STOCHASTIC COMPUTATION

SAMSI PROGRAM: FINAL REPORT

March 2, 2004

OVERVIEW & GOALS

The inaugural SAMSI program on Stochastic Computation focussed on synthesis and developmental research in four inter-related areas:

- *Stochastic computation in problems of model and variable selection -*
  - Computing hard likelihoods and Bayes' factors for model comparison and selection
  - Interface research with the SAMSI Inverse Problems program (Fall 2002)
  - Synthesis of existing stochastic computational approaches to model uncertainty and selection
- *Stochastic computation in inference and imputation in contingency tables analysis -*
  - Perfect simulation approaches to "missing data" problems
  - Synthesis of existing Markov chain simulation methods - "local" and "global" move approaches
  - Experiments with approaches in large, sparse tables
  - Studies in application in genetics and other areas
- *Stochastic computation in analysis of large-scale graphical models -*
  - Monte Carlo and related stochastic search methods for sparse, large-scale graphical models
  - Model definition and specification for sparse models
  - Studies in applications in genomics (large-scale gene expression studies)
- *Stochastic computation in financial mathematics, especially in options pricing models -*
  - Monte Carlo methods in specific options pricing models, and
  - Sequential Monte Carlo methods and particle filtering in stochastic volatility models.

StoCom aimed to explore the use of methods of stochastic computation in these key areas of statistical modelling. The core research focuses included studies of the performance characteristics of current stochastic computational methods, refinements and extensions of existing approaches, and development of innovative new approaches. In two of the areas, in particular, there was an explicit focus on the development of interactions between statisticians (coming from methodological and applied perspectives) and theoretical probabilists and mathematicians working on related problems. One key example was the component on imputation in contingency tables, with an interest in connecting Markov chain methods from statistics with perfect sampling from probability and algebraic approaches from mathematics. Another example was in modelling in financial pricing studies where statistical and mathematical "schools" have had limited interactions.

An over-arching theme of StoCom was to explore and develop an improved understanding of shortcomings of current approaches to exploring high-dimensional structured spaces (spaces of multiple models in variable selection, spaces of structured contingency table parameters and missing data, spaces of decomposable graphical models) and fitting important classes of stochastic models (contingency tables, graphical models, financial models). Research papers and software was anticipated. The earlier (SAMSI proposal stage) idea of a book on the program results was dropped in favour of publication of research papers, particularly anticipating the development of a broader range of research innovation - as opposed to review and synthesis - than initially conceived. The goals specifically included the invention, evaluation and development of new computational methods (Markov chain Monte Carlo, stochastic search, sequential methods, annealing methods, perfect sampling methods, and others) to address challenges and advance methodology for model fitting and exploration. Issues of scalability (especially in graphical models and contingency tables) were to be highlighted, as was explicit consideration of impact on applications in various fields.

The earlier/original proposal for StoCom was revised and focused via definition of specific areas for research: the need for new theoretical developments, developments in modelling and algorithms, and in potential applications in genetics and social sciences (contingency tables), genomics (graphical models), finance

(options pricing), and many other fields. Initiation of new research focused specifically on computation and the use of distributed processing was an additional area of definition introduced in fall 2002.

Core personnel and human resources goals were in line with the core focus of SAMSI on involving PhD students and postdoctoral fellows from statistics, probability and mathematics, and on stimulating personal and collaborative research across these sub-disciplines, with the expectation of continued and growing collaborations post-program as these researchers move on with an interdisciplinary mathematical sciences orientation to their early research career paths. The focus on stimulating increased interactions among statisticians and probabilists in addressing theoretical and algorithmic aspects of stochastic computation was key, and one core goal within that was the aim of connecting theoreticians (on the algebraic side of contingency tables analysis, in perfect sampling, in financial modelling, and other areas) with statisticians more concerned with model refinement and application.

OUTCOMES OF RESEARCH

A full and detailed description of the research, personnel activities, outcomes and ongoing/resulting research and collaborations follows. This describes all aspects of the program, including discussion of specific research activities and projects that defined the path to meeting core program goals while also broadening scope on innovative research not earlier anticipated. Papers and software developed during StoCom and from research initiated as a result are detailed, as are some grant proposals written either directly as a result of research initiated during StoCom, or having a substantial bearing on the StoCom work.

