BayesR Program Documentation

Gerhard Moser
November 21, 2014

Introduction

The BayesR software implements a Bayesian mixture model for the analysis of complex traits using Markov chain Monte Carlo (MCMC). It simultaneously identifies associated SNPs, estimates the genetic variance explained by SNPs, describes the genetic architecture of the trait and predicts phenotype from SNP genotypes.

Download

The software is distributed under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version. You can download the software from http://www.cnsgenomics.com/software/ or https://github.com/syntheke/bayesR.

Compiling and running the programs

The software consist of three FORTRAN95 source code files:

BayesR: main program

RandomDistributions.f90: module containing various random number generators

baymods.f90: support module for BayesR containing common variables and routines

The program can be compiled with a FORTRAN95 compiler on a unix operating system using the following command:

```
gfortran –o bayesR RandomDistributions.f90 baymods.f90 bayesR.f90
```

The GNU compiler was assumed above for convenience. Using optimization flags can produce faster code. We found that using the Intel Fortran compiler generates code that runs at least twice as fast compared to gfortran.

Assuming the executable file is in the path of the user, the program can then be run using:

```
bayesR –bfile [prefix] –out [prefix]
```

where "`–bfile [prefix]`" specifies the PLINK binary ped files prefix and "`-out [prefix]`" specifies the output file prefix. This runs BayesR with its default options, but the user can change most parameters.

Options

Use the "`–help`" option to display a list of options with a short description and their default values (if any).

```
./a.out -h
argument type        description                                 default
-bfile   [prefix]    prefix PLINK binary files
-out     [prefix]    prefix for output
-n       [num]       phenotype column                            1
-vara    [num]       SNP variance prior                          0.01
-vare    [num]       error variance prior                        0.01
-dfvara  [num]       degrees of freedom Va                       -2.0
-dfvare  [num]       degrees of freedom Ve                       -2.0
-delta   [num]       prior for Dirichlet                         1
-msize   [num]       number of SNPs in reduced update            0
-mrep    [num]       number of full cycles in reduced update     5000
-numit   [num]       length of MCMC chain                        50000
-burnin  [num]       burnin steps                                20000
-thin    [num]       thinning rate                               10
-ndist   [num]       number of mixture distributions             4
-gpin    [num]       effect sizes of mixtures (% x Va)           0.0,0.0001,0.001,0.01
-seed    [num]       initial value for random number             0
-predict [flag]      perform prediction                          f
-snpout  [flag]      output detailed SNP info                    f
-permute [flag]      permute order of SNP                        f
-model   [filename]  model summary file (for prediction)
-freq    [filename]  SNP frequency file (for prediction)
-param   [filename]  SNP effect file (for prediction)
```

Data files

The program requires PLINK binary ped file format. It requires '*.bim and *.fam files to determine the number of SNPs and the number of individuals and a 'bed' file for the genotype information.

Genotype data

The program requires genotypes in PLINK binary format in default-SNP major mode. Since BayesR includes all genotypes in the model, samples missing a genotype call cannot simply be omitted. Missing genotypes are replaced by the mean genotype value of a given marker.

Phenotype data

The program reads "column 6" as the phenotype column from the PLINK *.fam file. A different phenotype column can be specified by using the "`–n [num]` "option, where "`–n 1`" uses the original 6th column (default), "`–n 2`" uses column 7 and so forth. Missing phenotypes (or phenotypes to be predicted) must be coded as 'NA'.

A priori information for variance components

Prior inverted-chi squared distribution can be specified for σ_g^2 and σ_e^2. Scale and degrees of freedom (df) for the variance components are required. "Flat" (improper) distributions can be specified by setting df to -2. It is also possible to specify the heritability of the trait by setting dfvara to -3.0 (i.e. `-dfvara -3.0`). In this case the scale parameter is treated as the heritability and the SNP-based variance is set (fixed) to σ_g^2=heritability*σ_p^2.

