Manual for BAMSE
Bayesian Association for Multiple SNP Effects

Anders Albrechtsen

November 12, 2007

Contents

0.1 Introduction

BAMSE is a software for performing association studies for unrelated individuals. We have developed a method based on Bayesian statistics that can model interactions for a large number of SNPs and environmental risk factors while accounting for the multiple testing problem. More specifically we have developed a Markov Chain Monte Carlo [1] method that allows for identification of sets of SNPs and environmental factors that when combined increase disease risk or change the distribution of a quantitative trait. In this method, combinations of genetic and environmental genetic factors define risk sets. Individuals with genotypes and environmental factors that are members of a risk set constitute a risk group that have modified distributions of disease risk or quantitative trait value. Phenotypic traits are modelled using normal distributions (for quantitative traits) or using a binomial distribution (for case/control data). A Markov chain is established with state space on the set of all possible risk sets and parameters (e.g. mean and variance) of trait value distributions for each risk set. The stationary distribution of the Markov chain is given by the posterior density of risk sets and trait value distributions. Statistical inferences are based on this posterior distribution. The Markov chain is simulated using a reversible jump [2] version of the Metropolis-Hastings algorithm, to jump between risk sets. This method differs fundamentally from previous approaches by entertaining non-linear models and by addressing the multiple testing problem in a computationally and statistically efficient manner. The present version of the program can successfully be employed on data sets of thousands of individuals and a couple of hundred SNPs.

0.1.1 Risk sets

Risk sets are combinations of SNPs and/or environmental factors. An example of a risk set could be
{SNP5=Heterozygote or homozygote, SNP67=Wild type}
or
{SNP45=Homozygote, age¿46}.
Here each risk set has two components. The individuals who fits the description of all the component is a risk set will constitute a risk group.

0.2 input data

The input file must be in the following format for k individuals, s SNPs, e environmantal parameters, and a adjustment factors:
$phe^1\ snp_1^1\ snp_2^1 \ldots snp_s^1\ env_1^1\ env_2^1 \ldots env_e^1\ adj_1^1\ adj_2^1 \ldots adj_n^a$
$phe^2\ snp_1^2\ snp_2^2 \ldots snp_s^2\ env_1^2\ env_2^2 \ldots env_e^2\ adj_1^2\ adj_2^2 \ldots adj_a^2$
$\vdots$
$phe^k\ snp_1^k\ snp_2^k \ldots snp_s^k\ env_1^k\ env_2^k \ldots env_e^k\ adj_1^k\ adj_2^k \ldots adj_a^k$

where phe^i is the phenotype (response variable) for the ith individual, snp_i is the ith SNP, env_i is the ith environmental factor and adj_i is the ith adjustment factor. The number of columns should be the sum of the phenotype + the number of SNPs + the number of environmental factors + the number of adjustment variables (1+s+e+a). See also Section 0.6 and the test files.

phe The phenotype - any number - no missing data is allowed. The trait must be normally distributed or binary (0 and 1)

snp The SNP - three categories (AA,Aa,aa) denoted 1, 2, 3. 0 is missing data. 1 mean homozygote for the A allele while 2 is heterozygotes. Which number represents homozygote for the minor allele, heterozygous is not important.

env The environmental factor - any number - no missing data allowed. Any affin transformation of the environmental factor will not change the results.

adj The adjustment factor - any number - no missing data is allowed. A linear relationship with the phenotype is assumed (like a covariate in a linear model).

