ON MODEL-BASED CLUSTERING FOR EQUITABLE AND SUSTAINABLE WELL-BEING AT LOCAL LEVEL: HOW MANY ITALIES?

Natalia Golini[1], Francesca Martella[2] and Antonello Maruotti[3]

[1] Department of Economics and Statistics "Cognetti de Martiis", University of Turin, (e-mail: natalia.golini@unito.it

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

[3] Department of Economic, Political Sciences and Modern Languages, Libera Università Maria Ss. Assunta, (e-mail: a.maruotti@lumsa.it)

ABSTRACT: The choice of an appropriate number of clusters is a key issue in model-based clustering framework. The most popular approaches are based on the information criteria. However, often the latter may likely overestimate the number of clusters even though a good density estimation is possible. Here, we provide a dynamic model-based clustering approach to identify homogeneous Italian NUTS3 areas based on their equitable and sustainable well-being indicators from 2004 to 2019. In particular, the proposed model allows NUTS3 areas to move between clusters over time and a local dimensional reduction within each cluster. The empirical results show a high heterogeneity among the NUTS3 areas, leading to a high number of clusters. Possible strategies for merging similar NUTS3 clusters are investigated.

KEYWORDS: dimensionality reduction, dynamic clustering, hidden Markov model, longitudinal data

1 Introduction

In Italy, the National Institute of Statistics (Istat) has developed a multidimensional approach to measure "equitable and sustainable well-being" (BES), having the aim to integrate the traditional economic indicators with the quality of life of people, environment, inequality and sustainability measures. These indicators, updated annually since 2004, are declined into 12 relevant domains. Recently, Istat has designed a system of equitable and sustainable well-being indicators at NUTS3 level*, i.e. at the 107 Italian provinces, to deepen the

*NUTS: Nomenclature of Territorial Units for Statistics; NUTS 3: small regions for specific diagnoses (https://ec.europa.eu/eurostat/web/nuts/background).

knowledge of the well-being distribution across Italy to assess inequalities across areas. Local indicators are consistent with the national BES measures.

This paper addresses the complex, often non-linear, correlation between the indicators, the heterogeneity characterizing the Italian NUTS3 areas and changes and shifts in society over time under a unified framework. We identify homogeneous NUTS3 areas which behave in a lifestyle-similar fashion while keeping track of changes over time. We consider a clustering approach because more structured than a suitable standard approach in socio-economic analyses. To accommodate the multivariate longitudinal structure of the data, we propose a parsimonious hidden Markov model (HMM) that allows NUTS3 areas to transit between clusters, i.e. different well-being levels, over time. In this respect, a first-order finite-state Markov chain has been used to consider the temporal dependence. Moreover, a factor model framework is considered to capture correlation among indicators. And finally, we allow such correlations to vary across clusters and times to make the model flexible enough to capture the longitudinal structure of the data. The model parameters have been estimated through an Alternating Expected Conditional Maximization (AECM; Meng & van Dyk, 1997) algorithm.

2 Data and methods

2.1 Data description

The motivating dataset is composed of 102 NUTS3 areas and 18 well-being selected indicators, declined in 7 domains, to monitor their dynamics during the period $2004-2019$. This choice was made to consider the largest number of Italian NUTS3 areas without missing data during the observational period. Accordingly, four domains (Economic well-being, Social relationships, Landscape and cultural heritage and Innovation, research and creativity) and five NUTS3 areas (Barletta-Andria-Trani, Enna, Fermo, Monza e della Brianza and Sud Sardegna) are excluded from our analysis. The data are freely available at the Istat website[†]. A descriptive analysis confirms that socioeconomic divergence between the North and South of Italy continues, with the North more productive, rich and with a good health system, and the South/Islands, where the economy is mainly based on tourism, with higher unemployment

[†]`https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-at-local-level`.
This database contains data and metadata for the period $2004-2020$.

rates. Moreover, each BES indicator is related to others differently: the correlation structure is rather heterogeneous, and patterns of nonlinear correlation are present.

