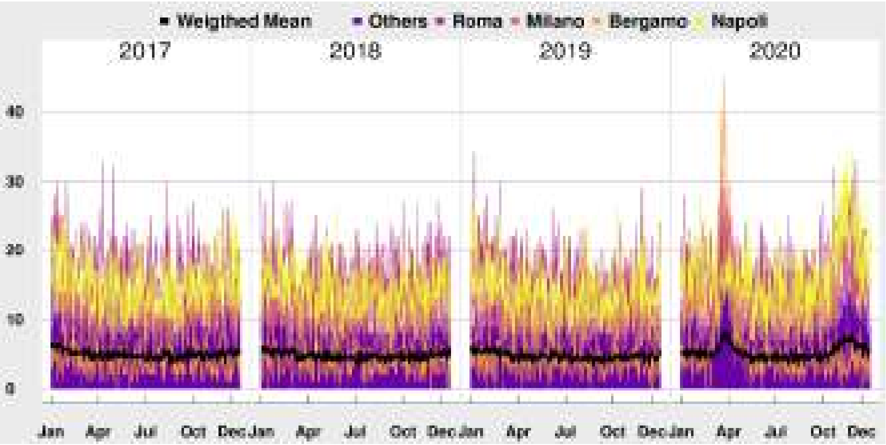

Fig. 1 Daily death counts during the last four years, for the Italian provinces. The plots refer to people aged between 50 and 69 years. For each province, death counts along the year are plotted in light blue: curves are overlaid one on top of the other to visualize their variability. The black solid line is the weighted mean number of deaths, where each province has a weight proportional to its population, while some selected provinces are highlighted in colour.

To set some notation, we denote the available death counts data as d_{iyt}, where i is a geographical index, identifying provinces or municipalities, y is the year and t is the day within year y. Moreover, we denote by T_{iy} the absolutely continuous random variable *time of death along the calendar year*, that models the instant of death of a person living in area i and passing away during year y. We hence consider the empirical discrete probability density of this random variable,

$$p_{iyt} = \frac{d_{iyt}}{\sum_t d_{iyt}} \qquad \text{for } t = 1, \ldots, 365$$

for each area i and year y. The family $\{p_{iy}\}_{iy}$ is the main focus of our analysis: we show these discrete densities in Fig. 2, with the same color choices of Fig. 1. It is clear that using densities provides a natural alignment of areas whose population differs significantly, providing complementary insights with respect to the absolute number of death counts: greater emphasis is given on the temporal structure of the phenomenon. For example, the astonishing behaviour of the province of Bergamo during the first pandemic wave in 2020, is now much more visible.

Empirical densities of daily mortality, provinces, 50-69 years

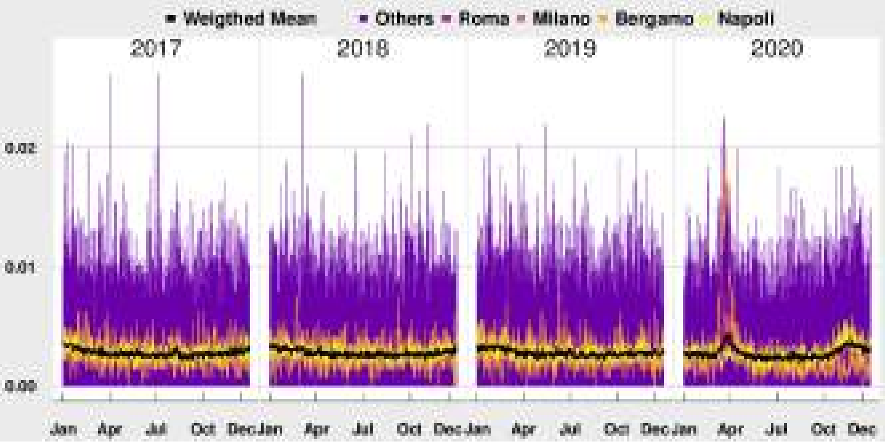

Fig. 2 Empirical densities of daily mortality, for people aged between 50 and 69 years, at the provincial scale. For each province, the empirical density of the daily mortality is plotted in light blue: densities are overlaid one on top of the other to visualize their variability. The black solid line is the weighted mean density, where the weight for each province has been set to be proportional to its population; some selected provinces are highlighted in colour.

In this talk, we will show results obtained by embedding a smoothed version of the $\{p_{iy}\}_{iy}$, i.e., an estimate $\{f_{iy}\}_{iy}$ of the continuous density functions of the $\{T_{iy}\}_{iy}$, in the Hilbert space $B^2(\Theta)$, called *Bayes space* [2, 4, 3], where Θ denotes the calendar year. This is the set (of equivalence classes) of functions

$$B^2(\Theta) = \{f : \Theta \to \mathbb{R}^+ \; s.t. \; f > 0, \; log(f) \in L^2(\Theta)\}$$

where the equivalence relation in $B^2(\Theta)$ is defined among *proportional* functions, i.e., $f =_{B^2} g$ if $f = \alpha g$ for a constant $\alpha > 0$. In [1], we also propose a preliminary exploration of the $\{p_{iy}\}_{iy}$ based on the *Wasserstein space* embedding, a very regular metric space of probability measures with a straightforward physical interpretation [5]. For the sake of brevity, we here focus on the analysis in $B^2(\Theta)$, which constitutes our main contribution.

$B^2(\Theta)$ is equipped with an Hilbert geometry, constituted by appropriate operations of sum, multiplication by a scalar, and inner product, which make it the infinite-dimensional counterpart of the Aitchison simplex used in standard compositional analysis [6, 7]: for this reason this space is considered the most suited Hilbert embedding for positive continuous density functions. The smoothed densities $\{f_{iy}\}_{iy}$

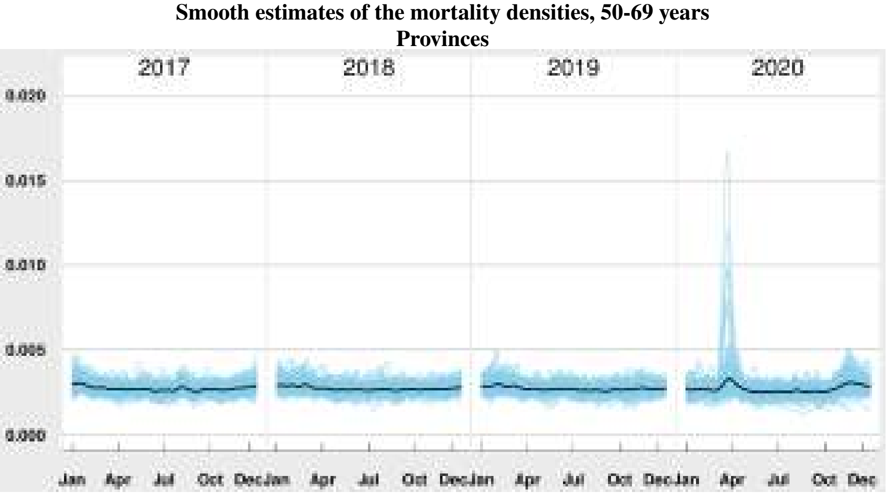

Fig. 3 Smooth estimates of the mortality densities over the 107 Italian provinces. The usual pattern of mortality is visible till 2019, while the functional process is completely different in 2020, with the two pandemic waves clearly captured by the estimated densities. The black thick lines represent the mean density, computed in B^2, with weights proportional to the population in each area.

are shown in Fig. 3: they are obtained by smoothing the $\{p_{iy}\}_{iy}$ via *compositional splines* [8, 9]. It is easy to see, by comparison with Fig. 2, how smoothing filters out a good amount of noise, much more than the case of the elderly class: this is fairly reasonable, since the death process is usually more noisy for younger age classes. From now on, the $\{f_{iy}\}_{iy}$ are analysed as a spatio-temporal functional random sample taking values in $B^2(\Theta)$. We briefly anticipate the results of such analysis:

1. The $\{f_{iy}\}_{iy}$ are decomposed, by means of a linear model formulated in $B^2(\Theta)$ [10], in a *predictable* and an *unpredictable* part, on the basis of mortality during previous years;
2. The unpredictable part is then analysed spatially in order to infer the main spatial correlation characteristics of the process; in particular, the impacts of the pandemic are investigated via functional variography [13, 14, 11, 12] and Principal Component Analysis in the B^2 space (SFPCA, [16]);
3. The results obtained at the provincial level are reduced to the municipality scale by *spatial downscaling* [15] techniques, obtaining smooth density estimates for each municipality. This provides continuous density at the municipality level, without directly smoothing the corresponding daily death process, which is quite irregular due to the reduced population of many municipalities. The spatial downscaling estimates, that are exclusively based on provincial data, are then compared with the actual measurements on municipalities, allowing for the identification of local anomalies.

Points 1 and 2 above are detailed in Section 2, while point 3 will be discussed during the talk. The reader is referred to [1] for full details on the analysis pipeline.

2 Some Results

The first step of the analysis of the random sample $\{f_{iy}\}_{iy}$, where i is indexing the 107 Italian provinces, is the formulation of a family of function-on-function linear models in $B^2(\Theta)$, extending classical models formulated in the L^2 case [17], namely

$$f_{iy}(t) = \beta_{0y}(t) + \langle\beta_y(\cdot,t), \overline{f}_{iy}\rangle_{B^2} + \epsilon_{iy}(t), \quad i = 1, \dots 107, \quad t \in \Theta, \tag{1}$$

where $\overline{f}_{iy} = \frac{1}{4}\sum_{r=y-4}^{y-1} f_{ir}$ is the B^2 mean of the observed densities in the four years preceding year y, functional parameters $\beta_{0y}(t), \beta_y(s,t)$ are defined in the B^2 sense, as well as the residual terms $\epsilon_{iy}(t)$ and all operations of summation and multiplication by a scalar. Model (1) is trying to explain the realization of the mortality density f_{iy} for a year y in a province i as a linear function of what happened in the same province during the preceding years. It is thus interesting to look at the following functional prediction errors:

$$\delta_{iy} = f_{iy} - \hat{f}_{iy} \tag{2}$$

where

$$\hat{f}_{iy}(t) := \hat{\beta}_{0y-1}(t) + \langle\hat{\beta}_{y-1}(\cdot,t), \overline{f}_{iy}\rangle_{B^2}. \tag{3}$$

The δ_{iy} are not the estimate $\hat{\epsilon}_{iy}$ of the residual of model (1): they rather represent the error committed in forecasting f_{iy} using the model fitted at year $y-1$. Thus, we can look at the densities δ_{iy} as the *unpredictable component* of f_{iy}, i.e., as a proxy of what happened at year y which could not be predicted by information available at the previous years, and analyze them under the spatial viewpoint. For example, we can look at the spatial heatmaps of the B^2 norms of the δ_{iy}, which are shown in Fig 4. It is clear, by looking at the magnitude of the error norms, that what happened during 2020 was to a large extent unpredictable, since almost all Italian

Prediction error norms and B^2 functional clustering, provinces, 50-69 years

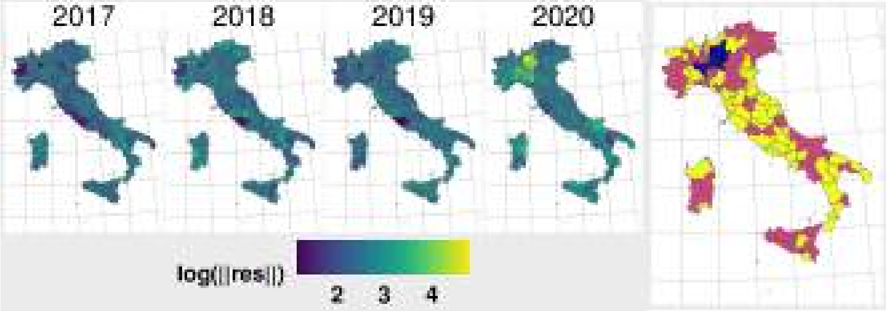

Fig. 4 First four panels, from the left: heatmaps of the B^2 norm of the prediction errors δ_{iy}, in logarithmic scale, for the elderly class. In 2020 the pandemic diffusion is clearly visible in northern Italy, while the prediction errors are generally higher on all provinces. Last panel: result of a K-mean B^2 functional clustering ($K = 3$) on the δ_{iy}, during 2020.

provinces are characterized by higher errors with respect to previous years. More significantly, in 2020 a clear spatial pattern can be noticed, at least during the first wave in northern Italy: a diffusive process, having at its core the provinces most gravely hit by the first pandemic wave, seems to take place in northern Italy. This pattern is, as reasonable, slightly less evident with respect to the case of the elderly class analysed in [1]. Going in this direction, we also show in Fig 4 the result of a K-means functional clustering, set in the B^2 space, of the δ_{iy} for the year 2020. We clearly identify provinces hit by the first wave (blue cluster), while the other two clusters behave irregularly: this is a neat distinction with people aged more than 70 years, where each cluster clearly identifies different kinds of pandemic behaviour (see [1]). For a more precise investigation of the spatial correlation structure of the process across different years, from the δ_{iy} we compute a *functional trace variogram* for each year: we show them for 2017 up to 2020 in Figure 5. Without entering into the details of the mathematical definition of variograms, we can look at the fitted curves in Figure 5 as follows. Distances are on the x-axis, while on the y-axis we have a function of the spatial correlation of the process: when the curve reaches its horizontal asymptote, it has reached the total variance of the process and we are beyond the maximum correlation length. In this perspective, it is immediate to infer that not only the total variance of the functional process δ_{iy} has sharply increased

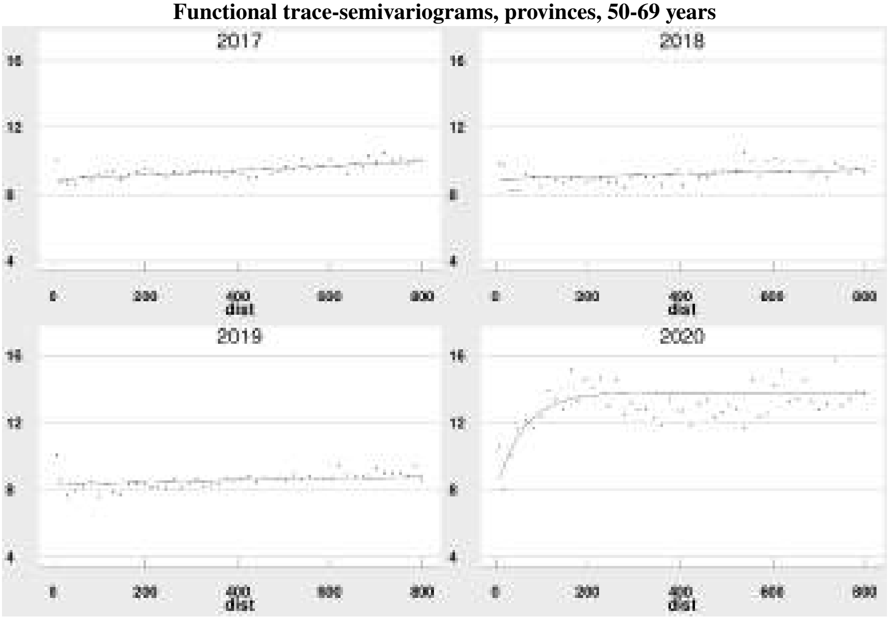

Fig. 5 Empirical trace-semivariograms for the prediction errors δ_{iy}, in people aged between 50 and 69 years. The purple lines are the corresponding fitted exponential models. Distances on the x-axes are expressed in kilometers. The last panel shows the 2020 severe perturbation of the spatial dependence structure of the process generating the prediction errors.

in 2020, but also a significant spatial correlation has manifested, compatibly with the presence of a pandemic. In the main work [1], we further deepen the connection between the pandemic and the upheavals in the spatial structure by means of Principal Component Analysis of the δ_{iy} in the Bayes space (SFPCA, [16]).

References

1. Scimone, R., Menafoglio, A., Sangalli, L. M., Secchi, P.: A look at the spatio-temporal mortality patterns in Italy during the COVID-19 pandemic through the lens of mortality densities. Spatial Stat. (2021) doi:10.1016/j.spasta.2021.100541
2. Egozcue, J., Díaz–Barrero, J., Pawlowsky-Glahn, V.: Hilbert space of probability density functions based on Aitchison geometry. Acta Mathematica Sinica. **22**, 1175-1182 (2006)
3. Pawlowsky-Glahn, V., Egozcue, J., Boogaart, K.: Bayes Hilbert spaces. Aust. New Zeal. J. Stat. **56**, 171-194 (2014)
4. Boogaart, K., Egozcue, J., Pawlowsky-Glahn, V.: Bayes linear spaces. SORT. **34**, 201-222 (2010)
5. Villani, C.: Topics in Optimal Transportation. American Mathematical Society (2003)
6. Aitchison, J.: The statistical analysis of compositional data. J. Roy. Stat. Soc. B Stat. Meth. **44**, 139-177 (1982)
7. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall, London (1986)
8. Machalová, J., Hron, K., Monti, G.: Preprocessing of centred logratio transformed density functions using smoothing splines. J. Appl. Stat. **43** (2015)
9. Machalová, J., Talská, R., Hron, K., Gába, A.: Compositional splines for representation of density functions. Comput. Stat. **36**, 1031-1064 (2021)
10. Talská, R., Menafoglio, A., Machalová, J., Hron, K., Fišerová, E.: Compositional regression with functional response. Comput. Stat. Data Anal. **123**, 66-85 (2018)
11. Menafoglio, A., Petris, G.: Kriging for Hilbert-space valued random fields: The operatorial point of view. J. Multivariate Anal. **146** (2015)
12. Menafoglio, A., Grujic, O., Caers, J.: Universal kriging of functional data: Trace-variography vs cross-variography? Application to gas forecasting in unconventional shales. Spatial Stat. **15**, 39-55 (2016)
13. Nerini, D., Monestiez, P., Manté, C.: Cokriging for spatial functional data. J. Multivariate Anal. **101**, 409-418 (2010)
14. Menafoglio, A., Secchi, P., Dalla Rosa, M.: A universal kriging predictor for spatially dependent functional data of a Hilbert space. Electronic Journal of Statistics **7**, 2209-2240 (2013)
15. Goovaerts, P.: Kriging and semivariogram deconvolution in the presence of irregular geographical units. Mathematical Geosciences. **40**, 101-128 (2008)
16. Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces. Comput. Stat. Data Anal. **94**, 330-350 (2016)
17. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer (2005)

Clustering Validation in the Context of Hierarchical Cluster Analysis: an Empirical Study

Osvaldo Silva, Áurea Sousa, and Helena Bacelar-Nicolau

Abstract The evaluation of clustering structures is a crucial step in cluster analysis. This study presents the main results of the hierarchical cluster analysis of variables concerning a real dataset in the context of Higher Education. The goal of this research is to find a typology of some relevant items taking into account both the homogeneity and the isolation of the clusters.Two similarity measures, namely the standard affinity coefficient and Spearman's correlation coefficient, were used, and combined with three probabilistic (*AVL*, *AVB* and *AV1*) aggregation criteria, from a parametric family in the scope of the *VL* (Validity Link) methodology. The best partitions were selected based on some validation indices, namely the global *STAT* levels statistics and the measures P(I2, Σ) and γ, adapted to the case of similarity coefficients. In order to evaluate the clusters and identify their most representative elements, the Mann and Whitney *U* statistics and the silhouette plot were also used.

Keywords: clustering validation, affinity coefficient, Spearman correlation coefficient, *VL* methodology

Osvaldo Silva (✉)
Universidade dos Açores and CICSNOVA.UAc, Rua da Mãe de Deus, 9500-321, Portugal, e-mail: osvaldo.dl.silva@uac.pt

Áurea Sousa
Universidade dos Açores and CEEAplA, Rua da Mãe de Deus, Portugal,
e-mail: aurea.st.sousa@uac.pt

Helena Bacelar-Nicolau
Universidade de Lisboa (UL) Faculdade de Psicologia and Institute of Environmental Health (ISAMB/FM-UL), Portugal, e-mail: hbacelar@psicologia.ulisboa.pt

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_37

1 Introduction

Cluster analysis or unsupervised classification usually concerns exploratory multivariate data analysis methods and techniques for grouping either a set of data units or an associated set of descriptive variables in such a way that elements in the same group (cluster) are more similar to each other than elements in different clusters [6]. Therefore, it is important to validate the results obtained, bearing in mind that, in an ideal situation, the clusters should be internally homogeneous and externally well separated or isolated. Thus, according to Silva et al. ([15], p. 136), there are some important questions, such as: "i) How to compare partitions obtained using different cluster algorithms? ii) Is it possible to join information from several approaches in the decision-making process of choosing the most representative partition?"