0.3 options file

The option file must be modified according to the input data.

```
500     the number of individuals
20      the number of SNPs
0       the number of environmental factors
1       the starting number of risk sets
14      the maximum allowed risk sets
0       the minimum allowed risk sets
4       the maximum allowed active components
1       the minimum allowed active components
0.5     the prior for the number of active components ~geo(PriorActive)
0.5     the prior for the number of risk sets ~geo(PriorRiskSet)
200     the maximum value for the standard deviation
100     the number kappa will be multiplied with
0       0: the empirical average will be used, 1: the midrange will be used
10      the starting value for the SD
1       1: prints the risk sets, 0: does not
100000  the number of iterations
0       the number of iterations discarded (the burn in)
100     the thinning rate
5       the ratio for updating the mean
0       0: the trait is normally distributed. 1: the trait is binary
0       the  number of permutations
1       1: risksets have a higher mean than the none risk set
0       the number of adjustment factors
0       range min
0       range max: if range min =0 and range max = 0...
0.90    this fraction of the individual used to calculate.........
0  0: use frequencies for prios for missing genotypes. 1: use custom priors
0.33  the cutoff for accepting a infered genotype.
```

NB! the maximum allowed active parameters must never exceed the number of environmental factors + the number of SNPs.

the number of individuals The number of individuals in the input file

the number of SNPs The number of SNPs in the input file

the number of environmental factors The number of environmental factors in the input file

the starting number of risk sets the number of risk sets in the first state of the algorithm

the maximum allowed risk sets integer ≥ 1.

the minimum allowed risk sets integer ≥ 0 and less or equal to the maximum allowed risk sets

the maximum allowed active components integer ≥ 1. This number is also the maximum allowed order of an interaction.

the minimum allowed active components integer ≥ 1 and less or equal to the maximum allowed active parameters

the prior for the number of active parameters geo(PriorActive) [0,1] The prior for the order of the interactions assuming a geometric distribution. Since the number order is finite (the maximum allowed active parameters) the distribution is normalized to sum to one. See figure figure 1 on page 8

the prior for the number of risk sets geo(PriorRiskSet) [0,1] The prior for the number of risk sets assuming a geometric distribution. Since the number order is finite (the maximum allowed risk sets) the distribution is normalized to sum to one. See figure figure 1 on page 8

the maximum value for the standard deviation some large enough number

the number kappa will be multiplied with a number ≥ 0. The higher the number the lower the prior is for the extreme phenotypes. If the number is low, e.g. 1, the prior for the means of the risk sets is fairly flat.

the empirical average will be used 1: the midrange will be used The prior for the means of the risk sets i assumed to be normally distributed. The mean for this normal distribution is either the average of the observed phenotypes or the median of the phenotypes

1: prints the risk sets, 0: does not Always choose 1 to get any output for further analysis.

the number of iterations An integer. The MCMC algorithm must be run for at least several thousands iterations. The more SNPs and environmental factors that are included the longer the algorithm must be run. Recommendation: First try to run the algorithm for $100,000$ iteration one or two times and then do some convergence diagnostics (See also section 0.6.1 on page 8).

the number of iterations discarded (the burn in) The burn in can be specified in the analysis of the output but can also be specified here (if no convergence diagnostic will be performed). Recommendation: chose 0 and remove the burn in later.

the tinning rate the rate of the sampling. For a long run time with large or complex data the thinning rate should be high e.g. 100 or 500

the ratio for updating the mean the number of times the mean is updated in each iteration. A high number can improve mixing but for complex data where many (¿2) risk sets is frequently sampled a high number will slow the algorithm severely.

0: the trait is normally distributed. 1: the trait is binary If the trait is binary the trait should consist of 0 and 1. No adjustment is allowed for the binary test.

the number of permutations The number of permutations to be performed. NB! start with a small number of permutation since large data sets can give memory problems. The permutations will give a p-value based on the posterior distributions of the number of risk sets e.i. the p-value for the global association to the trait. For large data sets permutation test can be very slow.

risk sets have a higher mean than the none risk set should the risk sets have a higher mean than the mean of the none-risk set. If this is chosen then the label switching problem will be removed.

the number of adjustment factors integer.

range min The minimal allowed mean for the risk sets. (see range max)

range max: if range min =0 and range max = 0... The maximal allowed mean for the risk sets. If both range min and range max are 0 the min and max is calculated from the data. (see this fraction..)

