DPR User Manual

Ping Zeng and Xiang Zhou

E-mail: pingzeng@umich.edu and xzhousph@umich.edu

October 2, 2016

Contents

1 Introduction

1.1 What is DPR

DPR is the software implementing the latent Dirichlet Process Regression (DPR) model for robust genetic prediction of complex traits with typed genotypes. While most existing GWAS prediction models use parametric priors, DPR makes use of a flexible non-parametric prior on the SNPs effect sizes. By using a non-parametric model, DPR is adaptive to a broad spectrum of genetic architectures and can achieve robust predictive performance for a variety of complex traits. DPR can be fitted using two complementary algorithms: the Monte Carlo Markov Chain (MCMC) algorithm, and the mean filed variational Bayesian (VB) approximation algorithm.

1.2 How to Cite DPR

Zeng Ping, Zhou Xiang. Robust Genetic Prediction of Complex Traits with the Latent Dirichlet Process Regression Models. 2016.

1.3 The DPR Model

DPR fits the following multiple linear regression model

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}, \mathbf{u} \sim N(0, \sigma_b^2 \sigma_e^2 \mathbf{K}), \boldsymbol{\varepsilon} \sim N(0, \sigma_e^2 \mathbf{I}_n), \quad (1)$$

where $\mathbf{y}$ is an n-vector of phenotypes measured on n individuals; $\mathbf{W}$ is an n by c matrix of covariates including a column of 1s for the intercept term; $\boldsymbol{\alpha}$ is a c-vector of coefficients; $\mathbf{X}$ is an n by p matrix of genotypes; $\boldsymbol{\beta}$ is the corresponding p-vector of effect sizes; $\boldsymbol{\varepsilon}$ is an n-vector of residual errors where each element is assumed to be independently and identically distributed from a normal distribution with variance σ_e^2. Using the stick-breaking construction of Dirichlet process [1-5], we specify an infinite normal mixture prior on β for SNP j

$$\beta_j \sim \pi_1 N(0, 0\times\sigma_e^2) + \sum_{k=2}^{\infty} \pi_k N(0, \sigma_k^2\sigma_e^2),$$
$$\pi_k = \nu_k \prod_{l=1}^{k-1}(1-\nu_l), \nu_k \sim \text{Beta}(1,\lambda). \quad (2)$$

For the hyper-parameters we consider limiting priors for $\boldsymbol{\alpha}$, σ_k^2, σ_b^2, σ_e^2, and λ. In the special case $\mathbf{K} = \mathbf{X}\mathbf{X}^T / p$, we can decompose $\mathbf{u} = \mathbf{X}\boldsymbol{b}$, which can be viewed as the combined effect of all small effects, with a prior on b for SNP j as

$$b_j \sim N(0, \sigma_b^2\sigma_e^2 / p). \quad (3)$$

Thus the total effect size for a given SNP j is $\beta_j + b_j$.

To fit the DPR model, we develop two complementary algorithms: one is based on the Markov Chain Monte Carlo (MCMC) algorithm, and the other is based on the variational Bayesian (VB) approximation. The MCMC sampling algorithm is accurate but computationally slow. The variational Bayesian algorithm is computationally fast, but can be less accurate.

1.4 Missing Data

1.4.1 Missing Genotypes

As in our previous algorithm GEMMA [6, 7], the tricks used in the DPR algorithm rely on having complete or imputed genotype data at each SNP. Individuals with missing phenotypes will not be included in DPR analysis. That is, DPR requires the user to impute all missing genotypes before fitting. This imputation step is arguably preferable than simply dropping individuals with missing genotypes, since it can improve the prediction accuracy. Therefore, for fitting DPR, missing genotypes are recommended to be imputed first. Otherwise, any SNPs with missingness above a certain threshold (default 5%) will not be analyzed, and missing genotypes for SNPs that do not pass this threshold will be simply replaced with the estimated mean genotype of that SNP. For predictions, though, all SNPs

will be used regardless of their missingness. Missing genotypes in the test set will be replaced by the mean genotype in the training set.

