# Hidden Markov and related discrete latent variable models: An application to compositional data

Francesco Bartolucci, Michael Greenacre, Silvia Pandolfi, and Fulvia Pennoni

**Abstract** We review the class of discrete latent variable models and we propose a new formulation of the hidden Markov model for compositional data. We illustrate some results of the analysis of the expenditures of the Spanish regions over several decades, showing that the approach is promising to cluster regions with different patterns linked to the composition of parts in the system over time. We give particular emphasis to the possible developments of discrete latent variable models that take inspiration from common problems of these models, such as the multimodality of the likelihood function and issues related to the choice of the number of support points of the latent variables.

## 1 Introduction

This chapter is focused on discrete latent variable (DLV) models, which include variables not directly observable and are assumed to follow a discrete distribution tailored to explain the association between observable variables [5]. Among these models, we mention in particular latent class [12], hidden Markov (HM) [3], and stochastic block models [15].

The inclusion in a statistical model of latent variables implies several advantages: (*i*) the capability of clustering units in different latent groups, also named components, subpopulations, classes or states, when the latent variables have a discrete distribution; (*ii*) the high degree of flexibility, so that complex dependence structures

Department of Economics
University of Perugia, Italy, e-mail: francesco.bartolucci@unipg.it

Department of Economics,
Universitat Pompeu Fabra, Spain, e-mail: michael.greenacre@upf.edu

Department of Economics
University of Perugia, Italy, e-mail: silvia.pandolfi@unipg.it

Department of Statistics and Quantitative Methods
University of Milano-Bicocca, Italy, e-mail: fulvia.pennoni@unimib.it

may be dealt with also in the presence of variables of a different nature; (*iii*) the straightforward interpretation when the latent variables correspond to explanatory factors that cannot be observed; and (*iv*) the possibility, under certain conditions, to interpret some parameters causally, as the model can consider certain forms of confounding. Specifically, we can account for the effect of unobservable variables affecting both the treatment and the response by latent variables having a suitable distribution. Parameter estimation is based on the maximum likelihood approach through the expectation-maximization (EM [10]) algorithm.

Our first aim is to illustrate new possible applications of these models for compositional data [13], which are nonnegative multivariate observations where relative rather than absolute information is relevant; thus, they represent a quantitative measurement of the parts of some total, expressed as proportions summing to 1. Secondly, we aim to discuss general issues, suggest possible solutions, and highlight new research frontiers regarding these models.

The chapter is organized as follows. First, we introduce a new development of the HM model for compositional data. Second, we show an application to analyze the annual capital stock wealth amount of different sectors in Spain. Finally, we conclude with some remarks on future directions of this research.

## 2 Hidden Markov model for compositional data

HM models find application in the analysis of time-series [23] and longitudinal data [3]. These models may be employed for clustering units in a dynamic fashion, where the same individual may move between clusters across time. In the context of longitudinal data, for a sample of $n$ individuals observed over $T$ time occasions we denote the vectors of response variables by $\boldsymbol{Y}_{it}$ with elements $Y_{ijt}$, $i = 1,\ldots,n$, $j = 1,\ldots,r$, $t = 1,\ldots,T$, where $r$ is the number of such variables. For every $i$, the sequence of discrete latent variables $U_{it}$ collected in the vector $\boldsymbol{U}_i$ is introduced; these variables are assumed to follow a Markov chain of first-order with $k$ states.

The HM model assumes that each time-specific vector of response variables $\boldsymbol{y}_{it}$, corresponding to parts of the composition that are proportions summing to 1, follows a Dirichlet distribution, although other choices are also possible [21]. The parameters of this distribution depend on the underlying discrete latent variable; in symbols, we have $\boldsymbol{Y}_{it}|U_{it} = u \sim \text{Dir}(\boldsymbol{\alpha}_u)$, where $\boldsymbol{\alpha}_u$ is the state-specific vector of parameters with $u = 1,\ldots,k$.

