Book of Short Papers of the 5th international workshop on

Models and Learning for Clustering and Classification

MBC^2 2020, Catania, Italy

Salvatore Ingrassia, Antonio Punzo, Roberto Rocci
(editors)

Ledizioni
The Innovative LEDIpublishing Company

Book of Short Papers of the 5th international workshop on

Models and Learning for Clustering and Classification

MBC^2 2020, Catania, Italy

Salvatore Ingrassia, **Antonio Punzo, Roberto Rocci**
(editors)

LEDIZIONI

Book of Short Papers of the 5th international workshop on Models and Learning for Clustering and Classification (MBC2 2020, Catania, Italy), Salvatore Ingrassia, Antonio Punzo, Roberto Rocci (editors)

Sito workshop
mbc2.unict.it

Ledizioni: settembre 2021

ISBN: 9788855265393

Ledizioni – LEDIpublishing
Via Antonio Boselli 10 – 20136
Milano, Italia

www.ledizioni.it

Indice

Impact of the COVID-19 pandemic on music: a method for clustering sentiments

Alessandro Albano, Mariangela Sciandra, Antonella Plaia, Irene Carola Spera

Abstract The outbreak of coronavirus disease 2019 (COVID-19) was highly stressful for people. In general, fear and anxiety about a disease can be overwhelming and cause strong emotions in adults and children. One way to cope with this stress consists in listening to music. Aim of this work is to understand if the music heard during the lock-down reflects the emotions generated by the pandemic on each of us. So, the primary goal of this work is to build two indices for measuring the anger and joy levels of the top streamed songs by Italian Spotify users (during the SARS-CoV-2 pandemic), and study their evolution over time. A Hierarchical Cluster Analysis has been applied in order to identify groups of weeks reflecting common musical sentiments, and a Beta regression model is used to validate the results of cluster analysis.

Key words: Covid-19, Hierarchical clustering, Beta regression, Anger index, Joy index

1 Introduction

The culture of the societies has always been characterized by art and music, since the time of the oral transmission of knowledge. Indeed, the music itself has the force to exorcise fear, anger and to instil emotions. Given this leading role, the music allows us to see how Italians faced the pandemic.
The first signal of the spread of the SARS-CoV-2 virus in Italy occurred on the 31st

Alessandro Albano
Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, Building 13, Italy e-mail: alessandro.albano@unipa.it

Mariangela Sciandra
Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, Building 13, Italy e-mail: mariangela.sciandra@unipa.it

of January 2020. From that time on, the Government restriction measures led to the quarantine, initially, of 11 different municipalities in northern Italy (on 22nd of February), and then the complete closure of all schools and universities on the 4th of March. It is well known that fear and anxiety about a disease can be overwhelming and cause strong emotions in adults and children. The goal of this paper is to understand if the bad emotional status of people affected the way they approached music. More specifically, to study how anger and joy levels in songs changed during the coronavirus pandemic and to evaluate this change over time. To this aim, firstly, two different indices of anger and joy respectively are built up by merging information from two different types of metadata: song audio features extracted from the Spotify Web API [2]- [5], and the anger/joy imparted by the lyrics extracted from Genius Web API [1]- [5]. Secondly, hierarchical clustering [7], based on Canberra distance [4], is applied in order to investigate the clustering structure over the weeks. Results from cluster analysis are validated through a Beta regression model, which analyses the relationship between clustered weeks and the proposed indices.

2 Anger and Joy indices

The study of "music and emotion" is a crucial aspect of understanding the psychological relationship between human affect and music. A song can instil emotions through sound and lyrics. The two proposed indices in this work are modified versions of the Lyrical and Sonic Anger index proposed by Oppenheimer [3]. They are derived by combining information about both the level of anger/joy due to the song's sound and the song's lyric. Among the song audio features available on the Spotify Web API, the following were used:

Energy: In a song, this measure varies from 0 to 1 and represents the percentage of intensity and activity.

Speechiness: It detects the presence of spoken words in a track. The more exclusively speech-like the recording, the closer to 1 the attribute value will be.

Valence: A measure from 0 to 1 describing the musical positiveness conveyed by a track.

Levels of anger/joy due to song's lyrics were measured through the two indices defined as it follows:

Pct angry: Percentage of angry words in the lyrics according to The NRC's lexicon of words.

Pct joy: Percentage of joyful words in the lyrics according to The NRC's lexicon of words.

The *Anger index* and the *Joy index* are defined as:

$$Anger = w_1 \cdot \sqrt{\text{pct angry} \times \text{speechiness}} + w_2 \cdot \sqrt{(1 - \text{valence}) \times \text{energy}} \qquad (1)$$

$$Joy = w_1 \cdot \sqrt{\text{pct joy} \cdot \text{speechiness}} + w_2 \cdot \sqrt{\text{valence} \times \text{energy}}, \tag{2}$$

where $w_1, w_2 \geq 0$ and $w_1 + w_2 = 1$; thus w_1 stands for the importance of the song's lyrics while w_2 represents the importance of the song's sound.
Both indices are built in such a way to vary from 0 to 1. Speechiness is used to follow the intuition that a song will be very angry (joyful) if it contains relatively many angry (joyful) words in large lyrics. Contextually, song polarity can be determined by looking at the balance between *valence* and *energy*.
In order to establish the weighting structure of the two indices, a sensitivity analysis was carried out to study the effect on the indices induced by different weighting schemes. The following weighting structures were compared:

1. $w_1 = 0.5$ $w_2 = 0.5$
2. $w_1 = 0.4$ $w_2 = 0.6$
3. $w_1 = 0.2$ $w_2 = 0.8$
4. $w_1 = 0.6$ $w_2 = 0.4$
5. $w_1 = 0.8$ $w_2 = 0.2$.

We investigated how the weighting structure influences the temporal proximity of weeks in clusters. The sensitivity analysis results show that the second weighting scheme ($w_1 = 0.4$ $w_2 = 0.6$) leads to highly cohesive clusters, which are made up of adjacent weeks (see Section 4). Therefore, we decided to give more importance to the emotional content imparted by audio features (60%) rather than the emotional content induced by the lyrics (40%).

3 The Italian music data

The dataset is made up of 340 records representing the top 20 streamed songs per week by the Italian Spotify users. The time period considered span from the 7th February to the 4th June (17 weeks). For each song, the number of weekly streams is counted. Streams are valid in Spotify when a song is played for over 30 seconds. First, lyrics corpus was preprocessed through tokenization and stop words deletion, then the *Anger and Joy indices* are computed as defined in Eq 1 - 2. The distributions of the two indices over the weeks (adequately weighted for the number of streams of each song) are shown in Fig.1, the interpretation of boxplots' color will become clear in section 4.
Fig.1 shows how the highest levels of *Anger* are associated with the lowest levels of *Joy*. In particular, from the week IV (28th Feb.-5th March) to the week VI (13th-19th March) the *Anger* index takes the highest values, with a peak in the week V (6th March -12th March). In the same week, the *Joy* index reaches a negative peak. It is reasonable, considering that they are the first three weeks of lock-down.

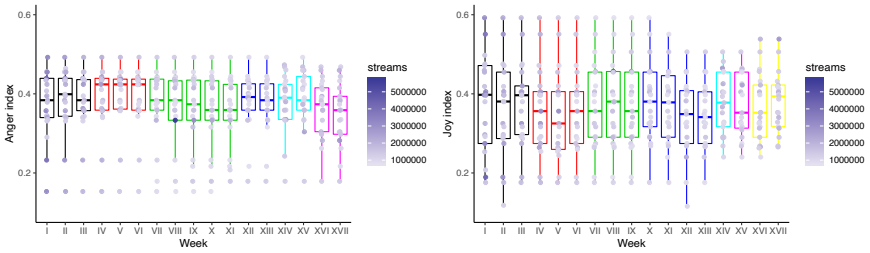


Fig. 1: Distributions of the *Anger* (left) and *Joy* (right) indices over the weeks.

4 Cluster analysis and validation

This section aims to detect the existence of weeks sharing similar sentiments by employing hierarchical cluster analysis. In general, the classification of observations into groups requires some methods for computing the distance or the (dis)similarity between each pair of observations. The choice of distance measures is a critical step in clustering since it influences the shape of the clusters. In our case, computing the distance between two probability density functions can be regarded as measuring the overlapping area between the two curves. After a deep exam of the measures proposed in the literature, we decided to use the *Canberra distance*, belonging to the L_1 family, whose generalised equation is given in the form:

$$d_{can} = \sum_{i}^{d} \frac{|P_i - Q_i|}{P_i + Q_i}, \tag{3}$$

where P_i and Q_i are the ith components of vectors **P** and **Q** respectively in an n-dimensional real vector space. In this case, **P** and **Q** represent the two discretized probability density functions under comparison.
The distance matrix between weeks was computed in `R` using the `philentropy` library. The number of clusters was selected by using the first quantile of the pairwise distances' distribution as a threshold. In Tab.1 and Fig.2, the resulting clusters are shown: as regards the *Anger index* six clusters are identified, while the clustering algorithm applied to the *Joy index* identifies seven groups.
In Fig.1 can be noticed that, although we haven't imposed any temporal constrain, all clusters are made up of adjacent weeks. Furthermore, as previously spotted in section 3, weeks IV, V and VI are effectively grouped in the same cluster either for *Anger* and *Joy* (Cluster 2 depicted in red).
Following, in order to validate the quality of the resulting clustered structure and verify if the clusters obtained are significantly different in terms of average *Anger* and *Joy*, two Beta models have been estimated as the indices are continuous and limited in the interval $[0,1]$. In each model, we set Cluster number 2 (corresponding to the first three weeks of lock-down) as a baseline, in order to compare it with

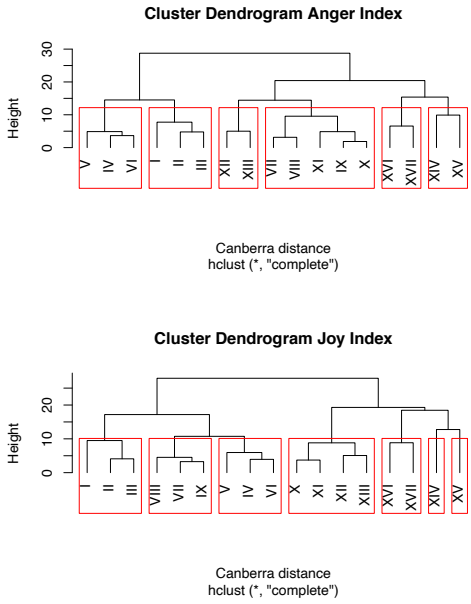


Fig. 2: Clusters dendrograms of the *Anger* and *Joy* indices.

Table 1: Cluster specifications

Cluster	Anger	Joy
Cluster 1	07/02- 27/02	07/02- 27/02
Cluster 2	28/02 - 19/03	28/02 - 19/03
Cluster 3	20/03- 23/04	20/03- 09/04
Cluster 4	24/04- 07/05	10/04- 07/05
Cluster 5	08/05 - 21/05	08/05 - 14/05
Cluster 6	22/05- 04/06	15/05- 21/05
Cluster 7	-	22/05- 04/06

all the other clusters. According to the model, Cluster number 2 can be regarded as the angriest and the least joyful one. The regression coefficients' p-values here have only a descriptive validity because they are estimated on the same data used to derive clusters. Therefore, they can only be interpreted as a confirmatory tool of an expected inferential result, also due to the high number of observations (streams).

Tab.4 shows the percentage changes, in the *Anger* and *Joy indices* of songs for with respect to Cluster 2. Cluster 6 (from 22th May to 5 th June) has the highest reduction of *Anger* equal to 8.4%, while Cluster 5 (from 8th May to 4th May) has the largest increase of *Joy* equal to 6.5%.

Table 2: Beta model estimated coefficients for *Anger* (left) and *Joy* (right)

Coef	Estimate	Std. Error	P-value
Intercept	-0.445	0.00003	$< 2e^{-16}$
Cluster 1	-0.065	0.00005	$< 2e^{-16}$
Cluster 3	-0.113	0.00005	$< 2e^{-16}$
Cluster 4	-0.054	0.00006	$< 2e^{-16}$
Cluster 5	-0.035	0.00006	$< 2e^{-16}$
Cluster 6	-0.140	0.00065	$< 2e^{-16}$

Coef	Estimate	Std. Error	P-value
Intercept	-0.593	0.00004	$< 2e^{-16}$
Cluster 1	0.065	0.00006	$< 2e^{-16}$
Cluster 3	0.071	0.00007	$< 2e^{-16}$
Cluster 4	0.022	0.00006	$< 2e^{-16}$
Cluster 5	0.100	0.00009	$< 2e^{-16}$
Cluster 6	0.083	0.00010	$< 2e^{-16}$
Cluster 7	0.096	0.00007	$< 2e^{-16}$

Table 3: Variation in *Anger* and *Joy* indices (%)

Cluster	Anger	Joy
Cluster 1	-4.2	+4.2
Cluster 3	-6.6	+4.6
Cluster 4	-3.3	+1.4
Cluster 5	-2.1	**+6.5**
Cluster 6	**-8.4**	+5.4
Cluster 7	-	+6.2

5 Conclusions

Two indices were proposed to measure the levels of anger and joy of Italian songs listened during the Covid-19 pandemic. A hierarchical clustering algorithm based on Canberra distance was applied, and the results were validated through a Beta regression model. Results show that weeks can be clustered into groups sharing similar sentiments: the first three weeks of quarantine (spanning from 28th February to 19th march) represent both the angriest and the least joyful cluster of weeks.

References

1. Genius WEB API
https://genius.com/api-clients
2. Spotify WEB API
https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/
3. Using Data to Find the Angriest Death Grips Song
https://towardsdatascience.com/angriest-death-grips-data-anger-502168c1c2f0
4. Cha, Sung-Hyuk: Survey on distance/similarity measures between probability density functions.City (2007)
5. Sciandra M., Spera I.: A model based approach to Spotify data analysis: a Beta GLMM. Journal of Applied Statistics (2020).
6. Drost, Hajk-Georg: Philentropy: information theory and distance quantification with R. Journal of Open Source Software (2018).
7. Johnson, Stephen C: Hierarchical clustering schemes. Psychometrika (1967).

A binned technique for scalable model-based clustering on huge datasets

Filippo Antonazzo, Christophe Biernacki & Christine Keribin

Abstract Clustering is impacted by the regular increase of sample sizes which provides opportunity to reveal information previously out of scope. However, the volume of data leads to some issues related to the need of many computational resources and also to high energy consumption. Resorting to binned data depending on an adaptive grid is expected to give proper answer to such green computing issues while not harming the quality of the related estimation. After a brief review of existing methods, a first application in the context of univariate model-based clustering is provided, with a numerical illustration of its advantages. The issues of a trivial multivariate extension are discussed and a marginal-binned strategy is proposed to estimate bivariate Gaussian diagonal mixtures.

Key words: Big Data, clustering, binned data, green computing.

1 Scalable clustering for huge datasets

Today, thanks to the technological development of the last decades, it is common to work on *huge datasets*, which are large collections of data whose volume (both of observations and attributes) is still growing. But, despite the enormous statistical information conveyed, any statistical analysis, such as clustering, conducted with

Filippo Antonazzo
Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France e-mail: filippo.antonazzo@inria.fr

Christophe Biernacki
Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France e-mail: christophe.biernacki@inria.fr

Christine Keribin
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay
91405 Orsay, France, e-mail: christine.keribin@universite-paris-saclay.fr

classical methods is difficult because it requests too much time, too much memory and too much energy. This is also in contrast with the current eco-friendly policies of many national governments and industries which are searching for methods able to do suitable statistical analysis without employing complex and wasteful technologies. We want to satisfy this need, proposing a method capable to analyse big data employing limited computational resources, like those of a standard laptop.

For the same reasons, scalable clustering algorithms for huge datasets flourished in literature during the last two decades. Some algorithms employ data-reduction techniques, like random subsampling [9] or data-compression through the use of sufficient statistics [14]. Other authors transform the space of analysis [11] or examine dense data units built imposing a grid on the original data [1]. It is also possible to reduce the number of operations, adopting particular data structure, such as trees [14] or graphs [9], or imposing some criteria [1] to prune irrelevant clusters that, thus, exit from the computational process. In addition, the problem of dimensionality is usually tackled down by performing clustering in subspaces of lower dimension [2].

The objective of the paper is to introduce scalability in model-based clustering [8], a statistical approach well appreciated because it enables a theoretically well-posed framework where formal criteria to assess the quality of the clustering are available. It is in this context that we will propose our novel method based on binned data, which, assuming observations with values belonging to a real space $\mathscr{X}$, correspond to a reduced dataset only containing the counts of observations in given regions of $\mathscr{X}$. In practice they usually appear as soon as it is impossible to collect data with infinite precision, like in [7] and [3], but we will use them with a different point of view. The key idea we defend is to group original data in order to obtain *artificially* binned ones and reduce the dimensionality of the problem working with them. We first consider the univariate case (where $\mathscr{X} = \mathbb{R}$) to introduce the notation and highlight, through a numerical example, how much promising is our method. Finally, we discuss how to extend it to the multivariate context, pointing out possible issues of trivial generalizations and presenting a new marginal-binned methodology able to cope with them in a restrictive bivariate diagonal scenario, as a final simulation shows.

2 Binned model-based clustering approach: univariate case

Let $\mathbf{x} = (x_1, \ldots, x_n)$, with $x_i \in \mathscr{X} = \mathbb{R}$, a raw sample of n observations arising from a univariate K-Gaussian mixture of density

$$f(x;\theta) = \sum_{k=1}^{K} \pi_k \phi(x;\mu_k,\sigma_k^2) \quad \pi_k > 0, \ \sum_{k=1}^{K} \pi_k = 1, \tag{1}$$

in which μ_k denotes the mean of the k-th component, σ_k^2 is its variance and θ is the vector that contains all the parameters, thus $\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$. The key-idea is to build a grid G made of $R << n$ cut points $(a_1, \ldots, a_R)$ that divides

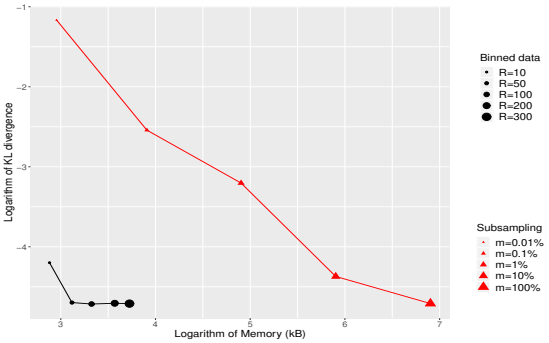


Fig. 1 Binned estimation of a simulated 3-class mixture: logarithm of Kullback-Leibler divergence between the true mixture distribution and the estimated one for different values of R and m in function of the required computer memory (logarithmic scale).

the real space $\mathbb{R}$ into $R+1$ intervals $[a_{r-1}, a_r[$, $r = 1, \ldots, R+1$, setting $a_0 = -\infty$ and $a_{R+1} = \infty$. In this way, binned data are stored in a vector $\mathbf{y} = (y_1, \ldots, y_{R+1})$, where each component is defined as

$$y_r = \#\{x_i : a_{r-1} \leq x_i < a_r\}. \tag{2}$$

As $R \ll n$, working with binned data instead of raw ones reduces the dimensionality of the problem and also proposes interesting theoretical questions. In fact, the binned statistical model is a multinomial one $M(n, p(\theta))$ with $p(\theta) = (p_1(\theta), \ldots, p_{R+1}(\theta))$, where $p_r(\theta) = \int_{a_{r-1}}^{a_r} f(x;\theta)dx$. It could be proved (result not provided here) that this model remains identifiable under certain (and weak) conditions on the grid G.

