Advanced statistical computing

Teachers: Carl Nettelblad, Behrang Mahjani

Salman Toor, Silvelyn Zwanzig

The guest lecturer for statistical analysis of large data sets will be announced later.

Many recent statistical methods are based on computationally intensive computer simulations. Also, new areas of statistical computing are emerging based on large amounts of data in several disciplines. A rapidly growing field is statistical analysis of data from life sciences. Such analysis can be made in fixed functionality tools or pipelines, but the statistical computing environment R has become very popular due to its greater flexibility.

Correctly used, R is a powerful resource where application-specific libraries can be combined with the use of current heterogeneous and distributed computing infrastructures. This course puts students existing practical experience on a more solid theoretical ground, allowing them to benefit from the knowledge of a world-class expert in the field in the final block of the course.

Goal of the course: This a new course covering some of the most recent applications of scientific computing to statistics, for an audience with some familiarity with R or other computing tools such as Python or Matlab. After finishing the course, the student can implement computationally intensive algorithms for statistical analysis in R, and optimize such software for use on modern computer resources. They also learn about the statistical challenges of dealing with large data sets, and how to use R for data-intensive computing. The course focuses on applications from life science.

Target audience: PhD students in Bioinformatics (and related fields), Mathematical Statistics, and Scientific Computing.

Prerequisite: Basic knowledge of mathematical statistics, linear algebra, and numerical methods. Some experience of programming in R, Python, or Matlab.

Credits: 5 or 7.5
This course consists of three blocks, at 2.5 credits each. The first block is not strictly necessary for students with significant knowledge of programming in R.

Number of scheduled occasions: 16 (11 lectures + 5 labs)

Exam Information: To pass each block, the student should finish the project of that block. The content of each project is related to the labs in that block.

Syllabus:

Block 1: Advanced programming in R (2.5 credits)
This block provides the students a deep knowledge of programming in R, which is crucial for working with large data sets. It also teaches the students how to write maintainable code giving reproducible scientific results.

Teachers: Carl Nettelblad, Behrang Mahjani, Silvelyn Zwanzig
Total number of lectures in this block: 4 lectures + 2 labs

1. Version control, GIT, Libraries, CRAN

2. Language Foundations (Data structures, Subsetting, Vocabulary, Style, Functions, OO field guide, Environments, Exceptions and debugging)

3. Functional programming (Functionals, Function operators, Metaprogramming)

4. Non-standard evaluation (Expressions, Domain specific languages)

Block 2: High performance programing in R (2.5 credits)
The goal of this block is to teach the students how to write and analyze a high performance code in R, which is essential in handling computationally intensive algorithms.

Teachers: Carl Nettelblad, Salman Toor, Behrang Mahjani (lab)
Total number of lectures in this block: 4 lectures + 2 labs

Part 1: Performant code in R (Performance, Profiling, Memory, Rcpp, R's C interface)
(Carl Nettelblad) (1 lecture + 1 lab)

Part 2: Parallelization in R

1. Parallel Computing (Explicit parallelism, implicit parallelism, GPUs)
(Carl Nettelblad) (1 lecture)

2. Big data on Cloud Computing (Hadoop, Spark) *(Salman Toor) (2 Lectures+ 1 lab)*

Block 3: Statistical and numerical methods for analysis of large data sets, with focus on bioinformatics applications (2.5 Credits)
This block showcases how to move your R usage from individual machines and modest-size datasets to the extremely large data sets that are becoming increasingly common. We cover both aspects of code and software architecture, and the statistical treatment, including some common mistakes.

Teachers: Guest lecturer
Total number of lectures in this block: 3 lectures + 1 labs

1. **Numerical precision in extremely large data sets**
2. **Packages for handling distributed and very large data in R**
3. **Statistical challenges when analyzing billions of data points**