1.4.2 Missing Phenotypes

All individuals will be used for calculating the genetic relatedness matrix **K**, so that the resulting relatedness matrix is still an $n \times n$ matrix regardless of how many individuals have missing phenotypes. For relatedness matrix calculation, because missingness and minor allele frequency for a given SNP are calculated based on analyzed individuals (i.e. individuals with no missing phenotypes and no missing covariates), if all individuals have missing phenotypes, then no SNPs and no individuals will be included in the analysis and the estimated relatedness matrix will be full of nan's.

2 Installing and Compiling DPR

If you have downloaded a binary executable, no installation is necessary. In some cases, you may need to use "chmod a+x DPR" before using the binary executable. In addition, notice that the end-of-line coding in Windows (DOS) is different from that in Linux, and so you may have to convert input files using the utility *dos2unix* or *unix2dos* in order to use them in a different platform.

If you want to compile DPR by yourself, you will need to download the source code, and you will need a standard C/C++ compiler such as GNU gcc, as well as the GSL and LAPACK libraries. You will need to change the library paths in the Makefile accordingly. A sample Makefile is provided along with the source code. For details on installing GSL library, please refer to http://www.gnu.org/s/gsl/. For details on installing LAPACK library, please refer to http://www.netlib.org/lapack/. The DPR software also relies on the Eigen library (http://eigen.tuxfamily.org/index.php?title=Main_Page), which is included in the source code folder.

3 Input File Format

DPR requires three input files containing genotypes, phenotypes and relatedness matrix. Genotype and phenotype files can be in two formats, either both in the PLINK binary ped

format [8] or both in the BIMBAM format [9]. Mixing genotype and phenotype files from the two formats (for example, using PLINK files for genotypes and using BIMBAM files for phenotypes) will result in unwanted errors. BIMBAM format is particularly useful for imputed genotypes, as PLINK codes genotypes using 0/1/2, while BIMBAM can accommodate any real values between 0 and 2 (and any real values if paired with "-notsnp" option).

3.1 PLINK Binary PED File Format

DPR recognizes the PLINK binary ped file format (http://pngu.mgh.harvard.edu/~purcell/plink/) for both genotypes and phenotypes. This format requires three files: *.bed, *.bim and *.fam, all with the same prefix. The *.bed file should be in the default SNP-major mode (beginning with three bytes). One can use the PLINK software to generate binary ped files from standard ped files using the following command:

```
plink --file [file_prefix] --make-bed --out [bedfile_prefix]
```

For the *.fam file, DPR only reads the second column (individual id) and the sixth column (phenotype). One can specify a different column as the phenotype column by using "-n [num]", where "-n 1" uses the original sixth column as the phenotype, and "-n 2" uses the seventh column, and so on and so forth. DPR will read the phenotypes as provided and will recognize either "-9" or "NA" as missing phenotypes.

DPR codes alleles as 0/1 according to the plink webpage on binary plink format (http://pngu.mgh.harvard.edu/~purcell/plink/binary.shtml). Specifically, the column 5 of the *.bim file is the minor allele and is coded as 1, while the column 6 of the *.bim file is the major allele and is coded as 0. The minor allele in column 5 is therefore the effect.

For prediction problems, one is recommended to list all individuals in the file, but label those individuals in the test set as missing (i.e. label as "-9" or "NA"). This will facilitate the use of the prediction function implemented in DPR.

3.2 BIMBAM File Format

DPR also recognizes BIMBAM file format (http://stephenslab.uchicago.edu/software.html), which is particularly useful for imputed genotypes as well as for general covariates other than SNPs. BIMBAM format consists of three files, a mean genotype file, a phenotype file, and an optional SNP annotation file. We explain these files in detail below.