Among the StoCom findings and results, the following are of particular note. First, in the area of large-scale variable selection and regression model search, StoCom introduced novel methods for rapid exploration of "model spaces" with efficient adaptive sampling algorithms that can substantially improve regression model search when faced with many candidate predictors. This was coupled with innovations in model theory (Bayesian prior distributions) for model uncertainty that are demonstrably superior, in decision theoretic terms, to previous approaches. Second, the contingency tables program produced significant advances in computational methods - an efficient sampling method for exact tests of Hardy-Weinberg genetic equilibrium, and a new Markov chain method for regression models for much larger problems than previously were possible. StoCom also impacted here in furthering development of the interconnections between algebraic research on lattice points and polytopes, with statistical applications in genetics and tabular census data. This connection has resulted in faster algorithms for counting and sampling tables with constraints, and has led to new mathematical research with future applications. Thirdly, in the area of large-scale graphical models, StoCom research innovations led to substantial advances in the capacity to explore very high-dimensional spaces of graphs, extended known approaches to general classes of graphs (as opposed to earlier constraints to limited, decomposable graphs), introduced radically effective novel approaches of searching for models - named Shotgun Stochastic Search - and advanced our understanding of the applied relevance of theory behind these models through studies of implications of "standard" prior distributions. This component also explored and generated studies of the approach in exploratory analysis of large-scale genomic data, and opened the way to more incisive applications in that field, as well as introducing distributed computation on Beowulf clusters to enable the application of new search methods. Fourthly, in the area of financial modelling, StoCom brought together statistical and mathematical researchers to define new approaches to stochastic computation in some of the most challenging stochastic volatility and pricing models, and defined new research directions from this collaboration in studies of volatility at multiple scales, among other topics.

On personnel and activities of engaged individuals, the workshops involved almost 200 researchers from statistics, mathematics and probability, as well as a number of disciplinary scientists. This included strong representation of women and many new/junior researchers and PhD students. The core research program actively involved six PhD students "on site" as well as four others remotely via collaborations, three postdoctoral fellows (one female) and a number of local junior/new faculty members. A number of these students are now working partly in areas related to StoCom for the completion of their thesis research, and others have or are about to graduate and continue to work in the core area, or closely allied areas. Several core research activities catalysed by StoCom now include junior researchers or graduating students, as described below. Details of career trajectories of graduating students and postdoctoral fellows moving into faculty positions, and the impact of StoCom in defining some or all of their ongoing research, are given, along with information about ongoing research, papers and grants arising.

ISSUES FACED

StoCom was conceived by the SAMSI proposal planners as an inaugural program aimed at establishing the broad, multi-faceted synthesis program concept. It was defined with a view to engaging multiple participants. The key challenge faced, however, was that of the timetable – from establishment/award of SAMSI to opening, this was simply not conducive to engaging as broad, diverse and (particularly) senior research group as anticipated and as desirable. That the program involved so many researchers in the core research was, in retrospect, quite notable. However, a number of individuals involved in the early planning stages, especially at senior levels, did not, ultimately, engage and participate. Some of this was simply a matter of lack of enough time to establish sabbatical or other leaves to make the commitment. A number that did participate on short or longer term visits were not so actively involved in the core research, but, rather, engaged more as "arms-length" commentators rather than as participants. This was, in some cases, beneficial; in others, at best a distraction. With a now established track and long-term planning systematised, this difficulty in ensuring "engagement" of more senior individuals involved in a core, synthesis field of research will be much less of an issue than it was for this kick-off program working under such initial time constraints and with substantial pressure to "succeed" in year one. This is particularly critical with respect to the engagement and mentoring of new researchers - StoCom was well endowed and attended by junior researchers - students, postdoctoral researchers and new faculty alike - as core participants, but also attracted additional junior individuals from the local community. One result of the numbers was an inflation of the mentoring role and burden on the two program directors (Clyde and West) and SAMSI Director Berger (also a StoCom participant) as a result. The research developments, innovations and outcomes, and the impact on junior research directions in terms of thesis, postdoctoral positions and new collaborative projects, is substantial and indeed remarkable, though some additional core, longer-term senior research presence would have added measurably. Again, this can be mainly seen as a "teething problem" that should be less of an issue with programs based on longer-term planning and with hindsight as to the need for focus on ensuring that more senior individuals seen as key in engagement in program planning discussions do maintain their roles throughout the program.

