CAVIARBF Manual

Updated 4/28/2015

Wenan Chen

Introduction

There are two executables if the code is successfully built: *caviarbf* and *model_search*. *caviarbf* is used to generate the Bayes factor file, which is used by *model_search* to calculate different model statistics, for example, the marginal posterior inclusion probability (PIP), i.e., the marginal probability of each SNP being causal.

Build and Test the Tool

To install the tool, simple run the following in a Unix-like system.

```
bash build.sh
```

To build it in a Windows system, install the Cygwin environment first and then run the above.

To test the Tool, run the following:

```
bash test.sh
```

Program *caviarbf*

A typical run

```
./caviarbf -z ./example/myfile.z -r ./example/myfile.LD -t 0 -a 0.1281429 -n
2000 -c 5 -o ./example/myfile.sigma0.1281429.bf
```

Type the following to get the help information

```
./caviarbf --help
```

```
USAGE:

   ./caviarbf  -z <string> -r <string> -t <integer> -a <numeric> -n
               <integer> -c <integer> -o <string> [--] [--version] [-h]


Where:

   -z <string>,  --zfile <string>
     (required)  input file for marginal test statistics

   -r <string>,  --rfile <string>
     (required)  input file for correlation matrix

   -t <integer>,  --prior-type <integer>
     (required)  prior type for variant effect size

        0: specify sigmaa
```

```
      1: specify the proportion of variance explained (pve)

-a <numeric>,  --prior-value <numeric>
  (required)  the prior value associated with the prior type

-n <integer>,  --sample-number <integer>
  (required)  the total number of samples in the data

-c <integer>,  --max-causal <integer>
  (required)  the maximal number of causal variants in the model

-o <string>,  --output <string>
  (required)  the output file name for Bayes factors

--,  --ignore_rest
  Ignores the rest of the labeled arguments following this flag.

--version
  Displays version information and exits.

-h,  --help
  Displays usage information and exits.


Calculate Bayes factors based on summary statistics
```

Input File Format for *caviarbf*

Marginal test statistics file: a text file with 2 or 3 columns separated by spaces. For example (an example used by CAVIAR),

snp1 1.2
snp2 5.1
snp3 1.9
cnv1 -6.1
cnv2 -3.2
cnv3 -4

The third column is optional, which specifies the variance of each variant. In this case, we assume the effect size is not related to the allele frequency of the variant. The result will be similar to BIMBAM. Without the third column of variance of each variant, we assume that the effect size is larger when the minor allele frequency is smaller.

Correlation matrix file: a matrix of the correlation among variants. For example (again an example used by CAVIAR)

1.000 0.840 0.050 0.050 0.001 -0.010
0.840 1.000 0.040 0.040 -0.010 -0.007
0.050 0.040 1.000 0.950 0.060 -0.001
0.050 0.040 0.950 1.000 0.065 0.003
0.001 -0.010 0.060 0.065 1.000 0.018

-0.01 -0.007 -0.001 0.003 0.018 1.000

Output File Format for *caviarbf*

The output file is a text file with Bayes factors. It has the same format as that from BIMBAM. Here is an example:

note:bf=log10(Bayes Factor), SNP IDs are from 1 ... m

bf	se	snp1	snp2	snp3	snp4	snp5
-0.591455	NA	1	NA	NA	NA	NA
+4.658971	NA	2	NA	NA	NA	NA
-0.127742	NA	3	NA	NA	NA	NA
+7.052327	NA	4	NA	NA	NA	NA
+1.289039	NA	5	NA	NA	NA	NA
+2.519908	NA	6	NA	NA	NA	NA
+9.619151	NA	1	2	NA	NA	NA
-0.485480	NA	1	3	NA	NA	NA
+6.794833	NA	1	4	NA	NA	NA
+0.954240	NA	1	5	NA	NA	NA
+2.145068	NA	1	6	NA	NA	NA
+4.577784	NA	2	3	NA	NA	NA
+12.322104	NA	2	4	NA	NA	NA
+6.055098	NA	2	5	NA	NA	NA
+7.274689	NA	2	6	NA	NA	NA
+84.266736	NA	3	4	NA	NA	NA
+1.568898	NA	3	5	NA	NA	NA
+2.617624	NA	3	6	NA	NA	NA
+7.998226	NA	4	5	NA	NA	NA
+9.658610	NA	4	6	NA	NA	NA
+3.923135	NA	5	6	NA	NA	NA

