Discrete Laplace mixture model with applications in forensic genetics

Mikkel Meyer Andersen

1 Department of Mathematical Sciences, Aalborg University, Denmark

Summary

This R package implements a model based on a mixture of multivariate discrete Laplace distributions that has applications in forensic genetics. The implementation consists of parameter estimation and various functionalities. The method and this package were (and still are) used by multiple groups for e.g. frequency estimation (Andersen, Eriksen, and Morling 2013; S Willuweit and Roewer 2015; S. Willuweit et al. 2018; Roewer and Willuweit 2018; Egeland, Kling, and Mostad 2016; Cereda et al. 2014; Cereda 2017), cluster analysis (Andersen, Eriksen, and Morling 2014), and mixture interpretation [Andersen et al. (2015); Taylor2018]. Below, background for the method and package is described.

Estimating haplotype frequencies is important in e.g. forensic genetics, where the frequencies are used to calculate the likelihood ratio for the evidential weight of a DNA profile found at a crime scene (Andersen, Eriksen, and Morling 2013; Steele and Balding 2015). Estimation is naturally based on a population model, motivating the investigation of the Fisher-Wright model (Fisher 1930; Wright 1931; Ewens 1972; Ohta and Kimura 1973) of evolution for haploid lineage DNA markers.

An exponential family (a class of probability distributions that is well understood in probability theory) called the ‘discrete Laplace distribution’ was described in (Andersen, Eriksen, and Morling 2013) that also illustrates how well the discrete Laplace distribution approximates a more complicated distribution that arises by investigating the well-known population genetic Fisher-Wright model of evolution by a single-step mutation process (Caliebe et al. 2010).

In (Andersen, Eriksen, and Morling 2013), it was also shown that the discrete Laplace distribution can be used to estimate haplotype frequencies for haploid lineage DNA markers (such as Y-chromosomal short tandem repeats), which in turn can be used to assess the evidential weight of a DNA profile found at a crime scene. This was done by making inference in a mixture of multivariate discrete Laplace distributions using the EM algorithm to estimate the probabilities of membership of a set of unobserved subpopulations and also estimate the central haplotypes of the subpopulations. The implementation was made efficient by avoiding to construct the model matrix explicitly which made it possible to perform the calculations on a normal computer.

This package implements the method described in (Andersen, Eriksen, and Morling 2013) as a freely available open source software R package using both R (R Core Team 2018) and C++ (Eddelbuettel and Balamuta 2017). The documentation of disclapmix consists of manual pages for the various available functions, articles describing how to perform contiguous analyses (vignettes), and unit tests.

I would like to thank Poul Svante Eriksen (Aalborg University, Denmark) for tips, hints, helpful discussions and help with implementation and debugging.
References