3.2.1 Mean Genotype File

This file contains genotype information. The first column is SNP id, the second and third columns are allele types with minor allele first, and the remaining columns are the posterior/imputed mean genotypes of different individuals numbered between 0 and 2. An example mean genotype file with two SNPs and three individuals is as follows:

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

DPR codes alleles exactly as provided in the mean genotype file, and ignores the allele types in the second and third columns. Therefore, the minor allele is the effect allele only if one codes minor allele as 1 and major allele as 0. One can use the following bash command (in one line) to generate BIMBAM mean genotype file from IMPUTE genotype files (http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html) [10]:

```
cat [impute filename] | awk -v s=[number of samples/individuals]  '{ printf $2 "," $4 "," $5; for(i=1; i<=s; i++) printf "," $(i*3+3)*2+$(i*3+4); printf "\n" }' > [bimbam filename]
```

Notice that one may need to manually input the two quote symbols ’. Depending on the terminal, a direct copy/paste of the above line may result in “-bash: syntax error near unexpected token ‘(’ ” errors.

Finally, the mean genotype file can accommodate values other than SNP genotypes. One can use the “-notsnp” option to disable the minor allele frequency cutoff and to use any numerical values as covariates.

3.2.2 Phenotype File

This file contains phenotype information. Each line is a number indicating the phenotype value for each individual in turn, in the same order as in the mean genotype file. Notice that only numeric values are allowed and characters will not be recognized by the software. Missing phenotype information is denoted as "NA". The number of rows should be equal to the number of individuals in the mean genotype file. An example phenotype file with five individuals and one phenotype is as follows:

```
1.2
NA
2.7
-0.2
3.3
```

One can include multiple phenotypes as multiple columns in the phenotype file, and specify a different column for association tests by using "-n [num]", where "-n 1" uses the original first column as phenotypes, and "-n 2" uses the second column, and so on and so forth. An example phenotype file with five individuals and three phenotypes is as follows:

```
 1.2   -0.3   -1.5
 NA    1.5    0.3
 2.7   1.1    NA
-0.2   -0.7   0.8
 3.3   2.4    2.1
```

For prediction problems, one is recommended to list all individuals in the file, but label those individuals in the test set as missing. This will facilitate the use of the prediction function implemented in DPR.

3.2.3 SNP Annotation File (optional)

This file contains SNP information. The first column is SNP id, the second column is its base-pair position, and the third column is its chromosome number. The rows are not required to be in the same order of the mean genotype file, but must contain all SNPs in that file. An example annotation file with four SNPs is as follows:

rs1, 1200, 1
rs2, 1000, 1
rs3, 3320, 1
rs4, 5430, 1

If an annotation file is not provided, the SNP information columns in the output file will have “-9” as missing values.

3.3 Relatedness Matrix File Format

DPR requires a relatedness matrix file in addition to both genotype and phenotype files. The genetic relatedness matrix can be supplied in two different ways: either use the original relatedness matrix, or use the eigen values and eigen vectors of the original relatedness matrix.

3.3.1 Original Matrix Format

DPR takes the original relatedness matrix file in two formats. The first format is an $n \times n$ matrix, where each row and each column corresponds to individuals in the same order as in the *.fam file or in the mean genotype file, and i^{th} row and j^{th} column is a number indicating the relatedness value between i^{th} and j^{th} individuals. An example relatedness matrix file with three individuals is as follows:

```
 0.3345   -0.0227    0.0103
-0.0227    0.3032   -0.0253
 0.0103   -0.0253    0.3531
```

The second relatedness matrix format is a three column “id id value” format, where the first two columns show two individual id numbers, and the third column shows the relatedness value between these two individuals. Individual ids are not required to be in the same order as in the *.fam file, and relatedness values not listed in the relatedness matrix file will be considered as 0. An example relatedness matrix file with the same three individuals above is shown below:

```
id1   id1    0.3345
id1   id2   -0.0227
id1   id3    0.0103
id2   id2    0.3032
```

id2 id3 -0.0253
id3 id3 0.3531

As BIMBAM mean genotype files do not provide individual id, the second format only works with the PLINK binary ped format. One can use "-km [num]" to choose which format to use, i.e. use "-km 1" or "-km 2" to accompany PLINK binary ped format, and use "-km 1" to accompany BIMBAM format.