Here is a numerical example to motivate the fundamental interest of our proposed "binned" method, which is compared to the subsampling strategy (depending on the subsample percentage m) on a simulation sample of $n = 10^6$ raw data i.i.d. arising from a univariate Gaussian mixture with three components. Binned data are created through a grid with the tuning parameter R. An EM algorithm [4] is performed respectively with different values of R and m (thus different candidate subsample and binned datasets). In Figure 1 it is possible to note that the loss of information (measured by the Kullback-Leibler divergence) induced by binning is much lower than that obtained with subsampling, even negligible if we use a grid moderately dense. This is in addition accompanied by an evident gain in terms of computer memory. Such promising results could be also obtained (but not displayed here) concerning gain in terms of algorithm running time or model selection behaviour.

3 Issues of a trivial multivariate extension

Once analyzed the univariate case, extending the method to a D-variate situation seems straightforward. Let $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{x}_i \in \mathscr{X} = \mathbb{R}^D$, a sample arising from a multivariate K-Gaussian mixture of density

$$f(\boldsymbol{x};\theta)=\sum_{k=1}^{K}\pi_k\phi(\boldsymbol{x};\mu_k,\Sigma_k)\quad \pi_k>0,\ \sum_{k=1}^{K}\pi_k=1, \tag{3}$$

where, for each component $k=1,\ldots,K$, $\mu_k=(\mu_{k1},\ldots,\mu_{kD})$ is the vector of means and Σ_k is the variance-covariance matrix, with diagonal $(\sigma^2_{k1},\ldots,\sigma^2_{kD})$. It is immediate to define a multivariate grid G building it as a Cartesian product between D one-dimensional grids. It means that $G=G_1\times\ldots\times G_D$, where each grid G_d has R_d cut points $(a_{d1},\ldots,a_{dR_d})$. Assuming that $R_d=R$, for $d=1,\ldots,D$, we can define a $(R+1)^D$-dimensional binned vector $\mathbf{y}=(y_1,\ldots,y_{(R+1)^D})$, where, for $r=1,\ldots,(R+1)^D$:

$$y_r=\#\{\boldsymbol{x}_i: 1+z_{i1}+z_{i2}(R+1)+z_{i3}(R+1)^2\ldots+z_{iD}(R+1)^{D-1}=r\},$$
$$\text{with}\quad z_{id}=l \text{ if } a_{dl}\le x_{id}<a_{d(l+1)},\ l=0,\ldots,R,\ \ \forall d=1,\ldots,D,$$

where $a_{d0}=-\infty$ and $a_{d(R+1)}=\infty$ for each $d=1,\ldots,D$.

Despite the relatively simple formalization, using such a grid is not feasible. Indeed, the following issues arise:

- It is impossible to obtain a manageable amount of binned data because the number of non-empty bins increases exponentially increasing the number of variables (proof not provided here).
- The related EM algorithm employs several multidimensional numerical integrations. Thus, it would become too complex in terms of computing time.

Consequently, we propose below a specific alternative strategy (called "marginal-binned") to estimate multivariate diagonal mixtures not affected by these problems. For simplicity, we will illustrate it in a restrictive bivariate scenario, where $\mathscr{X}=\mathbb{R}^2$, even if the proposal is more general.

4 A marginal-binned strategy for bivariate diagonal mixtures

Let consider a bivariate $(D=2)$ diagonal Gaussian mixture with K components. Thus, the variances Σ_k in (3) are diagonal and the vector of parameters is simply:

$$\theta=(\underbrace{\pi_1,\ldots,\pi_K}_{\pi},\underbrace{\mu_{11},\ldots,\mu_{K1},\sigma^2_{11},\ldots,\sigma^2_{K1}}_{\alpha_1},\underbrace{\mu_{12},\ldots,\mu_{K2},\sigma^2_{12},\ldots,\sigma^2_{K2}}_{\alpha_2}).$$

Denoting with $\mathbf{x}_1$ and $\mathbf{x}_2$ the first and the second component of a sample $\mathbf{x}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$, $\boldsymbol{x}_i\in\mathbb{R}^2$, and adopting a square grid $G=G_1\times G_2$ with $R_1=R_2=R$, we define:

- $\mathbf{y}_1$: binned data vector of $\mathbf{x}_1$ under G_1;
- $\mathbf{y}_2$: binned data vector of $\mathbf{x}_2$ under G_2.

It means that, for each $d=1,2$, $\mathbf{y}_d=(y_{d1},\ldots,y_{d(R+1)})$, where each component is defined as $y_{dr}=\#\{x_{di}: a_{d(r-1)}\le x_{di}<a_{dr}\}$. We name $\mathbf{y}_1$ and $\mathbf{y}_2$ as the *marginal counts* of $\mathbf{y}$. By construction, they are equivalent to the counts obtained by binning

the univariate marginals of the joint distribution. It can be observed that each of them is a binned data vector arising from a univariate mixture with density $f_d(x_d;\theta_d) = \sum_{k=1}^{K} \pi_k \phi(x_d;\mu_{kd},\sigma^2_{kd})$, with parameter $\theta_d = (\pi,\alpha_d)$.

Given the one-dimensional binned log-likelihoods $\ell_1(\theta_1;\mathbf{y}_1)$ and $\ell_2(\theta_2;\mathbf{y}_2)$, it is possible to obtain an estimate of θ maximizing their sum $cl(\theta;\mathbf{y}_1,\mathbf{y}_2) = \ell_1(\theta_1;\mathbf{y}_1) + \ell_2(\theta_2;\mathbf{y}_2)$. This method is not new in literature: in fact, it is known as *composite likelihood estimation*, firstly introduced in [6], who also gives interesting theoretical properties of the estimators obtaining by maximizing the *composite likelihood* $cl(\theta;\mathbf{y}_1,\mathbf{y}_2)$, like consistency and asymptotic distribution. Important contributions are given in [5] and [12], who furnished, in a composite likelihood framework, a specific formulation of the EM algorithm and an application with binned data, respectively. In a mixture model context, a similar approach is followed by [10], but in a problem involving discrete data, with a more complex formulation and without taking into account the computational and memory issues mentioned in Section 3.

Combining the ideas contained in [5] and [12], we developed a new marginal-binned EM algorithm maximizing $cl(\theta;\mathbf{y}_1,\mathbf{y}_2)$ (details not displayed here) and we tried it on simulated data sets of size $n = 10^6$, generated by different bivariate diagonal mixture models with, for simplicity, two components. In particular, it is interesting to show results obtained in a difficult scenario, where the two components are not well separated: this is useful to illustrate the goodness of the proposed methodology. These ones are depicted in Figure 2, where the 0.95 density ellipses for the real and the estimated densities (with $R = 40$) of the two components are shown. It is possible to note that they are very close, as well as the respective means, denoting a good quality of estimation, despite the difficulty of the situation. The outcomes regarding time and memory performances confirm the results of the univariate simulation presented in Section 2, thus they are not displayed here.

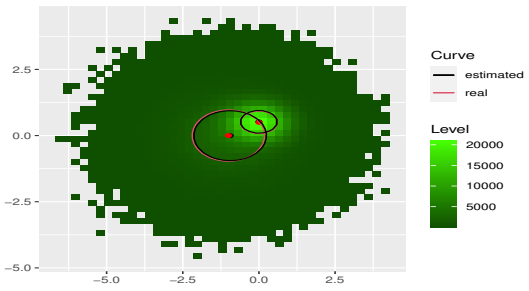


Fig. 2 0.95 density ellipses and means for the two components of the real density mixture (in red) and of the estimated one (in black). In background, the levelplot of the true density.

5 Ongoing works

The depicted methodology has proved to be efficient both from the point of view of statistical quality and computational resources management. But, some problems remain open. Firstly, it is impossible to estimate non-diagonal mixtures using only marginal counts. However, we wonder if it is possible to recover an acceptable trade-off between computational savings and clustering quality using our marginal-binned strategy. In the section dedicated to the multivariate scenario we did not mention the problem of model selection: in [13] it is possible to find some choice criteria specific for composite likelihood estimations but their calculation could be too burdensome. So, it is important to find a criterion demanding a lighter computational effort. Finally, the crucial point of the work is grid selection. We aim to find a criterion able to select the grid providing an optimal estimation (in terms of statistical quality) without neglecting the main purpose of this methodology: saving energetic resources.

References

1. Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, 94-105 (1998)
2. Böhm, C., Kailing, K., Kröger, P. & Zimek, A.: Computing clusters of correlation connected objects. In Proceedings of the 2004 ACM SIGMOD international conference on Management of data, 455-466 (2004)
3. Cadez, I. V., Smyth, P., McLachlan, G. J. & McLaren, C. E.: Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. Machine Learning, **47**(1), 7-34 (2002)
4. Dempster, A. P., Laird, N. M., & Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), **39**(1), 1-22 (1977)
5. Gao, X. & Song, P. X. K.: Composite likelihood EM algorithm with applications to multivariate hidden Markov model. Statistica Sinica, 165-185 (2011)
6. Lindsay, B. G.: Composite likelihood methods. Contemporary mathematics, **80**(1), 221-239 (1988)
7. McLachlan, G. J. & Jones, P. N.: Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics, 571-578 (1988)
8. McLachlan, G. J., & Peel, D.: Finite mixture models. John Wiley & Sons (2004)
9. Ng, R. T. & Han, J.: CLARANS: A method for clustering objects for spatial data mining. IEEE transactions on knowledge and data engineering, **14**(5), 1003-1016 (2002)
10. Ranalli, M., & Rocci, R.: Mixture models for ordinal data: a pairwise likelihood approach. Statistics and Computing, **26**(1-2), 529-547 (2016)
11. Sheikholeslami, G., Chatterjee, S. & Zhang, A.: Wavecluster: A multi-resolution clustering approach for very large spatial databases. In VLDB **98**, 428-439 (1998)
12. Whitaker, T., Beranger, B. & Sisson, S. A.: Composite likelihood methods for histogram-valued random variables. Statistics and Computing, 1-19 (2020)
13. Varin, C., Reid, N. & Firth, D.: An overview of composite likelihood methods. Statistica Sinica, 5-42 (2011)
14. Zhang, T., Ramakrishnan, R. & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, **1**(2), 141-182 (1997)

Beyond uncounfoundness in predicting counterfactuals: a machine learning approach

Gianluca Bontempi

Abstract This paper proposes a machine learning approach, called *Dependency-ToCounterFactuals* (D2CF) to estimate causal effects in a data-driven setting where neither uncounfoundness is taken for granted nor the causal structure is known. Such task is particularly challenging since the only source of information is the observational dataset, which has to be used both to infer (at least some parts) of a causal structural model and then to estimate a counterfactual quantity. A number of experiments performed on synthetic datasets shows the potential of our counterfactual algorithm with respect to state-of-the-art and naive methods.

1 Introduction

Suppose we collect a dataset recording a set of observed *treatments*, *contexts* and *outcome*. The problem of counterfactual learning is to predict the outcome variable for a given observed context if a treatment would have been assigned. Counterfactual reasoning has been covered extensively in experimental and observational studies by using the *potential outcomes* framework (also known as the Rubin Causal Model) [6]. This framework makes typically the (strong) assumption of *uncounfoundness* i.e. that the potential outcomes are independent of the observed treatments given a proper subset of the context, also called background. This assumption is untestable in practice though techniques exist to assess its plausibility from data [3]. This paper proposes D2CF, an approach for large dimensional settings which i) makes no assumption of uncounfoundness and ii) relaxes the two-population assumption by estimating causal effects whose treatment values do not necessarily belong to the observed dataset. The D2CF counterfactual learning algorithm relies on a number of theoretical results from the work of Pearl on counterfactuals [5] and combine them with a classification technique (D2C [1]) to infer

Machine Learning Group, Computer Science Department, ULB, Université Libre de Bruxelles,Brussels, Belgium, e-mail: gbonte@ulb.ac.be

causal structural information (e.g. the existence of a causal link) from observed data. Preliminary results on synthetic data are promising.

2 Counterfactual prediction

A r.v. $\mathbf{z}$ is a cause of another r.v. $\mathbf{y}$ if the distribution of $\mathbf{y}$ is different from the marginal one when we set the value of $\mathbf{z}$, i.e. $p(\mathbf{y}_z = y) = p(\mathbf{y} = y|\mathtt{do}(\mathbf{z} = z)) \neq p(\mathbf{y} = y)$ where $\mathbf{y}_z$ is called the *potential outcome* [6]. By $\mathbf{y}_z(x)$ we denote the outcome $\mathbf{y}$ when the treatment $\mathbf{z}$ is set to z and the observed context is x. In the classical case the domain of $\mathbf{z}$ is $\mathscr{Z} = \{\bar{z}, \underline{z}\}$ (e.g. in medical jargon the action $\underline{z}$ is typically called "control" and $\bar{z}$ the "treatment"). The *individualized treatment effect* (ITE) $I(x) = E[\mathbf{y}_{\bar{z}}(x)] - E[\mathbf{y}_{\underline{z}}(x)]$ is the difference of two expected potential outcomes for a specific context x (e.g. for a given patient). The estimation of this quantity in an observational setting is challenging for the following reasons: i) in an observational setting we do not have access to any measure from the interventional setting, ii) even if we were in an experimental setting (i.e. we intervene by setting $\mathbf{z} = \bar{z}$) we would know only the quantity $y_{\bar{z}(x)}$ (also called the *factual* outcome) but not the *counterfactual* term $y_{\underline{z}}(x)$ [2] and iii) we could be interested in estimating the treatment effect for treatment values $\underline{z}$ and $\bar{z}$ not contained in the observational set.

Consider now an observational setting where a dataset D_N records a set of N interventions, contexts and outcome. Suppose we want to make a counterfactual prediction, i.e. predict from the data the ITE value for an intervention on the treatment $\mathbf{z}$. The simplest approach is Direct modelling [4] where a conventional supervised learning model $\hat{y} = h(x,z)$ is fitted to the factual dataset $\langle x_i, z_i, y_i \rangle (i = 1, \ldots, N)$ and used to compute the estimate of the quantity $\mathrm{ITE}(x_i)$ when the set of possible interventions $\mathscr{Z} = \{\bar{z}, \underline{z}\}$ is composed of only two values (e.g. treatment vs. control).

This approach is naive since it makes the assumption that the observed (or empirical) factual distribution coincides with the unobserved counterfactual distribution. Causal inference literature showed that this is not the case and that the difference between those two distributions depends on two elements: i) the variation between the factual treatment assignment mechanism and the counterfactual assignment mechanism of interest, ii) the causal mechanism underlying the process.

2.1 The D2CF algorithm

The proposed algorithm (called D2CF) makes use of two well known causal properties ("Adjustment for direct causes" and the "Rule 2 of do-calculus" [5]) to avoid the limitations of the Naive approach. Let G the DAG model underlying our observations and suppose that $\mathbf{y}$ is a descendant of the variable $\mathbf{z}$. Let us define by $\pi_{\mathbf{y}}$ the set of parents of the outcome $\mathbf{y}$, by $\delta_{\mathbf{z}}$ the set of descendants of $\mathbf{z}$ and by $\delta'_{\mathbf{z}}$ the set $\mathbf{x} \setminus \delta_{\mathbf{z}}$. We have $\pi_{\mathbf{y}} = (\pi_{\mathbf{y}} \cap \delta_{\mathbf{z}}) \cup (\pi_{\mathbf{y}} \cap \delta'_{\mathbf{z}}) = \pi^0_{\mathbf{y}} \cup \pi^1_{\mathbf{y}}$ where $\pi^0_{\mathbf{y}}$ ($\pi^1_{\mathbf{y}}$) is the set

of parents of $\mathbf{y}$ which are (not) descendants of $\mathbf{z}$. From the causal properties of G it follows that i) $\mathbf{y}$ is conditionally independent of all non descendants of $\mathbf{y}$ given $\pi_{\mathbf{y}}$, ii) being $\mathbf{z}$ an ancestor of $\mathbf{y}$, $\mathbf{y}$ is conditionally independent of $\mathbf{z}$ given $\pi^1_{\mathbf{y}}$ in the graph $G_{\underline{\mathbf{z}}}$ where $G_{\underline{\mathbf{z}}}$ denotes the graph obtained by deleting all edges emerging from $\mathbf{z}$.

From the Rule 2 of do-calculus it follows $p(\mathbf{y}|\mathtt{do}(z), \pi^1_{\mathbf{y}}) = p(\mathbf{y}|z, \pi^1_{\mathbf{y}})$ The D2CF algorithm estimates the counterfactual ITE by estimating from data the term $p(\mathbf{y}|z, \pi^1_{\mathbf{y}})$. Since the causal model G is not known, we assume that a data-driven estimator (e.g. D2C [1]) is available to return the probability of existence of a causal link (e.g. parent, ancestor or descendant) between two variables. The D2CF algorithm relies on the following steps to estimate the potential outcomes $y_{\bar{z}}(x_i)$ and $y_{\underline{z}}(x_i)$ with $\bar{z} \neq z_i$ and $\underline{z} \neq z_i$:

1. Check whether $\mathbf{y}$ belongs to the descendants of $\mathbf{z}$. If this is not the case (i.e. estimated D2C probability lower than 0.5), return $y_{\bar{z}}(x_i) = y_{\underline{z}}(x_i) = y_i$ and exit;
2. For each variable in $\mathbf{x}$ estimate the probability of belonging to $\delta_{\mathbf{z}}$ (i.e. being descendants of $\mathbf{z}$);
3. For each variable in $\mathbf{x}$ estimate probability of belonging to $\pi_{\mathbf{y}}$ (i.e. being parents of $\mathbf{y}$);
4. Select the subset $\mathbf{x}' \subset \mathbf{x}$ of the S variables having the highest (estimated) probability of belonging to $\pi^1_{\mathbf{y}}$ (i.e. being parents of $\mathbf{y}$ without being descendants of $\mathbf{z}$);
5. Fit a model $h(x', z_i)$ to the observed dataset and compute the error $\varepsilon(x'_i) = y_i - h(x'_i, z_i)$;
6. For each variable in $\mathbf{x}$ estimate the probability of being parents of $\mathbf{z}$;
7. Select the subset $\mathbf{x}'' \subset \mathbf{x}$ of the S variables having the highest probability of being parents of $\mathbf{z}$;
8. Fit a model $h(x', x", z_i)$ to the observed dataset;
9. Compute $\hat{y}_{\bar{z}}(x_i) = E_{x"}[\mathbf{y}|\mathbf{z} = \bar{z}, x'_i] + \varepsilon(x'_i)$ where $E_{x"}[\mathbf{y}|\mathbf{z} = \bar{z}, x'_i] \approx \frac{\sum_{j=1}^{J} h(x'_i, x"_j, \bar{z})}{J}$ and $x"_j$ are J vectors sampled according to the marginal distribution of $\mathscr{X}'$
10. Compute analogously $\hat{y}_{\underline{z}}(x_i)$.