This paper presents the main results of a hierarchical cluster analysis of variables concerning a real dataset in the field of Higher Education, in order to find a typology taking into account relevant validation measures. Two similarity measures (standard affinity coefficient and Spearman's correlation coefficient) were used, and combined with a parametric family aggregation criteria in the scope of the *VL* methodology (e.g., [10, 11, 17]).

With regard to the validation of clustering structures, some validation indices were used for the evaluation of partitions and the clusters that integrate them, which are referred to in Section 2. The main results are presented and discussed in Section 3. Section 4 contains some final remarks.

2 Data and Methods

Data were obtained from a questionnaire administered to three hundred and fifty students who were attending Higher Education in a public university, after their informed consent. The questionnaire contains, among others, eleven questions related to academic life and the respective courses.

Several algorithms of hierarchical cluster analysis of variables were applied on the data matrix. The variables (items) are: T1-Participation, T2-Interest, T3-Expectations, T4-Accomplishment, T5-Job Outlook, T6- Teachers' Professional Competence, T7-Distribution of Curricular Units, T8- Number of weekly hours of lessons, T9-Number of hours of daily study, T10-School Outcomes and T11-Assessment Methods, which were evaluated based on a Likert scale from 1 to 5 (1-Totally disagree, 2- Partially disagree, 3- Neither disagree nor agree, 4- Partially agree, 5- Totally agree).

The Ascendant Hierarchical Cluster Analysis (AHCA) was based on the standard affinity coefficient [1, 17] and Spearman's correlation coefficient. In this paper both measures of comparison were combined with three probabilistic aggregation criteria (*AVL*, *AVB* and *AV1*), issued from the *VL* parametric family. This methodology, in the scope of Cluster Analysis, uses probabilistic comparison functions, between pairs of elements, which correspond to random variables following a unit uniform distribution. Besides, this approach considers probabilistic aggregation criteria, which can be interpreted as distribution functions of statistics of independent random variables, that are i.i.d. uniform on [0, 1] (e.g., [17]).

Let A and B be two clusters with cardinals, respectively, α and β, and let γ_{xy} be a similarity measure between pairs of elements, $x, y \in E$ (set of elements to classify). Concerning the family I of *AVL* methods (e.g., *SL*, *AV1*, *AVB*, and *AVL*), the comparison functions between clusters can be summarized by the following conjoined formula:

$$\Gamma(A, B) = (p_{AB})^{g(\alpha,\beta)} \tag{1}$$

where $\alpha = Card\ A$, $\beta = Card\ B$, $p_{AB} = max[\gamma_{ab} : (a, b) \in (A \times B]$, with $1 \leq g(\alpha, \beta) \leq \alpha\beta$, and γ_{xy}, establishing a bridge between *SL* and *AVL* methods which have a braking effect on the formation of chains. For example, $g(\alpha, \beta) = 1$ for *SL*, $g(\alpha,\beta)$=$(\alpha + \beta)/2$ for *AV1*, $g(\alpha,\beta)$=$\sqrt{\alpha\beta}$ for *AVB*, and $g(\alpha, \beta) = \alpha\beta$ for *AVL* (see [3, 17]).

The application of the two measures of comparison between elements (Spearman correlation coefficient and standard affinity coefficient), combined with the aforementioned aggregation criteria, aims to find a typology of items corresponding to the best partition among the best partitions obtained by the several algorithms, in order to verify if there are any substantial changes in the results. Therefore, some validation indices based on the values of the corresponding proximity matrices were used, namely the global levels statistics (*STAT*) [1, 10, 11] and the indices P(I2mod, Σ) and γ [8], adapted to this type of matrices [16], so that the choice of the best partition is judicious and based on the desirable properties (e.g., isolation and homogeneity of the clusters). Concerning the best partitions, the respective clusters and the identification of their most representative elements were based on appropriate adaptations of the Mann and Whitney *U* statistics [8] and of the silhouette plots [14] to the case of similarity measures.

Each level of a dendrogram corresponds to a stage in the constitution of the partitions hierarchy. Therefore, the study of the most relevant partition(s) is strictly related to the choice of the best cut-off levels (e.g., [6, 5])

According to Bacelar Nicolau [1, 2], the global levels statistics (*STAT*) values must be calculated for each of the $k = 1, nivmax$ levels of the corresponding dendrograms, designating them by $STAT(k)$. At each level k, $STAT(k)$ is the global statistics that measures the total information given by the pre-order associated to the corresponding partition, in relation to the initial pre-order associated with the similarity or dissimilarity measure. A “significant” level is considered to be one that corresponds to a partition for which the global statistics undergoes a significant increase in relation to the information provided by neighbouring levels, that is, a local maximum of the differences $DIF(k) = STAT(k) - STAT(k - 1)$, $k = 1, nivmax$.

2.1 Adaptation of the P (I2, Σ)

To evaluate the partitions, an appropriate adaptation of the index P (I2, Σ) [8] for the case of similarity measures was used, given by the following formula:

$$P(I2mod, \Sigma) = \frac{1}{c}\sum_{r=1}^{c} \frac{\underset{i \in C_r}{\Sigma}\ \underset{j \notin C_r}{\Sigma} s_{ij}}{n_r \times (N - n_r)} \tag{2}$$

where c is the number of clusters of the partition and s_{ij} is the value of the similarity measure between the element i belonging to cluster C_r and the element j belonging to another cluster. This index takes into account the number of clusters and the number of elements in each of the clusters and evaluates the isolation of clusters belonging to a given partition.

2.2 Goodman and Kruskal Index (γ)

The γ index, proposed by Goodman and Kruskal [7], has been widely used in cluster validation [9]. Comparisons are developed between all within-cluster similarities, s_{ij} and all between-cluster similarities s_{kl} [18]. A comparison is judged concordant (respectively discordant) if s_{ij} is strictly greater (respectively, smaller) than s_{kl}. The γ index is defined by:

$$\gamma = (S_+ - S_-)/(S_+ + S_-), \tag{3}$$

where S_+ (or S_-) is the number of concordant (respectively, discordant) comparisons. This index is a global stopping rule and it evaluates the fit of the partition in c clusters based on the homogeneity (high similarity between the elements within the clusters) and the isolation (low similarity of the elements between the clusters) of the clusters. Note that the higher the value of this index, the better is the adjustment of that partition.

The use of *STAT*, γ and P(I2mod, Σ) indices can help identifying the most significant levels of a dendrogram, taking into account both the homogeneity and the isolation of the clusters [15].

2.3 U Statistics (Mann and Whitney)

U statistics [12] are relevant for assessing the suitability of a cluster, combining the concepts of compactness and isolation. Thus, the “best” cluster is the one with the lowest values of global *U*-index, U_G, and local *U*-index, U_L [8]. In the present paper we used an appropriate adaptation of these indices to the case of similarity measures (for details, see [19]). Moreover, the clusters considered “ideal” are those for which U_G and U_L both take the value zero. Mann and Whitney’s *U* statistics are useful in decision making, in situations of uncertainty, both for the evaluation of the clusters and partitions.

2.4 Silhouette Plots

We also used an appropriate adaptation of the silhouette plots [14], which allows the assessment of compactness and relative isolation of clusters. The adaptation of this measure for the case of similarity measures, $Sil(i)$, considers the average of the similarities between an element i belonging to cluster C_r , which contains $n_r (\geq 2)$ elements, and all other elements that do not belong to this cluster (see [19]). The values of this measure $\{Sil(i) : i \in C_r\}$ lie between -1 and $+1$, with "values near +1 indicating that element strongly belongs to the cluster in which it has been placed" ([8], p. 205). In the case of a singleton cluster, $Sil(i)$ assumes the value zero [8] in the corresponding algorithm.

3 Results and Discussion

The best partitions provided by the dendrograms are shown in Table 1.

Table 1 The best partitions concerning the dendrograms.

Coefficient	Method	The best partition	Validation indices
Affinity	*AVL*	(T1, T3, T4, T5 ,T6, T7, T8, T10, T11), (T2, T9)	STAT=5.1301 γ= 0.8589 P(I2mod,Σ)=0.2077
	AV1/AVB	(T1, T3, T4 , T5, T6, T7, T8, T10, T11), (T2), (T9)	STAT=5.3453 γ= 0.8830 P(I2mod,Σ)=0.2049
Spearman	*AVL*	(T3, T4 ,T2 , T9) (T7, T11, T8), (T6, T10), (T1), (T5)	STAT=4.0152 γ= 0.8178 P(I2mod,Σ)=0.3896
	AV1/AVB	(T3, T4 ,T2 , T9, T6) (T7, T11, T8), (T1, T10), (T5)	STAT=4.05751 γ= 0.7317 P(I2mod,Σ)=0.38177

Figure 1 shows the dendrograms obtained, respectively, by the standard affinity coefficient (left side) and Spearman's correlation coefficient (right side), both combined with the *AVL* method.

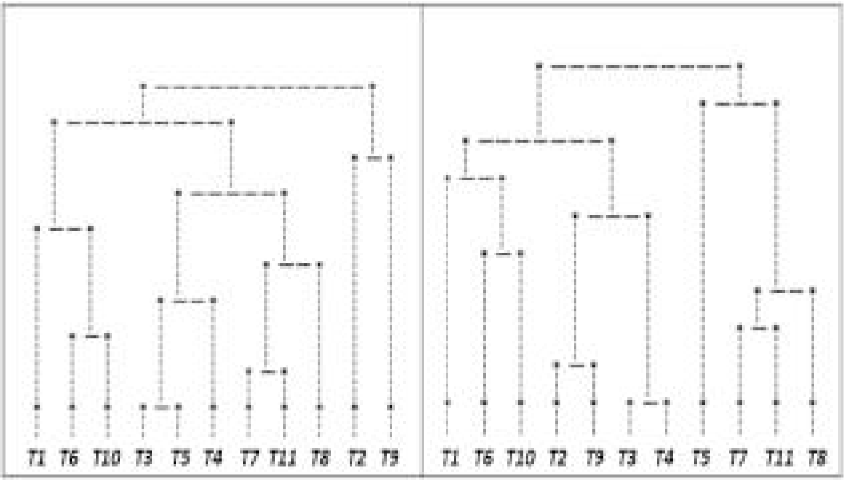

Fig. 1 Dendrograms based on standard affinity coefficient (left side) and Spearman's correlation coefficient (right side) - *AVL*.

The "best" partition obtained using the affinity coefficient and the *AVL* method is the partition into two clusters (level 9 of the aggregation process). The first cluster consists of nine items that highlight the importance of the teachers' professional competence, the structuring/content of the course and the future perspectives in relation to the career opportunities, mostly factors exogenous to the students. The second one is composed by two items (T2 and T9) which emphasize the role of interest in the study of Mathematics.

The algorithms in which the standard affinity coefficient was used are the ones that provided the best partitions and their hierarchies are the ones that remained closest to the initial pre-orders. In fact, in the case of Spearman correlation coefficient the values of *STAT* and γ indices are clearly lower than the previous ones. Moreover, the cluster {T1, T3, T4, T5, T6, T7, T8, T10, T11}, corresponding to the best partition provided by the combination of the standard affinity coefficient with the aggregation criteria *AVL*, *AV1* and *AVB*, presents (U_G =39 and U_L=4, both lower than those obtained for the cluster {T3, T4, T2, T9, T6} (U_G=65 and U_L=26) provided by the Spearman correlation coefficient combined, respectively, with *AV1* and *AVB* methods.

Focusing the attention on the two first partitions of Table 1, the only difference between them is that while the best partition provided by *AV1* and *AVB* methods contains the singletons T2 and T9, the best partition given by *AVL* joins these two singletons in the same cluster. The values of the numerical validation indices shown in Table 1 indicate that the best partition is the one provided by *AV1* and *AVB* methods. This conclusion is reinforced by the observation of the silhouette plot (see Figure 2), which indicates that the cluster joining T2 and T9, given by *AVL* method, includes the elements which have the two lowest values of *Sil* and *Sil* (T2) is negative

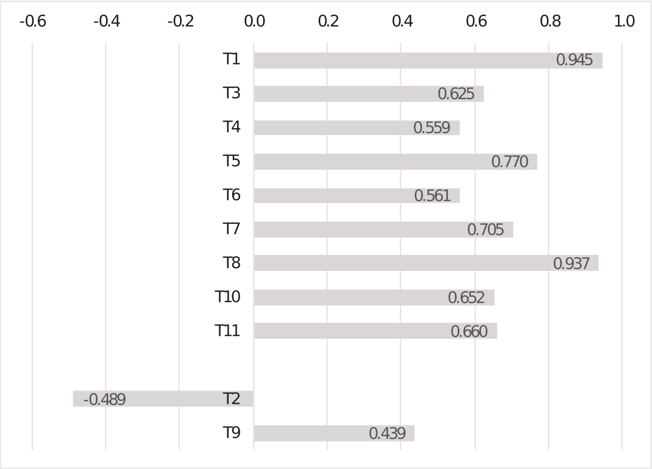

Fig. 2 Silhouette plot - standard affinity coefficient and *AVL* method.

(i.e., T2 does not fit very well in this cluster). Note that the silhouette plot cannot be used for the best partition, since it does not apply for singletons.

4 Final Remarks

This research was useful concerning the identification of relevant partitions of items in the context of Higher Education. In the cases where the affinity and the Spearman correlation coefficients were used, it was concluded that the probabilistic criteria *AV1* and *AVB* showed a higher agreement regarding the hierarchies of partitions obtained than the *AVL* method.

The validation measures *STAT*, γ and P(I2mod, Σ) help us to determine the best cut-off levels of a hierarchy of clusters, taking into account both the homogeneity and the isolation of the clusters. It should also be noted that if there is no absolute consensus between these three measures, the Mann and Whitney *U* statistics and the silhouette plot prove to be very useful, as we have seen with the application of this methodology to evaluate both the clusters and the partitions obtained.

Acknowledgements Funding. This work is financed by national funds through FCT – Foundation for Science and Technology, I.P., within the scope of the project «UIDB/04647/2020» of CICS.NOVA – Centro de Ciências Sociais da Universidade Nova de Lisboa.

References

1. Bacelar-Nicolau, H.: Analyse d'un Algorithme de Classification Automatique. Thèse de 3ème Cycle. ISUP, Paris VI (1972)
2. Bacelar-Nicolau, H.: Contributions to the Study of Comparison Coefficients in Cluster Analysis (in Portuguese). Univ. Lisbon (1980)
3. Bacelar-Nicolau, H.: On the distribution equivalence in cluster analysis. In: P. A., Devijver, & J. Kittler (eds.) Pattern Recognition Theory and Applications, NATO ASI Series, Series F. Computer and Systems Sciences, vol. 30, pp. 73-79. Springer - Verlag, New York (1987)
4. Bacelar-Nicolau, H., Nicolau, F. C., Sousa, Á., Bacelar-Nicolau, L.: Clustering of variables with a three-way approach for health sciences.Testing, Psychometrics, Methodology in Applied Psychology (TPM) (2014) doi: 10.4473/TPM21.4.56
5. Benzécri, J. P.: Analyse Factorielle des Proximités. Publication de l'Institut de Statistique de l' Universite de Paris (ISUP), XIII et XIV (1965)
6. De La Vega, W.: Techniques de la classification automatique utilisant un índice de ressemblance. Revue Française de Sociologie. **VIII**, 506–520 (1967)
7. Goodman, L. A., Kruskal, W. H.: Measures of association for cross-classifications. Journal of the American Statistical Association. **49**, 732–764 (1954)
8. Gordon, A. D.: Classification, 2nd Ed. Chapman & Hall, London (1999)
9. Hubert, L. J.: Some applications of graph theory to clustering. Psychometrika **39**(3), 283–309 (1974)) doi: 10.1007/BF02291704
10. Lerman, I. C.: Classification et Analyse Ordinale des Données. Dunod, Paris (1981)
11. Lerman, I. C.: Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering. Series: Advanced Information and Knowledge Processing. Springer-Verlag, Boston (2016)
12. Mann, H. B., Whitney, D. R.: On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 50–60 (1947)
13. Nicolau, F. C., Bacelar-Nicolau, H.: Some trends in the classification of variables. In: Hayashi et al. (eds.) Data Science, Classification and Related Methods, pp. 89-98. Springer,Tokyo (1998)
14. Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computation and Applied Mathematics. **20**, 53–65 (1987)
15. Silva, O., Bacelar-Nicolau, H.; Nicolau, F.: A global approach to the comparison of clustering results. Biometrical Letters **49**(2), 135–147 (2013) doi: 10.2478/bile-2013-0010
16. Silva, O., Bacelar-Nicolau, H., Nicolau, F. C., Sousa, Á.: Probabilistic approach for comparing partitions. In: Manca, R., McClean, S., Skiadas, C. H.(eds.) New Trends in Stochastic Modeling and Data Analysis, pp. 113-122. ISAST (International Society for the Advancement of Science and Technology), Athens (2015)
17. Sousa, Á., Silva, O., Bacelar-Nicolau, H., Nicolau, F. C.: Distribution of the affinity coefficient between variables based on the Monte Carlo simulation method. Asian Journal of Applied Sciences. **1**(5), 236–245 (2013a)
18. Sousa, Á., Tomás, L., Silva, O., Bacelar-Nicolau, H.: Symbolic data analysis for the assessment of user satisfaction: an application to reading rooms services. European Scientific Journal (ESJ). Special/Edition **3**, 39–48 (2013b)
19. Sousa, Á., Nicolau, F., Bacelar-Nicolau, H., Silva, O.: Cluster analysis using affinity coefficient in order to identify religious beliefs profiles. European Scientific Journal (ESJ). Special/Edition **3**, 252–261 (2014)

An MML Embedded Approach for Estimating the Number of Clusters

Cláudia Silvestre, Margarida G. M. S. Cardoso, and Mário Figueiredo

Abstract Assuming that the data originate from a finite mixture of multinomial distributions, we study the performance of an integrated *Expectation Maximization* (EM) algorithm considering *Minimum Message Length* (MML) criterion to select the number of mixture components. The referred EM-MML approach, rather than selecting one among a set of pre-estimated candidate models (which requires running EM several times), seamlessly integrates estimation and model selection in a single algorithm. Comparisons are provided with EM combined with well-known information criteria – e.g. the Bayesian information Criterion. We resort to synthetic data examples and a real application. The EM-MML computation time is a clear advantage of this method; also, the real data solution it provides is more parsimonious, which reduces the risk of model order overestimation and improves interpretability.

Keywords: finite mixture model, EM algorithm, model selection, minimum message length, categorical data

1 Introduction

Clustering is a technique commonly used in several research and application areas. Most of the clustering techniques are focused on numerical data. In fact, clustering

Cláudia Silvestre (✉)
Escola Superior de Comunicação Social, Campus de Benfica do IPL 1549-014 Lisboa, Portugal,
e-mail: `csilvestre@escs.ipl.pt`

Margarida G. M. S. Cardoso
BRU-UNIDE, ISCTE-IUL, Av. das Forças Armadas, 1649-026 Lisboa, Portugal,
e-mail: `margarida.cardoso@iscte-iul.pt`

Mário Figueiredo
Instituto de Telecomunicações, Portugal, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal,
e-mail: `mario.figueiredo@tecnico.ulisboa.pt`

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_38

methods for categorical data are more challenging [12] and there are fewer techniques available [11].

In order to determine the number of clusters, model-based approaches commonly resort to information-based criteria e.g., the *Bayesian Information Criterion* (BIC) [15] or the *Akaike Information Criterion* (AIC) [1]. These criteria look for a balance between the model's fit to the data (which corresponds to maximizing the likelihood function) and parsimony (using penalties associated with measures of model complexity), thus trying to avoid over-fitting. The use of information criteria follows the estimation of candidate finite mixture models for which a predetermined number of clusters is indicated, generally resorting to an EM (*Expectation Maximization*) algorithm [7]. In this work, we focus on determining the number of clusters while clustering categorical data, using an EM embedded approach to estimate the number of clusters. This approach does not rely on selecting among a set of pre-estimated candidate models, but rather integrates estimation and model selection in a single algorithm. Our new implementation to deal with categorical variables by estimating a finite mixture of multinomials, follows a previous version described in [16]. We capitalized on the work of Figueiredo and Jain [9] for clustering continuous data and extended it for dealing with categorical data. The embedded method is thus based on a *Minimum Message Length* (MML) criterion to select the number of clusters and on an EM algorithm to estimate the model parameters.

