piMASS user manual

Yongtao Guan
Baylor College of Medicine

Version 0.90
19 Janurary 2011
Revised on 25 March 2015
Updated on 28 Sept 2015

Contents

1 Copyright

2 What's in the package

You should find two executables: pimass-mac for Mac OS X (an earlier release named pimass) and pimass-lin for Linux. You should also find two directories: "input/" directory that contains example files, and "src" directory contains source files. And you should also find this tex file and the pdf file generated from it.

3 What the piMASS can do

The software is developed and maintained by Yongtao Guan (ytguan@gmail.com), based on the work in the paper https://projecteuclid.org/euclid.aoas/1318514285. Please refer to the paper for the details of the method. piMASS is designed to do the following:

- Perform multi-SNP association analysis either genome-wide, or on a small region.
- Estimate proportion of phenotypic variance explained by the genotypes (the second column in pref.path.txt).
- Estimate the posterior of the number of SNPs that have effects on phenotype (the fourth column in pref.path.txt).
- Report the marginal posterior inclusion probabilities of each SNP, a measure for the marginal association (in pref.mcmc.txt).

4 Input file formats

The users should prepare genotype files and phenotypes, described below. Both allow missing values. However, individuals who have either missing genotype or missing phenotype will be excluded from the study. Therefore, we recommend users to perform imputation on the original genotype

to fill in the missing values. An alternative, but less ideal, solution is to fill in a missing genotype with the mean genotypes of that SNP. For file that have multiple columns, different columns can be separated by either a coma, a blank space, or a semicolon.

4.1 Mean genotype file format

The genotype file is identical in format to the mean genotype file used in `BIMBAM`, not surprisingly. Each row of the file contains a SNP. The first column is the SNP ID, and the second and third columns are allele types with minor allele first. The rest columns are the mean genotypes of different individuals – numbers between 0 and 2 that represents the (posterior) mean genotype, or dosage of the minor allele. An example of mean genotypes file of two SNPs and three individuals follows.

```
rs1, A, T, 0.02, 0.80, 1.50
rs2, G, C, 0.98, 0.04, 1.00
```

4.2 Phenotype file format

In the phenotype input file, each line is a phenotype value for an individual. Missing phenotypes should be denoted as NA. The number of lines in the phenotype file should be equal to the number of individuals in genotype file (`N`), otherwise `piMASS` will either throw away the values after `N` rows or append "NA" to achieve total `N` rows. In either case, a warning will be printed. Users want to make sure that individuals are in the same order between genotype and phenotype files. An example phenotype file with 3 individuals is below.

```
1.2
NA
-0.3
```

If the phenotypes are binary (e.g. in a case-control study) then the format is the same, but each entry should be 0, 1 or NA. It does not matter cases are assigned 1 or 0.

4.2.1 Normal quantile transformation

For quantitative traits an important assumption underlying the methods implemented in `piMASS` is that the phenotype has a normal distribution within each genotype class. We suggest using a normal quantile transformation to transform the phenotype to be normal before running `piMASS` . For example, in `R`, this can be accomplished using `xtransformed = qqnorm(x,plot.it =F)$x`.

This quantile transformation does not fully solve the problem (it ensures that the phenotype is normal overall, but not necessarily normal within each genotype class). However, with the small effect sizes typical in genetic association studies it appears to be a simple sensible way to guard against strong departures from modeling assumptions. If you have other covariates that may be important predictors of phenotype (e.g. Age, Sex) we suggest first regressing the phenotype on these covariates using standard multiple linear regression software, and then running `piMASS` on the residuals from this regression (after applying a normal quantile transformation to these residuals).

4.3 SNP location file format

The file contains three columns: the first column is the SNP ID, the second column is its physical location, and the third column contains its chromosome number. Note, it is OK if the rows are not ordered according to position, but the file must contain all the SNPs in the genotype files. If the genotype files contain SNPs across different chromosome, `piMASS` will sort SNPs based on its chromosome and position. An example position file is below.

```
rs1, 1200,  1
rs2, 4000,  1
rs3, 3320,  1
```

Note that this file is strictly needed only if the order of the SNPs in the genotype file is not the same as the order of their physical locations along the chromosome, or if multiple genotype and phenotype files are used (see below). To align SNPs in the correct order is important for the sampling procedure because we propose exchange nearby SNPs to facilitate the mixing among correlated SNPs.