Step 1 checks whether the treatment $\mathbf{z}$ is indeed an ancestor of the outcome $\mathbf{y}$. Steps 2, 3 and 4 estimate the components of the sets $\delta_{\mathbf{z}}$, $\pi_{\mathbf{y}}$ and $\pi^1_{\mathbf{y}}$, respectively. Note that in Step 4 we consider the event "being parents of $\mathbf{y}$" independent of the event "being descendant of $\mathbf{z}$" and that in the experiments we set $S = 1$. Step 5 estimates from data the conditional distribution. Since this estimator could be biased (e.g. spurious paths related to errors in $\delta_{\mathbf{z}}$, $\pi_{\mathbf{y}}$ and $\pi^1_{\mathbf{y}}$), the steps 6-7-8-9 aim to correct the bias by taking advantage of the adjustment for direct causes.

3 Experimental results

This experimental session assesses D2CF vs. a number of naive and state-of-the-art approaches on several synthetic datasets generated for linear and non-linear DAG configurations of different sizes. For each observed sample x_i, z_i, y_i by simulation

we can create two counterfactual configurations where $\mathbf{z}$ is set to $\bar{z}$ and $\underline{z}$ respectively. The goal of the algorithms is to estimate from observational data the N ITE quantities $I(x_i) = \text{ITE}(x_i) = y_{\bar{z}}(x_i) - y_{\underline{z}}(x_i)$. The benchmarked algorithms are DIR1 (direct naive approach relying on a supervised learning fitting the dependency between the treatment variable (only) and the outcome), DIR2 (direct naive approach relying on a supervised learning fitting the dependency between all the variables (context and treatment variables) and the outcome), PROP (based on the computation of a propensity score) and ORACLE (having access to all the variables in the counterfactual configuration and to the causal graph, relying on supervised learning to fit the dependency between the parents of the outcome variable and the outcome). Note that ORACLE corresponds to an idealized configuration where all the causal information is available for the counterfactual prediction. The same learning algorithm (Random Forest) have been used to fit all the dependencies. The size N of the training sets is in the range $[250, 500]$. The relative error results for three different multivariate settings ($10 < n < 20$, $20 \leq n < 40$, $40 \leq n < 60$) are reported in 1, respectively.

R	$10 < n < 20$	$20 \leq n < 40$	$40 \leq n < 60$
DIR 1	0.674	0.686	0.7
DIR 2	0.884	0.932	0.959
PROP	0.675	0.687	0.703
D2CF	0.656	0.669	0.688
ORACLE	0.582	0.599	0.619

Table 1 Relative error results (the lower the better)

The experimental results show the the D2CF is consistently able to outperform the other data-driven algorithms in terms of better concordance and lower relative error. The gap between D2CF and ORACLE is a measure of the lost information due to the unavailability of the true graph. Such results on synthetic data, though preliminary, are promising. Future work will focus on the assessment of D2CF in real settings, notably uplift modelling in churn detection.

References

1. G. Bontempi and M. Flauder. From dependency to causality: A machine learning approach. *Journal of Machine Learning Research*, 16:2437–2457, 2015.
2. Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ArticleType: research-article / Full publication date: Dec., 1986 / Copyright 1986 American Statistical Association.
3. GW Imbens and DB Rubin. *Causal Inference*. Cambridge University Press, 2015.
4. Fredrik D. Johansson, Uri Shalit, and David A. Sontag. Learning representations for counterfactual inference. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 3020–3029. JMLR.org, 2016.
5. J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.
6. D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

A parsimonious parameterization of a nonnegative correlation matrix

Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria

Abstract Hierarchical relationships among manifest variables can be detected by analyzing their correlation matrix. To pinpoint the hierarchy underlying a multidimensional phenomenon, the Ultrametric Correlation Model (UCM) has been proposed with the aim of reconstructing a nonnegative correlation matrix via an ultrametric one. In this paper, we illustrate the mathematical advantages that a simple structure induced by the ultrametric property entails for the estimation of the UCM parameters in a maximum likelihood framework.

Key words: Ultrametric correlation matrix, parameterization of a correlation matrix, nonnegative correlation matrix, partitioned matrix

1 Introduction

Correlation matrices can be analyzed to detect hierarchical relationships among p manifest variables (MVs). A general correlation matrix has $p(p-1)/2$ parameters, each one representing the level of correlation between pairs of MVs. The model proposed by [2], called Ultrametric Correlation Model (UCM), provides a parsimonious representation of a nonnegative correlation matrix via an ultrametric one [3, pp. 58-59], while maintaining the relevant relations among MVs. The model aims

Carlo Cavicchia
Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands, e-mail: cavicchia@ese.eur.nl

Maurizio Vichi
University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome, Italy, e-mail: maurizio.vichi@uniroma1.it

Giorgia Zaccaria
University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome, Italy, e-mail: giorgia.zaccaria@uniroma1.it

at identifying consistent disjoint groups of MVs, each one representing a latent concept, and the hierarchical relationships among them. The non-negativity assumption turns out to be realistic in real applications (e.g., the g factor [8], the mental ability tests [1]) since many multidimensional phenomena are described by a set of variables that are concordant each other. By assuming that the variable space is partitioned into Q groups ($Q \in \{1,...,p\}$), each one associated with a latent concept, a $(p \times p)$ nonnegative correlation matrix is approximated in the UCM by

$$\mathbf{R}_{\mathrm{u}} = \mathbf{V}(\mathbf{R}_{\mathrm{B}} - \mathbf{I}_Q)\mathbf{V}' + \mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}' - \mathrm{diag}\big(\mathrm{dg}(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}')\big) + \mathbf{I}_p, \tag{1}$$

where $\mathbf{V}, \mathbf{R}_{\mathrm{W}}, \mathbf{R}_{\mathrm{B}}$ are the $(p \times Q)$ binary and row stochastic membership matrix, the $(Q \times Q)$ within-concept consistency matrix and the $(Q \times Q)$ ultrametric between-concept correlation matrix, respectively. $\mathbf{R}_{\mathrm{u}}$ turns out to be a $(2Q-1)$-ultrametric matrix which induces a hierarchy [2, Lemma 1 and Theorem 1] and it is associated with a parsimonious correlation structure. The ultrametric parameterization allows a decrease of the number of parameters needed to reconstruct a nonnegative correlation matrix. Indeed, $\mathbf{R}_{\mathrm{u}}$ can have as few as 1 parameter if $Q = 1$, or as many as $p-1$ parameters if $Q = p \geq 2$. Thus, the lower the number of the variable groups, the simpler the structure of the ultrametric correlation matrix.

In this short paper, we start to inspect the mathematical advantages that a simplified structure induced by the ultrametric property entails in the maximum likelihood estimation of the UCM under the assumption of Gaussian distributed data. We derive the main elements of the likelihood function, i.e., the simplified determinant and inverse of $\mathbf{R}_{\mathrm{u}}$, for some specific structures of the ultrametric correlation matrix. The results presented herein will be used, generalized and integrated in the extended paper along with the estimates of the UCM parameters in a maximum likelihood framework.

2 Multivariate normal distributions with the ultrametric correlation matrix

Let $\mathbf{X} = [X_1,...,X_p]'$ be a p-dimensional random vector with $\mathbf{X} \sim N_p(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $\mathbf{Y} = \mathrm{diag}(\mathrm{dg}(\boldsymbol{\Sigma}_X))^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}_X) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_Y)$, where $\mathrm{dg}(\mathbf{A})$ is the vector including the elements of the diagonal of a square matrix $\mathbf{A}$ and $\boldsymbol{\Sigma}_Y = \mathbf{R}_{\mathrm{u}}$ is the $(p \times p)$ ultrametric correlation matrix in Eq. (1). The number of parameters of $\mathbf{R}_{\mathrm{u}}$ to be estimated depends on $Q \leq p$. Under the i.i.d. assumption, the log-likelihood function for the data $\mathbf{y} = [\mathbf{y}_1,...,\mathbf{y}_n]'$, obtained from the aforementioned transformation of $\mathbf{x} = [\mathbf{x}_1,...,\mathbf{x}_n]'$, is

$$\ell(\mathbf{R}_{\mathrm{u}}; \mathbf{y}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{R}_{\mathrm{u}}| - \frac{n}{2}\mathrm{tr}(\mathbf{R}\mathbf{R}_{\mathrm{u}}^{-1}), \tag{2}$$

where $\mathbf{R}$ is the observed nonnegative correlation matrix.

Table 1 Ultrametric correlation structures.

Ultrametric Correlation Matrix	# parameters	Description
1-ultrametric correlation matrix	1	Constant correlation matrix[1]
3-ultrametric correlation matrix	$3+p$	2-block oblique correlation matrix
3-ultrametric correlation matrix with $\mathbf{R}_{\mathrm{B}}=\mathbf{I}_2$	$2+p$	2-block orthogonal correlation matrix
$(2Q-1)$-ultrametric correlation matrix	$2Q-1+p$	Q-block oblique correlation matrix[2]
$(2Q-1)$-ultrametric correlation matrix with $\mathbf{R}_{\mathrm{B}}=\mathbf{I}_{\mathbf{Q}}$	$Q+p$	Q-block orthogonal correlation matrix[2]
$(2Q-1)$-ultrametric correlation matrix with $\mathbf{R}_{\mathrm{W}}=\lambda\mathbf{I}_{\mathbf{Q}}$	$Q+p$	Q-block oblique correlation matrix with constant correlation within blocks[2]
$(2p-1)$-ultrametric correlation matrix	$p-1$	p-block correlation matrix[3]

[1] $\mathbf{V}=\mathbf{1}_p$.
[2] It is assumed $Q<p$.
[3] $\mathbf{V}=\mathbf{I}_p$.

Possible structures of the ultrametric correlation matrix $\mathbf{R}_{\mathrm{u}}$ are described in Table 1. They can be grouped in three main classes: 1-ultrametric correlation matrices, 3-ultrametric correlation matrices and $(2Q-1)$-ultrametric correlation matrices. The first one corresponds to an equicorrelation matrix in which a constant correlation occurs among MVs, i.e., $\mathbf{R}_{\mathrm{u}}=\lambda(\mathbf{1}_p\mathbf{1}_p'-\mathbf{I}_p)+\mathbf{I}_p$, where $\mathbf{1}_p$ is the p-dimensional vector of unitary elements and $\mathbf{I}_p$ is the identity matrix of order p. The second class contains two possible cases: (i) two-block oblique correlation matrix, where two groups of MVs have correlations within blocks equal to λ_1 and λ_2, respectively, and correlation between blocks equal to λ_3, i.e., Eq. (1) with $\mathbf{R}_{\mathrm{W}}=\mathrm{diag}([\lambda_1,\lambda_2]')$ and $\mathbf{R}_{\mathrm{B}}=\lambda_3(\mathbf{1}_2\mathbf{1}_2'-\mathbf{I}_2)+\mathbf{I}_2$; (ii) two-block orthogonal correlation matrix, where two groups of MVs have correlations within blocks equal to λ_1 and λ_2, respectively, and correlation among blocks equal to zero ($\lambda_3=0$), i.e., Eq. (1) with $\mathbf{R}_{\mathrm{W}}=\mathrm{diag}([\lambda_1,\lambda_2]')$ and $\mathbf{R}_{\mathrm{B}}=\mathbf{I}_2$. The third class contains four possible cases: (i) Q-block oblique correlation matrix, in which Q groups of MVs have correlations within blocks equal to the diagonal elements of $\mathbf{R}_{\mathrm{W}}$ and correlations between pairs of blocks equal to the off-diagonal elements of $\mathbf{R}_{\mathrm{B}}$, i.e. Eq. (1); (ii) Q-block orthogonal correlation matrix, in which Q groups of MVs have correlations within blocks equal to the diagonal elements of $\mathbf{R}_{\mathrm{W}}$ and zero correlation among them, i.e., Eq. (1) with $\mathbf{R}_{\mathrm{B}}=\mathbf{I}_Q$; (iii) Q-block oblique correlation matrix, with constant correlation λ within blocks and correlations between pairs of blocks equal to the off-diagonal elements of $\mathbf{R}_{\mathrm{B}}$, i.e., Eq. (1) with $\mathbf{R}_{\mathrm{W}}=\lambda\mathbf{I}_Q$; (iv) p-block correlation matrix, where $Q=p$, i.e., each group is composed of one MV, with correlations between pairs of MVs equal to the off-diagonal elements of $\mathbf{R}_{\mathrm{B}}$, i.e., Eq. (1) with $\mathbf{R}_{\mathrm{W}}=\mathbf{I}_p$.

In this section, we focus on three structures of $\mathbf{R}_{\mathrm{u}}$ shown in Table 1 - the 1-, 3- and $(2Q-1)$-ultrametric correlation matrix - illustrating the simplification of the main elements of Eq. (2) under the aforementioned parameterization of a nonnegative correlation matrix. For further details on the partitioned matrices which the following results are based on, see [4, 5].

2.1 Case 1: 1-ultrametric correlation matrix

If we assume that $Q = 1$, the 1-ultrametric correlation matrix can be written as $\mathbf{R}_{\mathrm{u}} = (1-\lambda)\mathbf{I}_p + \lambda \mathbf{1}_p \mathbf{1}_p'$, with $0 \leq \lambda < 1$. Thus, the determinant of $\mathbf{R}_{\mathrm{u}}$ is

$$\det(\mathbf{R}_{\mathrm{u}}) = [1+\lambda(p-1)](1-\lambda)^{p-1} \tag{3}$$

and its inverse - [3, p. 61] and [7] - is

$$\mathbf{R}_{\mathrm{u}}^{-1} = \frac{1}{1-\lambda}\left(\mathbf{I}_p - \frac{\lambda}{1+\lambda(p-1)}\mathbf{1}_p\mathbf{1}_p'\right). \tag{4}$$

2.2 Case 2: 3-ultrametric correlation matrix

If we assume that $Q = 2$, the 3-ultrametric correlation matrix can be written as $\mathbf{R}_{\mathrm{u}} = \lambda_3 \mathbf{V}(\mathbf{1}_2\mathbf{1}_2' - \mathbf{I}_2)\mathbf{V}' + \mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}' - \mathrm{diag}\big(\mathrm{dg}(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}')\big) + \mathbf{I}_p$, where $\mathbf{R}_{\mathrm{W}} = \mathrm{diag}([\lambda_1, \lambda_2]')$, $\lambda_1, \lambda_2, \lambda_3$ are the correlations within the first, the second group and between groups, respectively, with $0 \leq \lambda_3 \leq \lambda_s < 1$, $s = 1, 2$. $\mathbf{V}$ is assumed to have contiguous rows representing MVs which belong to the same group after an appropriate row permutation. The 3-ultrametric correlation matrix can be rewritten as

$$\mathbf{R}_{\mathrm{u}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{bmatrix},$$

where $\mathbf{A} = (1-\lambda_1)\mathbf{I}_{p_1} + \lambda_1 \mathbf{1}_{p_1}\mathbf{1}_{p_1}'$, $\mathbf{D} = (1-\lambda_2)\mathbf{I}_{p_2} + \lambda_2 \mathbf{1}_{p_2}\mathbf{1}_{p_2}'$, $\mathbf{B} = \lambda_3(\mathbf{1}_{p_1}\mathbf{1}_{p_2}')$ and p_1, p_2 represent the number of MVs in the first and the second group, respectively, s.t. $p_1 + p_2 = p$. It is worth noticing that the matrices $\mathbf{A}$ and $\mathbf{D}$ are 1-ultrametric (see Section 2.1). It follows that the determinant of $\mathbf{R}_{\mathrm{u}}$ is

$$\begin{aligned} \det(\mathbf{R}_{\mathrm{u}}) = \quad & \det(\mathbf{D})\det(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}') = [1+\lambda_2(p_2-1)](1-\lambda_2)^{p_2-1} \\ & \cdot \left\{ \left[\lambda_1 - \tfrac{p_2\lambda_3^2}{1-\lambda_2}\left(1 - \tfrac{p_2\lambda_2}{1+\lambda_2(p_2-1)}\right)\right] p_1 + (1-\lambda_1) \right\} (1-\lambda_1)^{p_1-1} \end{aligned} \tag{5}$$

and the inverse of $\mathbf{R}_{\mathrm{u}}$ is

$$\mathbf{R}_{\mathrm{u}}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{K} & \mathbf{N} \\ \mathbf{N}' & \mathbf{M} \end{bmatrix}, \tag{6}$$

where $\mathbf{K} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}')^{-1} = \left[(1-\lambda_1)\mathbf{I}_{p_1} + \left[\lambda_1 - \tfrac{p_2\lambda_3^2}{1-\lambda_2}\left(1 - \tfrac{p_2\lambda_2}{1+\lambda_2(p_2-1)}\right)\right]\mathbf{1}_{p_1}\mathbf{1}_{p_1}'\right]^{-1}$, $\mathbf{N} = -\mathbf{K}\mathbf{B}\mathbf{D}^{-1}$ and $\mathbf{M} = \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{B}'\mathbf{K}\mathbf{B}\mathbf{D}^{-1}$, $\mathbf{D}$ and $(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}')$ nonsingular.