3.3.2 Eigen Value and Eigen Vector Format

DPR can also read the relatedness matrix in its decomposed forms using "-d" and "-u" options. To do this, one should supply two files instead of one: one file containing the eigen values ("-d") and the other file containing the corresponding eigen vectors ("-u"). The eigen value file contains one column of n_a elements, with each element corresponds to an eigen value. The eigen vector file contains an $n_a \times n_a$ matrix, with each column corresponds to an eigen vector. The eigen vector in the i^{th} column of the eigen vector file should correspond to the eigen value in the i^{th} row of the eigen value file. Both files can be generated from the original relatedness matrix file by using the "-eigen" option in DPR. Notice that n_a represents the number of analyzed individuals, which may be smaller than the number of total individuals n.

4 Running DPR

4.1 An Example Dataset

If you downloaded the DPR source code recently, you will find an "example" folder containing a small GWAS example dataset. This data set comes from the heterogeneous stock mice data, kindly provided by Wellcome Trust Centre for Human Genetics on the public domain http://mus.well.ox.ac.uk/mouse/HS/, with detailed described in [11].

The data set consists of 1,904 individuals from 85 families, all descended from eight inbred progenitor strains. We selected two phenotypes from this data set: the percentage of CD8+ cells, with measurements in 1,410 individuals; mean corpuscular hemoglobin (MCH), with measurements in 1,580 individuals. A total of 1,197 individuals have both

phenotypes. The phenotypes were already corrected for sex, age, body weight, season and year effects by the original study, and we further quantile normalized the phenotypes to a standard normal distribution. In addition, we obtained a total of 12,226 autosomal SNPs, with missing genotypes replaced by the mean genotype of that SNP in the family. Genotype and phenotype files are in both BIMBAM and PLINK binary formats.

For demonstration purpose, for CD8, we randomly divided the 85 families into two sets, where each set contained roughly half of the individuals (i.e. inter-family split) as in [6]. Therefore, the phenotype files contain four columns of phenotypes. The first column contains the quantitative phenotypes CD8 for all individuals. The second column contains quantitative phenotypes CD8 for individuals in the training set. The third column contains quantitative phenotypes CD8 for individuals in the test set. The fourth column contains the quantitative phenotypes MCH for all individuals.

4.2 SNP filters

The are a few SNP filters implemented in the software.

- Polymorphism. Non-polymorphic SNPs will not be included in the analysis.
- Missingness. By default, SNPs with missingness below 5% will not be included in the analysis. Use "-miss [num]" to change. For example, "-miss 0.1" changes the threshold to 10%.
- Minor allele frequency. By default, SNPs with minor allele frequency below 1% will not be included in the analysis. Use "-maf [num]" to change. For example, "-maf 0.05" changes the threshold to 5%.
- Correlation with any covariate. By default, SNPs with r2 correlation with any of the covariates above 0.9999 will not be included in the analysis. Use "-r2 [num]" to change. For example, "-r2 0.999999" changes the threshold to 0.999999.
- Hardy-Weinberg equilibrium. Use "-hwe [num]" to specify. For example, "-hwe 0.001" will filter out SNPs with Hardy-Weinberg p values below 0.001.
- User-defined SNP list. Use "-snps [filename]" to specify a list of SNPs to be included in the analysis.

Calculations of the above filtering thresholds are based on analyzed individuals (i.e. individuals with no missing phenotypes and no missing covariates). Therefore, if all individuals have missing phenotypes, no SNP will be analyzed and the output matrix will be full of "nan"s.