4.4 Multiple genotype and phenotype files

In some cases it may be convenient to provide genotypes (and corresponding phenotypes) in multiple files. For example, in a genome-wide study, it may be helpful to have one genotype file containing the case data, and a second genotype file containing the control data. Or one genotype files containing individuals in the first stage of the study and the another contains the second.

When using multiple genotype files `piMASS` does not require that the same SNPs be present in both files (although if the same SNP is present in both files then the SNP identifier should be the same in both files, to convey this information). However, SNPs missing in one of the files will cause the SNP to be excluded from the study because of the missingness.

When using multiple genotype files, the user must also provide multiple phenotype files, with each phenotype file corresponding to the individuals in a genotype file. The exception to this is that, for a case/control study, the case phenotypes can be specified by `-p 1` and control phenotypes can be specified by `-p z`.

4.4.1 The strand issue

When merging genotypes from different studies, there arises the issue of whether or not the genotypes for a SNP were obtained on the same strand. In some cases this can be checked easily: for example, if a SNP in one study is A/G, and in the other is T/C, we infer that the two studies used different strands, and we can flip one of the SNPs to correct this. `piMASS` performs these kinds of flip automatically. However, if a SNP is A/T, or C/G, one cannot tell whether the strandedness is the same or different across studies without external information. In this situation, `piMASS` assumes that genotypes for a single SNP in multiple input files refer to the same strand.

Note: if genotypes at a SNP are not compatible with the SNP being bi-allelic, even after strand flips (as might happen when multiple genotypes are used, see below), then the SNP is considered to be "bad" and `piMASS` will exclude the SNP from the study.

5 Running the piMASS

First some general comments:

- piMASS is a command line based program. The command should be typed in a terminal window, in the directory in which piMASS executable exists.
- The command line should be all on one line: the line-break in the example is only because the line is too long to fit on one page.
- Unless otherwise stated, the "options" (-g -p -pos -o, etc.) are all case-sensitive.

Now we illustrate how to use piMASS through examples.

1. **A single genotype file and a single phenotype file**

   ```
   ./pimass -g input/test.mgt -p input/test.ph -w 10000 -s 100000 -o pref -num 10
   ```

   The command line will run MCMC with burn-in steps 1000 and sampling steps 100000, every 10 steps a sample will be recorded. The output file names will begin with pref.

2. **Multiple genotype files and multiple phenotype files**

   ```
   ./pimass -g cohort1.txt -p pheno1.txt  -g cohort2.txt \
     -p pheno2.txt  -w 100000 -s 1000000 -o pref  -pos pos.txt  -num 10
   ```

   This command line takes two genotype files and two phenotype files, merge them based on the SNP position files. piMASS will run 100k warm-up steps and follow by 1M sampling steps, every 100 steps record states sampled.

3. **Binary phenotypes**

   ```
   ./pimass -g case_mgt.txt -p 1 -g ctrl_mgt.txt -p z -pos pos.txt \
                    -o pref  -w 10000 -s 1000000 -num 100
   ```

   This command line asks piMASS to take two genotype files. The -p 1 assign all individuals in the matching genotype (case_mgt.txt in the example) as 1, and -p z assign all individuals in the matching genotype (ctrl_mgt.txt in the example) as 0. piMASS will run 10K warm-up steps, followed by 1M sampling steps, and every 100 steps piMASS records a sample in the state space.

4. **Setup other parameters**

   ```
   ./pimass -g cohort.txt -p pheno.txt  -w 10000 -s 100000 -o pref -num 10 \
           -hmin 0.01 -hmax 0.5 -pmin 1 -pmax 1000 -smin 1 -smax 200
   ```

   This command line is identical to the first command line in the list except that it specifies ranges for the hyper-parameters h and p and the restrictions on the number of SNPs in the model. Specifically, it specifies that minimum and maximum of the h is 0.01 and 0.5 respectively, the minimum and maximum of p is 1 out of total number of SNPs and 1000 out of total number of SNPs, and it restricts the number of SNPs in a model to be between 1 and 200.