2.3 Case 3: (2Q-1)-ultrametric correlation matrix with zero correlation among blocks of variables

If we assume that $Q = 2$ and $\lambda_3 = 0$, i.e., the correlation between the variable groups is equal to zero, $\mathbf{R}_{\mathrm{u}} = \mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}' - \mathrm{diag}\big(\mathrm{dg}(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}')\big) + \mathbf{I}_p$, where $\mathbf{R}_{\mathrm{W}} = \mathrm{diag}([\lambda_1, \lambda_2]')$. Then, the determinant of $\mathbf{R}_{\mathrm{u}}$ is

$$\det(\mathbf{R}_{\mathrm{u}}) = [1 + \lambda_1(p_1 - 1)]\,[1 + \lambda_2(p_2 - 1)]\,(1 - \lambda_1)^{p_1 - 1}(1 - \lambda_2)^{p_2 - 1} \tag{7}$$

and its inverse is

$$\begin{aligned}\mathbf{R}_{\mathrm{u}}^{-1} &= \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0}_{p_1,p_2} \\ \mathbf{0}_{p_2,p_1} & \mathbf{D}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{1-\lambda_1}\left(\mathbf{I}_{p_1} - \frac{\lambda_1}{1+\lambda_1(p_1-1)}\mathbf{1}_{p_1}\mathbf{1}'_{p_1}\right) & \mathbf{0}_{p_1,p_2} \\ \mathbf{0}_{p_2,p_1} & \frac{1}{1-\lambda_2}\left(\mathbf{I}_{p_2} - \frac{\lambda_2}{1+\lambda_2(p_2-1)}\mathbf{1}_{p_2}\mathbf{1}'_{p_2}\right) \end{bmatrix}. \end{aligned} \tag{8}$$

In order to generalize the latter case to Q groups with no correlation among them, $\mathbf{R}_{\mathrm{u}}$ can be rewritten as a block diagonal matrix

$$\mathbf{R}_{\mathrm{u}} = \begin{bmatrix} (1-\lambda_1)\mathbf{I}_{p_1} + \lambda_1\mathbf{1}_{p_1}\mathbf{1}'_{p_1} & \mathbf{0}_{p_1,p_2} & \dots & \mathbf{0}_{p_1,p_Q} \\ \mathbf{0}_{p_2,p_1} & (1-\lambda_2)\mathbf{I}_{p_2} + \lambda_2\mathbf{1}_{p_2}\mathbf{1}'_{p_2} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \mathbf{0}_{p_Q,p_1} & \dots & \dots & (1-\lambda_Q)\mathbf{I}_{p_Q} + \lambda_Q\mathbf{1}_{p_Q}\mathbf{1}'_{p_Q} \end{bmatrix},$$

with $p_1 + p_2 + \ldots + p_Q = p$. Thus,

$$\begin{aligned}\det(\mathbf{R}_{\mathrm{u}}) = {} & [1 + \lambda_1(p_1 - 1)] \cdot [1 + \lambda_2(p_2 - 1)] \cdot \ldots \cdot [1 + \lambda_Q(p_Q - 1)] \cdot (1 - \lambda_1)^{p_1 - 1} \\ & \cdot\ (1 - \lambda_2)^{p_2 - 1} \cdot \ldots \cdot (1 - \lambda_Q)^{p_Q - 1}\end{aligned} \tag{9}$$

and

$$\mathbf{R}_{\mathrm{u}}^{-1} = \begin{bmatrix} \frac{1}{1-\lambda_1}\left(\mathbf{I}_{p_1} - \frac{\lambda_1}{1+\lambda_1(p_1-1)}\mathbf{1}_{p_1}\mathbf{1}'_{p_1}\right) & \dots & \mathbf{0}_{p_1,p_Q} \\ \dots & \dots & \dots \\ \mathbf{0}_{p_Q,p_1} & \dots & \frac{1}{1-\lambda_Q}\left(\mathbf{I}_{p_Q} - \frac{\lambda_Q}{1+\lambda_Q(p_Q-1)}\mathbf{1}_{p_Q}\mathbf{1}'_{p_Q}\right) \end{bmatrix} \tag{10}$$

which is a block diagonal matrix, where each block is the inverse of a 1-ultrametric correlation matrix (see Section 2.1).

3 Conclusions and Further Developments

In this paper, a parsimonious parameterization of a nonnegative correlation matrix via an ultrametric correlation one is proposed. Moreover, we inspect the advantages

that a simple structure, induced by an ultrametric correlation matrix, entails in the maximum likelihood estimation of the Ultrametric Correlation Model parameters, assuming the normality of the data. The parameterization is studied to derive, in closed form, the equation of the determinant and inverse of an ultrametric correlation matrix in three cases, i.e., 1-ultrametric correlation matrix, 3-ultrametric correlation matrix and $(2Q-1)$-ultrametric correlation matrix with no correlation among groups of MVs. These elements are crucial in the maximum likelihood estimation of the Ultrametric Correlation Model parameters. The ultrametric correlation matrix allows a decrease of the number of parameters to be estimated compared to a general correlation matrix with $p(p-1)/2$ parameters. The generalization of the results herein to a $(2Q-1)$-ultrametric correlation matrix for estimating the Ultrametric Correlation Model in a maximum likelihood framework will be illustrated in an extended paper.

Our goal for future studies is also to introduce a test for correlation in order to pinpoint non-significant correlations in the ultrametric matrix; this can further reduce the number of parameters in the model. Furthermore, the ultrametric correlation matrix in Eq. (1) can be used to parameterize a nonnegative correlation matrix in Gaussian mixture models [6] when a multidimensional phenomenon is studied on observations coming from $G < +\infty$ sub-populations with a Gaussian distribution.

References

1. Bechtoldt, H.: An Empirical Study of the Factor Analysis Stability Hypothesis. Psychometrika **26**(4), 405–432 (1961)
2. Cavicchia, C., Vichi, M., Zaccaria, G.: The Ultrametric Correlation Matrix for Modelling Hierarchical Latent Concepts. Advances in Data Analysis and Classification **14**(4), 837–853 (2020)
3. Dellacherie, C., Martinez, S., San Martin, J.: Inverse M-Matrices and Ultrametric Matrices. Springer International Publishing, Lecture Notes in Mathematics (2014)
4. Lütkepohl, H.: Handbook of Matrices. John Wiley & Sons, England (1996)
5. Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics. John Wiley & Sons, Chicester (1988)
6. McLachlan, G., Peel, D.: Finite Mixture Models. John Wiley & Sons, New York (2000)
7. Miller, K.S.: On the Inverse of the Sum of Matrices. Math. Mag. **54**(2), 67–72 (1981)
8. Spearman, C.E.: The Abilities of Man: Their Nature and Measurement. Macmillan, New York (1927)

In-sample and cross-validated likelihood-type criteria for clustering selection

Luca Coraggio and Pietro Coretto

Abstract The selection of an optimal clustering solution is a long-standing problem. In this study, we focus on model-based clustering, where this problem amounts to choose the architecture of the mixture distribution. Decisions to be made pertain to: cluster prototype distribution; number of mixture components; (optionally) restrictions on the clusters' geometry. Typical solutions to aid these decisions use penalized model selection criteria, based on the observed likelihood function. We compare these techniques, which we refer to as in-sample methods, with cross-validation alternatives. The latter is rather popular in many data-driven applications, but is less explored for clustering problems. We analyse both classical methods such as BIC, AIC, AIC3 and ICL, and cross-validation schemes, defining the risk in terms of minus the log-likelihood function. The analysis makes use of the popular Iris dataset. We find that less explored alternatives like AIC3 and cross-validation methods yields better performances and deserve further study.

Key words: model based clustering, model selection, penalized likelihood, cross-validation.

1 Introduction

In this study we compare different criteria to select an optimal clustering solution among different candidate ones, in model-based clustering. Other studies investigated indexes used for model selection in this setting: [6, 13, 9] compare information

Luca Coraggio
Department of Economics and Statistics, University of Naples Federico II, e-mail: luca.coraggio@unina.it

Pietro Coretto
Department of Economics and Statistics, University of Salerno, e-mail: pcoretto@unisa.it

based criteria; for an extensive review see [7, Ch. 7]. With respect to these works, we add the less-explored cross-validation criteria to the comparison.

In model-based clustering it is assumed that data are generated from a finite mixture distribution with density

$$f(\cdot\,;\,\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k f_k(\cdot\,;\,\boldsymbol{a}_k),$$

where K is a finite positive integer indicating the number of components, and $\boldsymbol{\theta} = [p_1,\dots,p_K,\boldsymbol{a}_1^\top,\dots,\boldsymbol{a}_K^\top]^\top$ is the unknown parameter vector. Here f_k are densities representing the k-th *cluster*, $0 < p_k < 1$'s are mixing proportions, so that $\sum_{k=1}^{K} p_k = 1$, $\boldsymbol{a}_k$ is the parameter vector describing the cluster shape under f_k. Henceforth, for some finite integer $d > 0$, $f_k(\cdot\,;\,\boldsymbol{a}_k)$ is the Gaussian density with d-dimensional mean $\boldsymbol{\mu}_k$, and $d \times d$ covariance matrix $\boldsymbol{\Sigma}_k$, thus $\boldsymbol{a}_k = [\boldsymbol{\mu}_k^\top, \mathrm{vect}(\boldsymbol{\Sigma}_k)]^\top$, where $\mathrm{vect}(\boldsymbol{A})$ is the row-vector containing the elements of the upper triangle of the square matrix $\boldsymbol{A}$ including its diagonal. Let us indicate with $\mathscr{M}$ a set of candidate models; then, defining a member, m, of this set requires: (i) defining K; (ii) defining a parametrization for the covariance matrix $\boldsymbol{\Sigma}_k$, for $k = 1,\dots,K$. Let $\boldsymbol{a}_k = [\boldsymbol{\mu}_k^\top, \mathrm{vect}(\boldsymbol{\Sigma}_k^{(h)})]^\top$ be the parameters of the k-th component according to a certain parametrization h of the covariance structure. [4] proposed to decompose $\boldsymbol{\Sigma}_k^{(h)}$ into parameters describing geometrical notion of clusters' volume, orientation, and shape to reproduce different levels of model complexity.

Let $\boldsymbol{\theta}(m)$ be the parameter vector representing a candidate model $m \in \mathscr{M}$. Typically, the practice is to obtain an estimate, $\hat{\boldsymbol{\theta}}(m)$, via maximum likelihood (ML), for each member of $\mathscr{M}$. These estimates allow the selection of a model m^*, and its implied clustering, based on some optimality notion. In the context of Gaussian model-based clustering the choice of m^* is typically performed by optimizing an information-theoretic statistic, based on the log-likelihood function. In Section 2, we introduce some popular criteria used to inform this choice. In Section 3 we compare the different criteria using the Iris dataset.

2 Methodology

Let $X_n = \{x_1,\cdots,x_n\}$ be a sample of n data points in $\mathbb{R}^d$, for some finite integer $d > 0$. Let $z_{i,k}$ be the unobserved assignment, where $z_{i,k} = 1$ if x_i belongs to the k-th cluster and 0 otherwise. Let $K(m)$ and $h(m)$ be the values of K and h according to $m \in \mathscr{M}$. Define

$$l(\boldsymbol{\theta}(m)) = \sum_{i=1}^{n}\sum_{k=1}^{K(m)} \log(p_k f_k(x_i, \boldsymbol{a}_{k,h(m)})) \tag{1}$$

$$cl(\boldsymbol{\theta}(m)) = \sum_{i=1}^{n}\sum_{k=1}^{K(m)} z_{i,k}\log(p_k f_k(x_i, \boldsymbol{a}_{k,h(m)})) \tag{2}$$

where $l(\cdot)$ is the log-likelihood function under m, and $cl(\cdot)$ is the so called complete log-likelihood function. As mentioned in the introduction, let $\hat{\boldsymbol{\theta}}(m)$ the ML estimate of $\boldsymbol{\theta}(m)$, and let $\hat{z}_{i,k}$ be the maximum a posteriori estimates of $z_{i,k}$. Replacing $\hat{\boldsymbol{\theta}}(m)$ into (1), and $\hat{\boldsymbol{\theta}}(m)$ and $\hat{z}_{i,k}$ into (2), the corresponding sample estimates $\hat{l}(m)$ and $\hat{cl}(m)$ are obtained. Let ν_m be the number of free parameters in the model m, where ν_m increases with both $K(m)$ and the number of parameters required by the covariance parametrization $h(m)$. Among the criteria used for model selection, we consider sampling approximations of the *Bayesian Information Criterion* (BIC) of [10], the *Akaike Information Critirion* (AIC) of [1], a modified version of the AIC (AIC3) of [3] and the *Integrated Complete Likelihood Criterion* (ICL) of [2]. These are defined as:

$$AIC(m) = 2\hat{l}(m) - 2\nu_m, \qquad BIC(m) = 2\hat{l}(m) - \log(n)\nu_m,$$
$$AIC3(m) = 2\hat{l}(m) - 3\nu_m, \qquad ICL(m) = 2\hat{cl}(m) - \log(n)\nu_m.$$

According to these criteria, we select the model m^* that maximises one of the quantities above. Note that, although derived from different perspectives, these methodologies share the following form: "(complete) log-likelihood at the MLE $-$ penalty", where the penalty increases with model complexity, and sometimes decreases with n.

Another proposal, that is less explored, but still based on likelihood-type statistics, is the cross-validation (CV) method of [12]. In CV a risk measure $CV(m)$ is computed out-of-sample by splitting the available data, and a model m^* is chosen in order to optimize $CV(m)$. For a given m the CV works as follows: (i) a partition of X_n into a training-set, X^{tr}, and a test-set, X^{te}, is obtained; (ii) $\hat{\boldsymbol{\theta}}^{tr}(m)$ is estimated using the sample points in X^{tr}; (iii) $CV(m) = \hat{l}^{te}(m)/n$ is computed, where $\hat{l}^{te}$ is the estimated $\hat{l}(m)$ computed on X^{te} using $\hat{\boldsymbol{\theta}}^{tr}(m)$. In order to reduce the bias/variance of the CV, multiple splits are performed and the averaged value of $CV(m)$ is maximised. Specifically, ten-fold cross-validation amounts to partition X_n in subsets of 10% size, $X^{(i)}$, using each time 90% of the data for estimation and 10% for evaluation. That is, split $X_n = \bigsqcup_{i=1,\ldots,10} X^{(i)}$ and obtain a set of (ten) partitions: Π_{CV}. Monte Carlo cross-validation (MCCV) randomly shuffles the data X_n and then partitions it in two equal-size subsets, $X^{(tr,i)}$ and $X^{(te,i)}$. This is done M times obtaining a set of partitions Π_{MC}.

$$\Pi_{CV} = \left\{ \sqcup_{j\neq i} X^{(j)}, X^{(i)} \right\}_{i=1,\ldots,10}, \quad \Pi_{MC} = \left\{ X^{(tr,i)}, X^{(te,i)} \right\}_{i=1,\ldots,M}.$$

For every model m and every element of Π_{CV} (similarly Π_{MC}), a value of $CV(m)$ can be computed; denote with $CV_i(m)$ the value obtained according to the i-th element in Π_{CV} (Π_{MC}). The selected model is the one that achieves the maximum mean value,

$$\frac{1}{|\Pi_{CV}|} \sum_{i=1}^{|\Pi_{CV}|} CV_i(m),$$

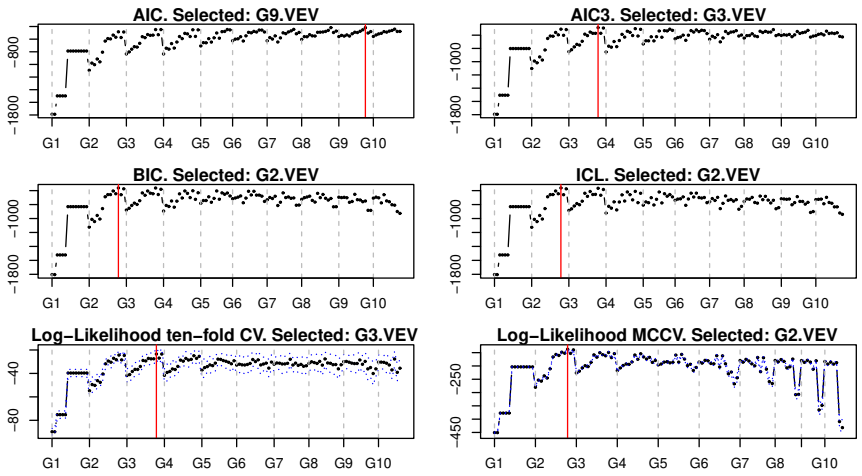


Fig. 1 x-axes show models $m \in \mathscr{M}$, ordered in terms of $K(m)$ first, and then by the number of free parameters implied by the covariance structure $h(m)$ (increasing complexity). Ticks group models by number of components $K(m)$. For example, models in between "G1" (included) and "G2" (excluded) are the 14 models with $K(m) = 1$, sorted in order of increasing free parameters, ν_m. For CV plots, 95%-confidence bands for the average $CV(m)$ are shown as well.

where $|\Pi_{CV}|$ indicates the cardinality of the set Π_{CV} (the formula for MCCV is obtained replacing Π_{CV} with Π_{MC}).

Note that this approach uses different data to estimate and evaluate the clustering solution. This does not happen for the other methods previously introduced, where the same information is used both to form the estimates, $\hat{\theta}(m)$, and to evaluate the obtained solution. We stress this point referring to AIC, AIC3, BIC and ICL as *in-sample* criteria.

3 Comparing methods on real data

We compare the methods introduced in Section 2 using the popular *Iris* dataset ([5]). This is a four dimensional dataset with $n = 150$ observations of Iris, divided in three different classes/groups based on plant species. The analysis employs the *mclust* R package ([11]) for parameters estimation. $\mathscr{M}$ includes a total of 140 Gaussian mixture models, obtained by combining the following:

- the number of mixture components, K, ranges from 1 to 10;
- 14 different covariance parametrizations, as indicated in [4], allowing for a wide variety of structures on clusters' shape and geometry.[1]

[1] These correspond to Mclust parametrizations: EII, VII, EEI, VEI, EVI, VVI, EEE, EVE, VEE, VVE, EEV, VEV, EVV, VVV.

For cross-validation, we compare two splitting methods:

- *10-fold CV*: the data set is randomly partitioned into 10 non-overlapping subsets (the folds), each used once as test-set while setting the remaining 9 folds as training-set;
- *Monte Carlo CV* (MCCV): the dataset is partitioned $T = 100$ times into two halves, one is used as training-set, the other is used as test-set.

Results for the 6 methods are summarized in Figure 1. There are two winning solutions. BIC, ICL and MCCV, select $K = 2$, ellipsoidal structures for both clusters with varying volume and orientation. This solution merges the overlapping groups corresponding to *versicolor* and *virginica* species, which might be still reasonable. AIC3 and 10-fold CV selects a solution with $K = 3$ and covariance structure as before. This solution achieves an *adjusted Rand index* = 0.9 (see [8]) where 3.3% of the points are misclassified in the strongly overlapping region between the *versicolor* and *virginica* species.

Based on our results, we conclude that AIC3 and 10-fold CV seem to have a superior performance. This is in contrast with results in [6, 13], where BIC achieves better results. In particular, the stronger penalty used in AIC3 seems to compensate the tendency of AIC to overestimate the number of groups. Cross-validation methods seem also promising, treating separately the estimation and evaluation phases of the solution.

Our analysis is limited in that it considers a single dataset. Thus, no conclusive statement can be made on the overall relative performances of the different criteria. Nonetheless, our results show that standard methods like BIC, AIC and ICL may select inferior solutions in some circumstances, where less popular alternatives like AIC3 and cross-validated likelihood outperform them. For this reason, we think that the latter criteria, especially out-of-sample ones, deserve more attention and need further investigation.

