CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE UNIVERSITY PRESS

ISSN 2704-601X (PRINT) - ISSN 2704-5846 (ONLINE)

SCIENTIFIC PROGRAM COMMITTEE

Giovanni C. Porzio (chair) (University of Cassino and Southern Lazio - Italy)

Silvia Bianconcini (University of Bologna - Italy)
Christophe Biernacki (University of Lille - France)
Paula Brito (University of Porto - Portugal)
Francesca Marta Lilja Di Lascio (Free University of Bozen-Bolzano - Italy)
Marco Di Marzio ("Gabriele d'Annunzio" University of Chieti-Pescara - Italy)
Alessio Farcomeni ("Tor Vergata" University of Rome - Italy)
Luca Frigau (University of Cagliari - Italy)
Luis Ángel García Escudero (University of Valladolid - Spain)
Bettina Grün (Vienna University of Economics and Business - Austria)
Salvatore Ingrassia (University of Catania - Italy)
Volodymyr Melnykov (University of Alabama - USA)
Brendan Murphy (University College Dublin -Ireland)
Maria Lucia Parrella (University of Salerno - Italy)
Carla Rampichini (University of Florence - Italy)
Monia Ranalli (Sapienza University of Rome - Italy)
J. Sunil Rao (University of Miami - USA)
Marco Riani (University of di Parma - Italy)
Nicola Salvati (University of Pisa - Italy)
Laura Maria Sangalli (Polytechnic University of Milan - Italy)
Bruno Scarpa (University of Padua - Italy)
Mariangela Sciandra (University of Palermo - Italy)
Luca Scrucca (University of Perugia - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)
Mariangela Zenga (University of Milan-Bicocca - Italy)

LOCAL PROGRAM COMMITTEE

Carla Rampichini (chair) (University of Florence - Italy)

Chiara Bocci (University of Florence - Italy)
Anna Gottard (University of Florence - Italy)
Leonardo Grilli (University of Florence - Italy)
Monia Lupparelli (University of Florence - Italy)
Maria Francesca Marino (University of Florence - Italy)
Agnese Panzera (University of Florence - Italy)
Emilia Rocco (University of Florence - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)

CLADAG 2021
BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE UNIVERSITY PRESS
2021

CLADAG 2021 BOOK OF ABSTRACTS AND SHORT PAPERS : 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021/ edited by Giovanni C. Porzio, Carla Rampichini, Chiara Bocci. — Firenze : Firenze University Press, 2021.
(Proceedings e report ; 128)

https://www.fupress.com/isbn/9788855183406

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-340-6 (PDF)
ISBN 978-88-5518-341-3 (XML)
DOI 10.36253/978-88-5518-340-6

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs
Front cover: Illustration of the statue by Giambologna, *Appennino* (1579-1580) by Anna Gottard

CLAssification and Data
Analysis Group (CLADAG)
of the Italian Statistical
Society (SIS)

FUP Best Practice in Scholarly Publishing (DOI https://doi.org/10.36253/fup_best_practice)
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Boards of the series. The works published are evaluated and approved by the Editorial Board of the publishing house, and must be compliant with the Peer review policy, the Open Access, Copyright and Licensing policy and the Publication Ethics and Complaint policy.

The online digital edition is published in Open Access on www.fupress.com.

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

This book is printed on acid-free paper
Printed in Italy

INDEX

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

Contributed Papers

MODEL SELECTION PROCEDURE FOR MIXTURE HIDDEN MARKOV MODELS

A. Abbruzzo [1], M.F. Cracolici [1] and F. Urso[1]

[1] Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy, (e-mail: antonino.abbruzzo@unipa.it, mariafrancesca.cracolici@unipa.it, furio.urso@unipa.it)

ABSTRACT: This paper proposes a model selection procedure to identify the number of clusters and hidden states in discrete Mixture Hidden Markov models (MHMMs). The model selection is based on a step-wise approach that uses, as score, information criteria and an entropy criterion. By means of a simulation study, we show that our procedure performs better than classical model selection methods in identifying the correct number of clusters and hidden states or an approximation of them.