As usual, each sequence of latent variables $U_{i1},\ldots,U_{iT}$ follows a Markov chain with initial and transition probabilities that, without covariates, are denoted by $\lambda_u = p(U_{i1} = u)$ and $\pi_{uv} = p(U_{it} = v|U_{i,t-1} = u)$, $t = 2,\ldots,T$, $u,v = 1,\ldots,k$, respectively. With unit-specific covariates, these probabilities are formulated by suitable logit parametrizations based on regression coefficients to account for the effect of such covariates. This formulation is based on the usual assumption that the response variables are conditionally independent given the latent variables, which is a form of *local independence* [3].

Regarding the parametrization of the Dirichlet distribution, we follow an approach that distinguishes the effects of the latent states on the expected value and on the variance (see also [16]). This amounts to normalize the parameters of the Dirichlet distribution so that their sum is equal to 1. Additionally, we introduce a new scale parameter corresponding to the sum of the parameters used in the traditional formulation of this distribution.

Assuming independence between subjects, the log-likelihood referred to the observed data can be written as $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i)$, where $\boldsymbol{\theta}$ is the vector of all model parameters, and $f(\boldsymbol{y}_i)$ is the manifest distribution of the observed compositional parts. The latter may be expressed as

$$f(\boldsymbol{y}_i) = \sum_{u_1=1}^{k} \sum_{u_2=1}^{k} \cdots \sum_{u_T=1}^{k} \lambda_{u_1} \left( \prod_{t=2}^{T} \pi_{u_{t-1}u_t} \right) \left[ \prod_{t=1}^{T} f(\boldsymbol{y}_{it} | U_{it} = u_t) \right].$$

In order to maximize $\ell(\boldsymbol{\theta})$, the EM algorithm relies on the complete-data log-likelihood expressed as $\ell^*(\boldsymbol{\theta})$; at the *E-step*, the algorithm computes the posterior expected value of $\ell^*(\boldsymbol{\theta})$ using certain forward-backward recursions [6, 22], and at the *M-step* it updates the model parameters. Initialization of the algorithm is a crucial aspect and requires careful consideration of different starting values. Model selection is related to the choice of the number of states when it is not known *a priori*, and it is usually performed according to the information criteria such as the Bayesian Information Criterion (BIC) [19]; see [1] among others. Suitable functions are implemented for the R environment extending those of the package `LMest` [4] to perform the estimation of the proposed HM model.

## 3 Analysis of spatio-temporal compositional data

We consider data provided by BBVA Foundation from Madrid[1] concerning the composition of capital stock wealth in different sectors of the Spanish economy. Data are recorded for an extended period from 1964 to 2019 [11]. In the present work, we mainly concentrate on the problem of the national data on the temporal scale, and we briefly mention how to broaden this to the more detailed spatio-temporal scale across the different autonomous regions of Spain.

For the spatio-temporal framework, the data are collected in vectors $\boldsymbol{y}_{it}$, $i = 1,\ldots,n$, $t = 1,\ldots,T$, where $n$ is the number of regions. The changing total amount across the years is, of course, important to analyze. Still, here it is the changing composition of the capital stock wealth that is of interest, namely the amounts of each year relative to their respective totals. Hence, compositional data are such that the sum of the elements of each compositional response vector is fixed at 1 or 100%. This has crucial implications in terms of data analysis. The $r = 15$ different sectors are considered for $T = 56$ years referred to the 17 autonomous communities and to the 2 autonomous cities (included together) of Spain. Trajectories of overall capi-

[1] https://www.fbbva.es/en/

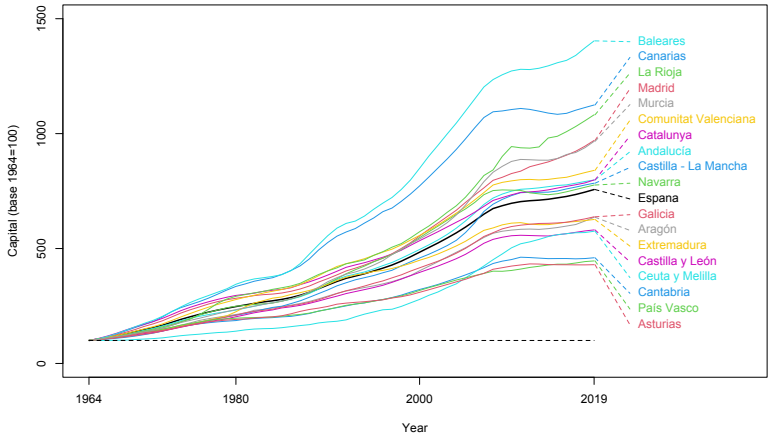


Fig. 1: *Trajectories of capital stock wealth, from 1964 to 2019, across the 18 autonomous communities of Spain, along with the trajectory for Spain (España) itself. To facilitate comparison, all amounts are set to 100 at the initial time point. Dashed lines are for indicating labels*

