Evaluation of Double Hierarchical Generalized Linear Model versus Empirical Bayes method — A comparison on Barley dataset

Author: Lei Wang

Supervisor: Majbritt Felleki

D-level in Statistics, Spring 2010

School of Technology and Business Studies

Dalarna University, Sweden

Contents

Evaluation of Double Hierarchical Generalized Linear Model versus Empirical Bayes method — A comparison on Barley dataset

Lei Wang

Supervisor: Majbritt Felleki

School of Technology and Business Studies, Högskolan Dalarna, SE-791 88, Falun, Sweden

ABSTRACT

Double hierarchical generalized linear models (DHGLM) have been applied in studies of genetics recently. It is a useful tool to model quantitative trait with respect to molecular marker effect, and thus Quantitative Trait Loci (QTL) analysis. Compared to Bayesian method to estimate marker effect and its variance, DHGLM has performed as well as Bayesian on simulated dataset of genetics. Hence, researchers are expecting to know how DHGLM will perform on real dataset. In this study, we apply DHGLM on famous Barley dataset, estimate marker-effect and its variance, and compare them to those of empirical Bayes (E-BAYES). All the possible QTL positions found by E-BAYES have been searched out also by DHGLM. Moreover, several new candidate markers imply QTL positions. DHGLM take advantage of fast computational algorithms and non-prior distribution of parameter requirement. It is efficient in genetic mapping and QTL analysis, and has great potential in plant and animal breeding.

KEY WORDS: DHGLM, marker-effect, variance component, QTL, Barley data

1. Introduction

The development of gene sequencing has opened up a completely new area in animal and plant breeding. Traditional method to investigate a quantitative trait is to make correlation analysis with phenotypic measurements and heritage relationship. The realization of mapping molecular markers provides an access to find out chromosome regions, even certain DNA fragments, which control a specific trait.

Quantitative trait loci (QTL) analysis plays an important role in revealing the genetic architecture of a trait. A QTL is a region of DNA that is associated with a particular phenotypic trait - these QTLs are often found on different chromosomes. Usually, QTLs can be identified based on molecular markers (for example, RFLP[1] marker). The main task of QTL analysis is to find out which markers are possible positions of QTLs.

QTL mapping studies usually genomic markers loci of large amount, which is much more than the number of observations. (Nengjun Yi, 2009) If the epistatic effects, which are interaction effects of different markers, are considered, the number of candidate markers will be too large to make regression analysis and variable selection. Classical generalized linear models cannot handle many correlated variables (marker effects) simultaneously. Another challenge is to estimate the variance of random effects, i.e., marker-effects, which explains the phenotypic variation caused by certain markers. Therefore, sophisticated methods are required to handle large number of marker effects and heterogeneity in the variance component of each marker.

In previous research, **empirical Bayesian method** (E-BAYES) is widely used to solve the heterogeneity and great improvement to classical GLM has been achieved. However, the prior distributions of parameters choosing and multi-dimensional quadrature usually make E-BAYES difficult to give a unified framework to broad class of models.

Hierarchical generalized linear models (HGLM), which are based on

[1] RFLP: restriction fragment length polymorphism. In experiment, DNA sample is broken into pieces in order to take some treatments, and such pieces are RFLPs. RFLP is one kind of molecular marker used in gene sequencing.

h-likelihood, can also give a good solution to heterogeneity in variance component. Moreover, Lee and Nelder (1996, 2006) and Lee et al. (2006) have shown how to model and make inferences using HGLM, without resorting to an E-BAYES framework. HGLM take the advantage of requiring neither priors nor multi-dimensional quadrature.

HGLM has already been applied in genetics. For example, Jaffrezic et al. (2000) used an HGLM for analysis of lactation curves with heterogeneous residual variances over time. Recently, Rönnegård et al. (2010) used DHGLM for fast variance component estimation in a model with genetic heterogeneity in the residual variance of an animal model. Rönnegård and Valdar (2010) have also suggested the use of DHGLM to detect variance-controlling QTL. The possible applications for DHGLM in studies of genetic heterogeneity in residual variance will be continued to be investigated. It is believed that DHGLM approach has a great potential for future use in genetics and application in plant and animal breeding.

The aim of this paper is to apply DHGLM on Barley data to find out QTLs, which controls the kernel weight of barleys. And compare the results to that of E-BAYES method, which was investigated by Xu (2007). As DHGLM has seldom been applied on real dataset until now, it is expected to see how this method will perform.