6 Output files

The `piMASS` , upon finishing, produces five output files in a directory named `output/`. Such a directory will created automatically in the path where the executable is located if it does not already exist. The names of output files begin with *prefix* specified with the `-o` option. We now describe the contents of these output files.

6.1 Log file: `prefix.log`

A log file contains the command line used, running log, and warnings generated. When sending in a bug report, it is important to include the log file as an attachment.

6.2 SNP information file: `prefix.snp.txt`

This file contains 10 columns, and each row contains information of a single SNP. The columns are rsID, minor allele, major allele, minor allele frequency, variance estimates of the SNP genotypes, chromosome, position, numerical ID, log10(BF), and the estimated effect size. Among them, the numerical ID is used in the `prefix.gamma.txt` to record which SNPs are in the model.

6.3 Sampling path file: `prefix.path.txt`

This file contains `9` columns, each row corresponding a sampled states (excluding gamma). The first column is the `h`, the second column is the re-estimated `h` based on sampling posterior effect sizes, the rest columns are `log10(p)`, number of SNPs, acceptance ratio for the local proposals, acceptance ratio for the long-range proposals, the derived prior on effect sizes, `log(BF)` of the model, and the log(prior probability of the model). Samples recorded approximate the posterior distribution of the state space. Of particular interest, the second column with header `hh` is a better estimate of the heritability than the first column.

6.4 Sampling path file: `prefix.gamma.txt`

The first column of the file is the number of SNPs in the model, the rest columns are SNP IDs, when the number of SNPs is smaller than the maximum allowed number of SNPs, NAs are appended.

6.5 The posterior inclusion estimates: `prefix.mcmc.txt`

For quantitative phenotypes, the file contains SNP ID, chromosome, position, estimates of the posterior inclusion probabilities based on simple counting, estimates of the posterior inclusion probabilities based on Rao-Blackwellization, the naive estimates of the posterior effect size, and Rao-Blackwellized estimates of the posterior effect size.

For binary phenotype (when `-cc` is used), the output file is different in that the two columns of the Rao-Blackwellized estimates are no longer there.

7 Bug fix

- Fix a bug with malloc that fail by very large datasets. (28 Sept 2015)
- Roll back to the corrected source code. (28 Sept 2015)

Appendix A: `piMASS` Options

Unless otherwise stated, *arg* stands for a string, *num* stands for a number.
File I/O related options:

- -g *arg* can use multiple times, must pair with -p.
- -p *arg* can use multiple times, must pair with -g. *arg* can be a file name; z or 1, which indicates the pairing genotype individuals have phenotype 0 or 1.
- -pos *arg* can use multiple times. *arg* is a file name.
- -o *arg* *arg* will be the prefix of all output files, the random seed will be used by default.
- -cc calc bf of logit regression on binary phenotype.

Parameter setup:

- -w(warm) *num* specify steps of warm up in MCMC.
- -s(step) *num* specify steps of sampling steps in MCMC.
- -num *num* specify thinning, record one states in every *num* steps.
- -r *num* specify random seed, system time by default.
- -hmin *num* specify minimum value for h; default value 0.001.
- -hmax *num* specify maximum value for h; default value 0.999.
- -pmin *num* specify minimum value for p; default value 1.
- -pmax *num* specify maximum value for p; default value 300.
- -smin *num* specify minimum number of SNPs in the model; default value 0.
- -smax *num* specify maximum number of SNPs in the model; default value 300.

Other options:

- -v(ver) print version and citation
- -h(help) print this help
- -exclude-maf *num* exclude SNPs whose maf is less than *num* , default 0.01.
- -exclude-nopos *num* exclude SNPs that has no position information, 1 = yes (default), 0 = no
- –silence no terminal output.