BAYESIAN HETEROSKEDASTIC GENERALIZED LINEAR MIXED MODELS FOR ANIMAL BREEDING APPLICATIONS

K. Kizilkaya and R. J. Tempelman

Dept. of Animal Science, Michigan State University, East Lansing, MI, USA

INTRODUCTION

The existence of heterogeneous genetic and residual variances across environments has been established for a number of production traits (e.g. San Cristobal *et al.*, 1993; Meuwissen *et al.*, 1996). The ignorance of such heteroskedascity can lead to biased predictions of breeding values, potentially favoring a disproportionate numbers of animals selected from high variance environments (Hill, 1984). Recently, residual heteroskedasticity extensions have been proposed for threshold model analysis of ordinal categorical data (Foulley and Gianola, 1996; Jaffrézic *et al.*, 1999). However, many of the proposed models invoke analytical approximations which appear tenuous, particularly for the analysis of categorical data. Furthermore, we perceive the lack of a unifying framework for structural modeling of heterogeneous variances in generalized linear mixed model (GLMM) analysis of either continuous production or categorical fitness traits. The objective of our study is to propose a Bayesian structural multiplicative model on residual variances for observed or augmented variables in heteroskedastic GLMM, concentrating on a cumulative probit threshold model analysis of ordinal data based on use of Markov Chain Monte Carlo (MCMC) methods.

MODELS AND METHODS

The heteroskedastic generalized linear mixed model. In a number of generalized linear mixed models, data augmentation schemes exist such that a n x 1 vector of either observed or augmented variables $\mathbf{L} = \{L_i\}_{i=1}^{n}$ can be written as function of fixed effects β and random effects $\mathbf{u}$. Examples where such augmented variables are useful include threshold models (Sorensen *et al.*, 1995) and censored data models (Sorensen *et al.*, 1998). We write this linear mixed model as:

$$\mathbf{L} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad [1]$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices and $\mathbf{e}$ is a vector of normally distributed residuals. Suppose that $\mathbf{e}$ is further partitioned according to specified levels of residual heteroskedasticity: $\mathbf{e}' = \left[\mathbf{e}_{11}' \quad \mathbf{e}_{12}' \quad \ldots \quad \mathbf{e}_{st}'\right]$ where $\mathbf{e}_{kl} \sim N\left(\mathbf{0}, \mathbf{I}_{n_{kl}}\sigma^2_{e_{kl}}\right)$ pertains to the n_{kl} x 1 subvector of residuals identified with the kth level of a "fixed effect" subclass (e.g. sex) for residual variances k = 1,2,…,s and the lth level of a "random effect" subclass (e.g. herd) for residual variances; l = 1,2,…,t. A clear interpretation of the meaning of "fixed" and "random" effects modeling of variances in an animal breeding context is provided by San Cristobal *et al.* (1993). However, in contrast to the additive models that have been most commonly presented for modeling log-variances, we propose a multiplicative model for residual variances as follows:

$$\sigma^2_{e_{kl}} = \bar{\sigma}^2_{e_k}\delta_l \text{ , } k=1,2,\ldots,s;\ l = 1,2,\ldots,t. \qquad [2]$$

Here $\bar{\sigma}^2_{e_k}$ is the residual variance identified with the *k*th level of a fixed effects subclass and $\delta_l > 0$ is a random multiplicative scaling factor unique to the *l*th level of a random effects subclass.

Either subjective or flat priors are typically specified for β whereas a structural multivariate Gaussian prior is specified for **u**, i.e. $\mathbf{u} \sim N(\mathbf{0},\mathbf{G}(\phi))$. Here **G** is a function of several unknown variances or variance-covariance matrices in ϕ. Inverted gamma densities, inverted Wishart densities or combinations thereof are typically specified as priors for ϕ, depending on whether or not there are multiple sets of random effects with specified covariances between various sets, such as between additive and maternal genetic effects (Jensen *et al.*, 1994). A subjective inverted-gamma, for conjugate convenience, or flat prior may be specified separately for each $\bar{\sigma}^2_{e_k}$, $k = 1,2,\ldots,s$, whereas a structural prior is used to model the random multiplicative scaling factors, δ_l, $l = 1,2,\ldots,t$. We conveniently choose this structural prior to be an inverted-gamma density with parameters α_e and α_e-1, i.e.

$$p(\delta_l|\alpha_e) \propto \frac{(\alpha_e - 1)^{\alpha_e}}{\Gamma(\alpha_e)} (\delta_l)^{-(\alpha_e+1)} \exp\left(-\frac{\alpha_e - 1}{\delta_l}\right); \quad l = 1,2,\ldots,t \qquad [3]$$

Note that $E(\delta_l)=1$ and $Var(\delta_l)=(\alpha_e-2)^{-1}$ such that as $\alpha_e \to \infty$, that random effects factor's influence on residual heteroskedasticity diminishes. Note that with the specification in [3], there is a borrowing of information across levels $l = 1,2,\ldots,t$ of the random factor, just as there is for random effects modeling of location parameters.