2. Materials and methods

2.1 Materials

The dataset is the well-known barley data from the North American Barley Genome Mapping Project (Tinker et al., 1996). Experimental populations of homozygous double-haploid (DH) lines were collected. There were seven agronomic traits (grain yield, days to heading, days to maturity, plant height, lodging severity, kernel weight, and test weight) in the original research. We just take one trait as response- the average value of kernel weight across environments - in our research. The reason to choose the kernel weight as the phenotype is to compare the results of DHGLM to that of empirical Bayesian method (Xu, 2007)

The DH population contained $N = 145$ lines, and each was grown in 25 different

environments. As mentioned above, the phenotype analyzed is the kernel weight. Mapping of more than 200 marker loci (mostly RFLP) was performed in the Harrington/TR306 DH population by Kasha et al. (1995). A subset of $j = 127$ markers was chosen to give a base map with relatively uniform coverage. The markers used in the QTL analysis were chosen at 2- to 5-cM[2] intervals. Since the experimental populations are homozygous DH lines, the genotypes of markers only have two kinds. Hence, the genotypes are coded as +1 for genotype A (one parent), -1 for genotype B (the other parent), and 0 for missing genotype. The total missing genotypes only accounted for less than 5% of all genotypes.

2.2 Methods

To find out QTL positions, it is intuitively to construct a model involving all the markers' effects, and then search for candidate positions according to marker effect and relative contribution to the total phenotypic variance. Treating each marker effect as a random effect u, the phenotype as response variable y, and the environmental effects as fixed effect β, naturally, we consider a generalized linear mixed model (GLMM). However, the heterogeneity in variance of marker effect violates the assumption of GLMM. Another shortage of GLMM is unavailable to investigate the correlation between variances of random effects. Besides Bayesian framework, double hierarchical generalized linear model (DHGLM) is used to handle those problems.

Hierarchical generalized linear models (HGLM), which is based on h-likelihood, is a synthesis of two widely-used existing model classed: GLMs and normal LMMs. Lee and Nelder (1996) extended GLMMs to HGLMs, in which the distributions of random effects are extended to conjugates of arbitrary distributions from the GLM family.

HGLMs are defined as follows:

(i) Conditional on random effect u, the response y follows a GLM family, satisfying

$$E(y|u) = \mu \text{ and } var(y|u) = \phi V(\mu),$$

[2] cM: centimorgan. It is a unit in defining genetic distance of molecular markers.

for which the kernel of the log-likelihood is given by

$$\sum\{y\theta - b(\theta)\}/\phi\ ,$$

where $\theta = \theta(\mu)$ is the canonical parameter. The linear predictor takes the form

$$\eta = g(\mu) = X\beta + Zu,$$

where β is the fixed effect, and u is the random effect.

(ii) The random effect u follows a distribution conjugate to a GLM family. That is the improvement compared to LMMs, which allow u follow normal distributions.

DHGLM is further extended from HGLM by allowing additional random effects in variance component. Lee and Nelder (2006) introduced a class of double HGLM (DHGLM) in which random effects can be specified in both the mean and the dispersion components. That is to say, we release the assumption that each random effect has the same variance, and further model variance component by including random effects also. The more complex structural DHGLM provide a new method for QTL analysis. Since different contributions of molecular markers to the total phenotypic variance give the proof of possible locations for QTL, estimating different markers' variance and then checking which markers contribute most is the key task to do. In DHGLM, we treat each marker effect as random effect in the mean model (the first level model), and model their variances in the dispersion model (the second level model), which also contains random effects. Similar to the BayesA method (Meuwissen et al. 2001) used in genomic selection, we model the data on two levels, but the estimation method for DHGLM is an iterative regression method that does not require Bayesian methodology.

For the barley data, the sample is N=145 DH lines, and the response y is the phenotype of kernel weight of each line. A total of j=127 markers are the random effects.

(i) The first level model is:

$$y = X\beta + Zu + e, \quad (1)$$

where β is an intercept, presents environmental effect in this case, u follows

$u_j \sim N(0, \lambda_j)$ for marker j, and residuals $e_i \sim N(0, \sigma^2)$ for observation i.

As mentioned in the material section, barley is double haploid and thus the coding for each marker is +1/-1, 0 for missing genotype. Hence, design matrix, Z, is a matrix of which the dimension is 145×127 and the elements are +1, -1, 0.

(ii) The second level model is:

The variance of marker effect, λ_j, modeled as

$$log(\lambda_j) = \alpha + b_j$$

with an intercept α, and where b_j is random effect in the dispersion model, which has a multivariate normal distribution, with an autocorrelation structure:

$$\begin{pmatrix} b_1 \\ b_2 \\ \mathbb{M} \\ b_{127} \end{pmatrix} \sim \left(0, \quad s_b^2 \begin{pmatrix} 1 & r & r^2 & \mathbf{K} \\ r & 1 & & \\ r^2 & & \mathbf{O} & \\ \mathbb{M} & & & 1 \end{pmatrix} \right), \quad 0 \le r \le 1$$

The correlation between b_k and b_l is $\rho^{|k-l|}$.

