

**11. Title: OpenMP and MCMC: Parallelizing the simulation of a single Metropolis Hastings Markov chain**  
**Proposer: Geoff Nicholls**

MCMC is a widely used Monte Carlo algorithm for generating samples from a given probability distribution. We simulate a Markov chain with the property that the equilibrium of the chain coincides with a given target distribution (commonly, some posterior probability distribution). In order to achieve equilibrium it is sometimes necessary to simulate many steps of the chain. It is common practice to run several simulations on independent computers from distinct start states, and thereby check that equilibrium has been reached in any one simulation. We would like to use parallelism in a different way, to achieve very long runs of a single chain. A simple algorithm has been given to do this (and this might be surprising, as a single simulation has serial correlation).

We would like to use the OpenMP protocol to test the efficiency of this scheme on parallel machines which are commonly available (for example dual core laptops). There are important applications in Geoscience, and Medical imaging: all cases where the evaluation of the likelihood is computationally demanding, so that each step of the MCMC simulation may take a second or more to evaluate. The focus of the project is on the algorithm, so we can to some extent choose the application to keep things simple.

Students can get a feeling for the problem by reading this MSc thesis from the University of Auckland. We will take a much simpler application than the two-phase flow case considered there. <http://www.stats.ox.ac.uk/~nicholls/linkfiles/papers/TCThesis.pdf>

Relevant skills are 1) knowledge of Metropolis Hasting MCMC, 2) basic C programming 3) some basic experience with numerical analysis, for example, elementary numerical methods for solving DE's or PDE's.

References:

Parallel Programming in C with MPI and OpenMP, Michael J. Quinn, ISBN 0072822562 / 9780072822564

C.P. Robert and G. Casella. "Monte Carlo Statistical Methods" (second edition). New York: Springer-Verlag, 2004.

**12. Title: The ABC's of migration.  
Proposer: Geoff Nicholls**

The "Approximate Bayesian Computation" (ABC) algorithm is a scheme for approximate Monte Carlo. It is attractively simple. For example, the rejection algorithm version for the posterior for discrete parameter  $x$  given discrete data  $y$ , the distribution  $p(x|y)=p(y|x)q(x)/c(y)$ , is as follows:

(a) draw  $z$  from  $q(\cdot)$  and then (b) simulate  $y'$  according to  $p(\cdot|z)$ . Finally (c) if  $y'=y$  accept  $x=z$  as a draw from  $p(x|y)$  otherwise repeat (a) and (b) until success at (c). Now this is exact. The trick is to define a distance  $d(y,y')$  between vectors in the space of the data  $y$ , and accept at (c) if  $d(y',y)$  is below some threshold. If the distance is chosen in the right way, the ABC algorithm generates samples with distribution close to the desired target  $p(x|y)$ . There is an MCMC version as well.

The framework is attractive, as it is very often easy to simulate the observation process  $p(y|x)$ . We would like to try this scheme for a genealogical model of migration called the Island Migration model. This is perhaps the simplest modification of the Kingman coalescent (itself a model of genealogies in a single panmictic population) to take into account population structure. We have some HIV DNA sequences which were sampled from different parts of the body of a patient. We are interested in recovering the migration-history of the ancestors of these viruses with the patient's body.

Relevant skills are a) some knowledge of the material in MS2a or MS2b  
b) some programming skills in R or MatLab c) some basic familiarity with Monte Carlo methods

References

For a fairly high-level discussion of the issues see for example G. Ewing, G. Nicholls, and A. Rodrigo, "Using Temporally Spaced Sequences to Simultaneously Estimate Migration Rates, Mutation Rate and Population Sizes in Measurably Evolving Populations" *Genetics*, December 1, 2004; 168(4): 2407 – 2420 (2004)  
[<http://www.genetics.org/cgi/content/abstract/168/4/2407>]

Simon Taveré (2004), "Ancestral Inference in Population Genetics",  
Lectures on Probability Theory and Statistics, Springer Berlin / Heidelberg, Volume 1837

Handbook of Statistical Genetics, Second Edition, Chapter 7,  
David J. Balding <<http://eu.wiley.com/WileyCDA/Section/id-302479.html?query=David+J.+Balding>> (Editor),  
M. Bishop <<http://eu.wiley.com/WileyCDA/Section/id-302479.html?query=M.+Bishop>> Editor)  
C. Cannings <<http://eu.wiley.com/WileyCDA/Section/id-302479.html?query=C.+Cannings>>  
(Editor), Wiley, 2003