Documentation for Software Package SBAMS

Xiaoquan Wen

July 23, 2013

1 Introduction

Software package SBAMS (**S**tructured **Ba**yesian **M**odel **S**election) implements the Bayesian statistical methods discussed in [1], and provides the functionality to perform structured Bayesian model selections in the Multivariate Linear Regression (MVLR) models. The relevant technical details are discussed in the paper and its supplementary materials. The purpose of this documentation is to explain the software implementation and corresponding input/output formats.

2 Approximate Bayes Factor Computation

2.1 R implementation

The file mvlr.R implements the approximate Bayes factor calculation for the MVLR model. The code is self-explanatory and it is convenient to use model comparisons in a set of candidate MVLR models.

2.2 C++ implementation

We also implement the C++ class MVLR, which can be conveniently re-used in the C++ implementations of other software packages. The public and private interfaces of the MVLR class are documented in the header file MVLR.h. The compilation of the object file requires GNU Scientific Library (GSL).

3 MCMC Algorithm for Bayesian Model Selection

We also implement the MCMC algorithm described in [1] to perform Bayesian model selection for MVLR model using the C++ programming language. The program, *sbams_mvlr*, can be used for some rather general purposes. Because the usual multiple linear regression model can be viewed a special case of MVLR, *sbams_mvlr* applies in the setting where a multiple linear regression suffices.

3.1 Installation

We provide pre-compiled binary executables for some UNIX based operating systems. Alternatively, one can directly download the source code and compile the desired binary executable.

To do so, a GNU C++ compiler,e.g. GNU g++, and GNU Scientific Library (GSL) are required. To download and install GSL on your system, please refer to http://www.gnu.org/gsl/ for details.

3.2 Input Files

There are two required input files containing data and prior information on effect sizes, respectively. There is also an optional file specifying user-defined priors on $\xi(\boldsymbol{\beta})$.

3.2.1 Data File Format

The data file contains a data matrix for response and covariate variables used in the MVLR. Each row of the data matrix has the following format,

```
header  variable_id  value_1  value_2 ... value_n
```

In particular, the "header" column takes values either "response" or "covariate", indicating the role of a particular variable in the MVLR; the "variable_id" column contains a string identifier, and the following columns contain the numerical values of the variable for all the subjects. The following data section is excerpted from a simulated data set:

```
response F -0.059   0.168 ...
response L  1.072  -2.986 ...
response T  1.955  -5.209 ...
covariate rs1  0.76 -0.24 ...
covariate rs2  0.47 -0.53 ...
```

```
covariate rs3 -0.17  0.83 ...
          ...
```

3.2.2 Prior Grid File Format

The prior information file, also known as "grid file", contains a set of grid values for the prior specification used in [1] (see also [3, 2]). In particular, this prior model assumes two parameters, (ϕ, ω), for the non-zero regression coefficients of a single covariate across multiple subgroups. The average effect of the covariate is defined by

$$\bar{\beta} \sim \mathrm{N}(0, \omega^2)$$

and conditional on the effect size is non-zero in a particular subgroup s, it assumes

$$\beta_s \sim \mathrm{N}(0, \phi^2).$$

Intuitively, ω defines the average prior effect size for the covariate and ϕ defines the level of prior heterogeneity across subgroups. Equivalently, it can be shown that quantity $\frac{\omega^2}{\omega^2+\phi^2}$ is the prior correlation coefficient between non-zero effects. Instead of using a single (ϕ, ω) pair for all covariates, we encourage users to specify a set of grid values and average the results over these different prior models.

The grid file contains a two-column data matrix: the first column always represents the heterogeneity parameter (ϕ) and the second column is used to specify the average effect size parameter (ω). Each row of the grid data matrix provides a unique prior model and different rows can be used to describe different prior heterogeneity levels and/or prior average effect sizes. The following sample of the grid file is used by [1] to perform multiple-tissue eQTL analysis. The grid uses four levels of average prior effects values, which provides a comprehensive coverage of many possible scenarios. The prior correlation for all prior models is set to 0.94, reflecting the fact that effect sizes of most tissue-consistent eQTLs show low degree of heterogeneity.

```
0.05 0.20
0.10 0.40
0.20 0.80
0.40,1.60
```

3.2.3 Optional Prior Configuration File Format

The prior configuration probability for each covariate in $\xi(\boldsymbol{\beta})$ can be specified by an optional parameter file. If this file is unspecified, the default prior will be applied in the analysis (please refer to Program Parameters Section for details).

