

STOCHASTIC COMPUTATION

A SAMSI Program from September, 2002 to May, 2003

The last decade has witnessed an essential revolution in statistical sciences based on developments in stochastic simulation and optimisation methods for scientific computation. One of the driving motivations for development of these methods has been for analysis of random effects (or multilevel, or hierarchical) models, a major class of mathematical/statistical models that arise naturally in a vast array of problems of complex stochastic systems.

A SAMSI initiative in this area is planned for 2002/03. This program will be a *synthesis program* whose goals are assessment and synthesis of stochastic simulation methods in central classes of probabilistic/statistical models widely used in scientific modelling, with graphical models as an over-arching framework. Specific model structures will be drawn from applications in various fields, including bioinformatics and areas that intersect the concurrent Inverse Problems SAMSI program. One major theme is to promote and develop more substantial interactions between probabilists and statisticians working on stochastic simulation theory and methods; the recent emergence of methods of perfect simulation represents a major opportunity for catalyzing such interactions.

The specific topics of program focus, from the general arena outlined below, will be finalized during ‘formation meetings’ to be held from October 2–4, 2002. The specific topics will depend heavily on senior research leadership contributed by visiting scientists.

STOCHASTIC COMPUTATION: THEORY & METHODOLOGY

Samples drawn stochastically from a probability distribution are used today for a vast array of mathematical and statistical calculations, and effectively define the standards of modern statistical computation. The last decade or so has seen *Markov Chain Monte Carlo* (MCMC) and related methods become established as central tools for statistical analysis of complex models. MCMC simulation uses Markov Chain ‘walks’ over the domain of the distribution that, together with a variety of techniques, can be used to obtain samples from the distributions or perform calculations (such as expectations) involving the distribution. Basic importance sampling and other direct/exact methods also fall under this heading and still play an important role in many situations. MCMC methods allow for sampling from complicated, high dimensional distributions and for statistical analysis of models of enormous complexity.

Among the many aspects of MCMC methodology that represent areas of need for research overview and synthesis include reversible jump MCMC methods and related approaches to dealing with multiple models of differing dimension; approximation algorithms to address computational intractabilities; strategies based on combining information from multiple parallel Markov chains; the use of concepts and methods of perfect sampling; variants of importance sampling; adaptive and evolutionary methods; auxiliary variable techniques; simulation and search algorithms for statistical model selection and variable selection; analysis using sequential Monte Carlo methods.

Recent years have seen rapid growth in the development of novel theory and methods for ‘perfect sampling,’ including methods based on couple Markov chains and others, that may eventually lead to substantially more efficient and accurate stochastic computational algorithms than current MCMC methods. Much of the work in these areas has yet

to percolate through to analysis of the kinds of complex mathematical models routinely used by statisticians and applied mathematicians, and there is a pressing need to promote broader and deeper interactions between the ‘theorists,’ largely in probability and applied probability communities, and the ‘methodologists’ in statistics and applied mathematics.

The above example of perfect sampling touches on a general community need. The methodological developments in stochastic computation, especially in MCMC and associated methods, have far outstripped theoretical developments. The result is a very considerable need for fundamental theoretical attention to the area, and advances in our theoretical understanding of algorithms in key aspects that include convergence bounds for iterative simulation methods and algorithmic optimisation, perfect sampling and its variants, MCMC design and optimisation in highly structured, high-dimensional probability distributions (e.g., graphical models and networks), and sequential simulation/particle filtering methods.

STOCHASTIC COMPUTATION: MODEL CONTEXTS

An over-arching framework is graphical models of multivariate probability distributions; data and parameter variables are nodes in the graph and arcs represent possible dependencies between nodes. Analysis aims to define and estimate the sets of conditional probability distributions node-by node. In realistic models, these graphical models have many nodes and the modelling of conditional distributions at nodes involves problems of variable and model selection. Analytic approximations (e.g., mean-field and variational approximations) are useful, as are Monte Carlo methods including variants of importance sampling. Producing stochastic samples by sequentially visiting nodes, individually or in subsets, can be a challenging enterprise in other than small-scale models and with realistic structure in the dependencies among variables represented by the graph. In addition to approximations, model and variable selection, research areas of interest include auxiliary variable methods, the development of perfect sampling and coupling methods for classes of graphical models, sequential simulation approaches, and approaches using multiple Markov chains.

Subclasses of graphical models with ‘simpler’ structure arise in areas of application of stochastic computation. These include hierarchical models, in which unknown parameters corresponding to individuals in a population are themselves modeled as arising from a ‘population distribution.’ These models have been a staple of statistical analysis for many years, and the advent of MCMC computational techniques, which greatly simplify the analysis of such models, has caused an explosion of interest in and application of such models within a wide variety of disciplines. Such models now go under numerous names, including ‘hierarchical models,’ ‘mixed models,’ and ‘multilevel models.’ The general field also covers structured models with many hidden or ‘latent’ variables, whose increasing application in scientific and socio-economic modeling is enabled by developments in stochastic computation. Hierarchical/multi-level random-effects models have graphical representations and can sometimes be represented as trees with data and parameter variables comprising nodes. Computational challenges arise due to complex forms of conditional distributions at nodes, and increasing needs to deal with very high-dimensional models and parameters. Here too there is interest in moving some of the new theory of perfect sampling into more practical contexts.