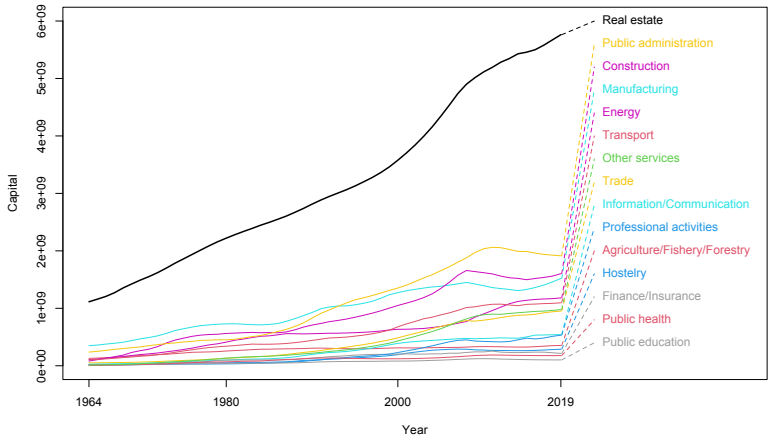


Fig. 2: *Trajectories of capital stock wealth, measured in thousands of euros, from 1964 to 2019, for the whole of Spain*

Table 1: *Number of states, maximum likelihood, number of parameters, and BIC index for HM models*

| $k$ | $\ell(\hat{\boldsymbol{\theta}})$ | #par | *BIC* |
|---|---|---|---|
| 1 | 36,430.97 | 15 | -72,818.59 |
| 2 | 38,229.40 | 33 | -76,363.43 |
| 3 | 39,835.10 | 53 | -79,517.02 |
| 4 | 40,533.36 | 75 | -80,849.94 |
| 5 | 41,294.41 | 99 | -82,302.67 |
| 6 | 42,068.90 | 125 | -83,776.51 |

tal stock wealth by Spanish regions are reported in Figure 1, while those of capital stock wealth by sector are depicted in Figure 2.

The nature of the data suggests basing the analysis on region-year compositions across sectors [13]; therefore, capital stock wealth amounts across sectors are divided by their total so as to fix the sum to 1. For this analysis, we adopt the model illustrated in Section 2. To perform model selection, this model is estimated for $k = 1, \ldots, 6$, where 6 is the maximum number of states that we consider reasonable in order to avoid an excessive model complexity. Results in terms of maximum log-likelihood, number of parameters, and corresponding values of BIC are shown in Table 1. A model with $k = 5$ latent states is adopted for interpretability (despite not being optimal in terms of BIC); this approach is usually adopted in applications to complex and high-dimensional data, where this index may continue to decrease as additional states are added until a very large value of $k$. In such a situation, it is advisable to choose accounting the value of $k$ also for the interpretability of the latent states.

Estimated means of the compositions given the latent state are visualized through the heat map depicted in Figure 3. The latent states are ordered increasingly based on the average proportion of capital stock wealth allocated to real estate, with regions in the first subpopulation characterized by the lowest part of capital stock wealth devoted to this sector. On the other hand, it is not possible to detect a clear ordering of the latent states with respect to the composition of capital invested in the other sectors and then a specific interpretation is in order for each of these subpopulations. Estimates of the initial and transition probabilities are shown in Table 2. It is noteworthy that the majority of regions belong to the last class, characterized by the highest proportion of capital stock wealth in real estate, while no regions are initially assigned to the first three classes. Furthermore, regarding transition probabilities, there is a very high persistence in the same subpopulation. However, it can be observed that the lower triangular elements of the transition matrix in Table 2 are generally larger than the upper triangular elements, which suggests a pattern where some regions tend to transition from the last two classes to the first four classes.

