Banff International Research Station

for Mathematical Innovation and Discovery

Challenges and Advances in High Dimensional and High Complexity Monte Carlo Computation and Theory
Mar 18 - Mar 23, 2012

MEALS

*Breakfast (Buffet): 7:00 – 9:00 am, Sally Borden Building, Monday – Friday
*Lunch (Buffet): 11:30 am – 1:30 pm, Sally Borden Building, Monday – Friday
*Dinner (Buffet): 5:30 – 7:30 pm, Sally Borden Building, Sunday – Thursday
Coffee Breaks: As per daily schedule, in the foyer of the TransCanada Pipeline Pavilion (TCPL)
***Please remember to scan your meal card at the host/hostess station in the dining room for each meal.**

MEETING ROOMS

All lectures will be held in the new lecture theater in the TransCanada Pipelines Pavilion (TCPL). LCD projector, overhead projectors and blackboards are available for presentations.

SCHEDULE

Sunday

16:00	Check-in begins (Front Desk – Professional Development Centre - open 24 hours)
17:30-19:30	Buffet Dinner
20:00	Informal gathering in 2nd floor lounge, Corbett Hall (if desired) Beverages and small assortment of snacks are available on a cash honor system.

Monday

7:00-8:45	Breakfast
8:45-9:00	Introduction and Welcome by BIRS Station Manager, TCPL
	Morning Session Chair: Antonietta Mira
9:15-10:00	Roland Assaraf, "Correlated sampling without re-weighting: computing properties with size-independent variances"
10:00-10:30	Coffee Break, TCPL
10:30-11:30	Michele Parrinello, "Sampling complex distributions in physics, chemistry and biology"
11:30-13:00	Lunch
13:00-14:00	Guided Tour of The Banff Centre; meet in the 2nd floor lounge, Corbett Hall
14:00	Group Photo; meet in foyer of TCPL (photograph will be taken outdoors so a jacket might be required).
	Afternoon Session Chair: Yuguo Chen
14:15-15:00	David van Dyk, "Computational challenges with complex data for complex astrophysics"
15:00-15:30	Coffee Break, TCPL
15:30-16:15	Chiara Sabatti, "Bayesian models for rare genetic variants"
16:15-17:00	Dawn Woodard, "Efficiency of Markov chain methods for genomic motif discovery"
17:30-19:30	Dinner

Tuesday

7:00-9:00 Breakfast

Morning Session Chair: Radu Craiu

9:00-9:45 Eric Moulines, “Some progresses in the simulation of multimodal distributions”

9:45-10:30 Nando de Freitas, “Adaptive MCMC for high dimensional and high complexity problems”

10:30-11:00 Coffee Break, TCPL

11:00-11:45 Jeffrey Rosenthal, “Adapting Metropolis algorithms and Gibbs samplers”

11:45-13:30 Lunch

Afternoon Session Chair: Jeffrey Rosenthal

13:30-14:15 Christian Robert, “Approximate Bayesian Computation for model selection”

14:15-15:00 Francois Perron, “Bayesian estimation of copulas based on ranks and ABC”

15:00-15:30 Coffee Break, TCPL

15:30-16:15 Scott Sisson, “Approximate Bayesian Computation in high dimensions”

16:15-17:00 Zhiqiang Tan, “A sampling algorithm via tempering, importance subsampling and Markov chain moving”

17:30-19:30 Dinner

Wednesday

7:00-9:00 Breakfast

Morning Session Chair: Yuguo Chen

9:00-9:45 Gareth Roberts, “Sequential importance sampling for irreducible diffusions”

9:45-10:30 Ian Dinwoodie, “Comparing discrete dynamics over finite fields”

10:30-11:00 Coffee Break, TCPL

11:00-11:45 Jun Liu, “On two ideas in sequential Monte Carlo methods”

11:45-13:30 Lunch

Free Afternoon

17:30-19:30 Dinner

Thursday

7:00-9:00 Breakfast

Morning Session Chair: Duncan Murdoch

9:00-9:45 Helene Massam, “Bayes factors and the geometry of discrete loglinear models”

9:45-10:30 Faming Liang, “Bayesian subset modeling for high dimensional generalized linear models and its asymptotic properties”

10:30-11:00 Coffee Break, TCPL

11:00-11:45 James Hobert, “Convergence rate results for two Gibbs samplers”

11:45-13:30 Lunch

Afternoon Session Chair: Radu Craiu

13:30-14:15 Krzysztof Latuszynski, “Why does the Gibbs sampler work on hierarchical models?”