The first line is a comment. The second line is the header. The first column is the Bayes factor. The rest columns indicate which SNPs/variants are in the model. NA means not in the model.

Recommendations about parameters

For -t option, 0 (setting sigmaa) is fully tested. 1 is in an experimental stage. When using sigmaa, 0.1 seems to be a generally good value for -a. Other values can be tried include 0.2, 0.4 as recommended by BIMBAM. When running on many SNPs, be careful not to set the -c option too large because it may take too much time and also space to store the output. You can start with 1 or 2 and increase it by 1 in each trial. If two settings show similar results, then there is no need to increase it further.

Program *model_search*

A typical run

```
./model_search -i ./example/pref4.multi.txt -m 50 -p 0 –
o ./example/pref4.multi.txt.prior0
```

Type the following to get the help information

```
./model_search --help
```

```
USAGE:

   ./model_search  {-p <numeric>|-f <string>} -i <string> -m <integer> [-s]
                   [-e] [-x] -o <string> [--] [--version] [-h]


Where:

   -p <numeric>,  --prior <numeric>
     (OR required)  the prior probability of each SNP being causal, in the
     range [0, 1)

     if it is 0, the prior probability will be set to 1 / m,

     where m is the number of variants in the region
         -- OR --
   -f <string>,  --prior-file <string>
     (OR required)  the file name specifying the prior probabilities of
     variants


   -i <string>,  --input <string>
     (required)  input file storing Bayes factors

   -m <integer>,  --snp-number <integer>
     (required)  the total number of variants in the data

   -s,  --stepwise
     output stepwise result with rho confidence level

   -e,  --exhaustive
     output exhaustive search result with rho confidence level

   -x,  --mixed
     output result first using exhaustive and then stepwise search with rho
     confidence level

   -o <string>,  --output <string>
     (required)  output file prefix

   --,  --ignore_rest
     Ignores the rest of the labeled arguments following this flag.

   --version
     Displays version information and exits.

   -h,  --help
     Displays usage information and exits.


   Search models and output probabilities based on Bayes factors
```

Input File Format for *model_search*

Bayes factor file: The output file from *caviarbf*. Since the format is the same as the output from BIMBAM, it can also be used to process BIMBAM output Bayes factors.

Prior probability file: This file can be used to specify different priors of being causal for different variants. Each row corresponds to a variant in the region. For example:

0.02
0.03
0.02
0.05
0.02
0.01

If all the priors are the same, we can also use -p to specify the common prior directly.

Ouput File Format for *model_search*

Posterior inclusion probability (PIP) file: The file has a .marginal suffix. It has two columns. The first column is the index of each variant in the data. The second is the PIP in a descending order.

Statistics file: This file has a .statistics suffix, and contains some statistics information. The first is the ratio between the likelihood of the data averaging over all models to the likelihood of the data under the global null model: no variant is causal. The second is the probability of at least 1 causal variant in the region. The third is the Bayes factor of the region (all alternative models vs. the global null model).

ρ-confidence level file: This file has a .stepwise suffix. It has the same format as the PIP file. Each row shows the ρ-confidence level when including the current variant and all variants above this row. To generate this file you need to specify the -s option.

Other experimental output files: .exhaustive file outputs the best ρ-confidence level by exhaustive search of models for each model size. Due to the time cost, we only do that until model size 4. This needs the -e option to generate it. .exhaustivestepwise file outputs the ρ-confidence level by first applying the exhaustive search up to model size 4 and then switching to a stepwise search. This needs the -x option.