References

1. AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. Akadémiai Kiadó, Budapest, 1973, pp. 267–281.
2. BIERNACKI, C., CELEUX, G., AND GOVAERT, G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence 22*, 7 (2000), 719–725.
3. BOZDOGAN, H. Determining the number of component clusters in the standard multivariate normal mixture model using model-selection criteria. Tech. rep., ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF QUANTITATIVE METHODS, 1983.
4. CELEUX, G., AND GOVAERT, G. Gaussian parsimonious clustering models. *Pattern recognition 28*, 5 (1995), 781–793.
5. FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*, 2 (1936), 179–188.

6. FONSECA, J. R. The application of mixture modeling and information criteria for discovering patterns of coronary heart disease. *Journal of applied quantitative methods 3*, 4 (2008), 292–303.
7. FRUHWIRTH-SCHNATTER, S., CELEUX, G., AND ROBERT, C. P. *Handbook of mixture analysis*. CRC press, 2019.
8. HENNIG, C., MEILA, M., MURTAGH, F., AND ROCCI, R. *Handbook of cluster analysis*. CRC Press, 2015.
9. HUANG, T., PENG, H., AND ZHANG, K. Model selection for gaussian mixture models. *Statistica Sinica* (2017), 147–169.
10. SCHWARZ, G. Estimating the dimension of a model. *Ann. Statist. 6*, 2 (1978), 461–464.
11. SCRUCCA, L., FOP, M., MURPHY, T. B., AND RAFTERY, A. E. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal 8*, 1 (2017), 205–233.
12. SMYTH, P. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and computing 10*, 1 (2000), 63–72.
13. STEELE, R. J., AND RAFTERY, A. E. Performance of bayesian model selection criteria for gaussian mixture models. *Frontiers of statistical decision making and bayesian analysis 2* (2010), 113–130.

Outlier and novelty detection for Functional data: a semiparametric Bayesian approach

Francesco Denti, Andrea Cappozzo, and Francesca Greselin

Abstract Given a sample of unlabeled observations, the goal of a novelty detection method is to identify which units substantially deviate from the observed labeled patterns. Therefore, in a model-based framework, it is firstly of paramount importance to learn the components that correspond to the manifest groups in the training set. Secondly, one needs to take into account the lack of knowledge regarding the statistical novelties. Thirdly, contaminated elements in the known classes could greatly jeopardize the identification of new groups. Motivated by these challenges, we propose a two-stage Bayesian non-parametric novelty detector. At stage one, robust estimates are extracted from the training set and, subsequently, such information is employed to elicit informative priors within a flexible semiparametric mixture. This general paradigm can be easily adapted to complex modeling frameworks: we provide here an application to functional data from a food authenticity study.

Key words: Bayesian mixture model, Dirichlet Process Mixture Model Functional data, Minimum Regularized Covariance Determinant

Francesco Denti
Department of Statistics and Computer Science, University of California, Irvine
e-mail: fdenti@uci.edu

Andrea Cappozzo
Department of Mathematics, Politecnico di Milano
e-mail: andrea.cappozzo@polimi.it

Francesca Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca
e-mail: francesca.greselin@unimib.it

1 Introduction

A model for novelty detection can be seen as a supervised classifier, trained on a fully-labeled training set, that allows the presence of new classes in the test set, not previously observed among the training units. This framework is different from classical classification methods, where the learning units are assumed to be realizations from all the possible sub-groups contained in the target population. Moreover, traditional classifiers take the observations in the training set as a reliable source of information, namely outlier-free and label noise-free. Unfortunately, this scenario is not always the case: to face this issue, [4] has introduced an adaptive classifier that employs two algorithms (transductive and inductive) for reliable inference.
Building upon these ideas, we propose a two-stage procedure in a nonparametric Bayesian framework for simultaneously dealing with outliers, label noise, and hidden classes that may be present in the test set.
In the first phase, our model extracts the known patterns from the training dataset using the Minimum Regularized Covariance Determinant (MRCD) estimator [3]. The MRCD procedure is characterized by a parameter $\eta^{MRCD} \in [0.5, 1]$. This parameter determines the proportion of N observations to consider, such that the determinant of a convex combination of a target matrix and the sample covariance matrix is minimal. Once these $\lfloor \eta^{MRCD} N \rfloor$ observations are detected, they are used to estimate the location $\boldsymbol{\mu}^{MRCD}$ and dispersion $\boldsymbol{\Sigma}^{MRCD}$ parameters. This estimator ensures the required robustness by trimming out the problematic observations, and in addition it can be also applied to "small N big p" data problems.
In the second phase, we use insights from the training set to elicit informative priors, modeling the data in the test set with a Bayesian mixture of known groups plus a novelty term. To reflect the lack of knowledge on the novelty term, we resort to a Dirichlet Process mixture model. The adoption of such nonparametric prior overcomes the problematic specification of the number of mixture components for the novel group. The modeling framework that we introduce in Section 2 is flexible and generalizable to data of different nature. To prove this claim, we will apply our method to a functional dataset, a type of data object that has become increasingly popular in recent applications.

2 Model Specification

Let the test observations to be noisy realizations $y_m(t)$, $m = 1, \dots, M$ of a univariate stochastic process $\mathscr{X}(t), t \in \mathscr{T}$ with $\mathscr{T} \subset \mathbb{R}$. Let $\boldsymbol{\Theta}_m(t) = \left(f_m(t), \sigma_m^2(t)\right)$ denote the vector comprising a smooth mean function $f_m : \mathscr{T} \to \mathbb{R}$ and the measurement noise function $\sigma_m^2 : \mathscr{T} \to \mathbb{R}^+$ for the m-th curve in the test set at instant t. We assume $f_m(t)$ and $\sigma_m^2(t)$ to be independent $\forall t$. Our model can be succinctly specified as follows:

$$
\begin{aligned}
y_m(t)|\boldsymbol{\Theta}_m(t) = N(f_m(t), \sigma_m^2(t)), \quad \Theta_m(t) \mid \tilde{p} \overset{i.i.d.}{\sim} \tilde{p} \\
\tilde{p} = \sum_{j=1}^{J} \pi_j \delta_{\boldsymbol{\Theta}_j} + \pi_0 \left[\sum_{h=1}^{+\infty} \omega_h \delta_{\boldsymbol{\Theta}_h^{nov}} \right], \\
(\pi_0, \pi_1, \ldots, \pi_J) \sim Dir(a_0, a_1, \ldots, a_J) \\
\boldsymbol{\omega} \sim SB(\gamma), \quad \boldsymbol{\Theta}_j \sim P_j^{Tr}, \qquad \boldsymbol{\Theta}_h^{nov} \sim H,
\end{aligned}
\tag{1}
$$

where SB represents the usual Stick-Breaking representation [10] and H is the base measure of a Dirichlet Process $DP(\gamma, H)$. According to model (1), the parameter $\boldsymbol{\Theta}_m(t)$ is either sampled from one of the J known classes, or from a Dirichlet Process mixture model that describes the novel term. Therefore, the specification of the prior distributions of the known classes P_j^{Tr} is crucial.

We propose informative priors for $\boldsymbol{\Theta}_j = (f_j(t), \sigma_j^2(t))$ such that their realizations are concentrated around $\bar{f}_j$ and $\bar{\sigma}_j^2$, the estimates of the mean and variance functions in each observed class $j = 1, \ldots, J$ obtained from the training set. We smooth each training curve x_n, $n = 1, \ldots, N$, via a weighted sum of L basis functions $x_n(t) = \sum_{l=1}^{L} \rho_{nl} \phi_l(t)$, where $\phi_l(t)$ is the l-th basis evaluated at t and ρ_{nl} its associated coefficient. Promising results were obtained by adopting B-spline basis functions. Using this representation, we obtain J matrices of coefficients, each of dimension $N_j \times L$, where N_j denotes the sample size of the j-th known class. By treating them as multivariate entities, as done in [1], we apply the MRCD estimator obtaining the following robust estimates:

$$
\bar{f}_j(t) = \sum_{l=1}^{L} \hat{\rho}_{jl}^{MRCD} \phi_l(t), \quad \bar{\sigma}_j^2(t) \quad = \frac{1}{N_j - 1} \sum_{n: l_n = j} \left(x_n(t) - \bar{f}_j(t) \right)^2
$$

where $\hat{\rho}_{jl}^{MRCD}$ is the robust location estimate computed via $MRCD$ on the $N_j \times L$ matrix of coefficients and l_n denotes the class label of the n-th unit, $j = 1, \ldots, J$. Similarly, we adopt a basis representation to induce a flat prior as the base measure H. The posterior distribution of model 1 is not available in closed form. Therefore, to conduct posterior inference, we need to rely on MCMC simulation techniques. In particular, for this paper, we employ a Blocked Gibbs Sampling scheme. We follow Ishwaran and James [6] and truncate at a level K the infinite sum that models the Dirichlet Process on the novelties. The selection of K needs to balance the trade-off between model flexibility and computational complexity. As a rule of thumb, we set K to be slightly larger than the highest number of novelties that we expect to find in the test set. Precise truncation error bounds can be found in Ishwaran and James [6]. Alternatively, it is also possible to adopt different sampling schemes: for example, one could employ a Slice-sampler algorithm [7] or, if dealing with large datasets, rely on approximate inference using, for example, mean-field Variational Bayes techniques [2]. These are both research avenues that we are currently exploring.

3 Application

We employ our model in the context of a food authenticity study, where the aim is to characterize unknown food samples by identifying their type and/or provenance. In particular, we test our model on the benchmark dataset of Near Infrared Spectra (NIR) of meat varieties [8]. The considered data contain the electromagnetic spectrum for a total of 231 homogenized meat samples, with absorbance values recorded at $p = 1050$ wavelengths. The dataset encompasses five different meat types, including 32 beef, 55 chicken, 34 lamb, 55 pork, and 55 turkey units. We randomly partition the samples into training and test sets. We let the entire beef group to be unobserved, fully allocating it in the test set as a novel class. Moreover, we manually adulterate four observations in the test set, mimicking the spectra modifications performed in [5]. In our application, we consider the beef subpopulation and the adulterated units as novelty groups. We approximate each training unit via a linear combination of $L = 100$ B-spline functions, and we apply the MRCD estimator to obtain robust group-wise estimates for the splines coefficients. These quantities, which are linearly combined with the B-spline bases, account for the training atoms $\boldsymbol{\Theta}_j$, $j = 1, \dots, 4$ specified in Equation (1). A value of $\eta_{MRCD} = 0.75$ is considered for the analysis. Once $\hat{\boldsymbol{\Theta}}_j$, $j = 1, \dots, 4$ are collected, the Bayesian nonparametric model is applied to the test units. The resulting confusion matrix is reported in Table 1, where we juxtapose for comparison the classification obtained via the one-class Support Vector Method (one-class SVM) [9], the baseline classifier in the novelty detection literature. One can appreciate how the novel class, as well as the adulterated units (labeled as "Outliers" in the table), are successfully captured by the novelty component, while the same does not happen for the one-class SVM. Figure 1 reports the estimated posterior co-clustering matrix estimated on the test units, whose entries are defined as the proportion of times the algorithm assigns a pair of observations to the same group across all the iterations of the MCMC chain. The true data partition is adequately well recovered by our methodology, with only some misclassification displayed within the poultry (chicken and turkey) superset. Notice that both the beef observations and the outliers are clustered together according to the co-clustering matrix: this happens because they are both captured by the flexible novelty term. Yet our model, thanks to its probabilistic-based framework, is also able to identify specific sub-structures within such component: from the co-clustering matrix of Figure 1, the best partition for the novelty subset can be recovered minimizing, for example, the associated Variation of Information loss criterion [11]. In this way, the obtained within-novelties clustering discriminates between beef units and outliers, successfully performing anomaly and novelty detection. In conclusion, to the best of our knowledge, the introduced semi-parametric Bayesian classifier is the first flexible novelty detection method specifically tailored for dealing with functional objects.

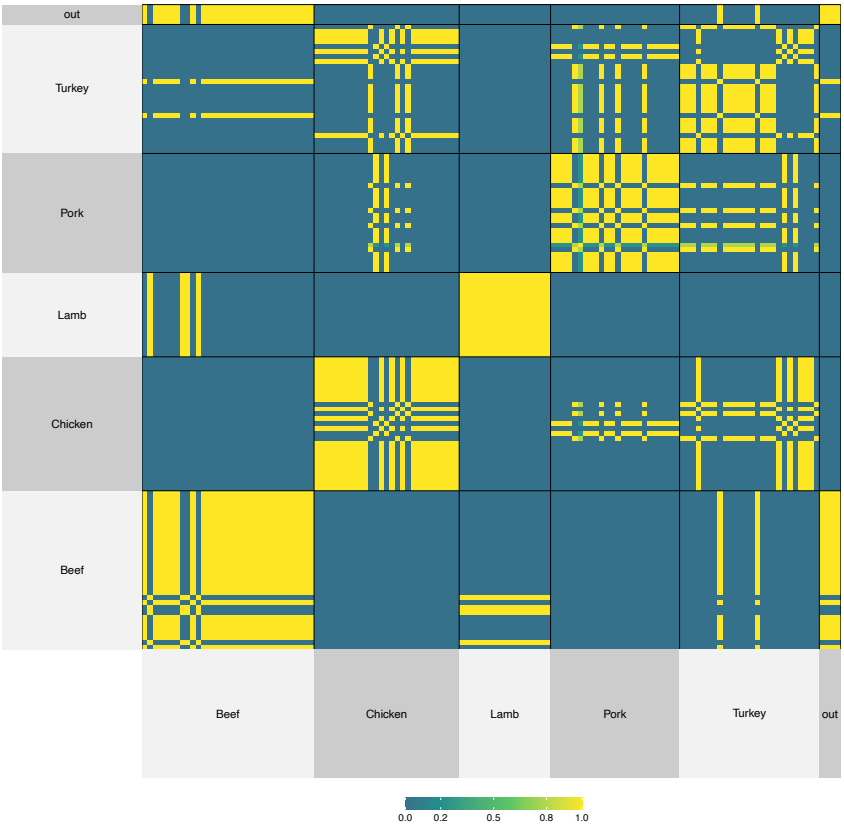


Fig. 1: Estimated posterior co-clustering matrix from the semi-parametric Bayesian classifier. Computed on the test units, meat dataset.

References

[1] C. Abraham, P. A. Cornillon, E. Matzner-Løber, and N. Molinari. "Unsupervised curve clustering using B-splines". In: *Scandinavian Journal of Statistics* 30.3 (2003), pp. 581–595.

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[3] K. Boudt, P. J. Rousseeuw, S. Vanduffel, and T. Verdonck. "The minimum regularized covariance determinant estimator". In: *Statistics and Computing* 30.1 (2020), pp. 113–128.

Table 1: Confusion matrices for the semi-parametric Bayesian classifier and one-class SVM on the test set, meat dataset.

Semi-parametric Bayesian classifier

	Truth					
Classification	Beef	Chicken	Lamb	Pork	Turkey	Outliers
Novelty	28	0	0	0	2	4
Chicken	0	18	0	0	6	0
Lamb	4	0	17	0	0	0
Pork	0	5	0	20	3	0
Turkey	0	4	0	4	15	0

One-class SVM

	Truth					
Classification	Beef	Chicken	Lamb	Pork	Turkey	Outliers
Novelty	27	5	8	14	4	0
Known patterns	5	22	9	10	22	4

[4] C. Bouveyron. “Adaptive mixture discriminant analysis for supervised learning with unobserved classes”. In: *Journal of Classification* 31.1 (2014), pp. 49–84.

[5] J. A. Fernández Pierna and P. Dardenne. “Chemometric contest at ‘Chimiométrie 2005’: A discrimination study”. In: *Chemometrics and Intelligent Laboratory Systems* 86.2 (2007), pp. 219–223.

[6] H. Ishwaran and L. F. James. “Gibbs Sampling Methods for Stick-Breaking Priors”. In: *Journal of the American Statistical Association* 96.453 (2001), pp. 161–173.

[7] M. Kalli, J. E. Griffin, and S. G. Walker. “Slice sampling mixture models”. In: *Statistics and Computing* 21.1 (2011), pp. 93–105.

[8] J. McElhinney, G. Downey, and T. Fearn. “Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats”. In: *Journal of Near Infrared Spectroscopy* 7.3 (1999), pp. 145–154.

[9] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. “Support vector method for novelty detection”. In: *Advances In Neural Information Processing Systems* 12 (2000), pp. 582–588.

[10] J. Sethuraman. “A constructive definition of Dirichlet Process prior”. In: *Statistica Sinica* 4.2 (1994), pp. 639–650.

[11] S. Wade and Z. Ghahramani. “Bayesian Cluster Analysis: Point estimation and credible balls (with Discussion)”. In: *Bayesian Analysis* 13.2 (2018), pp. 559–626.

Lasso-penalized clusterwise linear regression modeling with a two–step approach

Roberto Di Mari, Roberto Rocci and Stefano Antonio Gattone

Abstract Available approaches for the computation of lasso-penalized estimators of clusterwise linear regression models are time consuming and/or require approximate schemes. This makes the computation of the final solution, as well as the tuning of the penalty necessary to select the model, particularly cumbersome. To ease such computation, we introduce: 1) an expectation maximization algorithm with closed-form updates, that uses efficient formulas that are available for linear regression; 2) a new strategy to select the model based on a Least-Angle-RegreSsion (LARS) grid using standard information criteria.

Key words: Clusterwise linear regression, penalized likelihood, feature selection

1 Introduction

In most practical applications, regression problems deal with situations where the number of candidate covariates is large. The Lasso, proposed by Tibshirani (1996), has become very popular for the estimation of high-dimensional linear regression models where a sparse solution, a vector of regression coefficients with a relatively small number of non-zero elements, is desirable. This makes the com-

Roberto Di Mari
Department of Economics and Business, University of Catania, Italy
e-mail: roberto.dimari@unict.it

Roberto Rocci
Department of Statistical Sciences, University of Rome La Sapienza, Italy
e-mail: roberto.rocci@uniroma1.it

Stefano Antonio Gattone
Department of Philosophical and Social Sciences, Economics and Quantitative Methods, University G. d'Annunzio, Chieti-Pescara, Italy
e-mail: gattone@unich.it

putation of the final solution, as well as the tuning of the penalty necessary to select the model, particularly cumbersome. We contribute to this literature in two ways. First, by formulating an efficient procedure based on the EM algorithm with M-step updates in closed form. Second, by generalizing the use of the LARS algorithm (Efron et al., 2004) to efficiently obtain all cluster-specific lasso solutions given the estimated posterior probabilities. In this way, we build up a grid, that we call *informated*, on which we select the final solution by picking up the combination of cluster-specific lasso solutions that minimize BIC. The structure of the paper is laid out as follows. We present the methodology (Section 2), that we evaluate on a set of simulation conditions (Section 3) and conclude with some final comments.

