Proposed Course Syllabus

A. General information

Course Title: Advanced Statistical Methodology for the Computational Biosciences

Course Number: BS 850[1]

Pre-requisites: CAS MA 581/582, or equivalent (by permission of the instructor). Exposure to, and/or willingness to learn, hands-on computing at the level of BS 805/BS 822 is required.

Credits: 4

Instructor: Mayetri Gupta, Ph. D., Assistant Professor of Biostatistics (gupta@bu.edu)

B. Course goals, rationale, and teaching methods

Rationale:
With the dramatic increase in data generation in a variety of modern scientific fields such as molecular biology, genetics, and medical imaging due to rapid technological developments, the field of statistics has undergone a major change, as new and novel techniques of statistical modeling and data analysis are continually required. As more complex models are developed, classical statistical methods often fail in the face of huge dimensionalities and latent correlation structures in the data- Bayesian computing methods, especially Monte Carlo methods, have provided an invaluable tool to address many of these issues successfully, and this field has expanded rapidly over the last decade and a half.

Goals:
The goal of this course is for the student to develop a thorough understanding and set of skills in advanced statistical and computational methods used in current scientific, especially biological, applications. The objective is also to make students aware of the possibilities (and limitations) of various statistical computing methods and areas for improvement. In the first half of the course, students will be given exposure to important topics in modeling and simulation in increasingly complex scientific scenarios, with an emphasis on Bayesian modeling and computation tools, such as generation of random numbers, optimization methods, numerical integration, the EM algorithm, importance sampling, Gibbs sampler, Metropolis Hastings, auxiliary variable methods, data augmentation, reversible jump MCMC, and population-based Monte Carlo. From the midpoint of the course, emphasis will shift towards exposing students to important problems in the field of computational biology that require advanced statistical computing tools. Students will gain hands-on experience constructing, programming and implementing (in R/C/C++ or WinBUGS) computational techniques in real biological applications. Topics would include dynamic programming, hidden Markov models, multiple sequence alignment, phylogenetic tree reconstruction, gene regulatory network discovery and analysis of high-throughput array and sequencing data. Topics will be taught using a mixture of lectures, class discussions, critical readings and student presentations.

After completing this course, the student should in general be able to:

- Evaluate the appropriateness and feasibility of a particular statistical computing technique for a particular application.
- Develop and implement the appropriate statistical computing method for a given biological application.
- Explain and contrast the strengths and weaknesses of various statistical computing techniques for a particular application.
- Have a thorough grasp of the underlying principles which would be adequate to evaluate and develop novel techniques in scenarios that may arise in future applications.

[1] tentatively assigned

In particular, the student will be able to

- Develop, formulate and implement methods such as the EM algorithm, importance sampling, Gibbs sampler, Metropolis Hastings, auxiliary variable methods, data augmentation, reversible jump MCMC, and population-based Monte Carlo in the context of many important biological applications.
- Develop and implement computational tools for statistical models arising in computational biology and genomics, such as hidden Markov models, multiple sequence alignment, phylogenetic trees, and gene regulatory networks.
- Assess model fit and diagnostics from the implementation of the methods to a set of data collected from a particular scientific application.
- Summarize and present results of implementation through valid statistical inference, and compare and contrast the performance and limitations of various statistical computing methods in the application.

Course description (75 words)
This course discusses advanced statistical computing methods used in modern scientific investigation focusing on computational biology applications. Topics include random number generation, numerical optimization and integration, the EM algorithm, importance sampling, Gibbs sampling, Metropolis-Hastings and data augmentation algorithms, auxiliary variable methods, reversible jump and population-based Monte Carlo. We will discuss applications in genomics and computational biology, including dynamic programming, hidden Markov models, multiple sequence alignment, phylogenetic reconstruction, and gene regulation.

C. Required texts or other materials

Given the broad and advanced nature of the topics in this course, there is no one textbook that covers all the materials. Sections from the following books will be used as references, other articles for reading will be given out in class or downloadable from the course web site (All articles mentioned in the syllabus are either in the public domain or available to students for free through the Boston University Library web site). One copy of each of the textbooks (below) will be kept on reserve in the Biostatistics department library for student reference, and will also be available in the instructor's office.

[BDA] Gelman, Carlin, Stern, and Rubin (2003). Bayesian Data Analysis, Second Edition. (Chapman & Hall/CRC),

[JL] Liu (2008). Monte Carlo methods in Scientific Computing. (Springer)

Additional readings from books (below) and journal articles will be assembled as a course reader.