14:15-15:00 Duncan Murdoch, “Nearly perfect sampling”

15:00-15:30 Coffee Break, TCPL

15:30-16:15 Jose Blanchet, “Advances in efficient Monte Carlo for stochastic networks”

17:30-19:30 Dinner

Friday

7:00-9:00 Breakfast

9:00-10:30 Informal Discussions
10:30-11:00 Coffee Break, TCPL

11:30-13:30 Lunch

Checkout by 12 noon.

** 5-day workshop participants are welcome to use BIRS facilities (BIRS Coffee Lounge, TCPL and Reading Room) until 3 pm on Friday, although participants are still required to checkout of the guest rooms by 12 noon. **

Banff International Research Station
for Mathematical Innovation and Discovery

Challenges and Advances in High Dimensional and High Complexity Monte Carlo Computation and Theory
Mar 18 - Mar 23, 2012

ABSTRACTS

(in alphabetic order by speaker surname)

1. Roland Assaraf, Université Pierre et Marie Curie
   Title: Correlated sampling without re-weighting: computing properties with size-independent variances

Abstract: Many important quantities of chemical or physical interest are localized in space, that is, they do not depend on distant degrees of freedoms. Standard Quantum Monte Carlo estimators of such properties are local in space, but their variance scale linearly with the system size [D. Domin, R. Assaraf and W. Lester]. We show that the recently proposed generalized correlated sampling without re-weighting procedure [R. Assaraf, M. Caffarel and A. Kollias], extends the locality of the estimator to its variance. A proof is given for a large collection of non-interacting systems. We show that this property still holds for a large collection of strongly interacting systems, using simulations on non-metallic and metallic hydrogen chains.

2. Jose Blanchet, Columbia University
   Title: Advances in efficient Monte Carlo for stochastic networks

Abstract: Stochastic networks are multidimensional stochastic processes that evolve under constraints. They typically arise in queueing applications, where the constraints correspond to buffer capacities and system content; or in risk analysis, where the constraints represent contractual obligations and solvency requirements. In this talk we discuss some state-of-the art algorithms in stochastic networks, explain their complexity properties, and then flesh out some of the challenges and current work related to the computational analysis of stochastic networks. For example, we will discuss linear-time algorithms for large deviations of hitting time problems of multidimensional reflected Brownian motion, as well as algorithms for heavy-tailed models in the setting of stochastic risk networks.

3. Nando de Freitas, University of British Columbia
   Title: Adaptive MCMC for high dimensional and high complexity problems

Abstract: The talk will outline two recent advances on the application of Bayesian optimization and correlated bandits to automatically tune Markov chain Monte Carlo (MCMC) algorithms. The first part of the talk will describe an adaptive algorithm for equilibrium Monte Carlo sampling of binary-valued systems, which allows for large moves in the state space. This is achieved by constructing self-avoiding walks in the state space. As a consequence, many bits are flipped in a single MCMC step. The algorithm has several free parameters, which we show can be adapted with Bayesian optimization. The method performs remarkably well in a broad number of sampling tasks: toroidal ferromagnetic and

frustrated Ising models, 3D Ising models, restricted Boltzmann machines and chimera graphs arising in the design of quantum computers. The second part of the talk will describe the application of correlated bandits to automatically tune hybrid Monte Carlo (HMC). The talk will describe the use of cross-validation error measures for adaptation, which we believe are more pragmatic than many existing adaptation objectives. The new predictive measures take the intended statistical use of the model, whose parameters are estimated by HMC, into consideration. We apply the method to prediction and feature selection with multi-layer feedforward neural networks. The experiments with synthetic and real data show that the proposed adaptive scheme is not only automatic, but also does better tuning than human experts.