Dirichlet prior

The default is to use a uniform and almost uninformative prior for the mixture distribution with a pseudo-observation of 1 (SNP) for each class. Different priors can be specified using the "`–delta [num]`" option. For example, "`-delta 3,2,1`" specifies a prior with 3, 2 and 1 pseudo-observations for classes 1 to 3 of a 3-component mixture model, "`-delta 2`" sets the prior to 2 for all mixture components.

Mixture model

The BayesR model assumes that the true SNP effect is derived from a series of normal distributions. The default model uses 4 mixture distributions with SNP variances of 0, 0.0001, 0.001 and 0.01, so that the variance (S) of the *j*th SNP has 4 possible values: S1=0, S2=0.0001*Vg, S3=0.001*Vg, S4=0.01*Vg. Different mixture models can be specified using the "`–ndist [num]`" and "`–gpin [num]`" options. For example, "`–ndist 3 –gpin 0.0,0.001,0.05`" fits a 3 component mixture with SNP variances S1=0, S2=0.001*Vg, S3=0.05*Vg.

MCMC sampling

The default is to use a chain length of 50,000 samples ("`–numit`") with the first 20,000 samples ("`–burnin`") being discarded, and using every 10th sample ("`–thin`") for posterior inference. To improve mixing, one can use the option "`–permute`" to update SNP effects in random order.

Prediction

The basic usage for prediction analysis is:

```
bayesR –bfile [prefix] –out [prefix] –predict –model [filename]
–freq [filename] –param [filename]
```

where the flag "`–predict`" must be specified for prediction analysis, "`–bfile [prefix]`" specifies the PLINK binary ped files prefix, "`-out [prefix]`" specifies the output file prefix, "`-model [filename]`" specifies the file containing the estimated mean of the BayesR model (i.e prefix.model), "`-param [filename]`" specifies the file containing the estimated SNP effects (i.e. prefix.param) and "`-freq [filename]`" specifies the file containing the allele frequencies (i.e. prefix.frq). To be predicted phenotypes must be encoded 'NA'. For example specifying "`-n 2`" will predict phenotypes of individuals with 'NA' in column 7 of the PLINK *.fam file. The number and order of SNPs between training and validation data set must match.

Reduced update

By setting "`–msize 0`" (default), all SNPs are updated within a single iteration of the MCMC chain. To reduce computation time for larger data sets, the model size can be set to a value > 0. For example, when specifying "`–msize 500 -mrep 1000`" all SNP effects are only updated for the first 1,000 cycles. Thereafter, updating effect size continues until 500 SNPs with non-zero effects have been sampled within a cycle. The order of SNPs is randomly permuted in each MCMC cycle. The performance of the reduced MCMC scheme is not well tested.

Detailed SNP information

By specifying "`–snpout`" additional SNP information (potentially very large, see Output files) for each MCMC iteration is written to a file named 'prefix.snp'. The default option is to not output detailed SNP information.

Random seed

This is an integer number for seeding the random number generator. The default "`–seed 0`" seeds with the system clock.

Output Files

Log file

The file name prefix is as specified by "`–out [prefix]`". The suffix '.log' is appended to give the file name. This is a descriptive file and provides a summary of the run parameters used and the number of records processed.

Predicted phenotypes

This file outputs the predicted genomic values (GVs). The output prefix is used to give the file name 'prefix.gv'.

```
Example
 0.8452352E-01
  0.1995048E+00
 -0.1956367E+00
NA
 -0.1150557E+00
NA
…
```

Allele frequency

Contains allele frequency of the '2' allele. The suffix '.frq' is appended to the prefix. This file is required for scaling and centering genotypes for prediction analysis. The SNP order has to be the same as the genotype input file.

