BIO 248cd COURSE SYLLABUS

Course Title: Advanced Statistical Computing

Instructor: Robert Gray gray@hsph
Office: DFCI, Mayer 221
Telephone: 632-2446
Office hours: By appointment
Teaching Assistant: Mahlet Tadesse mtadesse@hsph
Course Web Page: http://biowww.dfci.harvard.edu/~gray/248-02

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.

Textbook: none—lecture notes available at

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

The primary computing environment for this course is Splus 6.0 (command Splus6 on hsph). A library of functions written for the course is available on hsph (and on the course web page). The library can be attached in Splus6 on hsph with the command

\[
\text{library(Slib6,lib.loc='/usr1/users/biostat/gray/bio248/','first=T)}
\]

Splus 6.0 has some new features not available in earlier versions, but it is also less efficient at some types of calculations. An Splus 3.4 (command Splus on hsph) version of the course library is also available. It can be attached in Splus with the command

\[
\text{library(Slib3,lib.loc='/usr1/users/biostat/gray/bio248/','first=T)}
\]

The library functions also work in R (http://lib.stat.cmu.edu/R/CRAN/), and an R version of the library can be attached in R with the command

\[
\text{library(bio248,lib.loc='/usr1/users/biostat/gray/bio248/')}\]

Although the differences between R and Splus seem to be decreasing with new releases of R, there are still many small differences and a few major ones, and the code used for some examples would require modification for R. One major difference is that the Splus functions nlminb() and nlmin(), used extensively in some examples, are not available in R, although there is a related function nlm(). There are also differences in the uniform random number generators in the two packages.

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).