2 Clustering with Finite Mixture Models

The literature on finite mixture models and their application is vast, including some books covering theory, geometry, and applications [8, 13, 3]. When applying finite mixture models to social sciences, the analyst is often confronted with the need to uncover sub-populations based on qualitative indicators.

2.1 Definitions and Concepts

Let $\mathbf{Y} = \{\underline{y}_i,\ i = 1, \ldots, n\}$ be a set of n independent and identically distributed (i.i.d.) sample of observations of a random vector, $\underline{Y} = [Y_1, \ldots, Y_L]'$. We assume $\underline{Y}$ follows a mixture of K components densities, $f(\underline{y}|\underline{\theta}_k)$ $(k = 1, \ldots, K)$, with probabilities $\{\alpha_1, \ldots, \alpha_K\}$, where $\underline{\theta}_k$ are the distributional parameters defining the k-th component and $\Theta = \{\underline{\theta}_1, \ldots, \underline{\theta}_K, \alpha_1, \ldots, \alpha_K\}$ the set of all the parameters of the model. The α values, also called *mixing probabilities*, are subject to the usual constraints: $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k \geq 0$, $k = 1, \ldots, K$. The log-likelihood of the observed set of sample observations is

$$\log f(\mathbf{Y}|\Theta) = \log \prod_{i=1}^{n} f(\underline{y}_i|\Theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k f(\underline{y}_i|\underline{\theta}_k). \qquad (1)$$

In clustering, the identity of the component that generated each sample observation is unknown. The observed data $\mathbf{Y}$ is therefore regarded as incomplete, where the missing data is a set of indicator variables $\mathbf{Z} = \{\underline{z}_1, ..., \underline{z}_n\}$, each taking the form $\underline{z}_i = [z_{i1}, ..., z_{iK}]'$, where z_{ik} is a binary indicator: z_{ik} takes the value 1 if the observation $\underline{y}_i$ was generated by the k-th component, and 0 otherwise. It is usually assumed that the $\{\underline{z}_i,\ i = 1, \ldots, n\}$ are i.i.d., following a multinomial distribution of K categories, with probabilities $\{\alpha_1, \ldots, \alpha_K\}$. The log-likelihood of complete data $\{\mathbf{Y}, \mathbf{Z}\}$ is given by

$$\log f(\mathbf{Y}, \mathbf{Z}|\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left[\alpha_k f(\underline{y}_i|\underline{\theta}_k)\right]. \tag{2}$$

2.2 Discrete Finite Mixture Models

Consider that each variable in $\underline{\mathrm{Y}}$, Y_l $(l = 1, \ldots, L)$ can take one of C_l categories. Conditionally on having been generated by the k-th component of the mixture, each Y_l is thus modeled by a multinomial distribution with n_l trials, C_l categories, and non-negative parameters $\underline{\theta}_{kl} = \{\theta_{klc},\ c = 1, \ldots, C_l\}$, with $\sum_{c=1}^{C_l} \theta_{klc} = 1$. For a sample $y_{il}(i = 1, \ldots, n)$ of Y_l, we denote as y_{ilc} the number of outcomes in category c, which is a sufficient statistic; naturally, $\sum_{c=1}^{C_l} y_{ilc} = n_l$. Thus, with $\underline{\theta}_k = \{\underline{\theta}_{k1}, \ldots, \underline{\theta}_{kL}\}$ and $\Theta = \{\underline{\theta}_1, \ldots, \underline{\theta}_K, \alpha_1, \ldots, \alpha_K\}$, the log-likelihood function, for a set of observations corresponding to a discrete finite mixture model (mixture of multinomials). This log-likelihood can be seen as corresponding to a missing-data problem, where the missing data has exactly the same meaning and structure as above. The log-likelihood of the complete data $\{\mathbf{Y}, \mathbf{Z}\}$ is thus given by

$$\log p(\mathbf{Y}, \mathbf{Z}|\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left(\alpha_k \prod_{l=1}^{L} \left[n_l! \prod_{c=1}^{C_l} \frac{(\theta_{klc})^{y_{ilc}}}{y_{ilc}!}\right]\right). \tag{3}$$

To obtain a *maximum-likelihood* (ML) or *maximum a posteriori* (MAP) estimate of the parameters of a multinomial mixture, the well-known EM algorithm is usually the tool of choice [7].

3 Model Selection for Categorical Data

Model selection is an important problem in statistical analysis [6]. In model-based clustering, the term *model selection* usually refers to the problem of determining the number of clusters, although it may also refer to the problem of selecting the structure of the clusters. Model-based clustering provides a statistical framework to solve this problem usually resorting to *information criteria*. Among the best-known information criteria we find BIC and AIC, their modifications - namely the consistent

AIC, (CAIC) and the Modified AIC (MAIC) - and also the Integrated Completed Likelihood (ICL) [14, 4]. They are all easily implemented, the final model being selected according to a compromise between its fit to data and its complexity. In this work, we use the *Minimum Message Length* (MML) criterion to choose the number of components of a mixture of multinomials. MML is based on the information-theoretic view of estimation and model selection, according to which an adequate model is one that allows a short description of the observations. MML-type criteria evaluate statistical models according to their ability to compress a message containing the data, looking for a balance between choosing a simple model and one that describes the data well. According to Shannon's information theory, if Y is some random variable with probability distribution $p(y|\Theta)$, the optimal code-length (in an expected value sense) for an outcome y is $l(y|\Theta) = -\log_2 p(y|\Theta)$, measured in bits (from the base-2 logarithm). If Θ is unknown, the total code-length function has two parts: $l(y, \Theta) = l(y|\Theta) + l(\Theta)$; the first part encodes the outcome y, while the second part encodes the parameters of the model. The first part corresponds the fit of the model to the data (better fit corresponds to higher compression), while the second part represents the complexity of the model. The message length function for a mixture of distributions (as developed in [2]) is:

$$l(y, \Theta) = -\log p(\Theta) - \log p(y|\Theta) + \frac{1}{2}\log|I(\Theta)| + \frac{C}{2}\left(1 - \log(12)\right), \tag{4}$$

where $p(\Theta)$ is a prior distribution over the parameters, $p(y|\Theta)$ is the likelihood function of mixture, $|I(\Theta)| \equiv |-E\left[\frac{\partial^2}{\partial\Theta^2}\log p(Y|\Theta)\right]|$ is the determinant of the expected Fisher information matrix, and C is the the number of parameters of the model that need to be estimated. For example, for the K mixture multinomial distributions presented in (3), $C = (K-1) + K\left(\sum_{l=1}^{L}(C_l - 1)\right)$. The expected Fisher information matrix of a mixture leads to a complex analytical form of MML which cannot be easily computed. To overcome this difficulty, Figueiredo and Jain [9] replace the expected Fisher information matrix by its complete-data counterpart $I_c(\Theta) \equiv -E\left[\frac{\partial^2}{\partial\theta^2}\log p(Y, Z|\Theta)\right]$. Also, they adopt independent Jeffreys' *priors* for the mixture parameters that is proportional to the square root of the determinant of the Fisher information matrix. The resulting message length function is

$$l(y, \Theta) = \frac{M}{2}\sum_{k:\,\alpha_k>0}\log\left(\frac{n\,\alpha_k}{12}\right) + \frac{k_{nz}}{2}\log\frac{n}{12} + \frac{k_{nz}(M+1)}{2} - \log p(y, \Theta) \tag{5}$$

where M is the number of parameters specifying each component (the dimension of each $\underline{\theta}_k$) and k_{nz} the number of components with non zero probability (for more details on the derivation of (5), see [9, 2]).

4 The MML Based EM Algorithm

In order to estimate a mixture of multinomials, we use a variant of the EM algorithm (herein termed EM-MML), which integrates both estimation and model selection, by directly minimizing (5). The algorithm results from observing that (5) contains, in addition to the log-likelihood term, an explicit penalty on the number of components (the two terms proportional to k_{nz}), and a term (the first one) that can be seen as a log-prior on the α_k parameters of Θ, that will directly affect the M-step.

E-step: The E-step of the EM-MML is precisely the same as in the case of ML or MAP estimation, since the generative model for the data is the same. Since we are dealing with a multinomial mixture, we simply have to plug the corresponding multinomial probability function yielding

$$\bar{z}_{ik}^{(t)} = \frac{\alpha_k \prod_{l=1}^{L} \left[n_l! \prod_{c=1}^{C_l} \frac{(\widehat{\theta}_{klc}^{(t)})^{y_{ilc}}}{y_{ilc}!} \right]}{\sum_{j=1}^{K} \alpha_j \prod_{l=1}^{L} \left[n_l! \prod_{c=1}^{C_l} \frac{(\widehat{\theta}_{jlc}^{(t)})^{y_{ilc}}}{y_{ilc}!} \right]}, \tag{6}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

M-step: For the M-step, noticing that the first term in (5) can be seen as the negative log-prior $-\log p(\alpha_k) = \frac{C-K+1}{2K} \log \alpha_k$ (plus a constant), and enforcing the conditions that $\alpha_k \geq 0$, for $k = 1, ..., K$ and that $\sum_{k=1}^{K} \alpha_k = 1$, yields the following updates for the estimates of the α_k parameters:

$$\widehat{\alpha}_k^{(t+1)} = \frac{\max\left\{0, \sum_{i=1}^{n} \bar{z}_{ik}^{(t)} - \frac{C-K+1}{2K}\right\}}{\sum_{j=1}^{K} \max\left\{0, \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} - \frac{C-K+1}{2K}\right\}}, \tag{7}$$

for $k = 1, ..., K$. Notice that, some $\widehat{\alpha}_k^{(t+1)}$ may be zero; in that case, the k-th component is excluded from the mixture model. The multinomial parameters corresponding to components with $\widehat{\alpha}_k^{(t+1)} = 0$ need not be further calculated, since these components do not contribute to the likelihood. For the components with non-zero probability, $\widehat{\alpha}_k^{(t+1)} > 0$, the estimates of multinomial parameters are updated to their standard weighted ML estimates:

$$\widehat{\theta}_{klc}^{(t+1)} = \frac{\sum_{i=1}^{n} \bar{z}_{ik}^{(t)} y_{ilc}}{n_l \sum_{i=1}^{n} \bar{z}_{ik}^{(t)}}, \tag{8}$$

for $k = 1, \ldots, K$, $l = 1, \ldots, L$, and $c = 1, \ldots, C_l$. Notice that, in accordance with the meaning of the θ_{klc} parameters, $\sum_{c=1}^{C_l} \widehat{\theta}_{klc}^{(t+1)} = 1$.

5 Data Analysis and Results

First, we evaluate the performance of the EM-MML algorithm on 10 synthetic data sets, over 50 runs. The data sets were originated from a mixture of 3 categorical variables (with 2, 3 and 4 levels) and 2 components. The correponding Sihouette index values illustrate the structures diversity: 0.099; 0.216; 0.217; 0.230; 0.713; 0.733; 0.746; 0.778; 0.805; 0.817. The obtained results are compared with those obtained from a standard EM algorithm combined with BIC, AIC, CAIC, MAIC, and ICL criteria.

The comparison resorts to a cohesion-separation measure and a concordance measure: the Fuzzy Silhouette index [5] of the clustering structure obtained and the Adjust Rand [10] between the same clustering structure and the original one. In Table 1 we can verify there are no significant differences between the EM-MML and the other criteria, except ICL which only recovers the very well separated structures. Regarding the number of clusters, EM-MML and MAIC are tied, recovering this number correctly for all data sets.The same is not true for the other criteria: AIC identifies 3 clusters in 3 data sets and 4 clusters once; in addition, BIC and CAIC could not find any cluster structure once and ICL was unable to do it for 4 data sets. In terms of computation time, since EM-MML does not require a sequential approach, it becomes clearly faster than the other criteria (Friedman test yields $\chi^2(5)$=2500 and p-value<0.01; Post hoc tests, with Bonferroni correction, only reveal statistically significant differences between the EM-MML and the other criteria).

Table 1 Criteria performance.

Criterion	Number of data sets	Fuzzy Silhouette: 95% CI Lower ; Upper Limits[a]	Adjusted Rand: 95% CI Lower ; Upper Limits[a]
AIC	10	0.430 ; 0.741	0.545 ; 0.867
BIC	9	0.622 ; 0.935	0.728 ; 1.000
CAIC	9	0.616 ; 0.931	0.732 ; 1.000
ICL	6	0.917 ; 0.948	1.000 ; 1.000
MAIC	10	0.568 ; 0.887	0.623 ; 0.950
EM-MML	**10**	**0.561 ; 0.891**	**0.594 ; 0.955**

[a] 1000 bootstrap samples were used to estimate the Confidence Intervals (CI).

Additional insight into the performance of EM-MML is obtained by applying it to a real data set referring to the 6th European Working Conditions Survey (2015), Eurofound working conditions survey. Note that these data are the most recent.

For the purpose of our experiment, we consider the aggregate data referring to 305 European regions and the answers to the following questions: Are you able to

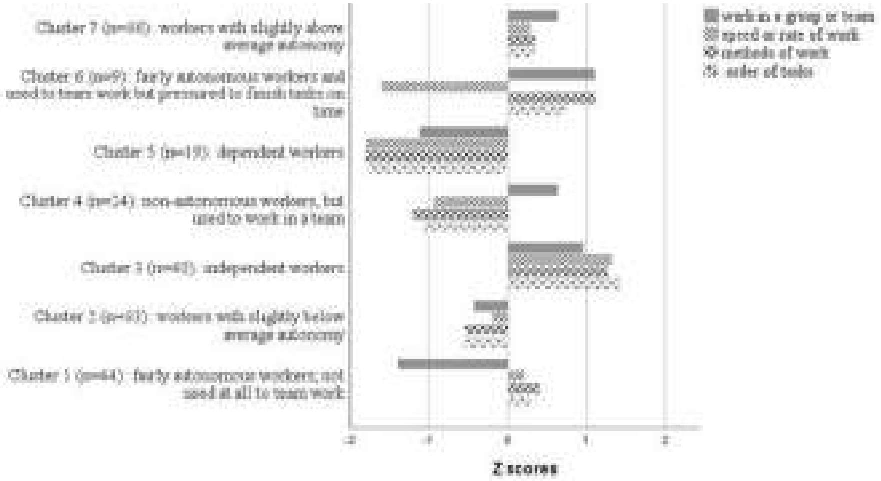

Fig. 1 Clusters' profile and their dimensions (n).

choose or change: a) your order of tasks; b) your methods of work; c) your speed or rate of work. Do you work in a group or team that has common tasks and can plan its work?

EM-MML selected 7 clusters, which is a smaller number than for the remaining criteria (ICL, BIC, CAIC, AIC and MAIC select 10, 12, 12, 15 and 15 respectively). This fact avoids estimation problems associated with very small segments and also improves the interpretability of the clustering solution.

The segments selected by EM-MML criterion are presented in Figure 1. Workers with slightly above average autonomy (cluster 7) live in several countries, but Ireland stands out, as well as Belgium, Germany, Netherlands, Switzerland, and the UK regions. Denmark, Estonia, Malta, and Norway are the countries where the most independent workers are found (cluster 3). The smallest cluster, 6, includes Sweden and a region of Greece and Kriti and Açores, a Greek and a Portuguese region, respectively. The cluster 5, where workers claim they have no autonomy, includes regions from many countries.

6 Discussion and Perspectives

In this work, a model selection criterion and method for finite mixture models of categorical observations was studied - EM-MML. This algorithm simultaneously performs model estimation and selects the number of components/clusters. When compared to information criteria, which are commonly associated with the use of the EM algorithm, the EM-MML method exhibits several advantages: 1) it easily recovers the true number of clusters in synthetic data sets with various degrees of separation; 2) its computations times are significantly lower than those required by standard approaches resorting to the sequential use of EM and an information criterion; 3) when applied to a real data set it produces a more parsimonious solution, thus easier to interpret. An additional advantage of this approach that stems from obtaining more parsimonious solutions is that such solutions have a higher number of observations per cluster, thus helping to overcome eventual estimation problems.

The performance of the EM-MML is encouraging for selecting the number of clusters, and the same criterion was already used for feature selection [17]. However, future research is required, namely considering data sets with different numbers of clusters and high dimensional data.

Acknowledgements This work was supported by Fundação para a Ciência e Tecnologia, grant UIDB /00315/2020.

References

1. Akaike, H.: Maximum likelihood identification of Gaussian autorregressive moving average models. Biometrika. **60**, 255–265 (1973)
2. Baxter, R. A., Olivier, J. J.: Finding overlapping components with MML. Stat. Comput. **10**(1), 5–16 (2000)
3. Bouguila, N., Fan, W.: Mixture Models and Applications (Unsupervised and Semi-Supervised Learning). Springer Nature Switzerland AG, Switzerland (2020)
4. Bozdogan, H.: Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In: Bozdogan, H. (eds.) Proceedings of the First US/Japan Conf. Frontiers of Stat. Modeling, pp.69–113. Boston: Kluwer Academic Publishers (1994)
5. Campello, R. J., Hruschka, E. R.: A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Set. Syst., **157**(21), 2858–2875 (2006)
6. Celeux, G., Martin-Magniette, M. L., Maugis-Rabusseau, C., Raftery, A. E.: Comparing model selection and regularization approaches to variable selection in model-based clustering. J. Soc. Fr. Statistique. **155**(2), 57–71 (2014)
7. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood estimation from incomplete data via the EM Algorithm. J. R. Stat. Soc. **39**, 1–38 (1997)
8. Everitt, B. S., Hand, D.: Finite Mixture Distributions. Chapman and Hall, New York (1981)
9. Figueiredo, M. A. T., Jain, A. K.: Unsupervised learning of finite mixture models. IEEE T. Pattern Anal. **24**, 381–396 (2002)
10. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 93–218 (1985).
11. Kumar, P., Kanavalli, A.: A similarity based K-means clustering technique for categorical data in data mining application. Int. J. Intell. Eng. Syst. (2021) doi: 10.22266/ijies2021.0430.05
12. Lee, C., Jung, U.: Context-based geodesic dissimilarity measure for clustering categorical data. Appl. Sciences (2021) doi: 10.3390/app11188416
13. McLachlan, G. J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
14. Novais, L., Faria, S.: Selection of the number of components for finite mixtures of linear mixed models. J. Int. Math. **24**(8), 2237–2268 (2021)
15. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
16. Silvestre, C., Cardoso, M. G. M. S. and Figueiredo, M.: A clustering view on ESS measures of political interest: an EM-MML approach. NTTS - New Techniques and Technologies for Statistics (2017).
17. Silvestre, C., Cardoso, M. G. M. S. and Figueiredo, M.: Feature selection for clustering categorical data with an embedded modeling approach. Expert Syst. **32**(3), 444–453 (2014).

Typology of Motivation Factors for Employees in the Banking Sector: An Empirical Study Using Multivariate Data Analysis Methods

Áurea Sousa, Osvaldo Silva, M. Graça Batista, Sara Cabral, and Helena Bacelar-Nicolau

Abstract Leadership has been considerate as a competitive advantage for organizations, contributing to their success and effective and efficient performance. Motivation, on the other hand, is assumed as a basic competence of leadership. Therefore, the main purpose of this paper is to know the perceptions of bank employees on the main motivational factors in the organizational context. Data analysis was performed based on several statistical methods, among which the Categorical Principal Component Analysis (CatPCA) and some agglomerative hierarchical clustering algorithms from *VL* (*V* for Validity, *L* for Linkage) parametrical family, applied to the items that aim to assess the aspects most valued by bankers in the work context. The CatPCA allowed to extract four principal components which explain almost 70% of the total data variance. The dendrograms provided by the hierarchical clustering algorithms over the same data, exhibit four main branches, which are associated with different main motivational factors. Moreover, CatPCA and clustering results show an important correspondence concerning the main motivations in this sector.

Keywords: leadership, welfare, motivational factors, CatPCA, cluster analysis

Áurea Sousa (✉)
Universidade dos Açores and CEEAplA, Rua da Mãe de Deus, 9500-321, Portugal,
e-mail: aurea.st.sousa@uac.pt

Osvaldo Silva
Universidade dos Açores and CICSNOVA.UAc, Rua da Mãe de Deus, Portugal,
e-mail: osvaldo.dl.silva@uac.pt

M. Graça Batista
Universidade dos Açores and CEEAplA, Rua da Mãe de Deus, Portugal,
e-mail: maria.gc.batista@uac.pt

Sara Cabral
Universidade dos Açores, Rua da Mãe de Deus, Portugal, e-mail: sara_crc@hotmail.com

Helena Bacelar-Nicolau
Universidade de Lisboa (UL) Faculdade de Psicologia and Institute of Environmental Health (ISAMB/FM-UL), Portugal, e-mail: hbacelar@psicologia.ulisboa.pt

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_39

1 Introduction

Motivation has always been subject of analysis by the scientific community, as numerous definitions have emerged. For Robbins and Judge ([21], p. 184), motivation is defined as "the processes that account for an individual's intensity, direction, and persistence of effort toward attaining a goal". These three indicators are assumed to be key-factors of motivation: intensity describes the individual's effort to achieve the proposed goals; this effort should go in a direction that benefits the organization; and, finally, the persistence with which the individual is able to maintain that effort. In this context, the individual's behavior is determined by what motivates them, which is why their performance results not only from ability and skills, but also from motivation. Moreover, motivation is complex and influenced by innumerable variables, considering the diverse needs and expectations that individuals try to satisfy in different ways [15]. Moreover, different leadership practices may lead to better or worse motivational responses from employees.