The remaining specifications in this heteroskedastic GLMM would depend upon the first (data sampling) stage of the *n* x 1 data vector **y**. For normally distributed data, [1] would suffice (i.e. **y** = **L** or observed data = augmented variables) such that no other parameters are needed, whereas for a cumulative probit threshold model for ordinal data with *C* ordinal categories, numbered $j = 1,2,\ldots,C$, we would specify the first stage of our hierarchical model using Sorensen *et al.* (1995):

$$p(\mathbf{y} \mid \mathbf{L}, \boldsymbol{\tau}) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{C} 1(\tau_{j-1} < L_i < \tau_j) 1(Y_i = j) \right\} \qquad [4]$$

Here $\boldsymbol{\tau} = [\tau_1 \quad \tau_2 \quad \ldots \quad \tau_{C-1}]'$ denotes a vector of unknown threshold parameters that delimit the augmented variables **L** into their respective observed data bins **y** with 1(.) denoting the indicator function having value 1 if the condition within the function is true, and 0 otherwise. Furthermore, $\tau_o = -\infty$ and $\tau_C = +\infty$. Note that we adopt the parameterization of Sorensen *et al.* (1995) in which the residual variance is explicitly modeled such that only *C*-3 parameters in τ are uniquely identifiable. That is, our heteroskedastic threshold model, as per Foulley and Gianola (1996), does not readily extend to the modeling of binary outcomes.

The joint posterior density of **L** (if required), β, **u**, $\left\{\bar{\sigma}^2_{e_k}\right\}_{k=1}^{s}$, $\{\delta_l\}_{l=1}^{t}$, ϕ, α_e and any other parameters necessary for the GLMM in question (e.g. τ in a cumulative probit threshold model) is simply specified as the product of the various stages of the hierarchical model. A MCMC inference strategy requires determination of and sampling from the full conditional distributions (FCD) of each parameter or groupings thereof. It can be readily shown that the FCD of β and **u** is multivariate normal, and the FCD of individual elements of **L** are truncated normal. Furthermore, it can also be readily shown that the FCD of each of $\bar{\sigma}^2_{e_k}$, k = 1,2,…,s and each δ_l, l = 1,2,…,t are inverted-gamma, whereas the FCD of components of ϕ are either inverted-Wishart or inverted-gamma, again depending on whether unknown covariances are specified in ϕ or not. The FCD for α_e is not readily recognizable such that a Metropolis Hasting sampling update seems necessary. Although the FCD for τ is recognizable, as demonstrated by Sorensen *et al.* (1995), we advocate the Metropolis-Hastings update proposed by Cowles (1996) since it demonstrates superior mixing properties.

Simulation study. A simple mixed effects model was used to generate augmented variables **L** for n=2500 progeny from 50 unrelated sires based on the following linear model

$$L_{ijkl}=\mu+sex_i+herd_j+sire_k+e_{ijkl} \qquad [5]$$

Here μ=0.5, $\{sex_i\}_{i=1}^{2}$ (sex_1=-0.5 and sex_2=0.5) represents a 2x1 vector of fixed sex effects. Furthermore, $\left\{herd_j\right\}_{j=1}^{100}$~N(**0**,**I**$\sigma^2_h$) represents a 100 x 1 vector of random effects and $\{sire_k\}_{k=1}^{50}$ ~ N(**0**,**I**σ^2_s) represents a 50 x 1 vector of independent random effects with σ^2_h = 0.25 and σ^2_s = 0.10. Finally, $\mathbf{e}_{ij}$ = $\{e_{ijkl}\}$ ~ N(**0**, $\mathbf{I}_{n_{ij}}\bar{\sigma}^2_{e_i}\delta_j$) represents the vector of residuals associated with the n_{ij} records from sex i and herd j where $\bar{\sigma}^2_{e_1}=1$, $\bar{\sigma}^2_{e_2}=1.25$ and δ_j ~ Gamma(α_e, α_e-1), j = 1,2,...,100. Two replicated datasets from each of two different populations or different values of α_e were generated: 1) α_e=3 and 2) α_e=50. These values represent extreme and mild levels of residual heteroskedasticity, respectively, across herds. Levels of fixed and random effects were randomly assigned together for data generation. Augmented data **L** was mapped to ordinal data **y** based on C=4 categories with τ_1=–0.50, τ_2=1.00 and τ_3=2.00 in both populations. Both **L** and **y** were analyzed using the appropriate GLMM based on both homogeneous and heterogeneous residual variance models in order to assess the ability of the Deviance Information Criterion (DIC) (Spiegelhalter *et al.*, 2002) to correctly choose the right model.