KEYWORDS: model selection, clusters, hidden states, entropy-based scores, information criteria

1 Introduction

In many research fields, we deal with data whose independent units present one or more categorical sequences that represent the evolution of a specific feature over time (longitudinal data). Thus, it is necessary to define suitable methods capable of modelling an evolving process by describing some unknown variables that influence the observed sequences. Latent class models such as MHMMs can be used to analyse longitudinal data under the assumptions that (i) the sequences follows a latent Markov process and that (ii) the population is heterogeneous (Vermunt *et al.*, 2008; Bartolucci & Pandolfi, 2015). These models present two latent levels: one related to the hidden states of the discrete-time Markov chain and one representing the population's subgroups. The identification of the number of clusters and hidden states can be achieved, according to the literature on Mixture and Hidden Markov models, by fitting different models to the data and then selecting the model by using the results of information criteria (IC) such as AIC and BIC or classification criteria based on entropy (Dias *et al.*, 2009; Crayen *et al.*, 2012). However, these criteria tend to underestimate or overestimate these numbers (Wang & Chan, 2011). Here, we define a model selection procedure that combines IC and an entropy criterion to balance their limitations. Performing a simulation study, we show that

the proposed procedure exhibits promising results compared to the classical techniques.

2 Mixture Hidden Markov models

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})$ be the generic i-th sequence of length T with $\text{card}|Y_i| = R$, $U_i = (U_{i1}, U_{i2}, \ldots, U_{iT})$ the i-th hidden random vector with $\text{card}|U_i| = S$ and assume n independent sequences. Let $M = \{M^1, M^2, \ldots, M^K\}$ be a set of Hidden Markov Models, where $\Theta^k = \{\pi^k, A^k, B^k\}$ is the set of parameters for each sub-models M^k, related to each sub-population $k = 1, \ldots, K$. For each sequence Y_i, we define the prior cluster probabilities that the model parameters are the ones related to the k-th sub-model M^k as $P(M^k) = w_k$. Then, the log-likelihood is

$$\ell(\Theta; Y) = \sum_{i=1}^{n} \log P(Y_i|\Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_{ik} \sum_{u} \pi^k_{u_1} b^k_{u_1}(y_{i1}) \prod_{t=2}^{T} a^k_{u_{t-1},u_t} b^k_{u_t}(y_{it}) \right), \tag{1}$$

where the hidden state sequences $u = (u_1, u_2, \ldots, u_T)$ take all possible combinations of values in the hidden state space S and where y_{it} are the observations of subject i at time t, $\pi^k_{u_1} = P(u_1 = s|\Theta^k)$ with $s \in \{1, \ldots, S^k\}$ is the initial probability of the hidden state at time $t = 1$ in sequence u for cluster k; $a^k_{u_{t-1},u_t} = P(u_t = j|u_{t-1} = i, \Theta^k)$ with $i, j \in \{1, \ldots, S\}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t in cluster k; and $b^k_{u_t}(y_{it}) = P(y_{it} = r|u_t = s, \Theta^k)$ with $s \in \{1, \ldots, S\}$ and $r \in \{1, \ldots, R\}$ is the probability that the hidden state of subject i at time t emits the observed state at t in cluster k. Parameters can be estimated by means of the Expectation-Maximization; and the log-likelihood is calculated by using the forward-backward algorithm.

3 Proposed model selection procedure

Our proposed procedure combines IC and entropy for identifying MHMMs models on the basis of both goodness-of-fit and degree of class separation. Hence, the procedure consists of two stages. Firstly, we estimate models with different number of clusters and states, for each model the IC value is calculated and the models having these values below a predetermined threshold (the mean of the IC) are selected. At the second stage, an entropy criterion is used to identify among the models selected at the first-stage the one with the best degree of separation between classes (clusters and states). At the second stage,

when dealing with MHMMs, it is necessary to define a criterion that takes into account two levels of entropy: the first $\mathrm{En}_1(S)$ relating to the classification of observations in latent states and the second $\mathrm{En}_2(K)$ concerning the degree of separation between clusters.

$$\mathrm{E}_{new}(S,K) = 1 - \frac{1}{2n}\left[\frac{\mathrm{En}_1(S)}{T\log S} + \frac{\mathrm{En}_2(K)}{\log K}\right] \tag{2}$$

where

$$\mathrm{En}_1(S) = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{s=1}^{S_k} P(u_{it}=s|Y_i,M^k)\log P(u_{it}=s|Y_i,M^k),$$

$$\mathrm{En}_2(K) = \sum_{i=1}^{n}\sum_{k=1}^{K} P(M^k|Y_i)\log P(M^k|Y_i).$$

$P(M^k|Y_i)$ is the posterior probability that the given i-th observed sequence has been generated by the k-th model; $P(u_{it}=s|Y_i,M^k)$ is the posterior probability that the t-th element in the i-th hidden sequence takes the s-th hidden states given the observed sequence Y_i and that the sequence has been generated by the model related to the k-th cluster. The $S=\sum_{k=1}^{K} S^k$ is the total number of hidden states in all the K clusters. E_{new} takes value from 0 to 1. Values close to 1 are related to low entropy and a good degree of class separation, values close to 0 are related to a high entropy level and unreliable classification.