Consider a MVLR model with p candidate covariates and s subgroups with sample size n, i.e., the response matrix Y is with the dimension $n \times s$.

The configuration prior is specified in a hierarchical way: first the quantity "pi1" defines the prior probability $\Pr(\xi(\beta_i) \neq 0)$ for $i = 1, ..., p$, i.e., the probability of a covariate having at least one non-zero coefficient in all the subgroups. By default, this value is set to $1/p$. Secondly, the conditional prior probabilities of $2^s - 1$ possible non-null configurations for each covariate, $\Pr(\xi(\beta_i) = j \mid \xi(\beta_i) \neq 0) = \eta_j$, where j is an integer in the range of $[1, 2^s - 1]$ and its binary expansion indicates the configuration. For example, for three subgroups, $j = 6$ indicates a configuration (110), i.e., the configuration corresponds to that regression coefficients are non-zero in subgroup 2 and 3. By default, all non-zero configurations are assigned equal probability, i.e, $\frac{1}{2^s-1}$.

These parameters are specified by two separate lines in the optional prior configuration file. The following is an example for $s = 2$:

```
pi1  0.001
config 0.2 0.3 0.50
```

This indicates the prior probability $\Pr(\xi(\beta_i) \neq 0) = 0.001$ and the conditional probability for configuration 1, 2 and 3 (corresponding to $(01), (10)$ and (11), respectively) are $0.2, 0.3$ and 0.5.

It should be noted that not both of the lines are required in the file, if an entry is missing from the file, the default value will be used in the program.

3.3 Running Program

Once the binary executable is in place and required data and prior grid files are prepared, the program can be initiated by the following command from a terminal:

```
./sbams_mvlr  -d data_file -g grid_file [options] [-o output_file]
```

The data and grid files, described in the previous section, are required command line arguments. Options and output file are optional, if not specified, the program assumes default parameter settings and output to the screen (stdout).

3.4 Program Parameters and Options

This section lists all the options that can be specified on the command line.

3.4.1 Input/Output Options

- *-d data_file*: required, specification of the data file.
- *-g grid_file*: required, specification of the prior grid file.
- *-p config_file*: optional, specification of the prior configuration file, see section 3.2.3 for details.
- *-o output_file*: optional, specification of the output file. By default, the program output goes to screen.

3.4.2 MCMC Control Options

- *-b burnin_steps*: optional, specification of burn-in steps. By default, the program runs 25,000 burnin-steps.
- *-r rep_steps*: optional, specification of total MCMC samples drawn after burn-in. This is set to 50,000.

3.4.3 Miscellaneous Options

- *-abf shrinkage_parameter*: optional. This option is used to determine the shrinkage parameter (ranging from $[0, 1]$) in computing the estimate of error covariance matrix (α values in [1]). The default value is set to 0.5.
- it -ens expected_signals: optional. Prior expectation of non-zero covariates. This option provides an alternative way to set "pi1" parameter (described in section 3.2.3),i.e, $\Pr(\xi(\boldsymbol{\beta}_i) \neq 0)$, which is computed as the proportion of the prior expected signals over all candidate covariates. By default, the expected number of signals is set to 1.

3.5 Output from Program

During the MCMC runs, the program outputs its selection paths to the screen through stderr. For example,

```
10    -1.424    [380:(1)]
11    -1.424    [380:(1)]
12    -1.424    [380:(1)]
13    -2.535    [380:(3)]
14    -2.535    [380:(3)]
```

```
15    -2.535   [380:(3)]
16    -2.535   [380:(3)]
17    -2.535   [380:(3)]
```

The first column indicates the number of iterations, the second column displays a quantity indicating the likelihood of current selection. The third column shows the current selected model (380 is the order of the covariate selected and the number in the parenthesis indicating the selected configuration).

In the final output, the program displays a list of models assessed during the MCMC runs along with the marginal inclusion probabilities of each covariate.

References

[1] Wen, X. "Bayesian Model Selection in Complex Linear Systems, as Illustrated in Genetic Association Studies", submit to Biometrics.

[2] Flutre, T., Wen, X, Pritchard, J., Stephens, M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLoS Genet. 2013 May;9(5):e1003486.

[3] Wen, X. and Stephens, M. "Bayesian Methods for Genetic Association Analysis with Heterogeneous Subgroups: from Meta-Analyses to Gene-Environment Interactions" *arXiv pre-print: 1111.1210*