The main purpose of this paper is to analyse the perceptions of bank employees who work in the banks that operate in the Autonomous Region of the Azores on the main motivational factors in the organizational context. Our study also intends to perform a reduction of the dimensionality of the data and to find a typology of a set of items that was used to evaluate the latent variable "Motivation", regarding the most valued aspects in the work context. Thus, Section 2 concerns the materials and methods of research. Section 3 presents and discusses the main results of this study. Finally, Section 4 contains the main conclusions.

2 Materials and Methods

This study was based on a quantitative approach, using a validated questionnaire, which can be found in Cabral [7]. The sample consists of 202 bank employees (51.0 % male and 49.0 % female) of the Autonomous Region of the Azores (response rate: 6.4%). Most respondents are 36 years old or older (60.9%) and have higher education (56.7%).

The present study refers to a subset of twenty-seven items used to evaluate the latent variable "Motivation" in work context, namely: 1 - The opportunity for career advancement, 2 - Have greater responsibility, 3 - The feeling of being involved in decision making, 4 - A job that gives you prestige and status, 5 - Have an interesting and challenging job, 6 - The recognition and appreciation of others for the accomplished work, 7 - Have a good relationship with your colleagues, 8 - Have a good relationship with your superiors, 9 - A work environment where there is trust and respect, 10 - The loyalty of superiors towards the collaborators, 11 - Team spirit, 12 - Sense of belonging to the organization, 13 - An adequate discipline, 14 - There is equality of treatment and opportunities between the various employees, 15 - Earn respect and esteem of your colleagues and superiors, 16 - Professional development, 17 - Salary appropriate to the professional functions, 18 - A stable job that gives you security, 19 - Good working conditions, 20 - Balance between personal and professional life, 21 - Being able to express your opinion and ideas without fear of reprisals, 22 - Availability to solve problems/personal situations, 23 - Have a fair and adequate system of objectives and incentives, 24 - Being rewarded for overtime work, 25 - Being pressured to achieve the proposed objectives, 26 - Ability to handle pressure at work, and 27 - Appropriate training to the professional functions.

For each item, respondents could pick only one of six modalities of response according to their level of agreement or disagreement with the items that assess motivation: Totally disagree; Disagree most of the time; Slight disagree; Slight agree; Agree most of the time, and Totally agree. In this study, Categorical Principal Components Analysis (CatPCA), using the Varimax rotation method with Kaiser Normalization; and some agglomerative hierarchical clustering algorithms (AHCA) were used. Data analysis was performed using the packages IBM SPSS Statistics 26 and CLUST11 [19].

Principal Components Analysis (PCA) aims to reduce the dimensionality of the original data so that "the first few dimensions account for as much of the available information as possible" ([9], p. 83), assuming linear relationships among numeric variables. Each principal component is uncorrelated with all others, and it is expressed as a linear combination of the original variables. CatPCA optimally quantifies categorical (ordinal or nominal) variables and can handle and discover nonlinear relationships between variables (e.g., [12]). In the present study, we applied the CatPCA due to the ordinal nature of the items under analysis.

The goal of a clustering algorithm is to obtain a partition, where the elements within a cluster are similar and elements (objects/individuals/groups of individuals or variables) in different groups are dissimilar, identifying natural clustering structures in a data set (e.g., [8]). Agglomerative clustering algorithms usually start with each element to sort into its own separate cluster of size 1 (singleton). At each step, the algorithms find the two "closest" clusters, taking into account the aggregation criterion, and join them. The process continues until a cluster containing all elements to classify is obtained. The AHCA of the set of items was based on the affinity coefficient as a measure of comparison between elements, combined with two classic (Single-Linkage (*SL*) and Complete-Linkage (*CL*)) and a family of probabilistic *VL* (*V* for Validity, *L* for Linkage) aggregation criteria (e. g., [1, 2, 3, 10, 11, 16, 17, 18, 22]).

According to Ng et al. ([20], p. 849), "the task of finding good clusters has been the focus of considerable research in machine learning and pattern recognition". However, the identification of the best partitions using validation indices is also of crucial importance. Therefore, a pertinent question arises: "How well does the partition fit the data?" ([8], p. 505). On what validation of results is concerned, the identification of the best partitions in the present study was based on the global level statistics, *STAT* [1, 10, 11]. The global maximum *STAT* value indicates the best cut-off level of a dendrogram and the local maxima *STAT* differences indicate the most significant levels.

The affinity coefficient between two distribution functions was introduced by Matusita in 1951 (e.g., [13, 14]). Bacelar-Nicolau extended it to the non-supervised classification field as a similarity measure between profiles. Let *V* be a set of p variables, describing a set *D* of *N* statistical data units (individuals), so that each of the $N \times p$ cells of the corresponding data table X contains one single non-negative real value x_{ik} (i = 1,..., *N*; $k = 1, \ldots, p$) which denotes the value of the k-th variable on the i-th individual. The standard affinity coefficient $a(k, k')$ between a pair of variables, V_k and V'_k (k, k' = 1,..., p) is given by formula (1), where $x_{.k} = \Sigma_{i=1}^{N} x_{ik}$, $x_{.k'} = \Sigma_{i=1}^{N} x_{ik'}$.

$$a(k, k') = \Sigma_{i=1}^{N} \sqrt{\frac{x_{ik}}{x_{.k}} \frac{x_{ik'}}{x_{.k'}}} \tag{1}$$

The coefficient (1) is a symmetric similarity coefficient which takes values in [0,1] (1 for equal or proportional vectors and 0 for orthogonal vectors). Note that its mathematical formula corresponds to the inner product between the square root column profiles associated with those variables and measures a monotone tendency between column profiles. In the particular case of binary variables, the affinity coefficient coincides with the well-known Ochiai coefficient. Furthermore (e.g., [4, 6]), it is related to the Hellinger distance d by the relation $d^2 = 2(1 - a)$, which has been used in the context of spherical factor analysis by Michel Volle. Later on, the standard affinity coefficient was extended to the clustering of statistical data units or variables, mainly in a three-way approach (e.g., [3, 4, 5, 6]). The computation of the standard affinity coefficient between individuals can be performed by previously transposing the data matrix and then applying formula (1).

The probabilistic aggregation criteria on the scope of *VL* methodology can be interpreted as distribution functions of statistics of independent random variables, that are i.i.d. uniform on [0, 1] (e.g., [3, 17]). The *SL* aggregation criterion can lead to very long clusters (chaining effect). On the other hand, the *AVL* (Aggregation Validity Link) has a tendency to form equicardinal clusters with an even number of elements. The comparison functions between a pair of clusters, A and B, concerning the family I of *AVL* methods can be generated by the following conjoined formula (e.g., [17, 10, 11]):

$$\Gamma(A, B) = (p_{AB})^{g(\alpha,\beta)} \tag{2}$$

with $\alpha = Card\ A$, $\beta = Card\ B$, $p_{AB} = max[\gamma_{ab} : (a, b) \in (A \times B]$, with $1 \leq g(\alpha, \beta) \leq \alpha\beta$, and γ_{xy} is a similarity measure between pairs of elements, x and y, of the set of elements to classify (e.g., $g(\alpha, \beta) = 1$ for *SL*, $g(\alpha, \beta) = \alpha\beta$ for *AVL*). Note that varying $g(\alpha, \beta)$ with $1 < g(\alpha, \beta) < \alpha\beta$, a sort of compromise can be built between *SL* and *AVL* methods (e.g., $g(\alpha, \beta)$=$(\alpha + \beta)/2$ for *AV1*). Thus, $\Gamma(A, B)$ will be "more polluted by the chain effect when $g(\alpha, \beta)$ remains near 1, and more contaminated by the symmetry effect as long as $g(\alpha, \beta)$ is in the neighbourhood of $\alpha\beta$" ([17], p. 95). Among the criteria that establish a compromise between *AVL* and *SL* methods, stands out the *AV1* method, whose behavior is very similar to that of *AVL* and often provides, at its cut-off level, a partition better adjusted to the preorder than the "best" classification obtained by *AVL*.

3 Main Results and Discussion

Concerning the CatPCA, the best solution comprises four principal components, and the percentage of variance accounted for (PVAF) across these components is almost 70% (about 69%) of the data's total variance. All extracted components have eigenvalues above 1. Moreover, the first three main components have a very good internal consistency and the fourth component has an acceptable internal consistency, as shown by the values of the Cronbach's Alpha coefficient (see Table 1).

Table 1 Rotated component loadings of the 4-component solution - Motivational factors.

Items	PC1	PC2	PC3	PC4
M1	0.213	0.351	**0.699**	0.166
M2	0.197	0.044	**0.794**	0.211
M3	0.248	0.148	**0.763**	-0.018
M4	-0.028	0.098	**0.482**	0.442
M5	0.354	0.219	**0.674**	0.037
M6	**0.522**	0.214	0.425	0.095
M7	**0.837**	0.110	0.193	-0.114
M8	**0.774**	0.151	0.244	0.099
M9	**0.778**	0.227	0.183	-0.125
M10	**0.783**	0.269	0.227	-0.043
M11	**0.757**	0.259	0.223	-0.103
M12	**0.798**	0.155	0.227	-0.035
M13	**0.708**	0.213	0.341	0.070
M14	0.486	**0.511**	0.372	-0.257
M15	**0.775**	0.263	0.252	0.041
M16	0.432	0.364	**0.665**	0.035
M17	0.289	**0.708**	0.410	-0.046
M18	0.462	**0.641**	0.097	-0.247
M19	**0.548**	0.532	0.211	-0.034
M20	0.503	**0.609**	0.074	-0.223
M21	**0.684**	0.401	0.070	0.074
M22	**0.678**	0.399	0.019	0.054
M23	0.295	**0.770**	0.284	0.102
M24	0.174	**0.835**	0.176	-0.011
M25	0.019	-0.012	0.233	**0.864**
M26	-0.038	-0.146	0.035	**0.896**
M27	**0.543**	0.458	0.230	0.227
Eigenvalue (VAF)	7.988	4.417	4.066	2.138
Percentage accounted (PVAF)	29.59	16.36	15.06	7.92
Cronbach's Alpha	0.950	0.934	0.919	0.610

The most important items for the first dimension are items M6, M7, M8, M9, M10, M11, M12, M13, M15, M19, M21, M22, and M27, which are related to human relationships/interactions with colleagues and hierarchical superiors, so it is called

“Psychological well-being/Interpersonal relationships”. This dimension explains the highest proportion of data variance (29.59%).

Concerning the second dimension, the items M14, M17, M18, M20, M23, and M24 are the most important, so this dimension was designated “Remuneration, job stability and incentive system”. The most relevant items regarding the third dimension are M1, M2, M3, M4, M5, and M16; so, this dimension was called “Career progression/Professional achievement”. Finally, the most important items for the fourth dimension are M25 and M26 related to “Fulfilment of the proposed objectives and the timings to achieve them”.

Regarding the AHCA of the same set of items, and considering the best cut-off levels, the results of the present study are summarized in Table 2.

Table 2 The best partition - Standard affinity coefficient.

Method	The best partition	*STAT*	Cut-off level
SL/CL	{M1, M2, M3, M5, M8, M10, M11, M12, M13, M15, M14, M16, M18, M19, M22, M20, M6, M23, M27, M24, M21}; {M4}; {M9}; {M7}; {M25}; {M26}; {M17}	15.8858	20
AV1	{M1, M2, M3, M6, M27, M21, M5, M23, M24, M8, M15, M14, M16, M10, M13, M11, M12, M18, M19, M20, M22}; {M4, M25, M26}; {M7}; {M9}; {M17}	15.6490	22

According to the *STAT* values, the best partitions were obtained by the classic *SL/CL* and the probabilistic *AV1* methods (see Table 2). All dendrograms highlighted four main branches, which are associated with different motivational factors ("Career progression"; "Psychological well-being / Interpersonal relationships"; "Organizational environment and working conditions"; "Conformity with objectives and time to reach them"), bringing new information, and identifying some singletons, as shown in Figure 1.

4 Conclusion

Organizations and their leaders have become increasingly aware of the importance of their employees being well and that negative feelings can negatively affect productivity. Thus, it is essential to ensure the well-being of employees, taking into account the main motivational factors identified in this study. CatPCA made it possible to extract four principal components (dimensions), which explain almost 70% of the total variance of the data, which were designated, respectively, by “Psychological well-being/Interpersonal relationships”; “Remuneration, job stability and incentive system”; “Career progression/Professional achievement”; and “Fulfilment of objectives and timings to achieve them”. Regarding the AHCA of the items that

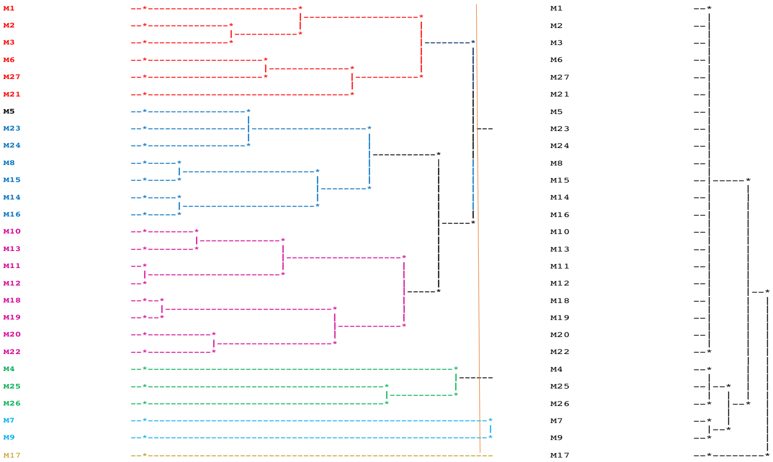

Fig. 1 Dendrogram - Standard affinity coefficient + *AV1*.

assess motivation, the dendrograms highlight four main branches, which are associated with different motivational factors called "Career progression"; "Psychological well-being / Interpersonal relationships"; "Organizational environment and working conditions"; and "Conformity with objectives and time to reach them". They carried new information and identify some singletons as well. Comparing the dendrograms, we conclude that the clusters referring to the best partitions are quite similar, with observed differences mainly concerning the few singletons. Moreover, the effective and fruitful correspondence between the AHCA and the CatPCA results may help to better understand the main types of factors identified. In fact, the four main branches of all dendrograms are related to motivational factors which corresponding interpretation are in consonance with those identified through CatPCA.

Acknowledgements This paper is financed by Portuguese national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., project number UIDB/00685/2020.

References

1. Bacelar-Nicolau, H.: Contributions to the Study of Comparison Coefficients in Cluster Analysis (in Portuguese). Univ. Lisbon (1980)
2. Bacelar-Nicolau, H.: The affinity coefficient in cluster analysis. In: Bekmann, M. J. et al. (eds.). Methods of Operations Research, pp. 507-512. Verlag Anton Hain, Munchen (1985)

3. Bacelar-Nicolau, H.: Two probabilistic models for classification of variables in frequency tables. In: Bock, H._H. (ed.) Classification and Related Methods of Data Analysis, pp. 181-186. Elsevier Sciences Publishers B.V., North Holland (1988)
4. Bacelar-Nicolau, H.: The affinity coefficient. In: Bock, H.-H. and Diday, E. (eds.) Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data, Series: Studies in Classification, Data Analysis, and Knowledge Organization, pp. 160-165. SpringerVerlag, Berlin (2000) doi: 10.1007/978-3-642-57155-8
5. Bacelar-Nicolau, H., Nicolau, F.C., Sousa, Á., Bacelar-Nicolau, L.: Measuring similarity of complex and heterogeneous data in clustering of large data sets. Biocybernetics and Biomedical Engineering, PWN-Polish Scientific Publishers Warszawa. **29**(2), 9-18 (2009)
6. Bacelar-Nicolau, H., Nicolau, F. C., Sousa, Á., Bacelar-Nicolau, L.: Clustering of variables with a three-way approach for health sciences. Testing, Psychometrics, Methodology in Applied Psychology (TPM) (2014) doi: 10.4473/TPM21.4.56
7. Cabral, S.: O Impacto da Liderança na Motivação dos Colaboradores do Setor Bancário na Região Autónoma dos Açores. Universidade dos Açores, Ponta Delgada (2018)
8. Gurrutxaga, I., Muguerza, J., Arbelaitz, O., Pérez, J.M., Martín, J.I.: Towards a standard methodology to evaluate internal cluster validity indices. Pattern Recognit. Lett. **32**, 505-515 (2011)
9. Lattin, J. M., Carrol, J. D., Green, P. E.: Analyzing Multivariate Data (1st ed.). Thomson Brooks, Cole (2003)
10. Lerman, I. C.: Classification et Analyse Ordinale des Données. Dunod, Paris (1981)
11. Lerman, I. C.: Foundations and methods in combinatorial and statistical data analysis and clustering. Series: Advanced Information and Knowledge Processing. Springer-Verlag, Boston (2016)
12. Linting, M., Meulman, J. J., Groenen, P. J., van der Koojj, A. J.: Nonlinear principal components analysis: Introduction and application. Psychol. Meth. **12**(3), 336–358 (2007) doi: 10.1037/1082-989X.12.3.336
13. Matusita, K.: On the theory of statistical decision functions. Ann. Inst. Stat. Math. **3**, 1-30 (1951)
14. Matusita, K.: Decision rules, based on distance for problems of fit, two samples and estimation. Ann. Inst. Stat. Math. **26**, 631-640 (1995)
15. Mullins, L. J.: Management and Organizational Behavior. Prentice Hall, England (2005)
16. Nicolau, F. C., Bacelar-Nicolau, H.: Nouvelles méthodes d'agrégation basées sur la fonction de répartition. In: Collection Séminaires INRIA 1981, Classification Automatique et Perception par Ordinateur, pp. 45-60. Rocquencourt (1982)
17. Nicolau, F. C., Bacelar-Nicolau, H.: Some trends in the classification of variables. In: Hayashi et al. (eds.). Data Science, Classification and Related Methods, pp. 89-98. Springer, Tokyo (1998)
18. Nicolau, F. C., Bacelar-Nicolau, H.: Teaching and learning hierarchical clustering probabilistic models for categorical data. In: Proc. 54th Session of the International Statistical Institute (IASE at ISI, IPM-71). Berlin, Germany (2003) `https://iase-web.org/documents/papers/isi54/3654.pdf.Cited25Jan2022`
19. Nicolau, F. C., Bacelar-Nicolau, H., Sousa, F., Sousa, Á., Silva, O.: CLUST11: Cluster Analysis Software - Standard and *VL* Probabilistic Approaches. LEAD, FP-UL (2011)
20. Ng., A. Y., Jordan, M. I., Weiss, Y. In: On spectral clustering: Analysis and an algorithm (2002) `https://proceedings.neurips.cc/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf.Cited25Jan2022`
21. Robbins, S. P., Judge, T. A.: Organizational Behavior. Pearson Education, Inc., New Jersey (2015)
22. Sousa, Á., Silva, O., Bacelar-Nicolau, H., Nicolau, F. C.: Distribution of the affinity coefficient between variables based on the Monte Carlo simulation method. Asian. J. Appl. Sci. **1**(5), 236-245 (2013)

A Proposal for Formalization and Definition of Anomalies in Dynamical Systems

Jan Michael Spoor, Jens Weber, and Jivka Ovtcharova

Abstract Although many scientists strongly focus on anomaly detection in different applications and domains, there currently exists no universally accepted definition of anomalies and outliers. Using an approach based on control theory and dynamical systems, as well as a definition for anomalies as described by philosophy of science, the authors propose a generalized framework viewing anomalies as key drivers of progress for a better understanding of the dynamical systems around us. By mathematically defining anomalies and delimiting deviations within expectations from completely unforeseen instances, this paper aims to be a contribution to set up a universally accepted definition of anomalies and outliers.