4. Ian Dinwoodie, Portland State University
   Title: Comparing discrete dynamics over finite fields

Abstract: Some biological network models can be approximated as dynamical systems with a discrete state space in a finite field. Then comparisons of hypothetical models are possible using a combination of methods of sequential importance sampling and commutative algebra. An example will be shown.

5. James Hobert, University of Florida
   Title: Convergence rate results for two Gibbs samplers

Abstract: After a brief review of Markov chain CLTs, geometric ergodicity and geometric drift conditions, I will describe new convergence rate results for two block Gibbs samplers. The first is a recently developed algorithm for exploring the intractable posterior density associated with a Bayesian quantile regression model. The second is a well-known Gibbs sampler for the Bayesian version of the general linear mixed model. (This is joint work with K. Khare and J. Roman.)

6. Krzysztof Latuszynski, University of Warwick
   Title: Why does the Gibbs sampler work on hierarchical models?

Abstract: Gibbs sampling and related MCMC have been incredibly successful in Bayesian statistics during the last 20 years. However, a theoretical understanding of exactly why the methods work so well on important classes such as hierarchical models still eludes us. This talk will shed some light on the problem by describing qualitative convergence results for Gibbs samplers on hierarchical models such as uniform, geometric and sub geometric ergodicity. (joint work with Omiros Papaspiliopoulos, Natesh S. Pillai and Gareth O. Roberts)

7. Faming Liang, Texas A&M University
   Title: Bayesian subset modeling for high dimensional generalized linear models and its asymptotic properties

Abstract: We propose a new prior setting for high dimensional generalized linear models (GLMs), which leads to a Bayesian subset regression (BSR) with the maximum a posteriori model coinciding with the minimum extended BIC model. Under mild conditions, we establish consistency of the posterior. Further, we propose a variable screening procedure based on the marginal inclusion probability and show that this procedure shares the same properties of sure screening and consistency with the existing sure independence screening (SIS) procedure. However, since the proposed procedure makes use of joint information of all predictors, it generally outperforms SIS and its improved version in real applications. For subject classification, we propose a selected Bayesian classifier based on the goodness-of-fit test for the models sampled from the posterior. The new classifier is consistent and robust to

the choice of priors. The numerical results indicate that the new classifier can generally outperform the Bayesian model averaging classifier and the classifier based on the high posterior probability models. We also make extensive comparisons of BSR with the popular penalized likelihood methods, including Lasso, elastic net, SIS and ISIS. Our numerical results indicate that BSR can generally outperform the penalized likelihood methods: The models selected by BSR tend to be sparser and, more importantly, of higher generalization ability. In addition, we find that the performance of the penalized likelihood methods tend to deteriorate as the number of predictors increases, while this is not significant for BSR.

8. Jun Liu, Harvard University
   Title: On two ideas in sequential Monte Carlo methods

Abstract: Based on the principles of importance sampling and resampling, Sequential Monte Carlo (SMC) encompasses a large set of powerful techniques dealing with complex stochastic dynamic systems. Many of these systems possess strong memory, with which future information can help to sharpen the inference about the current state. By providing theoretical justification of several existing algorithms and introducing several new ones, we study systematically how to construct efficient SMC algorithms to take advantage of future information without creating substantial computational burden. The main idea is to allow lookahead in the Monte Carlo process so that future information can be utilized in weighting and generating Monte Carlo samples, or resampling from samples of the current state. This is based on the joint work with Rong Chen and Ming Lin. Another idea regards the problem of sampling from a distribution defined on a "high resolution" space. In this situation we often have a sequence of spaces with increasingly high resolutions (or complexity) that approach the target space. We can use lower resolution spaces as stepping stones and design a sequential rejection control sampler (SRCS) to effectively sample in this setting. We show how this method can be applied to a Bayesian model to deal with the market-basket data mining problem.