When $\rho = 1$, the model is linear mixed model for λ_j is constant for all j. The models for $\rho = 0$ and $0 < \rho < 1$ are referred as DHGLM and smoothed DHGLM, respectively. The chosen value of ρ depends on the density of markers and the level of linkage.

The overall phenotypic variance (of the current population) is expressed as

$$\sigma_y^2 = \sigma^2 + \sum_{j=1}^{127} u_j^2 \sigma_{Z_j}^2$$

where $\sigma_{Z_j}^2$ is the variance of j's column of design matrix Z, which presents variance of Z_j across individuals.

The proportion of phenotypic variance explained by a particular QTL effect is expressed by

$$h_j^2 \approx \sigma_{Z_j}^2 \frac{u_j^2}{\sigma_y^2}$$

The value of h_j^2 has the meaning of heritability, which presents the status of a

specific marker in heritage.

2.3 Algorithm

In estimating all the parameters$(\beta, u_j, \sigma^2, \lambda_j, \sigma_b^2)$, we construct weighted linear models and use iterative weighted least squares (IWLS). The algorithm process is as follows:

1° Initiate $\sigma^2, \lambda_{1,} \ldots, \lambda_q$, and σ_b^2.

2° Fit a linear model with weights

$$\begin{pmatrix} I_N \frac{1}{s^2} & & 0 \\ & \frac{1}{l_1} & \\ 0 & & \mathbf{O} \\ & & \frac{1}{l_q} \end{pmatrix} \text{ on } \begin{pmatrix} y \\ 0_q \end{pmatrix} \sim \begin{pmatrix} X & Z \\ 0 & I_q \end{pmatrix} \begin{pmatrix} b \\ u \end{pmatrix}, \text{ here } q = 127$$

This model is an augmented model in the first level of DHGLM.

3° Let $d_1, \ldots d_{N+q}$ be the squared residuals, which follows χ^2 distribution, and $h_1, \ldots, h_{N+q}$ the hatvalues from the model in 2°.

4° Fit a Gamma GLM with log link and weights

$$\begin{pmatrix} \frac{1-h_1}{2} & & 0 \\ & \mathbf{O} & \\ 0 & & \frac{1-h_N}{2} \end{pmatrix} \text{ on } \begin{pmatrix} \frac{d_1}{1-h_1} \\ \mathbb{M} \\ \frac{d_N}{1-h_N} \end{pmatrix} \sim 1_N, \text{ where } 1_N \text{ is a vector of 1.}$$

From this step, we aim to get estimation of σ^2, which is the intercept term in the model.

5° Calculate working response $z_j = log(\lambda_j) + \frac{1}{\lambda_j}\left(\frac{d_{N+j}}{1-h_{N+j}} - \lambda_j\right), j = 1, \ldots, q$. The formula for these responses comes from IWLS, because $\frac{d_{N+j}}{1-h_{N+j}}$ follows a Gamma distribution with mean λ_j, and we use a log-link. Then, after the transformation above, z_j approximately follows a normal distribution. Fit a linear model with weights

$$\begin{pmatrix} \frac{1-h_{N+1}}{2} & & 0 & \\ & \mathbf{O} & & \\ 0 & & \frac{1-h_{N+q}}{2} & \\ & & & I_q \frac{1}{s_b^2} \end{pmatrix} \text{ on } \begin{pmatrix} z \\ 0_q \end{pmatrix} \sim \begin{pmatrix} 1_q & L \\ 0 & I_q \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

And this is the augmented model of second level in DHGLM.

6° Let $\tilde{d}_1, \ldots, \tilde{d}_{2q}$ be the squared residuals, and $\tilde{h}_1, \ldots, \tilde{h}_{2q}$ the hatvalues from the model in 5°.

7° Fit a Gamma GLM with log link and weights

$$\begin{pmatrix} \frac{1-\tilde{h}_{q+1}}{2} & & 0 & \\ & \mathbf{O} & & \\ 0 & & \frac{1-\tilde{h}_{2q}}{2} & \end{pmatrix} \text{ on } \begin{pmatrix} \frac{\tilde{d}_{q+1}}{1-\tilde{h}_{q+1}} \\ \vdots \\ \frac{\tilde{d}_{2q}}{1-\tilde{h}_{2q}} \end{pmatrix} \sim 1_q \text{, where } 1_q \text{ is a vector of 1.}$$

Form this step, we get the estimation of σ_b^2.