Keywords: anomaly detection, outlier analysis, dynamical systems

1 Introduction

Anomalies, often interchangeably called outliers [1], are of key interest in explorative data analysis. Therefore, anomaly detection finds application in many different scientific fields, i.e., in social science, economics, engineering, and medical science [2]. In particular, research in these domains regarding databases, data mining, machine learning or statistics focuses strongly on anomaly detection [3]. Despite the wide

Jan Michael Spoor (✉)
Institut für Informationsmanagement im Ingenieurwesen (IMI), Karlsruhe Institute of Technology, Karlsruhe, Germany, e-mail: jan.spoor@kit.edu

Jens Weber
Team Digital Factory Sindelfingen, Mercedes-Benz Group AG, Sindelfingen, Germany, e-mail: jens.je.weber@mercedes-benz.com

Jivka Ovtcharova
Institut für Informationsmanagement im Ingenieurwesen (IMI), Karlsruhe Institute of Technology, Karlsruhe, Germany, e-mail: jivka.ovtcharova@kit.edu

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_40

range of anomaly detection, there is currently no universally accepted definition of what an outlier or anomaly is [2], and the mathematical definition depends on the selected method to find these anomalies [4].

The authors previously proposed an applied framework to formalize anomalies within the context of control theory and dynamical systems [5]. In this publication, the idea is discussed in more depth, and a generalization of the framework is proposed to extend its application area to more domains since dynamical systems are relevant in engineering and science [6] as well as in management science and economics [7]. Furthermore, the proposed definition of anomalies should also be applicable outside of the context of control theory and aims to be a contribution to set up a universally accepted definition of anomalies and outliers.

When controlling or simulating dynamical systems, a measurement and prediction process is used. Anomalies occur in this process as substantial deviations of a measured system state (an actual value) from an expected system state (a planned value) [5]. Despite simulation and planning effort, these deviations still occur. While some deviations fall within an acceptable range and within the expectations of normal system behavior, other anomalies are completely unforeseen and do not fit the set-up and expectations of the system. Three sequential questions are derived to further investigate the nature of anomalies within dynamical systems:

1. What distinguishes unforeseen system states from regular system behavior?
2. How can unforeseen system states or errors occur despite simulation?
3. How can unforeseen system states be analyzed and transferred to a standard model of a system's behavior?

2 Definition of Anomalies for Dynamical Systems

2.1 Definitions of Anomalies and Outliers

In general, it is assumed that anomalies are somehow visible within the data of the observed systems. This is also clearly stated by the definition of an outlier or anomaly as data points with a substantial deviation from the norm since this requires a normal state of the system and a measurable deviation [8]. Furthermore, the anomaly detection requires existence and knowledge of a normal state, a definition of a deviation, a metric, and a threshold measure of distance. This threshold measure of distance uses the selected metric. All distances between the norm and the data points, which are either above (in case of distance measures) or below (in case of similarity measures) the defined threshold, are assumed to be non-substantial.

Therefore, in addition, the selection of an appropriate metric becomes an important tool to accurately describe an anomaly. Some authors claim that, in a practical application, the selection of a suitable metric might be more important than the algorithm itself. For example, if clusters are clearly separated within the examined dataset in context of the selected metric, clusters will be found independently of the used method or algorithm [9]. Other authors claim that the selected method for investigating clusters is of importance [10].

To summarize, there is no trivial definition of a normal state, a deviation, and when a deviation might be substantial. Some authors therefore describe the usefulness of an analysis only within the context of the goals of the analysis [11]. Outlier detection becomes more of a technical target than an actual scientific finding of something novel since the novelty is always defined within the technical target of the analysis. Alternatively, the normal model of the data defines an anomaly [1].

This results, for example, in approaches of regression diagnostics to exclude outliers and anomalous data prior to an analysis or to conduct the analysis along the standard model in a more robust way, which is less affected by anomalies [12]. Both approaches result in the maintaining of the normal model using anomalies as if they were less adequate or not at all representative of the data set.

Since anomalies are only relevant within a context, a typology of anomalies within different dataset contexts can be created. Thus, Foorthuis [13] proposes a typology along the following dimensions: types of data (qualitative, quantitative or mixed), anomaly level (atomic or aggregated) and cardinality of relationship (univariate or multivariate). Anomalies are, within this kind of typology, always dependent on the dataset and behave differently along the measured features, which have been classified as relevant for the specific analysis. The anomaly detection becomes a detection of unfitting, surprising values while maintaining the normal model.

2.2 Definition by Philosophy of Science

If the assumptions regarding normal states, deviation, and substantiality are dropped, it is possible to discuss anomalies on a more fundamental level for understanding our surroundings and the observations of them.

To do this, anomalies have to be placed in the historic context of science and research. Since anomaly detection as a discipline of data science is placed within the scientific context [14], anomaly detection can also be analyzed as part of the scientific method and therefore a comparison with the historical understanding of anomalies in the context of science becomes relevant. By definition of Kuhn [15], anomalies play an important role in the scientific discovery of novelties:

> Discovery commences with the awareness of anomaly, i.e., with the recognition that nature has somehow violated the paradigm-induced expectations that govern normal science. It then continues with a (...) exploration of the area of anomaly. And it closes only when the paradigm theory has been adjusted so that the anomalous has become the expected.

This statement describes scientific progress as a stepwise discovery and the placement of anomalies within a normal state by science. The discussed normal state is therefore dictated by current scientific knowledge, which encompasses the predictions of the currently available and widely used models and theories. An anomaly violates the normal state by violating the predictions of these models. The steps of scientific progress are then as follows:

1. Knowledge of the anomaly.
2. Stepwise acknowledgement of observations and conceptual nature of the anomaly.
3. Change of paradigm and methods to include the anomaly in the new models, often under resistance by the scientific community itself.

Therefore, different states of an anomaly exist as follows:

1. The anomaly is completely unknown.
2. The anomaly is neither described nor modeled but was observed.
3. The anomaly is not commonly recognized and placed within the standard model.

The states of anomalies correspond to the initially defined questions in the introduction regarding the delimitation of anomalous states from normal states, the exploration of the causes for anomalies, and the modeling and planning with the now known anomalies. If the states of anomalies are used to describe practical errors in engineering, error states of systems are not anomalies. This is the case because if error states are priorly classified as such, they are therefore already known and described. This corresponds to the idea that outliers or anomalies are created by a different underlying mechanism [16] and therefore imply an unknown system behavior, which needs modeling to better describe the system. In addition, this follows the assumption of a normal state in which anomalies simply derive from a normal model [1] since they are not part of the normal model. Also, this idea relates strongly to the discussion of the relation between novelty and anomaly detection [17].

To follow the definitions by Kuhn [15], science is driven by internal progress, limited by the current methods and available resources, while external targets, defined by stakeholders, e.g., society or companies, drive technicians. This description matches the idea that the usefulness of an analysis should be evaluated within the context of its goals [11] and distinguishes two types of anomalies: "Scientific" anomalies of a novel observation and "technical" anomalies as deviations from a predefined norm using a predefined measurement of substantiality.

"Scientific" anomalies might still result in unwanted system states, which then can result in some kind of error or critical system state. Nevertheless, not every "scientific" anomaly inevitably results in an error state and not every error state is a "scientific" anomaly. An anomaly is not a "scientific" anomaly if the error state is already documented or can be described by the standard model. In this case, the anomaly becomes a "technical" anomaly.

Using the philosophy of science definition of anomalies, the normal state is the prediction by the system model, the deviation is the difference between the prediction of the system state and the measured actual state of the system, and the substantiality is defined by the noise and precision of our predictions and measurement tools.

3 Proposed Framework for a Formalization of Anomalies

To separate "scientific" and "technical" anomalies, a formerly proposed framework [5] is generalized as illustrated in Fig. 2. and mathematically defined in this section.

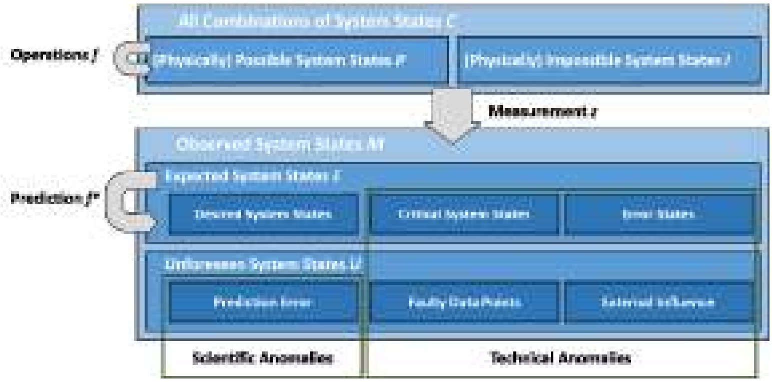

Fig. 1 Formalization of "scientific" and "technical" anomalies and system states.

Definition 1 (System State) There exists a multivariate description x_i of a state i with a finite number of features. For each feature j of state i a value x_{ij} exists, which is a realization of the feature space R_j. The value x_{ij} is the actual and precise state description of feature j at state i. Although there exists only a single true value x_{ij}, the value itself does not necessarily have to be a single data point but can be a multivariate or symbolic data value and can be of any data type.

$$\forall i \, \forall j \, \exists! \, x_{ij}, \quad x_{ij} \in R_j \tag{1}$$

The set C of all combinations of system state values with J features is given by:

$$C = \{x_i \mid \forall j \, \exists \, x_{ij} \in R_j\} = R_1 \times ... \times R_J \tag{2}$$

Definition 2 (Operation) An operation is an analytical function f which changes the system state from state i to the following state $i + 1$. Both states belong to the set of all combinations of system states C.

$$f : C \to C, \quad f(x_i) = x_{i+1} \tag{3}$$

There exists a finite set F of functions of endogenous state transformations. This set of functions is the scope of operations that can be performed. These functions are the fundamental functionality of a system, which can be performed without any external involvement. For all functions the following expression is applied:

$$g \in F \wedge f \in F : g \circ f \in F \tag{4}$$

Using the defined function space, a restriction of reachable system states via all functions from F is defined, resulting in the set of physically possible system states.

Definition 3 (Physically Possible System States) The relation f spans the complete space of state changes of a system using the entire scope of operations. The resulting space is the set of all possible system states. The physically possible system states are the possible realizations of x_i based on a starting point and if only functions from F are applied. The set P is a group with a neutral element of operations.

$$P = \{x_i \mid \forall f \in F : f(x_i) \in P\} \subseteq C \tag{5}$$

Definition 4 (Observed System States) Of the amount J of existing features of the system state, only an amount D of features is known with $D \leq J$. Since not all system states can be measured, a function z transforms the real system states and real operations of the system into observable system states and operations.

$$z : C \rightarrow M, \quad z(x_i) = x_{i^*} \tag{6}$$

Therefore, the set $M = R_1 \times ... \times R_D$ is the space of all observable and known system states. Function z is the measurement process.

Definition 5 (Observed Operations) Not all functions of the whole set of function F are known or observable when planning and operating a system.

$$F' \subseteq F \tag{7}$$

Additionally, only observable system states are modeled when operating a system. The observed operations of systems are therefore projections of a subsets of known operations of F and operate within the observed and known system states.

$$F^* = z(F') \tag{8}$$

The actual conducted operations f are always from the set of operations F, but the expectation and prediction utilize, due to lack of system knowledge, only $f^* \in F^*$.

$$f^* : M \rightarrow M, \quad f^*(x_{i^*}) = x_{i+1^*} \tag{9}$$

Therefore, all states applied in operation f^* are defined as expected system states.

Definition 6 (Expected System States) The system states, which are possible if only the observed and known operations of the set F^* are applied to all system states $x_{i^*} \in E$, are the expected system behavior.

$$E = \{x_{i^*} \mid \forall f^* \in F^* : f^*(x_{i^*}) \in E\} \subseteq M \tag{10}$$

The expected system states can be further split into desired system states, where the system is running most beneficially for its usage, a critical system state, where a possible error or rare system states are measured, and error states, which are system faults with operational risks involved as defined by Basel III [18]. Applied in engineering, this definition is compatible with the definition of DIN EN 13306 since the system is at risk of being unable to perform a certain range of functions without necessarily being completely inoperable [19]. All kinds of errors, warnings and non-beneficial system states are the "technical" anomalies within the contextual analysis of the data set.

Definition 7 (Unforeseen System States) The set of unforeseen system states U are therefore all measurable system states within the realm of observable system states but not within the expected system states:

$$U = M/E \tag{11}$$

"Scientific" anomalies in unforeseen system states are measured if the real operation f differs from f^* such that a prediction error occurs:

$$f^*(x_{i^*}) \in E, \quad f^*(x_{i^*}) \neq z(f(x_i)) \notin E \tag{12}$$

"Scientific" anomalies are part of the unforeseen system states. Another reason for unforeseen system states is a measurement of an impossible system state. Anomalies originated by physically impossible system states are to be distinguished from "scientific" anomalies since the reason for their occurrence follows a different mechanism. Thus, they are assigned to the "technical" anomalies.

Definition 8 (Physically Impossible System States) Physically impossible system states I are combinations of states in set C which are not reachable using function f:

$$I = C/P \tag{13}$$

Definition 9 (External Influence) Applying changes to the system, the feature space also changes. Consequently, the space of the physically possible system states changes. Previously impossible system states become possible system states.

Definition 10 (Faulty Data Points) If a measurement is conducted incorrectly, the measured values could be within the impossible system states. Faulty data points are therefore neither measurement noise nor imprecision, but should be systematically excluded. Note that faulty data points could be within the possible system space but need to be excluded either way.

4 Conclusion

It is concluded that the anomaly concept is often loosely defined and heavily depends on assumptions of a normal state, deviation, and substantiality. These definitions are often case-specific and influenced by the conducting researchers' choice. Therefore, a rigorous definition of anomalies is capable of further streamlining the discourse and increasing a common understanding of what kind of anomaly is described.

Using "technical" and "scientific" anomalies, further research will be conducted to set up models detecting both types of anomalies separately. Differences between observed and real system states and operations are a focus of further research to more precisely analyze the hidden processes of the "scientific" anomaly generation. Also, a more fundamental discussion of the philosophical definition of anomalies within the philosophy of science and its applications to anomaly detection in general should be conducted to further gain insight into the true nature of anomalies.

The authors plan to validate the concept by using the proposed definition and framework in exemplary applications within industrial processes. Furthermore, anomaly detection methods designed for applications in dynamical systems using the proposed framework are planned to be developed.

Acknowledgements The Mercedes-Benz Group AG funds this research. The research was prepared within the framework of the doctoral program of the Institut für Informationsmanagement im Ingenieurwesen (IMI) at the Karlsruhe Institute of Technology (KIT).

References

1. Aggarwal, C. C.: Outlier Analysis. Springer Science+Business Media, New York (2013)
2. Hodge, V. J., Austin, J.A.: Survey of outlier detection methodologies. Artif. Intell. Rev. **22**, 85-126 (2004)
3. Aggarwal, C. C., Sathe, S.: Outlier Ensembles. Springer, Cham (2017)
4. Wang, X., Wang, X., Wilkes M.: New Developments in Unsupervised Outlier Detection - Algorithms and Applications. Springer, Singapore (2021)
5. Spoor, J. M., Weber, J., Ovtcharova, J.: A definition of anomalies, measurements and predictions in dynamical engineering systems for streamlined novelty detection. Accepted for the 8th International Conference on Control, Decision and Information Technologies (CoDIT), Istanbul (2022)
6. Åström, K. J., Murray, R. M.: Feedback Systems - An Introduction for Scientists and Engineers. Princeton University Press, Princeton, New Jersey (2008)
7. Sethi, S. P., Thompson, G. L.: Optimal Control Theory - Applications to Management Science and Economics. Springer Science+Business Media, Boston, MA (2000)
8. Mehrotra, K. G., Mohan, C., Huang, H.: Anomaly Detection - Principles and Algorithms. Springer International Publishing, Cham (2017)
9. Skiena, S. S.: The Data Science Design Manual. Springer International Publishing, Cham (2017)
10. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer Science+Business Media, New York (2013)
11. Fahrmeier, L., Hamerle, A., Tutz, G. (ed.): Multivariate Statistische Verfahren. de Gruyter, Berlin (1996)
12. Rousseeuw, P. J., Leroy, A. M.: Robust Regression and Outlier Detection. John Wiley & Sons, Inc (1987)
13. Foorthuis, R.: On the nature and types of anomalies: A review of deviations in data. Int. J. Data Sci. Anal. **12**, 297-331 (2021)
14. Cuadrado-Gallego, J. J., Demchenko, Y.: The Data Science Framework: A View from the EDISON Project. Springer Nature Switzerland AG, Cham (2020)
15. Kuhn, T.: The Structure of Scientific Revolutions. 2nd ed. The University of Chicago Press, Chicago (1970)
16. Hawkins, D.: Identification of Outliers. Chapman and Hall (1980)
17. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. **41**(3) 15 (2009)
18. Bank for International Settlements: Basel Committee on Banking Supervision: International Convergence of Capital Measurement and Capital Standards (2006)
19. DIN Deutsches Institut für Normung e. V.: DIN EN 13306: Instandhaltung - Begriffe der Instandhaltung. Beuth Verlag GmbH, Berlin (2010)

New Metrics for Classifying Phylogenetic Trees Using K-means and the Symmetric Difference Metric

Nadia Tahiri and Aleksandr Koshkarov

Abstract The k-means method can be adapted to any type of metric space and is sometimes linked to the median procedures. This is the case for symmetric difference metric (or Robinson and Foulds) distance in phylogeny, where it can lead to median trees as well as to Euclidean Embedding. We show how a specific version of the popular k-means clustering algorithm, based on interesting properties of the Robinson and Foulds topological distance, can be used to partition a given set of trees into one (when the data is homogeneous) or several (when the data is heterogeneous) cluster(s) of trees. We have adapted the popular cluster validity indices of Silhouette, and *Gap* to tree clustering with k-means. In this article, we will show results of this new approach on a real dataset (aminoacyl-tRNA synthetases). The new version of phylogenetic tree clustering makes the new method well suited for the analysis of large genomic datasets.

Keywords: clustering, symmetric difference metrics, k-means, phylogenetic trees, cluster validity indices

1 Introduction

In biology, one of the most significant organizing principles is the "Tree of Life" (ToL) [12]. In genetic studies, there is evidence of an enormous number of branches, but even a rough estimate of the total size of the tree remains difficult. Many recent

Nadia Tahiri (✉)
Department of Computer Science, University of Sherbrooke, Sherbrooke, QC J1K2R1, Canada,
e-mail: Nadia.Tahiri@USherbrooke.ca

Aleksandr Koshkarov
Department of Computer Science, University of Sherbrooke, Sherbrooke, QC J1K2R1, Canada;
Center of Artificial Intelligence, Astrakhan State University, Astrakhan, 414056, Russia,
e-mail: Aleksandr.Koshkarov@USherbrooke.ca

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_41

representations of ToL have emphasized either the existence of deep evolutionary relationships [7] or the knowledge of a large and diverse variety of life, with an emphasis on Eukaryotes [8]. These approaches do not consider the dramatic evolution in our understanding of the diversity of life due to genomic sampling of previously unexplored environments.

As a result, Maddison in 1991 [11] was the first to formulate the idea of multiple consensus trees when he described his phylogenetic island method. He observed that island consensus trees can differ significantly from each other and are generally better resolved than the species-wide consensus tree. The most intuitive approach to discovering and clustering genes that share similar evolutionary histories is to cluster their genetic phylogenies. In this context, Stockham et al. in 2002 [18] proposed a tree clustering algorithm based on k-means [4, 9, 10] and the Robinson and Foulds quadratic distance [15]. Their clustering algorithm aims to infer a set of strict consensus trees, minimizing information loss. They proceed by determining the consensus trees for each set of clusters in all intermediate partitioning solutions tested by k-means. This makes the Stockham et al. algorithm very expensive in terms of execution time. More recently, Tahiri et al. in 2018 [19] proposed a fast and accurate tree clustering method based on k-medoids. Finally, Silva and Wilkinson in 2021 [17] introduced a revised definition of tree islands based on any tree-to-tree metric that usefully extends this notion to any set or multiset of trees and provided an interesting discussion of biological applications of their method.

In this context, the use of a method that infers multiple supertrees (i.e., a supertree clustering method) would help discover and cluster alternative evolutionary scenarios for several ToL subtrees.

The paper is structured as follows. In the next section, we introduce a new metric for k-means algorithm based on the Robinson and Foulds distance. The section 3 presents the simulation results (on a real dataset) obtained with our algorithm compared to other clustering methods. Finally, we discuss our contributions in section 4.

