

Symposium Program  
Monte Carlo in the New Millennium

Department of Statistics  
University of Florida

January 11–13, 2001

# Organizing Committee

Jim Booth, George Casella, Jim Hobert, Ranjini Natarajan, Brett Presnell.

## Contents

<b>Thursday, January 11</b>	<b>2</b>
<b>Friday, January 12</b>	<b>3</b>
<b>Saturday, January 13</b>	<b>4</b>
<b>Abstracts</b>	<b>5</b>
A Personal History of MCMC . . . . .	5
Scaled and Layered MultiShift Coupling for Perfect Simulation . . . . .	5
MCMC Sampler Convergence Rates for Hierarchical Normal Linear Models: A Simulation Approach . . . . .	5
ARE in MCMC (Asymptotic relative efficiency in Markov chain Monte Carlo) . . . . .	6
Bootstrap Tilting . . . . .	6
From Phylogenetic Inference to Counting zero-one Tables . . . . .	6
Testing Random Number Generators . . . . .	7
Perfect Sampling for Spatial Point Processes . . . . .	7
Hybrid Network-MCMC Methods for Conditional Logistic Regression . . . . .	7
Moralizing Perfect Sampling and Perfect Slice Sampling for Mixtures: Two Antagonistic Visions of Perfect Sampling . . . . .	8
Laplace approximations in MCMC algorithms . . . . .	8
Joint distributions, conditional distributions and the Gibbs sampler . . . . .	8
Categorical Data, Groebner Bases, and Exact Monte Carlo Sampling . . . . .	9
Hierarchical Models for Spatio-Temporal Processes . . . . .	9
<b>Poster Abstracts</b>	<b>10</b>
One-Sided Range Test for Testing Against an Ordered Alternative Under Heteroscedasticity . . . . .	10
A Stochastic Dictionary Model for Motif Discovery in Biological Sequences . . . . .	10
A Procedure for Generating Nonnormal Distributions for Monte Carlo Studies . . . . .	10
Discriminant Analysis with Missing Data Imputation . . . . .	11
Two-gene modeling of complex traits via Markov chain Monte Carlo . . . . .	11
Ordering and Improving Monte Carlo Markov Chains . . . . .	12
Improving Power of the Goodness of Fit tests through Penalized Disparity Measures . . . . .	12
Bayesian Risk Analysis of Radon Exposure Data from the Iowa Radon Lung Cancer Study . . . . .	13
A Robust I-Sample Analysis of Means Type Randomization Test for Variances for Unbalanced Designs . . . . .	13
Empirical Bayes Estimators for Borel-Tanner Distribution . . . . .	13

**Thursday, January 11**

7:00–10:00 p.m. Reception

Keene Faculty  
Center (Dauer Hall)

## Friday, January 12

8:30–9:00 a.m.	Breakfast	349 Reitz Union
9:00–10:20 a.m.	SESSION 1: OPENING SESSION Chair: Brett Presnell	349 Reitz Union
	Neil Sullivan, Dean    Opening Remarks	
	George Casella        A Personal History of MCMC	
	George Marsaglia      Testing Random Number Generators	
10:20–10:40 a.m.	Break	
10:40–Noon	SESSION 2: IMPORTANCE SAMPLING AND THE BOOTSTRAP Chair: Jim Booth	349 Reitz Union
	Tim Hesterberg        Bootstrap Tilting	
	Jun Liu                 From Phylogenetic Inference to Counting Zero-One Tables	
Noon–1:30 p.m.	Lunch	Arredondo Room (Reitz Union)
1:30–2:50 p.m.	SESSION 3: CONDITIONAL INFERENCE Chair: Sam Wu	349 Reitz Union
	Cyrus Mehta          Hybrid Network-MCMC Methods for Conditional Logistic Regression	
	Martin Wells          Categorical Data, Groebner Bases, and Exact Monte Carlo Sampling	
2:50–3:20 a.m.	Group Photograph & Break	
3:20-4:40 p.m.	SESSION 4: CONVERGENCE AND COMPATIBILITY Chair: Pam Ohman	349 Reitz Union
	Kate Cowles          MCMC Sampler Convergence Rates for Hierarchical Normal Linear Models: A Simulation Approach	
	J. Sethuraman         Joint Distributions, Conditional Distributions and the Gibbs Sampler	
6:00-7:30 p.m.	POSTER SESSION	Friends of Music Room (University Auditorium)