8° Update σ^2 from 4°, $\lambda_1, \ldots, \lambda_q$ from 5°, and σ_b^2 form 7°.

9° Iterate between $2^\circ - 8^\circ$ until convergence.

Then the estimate of random effect by IWLS is best linear unbiased predictor, denoted by BLUP.

The R code of function *dhglm*() is in Appendix and was written by Rönnegård and Shen (2009) on the simulated data from 2009 QTLMAS workshop.

3. Results and discussions

3.1 Results of E-BAYES method

As we need to compare the results of DHGLM to that of E-BAYES, which was presented in Xu' paper (2007), first we briefly show Xu's results of Barley data.

The model of E-BAYES (excluding epistatic effects) is the same as Model(1), which assumes marker effect as random effect. Assign a normal prior to the marker effect, that is $u_j \sim N(0, \lambda_j)$. And also assign $\sigma_j^2 \sim Inv-\chi^2(\tau, \omega)$ as the prior for variance component σ_j^2, where τ is the degree of freedom and $\omega > 0$ is the scale parameter.

The E-BAYES method with $\tau = -1,\ \omega = 0.0005$, and $\tau = -2,\ \omega = 0$, denoted by E-BAYES(-1, 0.0005) and E-BAYES(-2, 0) were used for the barley data analysis. Six positions close to markers nr: 2, 12, 37, 43, 75, 102, indicated QTL controlling the trait of kernel weight. According to E-BAYES(-1, 0.0005), the six markers explained 62% of the total variation for the trait. The corresponding proportion for method E-BAYES(-2, 0) is 59.75%. The details of estimated parameters of these two models are shown in Table 1.

3.2 Results of DHGLM and Smoothed DHGLM

Now we come to the results of the DHGLM methods. The IWLS algorithm took about less than one minute completing the whole process. That is much faster than MCMC of E-BAYES. The BLUP of DHGLM and Smoothed DHGLM are both close to 0 at positions with no QTL effects. (Figure 1), and give estimated marker-effects with larger magnitude at QTL positions. As the real positions of QTL are unknown, we take E-BAYES method as reference.

The possible positions found by DHGLM seems promising with E-BAYES, as we can see in Figure 1, the BLUP of marker nr: 12, 43, 102 are significantly lager than other positions. Even not exactly the same positions as Xu's results, the BLUP near the positions, denoted by triangle mark, also have larger absolute values than other

positions. Smoothed DHGLM gives more concentrated pattern of BLUP, hence the values of BLUP get smaller than non-smoothed DHGLM. It is expected that such changes occurred, for the aim to smooth is to decrease differences between neighbor markers, by considering the correlation of them.

Actually, near QTL positions, there is usually more than one point having large magnitude. That is the reason why we only conclude QTL locates near such markers, not exactly at these markers. Since the sign of $\hat{u}_j$ varies at possible QTL positions, it is not easy to point out the marker. The pattern of the estimated variance of random effect shows a clearer picture (Figure 2).

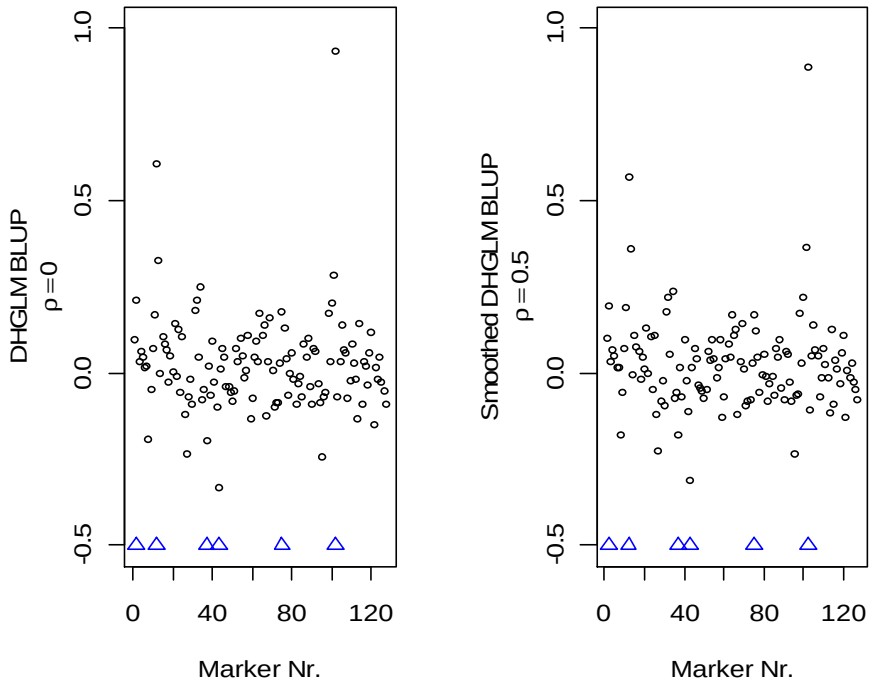

Figure 1: Marker-effect BLUP for the two models: DHGLM and Smoothed DHGLM. Barley data with possible QTL positions (from E-BAYES method) indicated by △.

