Summary statistics selection for ABC inference in *R*

Matthew A. Nunes[1,⋆], **David J. Balding**[2]

1. Department of Mathematics & Statistics, Lancaster University, Lancaster, U.K.
2. UCL Genetics Institute, University College London, London, U.K.

⋆Contact author: m.nunes@lancs.ac.uk

Keywords: Approximate Bayesian Computation, entropy minimization, summary statistics

For the purposes of statistical inference, high-dimensional datasets are often summarized using a number of statistics. Due to the intractable likelihoods involved in models for these complex datasets, this is often performed using Approximate Bayesian Computation (ABC), in which parameter inference is achieved by comparing the summaries of an observed dataset to those from simulated datasets under a chosen model (Beaumont et al., 2002).

The question of how best to summarize high-dimensional datasets to maximize potential for inference has been recently addressed in Nunes and Balding (2010). The authors propose two techniques to select a good set of summaries for ABC inference from a collection of possible statistics.

Since high-dimensional datasets arise in many areas of science, these methods could provide practical guidance to scientists on how to represent datasets for optimal inference. The minimum entropy (ME) and two-stage error prediction methods introduced in Nunes and Balding (2010) are implemented for the "rejection-ABC" algorithm (Tavaré et al., 1997) in the **ABCME** *R* package. The package also includes code to perform optional "regression adjustments" for the mean and variance of parameter values which can improve overall quality of inference of ABC algorithms (Fan and Yao, 1998; Yu and Jones, 2004; Beaumont et al., 2002). Some of the computational cost of the ABC algorithm and summary selection procedures is reduced through the use of C routines.

We provide an example of the techniques in the **ABCME** package, demonstrating posterior inference of the parameters from a coalescent model of population DNA sequences (Nordborg, 2007).

References

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian Computation in population genetics. *Genetics 162*(4), 2025–2035.

Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*(3), 645–660.

Nordborg, M. (2007). Coalescent theory. In D. J. Balding, M. J. Bishop, and C. C (Eds.), *Handbook of Statistical Genetics* (3rd ed.)., pp. 179–208. Wiley: Chichester.

Nunes, M. A. and D. J. Balding (2010). On optimal selection of summary statistics for Approximate Bayesian Computation. *Stat. Appl. Genet. Mol. Biol. 9*(1).

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997). Inferring coalescence times from DNA sequence data. *Genetics 145*(2), 505–518.

Yu, K. and M. C. Jones (2004). Likelihood-based local linear estimation of the conditional variance function. *J. Am. Stat. Assoc. 99*(465), 139–144.