## Saturday, January 13

8:30–9:00 a.m.	Breakfast	349 Reitz Union
9:00–10:20 a.m.	SESSION 5: PERFECT SAMPLING Chair: Ranjini Natarajan	349 Reitz Union
	<b>Christian Robert</b>	<b>Moralizing Perfect Sampling and Perfect Slice Sampling for Mixtures: Two Antagonistic Visions of Perfect Sampling</b>
	<b>Jem Corcoran</b>	<b>Scaled and Layered MultiShift Coupling for Perfect Simulation</b>
10:20–10:40 a.m.	Break	
10:40–Noon	SESSION 6: SPATIAL STATISTICS Chair: Alex Trindade	349 Reitz Union
	<b>Ian McKeague</b>	<b>Perfect Sampling for Spatial Point Processes</b>
	<b>Chris Wikle</b>	<b>Hierarchical Models for Spatio-Temporal Processes</b>
Noon–2:00 p.m.	Lunch	
2:00–3:20 p.m.	SESSION 7: EFFICIENCY OF MCMC ALGORITHMS Chair: Jim Hobert	349 Reitz Union
	<b>Judith Rousseau</b>	<b>Laplace Approximations in MCMC Algorithms</b>
	<b>Charles Geyer</b>	<b>ARE in MCMC (Asymptotic Relative Efficiency in Markov Chain Monte Carlo)</b>
4:30–8:30 p.m.	Barbecue	Chez Casella

## Abstracts

### A Personal History of MCMC

George Casella  
University of Florida

The title says it all.

### Scaled and Layered MultiShift Coupling for Perfect Simulation

Jem Corcoran  
University of Colorado

Perfect simulation or "coupling-from-the-past" involves running different sample paths of a Markov chain until they coalesce or meet. For a continuous state space, getting sample paths to meet can be a troublesome task. The most common approach involves finding a "minorization" condition that applies to all or part of the state space— this is often easier said than done. As an alternative, we consider an easy to use variant of a recently introduced coupling scheme known as "layered multishift coupling". Layered multishift coupling allows one to draw a common value from a distribution and a shifted version of that distribution. This variant algorithm allows one to draw common values from both shifted and re-scaled versions of the distribution as well as, in some cases, from two completely different distributions. We apply this in the simulation of some storage models and for sampling from the the auto-gamma distribution. In the latter example, a non-invertibility problem is circumvented.

### MCMC Sampler Convergence Rates for Hierarchical Normal Linear Models: A Simulation Approach

Kate Cowles  
University of Iowa

We present a straightforward method of approximating theoretical bounds on burn-in time for MCMC samplers for hierarchical normal linear models. An extension and refinement of Cowles and Rosenthal's (1998) simulation approach, it exploits Hodges's (1998) reformulation of hierarchical normal linear models. The method is illustrated with real datasets, involving a one-way variance components model, a growth-curve model, and a spatial model with a pairwise-differences prior. In all three cases, when the specified priors produce proper, unimodal posterior distributions, the method provides very reasonable upper bounds on burn-in time. In contrast, when the posterior distribution for the variance-components model can be shown to be improper or bimodal, the new method correctly identifies convergence failure while several other commonly-used diagnostics provide false assurance that convergence has occurred.

## ARE in MCMC (Asymptotic relative efficiency in Markov chain Monte Carlo)

Charles J. Geyer  
University of Minnesota

How can one say whether one MCMC scheme is better than another? Although several other criteria have been used in the literature, the criterion that relates to actual practice is ARE. Almost nothing theoretical can be said about the ARE of an MCMC scheme estimating one particular expectation. Much more can be said if we consider the ARE for all expectations for which ARE is defined (where there is a CLT with root  $n$  rate). This leads to a partial ordering of MCMC schemes we call "covariance ordering". We will explain what can be proved about and with this ordering.

Joint work with Antonietta Mira (U. of Insubria).

## Bootstrap Tilting

Tim Hesterberg  
MathSoft, Inc.