As already mentioned, the approach also allows assigning every region to a specific latent cluster over time, permitting the creation of maps to analyze the spatial pattern of the phenomenon under study. Just to give an idea, in Figure 4 we report these maps of the first and last available years.

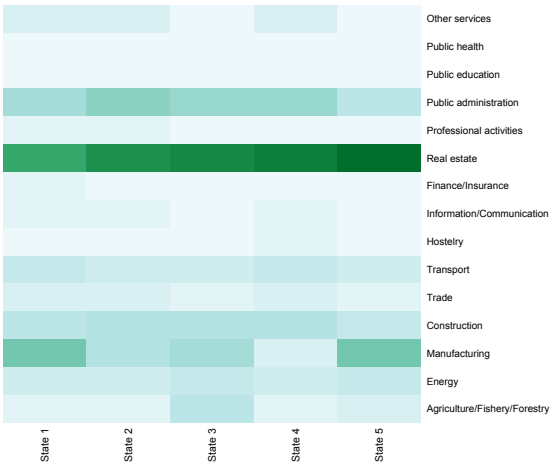


Fig. 3: *Heatmap of scaled cluster means under the HM model*

Table 2: *Estimated initial and transition probabilities under the HM model*

| Initial probabilities | | | | |
|---|---|---|---|---|
| $u=1$ | $u=2$ | $u=3$ | $u=4$ | $u=5$ |
| 0.000 | 0.000 | 0.000 | 0.221 | 0.779 |

| Transition probabilities | | | | | |
|---|---|---|---|---|---|
| $u$ | $v=1$ | $v=2$ | $v=3$ | $v=4$ | $v=5$ |
| 1 | 0.966 | 0.028 | 0.000 | 0.007 | 0.000 |
| 2 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| 3 | 0.007 | 0.035 | 0.958 | 0.000 | 0.000 |
| 4 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 |
| 5 | 0.027 | 0.004 | 0.027 | 0.005 | 0.937 |

## 4 Conclusions and further developments

In this final section, we mention some issues that are important to overcome the actual limitations of the discrete latent variable (DLV) models. First, even for relatively simple DLV models, the likelihood is typically multimodal, with a potentially huge number of modes. As a possible approach, we can consider choosing appropriate starting values for the estimation algorithm on the basis of deterministic and stochastic random rules [17], and then selecting the best solution among different attempts. Alternatively, we can use new tempered versions of the expectation-maximization (EM) algorithm, which consist in re-scaling the objective function depending on a

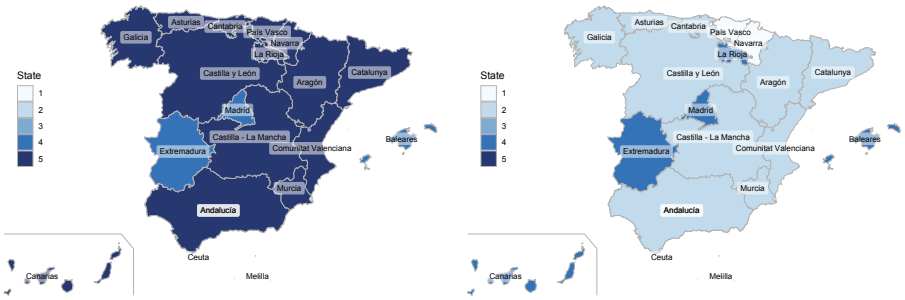


Fig. 4: *Regional maps showing the estimated distribution of the latent clusters built by the selected HM model for years 1964 and 2019*

variable, known as temperature, controlling in such a way the prominence of global and local maxima [8]; more recent advanced versions of the EM are also promising at this aim, such as those based on evolutionary algorithms [9].

Second, it is uncommon that in applications clearly separated clusters of units indeed exist. Therefore, the adopted DLV model may identify "spurious" clusters and model selection criteria may suggest many components with problematic interpretations. As a solution, a latent structure comprising a mixture of continuous distributions can be used, allowing the resulting model to serve as a compromise between discrete and continuous latent variable models [20]. Moreover, we can adopt model selection criteria that avoid a huge number of clusters and focus on the quality of clustering, such as the normalized entropy criterion [7].