4.3 Estimate Genetic Relatedness Matrix from Genotypes

4.3.1 Basic Usage

The basic usages to calculate an estimated genetic relatedness matrix with either the PLINK binary ped format or the BIMBAM format are:

```
./DPR -bfile [prefix] -gk [num] -o [prefix]
./DPR  -g [filename] -p [filename] -gk [num] -o [prefix]
```

where the "-gk [num]" option specifies which genetic relatedness matrix to estimate, i.e. "-gk 1" calculates the centered genetic relatedness matrix while "-gk 2" calculates the standardized genetic relatedness matrix; "-bfile [prefix]" specifies PLINK binary ped file prefix; "-g [filename]" specifies BIMBAM mean genotype file name; "-p [filename]" specifies BIMBAM phenotype file name; "-o [prefix]" specifies output file prefix. Notice that the BIMBAM mean genotype file can be provided in a gzip compressed format.

4.3.2 Detailed Information

DPR provides two ways to estimate the genetic relatedness matrix from genotypes, using either the centered genotypes or the standardized genotypes. We denote $\mathbf{X}$ as the $n \times p$ matrix of genotypes, x_i as its i^{th} column representing genotypes of i^{th} SNP, $\bar{x}_i$ as the sample mean and v_{x_i} as the sample variance of i^{th} SNP, and $\mathbf{1}_n$ as an $n \times 1$ vector of 1's. Then the two genetic relatedness matrices DPR can calculate are as follows:

$$G_c = \frac{1}{p}\sum_{i=1}^{p}(x_i - \mathbf{1}_n \bar{x}_i)(x_i - \mathbf{1}_n \bar{x}_i)^T,$$

$$G_s = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{v_{x_i}}(x_i - \mathbf{1}_n \bar{x}_i)(x_i - \mathbf{1}_n \bar{x}_i)^T.$$

Which of the two genetic relatedness matrix to choose will largely depend on the underlying genetic architecture of the given trait [6]. Specifically, if SNPs with lower minor allele frequency tend to have larger effects (which is inversely proportional to its genotype variance), then the standardized genotype matrix is preferred. If the SNP effect size does not depend on its minor allele frequency, then the centered genotype matrix is preferred. In our previous experience based on a limited examples, we typically find the centered genotype matrix provides better control for population structure in lower organisms, and the two matrices seem to perform similarly in humans.

4.3.3 Output Files

There will be two output files, both inside an output folder in the current directory. The prefix.log.txt file contains some detailed information about the running parameters and computation time, while the prefix.cXX.txt or prefix.sXX.txt contains an $n \times n$ matrix of estimated relatedness matrix.

4.4 Fit the Dirichlet Process Regression model

4.4.1 Basic Usage

The basic usages for fitting DPR with either the PLINK binary ped format or the BIMBAM format are:

```
./ DPR -bfile [prefix] -dpr [num] -o [prefix]
./ DPR -g [filename] -p [filename] -a [filename] - dpr [num] -o [prefix]
```

where the "-dpr [num]" option specifies which model to fit, i.e. "-dpr 1" fits DPR using variational Bayesian algorithm with fixed number of the normal components in the mixture prior, "-dpr 2" fits DPR using MCMC sampling with fixed number of the normal components in the mixture prior, "-dpr 3" fits DPR using MCMC sampling with adaptively selected number of the normal components in the mixture prior; "-bfile [prefix]" specifies PLINK binary ped file prefix; "-g [filename]" specifies BIMBAM mean genotype file name; "-p [filename]" specifies BIMBAM phenotype file name; "-a [filename]" (optional)

specifies BIMBAM SNP annotation file name; "-o [prefix]" specifies output file prefix. Notice that the BIMBAM mean genotype file can be provided in a gzip compressed format.

4.4.2 Detailed Information

Notice that a large memory is needed to fit DPR (e.g. may need 20 GB for a data set with 4000 individuals and 400,000 SNPs), because the software has to store the whole genotype matrix in the physical memory.

