Documentation for Software Package SBAMS

Xiaoquan Wen

June 21, 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 Input Files

There are two required input files containing data and prior information respectively.

3.1.1 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 ...
         ...
```

3.1.2 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 Running Program

3.3 Program Parameters

3.4 Input/Output Options

3.5 Program Options

3.6 Output from Program

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*