METRON (2015) 73:151–154
DOI 10.1007/s40300-015-0074-3

EDITORIAL

Latent variable models for the analysis of socio-economic data

Marco Alfó[1] · **Francesco Bartolucci**[2]

Published online: 20 August 2015

The present Special Issue is intended to cover statistical approaches based either on continuous or discrete latent variable models, with a particular focus on approaches combining the use of both types of latent variable within the same model. The aim is to give a (partial) picture of recent developments about these models and their applications in the social and economic sciences.

As it is well known, the term "latent variables" refers to variables that are not directly observable but are assumed to affect the observable ones in different ways. Latent variables are typically included in a statistical model to make it more flexible and, in particular, for:

- representing individual characteristics that cannot be directly measured (e.g., intelligence, satisfaction),
- accounting for unobserved heterogeneity (i.e., differences between units that cannot be explained on the basis of observable covariates only),
- modeling the dependence in hierarchically structured data: examples are clustered, longitudinal, and/or multivariate data where the association structure cannot be described by a simple multivariate distribution,
- modeling or accounting for measurement errors, such as in error-in-variables models,
- describing a formal grouping structure, as in model-based clustering.

These approaches have been developed and applied in different fields, such as Economics, Medicine, Psychology, Sociology, and include, as an important subclass, finite mixture models, where the latent variable(s) are assumed to be discrete. For an overview see McLachlan and Peel [16], Skrondal and Rabe-Hesketh [19], McCulloch et al. [15], Bartholomew et al. [2], and Bartolucci et al. [3], among others. A thoughtful essay on mixture models, with insights on inference, geometry, and applications is given by Lindsay [13].

✉ Marco Alfó
marco.alfo@uniroma1.it

Francesco Bartolucci
francesco.bartolucci@unipg.it

[1] Dipartimento di Scienze Statistiche, Sapienza Universitá di Roma, Rome, Italy

[2] Dipartimento di Economia, Universitá degli Studi di Perugia, Perugia, Italy

In the following, we provide a brief introduction to the articles included in this Special Issue. These articles cover different areas, such as model-based clustering, analysis of longitudinal data, capture–recapture studies, just to mention a few.

When discrete latent variables are considered for model-based clustering or latent class analysis, identifiability is one of the main issues together with the choice of the number of components, that is, categories of the discrete latent variable included in the model. Aitkin et al. [1] approach the problem by relying on posterior deviance distributions, based either on non-informative or diffuse priors. They illustrate the proposed approach by re-analyzing the galaxy data of Roeder [18], the psychiatric data of Berkhof et al. [5], and through a simulation study. Bertoletti et al. [6] discuss the same problem by adopting the Integrated Completed Likelihood (ICL) criterion of Biernacki et al. [7,8]. They show how, through the use of conjugate priors, an exact expression for ICL can be derived, avoiding any approximation. The approach is based on a greedy algorithm and is detailed for the special case of finite mixtures of multivariate Gaussian distributions. The issue of identifiability of mixtures of discrete probability distributions of the power series family is discussed in Böhning [9]. The author develops ideas based on the ratios of neighboring probabilities from such a family, leading to the so-called *ratio plot*, a graph which could be used as a diagnostic device to detect departures from homogeneity. He further explores the negative binomial and the beta binomial as models for zero-truncated count data arising from capture–recapture experiments, concluding that these models frequently suffer from boundary problems that may make them unreliable.

In some cases, the postulated model can be too complex to be estimated by a standard maximum likelihood approach. An instance is when, in analyzing multivariate data, profile-specific latent variables having a non-trivial distribution are introduced to model the association between outcomes. The likelihood function of the resulting model is rarely available in closed form and approximation techniques need to be employed to solve the high-dimensional integrals that are involved. To make estimation feasible in a reasonable amount of time, different approaches have been proposed in the literature. Florios et al. [11] discuss methods based on composite likelihood, where the *genuine* likelihood is replaced by a function based on lower order (marginal or conditional) log-likelihoods that involve only low-dimensional integrals. By a simulation study, the authors also show that a modified version of the estimator recently proposed by Vasdekis et al. [20] can lead to simpler computations and to a comparably high performance.

In the last decade, generalizations of mixed-parameters models to deal with data containing outliers have received an increasing interest; the same may be said for formulations of these models to be used when the main focus in not on the expected value of the conditional response distribution, given the covariates and the latent variables, but on different summary statistics of this distribution. In these cases, the quantile regression approach [12] has been advocated as a useful tool. In the past few years, several approaches have been presented in the literature proposing either marginal or conditional (on latent variables) quantile regression models for longitudinal data. An interesting overview of such approaches is provided by Marino and Farcomeni [14]; the authors give a detailed description of models that have appeared in both the econometric and the statistical literature. Relevant issues that are still open or have received only a limited attention in the past are also discussed, with the aim of providing the reader with a comprehensive review of currently developing topics in quantile regression.

The Special Issue also includes interesting papers that have a more applied cut. Bassi and Scarpa [4] discuss an application of models based on both continuous and discrete latent variables to analyze complex longitudinal data with a three-level hierarchy, entailing longitudinal

markers of day specific fertility. In particular, the heterogeneity between women and between cycles of the same woman is accounted for by using a multilevel latent class model for the discrete indicators which summarize the daily observations of cycle developments reported by the women participating in the study. Furthermore, a growth mixture model is employed to describe the evolution of the fertile phase of the women's cycle over time. Francis and Liu [10] discuss the use of latent variable models to assess escalation in crime seriousness. They propose a comparison of different methods, namely mixed-effects models, group-based trajectory models (which can be considered as a particular finite mixture model specification), and growth mixture models, where specific components of the two former approaches are joined together to create a very flexible model. The comparison is based on the analysis of a dataset concerning all England and Wales offenders born in 1953 and followed through 1999. Based on goodness-of-fit to the observed data, the authors show that growth mixture models outperform group-based trajectory models. It is also interesting to note that the authors show that R is the best software environment for modeling purposes in this field. Paas et al. [17] discuss multilevel latent class models for the segmentation of financial product portfolios with a longitudinal perspective. The authors analyze a dataset on consumer financial product portfolios across 14 EU countries based on three disaggregate cross-sectional databases for the years 1969, 1990, and 2003. In this way they overcome the issue of gathering panel data for such a long period of time for these countries. This approach makes it possible to evaluate similarities in segmentation structures across countries at different time-points using multiple cross-sectional datasets, providing indications about the development of consumer behavior over time, and identifying differences between countries.

We thank all the authors that have contributed to the realization of this Special Issue of *Metron*. Their enthusiasm has pushed us forward during the whole period we have spent handling the manuscripts and curating the Special Issue. Warm thanks go also to the reviewers, who provided the authors with valuable comments and suggestions, and have greatly contributed to the overall quality of this volume. A special thank goes to Giovanni Maria Giorgi, editor-in-chief of *Metron*, who has invited us and supported us in acting as guest editors. Thanks are also due to the editorial board and the staff of the journal for giving us the chance of putting together this piece of work.

To conclude, we would like to share a personal thought. While we were working on the realization of this volume, Bruce George Lindsay passed away; he died on May 5th, 2015. Recalling his huge research contributions and their impact on the literature on latent variable and mixture models would require a whole Special Issue in itself. Here, we want to publicly thank him for sharing his time and ideas with many of us, for being an incentive to put forward our research, and for writing many seminal papers that have represented, represent, and will represent milestones for our daily work. We are sure that all the authors, the reviewers, and the members of the editorial board share the same sentiment.

References

1. Aitkin, M.A., Vu, D., Francis, B.: A new Bayesian approach for determining the umber of components in a finite mixture. Metron (2015). doi:10.1007/s40300-015-0068-1
2. Bartholomew, D., Knott, M., Moustaki, I.: Latent Variable Models and Factor Analysis, 3rd edn. Wiley, New York (2011)
3. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. CRC Press, Boca Raton (2013)
4. Bassi, F., Scarpa, B.: Latent class modeling of markers of day-specific fertility. Metron (2015). doi:10.1007/s40300-015-0066-3

5. Berkhof, J., van Mechelen, I., Gelman, A.: A Bayesian approach to the selection and testing of mixture models. Statist. Sin. **13**, 423–442 (2003)
6. Bertoletti, M., Friel, N., Rastelli, R.: Choosing the number of clusters in a finite mixture models using an exact integrated completed likelihood criterion. Metron (2015). doi:10.1007/s40300-015-0064-5
7. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 719–725 (2000)
8. Biernacki, C., Celeux, G., Govaert, G.: Exact and Monte Carlo calculation of integrated likelihoods for the latent class models. J. Stat. Plan. Inference **149**, 719–725 (2010)
9. Böhning, D.: Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. Metron (2015). doi:10.1007/s40300-015-0071-6
10. Francis, B., Liu, J.: Modelling escalation in crime seriousness: a latent variable approach. Metron (2015). doi:10.1007/s40300-015-0073-4
11. Florios, K., Moustaki, I., Rizopoulos, D., Vasdekis, V.G.S.: A modified weighted pairwise likelihood estimator for a class of random effects models. Metron (2015). doi:10.1007/s40300-015-0070-7
12. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**, 33–50 (1978)
13. Lindsay, B.G.: Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics 5. Institute of Mathematical Statistics, Hayward (1995)
14. Marino, M.F., Farcomeni, A.: Linear quantile regression models for longitudinal experiments: an overview. Metron (2015). doi:10.1007/s40300-015-0072-5
15. McCulloch, C.E., Searle, S.R., Neuhaus, J.M.: Generalized, Linear and Mixed models, 2nd edn. Wiley, New York (2008)
16. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
17. Paas, L., Bijmolt, T.H.A., Vermunt, J.K.: Long-term developments of EU household financial product portfolios: a multilevel latent class analysis. Metron (2015). doi:10.1007/s40300-015-0067-2
18. Roeder, K.: Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. J. Am. Stat. Assoc. **92**, 894–902 (1990)
19. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. CRC Press, Boca Raton (2004)
20. Vasdekis, V., Rizopoulos, D., Moustaki, I.: Weighted pairwise likelihood estimation for a general class of random effects models. Biostatistics **15**, 677–689 (2014)