2 Methods

The k-means algorithm [9, 10] is a very common algorithm for data parsing. From a set of N observations $x_i, \ldots, x_N$ each one being described by M variables, this algorithm creates a partition in k homogeneous classes or clusters. Each observation corresponds to a point in a M-dimensional space and the proximity between two points is measured by the distance between them. In the framework of k-means, the most commonly used distances are the Euclidean distance, Manhattan distance, and Minkowski distance [4]. To be precise, the objective of the algorithm is to find the partition of the N points into k clusters in such a way that the sum of the squares of the distances of the points to the center of gravity of the group to which they are assigned is minimal. To the best of our knowledge, finding an optimal partition according to the k-means least-squares criterion is known to be NP-hard [13]. Considering this fact, several polynomial-time heuristics were developed, most of which have the time complexity of $O(KNIM)$ for finding an approximate partitioning solution, where K is the maximum possible number of clusters, N is the number of objects (for example, phylogenetic trees), I is the number of iterations in the k-means algorithm, and M is the number of variables characterizing each of the N objects.

A well-known metric of comparing two tree topologies in computational biology is the Robinson-Foulds distance (RF), also known as the symmetric-difference distance [15]. Moreover, the distance RF is a topological distance, which means that it does not consider the length of the edges of the tree. The formula of RF distance can be describe as $(n_1(T_1) + n_2(T_2))$, where $n_1(T_1)$ is the number of partitions of data implied by the tree T_1, but not the tree T_2 and $n_2(T_2)$ is the number of partitions of data implied by the tree T_2 but not the tree T_1. According to Barthélemy and Monjardet [1], the majority-rule consensus tree of a set of trees is the median tree of this set. This fact makes the use of tree clustering possible.

2.1 Silhouette Index Adapted for Tree Clustering

The first popular cluster validity index we consider in our study is the Silhouette width (SH) [16]. Traditionally, the Silhouette width of the cluster k is defined as follows:

$$s(k) = \frac{1}{N_k}\left[\sum_{i=1}^{N_k} \frac{b(i) - a(i)}{max(a(i), b(i))}\right], \tag{1}$$

where N_k is the number of objects belonging to cluster k, $a(i)$ is the average distance between object i and all other objects belonging to cluster k, and $b(i)$ is the smallest, over all clusters k' different from cluster k, of all average distances between i and all the objects of cluster k'.

We used Equations (2) and (4) for calculating $a(i)$ and $b(i)$, respectively, in our tree clustering algorithm (see also [19]). For instance, the quantity $a(i)$ can be calculated as follows:

$$a(i) = \left[\frac{\sum_{j=1}^{N_k} RF(T_{ki}, T_{kj})}{2n(T_{ki}, T_{kj}) - 6} + \xi\right] / N_k \ , \tag{2}$$

where N_k is the number of trees in cluster k, T_{ki} and T_{kj} are, respectively, trees i and j in cluster k, $n(T_{ki})$ is the number of leaves in tree T_{ki}, $n(T_{kj})$ is the number of leaves in tree T_{kj}, and ξ is a penalty function which is defined as follows:

$$\xi = \alpha \times \frac{Min(n(T_{ki}), n(T_{kj})) - n(T_{ki}, T_{kj})}{Min(n(T_{kj}), n(T_{kj}))} \ , \tag{3}$$

where α is the penalization (tuning) parameter, taking values between 0 and 1, used to prevent from putting to the same cluster trees having small percentages of leaves in common, and $n(T_{ki}, T_{kj})$ is the number of common leaves in trees T_{ki} and T_{kj}.

The formula for $b(i)$ is as follows:

$$b(i) = \min_{1 \le k' \le K, k' \ne k} \left[\frac{\sum_{j=1}^{N_{k'}} RF(T_{ki}, T_{k'j})}{2n(T_{ki}, T_{k'j}) - 6} + \xi \right] / N_{k'} \,, \tag{4}$$

where $T_{k'j}$ is the tree j of the cluster k', such that $k' \ne k$, and $N_{k'}$ is the number of trees in the cluster k'.

The optimal number of clusters, K, corresponds to the maximum average Silhouette width, SH, which is calculated as follows:

$$SH = \overline{s}(K) = \sum_{k=1}^{K} \left[s(k) \right] / K \,. \tag{5}$$

The value of the Silhouette index defined by Equation (5) ranges from -1 to +1.

2.2 *Gap* Statistic Adapted for Tree Clustering

It is worth noting that the *SH* cluster validity index (Equations (1) to (5)) do not allow comparing the solution consisting of a single consensus tree ($K = 1$; the calculation of *SH* is impossible in this case) with clustering solutions involving multiple consensus trees or supertrees ($K \ge 2$). This can be considered as an important disadvantage of the *SH*-based classifications because a good tree clustering method should be able to recover a single consensus tree or supertree when the input set of trees is homogeneous (e.g. for a set of gene trees that share the same evolutionary history).

The *Gap* statistic was first used by Tibshirani et al. [20] to estimate the number of clusters provided by partitioning algorithms. The formulas proposed by Tibshirani et al. were based on the properties of the Euclidean distance. In the context of tree clustering, the *Gap* statistic can be defined as follows. Consider a clustering of N trees into K non-empty clusters, where $K \ge 1$. First, we define the total intracluster distance, D_k, characterizing the cohesion between the trees belonging to the same cluster k:

$$D_k = \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \left[\frac{RF(T_{ki}, T_{kj})}{2n(T_{ki}, T_{kj}) - 6} + \xi \right] . \tag{6}$$

Then, the sum of the average total intracluster distances, V_K, can be calculated using this formula:

$$V_K = \sum_{k=1}^{K} \frac{1}{2N_k} D_k \,. \tag{7}$$

Finally, the *Gap* statistic, which reflects the quality of a given clustering solution including K clusters, can be defined as follows:

$$Gap_N(K) = E_N^*\left\{\log(V_K)\right\} - \log(V_K) \ . \tag{8}$$

where E_N^* denotes expectation under a sample of size N from the reference distribution. The following formula [20] for the expectation of $log(V_K)$ was used in our algorithm:

$$E_N^*\left\{\log(V_K)\right\} = \log(Nn/12) - (2/n)\log(K) \ , \tag{9}$$

where n is the number of tree leaves.

The largest value of the *Gap* statistic corresponds to the best clustering.

3 Results - A Biological Example

To illustrate the methods described above, we used a dataset from Woese et al. [22]. The aminoacyl-tRNA synthetases (aaRSs) are enzymes that attach the appropriate amino acid onto its cognate transfer RNA. The structure-function aspect of aaRSs has long attracted the attention of biologists [22, 6]. Moreover, the relationship of aaRSs to the genetic code is observed from the evolutionary view (the central role played by the aaRSs in translation would suggest that their histories and that of the genetic code are somehow intertwined [22]). The novel domain additions to aaRSs genes play an important role in the inference of the ToL.

We encoded 20 original aminoacyl-tRNA synthetase trees from Woese et al. [22] in Newick format and then split some of them into sub-trees to account for cases where the same species appeared more than once in the original tree. Our approach cannot handle data that includes multiple instances of the same species in the input trees. Thus, 36 aaRS trees with different numbers of leaves (including 72 species in total) were used as input of our algorithm (their Newick strings are available at: `https://github.com/tahiri-lab/PhyloClust`). Our approach was applied with the α parameter set to 1.

First, we implemented our new approach with the *Gap* statistic cluster validity index which suggested the presence of 7 clusters of trees in the data, thus suggesting a heterogeneous scenario of their evolution. Then, we conducted the computation using the *SH* cluster validity index and obtained 2 clusters of trees each of which could be represented by its own supertree. The first cluster obtained using *SH* included 19 trees for a total of 56 organisms, whereas the second cluster included 17 trees for a total of 61 organisms. The supertrees (see Figure 1) for the two obtained clusters of trees were inferred using the CLANN program [5]. Further, we decided to infer the most common horizontal gene transfers which characterized the evolution of gene trees included in the two obtained tree clusters. The method of [3], reconciling the species and gene phylogenies to infer transfers, was used for this purpose. The species phylogenies followed the NCBI taxonomic classification. These phylogenies were not fully resolved (the species phylogeny in Figure 1a contains 9 internal nodes

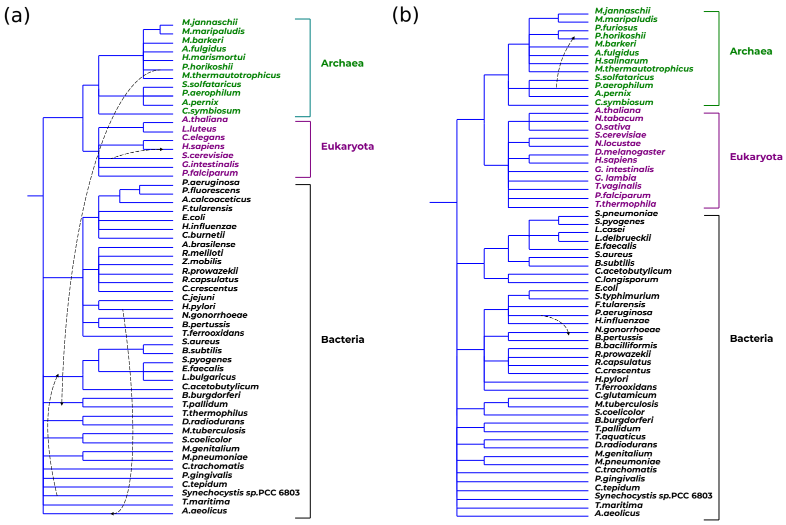

Fig. 1 Nonbinary species tree corresponding to the NCBI taxonomic classification are represented with (a) 56 species for cluster 1. The 4 HGTs (indicated by arrows) were found by the SH index for the first cluster; (b) 61 species with α equal 1 for cluster 2. The 2 HGTs (indicated by arrows) were found by the SH index with α equal 1 for the second cluster. We applied Most Similar Supertree Method ($dfit$) [5] implemented in CLANN Software with mrp criterion. This criterion is a matrix representation employing parsimony criterion.

with a degree higher than 3 and the species phylogeny in Figure 1b contains 10 internal nodes with a degree higher than 3).

We used the version of the HGT (Horizontal Gene Transfer) algorithm available on the T-Rex web site [2] to identify the scenarios of HGT events that reconcile the species tree and each of the supertrees. We choose the same root between species trees and supertrees: the root which split Bacteria to the clade of Eukayota and Archaea.

For the first cluster composed of 56 species, we obtained 40 transfers with 22 regular and 18 trivial HGTs. Trivial HGTs are necessary to transform a non-binary tree into a binary tree. We removed the trivial HGTs and selected between regular HGTs. The non-trivial HGTs with low representation are most likely due to the tree reconstruction artefacts. In Figure 1a, we illustrated only those HGTs that are most represented in the dataset.

We followed the same procedure for the second cluster composed of 61 species and obtained 42 transfers with 28 regular and 14 trivial HGTs that are not represented here. We selected only the most popular HGTs in the dataset. All other transfers are represented in Figure 1b.

The transfers link of *P. horikoshii* to the clade of *spirochetes* (i.e. *B. burgdorferi* and *T. pallidum*) was found by [3, 14]. The transfers of *P. horikoshii* to *P. aerophilum* were also found by [14]. These results confirmed the existing HGT of [3, 14].

4 Discussion

Many research groups are estimating trees containing several thousands to hundreds of thousands of species, toward the eventual goal of the estimation of the Tree of Life, containing perhaps several million leaves. These phylogenetic estimations present enormous computational challenges, and current computational methods are likely to fail to run even with datasets on the low end of this range. One approach to estimate a large species tree is to use phylogenetic estimation methods (such as maximum likelihood) on a supermatrix produced by concatenating multiple sequence alignments for a collection of markers; however, the most accurate of these phylogenetic estimation methods are extremely computationally intensive for datasets with more than a few thousand sequences. Supertree methods, which assemble phylogenetic trees from a collection of trees on subsets of the taxa, are important tools for phylogeny estimation where phylogenetic analyses based upon maximum likelihood (ML) are infeasible.

In this article, we described a new algorithm for partitioning a set of phylogenetic trees in several clusters in order to infer multiple supertrees, for which the input trees have different, but mutually overlapping sets of leaves. We presented new formulas that allow the use of the popular Silhouette and *Gap* statistic cluster validity indices along with the Robinson and Foulds topological distance in the framework of tree clustering based on the popular k-means algorithm. The new algorithm can be used to address a number of important issues in bioinformatics, such as the identification of genes having similar evolutionary histories, e.g. those that underwent the same horizontal gene transfers or those that were affected by the same ancient duplication events. It can also be used for the inference of multiple subtrees of the Tree of Life. In order to compute the Robinson and Foulds topological distance between such pairs of trees, we can first reduce them to a common set of leaves. After this reduction, the Robinson and Foulds distance is normalized by its maximum value, which is equal to $2n - 6$ for two binary trees with n leaves. Overall, the good performance achieved by the new algorithm in both clustering quality and running time makes it well suited for analyzing large genomic and phylogenetic datasets. A C++ program, called PhyloClust (Phylogenetic trees Clustering), implementing the discussed tree partitioning algorithm is freely available at `https://github.com/tahiri-lab/PhyloClust`.

Acknowledgements We would like to thank Andrey Veriga and Boris Morozov for helping us with the analysis of Aminoacyl-tRNA synthetases data. We also thank Compute Canada for providing access to high-performance computing facilities. This work was supported by Fonds de Recherche sur la Santé of Québec and University of Sherbrooke grant.

References

1. Barthelemy, J., Monjardet, B.: The median procedure in cluster analysis and social choice theory. Math. Soc. Sci. **1**, 235-267 (1981)
2. Boc, A., Legendre, P., Makarenkov, V.: An efficient algorithm for the detection and classification of horizontal gene transfer events and identification of mosaic genes. Algorithms From And For Nature And Life. pp. 253-260 (2013)
3. Boc, A., Philippe, H., Makarenkov, V.: Inferring and validating horizontal gene transfer events using bipartition dissimilarity. Syst. Biol. **59**, 195-211 (2010)
4. Bock, H.: Clustering methods: a history of k-means algorithms. Selected Contributions In Data Analysis And Classification. pp. 161-172 (2007)
5. Creevey, C., Fitzpatrick, D., Philip, G., Kinsella, R., O'Connell, M., Pentony, M., Travers, S., Wilkinson, M., McInerney, J.: Does a tree–like phylogeny only exist at the tips in the prokaryotes?. Proc. Roy. Soc. Lond. B Biol. Sci. **271**, 2551-2558 (2004)
6. Godwin, R., Macnamara, L., Alexander, R., Salsbury Jr, F.: Structure and dynamics of tRNA-met containing core substitutions. ACS Omega. **3**, 10668-10678 (2018)
7. Gouy, R., Baurain, D., Philippe, H.: Rooting the tree of life: the phylogenetic jury is still out. Phil. Trans. Biol. Sci. **370**, 20140329 (2015)
8. Hinchliff, C., Smith, S., Allman, J., Burleigh, J., Chaudhary, R., Coghill, L., Crandall, K., Deng, J., Drew, B., Gazis, R. et al.: Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl. Acad. Sci. Unit. States Am. **112**, 12764-12769 (2015)
9. Lloyd, S.: Least squares quantization in PCM. IEEE Trans. Inform. Theor. **28**, 129-137 (1982)
10. MacQueen, J. et al.: Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability. **1**, 281-297 (1967)
11. Maddison, D.: The discovery and importance of multiple islands of most-parsimonious trees. Syst. Biol. **40**, 315-328 (1991)
12. Maddison, D., Schulz, K., Maddison, W. et al.: The tree of life web project. Zootaxa. **1668**, 19-40 (2007)
13. Mahajan, M., Nimbhorkar, P., Varadarajan, K.: The planar k-means problem is NP-hard. International Workshop On Algorithms And Computation. pp. 274-285 (2009)
14. Makarenkov, V., Boc, A., Delwiche, C., Philippe, H. et al.: New efficient algorithm for modeling partial and complete gene transfer scenarios. Data Science And Classification. 341-349 (2006)
15. Robinson, D., Foulds, L.: Comparison of phylogenetic trees. Math. Biosci. **53**, 131-147 (1981)
16. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53-65 (1987)
17. Silva, A., Wilkinson, M.: On defining and finding islands of trees and mitigating large island bias. Syst. Biol. **70**6, 1282-1294 (2021)
18. Stockham, C., Wang, L., Warnow, T.: Statistically based postprocessing of phylogenetic analysis by clustering. Bioinformatics. **18**, S285-S293 (2002)
19. Tahiri, N., Willems, M., Makarenkov, V.: A new fast method for inferring multiple consensus trees using k-medoids. BMC Evol. Biol. **18**, 1-12 (2018)
20. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a dataset via the gap statistic. J. Roy. Stat. Soc. B Stat. Meth. **63**, 411-423 (2001)
21. Whidden, C., Zeh, N., Beiko, R.: Supertrees based on the subtree prune-and-regraft distance. Syst. Biol. **63**, 566-581 (2014)
22. Woese, C., Olsen, G., Ibba, M., Soll, D.: Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. **64**, 202-236 (2000)

On Parsimonious Modelling via Matrix-variate t Mixtures

Salvatore D. Tomarchio

Abstract Mixture models for matrix-variate data have becoming more and more popular in the most recent years. One issue of these models is the potentially high number of parameters. To address this concern, parsimonious mixtures of matrix-variate normal distributions have been recently introduced in the literature. However, when data contains groups of observations with longer-than-normal tails or atypical observations, the use of the matrix-variate normal distribution for the mixture components may affect the fitting of the resulting model. Therefore, we consider a more robust approach based on the matrix-variate *t* distribution for modeling the mixture components. To introduce parsimony, we use the eigen-decomposition of the components scale matrices and we allow the degrees of freedom to be equal across groups. This produces a family of 196 parsimonious matrix-variate *t* mixture models. Parameter estimation is obtained by using an AECM algorithm. The use of our parsimonious models is illustrated via a real data application, where parsimonious matrix-variate normal mixtures are also fitted for comparison purposes.

Keywords: matrix-variate, mixture models, clustering, parsimonious models

1 Introduction

The matrix-variate model-based clustering literature is expanding more and more over the last few years, as confirmed by the high number of contributions using finite mixture models for the modelization of matrix-variate data [1, 2, 3, 4, 5, 6, 7, 8]. This kind of data is arranged in three-dimensional arrays, and depending on the entities indexed in each of the three layers, different data examples might be considered [9]. In many of these applications, we observe a $p \times r$ matrix for each statistical

Salvatore D. Tomarchio (✉)
University of Catania, Department of Economics and Business, Catania, Italy,
e-mail: daniele.tomarchio@unict.it

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_42

observation. Thus, from a model-based clustering perspective, the challenge is to suitably cluster realization coming from random matrices.

One problem of matrix-variate mixture models is the potentially high number of parameters. To cope with this issue, [5] have recently proposed a family of parsimonious mixtures based on the matrix-variate normal (MVN) distribution. Nevertheless, for many datasets, the tails of the MVN distribution are often shorter than required. This has several consequences on parameter estimation as well as in the proper data classification [4, 7]. Therefore, in this paper we relax the normality assumption of the mixture components by using (in a parsimonious setting) the matrix-variate *t* (MVT) distribution. The MVT distribution has been used within the finite mixture model paradigm by [10] in an unconstrained framework. Here, to introduce parsimony in this model, (i) we use the eigen-decomposition of the two scale matrices of each mixture component and (ii) we allow the degrees of freedom to be tied across the groups. This produces the family of 196 parsimonious matrix-variate MVT mixture models (MVT-Ms) discussed in Section 2. Parameter estimation is implemented by using an alternating expectation-conditional maximization (AECM) algorithm [12]. In Section 3, our parsimonious MVT-Ms, along with parsimonious matrix-variate MVN mixture models (MVN-Ms) for comparison purposes, are fitted to a Swedish municipalities expenditure dataset. The differences in terms of fitting among the two families of models are illustrated. The estimated parameters and the data partition of the overall best fitting model are also commented. Finally, some conclusions are drawn in Section 4.