Bootstrap tilting confidence intervals could be the method of choice in many applications for reasons of both speed and accuracy. With the right implementation, tilting intervals are 37 times as fast as bootstrap BC- $a$  limits, in terms of the number of bootstrap samples needed for comparable simulation accuracy. Thus 100 bootstrap samples might suffice instead of 3700.

Tilting limits have other desirable properties — second-order accuracy, transformation invariance, and better finite-sample coverage and/or shorter intervals on average than competing procedures.

Bootstrap tilting also is useful for diagnostic purposes. They provide an immediate warning against the most common error in using the bootstrap.

## From Phylogenetic Inference to Counting zero-one Tables

Jun Liu  
Harvard University

In this talk we describe the basic sequential importance sampling (SIS) methodology, tracing its history in molecular simulation and reviewing its applications in diverse research fields. In particular, we will illustrate two recent success stories of the methodology, one for phylogenetic inference and another for counting the total number of zero-one tables with fixed margins. To a certain extent, we can claim that the table counting problem (including zero-one tables and the regular contingency tables) has been essentially solved by our SIS method.

The SIS method can be understood as a way to sequentially/recursively construct an importance sampling distribution for high-dimensional problems and it produces weighted multiple samples as its end result. With these multiple samples, new information can be easily "learned" by adjusting the associated importance weights. The recursive nature of state-space models make it ideal to develop nonlinear filters based on the SIS strategy. Since the importance weights tend to be more and more skewed as the system evolves, ideas of resampling, rejection sampling, and kernel density estimation are necessary for the control of Monte Carlo variations in SIS. Success stories of the method include energy minimization for polymer folding, target tracking, digital telecommunications, and table counting.

Based on joint work with Yuguo Chen.

## Testing Random Number Generators

George Marsaglia  
Florida State University

A Description of the Marsaglia Random Number CDROM with the Diehard Battery of Tests of Randomness, together with three new difficult-to-pass tests that many commonly used random number generators fail.

## Perfect Sampling for Spatial Point Processes

Ian McKeague  
Florida State University

This talk gives a review of perfect simulation algorithms that have been introduced recently for spatial point processes. The utility of these algorithms for implementing Bayesian inference in spatial cluster models (and for the Neyman-Scott cluster model in particular) is discussed. An application to data on leukemia incidence in upstate New York is presented.

## Hybrid Network-MCMC Methods for Conditional Logistic Regression

Cyrus R. Mehta  
Cytel Software Corporation and Harvard School of Public Health

Numerical algorithms for performing exact logistic regression are severely limited by the computational demands of enumerating all possible permutations of the binary response variable. Monte Carlo methods attempt to overcome this limitation by sampling from the reference set of all possible permutations. Two competing Monte Carlo methods are the network based independent sampling approach of Mehta, Patel and Senchaudhuri (JASA, 2000) and the Markov Chain Monte Carlo (MCMC) sampling approach of Foster, McDonald and Smith (JRSS, 1996). There are drawbacks to both approaches. The memory requirements of the network approach are often exceeded even for data sets of modest size. While MCMC imposes no such memory limitations it requires that the underlying Markov chain be ergodic, and that the sampling continue until the steady state is reached. When these requirements are violated MCMC produces incorrect answers. It is, moreover, very difficult to know in advance if these requirements are met. In this talk we present a hybrid network-MCMC algorithm that combines the strength of each approach and extends Monte Carlo methods beyond what could be achieved by each approach individually. We apply the method to toxicity data from a clinical trial of children in an intensive care unit.

This is joint research with Dr. Nitin Patel and Dr. Pralay Senchaudhuri.

## **Moralizing Perfect Sampling and Perfect Slice Sampling for Mixtures: Two Antagonistic Visions of Perfect Sampling**

Christian Robert

Université de Paris IX - Dauphine

In this talk, we will oppose two recent papers with Jim Hobert, and with George Casella, Kerrie Mengersen and Mike Titterton, to show that, while we can formally obtain a general representation of the stationary measure of a recurrent Markov chain as an infinite mixture, and thus achieve perfect sampling without coupling, the practical implementation of perfect sampling for realistic setups like mixtures of distributions is a very delicate enterprise. To illustrate this opposition, we will consider several perfect samplers, some of which rely on a marginalisation akin to Rao-Blackwellisation. We show that a genuine perfect slice sampler can be implemented for small sample sizes only and introduce an alternative perfect sampler based on a single backward chain, which can handle much larger sample sizes, but which is essentially equivalent to an accept-reject algorithm, a “non-result” also achieved by Corcoran and Tweedie (2000).