As assumed, possible QTL positions locate near markers, which have large random effects and contributions to total phenotypic variance. That is to say, the markers with large values of $\hat{u}_j$ and $\hat{\lambda}_j$ may indicate QTL positions. The pattern of the estimated variance of random effect gives evidence more clearly in Figure 2. A big peak,

which presents a large value of $\hat{\lambda}_j$, appears exactly at possible QTL positions, denoted by triangle marks. Besides peaks at denoted positions, several small peaks, which locate between marker 25 and 43, may also indicate QTL positions. The height of peaks gives a visible image of each marker's variance and the quantity of contribution to total variance.

Smoothed DHGLM gives broader peaks than DHGLM, and reduces the variances between QTL positions. As we can see in Figure 2, the dot-slashed line for $\rho = 0.5$ shows smaller variances at non-QTL positions than the solid line for $\rho = 0$. Since smoothing considers the correlation between neighbor markers, consequently it reduces the difference between them.

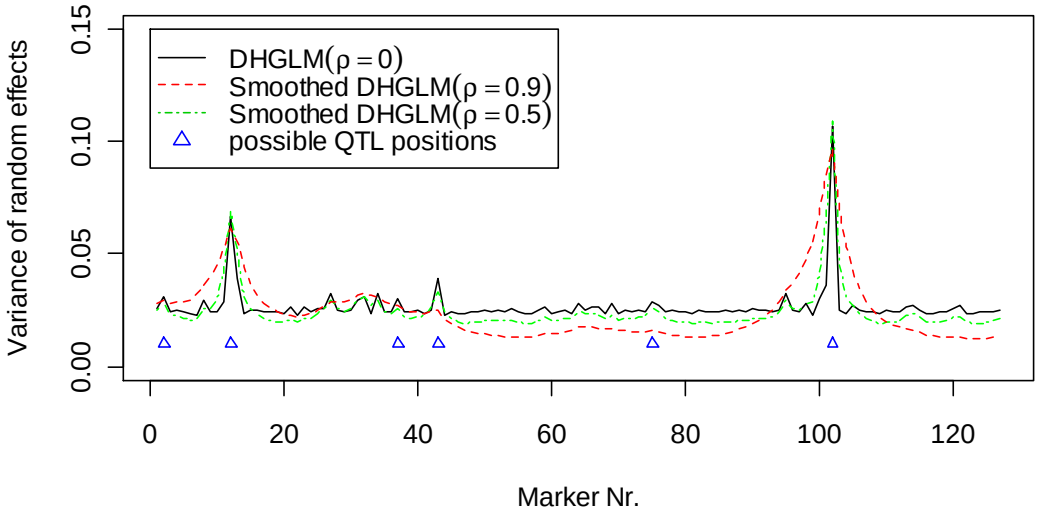

Figure 2: Estimated variance of random marker-effects for Barley data.

By comparing the curves of $\rho = 0.5$ and $\rho = 0.9$, with non-smoothing one $\rho = 0$, the slashed line for $\rho = 0.9$ seems over-smoothed, for the peaks are quite broad and small peak even disappear. One principle of choosing ρ is the density of markers; if markers are dense, it is better to set ρ close to 1, otherwise to 0. Effects of dense markers will be more influenced by the effects of surrounding neighbor markers, than those of spread markers. In addition, the variances will be more correlated. When $\rho = 1$, the model becomes LMM, for the variance of each marker is

identical. Here in our case, k=127 mapped markers covered a genome of 1500 cM (Xu, 2007). The density was not high, thus choosing ρ = 0.5 is better than ρ = 0.9 to describe the correlation between markers.

We have compared the QTL positions found by DHGLM and Smoothed DHGLM to E-BAYES, and correspond very well to each other. Now we calculate $\hat{u}_j$ and $\hat{h}_j^2$, and compare them to which of E-BAYES (Table 1).

Table 1: Comparing estimated parameters of DHGLM and Smoothed DHGLM, with E-BAYES on Barley dataset. Estimated marker effects ($\hat{u}_j$), and respective contributions $(\hat{h}_j^2)$ to total variance ($\hat{\sigma}_y^2$). $\hat{\beta}$ is fixed effect for environmental effect, $\hat{\sigma}^2$ is residual variance, $\hat{h}^2$ is the sum of six markers' contributions, and $*\hat{h}^2$ is total contribution of markers including new candidates: 8, 27, 34, 95.