9. Helene Massam, York University
   Title: Bayes factors and the geometry of discrete loglinear models

Abstract: A standard tool for model selection in a Bayesian framework is the Bayes factor which compares the marginal likelihood of the data under two given different models. We consider the class of hierarchical loglinear models for discrete data given under the form of a contingency table with multinomial sampling. We assume that the Diaconis-Ylvisaker conjugate prior is the prior distribution on the loglinear parameters and the uniform is the prior distribution on the space of models. Under these conditions, the Bayes factor between two models is a function of their prior and posterior normalizing constants under each model. These constants are functions of the hyperparameters (m, □), which can be interpreted respectively as marginal counts and the total count of a fictive contingency table, and of the data. Both constants are finite when □ is positive and m and the data are in the interior C of the support of the multinomial distribution. We will always assume that m which is a prior hyperparameter of our choice satisfies this condition and we study what happens when □→0 and the data is on the boundary of C. We will see that the behaviour of the Bayes factor is dictated by the dimension of the face of C to which the data belongs. We will also emphasize the role played by the characteristic function of C, a new object in the toolbox of exponential families.

10. Eric Moulines, Institut Telecom-Mines / Télécom ParisTech

Title: Some progresses in the simulation of multimodal distributions

Abstract: Sampling multimodal distributions remains a very challenging issue especially when the dimension of the space is large. Among the many approaches developed to address this problem, we focus in this talk on the so-called parallel tempering and its extension to interactive tempering (also named the "replica" method in statistical physics). We will discuss some preliminary results to automatically set the (many) different parameters appearing in these methods, in particular, the temperature levels and the energy rings. This provides a first attempt to make these methods really adaptive, which is badly needed in such case, because the tuning of this different parameters (which is critical) is particularly difficult. We will also report some computational results on several "difficult" simulation multimodal simulation problems as well as some theoretical convergence results (and many unsolved problems!). (Joint work with G. Fort, B. Miasojedow, P. Priouret, A. Shreck, M. Vihola)

11. Duncan Murdoch, University of Western Ontario
    Title: Nearly perfect sampling

Abstract: Propp and Wilson's (1996) coupling from the past (CFTP) algorithm caused excitement because it automatically selected the right number of samples to give draws exactly distributed according to the limiting distribution of a Markov chain. Unfortunately, it turned out to be quite difficult to apply in most real problems. In this talk I'll show that perfect sampling may be impractical, but nearly perfect sampling is not, and it may lead to insights that help in standard MCMC.

12. Michele Parrinello, Eidgenössische Technische Hochschule Zürich and Università della Svizzera Italiana, Lugano
    Title: Sampling complex distributions in physics, chemistry and biology

Abstract: In the study of physical systems like complex materials, proteins and chemical reactions one is often confronted with sampling complex distribution which depend on a large number of variable. In order to sample these distribution techniques like Metropolis Monte Carlo or Molecular Dynamics are used. However very of these distributions are characterized by several minima separated by very low probability regions. Physical examples of these standard sampling methods do no work and the system explores only the neighborhood of the initial state. To overcome this problem we have developed a method that we have called metadynamics that can overcome this problem. This is based on the identification of the relevant degrees of freedom of the system and in constructing an history dependent bias potential which allows overcoming the low probability region and sampling the full unbiased probability distribution. The choice of these relevant degrees of freedom can be done on the basis of physical consideration or by trial and error. More recently, using machine learning techniques, we have developed a method which we have named reconnaissance metadynamics which discovers automatically the relevant degrees of freedom. We shall present a number of application that demonstrate the power of the methods developed.