this fraction of the individual used to calcu... this value $(0, 1]$ is the fraction of the data that is used to calculate the minimal and maximal range of the risk sets means. If 1 all of the range of the data is used. If for example 0.9 is used then the range is between the 5% highest and the 5% lowest values of the observed trait.

use frequencies for prios for missing genotypes. 1: use custom priors If there is missing genotype the prios can be estimated using several method. If 0 is chosen their prios are estimated from the frequencies of the allele in the sample assuming Hardy Weinberg equilibrium. Custom prior estiamted from for example PHASE or fastPHASE can also be given. The format of the custom priors is the fastPHASE output format.

the cutoff for accepting a infered genotype a threshold for infering missing genotypes as observed if there probability is larger than this threshold. Only relevant if custon priors for the missing genotypes are used.

0.4 Running BAMSE

The program is written in C++ and can run on either window or Linux.

For window users get the bamse.exe file and the options file. Open the command prompt and go to the folder with the exe file and the options.txt file. For linux users you need the bamse file and the option file. Run the program from a terminal. The command line is:

```
bamse -o optionfilename inputfilenames > output.file
```

if the -o optionfilename is not provided then the program assumes the the optionfile is called option.txt and is located in the same folder. For Linux users the path must be given, e.g. ./bamse. Multiple inputfilenames can be given but they must all use the same optionfile.

If a custom prior for the missing genotypes are provided the path and name for this file must also be provided.

```
bamse -o optionfilename -m missingdatafilename inputfilenames > output.file
```

0.5 Analysing the results

The C++ program only gives the posterior distribution for the number of risk sets. Even though this is the primary measure for association it is probably of interest to see which factors and combinations of factors that cause this association. We have built an R package to read and interpret the C++ output.

0.5.1 using the R package

Install the R package from the local directory (file called BAMSE.zip for Windows and BAMSE_1.0.tar.gz for linux). Then import the package in R

```
>library(BAMSE)
```

A short description for the main function is given here and in Section 0.6 but there is a more thorough description in R function descriptions. The names and a short description can be seen by the command

```
>help(package=BAMSE)
```

and each function also has a description that can be accessed by typing ? and the function name e.g.

```
>?read.mcmc
```

where read.mcmc is the function name.

read in one or more outputs

```
>mcmc1<-read.mcmc("C:/BAMSE MANUAL/set5.mcmcres")
>mcmc2<-read.mcmc("C:/BAMSE MANUAL/set5.mcmcres2")
```

NB! If the result is very large this can take several minutes and requires a large amount of memory especially for window users. The output is stored in the objects here called mcmc1 and mcmc2. For a short description of the data just type the object name.

```
>mcmc1
```

0.5.2 Posterior probabilities

The measure of the association here is the posterior probabilities. This measure is in the range $(0, 1)$ and a value close to 0 is evidence against there being a association and a value close to 1 is evidence for there being a association. The BAMSE method gives three levels of association:

Posterior distribution of the number of risk sets This is the overall association measure. If any of the SNPs or environmental factors are associated with the trait, though marginal or though interaction, the posterior probability for their being a least one risk set will be high.

Posterior probability for a component being part of a risk set This posterior will give a probability for a component (SNP or environmental factor) being associated with the trait (though marginal or though interaction effects).

Posterior probability for a risk set or a cluster of risk sets This posterior will give a probability for a specific combination of components or multiple similar combinations of components being associated with the trait

The three measures does not need to be corrected for multiple testing. The association can also be shown as Bayes factors but this measure will need to be corrected for multiple testing.

0.6 Example

The test input files

set1.test 20 SNPs with no genetic effect

set2.test 20 SNPs with one SNP with a dominant effect

set3.test 20 SNPs with one pair of interacting SNPs

set4.test 20 SNPs with two pairs of interacting SNPs

set5.test 20 SNPs with one pairs of interacting SNPs and LD

All of the files have 500 individuals with a mean phenotype of 100 for unaffected individuals and an elevated phenotypes for affected.

We recommend a burn in of at least 1000 iterations and a run time of at least $50,000$. However the burn in does not need to be chosen when running the C++ program.