	DHGLM	**Smoothed DHGLM**	**E-BAYES**	
	(ρ = 0)	(ρ = 0.5)	(-1, 0.0005)	(-2, 0)
Marker	$\hat{u}_j(\hat{h}_j^2)$	$\hat{u}_j(\hat{h}_j^2)$	$\hat{u}_j(\hat{h}_j^2)$	$\hat{u}_j(\hat{h}_j^2)$
2	0.2098(0.0135)	0.1964(0.0124)	0.5101(0.0526)	0.5657(0.0647)
12	0.6081(0.1134)	0.5667(0.1031)	0.8436(0.1438)	0.8499(0.1460)
37	-0.1953(0.0118)	-0.1776(0.0102)	--	-0.3756(0.0285)
43	-0.3313(0.0323)	-0.3101(0.0296)	--	-0.3801(0.0292)
75	0.1757(0.0095)	0.1685(0.0091)	0.4317(0.0376)	0.3587(0.0260)
102	0.9342(0.2668)	0.8859(0.2511)	1.3849(0.3877)	1.2244(0.3031)
8	-0.1925(0.0113)	-0.1812(0.0105)	--	--
27	-0.2350(0.0157)	-0.2243(0.0150)	--	--
34	0.2483(0.0177)	0.2349(0.0166)	--	--
95	-0.2412(0.0178)	-0.2328(0.0173)	--	--
$\hat{\beta}$	42.4049	42.4100	42.6448	42.4952
$\hat{\sigma}^2$	0.6590	0.6717	0.0614	0.0002
$\hat{\sigma}_y^2$	3.2201	3.0774	--	--
$\hat{h}^2$	0.4474	0.4156	0.6218	0.5975
$*\hat{h}^2$	0.5099	0.4750		

The fixed effect, $\hat{\beta}$, which presents for environmental effect, does not differ a lot from of E-BAYES; that indicates DHGLMs fit well as E-BAYES. Each marker-effect and its contribution to total variance in DHGLM and Smoothed DHGLM are both smaller than E-BAYES (-1, 0.0005) and E-BAYES (-2, 0), as both $\hat{u}_j$ and $\hat{h}_j^2$ of

DHGLMs are smaller than of E-BAYES. One reason is that in DHGLM, we don't consider the epistatic effect of markers and assume marker-effect is independent from each other; while, in E-BAYES models, epistatic effect was included. Although the values of marker-effects and their variance components of DHGLM are not exactly the same as those of E-BAYES, that does not affect us searching for QTL positions. In E-BAYES results, epistatic effect was not mentioned at all, which indicated that epistatic effect in Barley data was not quite important. The total contribution of DHGLM is about 45%, and Smoothed DHGLM is 42%. Although the proportions are lower than E-BAYES, one thing should be concerned is that some possible QTL positions are not included in our comparison.

As we have noticed in Figure 2, several small peaks appear in non-marked places. We suppose those peaks may also indicate QTL positions. If we take the minimum of the six markers, which is No.75, as the baseline to search for candidate markers, four new markers: 8, 27, 34, 95 have larger values of $\hat{u}_j$ and $\hat{\lambda}_j$ than 75 in DHGLM case, and more new comers in Smoothed DHGLM. Thus, the new $\hat{h}^2$ gets larger when we include those four new markers' contributions.

The residual variance, $\hat{\sigma}^2$, of DHGLMs is 10 times larger than of E-BAYES (-1, 0.0005). That is not quite understandable that the residual variances of E-BAYES were so small: $\hat{\sigma}^2 = 0.0614$ for E-BAYES (-1, 0.0005), and $\hat{\sigma}^2 = 0.0002$ for E-BAYES (-2, 0).

Whether to do smoothing or not depends on the maker density and amount of linkage in the population. In the Barley dataset, the density of markers was not so high, and thus we choose a moderate value of ρ. Since there is not a general criterion of converting genetic relationship into statistical modeling, our choice of ρ is an experimental product, giving no proof.

4. Conclusion

From the comparisons and the analysis above, the accuracy of DHGLM and Smoothed DHLGM in QTL analysis is as efficient as E-BAYES. All QTL positions found by E-BAYES were searched out, and moreover, new candidate markers implying possible QTL positions appeared. Whether DHGLM can be assumed to be more efficient than E-BAYES, needs further investigation, since we do not know the criterion of marker selecting in E-BAYES on the Barley dataset.

DHGLM also take advantage of higher speed of computation than Bayesian method, and do not require prior distributions of parameters. A single algorithm, IWLS used by DHGLM is suitable for all new models. Hence, DHGLM is more flexible to handle sophisticated models and a wide class of distributions.

As genetics is evolving at a rapid pace with immense marker datasets appearing, the advantages of DHGLM mentioned above will be valuable for analyzing genetic data. We argue that DHGLM have a great potential for future use in genetics, as well as in other plant and animal breeding applications.

REFERENCES

Rönnegård, L., Lee, Y., 2009 Hierarchical generalized linear models have a great potential in genetics and animal breeding. *Proc. WCGALP ,Leipzig, Germany.*

Xu, S., 2007 An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63:** 513-521.

Yi, N., Banerjee, S. 2009 Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics* **181:** 1103-1113.

Lee, Y., Nelder, J.A. 2006 Double hierarchical generalized linear models. *Appl. Statist.* **55:** 139-185

Lee, Y., Nelder, J.A., Pawitan, Y. 2006 Generalized linear models with random effects. *Chapman $ Hall/CRC.*

Tinker, N.A., et al. 1996 Regions of the genome that affect agronomic performance in two-row barley. *Crop. Sci.* **36:** 1053-1062.

Wu, R., Ma, C., Casella, G. 2007 Statistical genetics of quantitative traits. *Springer.*

Shen, X., Rönnegård, L., Carlborg, Ö. 2010 Hierarchical likelihood opens a new way of estimating genetic values using genome-wide dense marker maps. *Proc. QTLMAS, Poznan, Poland (submitted).*

APPENDIX

```
#### A function of normal-normal-normal DHGLM ####
dhglm <- function(y, X, Z, X.disp = NULL, ac.mat = NULL, rho = 0, conv.crit = 1e-5,
max.iter = 200) {
## // rho = .9 is the default value for smoothed DHGLM
## // rho cannot be 1, otherwise chol() would not work on ac.mat
## PartI.SimulateData ##
rm(list=ls())
require(hglm)
N <- nrow(X)
p <- ncol(X)
k <- ncol(Z)
if (is.null(X.disp)) {
      X.disp <- rep(1, k)
      pd <- 1
} else {
      pd <- ncol(X.disp)
}
simulering <- FALSE

## autocorrelation matrix of the variance of the random effects
if (is.null(ac.mat)) ac.mat <- rho**(toeplitz(1:k) - 1)
L <- t(chol(ac.mat))
cat("chol(ac.mat) ready.\n")

if (simulering) {
      mu=mean(y)
      sigma2e<-var(y-mu)
      #Step function
      a.mark<-c(rep(0,40),rnorm(10,1,0),rep(0,40))
      e<-rnorm(1000,0,sqrt(sigma2e))
      y<-mu+Z%*%a.mark+e
      QTL.model<-hglm(y=y,X=X,Z=Z)
      plot((QTL.model$ranef),ylim=c(-2,3),xlab="Marker Nr.", ylab="Normal
BLUP")
}

## Part II: Implementation of GLM ##
##Step-I:Create Augmented Data Frame ##
Y <- c(y, rep(0, k))
nollor <- matrix(0, k, p)
Aug.Data <- data.frame(rbind(cbind(X, Z), cbind(nollor, diag(k))))
```

```
## Initialize the parameter values ##
b0 <- log(1)
g0 <- log(2)
sigma_b <- 1
phi <- exp(g0)
lamda <- rep(exp(b0),k)
CONV <- FALSE
niter <- 1

#### Start IWLS
while (!CONV & niter < max.iter) {
## Estimate the mean fixed and random effects ##
lm1 <- lm(Y ~ . - 1, data = Aug.Data, weights = c(rep(1/phi, N), 1/lamda))
q <- hatvalues(lm1)
dev <- (residuals(lm1))^2
dev_q <- dev/(1 - q)
beta<-lm1$coef[1]

## Estimate residual variance ##
gl1 <- glm(as.numeric(dev_q[1:N]) ~ 1, family = Gamma(link = log), weights = as.numeric((1 - q[1:N])/2))
phi0 <- as.numeric(exp(coef(gl1)))
## Estimate latent variable fixed and random effect ##
z <- c(as.numeric(dev_q[(N + 1):(N + k)]), rep(0, k))
z[1:k] <- log(lamda) + (z[1:k] - lamda)/lamda
nollor2 <- matrix(0, k, pd)
lm2 <- lm(z ~ . - 1, data = Aug.Data2, weights = c(as.numeric(1 - q[(N + 1):(N + k)])/2, 1/rep(sigma_b, k)))
q2 <- hatvalues(lm2)
dev2 <- (residuals(lm2))^2
dev2_q <- dev2/(1 - q2)

## Estimate latent variable random effect variance ##
gl2.2 <- glm(as.numeric(dev2_q[(k + 1):(k + k)]) ~ 1, family = Gamma(link = log), weights = as.numeric((1 - q2[(k + 1):(k + k)])/2))
sigma_b0 <- as.numeric(exp(coef(gl2.2)))
CONV <- max(abs(c(sigma_b, phi) - c(sigma_b0, phi0))) < conv.crit
cat("convergence:", max(abs(c(sigma_b, phi) - c(sigma_b0, phi0))), "\n")
niter <- niter + 1
sigma_b <- sigma_b0
phi <- phi0
lamda <- exp(lm2$fitted.values[1:k])
result <- list( beta = beta, phi = phi, sigma2_b = sigma_b, lambda = lamda, BLUP = coef(lm1)[(p + 1):(p + k)], niter = niter)
```

```
return(result)

## end of function dhglm()
}

#### Reading Barley dataset and apply dhglm() ####
rm(list=ls(all=T))
barGen<-as.matrix(read.table(file="gen_barley.csv",sep=","))
n<-145    #individual number
k<-127    #marker number
Z<-barGen
y<-read.table(file="phe_barley.csv",sep=",")
y <- as.numeric(as.matrix(y))
X<-matrix(1,n,1)

## calculate var(Z)
Z1<-matrix(0,1,k)
for(i in 1:k){
      Z1[1,i]<-var(Z[,i])
}
Z1

## REMEMBER to change rho's value in function dhglm()
## rho=0
m_1<-dhglm(y,X,Z)
names(m_1)
m_1$beta    #fixed effect
marker_variance1<-m_1$lambda
BLUP1<-m_1$BLUP
BLUP1_new<-(1:127)[abs(BLUP1)>=min(abs(BLUP1[c(2,12,37,43,75,102)]))]
BLUP1[BLUP1_new]

resVar1<-m_1$phi    ##residual variance sigma^2
sigma2_b1<-m_1$sigma2_b      ##variance of the second level, sigma^2 of b
y1_var<-resVar1+sum(BLUP1^2*Z1)

## calculate values of    h-square
h1_square<-Z1*(BLUP1^2)/y1_var
h1<-as.matrix(h1_square,1,127)
h1QTL<-h1[1,c(2,12,37,43,75,102)]
h1QTL_new<-h1_square[BLUP1_new]
sum(h1QTL_new)
sum(h1QTL)
```

```
## rho=0.9(0.5)
m_2<-dhglm(y,X,Z)
names(m_2)
marker_variance2<-m_2$lambda
BLUP2<-m_2$BLUP
resVar2<-m_2$phi
sigma2_b2<-m_2$sigma2_b
y2_var<-resVar2+sum(BLUP2^2*Z1)
h2_square<-Z1*(BLUP2^2)/y2_var
h2<-as.matrix(h2_square,1,127)
h2QTL<-h2[1,c(2,12,37,43,75,102)]
sum(h2QTL)


## MAP variance (Figure 2)
windows()
plot(marker_variance1,ylim=c(0,0.15),xlab="Marker Nr.",ylab="Variance of random effects",type="l",lty=1)
points(x=c(2,12,37,43,75,102), y=rep(0.01,6),pch=2,cex=0.7,col="blue")
points(marker_variance2,col="green",lty=4,type="l")
t1<-expression("DHGLM" (rho==0))
t2<-expression("Smoothed DHGLM" (rho==0.9))
t3<-expression("Smoothed DHGLM" (rho==0.5))
utils::str(legend(0,0.15,c(t1,t2,t3,"possibleQTLpositions"),pch=c(-1,-1,-1,2),lty=c(1,2,4,-1),col=c(1,2,3,4)))
## plot BLUP (Figure 1) ##
windows()
par(mfrow=c(1,2))
test<-paste("DHGLM BLUP,",expression(rho==0))
plot(BLUP1,ylim=c(-0.5,1),xlab="MarkerNr.",ylab="DHGLMBLUP",pch=1,cex=0.7)
points(x=c(2,12,37,43,75,102), y=rep(-0.5,6),pch=2,col="blue")
mtext(expression(rho==0),side=2,line=2,cex=1)
plot(BLUP2,ylim=c(-0.5,1),xlab="MarkerNr.",ylab="SmoothedDHGLMBLUP",pch=1,cex=0.7)
points(x=c(2,12,37,43,75,102), y=rep(-0.5,6),pch=2,col="blue")
mtext(expression(rho==0.5),side=2,line=2,cex=1)
```