13. Francois Perron, University of Montreal
    Title: Bayesian estimation of copulas based on ranks and ABC

Abstract: Assume that (X; Y) is a random couple having cumulative distribution function H with continuous margins F and G respectively. Sklar's decomposition yields a unique copula C such that H(x; y) = C(F(x);G(y)). Here F, G and C are the unknown parameters, the one of interest being the copula C. A fundamental property is the fact that the copula associated to

the distribution of (g(X); h(Y)) is the same as C, as long as g and h are strictly increasing functions. It is then natural to require that the inference on C be invariant with respect to strictly increasing transformations of the margins, that is, the estimation gives the same result whether it is based on the sample f(x_i; y_i), i=1,...,n or on f(g(x_i); h(y_i)), i=1,...,n. Therefore, the inference should be based on the bivariate ranks statistic f(rank(x_i); rank(y_i)), i=1,...,n, since it is maximal invariant here. We adopt a Bayesian approach, and we show that by using ranks, the prior on the margins has no effect on the inference for C. Moreover, the link between the Bayes estimator proposed here and the one based on the complete information f(x_i; y_i), i=1,...,n is discussed. Although the rank based Bayesian inference on C is not affected by the prior on the margins, we show through an example that in practice, some particular choices can be more convenient for the computations. We also discuss the problem of finding numerical results using simulations via ABC. Finally, we apply our methodology to treat the problem of estimating an archimedian copula using a Bayesian nonparametric approach.

14. Christian Robert, Universite Paris-Dauphine
    Title: Approximate Bayesian computation for model selection

Abstract: Approximate Bayesian computation (ABC), also known as likelihood-free methods, has become a standard tool for the analysis of complex models, primarily in population genetics but also for complex financial models. The development of new ABC methodology is undergoing a rapid increase in the past years, as shown by multiple publications, conferences and even softwares. While one valid interpretation of ABC based estimation is connected with nonparametrics, the setting is quite different for model choice issues. We examined in Grelaud et al. (2009) the use of ABC for Bayesian model choice in the specific of Gaussian random fields (GRF), relying on a sufficient property to show that the approach was legitimate. Despite having previously suggested the use of ABC for model choice in a wider range of models in the DIY ABC software (Cornuet et al., 2008), we present in Robert et al. (PNAS, 2011) theoretical evidence that the general use of ABC for model choice is fraught with danger in the sense that no amount of computation, however large, can guarantee a proper approximation of the posterior probabilities of the models under comparison. In a more recent work (Marin et al., 2011), we expand on this warning to derive necessary and sufficient conditions on the choice of summary statistics for ABC model choice to be asymptotically consistent. (Joint works with J.-M. Cornuet, A. Grelaud, J.-M. Marin, N. Pillai. and J. Rousseau)

15. Gareth Roberts, University of Warwick
    Title: Sequential importance sampling for irreducible diffusions

Abstract: This talk will present recent work on a sequential importance sampler which provides online unbiased estimation for irreducible diffusions (that is ones for which the reduction to the unit diffusion coefficient case by the Lamperti transform is not possible). For this family of processes, exact simulation (ie free from discretisation error) using recently developed retrospective simulation techniques is typically not possible. Thus the work significantly extends the class of discretisation error-free Monte Carlo methods available for diffusions. The methods are most useful in the multi-dimensional setting, where many interesting families of models (particularly in finance and population modelling) exhibit the irreducibility property. This is joint work with Paul Fearnhead, Krys Latuszynski, and Giorgos Sermaidis.

16. Jeffrey Rosenthal, University of Toronto
    Title: Adapting Metropolis algorithms and Gibbs samplers

Abstract: Markov chain Monte Carlo algorithms often require tuning, which is challenging in high dimension. One solution involves letting the computer automatically "adapt" the algorithm while it runs, to tune on the fly. However, such adaptive can easily destroy ergodicity if done naively. In this talk, we consider adapting the proposal distributions of Metropolis algorithms, and adapting the coordinate selection probability weights of Gibbs samplers. We present conditions which ensure ergodicity of the adaptive algorithms.

