Genetic Heterogeneity of Environmental Variance - estimation of variance components using Double Hierarchical Generalized Linear Models

L. Rönnegård[*,a,b], M. Felleki [a,b], W.F. Fikse [b] and E. Strandberg [b]

EAAP, Barcelona, August 2009, Session 53

[a]Statistics Unit, Dalarna University, SE-781 70 Borlänge, Sweden

[b]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden

* Corresponding author: lrn@du.se

Abstract

Previous studies have shown that environmental sensitivity (i.e. the capability of an animal to adapt to changes in the environment) may be under genetic control, which is essential to take into account if we wish to breed robust farm animals. Linear mixed models including a genetic effect explaining heterogeneity of the environmental variance have previously been used and parameters estimated using EM and MCMC algorithms. We propose the use of double hierarchical generalized linear models (DHGLM), where the squared residuals are assumed to be gamma distributed and the residual variance is fitted using a generalized linear model (GLM). The algorithm iterates between two sets of mixed model equations (MME), one on the level of observations and one on the level of variances. The method was applied to data on pig litter size with 10,060 records from 4,149 sows. The DHGLM was implemented using PROC REG in SAS and the algorithm converged within 7 days on a Linux server. The estimates were similar to those previously obtained using Bayesian methodology except for the correlation between the ordinary animal effects and the animal effects included in the residual variance. The correlation for DHGLM was calculated from the estimated BLUP values whereas the same correlation was included as a parameter in the Bayesian model.

To test whether the DHGLM approach gives unbiased estimates for the genetic correlation we simulated 10,000 observations with 10 levels (representing 10 sires) in the random effect and 1,000 observations per level. For each animal, an observation was generated as the sum of a fixed effect (2 levels), a random genetic effect (u) and a random residual. The residual effect was sampled from N(0,phi), where log(phi) was generated as the sum of a fixed effect (2 levels) and a random genetic effect (g). Both genetic effects (u and g) were negatively correlated and sampled from a multivariate normal distribution. We replicated the simulation 20 times and obtained estimates of variance components using DHGLM. The estimated variance components and correlations seems to be unbiased.

In the future we intend to develop the DHGLM methodology to include the genetic correlation as a parameter in the model.

Introduction

In linear mixed models it is usually assumed that the residual variance is the same for all observations. There might, however, be differences in residual variance between individuals and there may also be known explanatory variables controlling these differences. If there are random genetic effects in the model controlling the residual variance, we refer to this as *genetic heterogeneity*.

Why is this an important issue in genetics? Modern animal breeding require animals that are robust to environmental changes. Moreover, if there is genetic heterogeneity then traditional methods for predicting selection effects may not be sufficient [8, 3].

Methods have previously been developed to estimate the degree of genetic heterogeneity. San Cristobal-Gaudy et al. [10] developed an EM-algorithm. Sorensen & Waagepetersen [11] applied a Markov chain Monte Carlo algorithm to estimate the parameters in a similar model, which had the advantage of producing model checking tools based on posterior predictive distributions and model selection criteria based on Bayes factor and deviances. Wolc et al. [12] used mixed model methodology with the residuals modelled as a gamma Generalized Linear Model (GLM).

Are there similar problems in other areas of research? Linear models where random effects are specified in the residual variance part of the model have long been applied on financial time-series data. Two examples are "stochastic volatility models" [2] and "generalized autoregressive conditional heteroscedasticity (GARCH) models" [1], where the residual variance depends on random temporal financial shocks. These models have been estimated using EM- and MCMC-algorithms.

HGLM and hierarchical likelihood Recently, however, Lee & Nelder [6] developed the framework of double hierarchical generalized linear models (DHGLM). The parameters are estimated by iterating between a hierarchy of generalized linear models (GLM), where each GLM is estimated by iterative weighted least squares. DHGLM give model checking tools based on GLM theory and model selection criteria are calculated from the hierarchical likelihood (h-likelihood). A user-friendly version of DHGLM has been implemented in the statistical software package Genstat. To our knowledge, DHGLM has previously only been applied on data with relatively few levels in the random effects (less than 100) whereas models in animal breeding applications usually have a large ($>>$100) number of levels in the random effects since each individual have a random genetic effect.

