ABCreg documentation

Kevin Thornton

April 14, 2008

Introduction

This program implements the linear regression approach to “Approximate Bayesian Computation”, or ABC that is described in [1], with the tangent-tranformation of the data proposed by [2]. The code is written in C++ and is intended to automate the procedure as much as possible. It allows you to analyze a large number of data sets with one command.

Terms and Conditions

Citation

Please cite the following paper if you use `reg` in a publication:

Thornton, K. R. (2009) Automating approximate Bayesian computation by local linear regression. BMC Genetics 10: 35.

Requirements

In order to compile this code, you need the following on your Unix (Linux, OS X, etc.) system:

- A C/C++ compiler
- The GNU Scientific Library (`http://www.gnu.org/software/gsl`).

Compilation

To compile:

1. *tar xzf ABCreg.tar.gz*
2. *cd ABCreg*
3. *make*

Debug mode

To compile in debug mode, type “*make debug*” on your command line. Debug mode will run a bit more slowly, but it will also exit if anything “odd” happens during the calculations. Usually, the purpose of a debug mode is to help you out while modifying the code, and the normal compilation mode will be preferred for actual research.

Usage

To use the program, you need to generate two files. The first is the *data file*, which contains the list of summary statistics for the data set of interest. The summary statistics are to be present in one line per data set. The program will analyze multiple data sets if there are multiple lines present in the *data file*.

The second file required is called the *prior file*. It contains a list of parameters which have been randomly-sampled from the prior distributions you have chosen, and the corresponding summary statistics. ***Note that the summary statistics must appear in the same order in both the prior file and the data file!***.

Command-line options

`ABCreg` accepts the following command-line options. Values in brackets are to be supplied by the user:

- -p [*prior file*]
- -d [*data file*]
- -S [number of summary statistics]
- -P [number of parameters]
- -b [base name for output file]
- -t [tolerance]
- -T or -L
- -m [maxlines]

All of the options except -m are required. (You can used the program without -T or -L, but you'll likely get poor results.) The -m option puts a maximum on the number of lines to be read in from *prior file*, which can be useful if you find yourself with limited RAM available.

The -T option implements the tangent transformation from [2], and -L uses the standard natural-log transformation in the original [1] paper. You can use one of these or the other, but not both.

The -b option specifies the base name of the output file. For example, if you issue the command (to obtain posterior distributions for all data sets described in "datafile" for a model with 2 parameters, 4 summary statistics, and where the closest 0.001 of the prior, in terms of Euclidian distance, are used to generate posterior distributions):

```
./reg -p priorfile -d datafile -P 2 -S 4 -b output -t 0.001 -T
```

When the program is done, you will see output files created with the names

- output.0.tangent.post
- output.1.tangent.post
- output.2.tangent.post
- ...
- output.(n-1).tangent.post

These output files correspond to the posterior distributions obtained for each of the n lines present in *data file*.

Example

A complete example is provided in the file `exampleABC.sh`, which is provided with the source code. It is a simple example of estimating the scaled mutation rate parameter θ from the number of segregating sites. The example requires R (`http://www.r-project.org`) (with the `locfit` package installed) and Dick Hudson's `ms` (`http://home.uchicago.edu/~rhudson1`) to work.

If `R` and `ms` are installed in the user's `PATH`, executing "sh exampleABC.sh" in the ABCreg source directory (after compiling, of course) will run the entire analysis, which is implemented as a script in the `bash` language. See the comments in the script for documentation of each step.

References

[1] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

[2] G. Hamilton, M. Stoneking, and L. Excoffier. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *PNAS*, 102:746–7480, 2005.