DISCRETE LATENT VARIABLE MODELS: RECENT ADVANCES AND PERSPECTIVES

Francesco Bartolucci[1], Michael Greenacre[2],
Silvia Pandolfi[1] and Fulvia Pennoni[3]

[1] Department of Economics, University of Perugia, IT (e-mail: francesco.bartolucci@unipg.it, silvia.pandolfi@unipg.it)

[2] Department of Economics, Universitat Pompeu Fabra, ES (e-mail: michael.greenacre@upf.edu)

[3] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, IT (e-mail: fulvia.pennoni@unimib.it)

ABSTRACT: After a review of the class of discrete latent variable models in terms of formulation and estimation methods, recent advances and perspectives regarding these models are illustrated. We consider in detail the stochastic block model for social networks and models for spatio-temporal data. Among these developments, we discuss, in particular, the analysis of longitudinal compositional data about expenditures of the Spanish regions over several decades.

KEYWORDS: Compositional data, data augmentation, expectation-maximization algorithm, spatio-temporal modeling, variational inference.

1 Introduction

In general terms, latent variable models include variables not directly observable to describe the relation between observable variables. Among these models, those based on the assumption that the latent variables follow a discrete distribution, namely discrete latent variable (DLV) models, are nowadays commonly used (for a recent review, see Bartolucci *et al.*, 2022). With respect to models based on continuous latent variables, DLV models present some advantages, such as the flexibility and capability of clustering units in different latent groups, also named components, classes, or states. Obviously, there are also issues that may complicate the use of DLV models such as the selection of the number of support points of the discrete distribution of the latent variables and the multimodality of the likelihood function.

The first aim of this work is to provide a critical review of DLV models in terms of formulation and estimation methods. Regarding the first aspect, we describe recent proposals that can be used to deal with complex data struc-

tures such as social networks and spatio-temporal data. In particular, for the analysis of social networks we consider the stochastic block model and its extended versions that may be used in a longitudinal context where individuals are repeatedly observed in terms of social behavior. For the analysis of spatio-temporal data, we illustrate models based on latent variables which are specific to each site and time of observation. We also consider recent formulations which may be used to make causal inference on a certain policy or treatment and that conceive potential versions of the latent variables to properly define causal effects (Lanza *et al.*, 2013).

Regarding estimation, we show that both frequentist and Bayesian inferential approaches rely either on methods that directly assign the units to the different components or methods in which this explicit assignment is avoided. Among the methods of the first type, it is worth recalling those based on the maximization, with respect to the model parameters and the assignment of units to the components, of the so-called classification likelihood and the corresponding Bayesian methods based on Markov chain Monte Carlo (MCMC) algorithms (Gelman *et al.*, 2011) with data augmentation, where the latent variables are considered on the same footing as the model parameters. Estimation methods of the second type are instead based on popular algorithms such as the expectation-maximization (EM Dempster *et al.*, 1977) applied to find the maximum likelihood estimate of the parameters and corresponding MCMC algorithms for Bayesian inference. We also describe variational methods (see, among others, Daudin *et al.*, 2008), used for complex contexts, and in general we pay attention to the problem of scalability (Bartolucci *et al.*, 2018).

The second aim of the present work is to illustrate a new possible application of the DLV models to the analysis of temporal and spatio-temporal compositional data, as is briefly described in the following section.

2 Analysis of spatio-temporal compositional data

This development is motivated by the availability of a recent dataset about the composition of the annual investments in different sectors of the Spanish economy, for a long period that goes from 1964 to 2020 (García *et al.*, 2023). In the present work we concentrate mainly on the simpler problem of the national data on the temporal scale, mentioning later how to broaden this to the more detailed spatio-temporal scale across the different autonomous regions of Spain. The data are thus collected in the $m \times 1$ vectors $\mathbf{y}_t$, $t = 1, \ldots, T$, where T is the number of time occasions and m the number of sectors. For the spatio-temporal framework, the data would be in vectors $\mathbf{y}_{it}$, $i = 1, \ldots, n$, $t = 1, \ldots, T$,