2 The methodology

Let $\mathrm{y}_1,\ldots,\mathrm{y}_n$ be a sample of independent observations drawn from the response random variable Y_i, each observed alongside with a vector of J explanatory variables $\mathbf{x}_1,\ldots,\mathbf{x}_n$. Let us assume $\mathrm{Y}_i|\mathbf{x}_i$ to be distributed as a finite mixture of linear regression models, that is

$$f(\mathrm{y}_i|\mathbf{x}_i;\psi)=\sum_{g=1}^{G}\pi_g\phi_g(\mathrm{y}_i|\mathbf{x}_i,\sigma_g^2,\beta_g)=\sum_{g=1}^{G}p_g\frac{1}{\sqrt{2\pi\sigma_g^2}}\exp\left[-\frac{(\mathrm{y}_i-\mathbf{x}_i'\beta_g)^2}{2\sigma_g^2}\right], \quad (1)$$

where G is the number of mixture components and π_g, β_g, and σ_g^2 are the mixing proportion, the vector of $J+1$ regression coefficients that includes an intercept, and the variance term for the g-th cluster. The set of all model parameters is given by $\psi=\{(\pi_1,\ldots,\pi_G;\beta_1,\ldots,\beta_G;\sigma_1^2,\ldots,\sigma_G^2)\in\mathbb{R}^{G+(J+1)G+G}:\pi_1+\cdots+\pi_G=1,\pi_g>0,\sigma_g^2>0,\text{ for } g=1,\ldots,G\}$.
Based on the model of Equation (1), we can compute the posterior membership probabilities for each observation as follows

$$P(g|\mathrm{y}_i,\mathbf{x}_i)=\frac{\pi_g\phi_g(\mathrm{y}_i|\mathbf{x}_i;\sigma_g^2,\beta_g)}{\sum_{h=1}^{G}\pi_h\phi_h(\mathrm{y}_i|\mathbf{x}_i;\sigma_h^2,\beta_h)}, \quad (2)$$

and then classify each observation according, for instance, to fuzzy or crisp classification rules.
The negative of the log-likelihood function can be specified as

$$\ell(\psi)=-\sum_{i=1}^{n}\log\left\{\sum_{g=1}^{G}\pi_g\frac{1}{\sqrt{2\pi{\sigma_g}^2}}\exp\left[-\frac{(\mathrm{y}_i-\mathbf{x}_i'\beta_g)^2}{2\sigma_g^2}\right]\right\}, \quad (3)$$

which we minimize to estimate ψ either by means of direct optimization or with the perhaps more popular EM algorithm (Dempster et al., 1977).

Suppose that $\beta = \{\beta_1, \ldots, \beta_G\}$ is sparse, that is some or many of its elements are exactly equal to zero. ML estimates of β will therefore be close to but never exactly zero. In order to simultaneously estimate all model parameters and shrink to zero the insignificant regression coefficients, we propose minimizing the following negative lasso penalized log-likelihood function

$$p\ell(\psi) = \ell(\psi) + p(\psi), \tag{4}$$

where the penalty $p(\psi)$ is specified as follows

$$p(\psi) = \sum_{g=1}^{G} \pi_g \lambda_g \sum_{j=1}^{J} \left|\beta_{jg}\right|. \tag{5}$$

The computation of the maximum (penalized) likelihood estimates of model parameters can be done by using the usual EM algorithm with two modifications. First, the regression parameters are updated by estimating a weighted LASSO regression instead of a simple regression. For a given vector of lambdas, this update is available in closed form. Second, we note that, for penalty (5), the update for the mixing proportions is no longer $\widehat{\pi}_g = \frac{1}{n}\sum_{i=1}^{n} \widehat{P}(g|\mathrm{y}_i)$. Some authors suggest updates based on approximate schemes. Differently, we propose to parametrize the mixing proportions π_g, for $g = 1, \ldots, G$, with the following multinomial logit

$$\pi_g = \frac{\exp(\alpha_g)}{\sum_{h=1}^{G} \exp \alpha_h}, \tag{6}$$

where α_g are real–valued coefficients, for $g = 1, \ldots, G-1$, and α_G is set to zero for identification. Then, we update the α's by one Newton-Raphson step within the M-step, for which the gradient and the hessian of (4) can be computed in closed form.

In order to select the model, The proposed soft-LARS algorithm can be described in the following steps.

Step 0 - *Initialization*. Set λ_g, for $g = 1, \ldots, G$, to some pre-specified values > 0.

Step 1 - *Fit*. For fixed lambdas, estimate the model parameters of the mixture of lasso regressions.

Step 2 - *Model selection*. Select the best model for component g by using the BIC on the LARS grid of possible solutions given the estimated posterior weights.

Step 3 - *Refit (optional)*. Fit the unpenalized finite mixture by eliminating from the components the regressors corresponding to zero coefficients in step 2.

3 Simulation study

In order to asses the proposed methodology, we have carried out a simulation study. The data are generated from a 2-component clusterwise linear regression

with 5 regressors and component-specific intercepts. The simulation conditions are based on the setup of Khalili and Chen (2007). The factors that are considered are sample size (100 and 200), cluster size (0.5 and 0.5; 0.7 and 0.3; 0.9 and 0.1), magnitude of nonzero coefficients (*balanced* and *unbalanced*), and degree of collinearity between regressors (*uncorrelated* and *correlated*), with a total of 24 simulation conditions. For each simulation condition we have generated 250 data sets. For the comparison, we have included the non-penalized estimator of the clusterwise linear regression model as computed by the R package `flexmix` (Leisch, 2004) and the penalized lasso estimator used by Mortiera et al. (2015), also included in `flexmix` - we label them respectively *flex* and *flexlasso*. Of all 24 simulation conditions, we report the Mean Squared Error, the Adjusted Rand Index (ARI, Hubert and Arabie, 1985), the rate of correctly classified regression coefficients (zeros and nonzeros, avgCLASS) and the CPU time (in seconds). Results, for the sake of brevity, will be presented for the condition with $n = 100$ and equal class proportions, for both uncorrelated and correlated regressors. The mean values of all the target measures are reported in Table 1. Boxplots of ARI, MSE and avgCLASS are reported in Figure 1.

Method	ARI	MSE	avgCLASS	CPU time
	uncorrelated regressors			
flex	0.748	0.280	0.500	0.018
flexlasso	0.752	0.246	0.762	2.540
soft-LARS	0.771	0.198	0.905	0.004
	correlated regressors			
flex	0.648	0.470	0.500	0.020
flexlasso	0.657	0.394	0.745	2.850
soft-LARS	0.686	0.281	0.855	0.006

Table 1: Average results (over 250 samples). Simulation condition: $n = 100$, equal class proportions.

Results show that the presence of correlated regressors worsen all the target measures. The soft-LARS approach is the most accurate among the three considered approaches. This is the case both in terms of cluster recovery and accuracy of parameter estimate. The softLARS delivers also the highest rate of correctly classified zeros/nonzeros. These good performances are obtained with very short computing times whereas the CPU time of an entire run is on average about at least 400 times less than the penalized Lasso *flexlasso* and even quicker than the unpenalized method *flexlasso*. Results on the other simulation conditions are qualitatively in line with those reported in Table 1.

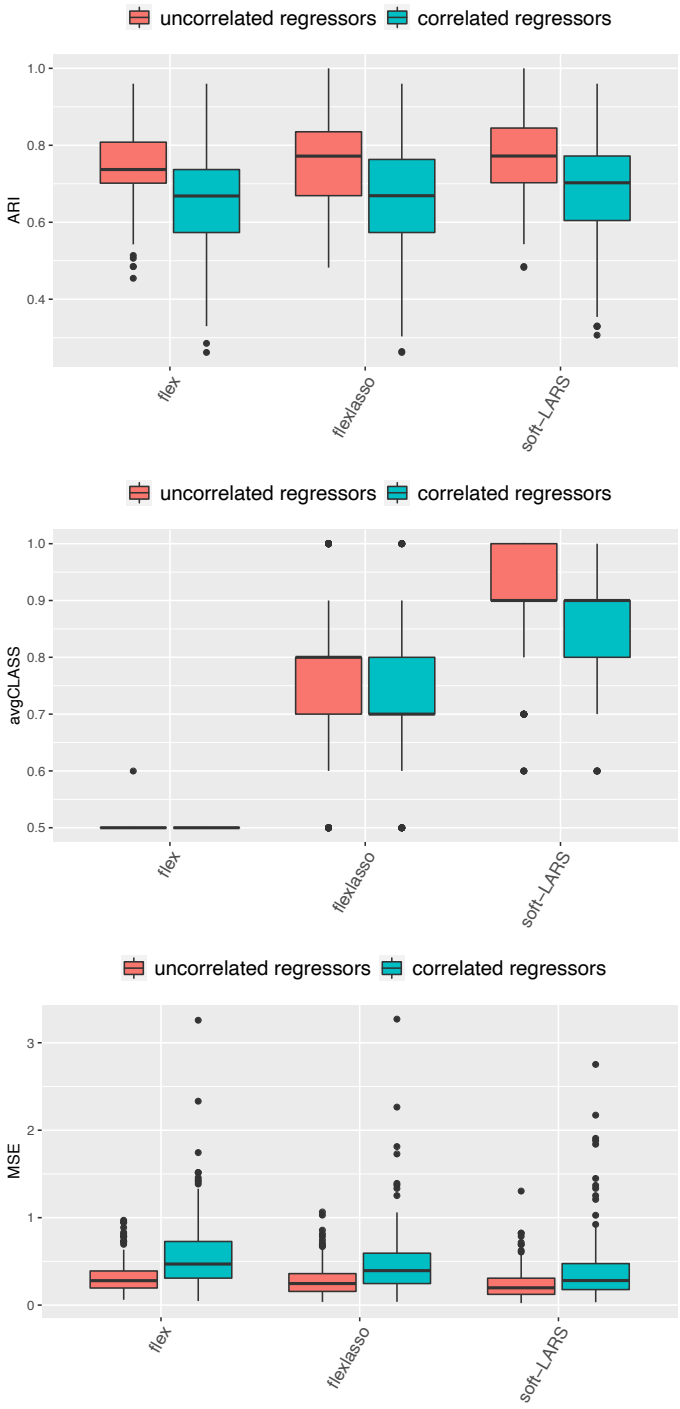


Fig. 1: Boxplot of ARI (top), avgClass (middle) and MSE (bottom) over 250 samples. Simulation condition: sample size $n = 100$, equal class proportions, uncorrelated regressors and correlated regressors.

4 Conclusions

In this chapter we have presented a penalized ML estimator of FMLR with a LASSO penalty. We observed from the simulation stud that our estimator based on the LARS search performs well - and better than its competitors. Future research might look into how to reformulate of the penalty to achieve equivariance of the penalized estimator, as well as how to handle the well–known issue of degeneracy of (conditional) Gaussian finite mixtures.

References

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39,** 1–22.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32,** 407–499.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2,** 193–218.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102,** 1025–1038.

Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent glass regression in r. *Journal of Statistical Software* **11,** 1–18.

Mortiera, F., Ouédraogo, D., Claeys, F., Tadesse, M. G., Cornu, G., Baya, F., Benedet, F., Freycon, V., Gourlet-Fleury, S., and Picard, N. (2015). Mixture of inhomogeneous matrix models for species-rich ecosystems. *Environmetrics* **26,** 39–51.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58,** 267–288.

Model-based clustering and first language acquisition

Massimo Mucciardi, Giovanni Pirrotta and Andrea Briglia

Abstract Language has been traditionally considered as a qualitative phenomenon that mainly requires hermeneutical methodologies in order to be studied, yet in recent decades - thanks to advances in data storage, processing and visualization - there has been a growing and fertile interest in analysing language by relying on statistics and quantitative methods. In light of these reasons, we think it is worthwhile to try to explore databases made up of transcripted infant spoken language in order to verify whether and how underlying patterns and recurrent sequences of learning stages work during acquisition. So, we think that model-based clustering method via the Expectation-Maximization (EM) algorithm can be useful to evaluate the development of linguistic structures over time in a reliable way.

Key words: First Language Acquisition, Model-Based Clustering, EM Algorithm, Phonetic Variation Rate, POS Tags

1 General Framework

First language acquisition can be studied and modeled by using statistical tools: experiments have shown how specific *innately biased statistical learning mechanisms* are activated during *in vitro* settings where children easily learn how to keep memory of the transitional probability between syllables to spot word' boundaries [1]. Computational methods and models have contributed to important advances in the understanding of language acquisition: corpus analysis is one of the most rigorous ways to account for pattern, regularities and learning stages in a sound and replicable procedure. The paper is organized as follows: section **2** describes the data structure;

Massimo Mucciardi e-mail: mucciard@unime.it, Andrea Briglia e-mail: abriglia@unime.it
Department of Cognitive Science, Education and Cultural Studies, University of Messina (Italy)

Giovanni Pirrotta e-mail: gpirrotta@unime.it
University of Messina (Italy)

section **3** briefly recalls the **Expectation Maximization** (EM) method, estimation strategy and data analysis. Finally, section **4** provides conclusions and suggestions for future research.

2 Data Structure

CoLaJE [2] is a database composed of seven children that have been videorecorded *in vivo* approximately one hour every month from their first year of life until they were five. In this exploratory research, statistical treatments have been tested only on one child (*Adrien*) because the transcriptions obtained from this corpus are the most complete. The data is transcripted in three forms: **CHI** is what the child says in the orthographic form, **PHO** what the child really says and **MOD** what he should have said according to the adult norm. To make the data uniform in a suitable form for automatic processing, we had to make trade-off like choices: child language is subject to interpretation difficulties by adults trying to decode it: in about 5% of the total number of occurrences, the number of words differs between the three main aforementioned forms in which sounds are coded: we decide to cut off these occurrences because they would have biased the final statistics, since the classification methods need to have an equal number of words related to the same phrase. The resulting data structure is a transformation from the video [3] into a statistically manageable database. In this respect, Code for the Human Analysis of Transcripts (CHAT) provides a standardized format for producing computerized transcripts of conversational interactions. By analyzing, cleaning, filtering and normalizing all the available original CHAT transcripts we aimed at producing one *corpus* composed of the overall amount of what the child said through the years. A total of **8214** annotated sentences containing more than 100 variables were collected. Some useful measures have been calculated such as: child age in years (time); Sentence Phonetic Variation Rate (**SPVR**) [8]: the **SPVR** is obtained by comparing *mod* and *pho* in order to measure how the relation between varied and correct form evolves over time. Then, we applied Part-Of-Speech Tagger (*POS Tags*), a software that reads text in a given language and assigns parts of speech to each word such as *noun*, *verb*, *adjective*. We used Stanza Core NLP engine [5] to tag all CHI words by using Universal Dependencies as a standard of reference for part-of-speech classification [11].

3 Data Analysis [1]

The EM algorithm is an iterative method relying on the assumption that the data is generated by a mixture of underlying probability distributions, where each component represents a separate group, or cluster. The method provides the optimal

[1] Some results are not shown due to lack of space, they are available upon request.

number of clusters in any empirical situation, by using a two step iterative algorithm: the (**E**) or expectation step and the (**M**) or maximization step. These two steps are repeated until a further increase in the number of clusters would result in a negligible improvement in the log-likelihood, namely a convergence. Accordingly, the program checks how much the overall fit improves in passing from one to two clusters (formed in all possible ways, and selecting the best), then from two to three, etc. If the error function calculated for the solution with K+1 clusters is not marked (e.g at least 5 percent better) more than the simpler solution with K clusters, then the solution with K clusters is considered ideal and retained [9] [10]. Considering the nature of the variables (count data) and assuming their independence, we use finite multivariate Poisson mixtures in the EM procedure. To extend previous research [8], we divide our database in strata considering 3 different age classes of the child (L=1.97 - 2.64; M= 2.71 - 3.39 H=3.46 - 4.33 expressed in years and months) and 3 classes of SPVR (L=$\leq$33; M=$>$33 and $\leq$66; H$>$66 expressed in percent). In total we get 9 strata (from LL to HH). By framing the analysis in this way, we turn model-based clustering via EM algorithm into a potentially interesting method that could provide a reliable way to observe linguistic structures development over time.

Table 1 provides three general indexes describing how child language is developing in quantity, quality and accuracy: these variables are represented respectively in, Child Total Words Tokenized **(CTWT)**, Child Total Distinct Words Tokenized **(CTDWT)** and Normalized Levenshtein Distance **(NLD)**. In particular **NLD** [4] is a string metric for calculating the edit distance between two given words, that means the number of deletion, insertion or substitutions of a single character needed to turn one word into the other. To obtain a realistic picture of the variation rate over a child's ages, we adjust the Levenshtein Distance by normalizing it: this means that the rate will be expressed in relative values, thus obtaining a result capable of comparing shorter and longer sentences We can observe the validity of **NLD** by the fact that it decreases over the three slot of ages as the child improves his language. In a coherent way, **CTWT**, the total number of words pronounced, increases and the **CTDWT**, the total number of different word types (proxy of an index of lexical diversity) increases as well with a similar rate. Table 2 and 3 summarize the main results obtained from clustering through a detailed overview on the most influential POS tags for each strata and its related clusters. In addition, the means of the POS are calculated in each strata **(PSM)**. We recall that the difference between **SPVR** and **NLD** is in the different way of quantifying the variation rate: SPVR counts as a variated form every word that is not pronounced exactly as it should have been pronounced (coarse-grained), while **NLD** gives a percentage of the number of letters by which the pronounced word differs from the target word (fine-grained). These general indexes have been calculated to test the soundness of our dataset: this was necessary because the following analysis and computations applied (parsing and EM) would inevitably be heavily biased by any error occurred in this initial step. Let's move on to comment on the clustering results in detail.