STOCHASTIC COMPUTATION

ACTIVITIES, PERSONNEL & OUTCOMES

1. ACTIVITIES OF NSF/SAMSI SPONSORED GRADUATE STUDENTS

- Chris Hans (100% SAMSI graduate student)
  Chris was a full participant in all StoCom workshops, and was fully engaged in collaborative research on statistical graphical modelling and stochastic computation, and in presentations at workshops and undergraduate outreach meetings. Beyond the period of SAMSI support (Sept 2002-May 2003) he was supported on non-SAMSI Duke research funds to continue research in this area. He has done so, continuing in collaboration on graphical models with Duke faculty and postdoctoral researchers, and also extending some of the ideas for large-scale model search from StoCom to novel regression model search that defines the core of his PhD thesis. In particular, he has developed approaches to ***shotgun stochastic search*** for generalized linear models under the notion of sparse models (a few relevant predictors) in the $p >> n$ paradigm. His work has led to dramatic results in applications in gene expression genomics. Additional papers are in process based on this immediate outcome and follow-on research from StoCom. Thus, Hans's PhD research and thesis (expected in spring/early summer 2005) will be almost wholly based on this StoCom-originated research.
- Christine Kohnen (100% SAMSI graduate student)
  Christine was involved in variable selection for linear and generalized linear models. She implemented a deterministic Bayesian best subset procedure in R that overcomes variable limitations of the Volinsky/Raftery S-Plus search, and that extends the previous program to include priors that are scale mixtures of normals. She also implemented several of new Metropolis-Hastings algorithms for stochastic search, as well a novel orthogonal data augmentation algorithm for Bayesian model averaging.
- Beom Lee (100% SAMSI graduate student)
  Beom worked on several aspects of Monte Carlo simulation, MCMC methods, and sequential importance sampling in stochastic volatility models on stock prices, interest rate term structures, and option prices. He worked with StoCom researchers, led by Chuanshu Ji, on an approximation method using central limit theorems in MCMC computation of option prices. He completed his PhD at UNC in June 2003, and is now on the faculty in statistics at the University of Alabama, and continuing his collaboration with Ji (UNC) on MCMC and stochastic volatility models.
- German Molina (100% SAMSI graduate student)
  German was a full participant in both the model selection and financial mathematics working groups. In the former he developed an adaptive stochastic search algorithm for model selection based on a rejection sampling version of sampling without replacement. In the financial models component German worked on MCMC algorithms for estimation of multiple volatility scales. His research on these projects contributed part of his PhD dissertation, which he presented and successfully defended in July 2003. He then moved to a research position in financial statistics (as Quantitative Research Analyst, Credit Suisse First Boston, Fixed Income Division, London, England) and is continuing his research involving stochastic computation in financial time series.
- Michael Nicholas (100% SAMSI graduate student)
  Nicholas worked on probability in sampling, Monte Carlo methods, and complexity. He developed a MATLAB implementations of our new sampling algorithms for Hardy-Weinberg equilibrium. Part of this involved implementing a known, efficient algorithm for hypergeometric sampling called the "ratio-of-uniforms" approach by Stadlober and Zechner. Post-StoCom, Nicholas is continuing on the PhD program in Mathematics at Duke.

Several non-SAMSI funded graduate students, in RTP and from elsewhere in the US, were involved in research during the StoCom program and then arising from the program. They include:

- Carlos Carvalho (Duke), who became involved in the graphical models component and contributed very significantly to the research and development on methods of stochastic computation in non-decomposable models, in particular. He is now working on extension of this work, together with Duke faculty advisers

and ex-StoCom postdocs Dobra and Jones, while also extending methods of shotgun stochastic search for regression to non-standard regression contexts. Thus, Carvalho's PhD research and thesis (expected in spring/early summer 2006) will be very heavily based on this StoCom-originated research, with a range of applications in genomics.

- Ruriko (Rudi) Yoshida and Raymond Hemmecke (Mathematics, U.C. Davis, working with Professor Jesus DeLoera on lattice point enumeration arising from the contingency tables component through collaboration with core program faculty). Yoshida recently completed her PhD and will move to Duke University in 2004 to begin a postdoctoral fellowship in mathematics related to areas arising from this collaboration. The Duke connection arose primarily through her involvements in StoCom.
- Seth Sullivant (Mathematics, U.C. Berkeley), through involvement in the StoCom workshops, developed and continues a collaboration with Ian Dinwoodie and Yuguo Chen on novel stochastic methods for contingency tables – this represents a new collaboration involving junior researchers at the statistics:probability/mathematics interfaces.
- Sean Han (Mathematics, NCSU) worked with StoCom researchers German Molina and Professor Fouque to develop MCMC method for estimating stochastic volatility models with multiple stochastic volatility scales, and his thesis research included substantial coverage of this area. He graduated at NCSU with PhD (Mathematics) in July 2003, and is now a postdoc at IMA (Minneapolis). Han continues to work with Fouque on stochastic volatility models.

2. RESEARCH ACTIVITIES

StoCom activities included regular weekly meetings of the four topic research groups, and continual research interactions throughout each week. The program also included two major workshops (Sept/Oct 2002; June 2003) and a series of four "StoCom Days", i.e., a mini-workshop on each of the four topic areas, in late January and mid-late February 2003. Details of workshops and activities are available at the StoCom website, `www.samsi.info/sto.html`

Since this inaugural program, the concept of mid-program "mini-workshops" has been adopted as a model for year-long SAMSI programs.

2.1 Model and Variable Selection

The focus of this program was a synthesis of computational methods for model choice, with exploration of best current practice among existing methods, identification of their limitations and development of new methodology. Computation in model choice (selection of a single model or choosing to "average" over many models) involves calculation of marginal likelihoods (integration) for each model, and, in large problems where enumeration is prohibitive, search algorithms to identify a subset of models in order to narrow down comparisons to a manageable set. Research focused on the following areas:

- The comparison of current and newly proposed stochastic search algorithms for variable selection in linear regression models, including Gibbs samplers, Metropolis-Hastings algorithms with informative proposal distributions, deterministic search based on branch and bound algorithms, stochastic search using sampling without replacement, and perfect sampling. A primary focus was on approaches that scale-up with the dimension of the problem.
- The choice of prior distributions. Current default prior distributions for linear models, such as $g-$priors, lead to tractable computations, but not necessarily consistent model selection. The program explored scale mixtures of $g-$priors to create new classes of objective prior distributions with both desirable theoretical properties and computational tractability. We developed improved algorithms for implementing model selection with Cauchy distributions, one of the recommended scale mixtures that has not been widely utilized in practice.
- Methods for computing intractable marginal likelihoods and model probabilities. Comparison of existing methodology such as reversible jump algorithms (that do not require enumeration of models) to approaches that have typically required enumeration of all models. The latter includes Laplace approximations, importance sampling, ratio importance sampling, computation of marginal likelihoods using MCMC samples from each model or from the full model (in the case of variable selection). Continuing work beyond the end of the StoCom program continues development of innovative algorithms

that combine the novel search strategies proposed in linear models with the above marginal likelihoods computations to extend these approaches to model spaces where enumeration of models is not feasible.
- Software in R for implementing model selection/model averaging in linear models and extensions is available at the StoCom web-site.

2.2. Contingency Tables

The goals were to bring new computational methods to solve old and new problems of statistical inference for contingency tables. The group's expertise is in three areas: Markov Monte Carlo methods, "perfect" sampling methods, and sequential importance sampling (SIS). These three areas have seen theoretical breakthroughs in recent years.

The first research project was on sampling from tables of genotype data for testing Hardy-Weinberg equilibrium, as in the well-known paper of Guo and Thompson. We were interested in solving two problems with existing approaches: the complexity problem when the total number of individuals in the table is on the order of 10^4, and the "irreducibility" problem when certain allele combinations are lethal and force zero entries in the table. We were able to solve both of these problems in our paper "Monte Carlo Algorithms for Hardy-Weinberg proportions."

A second main project was on contingency tables with ordered covariates. A special case would be logistic regression. Our motivating data set is the well-known Ille-et-Verlaine oesophageal cancer data stratified by age in Breslow and Day (1980), where the data at each age level is quite sparse, and has two ordered covariates, tobacco level and alcohol level. Generalized linear models can be used to fit models of odds-ratios, but the sparsity of these tables makes these approximations unreliable. The problem for us is to develop sampling methods for tables with fixed sufficient statistics. Work has been done on Markov chains for this problem by Diaconis, Graham, and Sturmfels that involves the number theoretic concept of "homogeneous partition identity." But this work is not practical because the number of moves in the Markov chain grows exponentially and is not computable for large tables. A current goal is to find more practical Markov chains or other perfect sampling methods, and this project is ongoing.

StoCom catalysed a collaboration between Ian Dinwoodie (from probabilty), new researcher Yuguo Chen (statistics) and graduate student Seth Sullivant (U.C. Berkeley, mathematics) that continues beyond the end of the program and seems likely to develop significantly in the next year or two. This is in the area of "algebraic sequential importance sampling", which aims to bring methods from commutative algebra to to aid in the construction of sequential Monte Carlo techniques.

2.3 Graphical Models

This component involved a range of activities in the exploration, evaluation, synthesis and further development of effective computational methods for fitting graphical models, especially with a view to scaling up to large-scale models. Research continued throughout the StoCom period and well into Fall 2003, as the graphical models group remained largely intact – the two postdocs with affiliation to this program component moved to Duke University in fall 2004. Among the research foci of the program and of follow-on research are the following areas - substantial progress has been made across these areas, as represented in the papers and software arising and also the ensuing research trajectories of postdocs and graduate students.

- The synthesis and evaluation of current stochastic simulation methods for decomposable graphical models, including variants of Gibbs and Metropolis-Hastings methods. Various modifications of MCMC methods have been developed and evaluated.
- The extension of such methods to non-decomposable models, including implementation of approaches using importance sampling in graphical model space, and the development of approaches linked to complete-conditional regressions, hierarchical/triangular regressions and directed graphical models. New approaches developed this way were investigated, especially with regard to scale-up to higher-dimensions.
- The exploration and development of annealing and optimisation ideas to generate hybrid methods, especially methods for more rapidly generating large numbers of high probability candidate graphs. New approaches were defined based on novel methods we have termed *shotgun stochastic search* that have been found to be enormously effective in moderate dimensional problems where MCMC simply fails. This innovation led to related developments of such methods in problems of regression variable selection and uncertainty, also with substantial consequences.

- Theoretical and modelling questions that are critical to applications, including the specification of priors over graph space, tuning parameter evaluation, and measures of complexity.
- Aggressive development of distributed processing for implementation of stochastic computation algorithms for large-scale graphical models, utilising computer (Beowulf) clusters. Serial and cluster-relevant code developed for a range of graphical modelling computations are a deliverable of the program - as referenced below.
- Studies with real and simulated data, including a range of gene expression data sets. A key focus was on scale-up of models and computations to higher dimensions. Results in synthetic examples and applications in gene expression analysis confirm the major utility of some of the new approaches developed under this program, and continue to be core focus for research post-StoCom.