A special class of hierarchical models is that of state-space (or hidden Markov) models, including dynamical systems models with time-evolving states. Beyond restricted classes of models (small discrete parameter spaces, or analytically tractable cases) computation via stochastic or related numerical methods is required. Practical models involve continuous multivariate parameters. MCMC methods and related methods of sequential simulation (sequential importance sampling, and particle filtering) exist and are being developed for such models, though little exists in the way of general theory to guide methodological work. Sequential analysis naturally invites the introduction of ideas of perfect simulation as well as refined resampling schemes for sequential importance sampling.

STOCHASTIC COMPUTATION: APPLICATION CONTEXTS

Various application fields may generate candidate models and data sets for development and evaluation of computational methods in this program. One already mentioned is the general area of hierarchical, random effects modelling, which will feature heavily in the concurrent Inverse Problems program and from which model contexts will be generated to implement and evaluate approaches to stochastic computation.

An additional broad field is bioinformatics. New and traditional areas of genetics and genomics are experiencing growth of interest in methods of stochastic computation for simulation of both biological processes and for technical use in analysis of specific models. Key examples include the simulation of phylogenetic trees based on sequence data, a broad and important area of evolutionary genomics and one now moving more seriously into application. Challenges here include the development of effective methods for simulation of evolutionary trees and for estimation and evaluation of statistical models of such trees. MCMC and importance sampling methods already play important roles, and sequential simulation are natural in this context. Related studies arise in more traditional areas of population genetics including pedigree studies. A quite different class of applications arises in functional genomics in studies of gene (and protein) expression data with the goals of developing insights into underlying genetic pathways and networks. Probabilistic graphical network models of such data, both observational and from controlled experimentation with gene modifications, provide natural approaches to inference in these problems. The general challenges to stochastic computation that arise in network models then need addressing, including particularly the needs to develop model and variable selection tools at nodes (genes) of the network, and the challenges of scaling up to very many nodes.

STOCHASTIC COMPUTATION: LOGISTICS

Program Leadership: The Program Leaders Committee consists of Merlise Clyde (Duke), Jun Liu (Harvard), Michael Newton (Wisconsin), and Mike West (Duke), with an extended Working Group that includes Jean Pierre Fouque (North Carolina State, mathematics), Alan Gelfand (Connecticut, statistics), David Heckerman (Microsoft Research, statistics and machine learning, also representing the SAMSI National Advisory Committee), Mark Huber (Duke, mathematics and statistics), Greg Lawler (Cornell and Duke, probability), John Monahan (North Carolina State, statistics) and Scott Schmidler (Duke, bioinformatics).

Program Schedule and Activities: The formal program will run from September 2002 to May 2003 inclusive. Once a small number of key program participants are identified (during Spring 2002), these participants will work during mid-2002 with the Program Committee to refine the areas of program focus within the broader agenda outlined above. Initial visitors will arrive at SAMSI in August 2002 and begin working more actively with the Working Group. The formal schedule from then on will be:

- *September 25 – October 4, 2002:* 10 day conference, attended by all program participants (including those whose longer-term visits may start later) featuring:
 - 3 days of tutorials, aimed at students, new researchers and crossover communication between mathematicians, probabilists and statisticians;
 - 4 days of research workshop on advanced computational techniques;
 - 3 days of ‘formation meetings’ of all participants with the Working Group, to finalize identification of specific topics to be developed via in-depth investigations as core of the program and to generally plan and organize the efforts during the program.
- *January 29–February 1, 2003:* ‘progress report’ workshop and conference
- *May 28–31, 2003:* final 5 day workshop to synthesize the findings of the working groups and lay the groundwork for the final report.

Program Participants:

- Working Group and various local researchers
- SAMSI visitors, including funded project participants and SAMSI fellows from probability, applied mathematics, computer science and statistics
- SAMSI post-doctoral fellows
- SAMSI PhD student research assistants
- Representatives of NISS industrial, governmental, national lab and academic affiliates

Program Deliverables:

- Book comprising several chapters reporting on research, overview and synthesis results in several key areas targeted by the program
- An evolving program web site that will comprise a “virtual laboratory” that records program methods, models, data and experiences
- Postdoctoral fellows and graduate students involved in collaborative work with key leaders in stochastic computation, and rapid exposure to hot topics and new areas that is expected to lead to follow-on research
- Multiple new relationships between sub-fields of the mathematical sciences through the focus on developing interactions to forge cross-over communication, especially between ‘theory’ and ‘methodology’ in stochastic computation