The instructor will use Blackboard to provide students with a copy of the slides (and computer output, if applicable) at least 24 hours before each class. Students will be expected to download this material and have it ready to use in class.

Other books from which certain chapters may be provided for reading are:

[DEKM] Durbin, Eddy, Krogh, and Mitchison (1999). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. (Cambridge University Press)

Required software

The statistical software which will be used in the course (including all homework assignments and projects) are R and WinBUGS, which are both freely downloadable, open source software. Instructions on downloading, installation, and usage of the software will be provided in class.

D. Course requirements and Student Evaluation

Assignments:

1. Homeworks (40% of the course grade)- to be assigned every 2-3 weeks, that will allow the students to apply what they have learned in each class. These may involve algorithm formulation and implementing techniques presented in class through writing computer code, and applying it to the analysis of a data set provided by the instructor. Students will be required to state explicitly the statistical models that are applied, to state appropriate tests of hypotheses, to describe the results and to draw conclusions from the results. Students can consult with each other in regards to the homework but the submitted version should be entirely their own work (written without help from others).

2. Course Project (40% of the course grade). Students may propose their own projects, or consult with the instructor for suggestions. The course project will ideally implement and apply some of the computational techniques learnt in class to a real biological problem, although many variations to this may be possible. Students will be asked to work in pairs or a group of 3 on the final project. Each project should accomplish a number of tasks that match the number of students in the group. The task assignment within the group will be left up to the students to decide, but students are responsible for communicating the task assignment to the instructor at the time of submitting the project proposal. Each group will submit a written report that summarizes their analysis and explicitly states each student's contribution to the analysis and write-up. Students are highly encouraged to start thinking about topics early in the semester and discuss their ideas with the instructor. The project proposal will be due at the end of Class 10, and the final project (in the form of a 8-10 page report) due the Monday following the final class (May 4, 2009). Innovativeness and depth of vision will be evaluated upon, as well as technical skills.

   The grading for the course project will be based upon the following criteria:

   - Approach and expertise (50%): Does the approach reflect upon knowledge and comprehension acquired during the course of the semester, demonstrated by the appropriateness of the methodology used to solve the specific task? Is the implementation of the analysis procedure sound? Does the student demonstrate ability in developing and using software for analysis and interpreting the output correctly?
   - Innovation(10%): Is there creativity demonstrated in solving the problem? Are there aspects of the problem that require more in-depth thinking than the straightforward application of the techniques discussed in class, and has there been some demonstrated effort to address these issues?
   - Critical evaluation (20%): Have potential pitfalls to the proposed solution to the problem, and alternative approaches been considered? Have the implemented approaches been tested for accuracy and robustness (if applicable)?
   - Inference (20%): Is enough evidence presented for the conclusions drawn, and is the evidence synthesized well? Are the ideas, goals, techniques used and conclusions reached stated clearly and concisely?

   In addition, to ensure fair grading, students' performance will be weighted by the difficulty of tasks so that a student who performs well in an easy task will not receive a better grade than a student who performs moderately well in a more difficult task.

3. In-class presentations (20% of the course grade). Students will have to summarize and discuss a chosen set of 1 or 2 journal articles from a list of articles assigned by the instructor. Evaluation will be based on comprehension of the article (30%), critical evaluation (30%), clarity of presentation (20%), and participation in discussions of articles presented by other groups (20%).

E. Class session topics

January 16 *Class 1. Review of basics of Bayesian modeling.*

Students will be introduced to concepts involved in Bayesian modeling and data analysis.

Topics will include: Setting up a probability model; using probability theory and Bayes rule; simple examples: binomial, normal, and exponential families; non-informative priors; exchangeability. Simple regression models. Inference from large samples and comparison to standard non-Bayesian methods. Many examples will be discussed to illustrate the basic principles. Basic multi-parameter and multi-level models, the multivariate normal distribution, multinomial models, Regression and hierarchical models.

Learning objectives:

After this session, students should be able to:

- Formulate a simple Bayesian probability model, and construct posterior distributions for inference.
- Enunciate differences between the frequentist and Bayesian approaches for simple problems, in terms of formulation and results.
- Lay out an analysis plan for standard single and multi-parameter Bayesian models.

January 23 *Class 2. Introduction to Bayesian computing methods.*

Students will be introduced to the concept of using simulation to summarize posterior distributions. There will be an introduction to iterative sampling methods- Metropolis-Hastings and Gibbs sampling approaches. We will discuss basic convergence diagnostics, summarizing results of a Bayesian analysis, simulation consistency and standard errors. We will also touch upon prior and posterior model checking and sensitivity analyses, to be discussed further in the next class.