Inference in DHGLM is based on the h-likelihood theory developed by Lee & Nelder [5] and is a direct extension of the HGLM algorithm proposed in the same paper which is explained in detail in Lee, Nelder & Pawitan [7]. HGLMs have previously been applied in genetics (e.g. [4, 9]). A major advantage of these models is that the studied phenotypic trait may have any distribution belonging to the exponential family of distributions (e.g. normal, binomial, poisson, gamma). Also multiple-trait models with a combination of these distributions is possible. DHGLM is a natural and exciting extension of HGLM.

Aim The aim of this paper is to examine the potential use of DHGLM in animal breeding applications. We test the DHGLM methodology both on simulated data and on the field data previously analyzed by Sorensen & Waagepetersen [11].

Material and Methods

Linear mixed models and HGLM

Lee & Nelder [5] showed that linear mixed models can be fitted using a hierarchy of GLM by using an augmented linear model. The linear mixed model

$$y = Xb + Zu + e$$

$$V = ZZ^T\sigma_u^2 + \mathbf{I}\sigma_e^2$$

may be written as an augmented weighted linear model:

$$y_a = \mathbf{T}_a\delta + e_a \tag{1}$$

where:

$$y_a = \begin{pmatrix} y \\ \mathbf{0}_q \end{pmatrix}$$

$$\mathbf{T}_a = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}$$

$$\delta = \begin{pmatrix} b \\ u \end{pmatrix}$$

$$e_a = \begin{pmatrix} e \\ -u \end{pmatrix}$$

Here, q is the number of columns in $\mathbf{Z}$, $\mathbf{0}_q$ is a vector of zeros of length q, and $\mathbf{I}_q$ is the identity matrix of size $q \times q$. The variance-covariance matrix of the augmented residual vector is given by:

$$V(e_a) = \begin{pmatrix} \mathbf{I}\sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q\sigma_u^2 \end{pmatrix}$$

This weighted linear model gives the same estimates of the fixed and random effects (b and u respectively) as Henderson's mixed model equations.

The estimates from weighted least squares are given by:

$$\mathbf{T}_a^t W^{-1}\mathbf{T}_a\hat{\delta} = \mathbf{T}_a^t W^{-1} y_a$$

where $W \equiv V(e_a)$.

The two variance components are estimated iteratively by applying a gamma GLM to the residuals e_i^2 and u_i^2 with intercept terms included in the linear predictors. The leverages h_i for these models are calculated from the diagonal elements of the hat matrix:

$$\mathbf{H}_a = \mathbf{T}_a(\mathbf{T}_a^t W^{-1}\mathbf{T}_a)^{-1}\mathbf{T}_a^t W^{-1} \tag{2}$$

So far we have assumed that the random effects u_i are iid, but this does not restrict the model because a covariance structure of u may be included implicitly by modifying the incidence matrix Z [8]. If we have an animal model, for instance, with relationship matrix $\mathbf{A}$ then we can include this by premultiplying the incidence matrix Z with the choleski factorization of $\mathbf{A}$.

Double HGLM

By applying the augmented model approach of eq. 1 also to the dispersion part of the model we obtain a double HGLM (DHGLM). The model for the residual variance is given by:

$$log(\mu_d) = \mathbf{X}_{\mathrm{d}} b_d + Z_d u_d \tag{3}$$

where u_d is a random effect with $V(u_d) = \mathbf{I}\sigma_d^2$.

Re-writing this model as an augmented model with the augmented response vector d_a consisting of the deviances d from model 1 and augmented values ψ:

$$d_a = \begin{pmatrix} d \\ \psi \end{pmatrix}$$

$$E(d_a) = \mu_d^*$$

$$log(\mu_d^*) = \mathbf{T}_a^* \delta^* \tag{4}$$

where $\mathbf{T}_a^* = \begin{pmatrix} \mathbf{X}_d & \mathbf{Z}_d \\ 0 & \mathbf{I} \end{pmatrix}$ andψ is the (unconditional) expectation of the random effects.

Model 1 is used for modelling the mean part of the model, whereas the residual variance now depends on the linear predictor of the dispersion in eq. 4. Let Σ be a diagonal matrix having elements equal to the predicted values $exp(\mathbf{T}_a^*\hat{\delta}^*)$ and $V(u_d) = \mathbf{I}\sigma_d^2$. The vector of individual deviances d^* obtained from eq. 4 is subsequently used to estimate σ_d^2 by fitting a gamma GLM to the response $d_i^*/(1-h_i^*)$ where h_i^* are the corresponding leverages.

