Abstract

The present thesis is a compendium of *three* original works produced between 2014 and 2018. The papers have a common link: they are different contributions made by compositional data analysis to the study of the models based on mixtures of probability distributions. In brief, we could say that *compositional data analysis* is a methodology that consists of studying a sample of measures that are strictly positive from a relative point of view. *Mixtures of distributions* are a specific type of probability distribution defined to be the convex linear combination of other distributions.

In the first work that makes up this thesis, the available options for defining mixture of probability distributions within the sample space of compositional data (simplex) are analysed, considering their specific algebraic structure. Among the different available options, it is observed that either mixtures of distribution are not well defined in the simplex, or that they were not rich enough in terms of their capacity to model sets of real compositional data, leading us to consider the log-ratio approach as a tool to solve existing problems. By means of compositional data analysis based on log-ratios, a method for constructing mixtures of distributions that are well defined in the simplex and are as rich as existing distributions for modelling real multivariate data is proposed.

Generally, the models based on mixtures of distributions are adjusted with the EM algorithm, which obtains the parameters of the distributions that intervene in the mixture and the parameters of the mixture itself. Apart from these parameters, the algorithm also calculates the probability that each of the observations has been generated by each of the components that make up the mixture distribution. These probabilities, called posterior probabilities, allow for classifying each of the observations in the most probable component, making this process a very popular clustering method. Some authors have proposed using these probabilities not only to cluster observations, but also to define a hierarchical structure of the components of mixture distribution. In the second work that makes up this thesis, a

model is presented that integrates all the proposals found in the literature that base the construction of this hierarchy on the vectors of posterior probabilities. Apart from this new integrating model, new methods for creating hierarchies using coherent measures for vectors of probabilities, from a compositional point of view, are introduced

The most frequent mixtures emerge from putting a categorical distribution together with another probability distribution, which is generally defined in the real space. Thus, by means of considering the categorical distribution compounded with a function of probability distribution, we obtain a finite mixture of distributions of this specific distribution with weights given by the parameters of the categorical distribution. In this case, it is said that the categorical distribution is the weighting distribution; the other distribution is called the kernel distribution. This process can be carried out whenever there is a mechanism that allows the parameters of the kernel distribution to be defined from the observed values of a random variable following the weighting distribution. More specifically, if we consider the multinominal distribution as the kernel distribution and the logarithm quotient-normal distribution in the simplex as the *mixing distribution*, we will have what is known as the logarithm quotient-normal-multinominal probability distribution. In the third and last work of this compendium, different properties of this distribution are derived, a new method for estimating the parameters of the distribution is presented and the capacity improvement for modelling counting data compared with the Dirichlet-multinomial distribution, one of the most popular in this context, is demonstrated.

Publications

- Comas-Cufí, M., Martín-Fernández, J.A., Mateu-Figueras, G. (2016), **Log-ratio methods in mixture models for compositional data sets**. Statistics and Operations Research Transactions, 40 (2), pp. 349–374.

- Comas-Cuf´ı, M., Mart´ın-Fern´andez, J.A., Mateu-Figueras, G. (2017), **Merging the components of a finite mixture using posterior probabilities**. Statistical Modelling (in press)

- Comas-Cufí, M., Martín-Fernández, J.A., Mateu-Figueras, G., Palarea-Albaladejo, J. (2018), **Modelling count data with the logistic-normal-multinomial distribution**. Statistics and Operations Research Transactions (submitted)