FINITE MIXTURE MODELS: A SYSTEMATIC REVIEW

José G. Dias [1]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, (e-mail: jose.dias@iscte-iul.pt)

ABSTRACT: Finite mixture models have been used in many fields for different purposes and under different names. Other non-exact names are model-based clustering and latent class models. This presentation gives an overview of this area. In particular, it summarizes the research that has been published both theoretical papers and in applications. A systematic literature review using the PRISMA methodology was used. The text mining analysis then identifies topics in the literature. Results show an explosion of the research in the field since 2000. Social and health sciences are the most prominent application areas, mainly focused on the detection of unobserved heterogeneity.

KEYWORDS: finite mixture models, latent class models, discrete latent variables, model-based clustering, systematic literature review.

Finite mixture (FM) models and related latent variable models are over one hundred years old. The origin of the FM model is usually attributed to Newcomb and Pearson (i.e., Newcomb, 1886; Pearson, 1894). Stigler, 1986, however, at least traces its origin back to the analysis of conviction rates by Poisson in the second quarter of the nineteenth century. Since 2000, the use of these models has grown exponentially. In the past few decades, advances in computer technology, FM modeling has proven to be a powerful tool for the analysis of a wide range of empirical problems. For instance, in the social sciences, which have a long tradition of latent class (LC) models, following the seminal work by Lazarsfeld and refinements notably by Goodman and Clogg (see, e.g., Goodman, 1974 and Clogg, 1995), more sophisticated models are gaining popularity. McLachlan & Peel, 2000 provide a good overview of the field until 2000. The exponential growth in the use of these models over the past two decades clearly shows that they are directly related to the democratization of statistical computation using fast personal computers (PCs) and increasing availability of software for their estimation.

This work presents an overview of the field using a systematic literature review. In addition to searching for articles using keywords to retrieve papers, we also used papers citing well-known references in the field (e.g., Titterington *et al.* , 1985; McLachlan & Peel, 2000; Scrucca *et al.* , 2016). The extraction and selection of papers from the Web of Science follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology. A total of 38,997 papers were included in the analysis. Topic analysis, a special case of text mining, is used to identify topic clusters in the corpus.

Results show the diverse use of FMs in the literature. Most publications use FMs to identify clusters. However, in other applications and contexts, topics cover density estimation, defining prior probabilities in Bayesian statistics, discrete latent variables, the golden standard problem, speech modeling, imagine analysis, longitudinal and trajectory analysis, or social class analysis. This research establishes a typology in the field of FM methodology and shows its wide range and flexible use in statistical modeling.

References

CLOGG, C. C. 1995. Latent class models. *Pages 311–359 of:* ARMINGER, G., CLOGG, C.C., & SOBEL, M.E. (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum.

GOODMAN, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**(2), 215–231.

MCLACHLAN, G.J., & PEEL, D. 2000. *Finite Mixture Models*. New York: John Wiley & Sons.

NEWCOMB, S. 1886. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, **8**(4), 343–366.

PEARSON, K. 1894. Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, **185**, 71–110.

SCRUCCA, L., FOP, M., MURPHY, T. B., & RAFTERY, A. E. 2016. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, **8**(1), 289–317.

STIGLER, S.M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA and London: Belkap Press of Harvard University Press.

TITTERINGTON, D.M., SMITH, A.F.M., & MAKOV, U.E. 1985. *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons.