Hierarchical generalized linear models have a great potential in genetics and animal breeding

L. Rönnegård[*] and Y. Lee[†]

Introduction

Genetics applied to livestock production is evolving at a rapid pace with marker data sets increasing to immense proportions. More, and different types of, data is expected as whole genome sequencing is likely to be applied widely within the near future. This requires new statistical modeling tools and inferences that are flexible to different kinds of data and are also fast enough to enable the analyses of large data. We expect that hierarchical generalized linear models (HGLMs) based on the h-likelihood will be such a tool.

HGLMs consist of the three objects, namely the data, fixed unknown constants (parameters) and unobserved random variables (unobservables). Traditional Bayesian models consist of the two objects, the data and unobservables, while frequentist's (or Fisher's) models consist of the data and parameters. By allowing all three objects in the statistical modeling it is possible to describe various features in the data, for example, within-subject correlation in longitudinal studies, smooth spatial and temporal trends, function fittings, and factor analysis, heteroskedasticity, heavy-tailed distributions, robust modelings and sparse variable selections. In the statistical literature unobservables appear with various names such as random effects, latent processes, factor, missing data, unobserved future observations, potential outcomes etc. Handling of such unobservables is the key to new extended likelihood inferences. Lee and Nelder (1996, 2006) and Lee et al. (2006) have shown how to model and make inferences using the h-likelihood. Inferences about unobservables can be made without resorting to an empirical Bayes framework (Lee and Nelder, 2010). A single algorithm, iterative weighted least squares, can be used throughout all new models and requires neither prior distributions of parameters nor multi-dimensional quadrature. The h-likelihood plays a key role in the synthesis of the computational algorithms needed for broad class of new models.

Examples of HGLM applied in genetics. Jaffrezic et al. (2000) used an HGLM for analysis of lactation curves with heterogeneous residual variances over time. Noh et al (2005) modeled heavy tailed distributions for random effects to take ascertainment into account in human QTL studies. Noh et al. (2006a) used HGLM to minimize bias in heritability estimation for binary traits in human family data. HGLM has also been successfully applied in survival analysis with random effects (Noh et al. 2006b). 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 (see also Felleki and Chalkias

[*] Dalarna University, SE-78170 Borlänge, Sweden

[†] Seoul National University, NS40, San56-1, Shin Lim-Dong, Kwan Ak-Ku, Seoul 151-747, Korea

(2010)). Rönnegård and Valdar (2010) have also suggested the use of DHGLM to detect variance-controlling QTL.

Possible future applications of HGLM in genetics and animal breeding. Variance component estimation plays a central role in animal breeding and is expected to do so in the future. We presume that HGLM will be used for improving variance component estimation in binary traits, count traits and survival traits; where severely biased variance component estimates can be produced by current standard software for generalized linear mixed models. The HGLM specification also enables joint modeling of multiple trait having different distributions.

The possible applications for DHGLM in studies of genetic heterogeneity in residual variance will be continued to be investigated. Furthermore, not only does DHGLM enable modeling of random effects in the residual variance, but also random effects in the variance for the random effects. Such models have previously been studied extensively in e.g. genomic selection using Bayesian methodology (Meuwissen et al. 2001), whereas the DHGLM approach may give new computational possibilities to fit models with a large number of markers.

Aim. We argue that (D)HGLM modeling has a great potential for future use in genetics and animal breeding. To illustrate this, we present results from an association analysis where we apply a DHGLM on simulated data from the 2009 QTLMAS workshop.

Material and methods

Association analysis using DHGLM. We perform an association analysis on the simulated data from the 2009 QTLMAS workshop (Coster et al. (2010)) using a DHGLM. 1000 individuals with phenotypes and 453 SNP-markers on five 1 Morgan long chromosomes were simulated for a growth trait. We analyzed the phenotype observed on the third occasion at 265 days of age. Six QTL controlling this trait were simulated close to markers nr: 36, 98, 174, 276, 337, 432, where the one close to marker nr 36 explained more than 50% of the genetic variation and the other QTL explained ~10% each (with the QTL close to marker nr 432 having smallest effect). 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. The SNP-effects u were modeled as random with $u_j \sim N(0,\lambda_j)$ for marker j, with phenotypes y as: $y = \mu + Zu + e$, with intercept μ and residuals $e_i \sim N(0,\sigma^2)$ for observation i. The variance of the SNP-effects, λ_j, was modeled as $\log(\lambda_j) = \alpha + b_j$ with intercept α and where b_j is the j:th element in a multivariate normal random effect with an autocorrelation structure. Hence, the correlation between b_k and b_l was $\rho^{|k-l|}$. For $\rho = 1$, λ_j is constant for all j and we have a linear mixed model. The models for $\rho = 0$ and $\rho = 0.9$ are referred to as *DHGLM* and *smoothed DHGLM*, respectively. As opposed to Bayesian methods, the HGLM approach does not require specification of priors and the estimation is performed using iterative weighted least squares. Hence, a simple regression method is used iteratively to obtain the estimates of u_j, λ_j, α and b_j.

Results and discussion

The DHGLM BLUP, compared to Normal BLUP, were closer to 0 at positions with no QTL effects (Figure 1), and give estimated SNP-effects with larger magnitude at QTL positions. Most interesting is the pattern of the estimated variance at different chromosomal positions (Figure 2), with peaks corresponding very well to the simulated QTL locations. Smoothed DHGLM gives broader peaks but reduces the variances down to 0 between QTL positions. For $\rho = 0$ the model is very similar to BayesA, whereas the computations are much faster. Whether the DHGLM or Smoothed DHGLM model is the one to prefer on real data probably depends on the marker density, amount of linkage disequilibrium in the population, distance between QTL and the distribution of the QTL effects.

Conclusion

We have argued that HGLM and DHGLM have a great potential in animal breeding and genetics due to its computational speed and modeling flexibility. A DHGLM was used to locate five out of six QTL (simulated in the 2009 QTLMAS data set) by means of a novel method for association analysis, where smoothing was performed on the level of the variances of the SNP-effects. We find the results from this analysis promising for future applications in genetics. Moreover, DHGLM can be fitted for large data sets (Rönnegård et al. (2010)), which enables fast whole genome association analysis with dense SNP-markers.

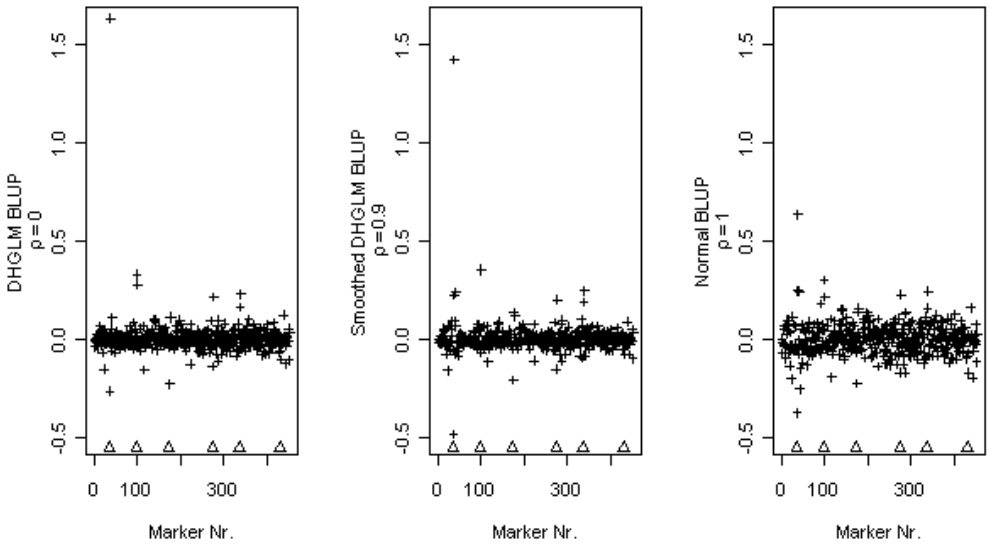

Figure 1: SNP-effect BLUP for the models: DHGLM, Smoothed DHGLM and a linear mixed model with constant variance. Simulated data from QTLMAS '09 with QTL positions indicated by Δ.

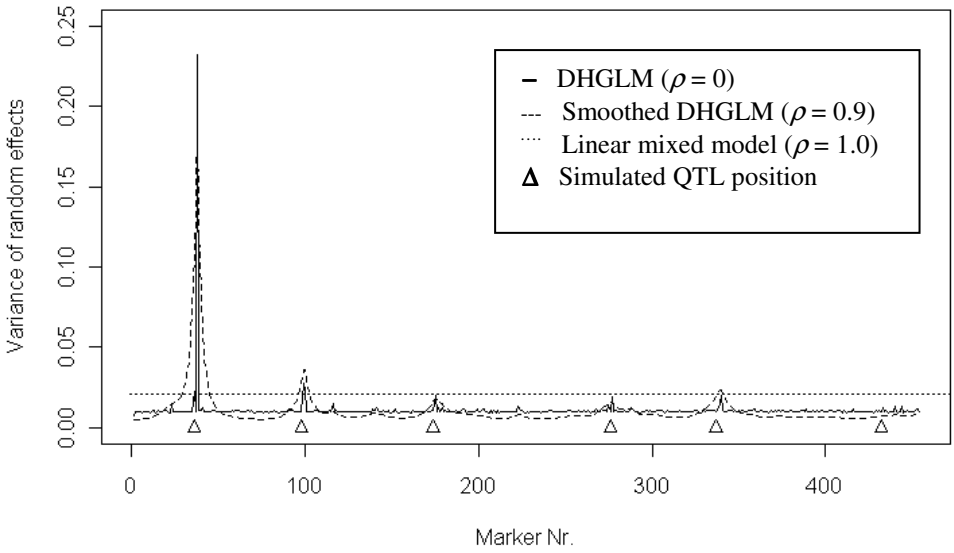

Figure 2: Estimated variance of random SNP-effects for QTLMAS '09 workshop data.

References

Coster, A. et al. (2010) QTLMAS 2009: Simulated data *submitted*

Felleki, M. and Chalkias, H. (2010) *Proc. WCGALP, Leipzig, Germany* (submitted)

Jaffrezic, F., White, I.M.S., Thompson, R., Hill, W.G. (2000) *J Dairy Sci,* 83:1089-1093

Lee, Y., Nelder, J. A. (1996) *J R Statist Soc B,* 58:619-678.

Lee, Y., Nelder, J. A. (2006) *App Stat,* 55:139-185.

Lee, Y., Nelder, J. A. (2010) *Stat Sci, to appear*.

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

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001) *Genetics,* 157:1819-1829

Noh, M., Lee, Y., Pawitan Y. (2005) *Genet Epidem,* 29:68-75.

Noh, M., Yip, B., Lee, Y., Pawitan Y. (2006a) *Genet Epidem,* 30:37-47.

Noh, M., Ha, I.D., Lee, Y. (2006b) *Statist Med,* 25:1341-1354

Rönnegård, L., Felleki, M., Fikse, F. et al. (2010) *submitted*

Rönnegård, L., Valdar, W. (2010) *submitted*