## **Laplace approximations in MCMC algorithms**

Chantal Guihenneuc, Judith Rousseau

We consider a latent variable model on longitudinal data. To model simultaneously the latent process and the link between the observations and the latent process, we use a Bayesian hierarchical structure (as in Guihenneuc et al (2000)). The posterior distribution is obtained via MCMC algorithms. The number of parameters to be sampled by the algorithm is then of the same order than the number of observations and so requires a huge running time. We propose Laplace approximations of the posterior distribution to get rid of the nuisance parameters and thus to improve on this running time. We prove that the stationary distribution of the algorithm (the modified target distribution) is close to the true target distribution (which is the posterior distribution of the parameter of interest). We also present simulations which illustrate the good behaviour of the approximation.

## **Joint distributions, conditional distributions and the Gibbs sampler**

Krishna B. Athreya and Jayaram Sethuraman

Cornell University and Florida State University

Let  $P$  and  $Q$  be transition functions on  $S_1 \times S_2$  and  $S_2 \times S_1$ , respectively. This paper explores conditions for the existence and uniqueness of a joint distribution  $\pi$  with conditional distributions  $P$  and  $Q$  as well as the convergence of the associated Gibbs sampler to this  $\pi$ . Roughly speaking, what is needed is a multiplicative condition on  $P$  and  $Q$  with appropriate integrability and an irreducibility condition on  $R = PQ$ . Examples are given to illustrate the consequences of the violation of some of these conditions and to demonstrate that the mere convergence of the Gibbs sampler does not insure the uniqueness of the joint distribution. It is also shown that Markov chains arising in Gibbs sampling are necessarily aperiodic. Similar results are obtained for the case of more than two variables.

## Categorical Data, Groebner Bases, and Exact Monte Carlo Sampling

George Casella and Martin T. Wells

Exact inference in contingency tables, and other categorical data problems, is accomplished by calculating a reference distribution that typically involves the enumeration of samples satisfying certain constraints. The landmark paper of Diaconis and Sturmfels was the first to connect the algebraic geometry, notable theory of Groebner bases, with the solutions of constrained, discrete problems. They focused on MCMC, and made connections between Groebner bases and Markov bases, that is, they showed how to construct a random walk of the set of all tables that was guaranteed to reach every table. We propose a more direct MCMC approach that takes advantage of the fact that the network representation of the data is a zero-dimensional ideal.

## Hierarchical Models for Spatio-Temporal Processes

Christopher K. Wikle

University of Missouri-Columbia

Spatio-temporal processes are ubiquitous in the environmental and physical sciences. This is certainly true of atmospheric and oceanic processes, which typically exhibit many different scales of spatial and temporal variability. The complexity of these processes and large number of observation/prediction locations preclude the use of traditional covariance-based space-time statistical methods. Alternatively, we have investigated conditionally-specified (i.e., hierarchical) spatio-temporal models. These methods offer several advantages over traditional approaches:

- Physical and dynamical constraints are easily incorporated into the conditional formulation, so that the series of relatively simple, yet physically realistic, conditional models leads to a much more complicated joint space-time covariance structure than can be specified directly.
- Spectral representations naturally allow spatial nonstationarities, reduce dimensionality, and introduce sparseness, and are easily included in the conditional framework. This greatly facilitates computation with massive data sets.

These models would be very difficult or impossible to implement without MCMC. However, their high-dimensionality and complexity can still present many problems, both practical and theoretical.

## Poster Abstracts

### One-Sided Range Test for Testing Against an Ordered Alternative Under Heteroscedasticity

Shun-Yi Chen  
Tamkang University

In a one-way fixed effects analysis of variance model, when normal variances are unknown and possibly unequal, a one-sided range test for testing the null hypothesis  $H_0 : \mu_1 = \dots = \mu_k$  against an ordered alternative  $H_a : \mu_1 \leq \dots \leq \mu_k$  by a single-stage and a two-stage procedure, respectively, is proposed. The critical values under  $H_0$  and the power under a specific alternative are calculated. Relation between the one-stage and the two-stage test procedures is discussed. A numerical example to illustrate these procedures is given.

