Call the program on a linux system as

./fbayesWG

(‘W’ stands for optional weighting of SNPs and ‘G’ for optional, iterative updating of gamma.) The algorithm depends on the settings in inputWG.par and nuisance.par.

**Input parameters (for general call of the program)**

snpfile CHARACTER; filename, file contains design matrix of SNP genotypes X (0,1,2 for homzogygous minor allele, heterozygous, homozygous major allele, resp.)

**X has dimension nxp, n=number of individuals, p number of SNPs**

datafile CHARACTER; filename, data file has n rows and several columns with individual observations and other covariates or factor levels

skiplines INTEGER; number of lines to skip in datafile (because of header or the like)

posobs INTEGER; column in datafile with the phenotype to be analysed

weightFile CHARACTER; optional weighting of SNPs is possible (which is realised similarly to a weighted regression approach); if applicable, enter a filename (file contains one column of dimension p containing the weights>=0; 0=exclude SNP; 1=neutral weight), **otherwise write some non-existing file name** (e.g. ‘xx’ but not empty)

outPrefix CHARACTER; prefix for all output files, see below

maxit INTEGER; in case of non-convergence of the algorithm, the maximum number of iterations to run

stepnum INTEGER; if stepnum>0 then a credible interval for each SNP effect is calculated (very time-consuming), stepnum devides the range of plausible values for genetic effects in intervals of equal size; then a grid search is applied to determine the limits of credibility interval; the precision CI depends on stepnum – the larger the better

INFERENCES ON SIGNIFICANCE CAN BE DONE WITHOUT CACULATING CREDIBLE INTERVALS; measure of evidence (similar to p-value) is calculated anyway

reparamMethod INTEGER; reparametrisation of genotype codes; important when non-additive effects have to be estimated; 1: Cockerham’s model, 2: Alvarez-Castro & Carlborg (depends on genotype frequencies rather than on allele frequencies), 3:Zeng et al. which is Cockerham’s model up to a constant for dominance effects; for a setting different from 1,2,3 nothing is done with the genotype codes – this could be useful for a plain regression analysis

ALLELE and GENOTYPE FREQUENCIES ARE CALCULATED FROM THE SAMPLE. This could be a critical issue if you deal with heavily unbalanced data with respect to family structure or the like.

solveNuisance Logical; (notation **.true.** or **.false.**) whether nuisance effects should be solved (done by iterative BLUE)

solveDom Logical; whether dominance effects should be solved

solveEpi Logical; whether epistatic effects should be solved (if solveDom=.false. then only addxadd interaction is considered otherwise all four possible ways of interaction)

maf REAL; minor allele frequency is used for filtering SNPs

CHOOSE MAF LIMIT OF AT LEAST 0.05 WHEN WORKING WITH REPARAMMETHOD 1 OR 3, OTHERWISE – IN CASE OD DOMxDOM INTERACTION IS INVOLVED – ALGORITHM MAY COLLAPSE BECAUSE OF BAD CONDITIONING

nonZeroMain REAL; prior guess for proportion of non-zero additive or dominance effects

nonZeroEpi REAL; prior guess for proportion of non-zero non-additive effects

gammaFix Logical; whether gamma (proportion of non-zero genetic effects) is fixed in the analysis; if gammaFix=.false. then gamma is updated iteratively depending on the posterior density p(gamma|y) for each SNP

lambdaFix Logical; whether lambda (variance parameter of a genetic effect) is fixed; **not used yet**

writeFreq Logical; whether allele frequencies should be written in file ‘alleleFreq.txt’

errorlevel REAL; type-I error for inferring significance of genetic effects based on measure of evidence

**Input parameters (only if nuisance effects should be estimated as well, otherwise it is neglected)**

nEff INTEGER; number of factors and covariates to be specified

IF MORE THAN 10 EFFECTS ARE REQUIRED, THE SOURCE CODE HAS TO BE CHANGED

columnDat list of INTEGER of length nEff; columns in datafile containing factors or covariates

effectType list of CHARACTER of length nEff; I: factor levels are integer; A: factor levels are alphanumeric values; R regression of covariate; R2 first AND second order regression on covariate (up to R9); order of entries as in columnDat

setZero list of INTEGER of length nEff; which factor level is treated as baseline (typically the first); does not matter in case of regression; order of entries as in columnDat