- VERB. We can see that VERB occupies an increasing important role in development: it is almost absent in the earlier age strata (**PSM** = L 0.02; M 0.25; H 0.18), it develops sharply in median age strata (**PSM** = 0.16; 0.62; 0.44) while it is present

in almost any sentence in the upper age strata (**PSM =** (0.79; 1.02; 0.67): it is clear also that VERB causes an increase in the error rate, as their values are higher in higher error rate strata (more than 33 percent). We can further explain the fact that VERB is higher in the LM, MM and HM strata by looking at the **CTWT** and **CTDWT** in the corresponding cells in table 1: they both have higher values as compared to the other strata: this because in these strata sentences are longer than the others and - *a fortiori* - they contain more verbs. If we want to know which specific verbs occur in the different clusters of a given strata, it is possible to observe the POS Cluster Mean **(PCM)** (values not shown) and read which kind of sentences have been placed in a specific cluster: from our results, it is possible to see how complex verbs (past and future forms, even in combination with auxiliaries) appear in later age clusters where **PCM** is higher than 0.5 while common verbs such as "to do", "to be", "to say", "to like" occur mainly in their present form in both low and high valued **PCM** in earlier strata clusters without any significant distribution detected. This difference in clustering is probably due to the fact that a two years old child essentially expresses himself through 1-2 words per sentence, so it is hard to divide something that already represents a unit in itself. When the child is four year old the clustering procedure divides in a much clearer way the corpus, helped by the fact that sentences are longer and grammatically richer. **- Morphosyntactic coherence**. If we look at the single sentence [7], we can observe that morphosyntactic coherence is higher in HL, HM clusters compared to those in L layers, which is in line with Parisse's results, we can also observe that the parts of the speech PRON, VERB, SCONJ - which could be considered as markers of longer sentences - increase their importance (see the PSM in table 2 and 3) along the age progression. Here below a couple of example[2]: *escargot tout chaud* **(CHI)** - ɛskaʁgo tu ʃo **(PHO)** - didago to so **(MOD)** in MH strata; *une souris verte* **(CHI)** - yn suʁi vɛʁtə **(PHO)** - yn ʦoji vatə **(MOD)** in HH strata. In the first, morphosyntactic coherence is expressed in a coherent way in the masculine form, but the pronoun has not been pronounced while in the second sentence the pronoun is correctly there and it is morphosyntactically coherent with the feminine form centered on the noun. We would then say that model-based clustering via EM seems capable to sort syntactically analogous sentences that are part of different error and age classes in a sufficiently precise way. **- NOUN, PROPN and PRON**. We can show how children develop a more abstract and adult-like way to referring to entities by pointing out the evolution of the values of PRON and the sum of the values of NOUN and PROPN: for L 0.02 vs 0.49, 0.20 vs 0.79, 0.09 vs 0.79; for M 0.13 vs 0.25, 0.70 vs 0.55, 0.41 vs 0.39; for H 1.14 vs 0.45, 1.48 vs 0.58, 0.74 vs 0.33. It is clear how children progressively learn to properly use pronouns instead of using nouns: this is reflected and confirmed in the fact that sentences are on average longer and thus children use anaphora in order to avoid the repetition of the noun or proper noun to indicate the main subject of the sentence. These results are in line with current literature on the acquisition of pronouns in French [6].

[2] **PHO** and **MOD** are the equivalent of the line in standard orthographic form **CHI** but have been translitterated in IPA (International Phonetic Alphabet). See for more details https://www.internationalphoneticalphabet.org/.

4 Conclusion

There are of course exceptions to these grouping tendencies but, besides that, we would suggest that these preliminary results represent a fair attempt to visualize child language development through clusters of words grouped by several criteria (age, grammatical properties, correct pronounciation). Until now, we can cautiously say that in this first stage of research the model-based clustering via EM algorithm can provide us some mild descriptions in the classification of POS tags. In other words, the unsupervised automatic procedure seems to be able to confirm a general grammatical development over time. This because cluster memberships are made up of grammatical categories that are differently learnt at different ages. Next step will be to focus on particular POS tags development over time by scanning every cluster and looking to confirm more specific learning tendencies.

Table 1: Corpus index by strata

Corpus index	**LL**	**LM**	**LH**	**ML**	**MM**	**MH**	**HL**	**HM**	**HH**
NLD	0.01	1.04	2.27	0.04	0.84	1.88	0.11	0.69	1.47
CTWT	1.52	2.52	1.54	1.88	3.67	2.34	4.54	5.43	3.01
CTDWT	1.19	2.09	1.26	1.53	3.10	1.98	3.69	4.48	2.49
# of sentences	611	184	914	851	626	1136	1762	1242	888

Table 2: Clustering results by strata (# - clusters number in brackets - POS sorted for ANOVA post-hoc F-test (in bold) $p<0.05$)

Ordered POS	**LL** (3)	**PSM**	**LM** (2)	**PSM**	**LH** (4)	**PSM**	**ML** (5)	**PSM**	**MM** (3)	**PSM**
POS1	**INTJ**	0.13	**VERB**	0.25	**PRON**	0.09	**CCONJ**	0.05	**ADP**	0.18
POS2	**DET**	0.09	**PROPN**	0.04	**ADV**	0.36	**PRON**	0.13	**ADV**	0.65
POS3	**ADP**	0.01	ADV	0.59	**DET**	0.08	**NOUN**	0.22	**DET**	0.28
POS4	**NOUN**	0.47	NOUN	0.75	**VERB**	0.18	**AUX**	0.05	**SCONJ**	0.04
POS5	**SYM**	0.02	INTJ	0.18	**NOUN**	0.62	**VERB**	0.16	**CCONJ**	0.04
POS6	**ADV**	0.56	PRON	0.20	**INTJ**	0.06	**NUM**	0.04	**INTJ**	0.17
POS7	**PROPN**	0.02	DET	0.17	**PROPN**	0.05	**SYM**	0.02	**NOUN**	0.52
POS8	**PRON**	0.02	AUX	0.10	**AUX**	0.04	**ADV**	0.83	**ADJ**	0.09
POS9	**VERB**	0.02	NUM	0.07	**ADJ**	0.02	**DET**	0.09	**NUM**	0.04
POS10	**X**	0.02	CCONJ	0.05	**SCONJ**	0.00	**PROPN**	0.03	**PROPN**	0.04
POS11	**CCONJ**	0.02	ADP	0.03	**CCONJ**	0.01	**ADP**	0.03	**AUX**	0.28
POS12	**SCONJ**	0.01	X	0.03	**ADP**	0.01	**X**	0.03	**VERB**	0.62
POS13	**AUX**	0.01	ADJ	0.02	**NUM**	0.02	**INTJ**	0.18	**PRON**	0.70
POS14	**NUM**	0.10	SCONJ	0.02	**SYM**	0.00	**ADJ**	0.01	**SYM**	0.01
POS15	ADJ	0.00	SYM	0.00	X	0.00	SCONJ	0.01	X	0.00

Table 3: Clustering results by strata (# - clusters number in brackets - POS sorted for ANOVA post-hoc F-test (in bold) $p<0.05$)

Ordered POS	**MH** (3)	**PSM**	**HL** (4)	**PSM**	**HM** (5)	**PSM**	**HH** (5)	**PSM**
POS1	**PRON**	0.41	**PRON**	1.16	**NOUN**	0.55	**AUX**	0.26
POS2	**AUX**	0.20	**DET**	0.32	**DET**	0.47	**NOUN**	0.31
POS3	**NOUN**	0.31	**VERB**	0.79	**PRON**	1.48	**VERB**	0.67
POS4	**DET**	0.16	**NOUN**	0.42	**ADJ**	0.13	**DET**	0.20
POS5	**ADP**	0.11	**SCONJ**	0.15	**AUX**	0.37	**PRON**	0.74
POS6	**ADV**	0.38	**ADP**	0.23	**VERB**	1.02	**NUM**	0.09
POS7	**PROPN**	0.08	**AUX**	0.21	**ADP**	0.26	**ADJ**	0.09
POS8	**SCONJ**	0.02	**ADV**	0.73	**ADV**	0.67	**ADP**	0.12
POS9	**VERB**	0.44	**ADJ**	0.09	**SCONJ**	0.10	**ADV**	0.31
POS10	**INTJ**	0.06	**CCONJ**	0.12	**X**	0.02	**X**	0.03
POS11	**NUM**	0.03	**SYM**	0.02	**CCONJ**	0.11	**PROPN**	0.02
POS12	X	0.01	**NUM**	0.08	**NUM**	0.04	**SCONJ**	0.04
POS13	SYM	0.00	**X**	0.02	SYM	0.01	**CCONJ**	0.04
POS14	ADJ	0.10	**PROPN**	0.03	INTJ	0.15	**INTJ**	0.08
POS15	CCONJ	0.01	**INTJ**	0.16	PROPN	0.03	SYM	0.00

References

1. Saffran J. R., Aslin R. N., Newport E. L.: *Statistical learning by 8-Month-Old infants*. Science, vol. 274,. 1926-1928, (1996)
2. Morgenstern A., Parisse C.: *The Paris Corpus*. French language studies 22. 7-12. Cambridge University Press. Special Issue, (2012)
3. CoLaJE Corpus, retrieved from http://colaje.scicog.fr/index.php/corpus, (2020)
4. Damerau J. *A technique for computer detection and correction of spelling errors*. Communications of ACM, 7 (3):171-176, (1964)
5. Zhang Y.; Zhang Y.; Bolton J.; Manning C. D. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. In Association for Computational Linguistics (ACL) System Demonstrations, (2020)
6. Morgenstern A., Sekali M. *What can child language tell us about prepositions?*. Jordan Zlatev, Marlene Johansson Falck, Carita Lundmark and Mats Andrén. Studies in Language and Cognition, Cambridge Scholars Publishing, pp.261-275,ffhalshs-00376186, (2009)
7. Parisse C., Le Normand M. T. *How children build their morphosyntax: The case of French*. Journal of Child Language, Cambridge University Press (CUP), 27, pp.267-292., (2000)
8. Briglia A., Mucciardi M., Sauvage J.: *Identify the speech code through statistics: a data-driven approach*. Proceedings SIS 2020 (Pearson Editions), (2020)
9. Dempster A.P., Laird N.M., Rubin D.B.: *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society. Series B: Methodological 39: 1–38. (1977)
10. Witten, I.H., Frank, E.: *Data Mining*. Carl Hanser, München and Wien (2011)
11. Universal Dependencies, retrieved from https://universaldependencies.org/fr/pos/index.html

Mixture of factor analyzers for mixed-type data via a composite likelihood approach

Monia Ranalli and Roberto Rocci

Abstract A parsimonious modelling approach for clustering mixed-type (ordinal and continuous) data is presented. It is assumed that ordinal and continuous data follow a finite mixture of Gaussians that is only partially observed. We define a general class of parsimonious models for mixed-type data by imposing a factor decomposition on component-specific covariance matrices. Parameter estimation is carried out using a EM-type algorithm based on composite likelihood.

Key words: Mixture models, Factor analyzers, Composite Likelihood, EM algorithm, Mixed-type data

1 Introduction

Cluster analysis methods are used to find subgroups in a population. Different clustering methods exist, mainly divided into dissimilarity-based, such as k-means, and model-based. The latter are techniques for estimating group memberships usually based on a parametric finite mixture. In this literature, the finite Gaussian mixture model is the most commonly used [7] for clustering continuous data. The idea is to interpret each mixture component as a sub-population, i.e. cluster. It can be extended to mixed-type data (continuous and ordinal variables) following the underlying variable approach (URV, [3, 4, 12]) by assuming that the ordinal variables are some variates of the mixture only partially observed (see e.g. [14, 1]).
In this framework two main issues closely related should be faced with when the dimensionality of the data is high: the number of parameters increases exponentially;

Monia Ranalli
Sapienza University of Rome, Piazzale Aldo Moro 5, Rome e-mail: monia.ranalli@uniroma1.it

Roberto Rocci
Sapienza University of Rome, Piazzale Aldo Moro 5, Rome e-mail: roberto.rocci@uniroma1.it

a large number of ordinal variables makes the full maximum likelihood estimation infeasible.
To solve the first issue, the model should be more parsimonious in terms of number of parameters to estimate. At this aim, appropriate reparameterizations need to be assumed for the covariance matrices. Accordingly, we define a general class of parsimonious mixture models for mixed-type data by imposing a factor decomposition on component-specific covariance matrices. The loadings and variances of error terms of the factor model may be constrained to be equal or unequal across mixture components [9, 8, 2].
As regard the second issue, we note that the maximum likelihood estimation is rather complex because it requires the computation of many high dimensional integrals, due to the presence of ordinal variables. The problem is usually solved by substituting the likelihood function with a surrogate function. More precisely, we replace the full likelihood with the composite likelihood [5], defined as the product of m-dimensional marginal or conditional events. In the current work, we present the model estimation considering the product of all possible sub-likelihoods based on two ordinal and all continuous variables. However, as long as sparsity is not a problem and computations are feasible, it is possible to use a higher m, including more ordinal variables. Under some regularity conditions [11], composite estimators are quite efficient [5, 15] even if less than the full maximum likelihood estimators, but much more efficient in terms of computational complexity. Model parameter estimates can be computed by an EM-type algorithm based on the complete-data composite log-likelihood.
Further details will be given in the extended version of the paper along with simulations and real data results to show the effectiveness of the proposal.

2 Model

Let $\mathbf{y}^{\bar{Q}} = [y_1, \ldots, y_{P-Q}]$ and $\mathbf{x} = [x_{P-Q+1}, \ldots, x_P]$ be $P-Q$ continuous variables and Q ordinal variables, respectively. The associated categories for each ordinal variable are denoted by $c_i = 1, \ldots, C_i$ with $i = \bar{Q}+1, \ldots, P$, where $\bar{Q} = P-Q$.
Following the underlying response variable approach, the observed variables $\mathbf{x}$ are considered as a discretization of continuous latent variables $\mathbf{y}^{Q} = [y_{\bar{Q}+1}, \ldots, y_P]$. The latent relationship between $\mathbf{x}$ and $\mathbf{y}^{Q}$ is explained by a threshold model defined as follows,

$$\gamma^{(i)}_{c_i-1} \leq y_i < \gamma^{(i)}_{c_i} \Leftrightarrow x_i = c_i,$$

where $-\infty = \gamma^{(i)}_0 < \gamma^{(i)}_1 < \ldots < \gamma^{(i)}_{C_i-1} < \gamma^{(i)}_{C_i} = +\infty$ are the thresholds defining the C_i categories.
According to our proposal $\mathbf{y} = [\mathbf{y}^{\bar{Q}}, \mathbf{y}^{Q}]$ follows a mixture of factor analyzers [9, 8, 2]

$$f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi(\mu_g, \Lambda_g \Lambda_g' + \Psi_g)$$

where ϕ is the multivariate normal density, Λ_g is the $P \times K$ matrix of factor loadings, and Ψ_g is the diagonal matrix of uniqueness. The loadings and uniqueness terms may be constrained to be equal or unequal across mixture components. The result of imposing, or not, each of these three constraints is the family of eight parsimonious Gaussian mixture models (PGMMs) that are described in Table 1 [9]. Each member of this family of models has a number of covariance parameters that is linear in data dimensionality. By assuming a common covariance structure, an even more parsimonious model can be used.

Table 1 The covariance structure of each latent parsimonious Gaussian mixture model

Model ID	$\Lambda_g = \Lambda$	$\Psi_g = \Psi$	Isotropic	Covariance Structure
CCC	C	C	C	$\Sigma_g = \Lambda\Lambda' + \psi\mathbf{I}_P$
CCU	C	C	U	$\Sigma_g = \Lambda\Lambda' + \Psi$
CUC	C	U	C	$\Sigma_g = \Lambda\Lambda' + \psi_g\mathbf{I}_P$
CUU	C	U	U	$\Sigma_g = \Lambda\Lambda' + \Psi_g$
UCC	U	C	C	$\Sigma_g = \Lambda_g\Lambda_g' + \psi\mathbf{I}_P$
UCU	U	C	U	$\Sigma_g = \Lambda_g\Lambda_g' + \Psi$
UUC	U	U	C	$\Sigma_g = \Lambda_g\Lambda_g' + \psi_g\mathbf{I}_P$
UUU	U	U	U	$\Sigma_g = \Lambda_g\Lambda_g' + \Psi_g$

For a random i.i.d. sample of size N, the log-likelihood is

$$\ell(\theta) = \sum_{n=1}^{N} \log \left[\sum_{g=1}^{G} p_g f(\mathbf{y}_n^{\bar{Q}}; \mu_g^{\bar{Q}}, \Sigma_g^{\bar{Q}\bar{Q}}) \pi_n \left(\mu_{n;g}^{Q|\bar{Q}}, \Sigma_g^{Q|\bar{Q}}, \gamma \right) \right] \tag{1}$$

(2)

where

$$\Sigma_g^{\bar{Q}\bar{Q}} = \mathrm{V}(\mathbf{y}^{\bar{Q}} \mid g),$$

$$\Sigma_g^{Q|\bar{Q}} = \mathrm{V}(\mathbf{y}^{Q} \mid \mathbf{y}^{\bar{Q}}, g)$$

and

$$\pi_n \left(\mu_{n;g}^{Q|\bar{Q}}, \Sigma_g^{Q|\bar{Q}}, \gamma \right) = \int_{\gamma_{c_1-1}^{(\bar{Q}+1)}}^{\gamma_{c_1}^{(\bar{Q}+1)}} \cdots \int_{\gamma_{c_P-1}^{(P)}}^{\gamma_{c_P}^{(P)}} \phi(\mathbf{y}^{Q}; \mu_{n;g}^{Q|\bar{Q}}, \Sigma_g^{Q|\bar{Q}}) d\mathbf{y}^{Q}$$

is the probability of response pattern $\mathbf{x}_n$ in the g-th component mixture with mean and covariance matrix conditioned on the continuous variables. This likelihood causes non trivial computational problems due to the presence of multidimensional integrals.

3 Estimation

As suggested in [14] and references therein, a composite likelihood approach could be adopted. It allows us to simplify the problem by replacing the full likelihood with a surrogate function based on m-dimensional marginals. It is a robust estimation method and its estimators have been proven to be consistent, asymptotically unbiased and normally distributed, under some mild regularity conditions [5, 15, 11]. In general they are less efficient than the full maximum likelihood estimators, or estimators obtained with a higher m, but in many cases the loss in efficiency is very small or almost null [5, 6].
In the sequel we refer to the case based on $Q(Q-1)/2$ marginal distributions each of them composed of two ordinal variables and $\bar{Q}$ continuous variables.

$$c\ell(\theta) = \sum_{n=1}^{N} \sum_{i_o=1}^{Q-1} \sum_{j_o=i_o+1}^{Q} \sum_{c_{i_o}=1}^{C_{i_o}} \sum_{c_{j_o}=1}^{C_{j_o}} \log \left[\sum_{g=1}^{G} p_g f(\mathbf{y}_n^{\bar{Q}}, \mu_g^{\bar{Q}}, \Sigma_g^{\bar{Q}\bar{Q}}) \right.$$
$$\left. \pi_{c_{i_o} c_{j_o}}^{(i_o j_o)}(\mu_{n;g}^{(i_o j_o|\bar{Q})}, \Sigma_g^{(i_o j_o|\bar{Q})}, \gamma^{j_o}) \right],$$

where $\pi_{c_{i_o} c_{j_o}}^{(i_o j_o)}(\mu_{n;g}^{(i_o j_o|\bar{Q})}, \Sigma_g^{(i_o j_o|\bar{Q})}, \gamma^{j_o})$ is the conditional probability of variables i and j being in category c_{i_o} and c_{j_o}, respectively, given all the $\bar{Q}$ continuous variables. Under the model, this conditional probability is obtained by integrating the density of a bivariate normal distribution with parameters $(\mu_{n;g}^{(i_o j_o|\bar{Q})}, \Sigma_g^{(i_o j_o|\bar{Q})}, \gamma^{j_o})$ between the corresponding threshold parameters. The computation of parameter estimates is carried out using simultaneously a standard EM algorithm on each sub-likelihood having the same set of parameters.