Algorithm overview

The fitting algorithm is implemented by:

1. Initialize Σ, $\sigma_u^2$and σ_d^2
2. Fit the model for the mean using eq. 1 (i.e. Henderson's mixed model equations) and calculate the leverages h_i for the augmented model.

3. Calculate the variance of the random effects in the mean model σ_u^2 by fitting a gamma GLM to the response $\hat{u}^2/(1-h_i)$.

4. Calculate the residual variances in Σ from the dispersion model 4. Calculate the corresponding leverages $h_i^*$and deviances d_i^*.

5. Calculate the variance of the random effects in the dispersion model σ_d^2 by fitting a gamma GLM to the response $d_i^*/(1-h_i^*)$.

6. Iterate steps 2-5 until conergence

We have described the algorithm for one random effect in the mean and dispersion parts of the model but extending the algorithm for several random effects is quite straight forward. We implemented the algorithm with two random effects using PROC REG in SAS. Hence, our implementation uses the augmented model approach with iterative least square fitting.

The described algorithm fits a double HGLM for a normally distributed trait y with normally distributed random effects u and u_d, whereas the general algorithm given by Lee and Nelder [7] allow a variety of distributions both for the outcome variable and the random effects.

Data and models for pig litter size

Pig litter size from 4149 sows were analyzed by Sorensen & Waagepetersen [11] and the data is described therein. The data includes 10060 records from the 4149 sows in 82 herds. Hence, there were repeated measurements on sows. The maximum number of parities was nine. The data included the following class variables: herd (82 classes), season (4 classes), type of insemination (2 classes), and parity (9 classes). The data is highly imbalanced with two herds having one observation and 13 herds with five observations or less. There were nine observations in the ninth parity.

Several models were analyzed by Sorensen & Waagepetersen [11] with an increasing level of complexity in the model for the residual variance and with the model for the mean $y = \mathbf{Xb}+\mathbf{Wp}+\mathbf{Za}+\mathbf{e}$ being unchanged. Here y is litter size (vector of length 10060), $\mathbf{b}$ is a vector including the fixed effects of herd, season, type of insemination and parity, and $\mathbf{X}$ is the corresponding design matrix (10060x94), $\mathbf{p}$ is the random permanent environmental effects (vector of length 4149),$\mathbf{W}$ is the corresponding incidence matrix (10060x4149) and $V(\mathbf{p}) = \mathbf{I}\sigma_p^2$, $\mathbf{a}$ is the additive genetic random effect, $\mathbf{Z}$ is the corresponding incidence matrix (10060x4149) and $V(\mathbf{a}) = \mathbf{A}\sigma_p^2$ where $\mathbf{A}$ is the additive relationship matrix. Hence the LHS of the mixed model equations is of size 8392x8392.

The residual variance $\mathbf{e}$ was modelled as follows.

Model I: Homogeneous variance

$$V(e_i) = exp(\tilde{b}_0)$$

where $\tilde{b}_0$ is a common parameter for all i.

Model II: Fixed effects in the linear predictor for the residual variance
In this model each parity and insemination type has its own value for the residual variance.

$$V(e_i) = exp(\tilde{\mathbf{x}}_i\tilde{\mathbf{b}})$$

where $\tilde{\mathbf{b}}$ is a parameter vector including effects of parity and type of insemination, and $\tilde{\mathbf{x}}_i$is the i: th row in the design matrix $\tilde{\mathbf{X}}$.

Model III: Fixed and random effects in the linear predictor for the residual variance

$$V(e_i) = exp(\tilde{\mathbf{x}}_i\tilde{\mathbf{b}}+\mathbf{w}_i\tilde{\mathbf{p}} + \mathbf{z}_i\tilde{\mathbf{a}})$$

where $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{a}}$ are random effects of peramanent environment and genetic additive values, respectively, and $\mathbf{w}_i$ and $\mathbf{z}_i$ are the i: th rows $\mathbf{W}$ and $\mathbf{Z}$. This is Model 4 in [11].

Results

Analysis of pig litter size data

The DHGLM estimates and Bayesian estimates (i.e. posterior mean estimates from [11]) were identical for the linear mixed model with homogeneous variance (Model I) and were very similar for Model II where fixed effects are included in the residual variance part of the model (Table 1). For Model III, with random effects in the residual variance part of the model, the DHGLM estimates deviated from the Bayesian estimates. This differences may be due to the fact that the genetic correlation ρ was not included as a parameter in the DHGLM approach. Alternatively, the difference could be caused by the fact that the Bayesian estimates are posterior distribution means and that the posterior distributions are skewed.

