SSS: Software for High-dimensional Bayesian Regression Model Search

by Chris Hans, Quanli Wang, Adrian Dobra and Mike West
hans@stat.ohio-state.edu
quanli@stat.duke.edu
adobra@stat.washington.edu
mike@stat.duke.edu

Overview

SSS is a suite of software suite implementing "shotgun stochastic search" for "large p" regression variable uncertainty and selection. The SSS theory and methodology for regression models is described and exemplified in Hans, Dobra and West (2007) [1]. The general framework is that of regression with uncertainty about which predictors are in the model; model uncertainty is represented in terms of a prior variable inclusion probability to penalise model dimension. SSS explores the space of potentially very many models defined by subsets of predictor variables, guided by the (unnormalised) posterior model probabilities, and ranks and summarises sets of "top models" for assessment and prediction. The parallel implementation also provides some basic support for leave-one-out cross-validation analysis.

The current program includes both serial and parallel versions, implementing SSS for linear, binary (logistic) and survival (Weibull) regression model frameworks. Examples and case studies in [1], [2] and [3] illustrate the approach. SSS is freely available and written in C++. Executables, examples and detailed descriptions of the programs can be downloaded from:

`http://www.stat.osu.edu/~hans/sss/`
or
`http://xpress.isds.duke.edu:8080/sss/`

Serial Versions

Several versions of the serial-computing implementation of SSS are provided and can be run directly in Windows or Unix by taking text file inputs and producing summary text file outputs. Additionally, a Java graphical user interface (GUI) allows SSS to be run smoothly as a standalone program in Matlab version 7 and higher (see Figure 1). Provided in the serial suite are:

- a Windows executable;
- a Java GUI program for easy Matlab interface;
- 32- and 64-bit Linux executables;
- example input files and data;
- Matlab script files providing examples for running SSS using the GUI; and
- Matlab and R script files providing examples for summarizing SSS output.

Parallel Version

The parallel-computing version of SSS is also provided, along with source code, for use on a Unix cluster. The parallel program is similar to the serial version but takes advantage of parallelizable aspects of SSS, leading to a more efficient implementation for very high-dimensional problems. The parallel program uses the MPI (Message-Passing Interface) application programming interface to facilitate communication between nodes in the cluster. No other special libraries are required. Included in the download are:

- C++ source code and compiling instructions;
- Example input files and data; and
- Matlab and R script files providing examples for summarizing SSS output.

References

[1] Hans, C., Dobra, A. and West, M. (2007). "Shotgun stochastic search for "large p" regression." *Journal of the American Statistical Association*, to appear.

[2] Hans, C. and West, M. (2006). "High-dimensional Regression in Cancer Genomics." *ISBA Bulletin*, **13**(2), 2–3.

[3] Rich, J., Hans, C., Jones, B., Iversen, E., McClendon, R., Rasheed, B., Dobra, A., Dressman, H., Bigner, D., Nevins, J. and West, M. (2005). "Gene expression profiling and genetic markers in glioblastoma survival." *Cancer Research*, **65**, 4051–4058.

Figure 1: SSS Matlab GUI