2.4 Financial Mathematics

The focus in this group was on Monte Carlo methods in the context of stochastic volatility (SV) models. Different backgrounds of the participants (probability, statistics and applied math) stimulated active interactions among different approaches and disciplines. Specific activities that defined the research were:

- Understanding of pros and cons of various algorithms. Many MCMC/SIS algorithms were reviewed, especially particle filters and some multi-move Metropolis methods. They are considered efficient when fitting historical SV models by only using underlying assets (no derivatives involved). These methods were explored in application a new class of models – multi-factor SV models with several well separated time scales, e.g. scales corresponding to fast (weekly) and slow (yearly) mean-reversion cycles of volatility. Extensive simulations and empirical studies were performed and written-up for publication.
- Improvement of calibration of SV models based on both stock and option data. For the task of option pricing in the context of SV models, it is inevitable to fit related risk-neutral dynamics by using option data in addition to stock prices. Much more intensive MCMC/SIS computation is required to perform numerical integration for an option price expressed as a conditional expectation under a risk-neutral measure. A key parameter – the volatility risk premium – has to be estimated as part of the Monte Carlo algorithm. Progress on this topic was made and led to follow-on research.

3. PAPERS, GRANT PROPOSALS & SOFTWARE

Core StoCom Research:

Papers listed below represent research that was initiated during StoCom, or part of the program research that has continued, and that is either now published, in press or under final journal review.

1. Y. Chen, I.H. Dinwoodie, A. Dobra and M. Huber (2003).
   Lattice Points, Sampling, and Contingency Tables Submitted to *Contemporary Mathematics*.
2. M. Clyde and E. George (2004)
   Model Uncertainty. *Statistical Science* (to appear).
3. A. Dobra, B. Jones, C. Hans, J.R. Nevins and M. West (2003)
   Sparse Graphical Models for Exploring Gene Expression Data. *Journal of Multivariate Analysis* (to appear).
4. A. Dobra, C. Tebaldi and M. West (2003)
   Bayesian Inference for Incomplete Multi-way Tables. Submitted for publication.
5. M. Huber, Y. Chen, A. Dobra, M. Nicholas and I.H. Dinwoodie (2003)
   Monte Carlo Algorithms for Hardy-Weinberg Proportions. Submitted to *Biometrics*.
6. C. Ji and B. Lee (2004)
   Central Limit Theorems in Computation of Option Prices with Stochastic Volatility Models. Submitted to *Mathematical Finance*.
7. B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter and M. West (2004)
   Experiments in Stochastic Computation for High-Dimensional Graphical Models. Submitted to *Statistical Science*.

8. B. Lee and C. Ji (2004)
   MCMC Calibration of Stochastic Volatility Models. Submitted for publication.
9. G. Molina, C.-H. Han and J.-P. Fouque (2003)
   MCMC Estimation of Multiscale Stochastic Volatility Models. Submitted to *Journal of Applied Econometrics*.

Core StoCom Research: Reports in preparation:

Papers listed below represent research that was initiated during StoCom, or part of the program research that has continued, and that is in the final write-up stage:

10. M. Clyde, G. Molina and M. Littman (2004)
    Bayesian Adaptive Stochastic Sampling for Variable Selection.
11. A. Cheng and C. Ji (2004)
    Gaussian Approximations in Option Pricing and Calibration of Stochastic Volatility Models.
12. C. Hans, A. Dobra and M. West (2004)
    Shotgun Stochastic Search for Regression Model Uncertainty and Exploration.
13. F. Liang, R. Paulo, G. Molina, M. Clyde and J. Berger (2004)
    Gaussian Hyper-Geometric and Other Mixtures of $g-$Priors for Bayesian Variable Selection.
14. R. Paulo, G. Molina, C. Kohnen, M. Clyde and J. Berger (2004)
    Stochastic Computation in Bayesian Variable Selection.
15. F. Wong, C. Carter and R. Kohn (2004)
    Testing for Structure in the Inverse Covariance Matrix.

Additional Papers:

Papers listed below represent work of StoCom participants and collaborators that was impacted by the program and completed during their participation:

16. C. Carter, F. Wong and R. Kohn (2003)
    Efficient Estimation of Covariance Selection Models. *Biometrika*, 90, 809-830.
17. C. Carter, F. Wong and R. Kohn (2003)
    A General Approach to Constructing Parameter Selection Priors. Submitted for publication.
18. Y. Chen and D. Small (2004)
    Testing the Rasch Model via Sequential Importance Sampling. *Psychometrika* (to appear).
19. E. Cripps, C. Carter and R. Kohn, R. (2004)
    Variable Selection and Covariance Selection in Multivariate Regression Models. *Handbook of Statistics: Bayesian Statistics: Modeling and Computation*, (eds: C. R. Rao and D. K. Dey), Elsevier Press (to appear).
20. I.H. Dinwoodie, L.F. Matusevich and E. Mosteig, E. (2004)
    Transform Methods for the Hypergeometric Distribution. *Statistics and Computing* (to appear).
21. I.H. Dinwoodie (2003).
    Estimation of Parameters in a Network Reliability Model with Spatial Dependence. Submitted to *Mathematics of Operations Research.*
22. I.H. Dinwoodie and B. MacGibbon (2003)
    Exact Analysis of a Paired Sibling Study Submitted for publication.