In default, DPR does not require the user to provide a relatedness matrix explicitly. It internally calculates and uses the centered relatedness matrix. Of course, one can choose to supply a relatedness matrix by using the "-k [filename]" option. In addition, DPR does not take covariates file when using DPR for prediction. The option "-dpr 1" fits DPR using variational Bayesian algorithm with fixed number of the normal components in the mixture prior, "-dpr 2" fits DPR using MCMC sampling with fixed number of the normal components in the mixture prior, "-dpr 3" fits DPR using MCMC sampling with adaptively selected optimal number (in the sense of smallest DIC) of the normal components in the mixture prior. For MCMC based methods, one can use "-w [num]" to choose the number of burn-in iterations that will be discarded, and "-s [num]" to choose the number of sampling iterations that will be saved. In addition, one can use "-m [num]" to select the number of top marginal SNPs to be included in the model (i.e. SNPs that have additional larger effects) to save computational time. To further improve computation, one can use "-t [num]" to update those non-selected, likely unimportant SNPs once every *t* iterations. One can use "-nk [num]" to choose the fixed truncated number of mixture normal distributions for DPR, and use "-mnk [num]" to choose the maximum truncated number of mixture normal distributions for the adaptive DPR. It is up to the users to decide these values for their own data sets, in order to balance computation time and computation accuracy.

The genotypes, phenotypes, as well as the relatedness matrix will be centered when fitting the models. The estimated values in the output files are thus for these centered values. Therefore, proper prediction will require genotype means and phenotype means from the individuals in the training set, and one should always use the same phenotype file (and the same phenotype column) and the same genotype file, with individuals in the test

set labeled as missing, to fit DPR and to obtain predicted values described in the next section.

4.4.3 Output Files

There will be two output files, all inside an output folder in the current directory. The prefix.log.txt file contains some detailed information about the running parameters and computation time. The prefix.param.txt contains the posterior mean estimates for the effect size parameters. An example le with a few SNPs is shown below:

chr	rs	ps	n_miss	b	beta	gamma
1	rs3094315	792429	0	-1.122508e-04	-2.673970e-04	1
1	rs4040617	817376	0	-7.706889e-05	0.000000e+00	0
1	rs2980300	819185	0	-5.200171e-05	0.000000e+00	0
1	rs4075116	825852	0	6.556617e-05	0.000000e+00	0
1	rs9442385	832343	0	-9.895405e-06	4.615158e-06	1

Notice that the beta column contains the posterior mean estimate for β_j ($\beta_j = \sum_{k=1}^{nk} \beta_{jk}\gamma_{jk}$), so the gamma column is always 0 or 1. Therefore, in the special case $\mathbf{K} = \mathbf{X}\mathbf{X}^T / p$, the total effect size estimate is $\hat{\beta}_j + \hat{b}_j$. Here b (i.e. $\hat{b}_j$) is given in equation (3) and beta (i.e. $\hat{\beta}_j$) is given in equation (1).

4.5 Predict Phenotypes Using Output from DPR

4.5.1 Basic Usage

The basic usages for association analysis with either the PLINK binary ped format or the BIMBAM format are:

```
./DPR -bfile [prefix] -epm [filename] -emu [filename] -predict -o [prefix]
./DPR -g [filename] -p [filename] -epm [filename] -emu [filename] -predict -o [prefix]
```

where the "-bfile [prefix]" specifies PLINK binary ped file prefix; "-g [filename]" specifies BIMBAM mean genotype file name; "-p [filename]" specifies BIMBAM phenotype file

name; "-epm [filename]" specifies the output estimated parameter file (i.e. prefix.param.txt file from DPR); "-emu [filename]" specifies the output log file which contains the estimated mean (i.e. prefix.log.txt file from DPR); "-o [prefix]" specifies output file prefix.

4.5.2 Detailed Information

DPR will obtain predicted values for individuals with missing phenotype, and this process will require genotype means and phenotype means from the individuals in the training set. Therefore, use the same phenotype file (and the same phenotype column) and the same genotype file, as used in fitting DPR. In the special case $\mathbf{K} = \mathbf{X}\mathbf{X}^T / p$, the predicted values for individual l is calculated as $X_l^{*T}(\hat{\boldsymbol{b}}+\hat{\boldsymbol{\beta}})+\hat{\mu}$, where X_i^* is the p-vector genotype for a new individual l and $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{\beta}}$ are the estimated p-vector SNPs effect sizes. Here, unlike in previous sections, all SNPs that have estimated effect sizes will be used to obtain predicted values, regardless of their minor allele frequency and missingness. SNPs with missing values will be imputed by the mean genotype of that SNP in the training data set.