The data is unbalanced with few observations within some herds. This is reflected in the leverage plot (Figure 1) as some leverages are equal to 1.0. Although the data was quite unbalanced and the algorithm was not computationally optimized the algorithm converged within 7 days on a Linux server.

Table 1 Comparison between DHGLM and Bayesian (S&W 2003) estimates for three models

Model		σ_a^2	σ_p^2	$\tilde{b}_0$	$\tilde{\delta}_{ins}$	$\tilde{\delta}_{par}$	$\sigma_{\tilde{a}}^2$	$\sigma_{\tilde{p}}^2$	ρ
I	DHGLM	1.40	0.60	2.00					
	S&W 2003	1.40	0.60	2.00					
II	DHGLM	1.39	0.72	1.86	-0.15	0.32			
	S&W 2003	1.37	0.71	1.87	-0.15	0.34			
III	DHGLM	1.38	0.68	1.83	-0.16	0.32	0.02	0.006	-0.04*
	S&W 2003	1.62	0.60	1.77	-0.17	0.35	0.09	0.06	-0.62

$\tilde{b}_0$ is the mean in the model for the residual variance

$\tilde{\delta}_{ins}$ is the fixed effect of insemination (in the model for the residual variance)

$\tilde{\delta}_{par}$ is the fixed effect for the difference in first and second parity (in the model for the residual variance)

*Correlation between realised BLUP of a and $\tilde{a}$ weighted by their reliabilities

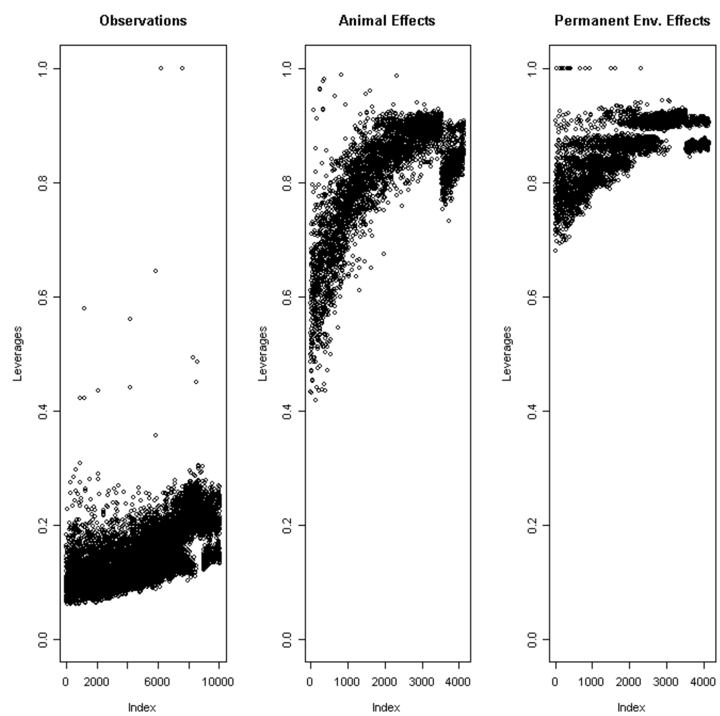

Figure 1 Leverages h_i for the mean part of the model. For the random (animal and permanent environmental) effects the reliabilities of the estimated BLUP are equivalent to $1 - h_i$.

Simulations

To test whether the DHGLM approach gives unbiased estimates for the genetic correlation we simulated 10,000 observations. The number of levels in the random effect was either 10 or 100, and the simulated genetic correlation was either 0 or -0.5. For each animal, an observation was generated as the sum of a fixed effect (2 levels), a random genetic effect (a) and a random residual. The residual effect was sampled from N(0,ϕ), where log(ϕ) was generated as the sum of a fixed effect (2 levels) and a random genetic effect ($\tilde{a}$). The genetic effects (a and $\tilde{a}$) were sampled from a multivariate normal distribution. We replicated the simulation 20 times and obtained estimates of variance components using DHGLM. The estimated variance components and correlations seems to be unbiased (Table 2).