Learning objectives:

After this session, students should be able to:

- Lay out an analysis plan for a simple Bayesian model using an iterative sampling approach.
- Assess the results of iterative sampling through convergence diagnostics, and model checks.

January 30 *Class 3. Statistical programming in R and WinBUGS.*

Topics: How to download and install the R programming language, and WinBUGS, and conduct basic data processing and analyses. Importing data from other formats (SAS and Excel) into R. Implementation of simple Bayesian analyses in R: programming, checking convergence, and obtaining posterior summaries. Students will be introduced to the concepts involved in designing and programming a simulation study. Tradeoffs between using R and WinBUGS.

Learning objectives:

After this session, students should be able to:

- Download and install R and WinBUGS.
- Read data sets from set of different formats into R.
- Implement and run a simple Bayesian analysis in R or WinBUGS, and summarize results.
- Design and program a simulation study to assess model sensitivity and carry out model checking and diagnostics.

February 6 *Class 4. Optimization, regularization, and quadrature.*

Topics: General algorithms for optimization and quadrature (numerical integration). Non-iterative and iterative methods. Newton's method, scoring, solutions to non-linear equations. Implementation of techniques in R. When is optimization/quadrature important in a Bayesian framework? We will discuss Bayesian inference as regularization, and weakly informative priors.

Learning objectives:

After this session, students should be able to:

- Formulate an optimization or quadrature-based solution for a particular problem.
- Program simple optimization algorithms for a particular problem, such as Newton's method and scoring, in R.
- Implement an analysis using optimization or quadrature techniques in R.
- Conceive the duality of a Bayesian analysis with a specified prior as regularization

February 13 *Class 5. Iterative sampling methods and missing data problems.*

Topics: The expectation-maximization (EM) and EM-type algorithms, and applications to various linear models. Convergence of iterative computational methods, examples. Missing data problems and data augmentation framework. Latent variable models; Student-t and hierarchical models for robust inference. Markov chain Monte Carlo methods: Metropolis-Hastings and the Gibbs sampler, Data augmentation.

Learning objectives:

After this session, students should be able to:

- Formulate a statistical problem in terms of a missing data framework and design an EM, Gibbs sampling or data augmentation (DA) algorithm for inference, and implement the algorithm in R.
- Critically evaluate the performance of an MCMC, DA or EM algorithm in terms of convergence, and provide summaries for convergence and inference.

February 20 *Class 6. Modeling and computation in genomics I. Computational biology basics*

Topics: Genomics and molecular biology basics. Methodology of sequence alignment, Dynamic programming methods and pairwise alignment algorithms. Global and local alignment. The BLAST algorithm. Bayesian methods for pairwise alignment.

Learning objectives:

After this session, students should be able to:

- Summarize the basic principles of molecular biology, including the Central Dogma.
- Design and implement simple dynamic programming methods for global and local sequence alignment.

February 27 *Class 7. Bayes factors and marginal likelihoods through analytic and Monte Carlo methods.*

Topics: Bayes factors for model choice. Approximations based on posterior modes. Normal and Laplace approximations to the likelihood and posterior distribution. Non-iterative Monte Carlo: importance sampling, rejection sampling, sequential methods. Estimation of normalizing constants when the marginal likelihood is intractable. Examples from computational biology and genetics will be discussed.

Learning objectives:

After this session, students should be able to:

- Derive Laplace approximations for basic Bayesian probability models.
- Design and implement an importance or rejection sampling-based analysis plan for a single or two-parameter statistical model.
- Summarize analytical methods for computing the Bayes factor in tractable models in the exponential family of distributions, and approximate (analytical or computational) methods for the Bayes factor for intractable likelihoods.

March 6 *Class 8. Modeling and computation in genomics II: Hidden Markov models in biology.*

Topics: Hidden Markov models for biological sequences and their estimation. Forward-backward algorithm, recursions, Viterbi and Baum-Welch algorithms. Multiple sequence alignment, gene-finding, connections to Bayesian probability models and Monte Carlo approaches.

Learning objectives:

After this session, students should be able to:

- Formulate a hidden Markov model for some simple biological problems in sequence analysis.
- Design and implement methods for optimization (Viterbi/Baum-Welch) or posterior sampling (data augmentation with recursive techniques).
- Design and implement a statistical method for multiple sequence alignment.