4.5.3 Output Files

There will be two output files, both inside an output folder in the current directory. The prefix.log.txt file contains some detailed information about the running parameters and computation time, while the prefix.prdt.txt contains a column of predicted values for all individuals. In particular, individuals with missing phenotypes will have predicted values, while individuals with non-missing phenotypes will have "NA"s.

5 Options

File I/O Related Options

- -bfile [prefix] specify input plink binary file prefix; require .fam, .bim and .bed files
- -g [filename] specify input bimbam mean genotype file name
- -p [filename] specify input bimbam phenotype file name
- -n [num] specify phenotype column in the phenotype file (default 1); or to specify which phenotypes are used in the mvLMM analysis

- -a [filename] specify input bimbam SNPs annotation file name (optional)
- -k [filename] specify input kinship/relatedness matrix file name
- -km [num] specify input kinship/relatedness matrix file type (default 1, valid value 1 or 2)
- -d [filename] specify input eigen value file name
- -u [filename] specify input eigen vector file name
- -c [filename] specify input covariates file name (optional); an intercept term is needed in the covariates file
- -epm [filename] specify input estimated parameter file name
- -emu [filename] specify input log file name containing estimated mean
- -snps [filename] specify input snps file name to only analyze a certain set of snps; contains a column of snp ids
- -o [prefix] specify output file prefix (default "result")

SNP Quality Control Options

- -miss [num] specify missingness threshold (default 0.05)
- -maf [num] specify minor allele frequency threshold (default 0.01)
- -r2 [num] specify r-squared threshold (default 0.9999)
- -hwe [num] specify HWE test p value threshold (default 0; no test)
- -notsnp minor allele frequency cutoff is not used and so all real values can be used as covariates

Relatedness Matrix Calculation Options

- -gk [num] specify which type of kinship/relatedness matrix to generate (default 1; valid value 1-2; 1: centered matrix; 2: standardized matrix.)

Eigen Decomposition Options

- -eigen specify to perform eigen decomposition of the relatedness matrix

DPR Model Options

- -dpr [num] specify algorithm choice (default 1; valid value 1-3; 1: VB algorithm; 2: MCMC algorithm; 3 adaptive DPR)
- -nk [num] specify fixed number of normal components in the mixture prior (default 4; valid value≥2)

- -mnk [num] specify maximum number of normal components in the mixture prior for the adaptive DPR, based on DIC the optimal nk is selected among 2~mnk (default 6; valid value≥3)
- -m [num] specify number of top marginal SNPs to be included in the model (default 500; i.e. SNPs that have additional larger effects)
- -t [num] update those non-selected, likely unimportant SNPs once every t iterations (default 10)
- -w [num] specify burn-in steps for MCMC sampling (default 10000)
- -s [num] specify sampling steps for MCMC sampling (default 10000)
- -sp [num] specify iterations to select the optimal nk in the adaptive DPR; the burn-in steps are equal to w×sp and the sampling steps are s×sp (default 0.1; suggested value between 0~1)

Prediction Options

- -predict specify to perform prediction

6 An axample for the mouse data

To fit a quantitative trait (i.e. CD8) using DPR with VB algorithm

./bin/DPR -g ./example/mouse_hs1940.geno.txt.gz -p ./example/mouse_hs1940.pheno.txt -n 2 -a ./example/mouse_hs1940.anno.txt -k ./example/mouse_hs1940.cXX.txt -dpr 1 -nk 4 -o mouse_hs1940_CD8_vb

Explain:

"-g" specifies BIMBAM genotypes, "-p" specifies phenotypes, "-a"specifies annotation file, "-k" specifies relatedness matrix, "-dpr 1" specifies fitting DPR using VB algorithm, "-nk 4" specifies four normal components included in into the mixture prior, "-o" specifies the output file.