For linux the command could look like this

```
./bamse set1.test>set1.mcmcres
```

where set1.mcmcres is the name of the output file. Using the default option file this should take less than a minute.

The posterior distribution for the number of risk sets can be seen at the bottom of the result file. If the posterior probability for there being at least one risk set is low (¡0.3) there is no association with the phenotype (1-posterior for there being 0 risk sets). If it is higher the result can be further explored using the R package called BAMSE.

The prior can be visualized using the plot.priors function (See Figure 1)

```
plot.priors(mcmc1,col=2)
```

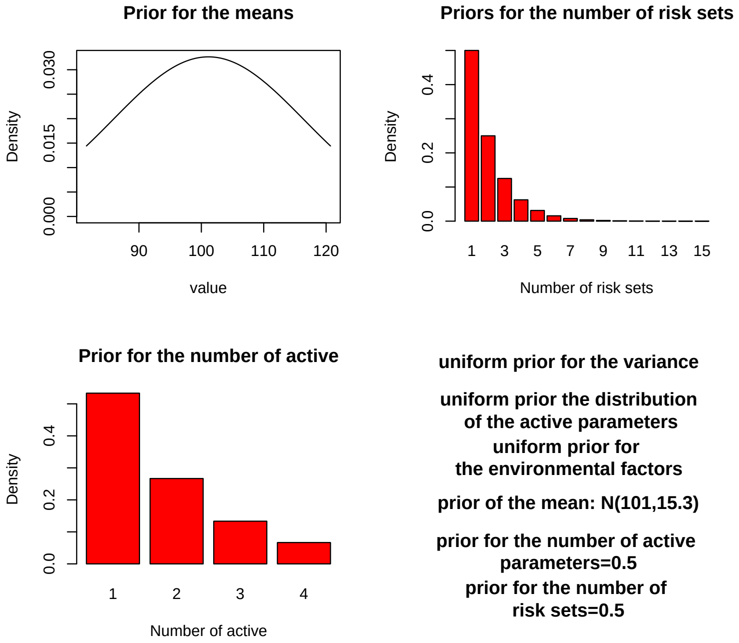

Figure 1: The priors (command:plot.priors(mcmc1,col=2))

0.6.1 Convergence diagnostic

To perform convergence diagnostic run the C++ program twice on the same data using the same settings. Try for example set4 and load the data into R as described in Section 0.5.1 so that the mcmc object is called mcmc1.

```
>library(BAMSE)
>mcmc1<-read.mcmc("C:/BAMSE MANUAL/set4.mcmcres")
>mcmc2<-read.mcmc("C:/BAMSE MANUAL/set4.mcmcres2")
```

It is very important that the chains have converged to stationarity. To view the parameters that have been sampled in each iteration type

```
>plot.mics(mcmc1)
```

or

```
>plot.mics(mcmc1,mcmc2)
```

For a more formal measure of convergence use Gelman and Rubin's diagnostic that gives a potential scale reduction factor. A low factor (e.g.¡1.05) give confidence that the chains have converged and are sampling from the stationary distribution.

```
>mcmc.diag(mcmc1,mcmc2,burnin=10000)
```

0.6.2 Association evaluation

To show graphic of the result try

```
>plot.standard(mcmc1,burnin=10000)
```

to show the likelihood before and after removing the burn in, show the posterior distribution of the number of risk sets and the posterior probability for each if the SNP and environmental parameters being part of a risk set. This function uses the functions plot.like, plot.riskset, and plot.visit.

Finally the most frequently sampled risk sets can be seen by the command

```
>mcmc.risksets(mcmc1,thres=0.1)
```

where the threshold defines how frequent the risk set has to be sampled.