SNP effects

The suffix 'param' is appended to the output prefix. The SNP order is the same as the genotype input file. This file contains mean posterior estimates for each individual SNP:

PIP1..k: Posterior inclusion probabilities of the SNP in mixture classes 1 to k

beta: SNP effect

```
Example
  PIP1        PIP2        PIP3        PIP4        beta
  0.9900000E+00   0.1000000E-01   0.0000000E+00   0.0000000E+00   0.5077715E-04
  0.9860000E+00   0.1400000E-01   0.0000000E+00   0.0000000E+00  -0.3132155E-04
  0.9780000E+00   0.2200000E-01   0.0000000E+00   0.0000000E+00   0.1003181E-05
  0.9820000E+00   0.1400000E-01   0.4000000E-02   0.0000000E+00   0.9343397E-04
  0.9760000E+00   0.2400000E-01   0.0000000E+00   0.0000000E+00   0.8146890E-05
  0.9780000E+00   0.2000000E-01   0.2000000E-02   0.0000000E+00   0.2624734E-04
```

Model summary

The suffix ‘model’ is appended to the output prefix. This file contains means of the posterior samples of model parameters:

Mean:	intercept
Nsnp:	number of SNPs in model
Va:	genetic variance explained by SNPs
Ve:	residual variance
Nk1,…,Nkk:	number of SNPs in mixture components 1 to *k*
Pk1,…,Pkk:	proportion of SNPs in mixture component 1 to *k*
Vk1,…,Vkk:	sum of squares of SNP effects in mixture component 1 to *k*

```
Example
Mean        0.4849967E-02
Nsnp        0.5185670E+04
Va          0.2000147E+00
Ve          0.7983887E+00
Nk1         0.2826683E+06
Nk2         0.4893394E+04
Nk3         0.2623860E+03
Nk4         0.2989000E+02
Pk1         0.9819629E+00
Pk2         0.1701720E-01
```

Posterior samples

File ‘prefix.hyp’ gives posterior parameter estimates for each MCMC sample:

Replicate:	iteration number
Nsnp:	number of SNPs in model
Va:	genetic variance explained by SNPs
Ve:	residual variance
Nk1,…,Nkk:	number of SNPs in mixture components 1 to *k*
Vk1,…,Vkk:	sum of squares of SNP effects in mixture component 1 to *k*

```
Example
Replicate       Nsnp          Va            Ve          Nk1          Nk2
      5010       3415   0.3393990E+00   0.6650676E+00     284439         2724
      5020       3127   0.3829503E+00   0.6437855E+00     284727         2422
      5030       3089   0.2489850E+00   0.7413293E+00     284765         2491
      5040       3448   0.4474014E+00   0.5840653E+00     284406         2639

Nk3         Nk4          Vk1           Vk2           Vk3           Vk4
688           3   0.0000000E+00  0.9341668E-01  0.2401173E+00  0.1302938E-01
703           2   0.0000000E+00  0.9547626E-01  0.2713150E+00  0.1900638E-01
586          12   0.0000000E+00  0.6552167E-01  0.1472237E+00  0.3183493E-01
805           4   0.0000000E+00  0.1125713E+00  0.3406390E+00  0.4072086E-02
```

Additional SNP information (optional)

Provides additional information for SNPs selected within the model (one line per iteration). Output is in sparse format: 'mixture class:SNP#:effect size'. The SNP number (SNP#) corresponds to the row number of the SNP in the PLINK "bim" file.

```
Example
3:65391:0.265405E-03 3:65442:0.741372E-02…
```

Example

A small example is provided using simulated data from the 14th QTL-MAS workshop (http://jay.up.poznan.pl/qtlmas2010/index.html). The analysis was performed using the command:

```
bayesR –bfile example/simdata –out simout –numit 10000 –burnin 5000 \
–seed 333
```

The results in file "simout.model" should look like this:

```
Mean        0.6844976E+01
Nsnp        0.7942960E+03
Va          0.4708471E+00
Ve          0.5293373E+00
Nk1         0.9236704E+04
Nk2         0.6231800E+03
Nk3         0.9661400E+02
Nk4         0.7450200E+02
Pk1         0.9207753E+00
Pk2         0.6204400E-01
Pk3         0.9654387E-02
Pk4         0.7526350E-02
Vk1         0.0000000E+00
Vk2         0.2880827E-01
Vk3         0.4424978E-01
Vk4         0.3927205E+00
```