4 Simulation study

We compare our procedure of modeling selection to other methods such as AIC, BIC, sample-size adjusted BIC (ssBIC) through a Monte Carlo simulation study. We define 24 scenarios considering 4 models having different number of clusters K and latent states $(S^1,S^2,\ldots,S^K)$, by varying the number of sequences $n \in \{200,2000\}$ and the state-dependent conditional probabilities $b^k_{u_t}(y_{it})$ to represent low, medium, and high levels of uncertainty in hidden states classification of observations. We generate 100 longitudinal datasets for each scenario for a total of 2400 datasets, the analysis is carried out by using the *R* package "seqHMM" (Helske & Helske, 2017). In Table 1 we report methods' success rate for $n=2000$, where success means identifying a model having the correct number of clusters K, and number of hidden states equal to the exact number or one from this number. The last column report the results of our procedure considering the AIC as the IC used at the first stage as it showed better results than other IC.

classification uncertainty	$(S^1, S^2, \ldots, S^K)$	BIC	AIC	ssBIC	E_{new}	Our Procedure
LOW	$(2,3)$	**1.00** -	0.91 (0.029)	0.96 (0.020)	0.85 (0.036)	0.85 (0.036)
	$(2,2,3)$	0.62 (0.048)	0.40 (0.049)	0.58 (0.049)	0.62 (0.048)	**0.66** (0.047)
	$(2,2,3,3)$	0.30 (0.046)	0.37 (0.048)	0.34 (0.047)	0.21 (0.041)	**0.60** (0.049)
	$(2,2,3,3,2)$	0.19 (0.039)	0.38 (0.048)	0.24 (0.043)	0.19 (0.039)	**0.48** (0.050)
MEDIUM	$(2,3)$	**1.00** -	0.86 (0.035)	**1.00** -	0.49 (0.050)	0.73 (0.044)
	$(2,2,3)$	0.20 (0.040)	0.40 (0.049)	0.29 (0.045)	0.41 (0.049)	**0.59** (0.049)
	$(2,2,3,3)$	0.02 (0.014)	0.35 (0.048)	0.08 (0.027)	0.10 (0.042)	**0.53** (0.049)
	$(2,2,3,3,2)$	0.00 (0.000)	0.12 (0.032)	0.00 (0.000)	0.29 (0.045)	**0.31** (0.046)
HIGH	$(2,3)$	**1.00** -	0.58 (0.049)	**1.00** -	0.46 (0.050)	0.62 (0.048)
	$(2,2,3)$	0.10 (0.030)	0.40 (0.049)	0.14 (0.035)	0.42 (0.049)	**0.59** (0.049)
	$(2,2,3,3)$	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.10 (0.030)	**0.28** (0.045)
	$(2,2,3,3,2)$	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.19 (0.039)	**0.25** (0.043)

Table 1: Results of the Monte Carlo study for $n = 2000$. Low, medium and high level of uncertainty in hidden states classification scenario

As we can see, the proposed procedure has a better performance than the classic IC-based model selection methods when the number of clusters is $K > 2$. We also note how, unlike these methods, it is less affected by an increase in the uncertainty of hidden states' classification.

References

BARTOLUCCI, F., & PANDOLFI, S. 2015. LMest: Latent Markov Models with and without Covariates. *R package version.*, **2**.

CRAYEN, C., EID, M., LISCHETZKE, T., COURVOISIER, D. S., & VERMUNT, J. K. 2012. Exploring dynamics in mood regulation—mixture latent Markov modeling of ambulatory assessment data. *Psychosomatic medicine.*, **74**(4), 366–376.

DIAS, J. G., VERMUNT, J. K., & RAMOS, S. 2009. Mixture hidden Markov models in finance research. *Pages 451–459 of: Advances in data analysis, data handling and business intelligence*. Springer.

HELSKE, S., & HELSKE, J. 2017. Mixture hidden Markov models for sequence data: The seqHMM package in R.

VERMUNT, J. K., TRAN, B., & MAGIDSON, J. 2008. Latent class models in longitudinal research. *Handbook of longitudinal research: Design, measurement, and analysis*, 373–385.

WANG, M., & CHAN, D. 2011. Mixture latent Markov modeling: Identifying and predicting unobserved heterogeneity in longitudinal qualitative status change. *Organizational Research Methods*, **14**(3), 411–431.