2 Methodology

2.1 Parsimonious Mixtures of Matrix-variate *t* Distributions

The probability distribution function (pdf) of a $p \times r$ random matrix $\mathcal{X}$ coming from a finite mixture model is

$$f_{\text{MIXT}}(\mathbf{X}; \mathbf{\Omega}) = \sum_{g=1}^{G} \pi_g f(\mathbf{X}; \mathbf{\Theta}_g), \tag{1}$$

where π_g is the gth mixing proportion, such that $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$, $f(\mathbf{X}; \mathbf{\Theta}_g)$ is the gth component pdf with parameter $\mathbf{\Theta}_g$, and $\mathbf{\Omega}$ contains all of the parameters of the mixture. In this paper, for the gth component of model (1), we adopt the MVT distribution having pdf

$$f_{\text{MVT}}(\mathbf{X}; \mathbf{\Theta}_g) = \frac{|\mathbf{\Sigma}_g|^{-\frac{r}{2}} |\mathbf{\Psi}_g|^{-\frac{p}{2}} \Gamma\left(\frac{pr+\nu_g}{2}\right)}{\left(\pi \nu_g\right)^{\frac{pr}{2}} \Gamma\left(\frac{\nu_g}{2}\right)} \left[1 + \frac{\delta_g\left(\mathbf{X}; \mathbf{M}_g, \mathbf{\Sigma}_g, \mathbf{\Psi}_g\right)}{\nu_g}\right]^{-\frac{pr+\nu_g}{2}}, \tag{2}$$

where $\delta_g\left(\mathbf{X};\mathbf{M}_g,\mathbf{\Sigma}_g,\mathbf{\Psi}_g\right) = \text{tr}\left[\mathbf{\Sigma}_g^{-1}(\mathbf{X}-\mathbf{M}_g)\mathbf{\Psi}_g^{-1}(\mathbf{X}-\mathbf{M}_g)'\right]$, $\mathbf{M}_g$ is the $p \times r$ component mean matrix, $\mathbf{\Sigma}_g$ is the $p \times p$ component row scale matrix, $\mathbf{\Psi}_g$ is the $r \times r$ component column scale matrix and $\nu_g > 0$ is the component degree of freedom. It is interesting to recall that the pdf in (2) can be hierarchically obtained via the matrix-variate normal scale mixture model when the mixing random variable W is a gamma distribution with scale and rate parameters set to $\nu_g/2$ [10]. Specifically, a hierarchical representation of MVT distribution can be given as follows

1. $W \sim \mathcal{G}\left(\nu_g/2, \nu_g/2\right)$,
2. $\mathbf{X}|W = w \sim \mathcal{N}(\mathbf{M}_g, \mathbf{\Sigma}_g/w, \mathbf{\Psi}_g)$,

where $\mathcal{G}(\cdot)$ is a gamma distribution and $\mathcal{N}(\cdot)$ denotes the MVN distribution. This representation will be convenient for parameter estimation presented in Section 2.2.

As discussed in Section 1, the mixture model in (1) may be characterized by a potentially high number of parameters. To address this concern, we firstly use the eigen-decomposition of the components scale matrices $\mathbf{\Sigma}_g$ and $\mathbf{\Psi}_g$. In detail, we recall that a generic $q \times q$ scale matrix $\mathbf{\Phi}_g$ can be decomposed as [11]

$$\mathbf{\Phi}_g = \lambda_g \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g', \tag{3}$$

where $\lambda_g = |\mathbf{\Phi}_g|^{1/q}$, $\mathbf{\Gamma}_g$ is a $q \times q$ orthogonal matrix whose columns are the normalized eigenvectors of $\mathbf{\Phi}_g$, and $\mathbf{\Delta}_g$ is the scaled ($|\mathbf{\Delta}_g| = 1$) diagonal matrix of the eigenvalues of $\mathbf{\Phi}_g$. By constraining the three components in (3), the following family of 14 parsimonious structures is obtained: EII, VII, EEI, VEI, EVI, VVI, EEE, VEE, EVE, VVE, EEV, VEV, EVV, VVV, where “E” stands for equal, “V” means varying and “I” denotes the identity matrix.

If we apply the decomposition in (3) to $\mathbf{\Sigma}_g$ and $\mathbf{\Psi}_g$, we obtain $14 \times 14 = 196$ parsimonious structures. However, to solve a well-known identifiability issue related to the scale matrices of matrix-variate distributions [1, 3, 5], we impose the restriction $|\mathbf{\Psi}_g| = 1$, which makes the parameter λ_g unnecessary, and reduces the number of parsimonious structures related to $\mathbf{\Psi}_g$ from 14 to 7: II, EI, VI, EE, VE, EV, VV. Thus, we have $14 \times 7 = 98$ parsimonious structures for the component scale matrices.

To further increase the parsimony of model (1), we also consider the option of constraining the component degrees of freedom ν_g. The nomenclature used is the same to that adopted for the scale matrices. This option, combined with that discussed above for the scale matrices, allows us to produce a total of $98 \times 2 = 196$ parsimonious MVT-Ms.

2.2 An AECM Algorithm for Parameter Estimation

To estimate the parameters of our family of mixture models, we implement an AECM algorithm. By using the hierarchical representation of Section 2.1, our complete data are $\mathbf{S}_c = \{\mathbf{X}_i, \mathbf{z}_i, w_i\}_{i=1}^N$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})'$, such that $z_{ig} = 1$ if observation i belongs to group g and $z_{ig} = 0$ otherwise, and w_i is the realization of W. Therefore, the complete-data log-likelihood can be written as

$$\ell_c\left(\boldsymbol{\Omega};\mathbf{S}_c\right) = \ell_{1c}\left(\boldsymbol{\pi};\mathbf{S}_c\right) + \ell_{2c}\left(\boldsymbol{\Xi};\mathbf{S}_c\right) + \ell_{3c}\left(\boldsymbol{\vartheta};\mathbf{S}_c\right), \tag{4}$$

where

$$\ell_{1c}\left(\boldsymbol{\pi};\mathbf{S}_c\right) = \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\ln\left(\pi_g\right),$$

$$\begin{aligned}\ell_{2c}\left(\boldsymbol{\Xi};\mathbf{S}_c\right) = \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\Bigg[&-\frac{pr}{2}\ln(2\pi) + \frac{pr}{2}\ln\left(w_{ig}\right) - \frac{r}{2}\ln|\boldsymbol{\Sigma}_g| - \frac{p}{2}\ln|\boldsymbol{\Psi}_g| \\ &-\frac{w_{ig}\delta_g(\mathbf{X};\mathbf{M}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\Psi}_g)}{2}\Bigg],\end{aligned} \tag{5}$$

$$\ell_{3c}\left(\boldsymbol{\vartheta};\mathbf{S}_c\right) = \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\left\{\frac{\nu_g}{2}\ln\left(\frac{\nu_g}{2}\right) - \ln\left[\Gamma\left(\frac{\nu_g}{2}\right)\right] + \left(\frac{\nu_g}{2}-1\right)\ln\left(w_{ig}\right) - \frac{\nu_g}{2}w_{ig}\right\},$$

with $\boldsymbol{\pi} = \left\{\pi_g\right\}_{g=1}^{G}$, $\boldsymbol{\Xi} = \left\{\mathbf{M}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\Psi}_g\right\}_{g=1}^{G}$ and $\boldsymbol{\vartheta} = \left\{\nu_g\right\}_{g=1}^{G}$.

Our AECM algorithm then proceeds as follows (notice that, the parameters marked with one dot are the updates of the previous iteration, while those marked with two dots are the updates at the current iteration):

E-step At the E-step we have to compute the following quantities

$$\ddot{z}_{ig} = \frac{\dot{\pi}_g f_{\mathrm{MVT}}\left(\mathbf{X}_i;\dot{\boldsymbol{\Theta}}_g\right)}{\sum_{h=1}^{G}\dot{\pi}_h f_{\mathrm{MVT}}\left(\mathbf{X}_i;\dot{\boldsymbol{\Theta}}_h\right)} \quad \text{and} \quad \ddot{w}_{ig} = \frac{pr+\dot{\nu}_g}{\dot{\nu}_g + \dot{\delta}_g\left(\mathbf{X}_i;\dot{\mathbf{M}}_g,\dot{\boldsymbol{\Sigma}}_g,\dot{\boldsymbol{\Psi}}_g\right)}. \tag{6}$$

There is no need to compute the expected value of $\ln\left(W_{ig}\right)$, given that we do not use this quantity to update ν_g.

CM-step 1 At the first CM-step, we have the following updates

$$\ddot{\pi}_g = \frac{\sum_{i=1}^{N}\ddot{z}_{ig}}{N} \quad \text{and} \quad \ddot{\mathbf{M}}_g = \frac{\sum_{i=1}^{N}\ddot{z}_{ig}\ddot{w}_{ig}\mathbf{X}_i}{\sum_{i=1}^{N}\ddot{z}_{ig}\ddot{w}_{ig}}.$$

Because of space constraints, we cannot report here the updates of each parsimonious structure related to $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\Psi}_g$. However, they can be obtained by generalizing the results in [5]. The only differences consist in the updates of the row and column scatter matrices of the gth component, that are here defined as

$$\ddot{\mathbf{W}}_g^R = \sum_{i=1}^{N}\ddot{z}_{ig}\ddot{w}_{ig}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_g\right)\dot{\boldsymbol{\Psi}}_g^{-1}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_g\right)',$$

$$\ddot{\mathbf{W}}_g^C = \sum_{i=1}^{N}\ddot{z}_{ig}\ddot{w}_{ig}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_g\right)'\ddot{\boldsymbol{\Sigma}}_g^{-1}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_g\right).$$

CM-step 2 At the second CM-step, we firstly define the "partial" complete-data log-likelihood function according to the following specification

$$\ell_{pc}\left(\boldsymbol{\Omega}; \mathbf{S}_{pc}\right) = \ell_{1c}\left(\boldsymbol{\pi}; \mathbf{S}_{pc}\right) + \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig} \ln f_{\text{MVT}}(\mathbf{X}_i; \boldsymbol{\Theta}_g), \tag{7}$$

where "partial" refers to fact that the complete data are now defined as $\mathbf{S}_{pc} = \{\mathbf{X}_i, \mathbf{z}_i\}_{i=1}^{N}$. Then, $\ddot{\nu}_g$ is determined by maximizing

$$\sum_{i=1}^{N} \ddot{z}_{ig} \ln f_{\text{MVT}}(\mathbf{X}_i; \ddot{\boldsymbol{\Theta}}_g) \quad \text{or} \quad \sum_{i=1}^{N}\sum_{g=1}^{G} \ddot{z}_{ig} \ln f_{\text{MVT}}(\mathbf{X}_i; \ddot{\boldsymbol{\Theta}}_g),$$

over $\nu_g \in (0, 100)$, depending on the parsimonious structure selected, i.e. V or E, respectively. Notice that, an higher upper bound could also have been selected for the maximization problem but, with the already chosen value, the differences between an estimated MVT distribution and the nested MVN distribution would be negligible. Furthermore, when a heavy-tailed distribution approaches to normality, the precision of the estimated tailedness parameters is unreliable [4].

3 Real Data Application

Here, we analyze the `Municipalities` dataset contained in the **AER** package [13] for the R statistical software. It consists of expenditure information for $N = 265$ Swedish municipalities over $r = 9$ years (1979–1987). For each municipality, we measure the following $p = 3$ variables: (i) total expenditures, (ii) total own-source revenues and (iii) intergovernmental grants received.

We fitted parsimonious MVT-Ms and MVN-Ms for $G \in \{1, 2, 3, 4, 5\}$ to the data, and for each family of models the Bayesian information criterion (BIC) [14] is used to select the best fitting model. According to our results, we found that the best among MVN-Ms has a BIC of -82362.61, a VVV-EE structure and $G = 4$ groups, while the best among MVT-Ms has a BIC of -82701.59, a VVE-EE-V structure and $G = 3$ groups. Thus, the overall best fitting model is that selected for MVT-Ms. The MVN-Ms seem to overfit the data, given that an additional group is detected. This is not an unusual behavior, given that the tails of normal mixture models cannot adequately accommodate deviations from normality, and additional groups are consequently found in the data [4, 7, 15]. Anyway, the best fitting models of the two families agree in finding varying volumes and shapes in the components row scale matrices and equal shapes and orientations in the components column scale matrices.

Figure 1 illustrates the parallel coordinate plots of the data partition detected by the VVE-EE-V MVT-Ms. The dashed lines correspond to the estimated mean for that variable, across the time, in that group. We notice that the first group contains municipalities having, on average, slightly higher expenditures, an intermediate

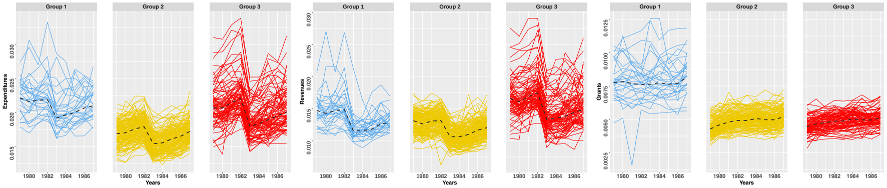

Fig. 1 Parallel coordinate plots of the data partition obtained by the VVE-EE-V MVT-Ms. The dashed lines correspond to the estimated means.

level of revenues and higher levels of intergovernmental grants than the other two groups. Furthermore, it seems to cluster several outlying observations, as confirmed by the estimated degree of freedom $\nu_1 = 3.75$, which implies quite heavy tails for this mixture component. The second group shows the lowest average levels of expenditures and revenues, but a similar amount of received grants to that of the third group. Interestingly, this group does not presents many outlying observations, as also supported by the estimated degree of freedom $\nu_2 = 10.95$. Lastly, the third group has the highest levels of revenues but, as already said, it is similar to the other two groups in the other variables. Also in this case, we have a moderately heavy tail behavior given that the estimated degree of freedom is $\nu_3 = 6.05$.

To evaluate the correlations of the variables with each other and over time, for the three groups, we now report the correlation matrices $\mathbf{R}_{(\cdot)}$ related to the covariance matrices associated to $\mathbf{\Sigma}_g$ and $\mathbf{\Psi}_g$:

$$\mathbf{R}_{\mathbf{\Sigma}_1} = \begin{bmatrix} 1.00 & 0.48 & 0.14 \\ 0.48 & 1.00 & -0.06 \\ 0.14 & -0.06 & 1.00 \end{bmatrix}, \mathbf{R}_{\mathbf{\Sigma}_2} = \begin{bmatrix} 1.00 & 0.55 & 0.18 \\ 0.55 & 1.00 & -0.07 \\ 0.18 & -0.07 & 1.00 \end{bmatrix}, \mathbf{R}_{\mathbf{\Sigma}_3} = \begin{bmatrix} 1.00 & 0.73 & 0.22 \\ 0.73 & 1.00 & -0.02 \\ 0.22 & -0.02 & 1.00 \end{bmatrix},$$

$$\mathbf{R}_{\mathbf{\Psi}_1} = \mathbf{R}_{\mathbf{\Psi}_2} = \mathbf{R}_{\mathbf{\Psi}_3} = \begin{bmatrix} 1.00 & 0.80 & 0.72 & 0.67 & 0.65 & 0.59 & 0.58 & 0.55 & 0.52 \\ 0.80 & 1.00 & 0.79 & 0.73 & 0.69 & 0.62 & 0.62 & 0.57 & 0.54 \\ 0.72 & 0.79 & 1.00 & 0.80 & 0.73 & 0.69 & 0.66 & 0.63 & 0.60 \\ 0.67 & 0.73 & 0.80 & 1.00 & 0.79 & 0.73 & 0.71 & 0.67 & 0.64 \\ 0.65 & 0.69 & 0.73 & 0.79 & 1.00 & 0.83 & 0.80 & 0.73 & 0.71 \\ 0.59 & 0.62 & 0.69 & 0.73 & 0.83 & 1.00 & 0.80 & 0.76 & 0.73 \\ 0.58 & 0.62 & 0.66 & 0.71 & 0.80 & 0.80 & 1.00 & 0.81 & 0.78 \\ 0.55 & 0.57 & 0.63 & 0.67 & 0.73 & 0.76 & 0.81 & 1.00 & 0.79 \\ 0.52 & 0.54 & 0.60 & 0.64 & 0.71 & 0.73 & 0.78 & 0.79 & 1.00 \end{bmatrix}.$$

When $\mathbf{R}_{\mathbf{\Sigma}_1}$, $\mathbf{R}_{\mathbf{\Sigma}_2}$ and $\mathbf{R}_{\mathbf{\Sigma}_3}$ are considered, we notice that, as it might be reasonable to expect, the correlations between total-expenditures and total-own source revenues or intergovernmental grants received are positive, despite they increase as we move from the first to the third group. Conversely, there exists a slightly negative correlation between total-own source revenues and intergovernmental grants received. However, there would be no great differences among the groups in this case. As concerns $\mathbf{R}_{\mathbf{\Psi}_1}$, $\mathbf{R}_{\mathbf{\Psi}_2}$ and $\mathbf{R}_{\mathbf{\Psi}_3}$, we observe that the correlation among the columns, i.e. between time points, decreases as the temporal distance increases. Furthermore, considering the dimensionality of these column matrices, it is readily understandable the benefit, in terms of number of parameters to be estimated, of an EE parsimonious structure with respect to a fully unconstrained model.

Finally, we analyze the uncertainty of the detected classification. This can be computed, for each observation, by subtracting the probability z_{ig} of the most likely group from 1 [16]. The lower the uncertainty is, the stronger the assignment becomes. The quantiles of the obtained uncertainties can be used to measure the quality of the classification. In this regard, we noticed that 75% of the observations have an uncertainty equal or lower than 0.05. However, we observed a maximum value of 0.50. This happens when groups intersect, since uncertain classifications are expected in the overlapping regions [17]. Relatedly, a more detailed information can be gained by looking at the uncertainty plot illustrated in Figure 2, which reports the (sorted) uncertainty values of all the municipalities. We see that the municipalities clustered in the first group, excluding a couple of cases, have practically null uncertainties. This applies to a lesser extent to the municipalities in the other two groups, given the slightly higher number of exceptions. For example, there are 15 observations (approximately 5% of the total sample size) that have uncertainty values greater than 0.3. However, and as said above, this is due to the closeness between the groups, which can be confirmed by looking at the parallel plots in Figure 1.

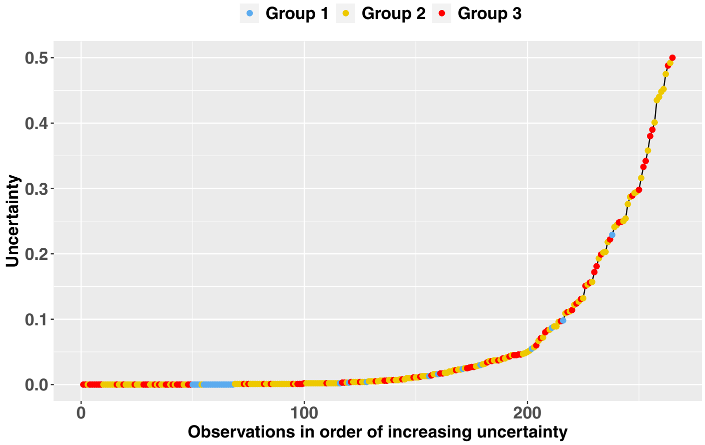

Fig. 2 Uncertainty plot for the `Municipalities` dataset.

4 Conclusions

One serious concern of matrix-variate mixture models is the potentially high number of parameters. Furthermore, many real data requires models having heavier-than-normal tails. To address both aspects, in this paper a family of 196 parsimonious mixture models, based on the matrix-variate *t* distribution, is introduced. The eigen-decomposition of the components scale matrices, as well as constraints on the components degrees of freedom, are used to attain parsimony. An AECM algorithm for parameter estimation has been presented. Our family of models have been fitted to a real dataset along with parsimonious mixtures of matrix-variate normal distributions. The results demonstrate the best fitting results of our models, and the overfitting tendency of matrix-variate normal mixtures. Lastly, the estimated parameters and data partition for the best of our models have been reported and commented.

Acknowledgements This work was supported by the University of Catania grant PIACERI/CRASI (2020).