Grant Proposals:

Grant proposals submitted as a direct result of new research initiated in StoCom, or partly responding to developments in StoCom involving core participants, include the following:

Grants Awarded:

1. NSF-NIH *Joint Program on Mathematical Biology* (funded for five years, to begin in July 2004)
   - The research, entitled *Modelling of Graphs, Networks and Trees for Genomic Applications: High-Dimensional Model Search*, includes statistical modelling and stochastic computation in large-scale model search (graphical models, regression models) and the application of such methods in various areas of computational biology.
   - The PI is Mike West (StoCom program leader) and other Co-PIs and investigators include StoCom co-director Merlise Clyde, StoCom postdocs Adrian Dobra and Beatrix Jones, and StoCom PhD students Carlos Carvalho and Chris Hans.

Grants Currently Under Review:

2. NSF (SCREMS) (currently under review)
   - Research proposal for a Beowulf cluster in support of statistical research involving stochastic computation in various areas.
   - The PI is StoCom co-director Merlise Clyde, and investigators include junior StoCom participant Yuguo Chen.
3. NSF (DMS) (currently under review)
   - Proposal entitled *Large-scale Model Averaging and Model Selection*, on research that includes theory, methods and computation for Bayesian regression model selection and uncertainty.
   - The PI is StoCom co-director Merlise Clyde, and Co-PI is StoCom junior participant Feng Liang.
4. NIH (NIGMS) *Biological Networks Modelling* (currently under review)
   - Research proposal includes large-scale statistical graphical models and stochastic search, directly deriving from StoCom research, in applications in biological network exploration and modelling.
   - Mike West (StoCom program leader) is a Co-PI, and additional named investigators are StoCom postdocs Adrian Dobra and Beatrix Jones.

Software:

- *GGM*, software implementations of stochastic computation (Metropolis Hastings MCMC and shotgun stochastic search) for model exploration and selection in Gaussian Graphical Models
  B. Jones, A. Dobra, C. Hans, C. Carvalho and M. West (2003) Available at the ISDS web site *http://www.stat.duke.edu/isds-info/software/*.
- Code for building an interface between *Cocoa* (the Italian algebra program) and Markov chains simulations for contingency tables
  A. Dobra and I. Dinwoodie (2003) Available at *http://www.stat.duke.edu/ adobra/MarkovBasesAlgebra*.
- *bayesreg*, R code for model selection and model averaging for regression using leaps and bounds algorithms
  C. Kohnen (2003) Available at the ISDS web site *http://www.stat.duke.edu/isds-info/software/*.
- *BASS*, R library for Bayesian Adaptive Stochastic Search - performs adaptive sampling without replacement for model search in linear regression models
  M. Clyde and M. Littman (2004) Available at *http://www.stat.duke.edu/isds-info/software/*.

4. INDUSTRIAL AND GOVERNMENTAL PARTICIPATION

- StoCom had some, albeit limited participation of government and industry-based SAMSI advisers in program formulation.
- The StoCom workshops attracted a number of participants from governmental labs and agencies.
- One specific spin-off was a collaborative research project that involved SAMSI/StoCom postdoc Beatrix Jones working on statistical research in collaboration with the Dr Scott Nichols and colleagues at the National Marine Fisheries Service (Pascagoula – Mississippi Laboratory). Following StoCom, Jones moved to Duke to continue her postdoc research in stochastic computation, graphical models and genomics as

her primary activity. She also developed in this spin-off research collaboration for a component of her time, leading to one initial paper. Jones is now continuing to work in these areas, and will move to take up her new faculty position in the Department of Information and Mathematical Sciences at Massey University, Auckland, New Zealand, in summer 2004.

5. UNDER-REPRESENTED GROUPS

Through repeat advertising and professional networks, extensive efforts were made to bring the StoCom program and its workshops to the attention of women researchers and members of underrepresented groups. The program involved a substantial number of women, especially among the new researchers and students both in the core program and in the workshops.