Each SNP or pair of SNPs can also be explored using

```
 plot.geno(mcmc1,nr.snp=c(20))
 plot.geno(mcmc1,nr.snp=c(1,20))
 plot.visit2(mcmc1)
 plot.visit2(mcmc1,nr.snp=20)
 plot.visit2(mcmc1,nr.snp=c(1,20))
```

In Test set 5 there is strong LD between the SNPs (see Figure 2). The data is simulated so that the individuals who are carriers of the minor allele at both SNP 1 and SNP 20 have a higher phenotype. Because of the LD it is hard to see which of the SNPs that are the functional SNPs. The p-values and the posterior probabilities for being part of a risk set can be seen in Figure 3. As seen in Figure 2 there is high LD for SNP 1 with SNP 7, 9, 17 and SNP 20 with SNP 2. Using the plot.visit function it can be seen that SNP 1 very often sampled with SNP 2 or SNP 20 (plot.visit2(mcmc1,1)) and SNP 20 is often sampled with SNP 7,9,17 or 20 (plot.visit2(mcmc1,20)).

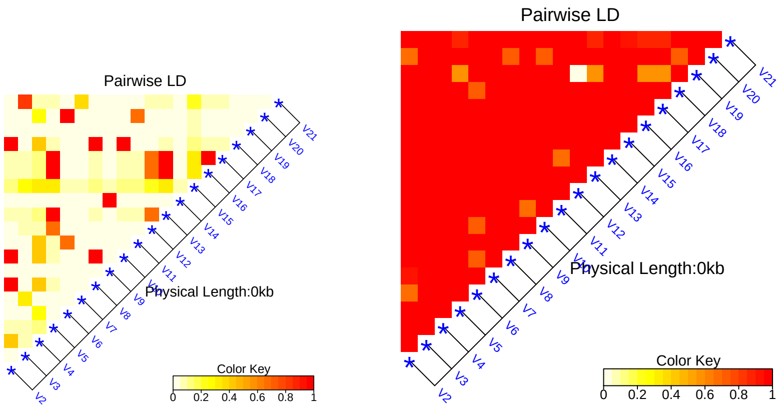

Figure 2: The LD patters of the twenty SNPs, measure as the squared correlation coefficient (r^2)(left) and as D' (right)

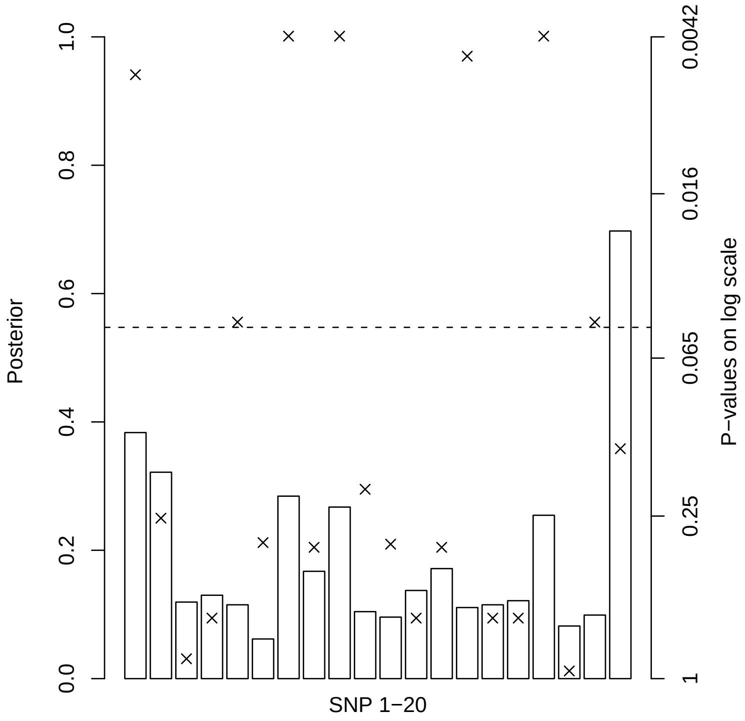

Figure 3: The posterior for being part of a risk set and the p-valued for the full model vs. the null model 2df (command:with.p(mcmc1,"C:/BAMSE MANUAL/set5.test"))

Bibliography

[1] W R Gilks, S Richardson, and D J Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1997.

[2] Peter J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.