Table 2 Estimated variance components in the model of the mean (σ_a^2) and the residual variance ($\sigma_{\tilde{a}}^2$) using DHGLM. The correlations (ρ) between the random effects were estimated retrospectively from the BLUP values. Mean (s.d.) of 20 replicates.

		Simulated values			Estimates		
No. clusters	Obs. per cluster	σ_a^2	$\sigma_{\tilde{a}}^2$	ρ	σ_a^2	$\sigma_{\tilde{a}}^2$	ρ
10	1000	1.0	0.5	0.0	1.06	0.44	-0.070
					(0.43)	(0.15)	(0.25)
100	100	1.0	0.5	0.0	0.99	0.51	-0.015
					(0.17)	(0.06)	(0.07)
10	1000	1.0	0.5	-0.5	1.00	0.48	-0.44
					(0.41)	(0.26)	(0.29)
100	100	1.0	0.5	-0.5	1.01	0.50	-0.42
					(0.17)	(0.07)	(0.07)

Discussion

We have shown that DHGLM is a feasible estimation algorithm for animal models. We implemented the algorithm using the simple regression algorithm PROC REG in SAS. The DHGLM algorithm iterates between weighted least squares and it should therefore be possible to develop a more computationallly efficient algorithm using standard numerical algorithms for least square problems. DHGLM estimation is available in the user-friendly environment of Genstat. We have been able to run DHGLM in Genstat for models with up to 5000 equations in the mixed model equations (results not shown). Hence, the Genstat version of DHGLM is suitable for sire models but not for animal models with a large number of individuals. A recently developed R package **hglm** is also available at www.larsronnegard.se, which allows for fixed effects in the residual variance.

The DHGLM approach also gives a possibility to analyze non-normal traits and it should be a good idea to fit a model to the pig litter size data where the dependent variable is poisson distributed. Important future development of the

DHGLM framework is to add ρ as a parameter of the model and to add model selection criteria based on the h-likelihood.

Acknowledgements

We thank Danish Pig Production for allowing us to use their data and Daniel Sorensen for providing the data. We thank Youngjo Lee (Seoul National University, South Korea) and Yudi Pawitan (Karolinska Institute, Stockholm) for valuable discussions on HGLM theory. This project is partly financed by the RobustMilk project, which is financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708. The content of this paper is the sole responsibility of the authors, and it does not necessarily represent the views of the Commission or its services.

References

[1] Bollerslev T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31:307-327.

[2] Harvey A.C., E. Ruiz and N. Shephard. 1994. Multivariate stochastic variance models. *Review of Economic Studies* 61:247-264.

[3] Hill W.G. and X.S. Zhang. 2004. Effects on phenotypic variability of directional selection arising through genetic differences in residual variability. *Genetical Research* 83:121-132.

[4] Jaffrezic F., I.M.S. White, R. Thompson and W.G. Hill. 2000. A Link Function Approach to Model Heterogeneity of Residual Variances Over Time in Lactation Curve Analyses. Journal of Dairy Science 83:1089–1093

[5] Lee Y. and J. A. Nelder. 1996. Hierarchical generalized linear models (with Discussion). *Journal of the Royal. Statistical Society B* 58:619-678.

[6] Lee Y. and J. A. Nelder. 2006. Double hierarchical generalized linear models (with discussion). *Applied Statistics* 55:139-185.

[7] Lee Y, J. A. Nelder and Y. Pawitan. 2006. *Generalized linear models with random effects*. Chapman and Hall, Boka Raton, U.S..

[8] Mulder H. A., P. Bijma and W.G. Hill. 2007. Prediction of breeding values and selection response with genetic heterogeneity of environmental variance. *Genetics* 175:1895-1910.

[9] Noh M., B. Yip, Y. Lee and Y. Pawitan. 2006. Multicomponent variance estimation for binary traits in family-based studies. *Genetic Epidemiology* 30:37-47.

[10] SanCristobal-Gaudy, M., J. M. Elsen, L. Bodin, and C. Chevalet. 1998. Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. *Genet. Sel. Evol.* 30:423–451.

[11] Sorensen D. and R. Waagepetersen. 2003. Normal linear models with genetically structured residual variance heterogeneity: a case study. *Genetical Research* 82:207-222.

[12] Wolc A., I. M. S. White, S. Avendano and W. G. Hill. 2009. Genetic variability in residual variation of body weight and conformation scores in broiler chickens. *Poultry Science* 88:1156-1161.