References

1. Gallaugher, M. P. B., McNicholas P. D.: Finite mixtures of skewed matrix variate distributions. Pattern Recognit. **80**, 83–93 (2018)
2. Melnykov, V., Zhu, X.: On model-based clustering of skewed matrix data. J. Multivar. Anal. **167**, 181–194 (2018)
3. Melnykov, V., Zhu, X.: Studying crime trends in the USA over the years 2000–2012. Adv. Data Anal. Classif. **13**(1), 325–341 (2019)
4. 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)
5. Sarkar, S., Zhu, X., Melnykov, V., Ingrassia, S.: On parsimonious models for modeling matrix data. Comput. Stat. Data Anal. **142**, 106822 (2020)
6. Tomarchio, S. D., McNicholas, P. D., Punzo, A.: Matrix normal cluster-weighted models. J. Classif. **38**(3), 556–575 (2021)
7. Tomarchio, S. D., Gallaugher, M. P. B., Punzo, A., McNicholas, P. D.: Mixtures of matrix-variate contaminated normal distributions. J. Comput. Gr. Stat. 1–9 (2022)
8. Tomarchio, S. D., Ingrassia, S., Melnykov, V.: Modelling students' career indicators via mixtures of parsimonious matrix-normal distributions. Aust. N. Z. J. Stat. 1–16 (2022)
9. Viroli, C.: Model based clustering for three-way data structures. Bayesian Anal. **6**(4), 573–602 (2011)
10. Doğru, F. Z., Bulut, Y. M., Arslan, O.: Finite mixtures of matrix variate t distributions. Gazi Univ. J. Sci. **29**(2), 335–341 (2016)
11. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**(5), 781–793 (1995)
12. Meng, X. L., Van Dyk, D.: The EM algorithm-an old folk-song sung to a fast new tune. J. Royal Stat. Soc. B. **59**(3), 511–567 (1997)
13. Kleiber, C., Zeileis, A.: Applied Econometrics with R. Springer-Verlag, New York (2008)
14. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
15. Gallaugher, M. P. B., Tomarchio, S. D., McNicholas, P. D., Punzo, A.: Multivariate cluster weighted models using skewed distributions. Adv. Data Anal. Classif. 1–32 (2021)
16. Fraley, C., Raftery, A. E.: Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. J. Classif., **20**(2), 263–286 (2003)
17. Tomarchio, S. D., Punzo, A.: Dichotomous unimodal compound models: application to the distribution of insurance losses. J. Appl. Stat. **47**(13-15), 2328–2353 (2020)

Evolution of Media Coverage on Climate Change and Environmental Awareness: an Analysis of Tweets from UK and US Newspapers

Gianpaolo Zammarchi, Maurizio Romano, and Claudio Conversano

Abstract Climate change represents one of the biggest challenges of our time. Newspapers might play an important role in raising awareness on this problem and its consequences. We collected all tweets posted by six UK and US newspapers in the last decade to assess whether 1) the space given to this topic has grown, 2) any breakpoint can be identified in the time series of tweets on climate change, and 3) any main topic can be identified in these tweets. Overall, the number of tweets posted on climate change increased for all newspapers during the last decade. Although a sharp decrease in 2020 was observed due to the pandemic, for most newspapers climate change coverage started to rise again in 2021. While different breakpoints were observed, for most newspapers 2019 was identified as a key year, which is plausible based on the coverage received by activities organized by the Fridays for Future movement. Finally, using different topic modeling approaches, we observed that, while unsupervised models partly capture relevant topics for climate change, such as the ones related to politics, consequences for health or pollution, semi-supervised models might be of help to reach higher informativeness of words assigned to the topics.

Keywords: climate change, Twitter, environment, time series, topic modeling

Gianpaolo Zammarchi (✉)
University of Cagliari, Viale Sant'Ignazio 17, 09123, Cagliari, Italy,
e-mail: gp.zammarchi@unica.it

Maurizio Romano
University of Cagliari, Viale Sant'Ignazio 17, 09123, Cagliari, Italy,
e-mail: romano.maurizio@unica.it

Claudio Conversano
University of Cagliari, Viale Sant'Ignazio 17, 09123, Cagliari, Italy, e-mail: conversa@unica.it

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_43

1 Introduction

Climate change is one of the biggest challenges for our society. Its consequences which include, among others, glaciers melting, warming oceans, rising sea levels, and shifting weather or rainfall patterns, are already impacting our health and imposing costs on society. Without drastic action aimed at reducing or preventing human-induced emissions of greenhouse gasses, these consequences are expected to intensify in the next years. Despite its global and severe impacts, individuals may perceive climate change as an abstract problem [1]. It is also a well-known fact that the level of information plays a crucial role in the awareness about a topic (e.g. healthy food [2] and smoking [3]) . Media represent a crucial source of information and can exert substantial effects on public opinion, thus helping to raise the awareness on climate change. For instance, media can explain climate change consequences as well as portraying actions that governments, communities and single individuals can take. For this reason, it is important to distinguish themes that might have gained popularity from those that may have seen a decrease of interest. Nowadays, social media have become a reliable and popular source of information for people from all around the world. Twitter is one of the most popular microblogging services and is used by many traditional newspapers on a daily basis. While we can hypothesize that in the last few years the media coverage on climate change might have risen, due for instance to international climate strike movements, the recent emergence of the coronavirus disease 2019 (COVID-19) pandemic might have led to a decrease of attention on other relevant topics.

Aims of this work were to: (1) assess trends in media coverage on climate change using tweets posted by main international newspapers based in United Kingdom (UK) and United States (US), and (2) identify the main topics discussed in these tweets using topic modeling.

2 Dataset and Methods

We downloaded all tweets posted from 2012 January 1^{st} to 2021 December 31^{st} from the official Twitter account of six widely known newspapers based in UK (The Guardian, The Independent and The Mirror) or US (The New York Times, The Washington Post and The Wall Street Journal) leading to a collection of 3,275,499 tweets. Next, we determined which tweets were related to climate change and environmental awareness based on the presence of at least one of the following keywords: “climate change”, “sustainability”, “earth day”, “plastic free”, “global warming”, “pollution”, “environmentally friendly” or “renewable energy”. We plotted the number of tweets on climate change posted by each newspaper during each year using R v. 4.1.2 [4].

We analyzed the association between the number of tweets on climate change and the whole number of tweets posted by each newspaper using Spearman’s correlation analysis. For each year and for each newspaper, we computed and plotted the differences in the number of posted tweets compared to the previous year, for either (a) tweets related to climate change and (b) all tweets. Finally, we used the changepoint R package [5] to conduct an analysis aimed at identifying structural breaks, i.e. unexpected changes in a time series. In many applications, it is reasonable to believe that there might be *m* breakpoints (especially if some exogenous event occurs) in which a shift in mean value is observed. The changepoint package estimates the breakpoints using several penalty criteria such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). We estimated the breakpoints using the Binary Segmentation (BinSeg) method [6] implemented in the package.

Lastly, we used tweets posted by The Guardian to perform topic modeling, a method for classification of text into topics. Preprocessing (including lemmatization, removal of stopwords and creation of the document term matrix) was conducted with tm [7] and quanteda [8] in R. We used two different approaches: 1) Latent Dirichlet Allocation (LDA) implemented in the textmineR R package [9]; and 2) Correlation Explanation (CorEx), an approach alternative to LDA that allows both unsupervised as well as semi-supervised topic modeling [10].

3 Results

3.1 Analysis of Tweet Trends and Breakpoints

Among 3,275,499 collected tweets, we identified 11,155 tweets related to climate change and environmental awareness. Figure 1A shows the number of tweets on climate change posted by each of the analyzed newspapers from 2012 to 2021, while Figure 1B the total number of tweets posted by each newspaper.

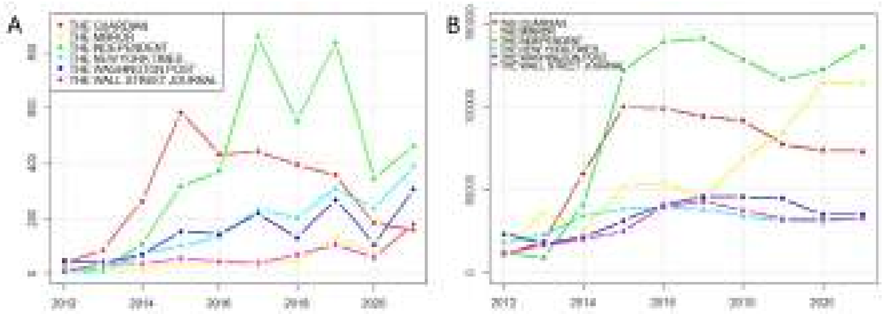

Fig. 1 Number of tweets on climate change (A) or total number of tweets (B) posted by the six newspapers from 2012 to 2021.

For the majority of newspapers, the number of tweets on climate change increased from 2014 to 2019, saw a sharp decrease in 2020, in correspondence of the emergence of the COVID-19 pandemic, and a subsequent rise in 2021. On the other hand, the

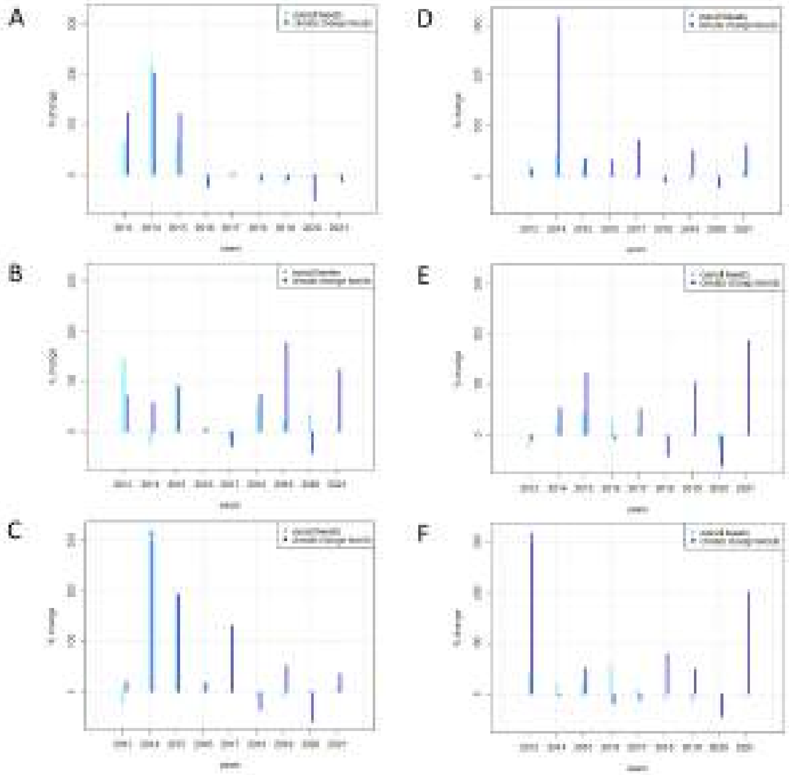

Fig. 2 Year-over-year percentage changes of overall tweets and tweets on climate change. A: The Guardian, B: The Mirror, C: The Independent, D: The New York Times, E: The Washington Post, F, The Wall Street Journal.

number of tweets on climate change posted by The Guardian showed a peak during 2015 and a subsequent decrease. However, it must be noted that The Guardian is also the newspaper that showed a more pronounced decrease in the overall number of tweets.

The number of tweets on climate change was significantly positively correlated with the overall number of tweets posted from 2012 to 2021 for four newspapers (The Guardian, Spearman's rho = 0.95, p < 0.001; The Mirror, Spearman's rho = 0.95, p < 0.001; The Independent, Spearman's rho = 0.76, p = 0.016; The Washington Post, Spearman's rho = 0.70, p = 0.031) but not for The New York Times (Spearman's rho = 0.18, p = 0.63) or The Wall Street Journal (Spearman's rho = 0.49, p = 0.15). Year-over-year percentage changes among either tweets related to climate change or all posted tweets can be observed in Figure 2.

Looking at Figure 2, we can observe a great variability in the posted number of tweets during the years, both for the total number of tweets and for the number of tweets on climate change. While the analysis aimed at identifying structural changes

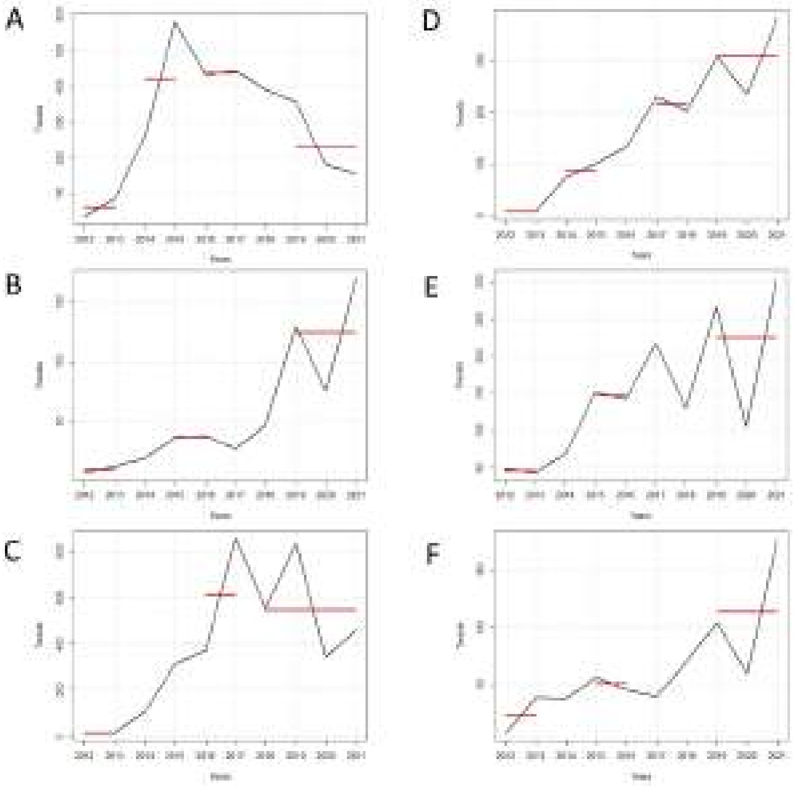

Fig. 3 Structural changes in the time series of tweets related to climate change. A: The Guardian, B: The Mirror, C: The Independent, D: The New York Times, E: The Washington Post, F, The Wall Street Journal. The red line represents the years between two breakpoints.

in the time series comprising tweets on climate change identified three or four breakpoints for all newspapers, wide variability was observed regarding the specific year in which these structural changes were identified (Figure 3). Despite the great variability, Figure 3 shows that even if a common breakpoint cannot be identified, 2019 was a key year for five out of six newspapers (except for The Independent).

3.2 Topic Modeling

Finally, we exploited the topic modeling approach to identify and analyze the main topics discussed by newspapers in their tweets. Due to space limitations, we focus only on The Guardian since this newspaper showed a trend in contrast with the others. Data comes from 2,916 tweets posted by The Guardian analyzed using LDA and CorEx. For LDA, a range of 5-20 unsupervised topics was tested, with the most interpretable results obtained with 10 topics (Table 1). The topic coherence ranged from 0.01 to 0.34 (mean: 0.13). For each topic, bi-gram topic labels were assigned with the labeling algorithm implemented in textmineR. We can observe that topics are related to politics or leaders (Topics 3, 7 and 10), environmental scientists or climate journalists (Topics 1 and 5), energy sources (Topics 4 and 8) and effects of climate change (Topics 2, 6 and 9). The intertopic distance map obtained with LDAvis is shown in Figure 4. The area of each circle is proportional to the relative prevalence of that topic in the corpus, while inter-topic distances are computed based on Jensen-Shannon divergence.

Table 1 Top terms for the ten topics identified with LDA.

dana_nuccitelli	air_pollution	barack_obama	renewable_energy	john_abraham
dana	pollution	fight	energy	john
dana_ nuccitelli	air	obama	renewable	trump
nuccitelli	air_pollution	trump	renewable_energy	australia
live	study	plan	uk	tackle
trump	finds	battle	sustainability	abraham
air_pollution	**donald_trump**	**fossil_fuel**	**extreme_weather**	**pope_francis**
pollution	trump	report	world	pollution
air	schoolstrike	fossil	paris	study
air_pollution	school	ipcc	leaders	tackling
uk	great	warns	talks	pope
tackle	donald	stop	deal	scientists

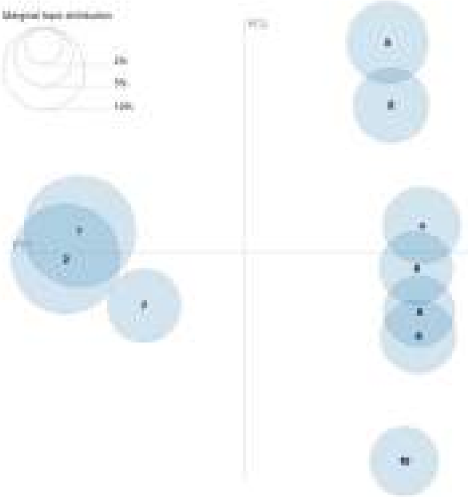

Fig. 4 Intertopic distance map.

Finally, we conducted a semi-supervised topic modeling analysis based on anchored words using CorEx. When anchoring a word to a topic, CorEx maximizes the mutual information between that word and the topic, thus guiding the topic model towards specific subsets of words. A model with 5 topics and three anchored words for each topic (Table 2) showed a total correlation (i.e. the measure maximized by CorEx when constructing the topic model) of 4.36. This value was higher compared to the one observed with an unsupervised CorEx analysis with the same number of topics (total correlation = 0.97, topics not shown due to space limits). Topics related to politics (Topic 3) and science (Topic 5) were found to be the most informative in our dataset based on the total correlation metric.

Table 2 Topics with anchored words and examples of tweets.

Topic	Topic words	Examples of tweets per topic
1	**school, strike, march,** schoolstrik, climatestrikeuk, ukschoolstrik, schoolstrikeclim, climatemarch, arabia, saudi	EPA wipes its climate change site day before march on Washington
2	**ocean, ice, environment,** john, dana, nuccitelli, air, abraham, sea, reed	Chasing Ice filmmakers plumb the 'bottomless' depths of climate change - new clip from @GuardianEco
3	**trump, obama, lead,** donald, barack, ivanka, brighton, repli, administr, pick	Trump administration pollution rule strikes final blow against environment
4	**plastic, fuel, oil,** fossil, compani, pictur, wast, big, bay, photo	Engaging with oil companies on climate change is futile
5	**studi, scientist, research,** find, link, say, show, death, prematur, speci	Microplastic pollution revealed 'absolutely everywhere' by new research

The anchored words are reported in bold.

4 Discussion

The present study aims to evaluate how some of the most relevant British and American newspapers have given space to the topic of climate change on their Twitter page in the last decade. Apart from The Guardian, which shows a decreasing trend in the number of tweets related to climate change, all the other newspapers showed an overall growing trend, except during 2020. During this year, the number of tweets related to climate change declined for all six newspapers. This was most probably due to the COVID-19 outbreak that was massively covered by all media. By analyzing the breakpoints in Figure 3, it is possible to observe that 2019 was a relevant year for climate change. This is plausible considering that, starting from the end of 2018, the strikes launched by the Fridays for Future movement to raise awareness on the issue of climate change, gained high media coverage.

Our topic modeling analysis showed that the main topics defined using unsupervised models such as LDA are mostly related to politics, environmental scientists, energy sources and effects of climate change. While unsupervised models capture relevant topics, using CorEx we found a semi-supervised model to be able to reach a higher total correlation, which is a measure of informativeness of the topics, compared to an unsupervised model with the same number of topics.

As future developments, we plan to extend our analyses to newspapers from other countries. We believe our work to be useful to gain more knowledge and awareness about the climate change topic and on how much space relevant newspapers have given to this issue on social media. Increasing the knowledge about the nature of the topics covered by newspapers will lay the basis for future studies aimed at evaluating public awareness on this highly relevant challenge.

References

1. Van Lange, P. A. M., Huckelba, A. L.: Psychological distance: How to make climate change less abstract and closer to the self. Curr. Opin. Psychol. **42**, 49–53 (2021)
2. Wakefield, M., Flay B., Nichter M., Giovino G.: Role of the media in influencing trajectories of youth smoking. Addiction, **98**, 79-103 (2003)
3. Dumanovsky, T., Huang, C. Y., Bassett, M. T., Silver, L. D.: Consumer awareness of fast-food calorie information in New York City after implementation of a menu labeling regulation. American Journal of Public Health, **12**, 2520-2525 (2010)
4. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2020). Available via `http://www.R-project.org`
5. Killick, R., Eckley, I. A.: changepoint: An R Package for changepoint analysis. J. Stat. Softw. **58**, 1–19 (2014)
6. Scott, A.J., Knott, M.: A cluster analysis method for grouping means in the analysis of variance, Biometrics, **30**, 507-512 (1974)
7. Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in R. J. Stat. Softw. **25**, 1–54 (2008)
8. Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A.: quanteda: An R package for the quantitative analysis of textual data. J. Open Source Softw. **3**, 774 (2018)
9. Jones, T., Doane, W.: Package ‘textmineR’. Functions for Text Mining and Topic Modeling (2021). Retreived from `https://cran.r-project.org/web/packages/textmineR/textmineR.pdf`
10. Gallagher, R., Reing, K., Kale, D., Ver Steeg, G.: Anchored correlation explanation: Topic modeling with minimal domain knowledge. Trans. Assoc. Comput. **5**, 529-542 (2017)

Index

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9