The Journal of Open Source Software

Discrete Laplace mixture model with applications in forensic genetics

Mikkel Meyer Andersen[1]

1 Department of Mathematical Sciences, Aalborg University, Denmark

DOI: 10.21105/joss.00748

Software

- Review
- Repository
- Archive

Submitted: 22 April 2018
Published: 04 June 2018

Licence
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

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 explicitely 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

Andersen, Mikkel Meyer, Poul Svante Eriksen, and Niels Morling. 2013. "The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies." *Journal of Theoretical Biology* 329:39–51. https://doi.org/10.1016/j.jtbi.2013.03.009.

———. 2014. "Cluster analysis of European Y-chromosomal STR haplotypes using the discrete Laplace method." *Forensic Science International: Genetics* 11:182–94. https://doi.org/10.1016/j.fsigen.2014.03.016.

Andersen, Mikkel Meyer, Poul Svante Eriksen, Helle Smidt Mogensen, and Niels Morling. 2015. "Identifying the most likely contributors to a Y-STR mixture using the discrete Laplace method." *Forensic Science International: Genetics* 15:76–83. https://doi.org/10.1016/j.fsigen.2014.09.011.

Caliebe, Amke, Arne Jochens, Michael Krawczak, and Uwe Rösler. 2010. "A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process." *Journal of Theoretical Biology* 266 (2):336–42. https://doi.org/10.1016/j.jtbi.2010.06.033.

Cereda, G. 2017. "Impact of Model Choice on LR Assessment in Case of Rare Haplotype Match (Frequentist Approach)." *Scandinavian Journal of Statistics* 44. https://doi.org/10.1111/sjos.12250.

Cereda, G., A. Biedermann, D. Hall, and F. Taroni. 2014. "An investigation of the potential of DIP-STR markers for DNA mixture analyses." *Forensic Science International: Genetics* 11. https://doi.org/j.fsigen.2014.04.001.

Eddelbuettel, D, and JJ Balamuta. 2017. "Extending *R* with *C++*: A Brief Introduction to *Rcpp*." *PeerJ Preprints* 5 (August):e3188v1. https://doi.org/10.7287/peerj.preprints.3188v1.

Egeland, Thore, Daniel Kling, and Petter Mostad. 2016. *Relationship Inference with Familias and R*. Academic Press. https://doi.org/10.1016/C2014-0-01828-X.

Ewens, W.J. 1972. "The sampling theory of selectively neutral alleles." *Theoretical Population Biology* 3:87–112.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

Ohta, T., and M. Kimura. 1973. "A Model of Mutation Appropriate to Estimate the Number of Electrophoretically Detectable Alleles in a Finite Population." *Genet. Res.* 22:201–4.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Roewer, L., and S. Willuweit. 2018. "Y-chromosomale STR-Analyse in der forensischen Praxis." *Rechtsmedizin* 28 (2):149–64. https://doi.org/10.1007/s00194-018-0229-7.

Steele, CD, and DJ Balding. 2015. *Weight of Evidence for Forensic DNA Profiles*. 2nd ed. Wiley. https://doi.org/10.1002/9780470867693.

Willuweit, S, and L Roewer. 2015. "The New Y Chromosome Haplotype Reference Database." *Forensic Science International: Genetics* 15:43–48. https://doi.org/10.1016/j.fsigen.2014.11.024.

Willuweit, S., K. Anslinger, G. Bäßler, M. Eckert, R. Fimmers, C. Hohoff, M. Kraft, et al. 2018. "Gemeinsame Empfehlungen der Projektgruppe "Biostatistische DNA-Berechnungen" und der Spurenkommission zur biostatistischen Bewertung von Y-chromosomalen DNA-Befunden." *Rechtsmedizin* 28 (2):138–42. https://doi.org/10.1007/s00194-018-0244-8.

Wright, S. 1931. "Evolution in Mendelian populations." *Genetics* 16:97–159.