### A Stochastic Dictionary Model for Motif Discovery in Biological Sequences

Mayetri Gupta and Jun S. Liu  
Harvard University

The biotechnology revolution, along with the human genome project, has led to a rapid growth of the large public databases of genetic data. One of the important problems in analyzing these huge sequence databases is the identification of sequence motifs through multiple local alignment, which often provides important clues to understanding the functional relationship between genes and how the expressions of the genes are regulated. We here build on the concept of a ‘Dictionary Model,’ which treats motifs in DNA (or protein) sequences as ‘words’ drawn with certain probabilities from Nature’s ‘dictionary’. We formulate this as a missing data problem — the missing data being both the identity of the ‘words’ and the ‘segmentations’ of the DNA sequence into the possible words. We then extend the concept of a ‘word’ to that of a probabilistic ‘word matrix’ which takes into account the possibility of non-exact occurrences of the motif in a DNA sequence. This stochastic dictionary model allows us to develop a Markov Chain Monte Carlo approach to estimate probabilities of the occurrences of the ‘words’ in our current dictionary, the unknown probabilistic word matrices, and identify likely motifs in any number of DNA sequences.

### A Procedure for Generating Nonnormal Distributions for Monte Carlo Studies

Todd C. Headrick, Ph.D.  
Southern Illinois University at Carbondale

A procedure is derived for simulating univariate and multivariate nonnormal distributions. Specified marginal or joint distributions are generated from fifth degree polynomial transformations of standard Gaussian pseudo-random deviates. The coefficients of the polynomials are determined by simultaneously solving a system of six equations for the first six standardized cumulants of the desired nonnormal distribution(s). The proposed procedure is bounded in terms various combinations of skew and kurtosis. Boundary calculations are subsequently made using Lagrange multiplier techniques. Examples of various nonnormal distributions (e.g., Gamma, Beta, chi-square, double exponential, logistic, etc.) are provided to demonstrate the data generation procedure. Monte Carlo results indicate that the proposed procedure yields excellent agreement between population parameters and average values of intercorrelation.

## **Discriminant Analysis with Missing Data Imputation**

Mi Ok Kim  
University of Illinois

Abstract : When discriminant analysis on an existing data with variable selection procedure requires variables some of which are missing on a test dataset, compromise is usually made that a test data set is investigated with rules established based on discriminant analysis on complete variables of the existing data. In some cases, such compromise obligate compromise in accuracy of the analysis because resulting rules have higher crossvalidation error rate. Missing data imputation provides a means to get around with missing data problem. Imputed values allow analysis to be based on the optimal set of variables that is decided by variable selection procedure. Simulation can provide means to check the precision. An example is presented where multiple variables are missing. Imputation consists of multivariate regression and vector-valued random drawing. A simulation of size 100 is included to provide precision of the analysis.

## **Two-gene modeling of complex traits via Markov chain Monte Carlo**

Yuqun Luo  
Ohio State University

Segregation Analysis of a disease focuses on elucidating the relation between a person's gene composition and his susceptibility to the disease.

Most common and chronic human genetic diseases, such as asthma, are complex traits, where more than one gene and environmental factors are involved in the etiology of the disease. It is thus natural to investigate two-gene models, where two genes act on the disease. It is vital to incorporate information from genetic markers (known chromosomal regions) in our analysis for such traits, to improve the accuracy of estimation for the large number of parameters of interest. In order to retain all information, the ability to analyze large, complex genealogies from homogeneous populations without breaking them down is also crucial. The inclusion of marker information and desire to analyze complex genealogies as a whole pose challenges beyond the limit of standard statistical methods.

We focus on developing a general approach to the segregation analysis for two-gene models, with the ability to incorporate marker information and handle complex pedigrees as a whole. A Bayesian Markov chain Monte Carlo approach is adopted to obtain accurate estimates of the parameters within a given class of models. Simulation studies and recent application to an asthma dataset has yielded encouraging results. We plan to propose and study criteria and algorithms to select between one-gene models and two-gene models, and among various classes of two-gene models.

