# Approximate Bayesian Computation: a simulation based approach to inference

Richard Wilkinson[1]⋆, Simon Tavaré

[1] University of Sheffield, Sheffield, Great Britain
⋆ r.d.wilkinson@shef.ac.uk

There is a large class of stochastic models for which we can simulate observations from the model, but for which the likelihood function is unknown. Without knowledge of the likelihood function standard inference techniques such as Markov Chain Monte Carlo are impossible, as the unnormalized likelihood function is explicitly required for the calculation of an acceptance rate. In this talk I shall introduce a group of Monte Carlo methods that can be used to perform inference for stochastic models from which we can cheaply simulate observations.

These likelihood-free Monte Carlo techniques usually go under the name of Approximate Bayesian Computation (ABC) [1], and although they are becoming popular in several application areas, such as genetics and ecology, they remain relatively unstudied by the statistics community. The basic idea is that we choose parameter values from a prior distribution, simulate a data set using these values, and then accept the parameters as a sample from the posterior distribution if the simulated data closely matches the field data. If the simulated data exactly matches the field data, then the parameter value used can be considered an independent draw from the posterior distribution. If the simulated data is only approximately equal to the field data, the the parameter value used can be considered as a draw from a distribution that approximates the true posterior distribution.

Approximate likelihood-free Markov Chain Monte Carlo algorithms [2] and approximate sequential importance sampling methods [3] have been suggested by other authors. I shall show how these approaches may be extended to take account of any known model structure. Many of the applications that have used ABC methods, such as epidemiology, population genetics and ecology, have stochastic models in which there is a hidden tree structure generating highly-correlated data. The likelihood-function is usually difficult to calculate, but a single simulation typically runs very quickly. Once we condition on the hidden tree structure we can compute the conditional likelihood for the other parts of the model. The ABC update can then be used to choose a new tree structure, and a standard MCMC step can be

used to update the other parts of the parameter space. This leads to considerable improvements in the efficiency and accuracy of the approach.

Finally, many important technical questions remain about ABC methods and in this talk I shall highlight some of these issues. The accuracy of the approximation has not been quantified, and there is almost no theory specifying how one should define the appropriate summary statistics and distance measures that are needed. For models that produce high-dimensional output it is necessary to summarize the data and compare a summary of the simulation output with a summary of the field data. If the summary is a sufficient statistic, in the sense that it contains all of the information about the parameter that is available in the data, then the ABC algorithm can be shown to produce meaningful output. However, sufficient statistics are not usually available and the effect of using non-suffienct statistics is not generally known. I shall illustrate some of these issues for a simple model where we can explicitly calculate the posterior distribution.


[1 ] M. A. Beaumont and W. Zhang and D. J. Balding (2002). Approximate Bayesian Computatation in Population Genetics. *Genetics* **162(4),** 2025-2035.

[2 ] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Science, USA* **100(26),** 15324-15328.

[3 ] S. A. Sisson, Y. Fan and M. M. Tanaka (2007). Sequential Monte Carlo without Likelihoods. *Proceedings of the National Academy of Science, USA* **104(6),** 1760-1765.