Course Title: Advanced Statistical Computing

Instructor: Robert Gray  
Office: DFCI, Mayer 221  
Telephone: 632-2446  
Office hours: By appointment  
Teaching Assistant: Elizabeth Brown  
Course Web Page: http://biowww.dfci.harvard.edu/~gray/248

Lectures: Monday & Wednesday, 8:30–10:20, Room FXB G-12

Prerequisites: BIO 235; Experience with Fortran or C programming

Audience: Intended primarily for doctoral students in Biostatistics. Familiarity with advanced statistical models, principles of statistical inference and some matrix theory is assumed, as are basic programming skills.

Grading: Several homework assignments (70%) and a take-home final (30%).

Textbook: none—Extensive notes will be used.

Some useful general references:


**DESCRIPTION**

This course will cover selected computational techniques useful in advanced statistical applications and statistical research. Topics to be covered are listed below. The goal of the course is not in depth study of basic algorithms and principles of numerical analysis, but to develop skills and knowledge useful in modern statistical methods.

**SCHEDULE (highly tentative)**

Lectures 1–2. Computer arithmetic, machine constants, computing sample moments.


Lectures 14–15. EM and related algorithms.


Lectures 20–22. Basic simulation methods. Generating uniform pseudo-random numbers, transformation methods, accept-reject methods, importance sampling, control variates, antithetic sampling.

Lecture 23. More on importance sampling. Numerical integration via importance sampling in Bayesian applications.


Lectures 26–29. Bootstrap methods. Resampling methods for variance estimation, bias reduction, hypothesis testing, and confidence intervals; iterated bootstrap; use of control variates, importance sampling and antithetic sampling to improve efficiency.

SOFTWARE

Computing for this course can be done in Splus. A library of functions written for the course will be available from the course web page and from the directory `/gray/bio248` on `hsph`.

Splus is not a very good environment for working with large matrices. It has a special `Matrix` library that can be used, but even these functions are not efficient (although they call efficient FORTRAN routines, they have so much overhead in the Splus code they still do not execute efficiently). MATLAB is a better alternative for these applications, but will not be used in the course examples or lectures, although it could be used for some of the exercises.

R ([http://lib.stat.cmu.edu/R/CRAN/](http://lib.stat.cmu.edu/R/CRAN/)) can also be used for much of the computation in the course, but there are some significant differences from Splus, such as that the `Matrix` library is not available in R (although there is a Matrix package that has some similar functionality), and some of the basic matrix functions have slightly different options and output; the `nlminb()` and `nlmin()` functions are not available in R, although there is a related function `nlm();` the `integrate()` function is not available in R, (there is an ‘integrate’ package available, but the functions are somewhat different than `integrate()` in Splus); there is a difference in how overflow in integer arithmetic is handled; and there are differences in uniform random number generators.

There are several references in the notes to LAPACK and the Basic Linear Algebra Subroutines (BLAS). These are collections of FORTRAN routines for performing numerical linear algebra. Compiled libraries are available on `hsph` in `/usr4/biostatistics/lib/lapack/libblas.a` and `/usr4/biostatistics/lib/lapack/liblapack.a` for FORTRAN, and `/usr4/biostatistics/lib/clapack/libblas.a` and `/usr4/biostatistics/lib/clapack/liblapack.a` for C versions (some care is needed with the C version, as they were created by an `f2c` translation of the FORTRAN code. The source code for these libraries can be obtained from the Netlib archive ([http://www.netlib.org](http://www.netlib.org)).