where n is the number of regions. The changing total amount invested across the years is, of course, important to analyze, but here it is the changing composition of the investments that is of interest, namely the amounts invested 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% (see Greenacre, 2021, for a recent review). This has crucial implications in terms of data analysis. Two approaches are presented here: first, an exploratory approach where the logratio transformation is used (Greenacre, 2018); and second, where the data are assumed to follow the Dirichlet distribution on the unit interval. For the logratio approach the simplest transformation is the so-called additive logratio transformation, where all compositional parts are expressed as a ratio with a fixed part, and then log-transformed. These transformed data can then be analyzed using existing approaches for multivariate interval-scale data, assuming multivariate normal distribution.

For the regional data at hand we formulate different models. The starting one is of hidden Markov type and does not account for the spatial dependence between the regions. It only accounts for temporal dependence. For every region, this model assumes that each time-specific vector of response variables $\mathbf{y}_{it}$, corresponding to parts of the composition, follows a Dirichlet distribution with parameters that depend on an underlying discrete latent variable. In symbols, we have

$$\mathbf{Y}_{it}|U_{it} = u \sim \text{Dir}(\boldsymbol{\alpha}_u),$$

where U_{it} is the underlying latent variable having support $\{1,\ldots,k\}$ and $\boldsymbol{\alpha}_u$ is the state-specific vector of parameters.

Moreover, each sequence of latent variables $U_{i1},\ldots,U_{iT}$ follows a Markov chain with initial probabilities and transition probabilities that, without covariates, are denoted by $\lambda_u = p(U_{i1} = u)$ and $\pi_{u|\bar{u}} = p(U_{it} = u|U_{i,t-1} = \bar{u})$, $t = 2,\ldots,T$. 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. Regarding the parametrization of the Dirichlet distribution, we follow an approach that separates the effects of the latent states on the expected value and on the variance (see also Maier, 2014).

We also consider a spatio-temporal model where, following recent approaches (e.g., Bartolucci & Farcomeni, 2022), the latent state of a region in a certain year may depend, not only on the previous state, but also on the state of the neighbor regions. More precisely, each latent variable U_{it} is modeled conditionally on $U_{i,t-1}$ and U_{jt}, $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the set of neighbors of region

i. Even in this case, multinomial logit parametrizations are adopted to include the effect of possible covariates. Again, we rely on the assumption of conditional independence between the response vectors given the latent variables that has an interesting interpretation and simplifies the estimation process.

References

BARTOLUCCI, F., & FARCOMENI, A. 2022. A hidden Markov space–time model for mapping the dynamics of global access to food. *Journal of the Royal Statistical Society, Series A*, **185**, 246–266.

BARTOLUCCI, F., BACCI, S., & MIRA, A. 2018. On the role of latent variable models in the era of big data. *Statistics & Probability Letters*, **136**, 165–169.

BARTOLUCCI, F., PANDOLFI, S., & PENNONI, F. 2022. Discrete latent variable models. *Annual Review of Statistics and Its Application*, **9**, 425–452.

DAUDIN, J.-J., PICARD, F., & ROBIN, S. 2008. A mixture model for random graphs. *Statistics and Computing*, **18**, 173–183.

DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**, 1–22.

GARCÍA, F. P., IVARS, M. M., RADOSELOVICS, J. F. G., CANDAU, E. B., & DOMÍNGUEZ, J. C. R. 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*.

GELMAN, A., JONES, A., & MENG, X. L. 2011. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: CRC Press.

GREENACRE, M. 2018. *Compositional Data Analysis in Practice*. Boca Raton, FL: CRC Press.

GREENACRE, M. 2021. Compositional data analysis. *Annual Review of Statistics and its Application*, **8**, 271–299.

LANZA, S. T., COFFMAN, D. L., & XU, S. 2013. Causal inference in latent class analysis. *Structural Equation Modeling*, **20**, 361–383.

MAIER, M. 2014. DirichletReg: Dirichlet regression for compositional data in R. *Institute for Statistics and Mathematics, Report 125*.