17. Chiara Sabatti, Stanford University
    Title: Bayesian models for rare genetic variants

Abstract: Recent genetic studies, based on resequencing technology, survey the entire spectrum of DNA variation in the enrolled subjects, resulting in a very large number of variable sites, where alleles different from the reference sequence are observed in a very small number of subjects. From a statistical viewpoint, we have a modest number of observations on a very large number of variables, most of which are very sparse in the sense that their observed realizations in the dataset are mostly equal to 0. I will describe some Bayesian models that can be particularly fruitful in analyzing these data, as well as the computational challenges that they pose.

18. Scott Sisson, University of New South Wales
    Title: Approximate Bayesian Computation in high dimensions

Abstract: Approximate Bayesian computation (ABC) has become an increasingly popular technique for the fitting of complex Bayesian models when the likelihood function is unavailable or computationally intractable. In the past 10 years it has received widespread application across many different scientific disciplines. ABC methods are simple to use (they were invented by Biologists), and have gradually increased in efficiency and sophistication as researchers have begun to understand their properties.
However ABC methods suffer greatly from the curse of dimensionality -- as the number of model parameters grows, their performance can deteriorate rapidly. This is problematic, as the popularity of ABC means that it is being applied to ever more complex (and higher-dimensional) problems. In this presentation, I will introduce ABC methods and illustrate both their practical utility and their poor performance in high dimensions. I will then discuss some useful ways in which standard ABC techniques may be extended into moderate and higher dimensions, demonstrating that there is a strong future for these methods in applied research.

19. Zhiqiang Tan, Rutgers University
    Title: A sampling algorithm via tempering, importance subsampling and Markov chain moving

Abstract: We present some ongoing work to develop a sampling algorithm, drawing on a number of ideas including tempering (creating a ladder of distributions), importance subsampling (using importance sampling to go from one distribution to another), and Markov chain moving (using Markov chain kernels to move locally). As a result, the algorithm has the promise of combining strength from simulated tempering, sequential importance sampling resampling (or particle filtering), and equi-energy sampling to achieve effective Monte Carlo sampling.

20. David van Dyk, Imperial College London
    Title: Computational challenges with complex data for complex astrophysics

Abstract: In recent years, technological advances have dramatically increased the quality and quantity of data available to astronomers. Newly launched or soon-to-be launched space-based telescopes are tailored to data-collection challenges associated with specific scientific goals. These instruments provide massive new surveys resulting in new catalogs containing terabytes of data, high resolution spectrography and imaging across the electromagnetic spectrum, and incredibly detailed movies of dynamic and explosive processes in the solar atmosphere. The spectrum of new instruments is helping scientists make impressive strides in our understanding of the physical universe, but at the same time generating massive data-analytic and data-mining challenges for scientists who study the resulting data. Models designed to approximate physical processes in astronomical sources may only be available as complex computer models. In order to fully model the data generation process we must embed these computer models into highly structured multi-level statistical models. Fitting such models requires sophisticated computational methods. In this talk I will discuss a number of specific problems from astronomy, the models that we designed to solve them, and the MCMC methods required to fit them. The computational techniques include dynamic transformations, judicious choices of prior distributions for nuisance parameters, sampling incompatible conditional distributions, collapsed and partially-collapsed Gibbs samplers, and parallel tempering with specifically tailored chains.

21. Dawn Woodard, Cornell University
    Title: Efficiency of Markov chain methods for genomic motif discovery

Abstract: We analyze the efficiency of a popular Gibbs sampling method used for statistical discovery of gene regulatory binding motifs in DNA sequences. We bound its convergence rate, and show that, due to multimodality of the posterior distribution, the rate of convergence often decreases exponentially as a function of the length of the DNA sequence. This implies that the run time of the algorithm grows exponentially in the sequence length, to attain a fixed accuracy. Our findings match empirical results, in which the motif-discovery Gibbs sampler has exhibited such slow convergence that it is used for finding modes of the posterior distribution (candidate motifs) rather than for obtaining samples from that distribution. We also give more general bounds on convergence rates of Markov chain methods used in Bayesian statistics, in the parametric i.i.d. setting and as a function of the number of observations.