March 20 *Class 9. MCMC for specialized problems: variable and model selection.*

Topics: Techniques for dealing with high-dimensional data. Variable and model selection for regression and multivariate Gaussian models; Stochastic search, spike and slab modeling; Bayesian additive regression trees.

Learning objectives:

After this session, students should be able to:

- Identify an appropriate model or variable selection technique for a high-dimensional data set.
- Design and implement a variable selection technique for a regression or multivariate Gaussian model through stochastic search or spike and slab modeling.

March 27 *Class 10. MCMC for specialized problems: Mixture models and clustering methods.*

Topics: Mixture models and clustering. Reversible jump MCMC. Bayesian nonparametrics, Dirichlet process models. Clustering in regression models.

Learning objectives:

After these sessions, students should be able to:

- Design an EM or MCMC algorithm for fitting a mixture model.
- Design a clustering algorithm using mixtures or a Dirichlet process model for in a multivariate Gaussian or regression framework.

April 3 *Class 11. Modeling and computation in genomics III. Statistical models in gene regulation.*

Topics: Statistical models for gene regulatory motif discovery. Combinatoric, likelihood-based and Bayesian approaches. Analysis of genomic tiling array data. Gene regulatory network analysis. Introduction to analysis of molecular evolution and phylogeny using DNA sequences.

Learning objectives:

After this session, students should be able to:

- Formulate and implement a statistical model and analysis approach (EM or Gibbs sampling) for a motif discovery problem.
- Formulate a phylogenetic model and design a Gibbs sampling-based approach for inference.

April 10 *Class 12. Advanced Monte Carlo for complex problems: Multilevel sampling and optimization.*

Topics: How Bayesian computation can get difficult. Multimodality, posterior correlations, spikes, intractable likelihoods. Adaptive sampling techniques. Multilevel sampling and optimization methods. Simulated tempering (ST), simulated annealing (SA).

Learning objectives:

After this session, students should be able to:

- Determine whether a more advanced MCMC technique is needed for a given problem, based on posterior correlations, shape of likelihood function, intractability.
- Design and implement an adaptive MCMC technique in a single or two-parameter statistical model.

April 17 *Class 13. Advanced Monte Carlo techniques for complex problems: Population-based Monte Carlo.*

Topics: Population-based Monte Carlo methods: parallel tempering (PT), and evolutionary Monte Carlo (EMC). Auxiliary variables and slice sampling, sequential Monte Carlo. Applications in genetics and genomics, Maximum likelihood and Bayesian approaches in phylogenetic analysis.

Learning objectives:

After this session, students should be able to:

- Design and implement a PT or EMC method for standard statistical models (e.g. regression-type models).
- Design and formulate a maximum likelihood or Bayesian approach for a phylogenetic reconstruction problem.

April 24 *Class 14. Modeling and computation in genomics IV: Using advanced MCMC for biological problems.*

Topics: Improved approaches in analysis of gene regulation using population-based Monte Carlo and other advanced sampling methods. Protein structure prediction. Graphical models and gene regulatory networks.

Learning objectives:

After this session, students should be able to:

- Discern situations where a more advanced Monte Carlo technique may be useful in improving statistical inference.
- Design and layout approaches for population-based Monte Carlo and adaptive methods in gene regulation, phylogenetic analysis, and protein structure prediction.
- Determine a graphical model structure for a given problem and construct an MCMC method for posterior inference.

May 1 *Session 15. Class presentations.*

Topics: In this class, students will make a short presentation (10-15 minutes) based on their chosen set of journal articles. Depending on the number of registered students, students may be asked to present in groups of two, or individually. The articles may be related to their class project, but this is not a necessity. Class participation in critical discussion of papers will also be noted for student evaluations.

Learning objectives:

After this session, students should be able to:

- Succinctly summarize the main findings of the article, and note its major advances and limitations.
- Critically assess the content of the article, and if applicable, propose future directions that may be of interest to pursue.

Readings:

- Based upon student selections from list given out by instructor.

Final project write-up due by 5 pm on May 4, 2009.

F. Statement regarding Academic Honesty

Students are responsible for knowing, and abiding by, the provisions of the GRS Academic Conduct Code, which is posted at `http://www.bu.edu/grs/academics/resources/adp.html`. Charges of academic misconduct will be brought to the attention of the Associate Dean for Education, who will review all such cases and decide upon the appropriate action. A student who is found guilty of academic misconduct may be subject to disciplinary action, up to and including dismissal from the School.