6. ROLES OF CORE PARTICIPANTS (FACULTY, VISITORS, POSTDOCS)

6.1 Model and Variable Selection

- Merlise Clyde (Senior faculty - Duke, statistics)
  Clyde was co-director of the StoCom program and leader of the component on model selection. She directed the research on this component, was involved in administration of the workshops and visitor activities, and had the lead role in developing manuscripts and web based records of this component.
- Jim Berger (SAMSI Director)
  Berger was involved in research in objective Bayes methodology in variable selection, and played a pivotal role in mentoring post-docs and graduate students.
- Joe Ibrahim (Senior faculty - UNC, biostatistics)
  Ibrahim contributed to work on the calculation of marginal likelihoods in the model selection component
- Bani Mallick (Senior faculty - Texas A& M)
  Mallick (short term visitor) contributed to the discussion and development of research topics, including perfect sampling in non-parametric high-dimensional wavelet regression models.
- Mark Huber (Junior faculty - Duke, mathematics and statistics)
  While primarily involved in the Contingency Table working group, Huber contributed new ideas for improving perfect sampling in orthogonal regression, such as wavelet regression and implementation of perfect sampling methods for general variable selection problems.
- Feng Liang (Junior faculty - Duke, statistics)
  Liang's research centered on objective prior specification for variable selection.
- Helen Zhao (Junior faculty - NCSU, statistics)
  Zhao's research in non-parametric regression methods bridges classical and Bayesian approaches, and Zhao was involved in most activities and meetings on the model selection component.
- Rui Paulo (75% SAMSI postdoc)
  Paulo was involved in research on objective Bayes prior specifications and was been instrumental in implementing methods for calculating marginal likelihoods.

6.2. Contingency Tables

- Ian Dinwoodie (Senior faculty - Tulane, mathematics – SAMSI Fellow)
  Dinwoodie was co-director of the contingency tables research component, covering research direction and administration. He generated much of the research emphasis at the interface between mathematics and statistics, and has a lead role in developing manuscripts and web based records of this component.
- Mark Huber (Junior faculty - Duke, mathematics and statistics)
  Huber was co-director of the contingency tables research component, covering research direction and administration. He generated much of the research emphasis at the interface between probability and statistics, and had a lead role in developing manuscripts and web based records of this component.
- Yuguo Chen (Junior faculty - Duke, statistics)
  Chen was involved in research in designing and implementing efficient sequential importance sampling methods for zero-one and contingency tables. He was actively involved in the collaboration on this component, and continues to work in the area.

- Adrian Dobra (15% SAMSI postdoc)
  Dobra was involved in research in modelling and computation in contingency tables, especially the development and implementation of MCMC and sequential importance sampling methods for higher-dimensional tables subject to observations on only selected margins. He played a critical and central role in this component, and was also involved in presentations at workshops and undergraduate outreach meetings.

6.3 Graphical Models

- Mike West (Senior faculty - Duke, statistics)
  West was director of the StoCom program and leader of the graphical models component. He provided overall research direction, as well as administration and mentoring for junior participants. West had the lead role in developing manuscripts and web based records of this component. West was also involved in the research, with Dobra, on the contingency tables component.
- Chris Carter (Senior faculty - Hong Kong, statistics – SAMSI Fellow)
  Carter was a visitor (year long) involved in the StoCom research on graphical models. His contributions built on his expertise in the area and he played a role in mentoring junior participants.
- Beatrix Jones (100% SAMSI postdoc)
  Jones was involved in research in modelling and computation, especially in connection with MCMC methods in decomposable and non-decomposable Gaussian graphical models, and the interfaces with applications in gene expression analysis. Jones also played a collaborative role in mentoring graduate student participants, and in presentations at workshops and undergraduate outreach meetings. Jones continued to develop her work and career path in these areas, and will take up a new faculty position (Massey University NZ) in summer 2004.
- Adrian Dobra (15% SAMSI postdoc)
  Dobra was critically and centrally involved in research in modelling and computation in graphical models, particularly in connection with manipulation of large-scale graphs, annealing methods and novel approaches to stochastic search in graphical models, the interfaces with approaches based on regression models, and cluster-based distributed computation. Dobra was also involved in presentations at workshops and undergraduate outreach meetings. Dobra has continued his work in this area, and is emerging as a lead researcher in graphical models and computation, with applications in computational biology. He currently holds positions as Research Assistant Professor of Molecular Genetics and Statistics at Duke University, and is a core faculty member of Duke's Computational and Applied Genomics Program.

6.4 Financial Mathematics

- Jean-Pierre Fouque (Senior faculty - NCSU, applied mathematics)
  Fouque directed the financial mathematics component, covering research direction and administration. He generated much of the research emphasis at the interface between mathematics and statistics, guided the research and mentored junior participants.
- Chuanshu Ji (Senior faculty - UNC, statistics)
  Ji was involved in collaborative StoCom research concerned with models of interest rate term structures, and option prices. He mentored junior researchers, including Beom Lee, on the development of theoretical as well as computational aspects of MCMC for option price problems.
- Yuguo Chen (Junior faculty - Duke, statistics)
  Chen interacted with other participants in the program, particularly in connection with exploration of methods of sequential importance sampling methods for filtering and smoothing problems in stochastic volatility models.
- Several other junior faculty were participants in meetings and research discussions of the StoCom financial models group, though not core participants. These included Tao Pang (Junior faculty - NCSU, applied mathematics), Sujit Ghosh (Senior faculty - NCSU, statistics) and Ana Valeva (Junior visiting faculty - Duke).

Note: Extensive information about the workshops of the program is available in the SAMSI Annual Reports for Years 2002-03 and 2003-04. In particular, these annual reports contain the programs, abstracts, and lists of workshop participants for each of the program workshops.