./bin/DPR -bfile ./example/mouse_hs1940 -n 2 -k ./example/mouse_hs1940.cXX.txt -dpr 1 -nk 4 -o mouse_hs1940_CD8_vb

Explain:

"-bfile" specifies plink files, i.e. mouse_hs1940.fam, mouse_hs1940.bim and mouse_hs1940.ped, "-n" specifies phenotypes using the 7th column of mouse_hs1940.fam, "-k" specifies relatedness matrix, "-dpr 1" specifies fitting DPR using VB algorithm, "-nk

4" specifies four normal components included in into the mixture prior, "-o" specifies the output file.

To fit a quantitative trait (i.e. CD8) using DPR with MCMC algorithm

./bin/DPR -bfile ./example/mouse_hs1940 -n 2 -k ./example/mouse_hs1940.cXX.txt -dpr 2 -nk 4 -w 10000 -s 10000 -o mouse_hs1940_CD8_mcmc

Explain:

"-bfile" specifies plink files, i.e. mouse_hs1940.fam, mouse_hs1940.bim and mouse_hs1940.ped, "-n" specifies phenotypes using the 7th column of mouse_hs1940.fam, "-k" specifies relatedness matrix, "-dpr 2" specifies fitting DPR using MCMC algorithm, "-nk 4" specifies four normal components included in into the mixture prior, "-w 10000" specifies 10000 burn-ins, "-s 10000" specifies 10000 samplings after burn-in, "-o" specifies the output file.

To fit a quantitative trait (i.e. CD8) using adaptive LDR

./bin/DPR -bfile ./example/mouse_hs1940 -n 2 -k ./example/mouse_hs1940.cXX.txt -dpr 3 -mnk 6 -sp 0.2 -w 10000 -s 10000 -o mouse_hs1940_CD8_ada

Explain:

"-bfile" specifies plink files, i.e. mouse_hs1940.fam, mouse_hs1940.bim and mouse_hs1940.ped, "-n" specifies phenotypes using the 7th column of mouse_hs1940.fam, "-k" specifies relatedness matrix, "-dpr 3" specifies fitting DPR using adaptive LDR, "-mnk 6" specifies the maximum number of normal components included in the mixture prior, "-sp 0.2" specifies samplings for this adaptive process, "-w 10000" specifies 10000 burn-ins, "-s 10000" specifies 10000 samplings after burn-in, "-o" specifies the output file. More specifically, this command will repeat fitting MCMC sampling for nk=2 to nk=6, for each nk (i.e. 2, 3, 4, 5 and 6), the burn-in number is 2000=10000×0.2 (w×sp), the MC sampling number is also 2000=10000×0.2 (s×sp). After the optimal nk (say nk*) is selected based on the smallest DIC across nk=2 to nk=6, this command continuous perform DPR using MCMC sampling with nk* normal components, and now the burn-in number is 10000, the MC sampling number is 10000.

Reference

1. Ferguson TS. A Bayesian analysis of some nonparametric problems. Ann Stat. 1973; 1: 209-230.
2. Andrews DF, Mallows CL. Scale mixtures of normal distributions. J R Stat Soc Ser B. 1974; 36: 99-102.
3. Sethuraman J. A constructive definition of Dirichlet priors. Stat Sinica. 1994; 4: 639 - 650.
4. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc. 2001; 96.
5. Blei DM, Jordan MI. Variational inference for Dirichlet process mixtures. Bayesian Anal. 2006; 1: 121-143.
6. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013; 9: e1003264.
7. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44: 821-824.
8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559-575.
9. Guan Y, Stephens M. Practical Issues in Imputation-Based Association Mapping. PLoS Genet. 2008; 4: e1000279.
10. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet. 2009; 5: e1000529.
11. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet. 2006; 38: 879-887.