Furthermore, accounting for informative missing data and dropout, especially in longitudinal data analysis, may be of primary interest. In this regard, we can consider implementing a joint or shared parameter model where the same latent variables govern the longitudinal process and time to dropout [2]. We can also consider an extra absorbing state and joint modeling missing data as proposed in [18].

Finally, specifically regarding the proposed application illustrated above, it is worth mentioning that as an alternative to the Dirichlet distribution, the HM models can be applied to the compositional data transformed to logratios [14], which are then treated as interval-scale variables with a multivariate normal distribution. These results can be compared with those obtained in the present study. Furthermore, it is also possible to extend the model by considering a spatio-temporal structure, where the latent state of a region in a certain year may depend not only on the previous state but also on the state of neighboring regions.

**Acknowledgements** F. Bartolucci, S. Pandolfi and F. Pennoni acknowledge the financial support from the grant "*Hidden Markov Models for Early Warning Systems*" of Ministero dell'Università e della Ricerca (PRIN 2022TZEXKF) funded by European Union - Next Generation EU.

## References

1. Bacci S, Pandolfi S, Pennoni F. 2014. A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification* 8:125-145
2. Bartolucci F, Farcomeni A. 2015. A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics* 71:80-89
3. Bartolucci F, Farcomeni A, Pennoni F. 2013. *Latent Markov Models for Longitudinal Data.* Boca Raton, FL: Chapman & Hall/CRC Press
4. Bartolucci F, Pandolfi S, Pennoni F. 2017. LMest: An R package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software* 81:1-38
5. Bartolucci F, Pandolfi S, Pennoni F. 2022. Discrete latent variable models. *Annual Review of Statistics and Its Application* 9:425-452
6. Baum LE, Petrie T, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* 41:164-171
7. Biernacki C, Celeux G, Govaert G. 1999. An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters* 20:267-272
8. Brusa L, Bartolucci F, Pennoni F. 2023. Tempered expectation-maximization algorithm for the estimation of discrete latent variable models, *Computational Statistics* 38:1391-1424
9. Brusa L, Pennoni F., Bartolucci F. 2024. Maximum likelihood estimation for discrete latent variable models via evolutionary algorithms, *Statistics and Computing* 34:1-15
10. Dempster A, Laird N, Rubin D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39:1-22
11. García F, Ivars M, Radoselovics J, Candau E, Domínguez J. 2023, El stock de capital en España y sus comunidades autónomas: Análisis de los cambios en la composición de la inversión y las dotaciones de capital entre 1995 y 2022. *Documentos De Trabajo, Fundación BBVA*
12. Goodman L. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215-231
13. Greenacre M. 2019. *Compositional Data Analysis in Practice.* Boca Raton, FL: Chapman & Hall/CRC Press
14. Greenacre, M. 2021. Compositional data analysis. *Annual Review of Statistics and Its Application* 8:271-299
15. Holland PW, Laskey KB, Leinhardt S. 1983. Stochastic blockmodels: First steps. *Social Networks* 5:109-137
16. Maier, M. 2014. DirichletReg: Dirichlet regression for compositional data in R. *Research Report Series / Department of Statistics and Mathematics No. 125*
17. Maruotti A, Punzo A. 2021. Initialization of hidden Markov and semi-Markov models: A critical evaluation of several strategies. *International Statistical Review* 89:447-480
18. Pandolfi S, Bartolucci F, Pennoni F. 2023. A hidden Markov model for continuous longitudinal data with missing responses and dropout. *Biometrical Journal* 65:1-25
19. Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6:461–464
20. Skrondal A, Rabe-Hesketh S. 2007. Latent variable modelling: A survey. *Scandinavian Journal of Statistics* 34:712-745
21. Smithson M, Shou Y. 1978. Flexible CDF-quantile distributions on the closed unit interval, with software and applications. *Communications in Statistics, Theory and Methods* 11:3876-3898
22. Welch LR. 2003. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter* 53:10-13
23. Zucchini W, MacDonald IL, Langrock R. 2017. *Hidden Markov Models for Time Series: An Introduction Using R.* Boca Raton, FL: CRC press