**Output files**

LOCI ARE NUMBERED FROM 1:2p IF DOMINANCE EFFECTS ARE CONSIDERED.

outPrefix.eff

* loc\_1: first locus
* loc\_2: second locus in case of epistatic effects, otherwise equal to loc\_1; if index is larger than p (=number of SNPs), subtract p to obtain true index of locus (using an index from 1..2p helps to distinguish additive from dominance gene action); type of effect:
  + loc\_1 1..p additive
  + loc\_1 p+1..2p dominance
  + loc\_1 1..p and loc\_2 1..p additive x additive
  + loc\_1 1..p and loc\_2 p+1..2p
    - loc\_2-p>loc\_1 additive x dominance
    - loc\_2-p<loc\_1 dominance x additive
  + loc\_1 p+1..2p and loc\_2 p+1..2p dominance x dominance
* gPostExp: estimated genetic effect
* gamma: probability of non-zero genetic effect
* measOfEvid: measure of evidence (Bayesian analogue to p-value)
* sign: 0/1 for non-significant/significant genetic effects based on preset errorlevel
* cred\_t1: if applicable, lower bound of credible interval, otherwise -999
* cred\_t2; if applicable, upper bound of credible interval, otherwise 999

Note that in case of epistasis only those effects are stored which are absolutely larger than 0.0001 to save memory space.

outPrefix.res

* first column contains yHat=genetic effects+fixed effects
* second column contains residuals resid=y-yHat

outPrefix.log

File contains information about filtering of SNPs due to MAF setting

**Additional output file only if solveNuisance=.true.**

coding\_of\_level.txt

* factor: number of factor (1..nEFF); 0 for general mean
* level\_numeric: internal value of factor level
* level\_original: original value of factor level corresponding to datafile

outPrefix.nui

File contains estimated fixed effects (order corresponds to coding\_of\_levels.txt)

**A note on extensions of fastBayes which are not published**

1. The current version of fastBayes includes the option of estimating the proportion of non-zero genetic effects (i.e. gamma) as well. For this purpose, the prior density of gamma was specified as

p(gamma) = (a+1)(1-gamma)a

similar to Scott & Berger (2006) J Statist Plann Inference 136:2144 – 2162, because it is assumed that gamma is likely close to zero. Based on the prior setting, the posterior distribution p(gamma|y), can be calculated analytically from which the estimate of gamma is assessed as conditional expectation, similar to the estimation of a genetic effect. As all theory is based on a one-locus model, the final gamma (in the current iteration) is estimated as mean from p 1-locus-model estimates. In total, there are three specific gammas: for additive, dominance and all epistatic effects.

1. Inferences on significance are based on the measure of evidence, which require the calculation of the posterior distribution of a genetic effect Pr(g|y). This approach can be referred to Pereira and Stern (1999) Entropy 1:99-110.

**A main concern is whether to fix gamma or not. With a bad prior guess (nonZeroMain or nonZeroEpi), the algorithm hardly converges. With the prior assumption as mentioned above, estimates of gamma, however, only little change during iterations! I am still thinking about alternative prior choices. When dealing with additive and/or dominance effects, my current suggestion is to define a plausible range for gamma (e.g. a grid over 1d-6 to 1d-1) and to let a cross-validation decide what the most appropriate value for gamma is. For this purpose fastBayes is used with setting gammaFix=.true., and nonZeroMain coincides with the value on the grid. Afterwards fastBayes is run using the best guess and the setting gammaFix=.false.. This 2-step approach can also be done in case of epistasis, but I have never tried.**

Due to the prior settings, fastBayes works most reliably with standardised phenotypic observations.

Good luck!

Dörte Wittenburg