RESULTS AND DISCUSSION

The 95% equal-tailed posterior probability intervals (PPI) are given for each of the dispersion parameters in Table 1 for each of the four replicate datasets, two from each α_e population.

Table 1. 95% Equal-Tailed Posterior Probability Intervals for dispersion parameters under a heteroskedastic cumulative probit threshold model of simulated ordinal data (and corresponding intervals under linear mixed model analysis of augmented variables)

	$\alpha_e = 3$		$\alpha_e = 50$	
Parameters	Replicate 1	Replicate 2	Replicate 1	Replicate 2
σ_s^2	0.08–0.21	0.07-0.20	0.06-0.18	0.12-0.30
	(0.08-0.21)	(0.08-0.21)	(0.07-0.19)	(0.12-0.31)
σ_h^2	0.17-0.35	0.20-0.40	0.17-0.35	0.21-0.43
	(0.17-0.33)	(0.22-0.42)	(0.16-0.32)	(0.23-0.44)
$\bar{\sigma}_{e1}^2$	0.78-1.14	0.85-1.24	0.87-1.12	0.98-1.19
	(0.76-1.05)	(0.87-1.20)	(0.87-1.04)	(0.99-1.18)
$\bar{\sigma}_{e2}^2$	0.95-1.41	0.91-1.36	1.15-1.53	1.07-1.44
	(0.93-1.29)	(1.07-1.47)	(1.16-1.38)	(1.10-1.32)
α_e	2.58-5.07	2.55-4.84	16.8-2458	10.2-534
	(2.53-4.33)	(2.53-4.23)	(21.2-724)	(19.0-439)

The PPI in Table 1 indicate generally satisfactory coverage properties. The DIC correctly chose the heterogeneous over the homogeneous variance model for each of the replicate datasets. We also found that DIC correctly chose the homoskedastic over the heteroskedastic model in homoskedastic simulated data, thereby indicating its value for model choice.

CONCLUSIONS

We have developed a hierarchical Bayes model that accounts for residual heteroskedasticity in augmented/observed variables in GLMM. Extensions to heterogeneous genetic variances and to multiple (rather than single) fixed and random multiplicative factors for a structural model are readily facilitated. We recently applied our model to first parity calving ease scores in beef cattle, determining significant evidence of residual heteroskedasticity across herds and sexes.

REFERENCES

Cowles, M..K. (1996) *Stat. Comput.* **6** : 101-111.

Foulley, J.L. and Gianola, D. (1996) *Genet. Sel. Evol.* **28** : 249-273.

Hill, W.G. (1984) *Anim. Prod.* **39** : 473-477.

Jaffrézic, F., Robert-Granié, C., and Foulley, J.L. (1999) *Genet. Sel. Evol.* **31**: 301-318.

Jensen, J., Wang, C.S., Sorensen, D.A., and Gianola, D. (1994) *Acta. Agric. Scand. Sect. A.* **44** :193-201.

Meuwissen, T.H.E., de Jong, G., and Engel, B. (1996) *J. Dairy Sci.* **79** : 310-316.

San Cristobal, M., Foulley, J.L., and Manfredi, E. (1993) *Genet. Sel. Evol*. **25** : 3-30.

Sorensen, D.A., Andersen, S., Gianola, D., and Korsgaard, I. (1995) *Genet. Sel. Evol.* **27** : 229-249.

Sorensen, D.A., Gianola, D., and Korsgaard, I. (1998) *Acta. Agric. Scand. Sect. A.* **48** : 222-229.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002) *J. Royal Stat. Soc. Ser. B.* (in press)