CAVIARBF Manual

Wenan Chen

Attention

The coding scheme in the calculation of the correlation (LD) matrix should be consistent with that of the z scores, i.e., the allele coded as 1 in the calculation of z scores needs to be coded as 1 in the calculation of the correlation (LD) matrix and vice versa. Because this is how everything is calculated when we have the full genotype data: everything calculated is based on the same coding scheme. Otherwise the results using summary statistics may be arbitrary.

One way to keep the consistency is to code alleles following the coding in a reference panel provided by an imputation software. For example when using IMPUTE2, we can align our alleles to the two a0 and a1 alleles, and a0 allele is coded as 0 and a1 allele is coded as 1.

Introduction

Now there is an R package to run multiple loci fine mapping with annotations. The following describes the input output file formats, including the new format for multiple loci and annotations which is only by the R package. Please see the R function help for more details on using the R pacakge.

For the C++ version of single locus fine mapping, 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.

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
```

Option -a now can support multiple values, for example, -a 0.1,0.2,0.4 to allow multiple values to be used and the final Bayes factor is an average among different values. This will be more robust when we are uncertain about the effect size. For quantitative trait, an exact Bayes factor can be calculated instead of an approximate Bayes factor. This is critical when the sample size is small or the effect size is large, for example, in expression quantitative trait loci (eQTL) analysis. To use the old approximate Bayes factors, use the option --appr. Use the approximate Bayes factor for a binary trait because no exact Bayes factor is available. The option -e can be used to add a small value to the diagonal of the correlation matrix. This can improve the performance when the correlation matrix is an approximation, such as from a reference panel.

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 for GWAS. Other values can be tried include 0.2, 0.4 as recommended by BIMBAM or use the robust multiple value version 0.1,0.2,0.4. For eQTL analysis, a multiple value option 0.1,0.2,0.4,0.8,1.6 is reasonable. 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
```

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.

Multiple loci fine mapping input files

The design of input files is similar to that used in PAINTOR. For each loci, there are three files: the marginal test statistics (z-score) file, the correlation matrix (LD) file and an annotation file. The format of the first two are the same as described above as in single locus fine mapping. The annotation file is simply a text file of a matrix, each row corresponds to a variant, each column corresponds to a type of annotation. The naming of these three files are as follows:

<z-score-file>

<z-score-file>.LD

<z-score-file>.annotations

Other suffixes of the the LD file and the annotation file can also be used as along as they are consistent among all files. Finally, a file containing the list of loci is required, each line is the z-score file name for each locus. All these files need to be in the same folder.

Multiple loci fine mapping output files

PIP file: It has a .marginal suffix. The format is the same as the single locus fine mapping output

PIP file with the variant ID: It has a .marginalz suffix. The first two columns are the same as that in the PIP file. The rest three columns are: the locus index, the variant ID and the z score.

Likelihood file: It has a .loglik suffix. It stores the relative log likelihood.

Annotation effect file: It has a .gamma suffix. The first column stores the annotation effect sizes (coefficients) for the intercept and each annotation. The second column is the coefficient of standardized annotations, i.e., the annotation is scaled to have unit variance. The first element of the second column is simply 0. Note that these coefficients are shrunken estimates if any penalized model is used.

Time cost: It has a .time suffix. It saves the running time. The column "elapsed" is usually used for measuring the time cost.

Log file: It has a .log suffix recording the running logs.

Individual annotation significance file: It has a .stats suffix. This is created when "topK" is used. The first column is the twice the difference of the log likelihood between the model including each annotation and the null model (only the intercept). The first row is the null model. The second and the third column are the estimated coefficients for the intercept and the annotation effect when including each annotation only in the model. The fourth column is the p-value by comparing the first column with a chi-square distribution of one degree of freedom. This is the significance of each annotation without considering other annotations.

Top K selected annotations: It has a .topk suffix. This is the sequentially selected top K relatively independent annotations using the "topK" option. There can be less than K annotations selected due to the correlation and p-value threshold constraints. Note that it might be overfittng or suboptimal when using these top selected annotations again for another round of fine mapping.