Publications and Technical Reports (added to the report later)

- Carter, C., F. Wong and R. Kohn *"Efficient Estimation Of Covariance Selection Models"* Biometrika, 90, 809-830. SAMSI 2003-12, March 10, 2003

- Chen, Y. and D. Small *"Testing the Rasch Model via Sequential Importance Sampling"* (2004) Psychometrika (to appear).

- Chen, Y., I.H. Dinwoodie, A. Dobra and M. Huber *"Lattice Points, Sampling, And Contingency Tables"* Contemporary Mathematics, 374: Integer Points in Polyhedra—Geometry, Number Theory, Algebra, Optimization, ed. A. Barvinok (2005) 65-78 (with Y. Chen, A. Dobra, and M. Huber.) SAMSI 2003-9, July 30, 2003

- Clyde, M. and E. George *"Model Uncertainty"* Statistical Science SAMSI 2003-16, December 2003

- Cripps, E., C. Carter and R. Kohn, R. *"Variable Selection and Covariance Selection in Multivariate Regression Models"* Handbook of Statistics: Bayesian Statistics: Modeling and Computation, (eds: C.R. Rao and D.K. Dey), Elsevier Press. SAMSI 2003-14, December 16, 2003

- Dinwoodie, I.H., L.F. Matusevich and E. Mosteig, E. *"Transform Methods for the Hypergeometric Distribution"* Statistics and Computing, 14 (2004), 287-297 (with E. Mosteig and L. Matusevich). SAMSI 2003-15, November 4, 2003.

- Dinwoodie, I.H. *"Estimation of Parameters in a Network Reliability Model with Spatial Dependence."* Submitted to Mathematics of Operations Research. ESAIM Probabilit\'es et Statistiques 9 (2005), 241-253. SAMSI 2003-11, October3, 2003

- Dinwoodie, I.H. and B. MacGibbon *"Exact Analysis of a Paired Sibling Study"* Computational Statistics 19 (2004), 525-534 (with B. MacGibbon). SAMSI 2003-10

- Dobra, A., B. Jones, C. Hans, J.R. Nevins and M. West (2003) *"Sparse Graphical Models for Exploring Gene Expression Data"* Journal of Multivariate Analysis 90: 196-212. SAMSI 2003-7, October 23, 2003

- Dobra, A., C. Tebaldi and M. West *"Bayesian Inference for Incomplete Multi-way Tables"* SAMSI 2003-2, February 27,2003
- Dobra, A., Tebaldi, C. and West, M. *"Data Augmentation in Multi-Way Contingency Tables with Fixed Marginal Totals"* Journal of Statistical Planning and Inference, 136, 355-372

- Huber, M., Y. Chen, A. Dobra, M. Nicholas and I.H. Dinwoodie *"Monte Carlo Algorithms for Hardy-Weinberg Proportions"* Submitted to Biometrics. SAMSI 2003-8, May 2003

- Ji, C. and B. Lee *"Central Limit Theorems in Computation of Option Prices with Stochastic Volatility Models"* Submitted to Mathematical Finance (2004)

- Jones, B., C. Carvalho, A. Dobra, C. Hans, C. Carter and M. West *"Experiments in Stochastic Computation for High-Dimensional Graphical Models"* Statistical Science 20:388-400 SAMSI 2004-1, January 9, 2004

- Jones, B. and M. West *"Covariance Decomposition for Undirected Gaussian Graphical Models"* Biometrika (92) 779-786 (2005)

- Lee, B. and C. Ji *"MCMC Calibration of Stochastic Volatility Models"* Submitted for publication. (2004)

- Molina, G., C.H. Han and J.P. Fouque *"MCMC Estimation of Multiscale Stochastic Volatility Models"* SAMSI 2003-6, June 3, 2003

- Rich, J.N., C. Hans, B. Jones, E.S. Iversen, R. E. McClendon, B.K. A. Rasheed, A . Dobra, H.K. Dressman, D.D. Bigner, J.R. Nevins and M.West *"Gene Expression Profiling And Analysis In Graphical Association Studies In Glioblastoma Survival"* Cancer Research, 65: 4051-5058 (2005)

Reports in Preparation

- Cheng, A. and C. Ji *"Gaussian Approximations in Option Pricing and Calibration of Stochastic Volatility Models"* (2004)

- Clyde, M., G. Molina and M. Littman *"Bayesian Adaptive Stochastic Sampling for Variable Selection"* (2004)

- Hans, C., A. Dobra and M. West *"Shotgun Stochastic Search for Regression Model Uncertainty and Exploration"* (2004)

- Liang, F., R. Paulo, G. Molina, M. Clyde and J. Berger *"Gaussian Hyper-Geometric and Other Mixtures of g– Priors for Bayesian Variable Selection"* (2004)

- Paulo, R., G. Molina, C. Kohnen, M. Clyde and J. Berger *"Stochastic Computation in Bayesian Variable Selection"* (2004)

- Wong, F., C. Carter and R. Kohn *"Testing for Structure in the Inverse Covariance Matrix"* (2004)