3.1 Classification, model selection and identifiability

As regards the classification, the observation is assigned to the component with the maximum fit according to CMAP criterion [13]. In a context of standard mixture models, the classification of the observations can be easily based on the MAP criterion. This means that the observation is assigned to the component corresponding to the maximum fit. In the same manner, the observations can be classified also under a composite likelihood framework: the observation is assigned to the component with the maximum fit. However, since the composite likelihood is constructed as the product of $Q(Q-1)/2$ sub-likelihoods, following the same principle, the fit of each observation is obtained by multiplying the corresponding $Q(Q-1)/2$ fits. In our experience, this does not cause any particular issue of numerical underflow. However, to handle potential numerical underflow, it is always possible to apply some numerical tricks based on the logarithm scale. In order to express the fit in terms of degree of membership, the fit of each observation can be normalized (i.e. it varies

between 0 and 1).
In the estimation procedure, we assume that the number of mixture components and the structure of covariance matrices are fixed. In practice, they are often unknown and thus, they have to be estimated through observed data. The best fitted model can be chosen by selecting the model minimizing the composite integrated classification likelihood (c-ICL) [13].

$$\text{c-ICL} = -2c\ell_c(\theta) + d\log N = -2c\ell(\theta) + 2EN(\hat{\mathbf{p}}) + d\log N \tag{3}$$

where $c\ell_c$ is the conditional expectation of the complete composite log-likelihood given the observed data, d is the number of parameters, while the second term is known as entropy of the fuzzy classification obtained in the E-step of the EM algorithm, $EN(\hat{\mathbf{p}})$.
A further important point of the proposed model that is worth to be discussed is parameter identifiability. To estimate both thresholds and component parameters if all the observed variables have three categories at least and when groups are know, we set the first two thresholds to 0 and 1, respectively [10]. Finally, the factorial reparameterization of component-specific covariance is not uniquely indentified. A possible identification constraint is to require that $\Lambda_g' \Psi_g^{-1} \Lambda_g$ is a diagonal matrix.
Further details will be given in the extended version of the paper along with simulations and real data results to show the effectiveness of the proposal.

References

1. EVERITT, B. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters 6*, 5 (1988), 305–309.
2. GHAHRAMANI, Z., AND HINTON, G. E. The em algorithm for mixtures of factor analyzers.
3. JÖRESKOG, K. G. New developments in lisrel: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity 24*, 4 (1990), 387–404.
4. LEE, S.-Y., POON, W.-Y., AND BENTLER, P. Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters 9*, 1 (1990), 91–97.
5. LINDSAY, B. Composite likelihood methods. *Contemporary Mathematics 80* (1988), 221–239.
6. MARDIA, K. V., KENT, J. T., HUGHES, G., AND TAYLOR, C. C. Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika 96*, 4 (2009), 975–982.
7. MCLACHLAN, G., AND PEEL, D. *Finite Mixture Models*. Wiley, 2000.
8. MCLACHLAN, G., PEEL, D., AND BEAN, R. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis 41*, 3 (2003), 379 – 388.
9. MCNICHOLAS, P. D., AND MURPHY, T. B. Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics 26*, 21 (2010), 2705–2712.
10. MILLSAP, R. E., AND YUN-TEIN, J. Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research 39*, 3 (2004), 479–515.
11. MOLENBERGHS, G., AND VERBEKE, G. *Models for discrete longitudinal data*. Springer Series in Statistics Series. Springer Science+Business Media, Incorporated New York, 2005.
12. MUTHÉN, B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika 49*, 1 (1984), 115–132.

13. RANALLI, M., AND ROCCI, R. Standard and novel model selection criteria in the pairwise likelihood estimation of a mixture model for ordinal data. *Analysis of Large and Complex Data. Studies in Classification,Data Analysis and Knowledge Organization. Editors: Adalbert F.X. Wilhelm Hans A. Kestler.* (2016).
14. RANALLI, M., AND ROCCI, R. Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis 110*, C (2017), 87–102.
15. VARIN, C., REID, N., AND FIRTH, D. An overview of composite likelihood methods. *Statistica Sinica 21*, 1 (2011), 1–41.

An application of matrix-variate mixtures to insurance data

Salvatore D. Tomarchio, Antonio Punzo and Luca Bagnato

Abstract Over the last years, there has been a growing interest in the analysis of matrix-variate data via mixture models. Quite often the tails of the matrix-variate normal distribution, used for the mixture components, are lighter than required, implying a bad fitting and the disruption of the underlying grouping structure. A solution to this issue consists in fitting mixtures of matrix-variate distributions with heavy tails. An example of such situation is here discussed by using a dataset concerning the non-life Italian insurance market. The fitting results of the matrix-variate normal mixture model are the worst, and the related data classification seems not realistic compared to the one produced by the heavy-tailed models.

Key words: Mixture models, Model-based clustering, Matrix-variate

1 Introduction

Matrix-variate data (also known as three-way data) have been increasingly discussed in the recent years, especially within the model-based clustering literature. Such data structure can occur from the observation of several attributes p measured in different situations r on a set of units N. Therefore, the data can be arranged in a three-way structure characterized by the following three dimensions: attributes (rows), situations (columns) and units (layers). In other terms, we have a $p \times r$ observed matrix for each statistical unit.

Salvatore D. Tomarchio
Università degli Studi di Catania, e-mail: daniele.tomarchio@unict.it

Antonio Punzo
Università degli Studi di Catania, e-mail: antonio.punzo@unict.it

Luca Bagnato
Università Cattolica del Sacro Cuore, e-mail: luca.bagnato@unicatt.it

In the model-based clustering literature, Viroli (2011) firstly introduced matrix-variate normal mixtures (MVN-Ms). However, for many real phenomena, the tails of the matrix-variate normal distribution, used for the mixture components, are lighter than required, with a direct effect on the corresponding mixture model in terms of fitting and disruption of the true grouping structure. The most commonly used solution to deal with such situations consists in relaxing the normality assumption of the mixture components. For example, Doğru et al. (2016) proposed mixtures of matrix-variate t distributions (MVt-Ms), and more recently Tomarchio et al. (2020) introduced mixtures of matrix-variate shifted exponential normal distributions (MVSEN-Ms) and mixtures of matrix-variate tail-inflated normal distributions (MVTIN-Ms).

In fashion of Tomarchio et al. (2020), all the aforementioned mixture models are herein considered. Specifically, they are fitted to a dataset concerning the non-life insurance consumption in Italy, across the 103 Italian provinces and over the years 1998–2002. Such data have been analyzed by Millo and Carmeci (2011) via a panel analysis, to assess the determinants of the non-life insurance consumption in Italy. The Italian territory has been always differentiated from the social, cultural, demographic and economic points of view. In detail, such distinctions mainly oppose the Central-Northern part of the country with the Southern and Insular one (see. e.g. Brunello and Cappellari, 2008; González, 2011). Relatedly, the insurance industry is also affected by such dichotomy (Millo and Carmeci, 2011). By rearranging the data in a three-way structure, we firstly evaluate if MVN-Ms provide an adequate fit to the data with respect to the heavy-tailed mixtures (MVt-Ms, MVSEN-Ms and MVTIN-Ms). Then, a comparison among the produced data classifications is done. In Section 2, some details about the considered mixtures are given, whereas Section 3 contains the insurance data analysis.

2 Matrix-variate mixture models

A $p \times r$ random matrix $\mathbf{X}$ arises from a parametric finite mixture model if its probability density function (pdf) can be written as

$$f(\mathbf{X};\boldsymbol{\Delta}) = \sum_{g=1}^{G} \alpha_g f(\mathbf{X};\boldsymbol{\Theta}_g), \tag{1}$$

where α_g is the mixture weight of the g-th component, with $\alpha_g > 0$ and $\sum_{g=1}^{G} \alpha_g = 1$, $f(\mathbf{X};\boldsymbol{\Theta}_g)$ is the pdf of the g-th component with parameters $\boldsymbol{\Theta}_g$, and $\boldsymbol{\Delta}$ contains all of the parameters of the mixture.

Here, we give the following four possibilities for the g-th mixture component, $g = 1,\ldots,G$, in (1):

- the MVN distribution, with $p \times r$ mean matrix $\mathbf{M}$, $p \times p$ row covariance matrix $\boldsymbol{\Sigma}$ and $r \times r$ column covariance matrix $\boldsymbol{\Psi}$:

$$f_{\text{MVN}}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi}) = \frac{\exp\left\{-\frac{\boldsymbol{\Omega}}{2}\right\}}{(2\pi)^{\frac{pr}{2}}|\boldsymbol{\Sigma}|^{\frac{r}{2}}|\boldsymbol{\Psi}|^{\frac{p}{2}}}, \tag{2}$$

where $\boldsymbol{\Omega} = \text{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\boldsymbol{\Psi}^{-1}(\mathbf{X}-\mathbf{M})'\right]$ denotes the squared Mahalanobis distance from $\mathbf{X}$ to the center $\mathbf{M}$ with respect to $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$. This notation will be used also for the subsequent distributions.

- The MVSEN distribution, with $p \times r$ mean matrix $\mathbf{M}$, $p \times p$ row scale matrix $\boldsymbol{\Sigma}$, $r \times r$ column scale matrix $\boldsymbol{\Psi}$ and tailedness parameter θ:

$$f_{\text{MVSEN}}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi},\theta) = \frac{\theta\exp(\theta)}{(2\pi)^{\frac{pr}{2}}|\boldsymbol{\Sigma}|^{\frac{r}{2}}|\boldsymbol{\Psi}|^{\frac{p}{2}}}\varphi_{\frac{pr}{2}}\left(\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}+\theta\right), \tag{3}$$

where $\varphi_m(z)$ is the Misra function (Misra, 1940), generalized form of the exponential integral function.

- The MVTIN distribution, with $p \times r$ mean matrix $\mathbf{M}$, $p \times p$ row scale matrix $\boldsymbol{\Sigma}$, $r \times r$ column scale matrix $\boldsymbol{\Psi}$ and tailedness parameter θ:

$$\begin{aligned} f_{\text{MVTIN}}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi},\theta) &= \frac{2\,(\pi)^{-\frac{pr}{2}}|\boldsymbol{\Sigma}|^{-\frac{r}{2}}|\boldsymbol{\Psi}|^{-\frac{p}{2}}}{\theta\,\delta(\mathbf{X};\boldsymbol{\Omega})^{\frac{pr}{2}+1}} \\ &\times\left[\Gamma\left(\frac{pr}{2}+1,(1-\theta)\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}\right)-\Gamma\left(\frac{pr}{2}+1,\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}\right)\right], \end{aligned} \tag{4}$$

where $\Gamma(a,z)$ denotes the upper incomplete gamma function.

- The MVt distribution, with $p \times r$ mean matrix $\mathbf{M}$, $p \times p$ row scale matrix $\boldsymbol{\Sigma}$, $r \times r$ column scale matrix $\boldsymbol{\Psi}$ and degrees of freedom ν:

$$f_{\text{MV}t}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi},\nu) = \frac{|\boldsymbol{\Sigma}|^{-\frac{r}{2}}|\boldsymbol{\Psi}|^{-\frac{p}{2}}\Gamma\left(\frac{pr+\nu}{2}\right)}{(\pi\nu)^{\frac{pr}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{\boldsymbol{\Omega}}{\nu}\right]^{-\frac{pr+\nu}{2}}, \tag{5}$$

where $\Gamma(a)$ denotes the gamma function.

Parameter estimation is carried out via several extensions of the expectation-maximization (EM) algorithm. For a complete description of the algorithms see Viroli (2011); Doğru et al. (2016); Tomarchio et al. (2020).

3 Insurance data application

The dataset is contained in the **splm** package (Millo and Piras, 2012) for the R computing environment. For this application, we consider the following three variables: (V1) real per capita premiums in 2000 euros, non-life insurance excluding mandatory motor third party liability, (V2) real per-capita GDP and (V3) real per-capita bank deposits. The reasons why we focused on this financial variables are: (1) they are nearly regularly used in the non-life insurance literature, and their influence on

the insurance industry has been widely discussed (see the references in Millo and Carmeci, 2011, for further details); (2) avoid an overparametrization of the models. Therefore for each province we have $p = 3$ variables and $r = 5$ years, resulting in a 3×5 matrix.

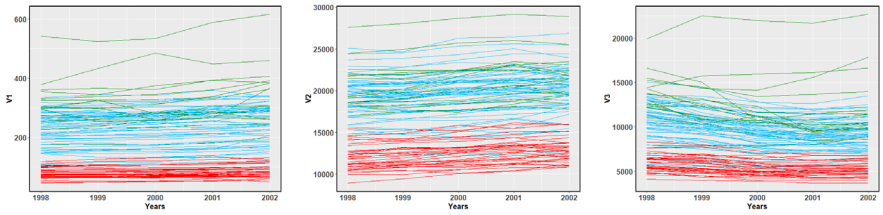
All the considered matrix-variate mixture models are then fitted to the data for $G \in \{1,2,3\}$, and for each model the Bayesian information criterion (BIC; Schwarz et al., 1978) is used to select G. With the formulation of the BIC used herein, the lowest its value, the better the model. The results are displayed in Table 1.

Table 1: Number of groups G selected by the BIC, along with the BIC values, for the considered matrix-variate mixture models.

Model	G	Value
MVN-Ms	3	20452.70
MSEN-Ms	2	20282.10
MVTIN-Ms	2	20274.66
MVt-Ms	2	20277.50

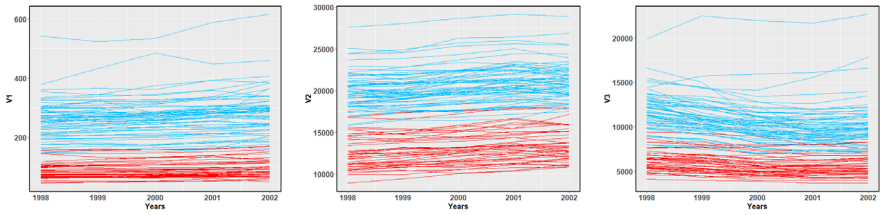
The first interesting result is that $G = 3$ for MVN-Ms, whereas for all the heavy-tailed mixtures $G = 2$. Furthermore, the MVN-Ms have the worst fitting performance, with a BIC value that is by far different from those of the other models. Relatedly, the BIC values of the heavy-tailed mixtures are close to each other, even if the best fitting model is the MVTIN-Ms. In addition, all the heavy-tailed mixtures produce the same data classification. A first comparison among the classifications produced can be done by looking at the parallel coordinate plots of the three variables, illustrated in Figure 1 for the MVN-Ms, and in Figure 2 for the MVTIN-Ms. In both figures, each color corresponds to a group, which are called Group 1 (red), Group 2 (cyan) and, only for MVN-Ms, Group 3 (green). As we can easily see in Figure 1, Group 3 is strongly overlapped to Group 2. A possible reason for this may be due to the tails of the matrix-normal distribution that are not heavy enough to model the observations which are relatively distant from the bulk of the data (especially for variables V1 and V3). Then, an additional mixture component is required. On the contrary, Group 3 is not present in Figure 2, where the two groups seem to have just a minimum level of overlap.

To better understand the classifications produced, a second graphical comparison is done by using the Italian political map. Specifically, in Figure 3 the Italian regions are bordered in yellow (islands excluded), while the internal provinces are delimited with the black lines and colored (as before) according to the estimated group memberships, both for MVN-Ms and MVTIN-Ms. Here, it is possible to see how Group 3 for MVN-Ms collects provinces spanning over several and different regions without a straightforward and reasonable justification. For example, the Lazio region in the Central Italy, have provinces belonging to all the three estimated groups, which is strange since they all share the same political and financial administration. Conversely, the two groups for MVTIN-Ms seem to almost perfectly divide Italy in two



(a) (b) (c)

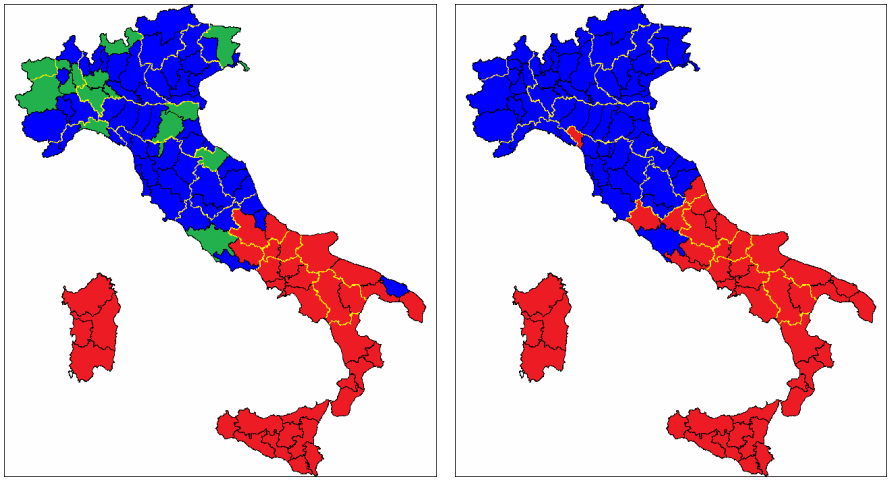
Fig. 1: Parallel coordinate plots for the best clustering solution according to the MVN-Ms.



(a) (b) (c)

Fig. 2: Parallel coordinate plots for the best clustering solution according to the MVTIN-Ms.

macro-regions, the Central-Northern Italy and the Southern-Insular Italy, which is in line with the findings of Millo and Carmeci (2011), where such separation is discussed. Furthermore, with the exclusion of three cases, all the provinces belonging to the same region are clustered together. The only exceptions concern the province of Rome (in the Lazio region), which due to its social-economic development is reasonably assigned to the Central-Northern Italy group, the province of Ascoli-Piceno (in the Marche region) and the province of Massa-Carrara (in the Toscana region). Therefore, considering the supporting literature and the interpretability of the results, it is reasonable to consider the classification produced by the MVTIN-Ms better than the one obtained via MVN-Ms.



(a) (b)

Fig. 3: Italian provinces colored according to the best clustering solutions for the MVN-Ms (a) and MTIN-Ms (b).

References

Brunello, G. and L. Cappellari (2008). The labour market effects of alma mater: Evidence from italy. *Economics of Education Review 27*(5), 564–574.

Doğru, F. Z., Y. M. Bulut, and O. Arslan (2016). Finite mixtures of matrix variate t distributions. *Gazi University Journal of Science 29*(2), 335–341.

González, S. (2011). The north/south divide in italy and england: Discursive construction of regional inequality. *European Urban and Regional Studies 18*(1), 62–76.

Millo, G. and G. Carmeci (2011). Non-life insurance consumption in italy: a sub-regional panel data analysis. *Journal of Geographical Systems 13*(3), 273–298.

Millo, G. and G. Piras (2012). splm: Spatial panel data models in R. *Journal of Statistical Software 47*(1), 1–38.

Misra, R. D. (1940). On the stability of crystal lattices. II. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 36, pp. 173–182. Cambridge University Press.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Tomarchio, S. D., A. Punzo, and L. Bagnato (2020). Two new matrix-variate distributions with application in model-based clustering. *Computational Statistics & Data Analysis 152*, 107050.

Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing 21*(4), 511–522.