## Ordering and Improving Monte Carlo Markov Chains

Antonietta Mira  
University of Insubria

The class of Markov chains having a specified stationary distribution is very large, so it is important to have criteria telling when one chain performs better than another for Markov chain Monte Carlo (MCMC) purposes. Efficiency of a MCMC sampler is measured by the asymptotic variance of the resulting estimates (absolute efficiency).

We introduce a few partial orders defined on this class and discuss ways of improving a given MCMC sampler relative to these orders.

In particular two strategies will be considered.

On general state spaces the *delaying rejection strategy* can be used to improve the performance of a given Metropolis-Hastings algorithm in the Peskun order.

On finite state spaces we introduce the idea of stationarity preserving and efficiency increasing *probability mass transfers* (PMT). Given a transition matrix by applying a sequence of PMT we obtain a Markov chain that estimates at best (with the smallest asymptotic variance up to a first order approximation) the mean (with respect to the stationary distribution of the chain) of a specific function of interest (relative efficiency).

## Improving Power of the Goodness of Fit tests through Penalized Disparity Measures

Surajit Ray  
Penn State University

The Pearson's chi-square and the log likelihood ratio chi-square are fundamental tools in goodness-of-fit testing and have been popular methods in applied statistics for a long time. Both of these are measures of discrepancies between the hypothesized model probabilities and the observed data proportions. Cressie and Read (1984) demonstrated that there exists a general family of divergences which includes both of the above two test statistics as special cases. This family is indexed by a single parameter, and it appears that divergences at either end of the scale are more powerful against alternatives of one type while being rather poor against alternatives of the other type. In this paper we present several alternative goodness-of-fit testing procedures which have reasonably high power at both kinds of alternatives and hence may be more preferred techniques. Some of these alternatives are modifications of the Cressie-Read statistics, while others are entirely new divergences. An extensive numerical and graphical study illustrates the advantage of the new methods.

## **Bayesian Risk Analysis of Radon Exposure Data from the Iowa Radon Lung Cancer Study**

Brian J. Smith and Mary Kathryn Cowles  
University of Iowa

We present a Bayesian approach to modeling the association between lung cancer risk and residential radon exposure measured with error. Markov chain Monte Carlo (MCMC) methods are used to fit our model to data from the Iowa Radon Lung Cancer Study. The MCMC methods are implemented with a fast and efficient computer algorithm that facilitates the analysis of large data sets. The disease and spatial radon process are jointly modeled so that both the measured radon concentrations and the risk information are used to estimate the true radon concentrations. We discuss the potential for improved exposure and risk estimates by comparing our methods and results with those from the frequentist analysis of the Iowa data.

## **A Robust I-Sample Analysis of Means Type Randomization Test for Variances for Unbalanced Designs**

Ping Sa and Peter Wludyka  
University of North Florida

An Analysis of Means (ANOM) type randomization test for testing the equality of I variances for unbalanced designs is presented. Randomization techniques for testing statistical hypotheses can be used when parametric tests are inappropriate. Suppose that I independent samples have been collected. Randomization tests are based on shuffles or rearrangements of the (combined) sample. Putting each of the samples “in a bowl” forms the combined sample. Drawing samples “from the bowl” forms a shuffle. Shuffles can be made with replacement (bootstrap shuffling) or without replacement (permutation shuffling). The test that is presented offers two advantages. It is robust to non-normality and it allows the user to graphically present the results via a decision chart similar to a Shewhart control chart. A Monte Carlo study is used to verify that the test exhibits excellent power when compared to other robust test.

## **Empirical Bayes Estimators for Borel-Tanner Distribution**

George Yanev  
University of South Florida

The Borel-Tanner probability distribution was derived by Borel (1942) and Tanner (1953) to characterize the distribution behavior of the number of customers served in a queuing system with Poisson input and constant service time. Later this probability distribution was applied in some models for random trees and branching processes. In the latter case one of the parameters can be interpreted as the offspring mean in a Galton-Watson process with Poisson reproduction law. We propose empirical Bayes estimators for this parameter.