2.2 Parsimonious hidden Markov models for longitudinal data

We consider an HMM for multivariate longitudinal data allowing the density of the observed process to follow a factorial model. In detail, the model is defined by an observed process $\{\mathbf{Y}_{it}, i = 1,\dots,n; t = 1,\dots,T\}$ and a hidden-dependent process $\{S_{it}, i = 1,\dots,n; t = 1,\dots,T\}$ defined on the cluster space $\{1,\dots,K\}$ such that $\Pr(S_{it} \mid S_{i1},\dots,S_{it-1}) = \Pr(S_{it} \mid S_{it-1})$. Regarding the observed process $\mathbf{Y}_{it} = \{Y_{it1},\dots, Y_{itP}\}$, Y_{itp} represents the p-th response variable given by the i-th units at time t $(i = 1,\dots,n; p = 1,\dots,P; t = 1,\dots,T)$ such that $f(\mathbf{Y}_{it}|\mathbf{Y}_{i1},\dots,\mathbf{Y}_{iT},S_{i1},\dots,S_{iT}) = f(\mathbf{Y}_{it}|S_{it})$. Moreover, we defined the initial probabilities $\pi_k = Pr(S_{i1} = k)$ $(i = 1,...,n; k = 1,...,K)$ and the transition probability matrix $\Pi = \{\pi_{k|j}\}$, where $\pi_{k|j} = Pr(S_{it} = k \mid S_{it-1} = j)$ $(i = 1,...,n; t = 1,...,T,; j,k = 1,...,K)$. In line with the idea proposed by Maruotti *et al.*, 2017, we assume that conditionally to the k-th cluster, the random vector $\mathbf{Y}_{it}$ is described by:

$$\mathbf{Y}_{it} = \mu_k + \Lambda_k \mathbf{f}_{itk} + \mathbf{e}_{itk}, \tag{1}$$

where $\mathbf{f}_{itk}$ is a q-dimensional vector of cluster-specific factors drawn from $\mathrm{N}_P(\mathbf{0}, \mathbf{I}_q)$, and $\mathbf{e}_{itk}$ is a p-dimensional vector of cluster-specific error terms drawn from $\mathrm{N}_P(\mathbf{0}, \Psi_k)$, where $\Psi_k = \mathrm{diag}(\psi_{k1},\dots,\psi_{kP})$, which is assumed to be independent of $\mathbf{f}_{itk}$. In other words, a unit i in cluster k follows a multivariate Gaussian density with cluster-dependent mean vector μ_k and covariance matrix $\Lambda_k\Lambda_k' + \Psi_k$. Notice that, by constraining whether $\Lambda_k = \Lambda$, $\Psi_k = \Psi$ and $\Psi_k = \psi_k \mathbf{I}_p$, a family of 8 different models can be derived. To fit the proposed models, we use the AECM algorithm and recursions widely used in the HMM literature. The simulation studies results have shown a very good model performance in terms of the accuracy of the parameter estimates, degree of agreement between two partitions, and the ability to detect the correct number of clusters.

3 Empirical results

We computed Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) for each of the eight fitted models and combination (K, q). All the information criteria select the

unconstrained model (volumes, shapes, and orientations of the clusters are variable among clusters) as the best solution and, in particular, BIC and ICL recommend the solution with $K = 15$ and $q = 4$ as the most reasonable one balancing fit and parsimony. The main results can be briefly summarized as follows: Italian NUTS3 areas are heterogeneous; the inferred clustering structure identifies homogeneous, well-separated *spatial* aggregations of NUTS3 areas, leading to *four Italies*; persistence is the norm, transitions across clusters are rare but still present; cluster-specific correlations among indicators are effectively observed; each cluster is strongly characterized by only a subset of well-being indicators.

4 Assessing separation between NUTS3 clusters

The estimated results show that Italy still has a significant way to go in achieving well-being convergence. The heterogeneity across NUTS3 areas is still relevant, leading to a high number of clusters. However, the fact that some clusters differ only by the values of a few indicators suggests that there may be opportunities to merge similar clusters to get a more accurate overall picture of well-being in Italy and keep the specificities for further policymakers interventions. On the basis of the most widely used approaches in the field (see Hennig, 2010; Baudry *et al.*, 2010; Melnykov, 2016 among others), this opportunity is investigated.

References

BAUDRY, JEAN-PATRICK, RAFTERY, ADRIAN E, CELEUX, GILLES, LO, KENNETH, & GOTTARDO, RAPHAEL. 2010. Combining mixture components for clustering. *Journal of computational and graphical statistics*, **19**(2), 332–353.

HENNIG, CHRISTIAN. 2010. Methods for merging Gaussian mixture components. *Advances in data analysis and classification*, **4**, 3–34.

MARUOTTI, A., BULLA, J., LAGONA, F., PICONE, M., & MARTELLA, F. 2017. Dynamic Mixture of factor analyzers to characterize multivariate air pollutant exposures. *Ann. Appl. Stat.*, **11**(3), 1617–1648.

MELNYKOV, VOLODYMYR. 2016. Merging mixture components for clustering through pairwise overlap. *Journal of Computational and Graphical Statistics*, **25**(1), 66–90.

MENG, X. L., & VAN DYK, D. A. 1997. The EM Algorithm? An Old Folksong Sung to a Fast New Tune. *J. R. Statist. Soc. B*, **59**(3), 511–567.