Estimation of breeding values for mean and dispersion, their variance and correlation using double hierarchical generalized linear models

M. FELLEKI[1,2*], D. LEE[3], Y. LEE[3], A. R. GILMOUR[4] AND L. RÖNNEGÅRD[1,2]

[1] *School of Technology and Business Studies, Dalarna University, 79188 Falun, Sweden*
[2] *Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, 75007 Uppsala, Sweden*
[3] *Department of Statistics, Seoul National University, Seoul 151-747, Korea*
[4] *School of Mathematics and Applied Statistics, Faculty of Informatics, University of Wollongong, Wollongong, NSW 2522, Australia*

Summary

The possibility of breeding for uniform individuals by selecting animals expressing a small response to environment has been studied extensively in animal breeding. Bayesian methods for fitting models with genetic components in the residual variance have been developed for this purpose, but have limitations due to the computational demands. We use the hierarchical (h)-likelihood from the theory of double hierarchical generalized linear models (DHGLM) to derive an estimation algorithm that is computationally feasible for large datasets. Random effects for both the mean and residual variance parts of the model are estimated together with their variance/covariance components. An important feature of the algorithm is that it can fit a correlation between the random effects for mean and variance. An h-likelihood estimator is implemented in the R software and an iterative reweighted least square (IRWLS) approximation of the h-likelihood is implemented using ASReml. The difference in variance component estimates between the two implementations is investigated, as well as the potential bias of the methods, using simulations. IRWLS gives the same results as h-likelihood in simple cases with no severe indication of bias. For more complex cases, only IRWLS could be used, and bias did appear. The IRWLS is applied on the pig litter size data previously analysed by Sorensen & Waagepetersen (2003) using Bayesian methodology. The estimates we obtained by using IRWLS are similar to theirs, with the estimated correlation between the random genetic effects being -0.52 for IRWLS and -0.62 in Sorensen & Waagepetersen (2003).

1. Introduction

In linear mixed models, it is often assumed that the residual variance is homogeneous. However, differences in the residual variance among individuals are quite common and it is important to include the effect for heteroscedastic residuals in models for traditional breeding value evaluation (Hill, 1984). Such models, having explanatory variables accounting for heteroscedastic residuals, are routinely used by breeding organizations today. The explanatory variables are typically non-genetic (Meuwissen *et al.*, 1996; Yang *et al.*, 2012), but genetic heterogeneity can be present and it is included as a random effect for the residual variance part of the model.

Product uniformity is often a desirable breeding goal. Therefore, we need methods to estimate both variance components and breeding values in the residual variance part of the model to be able to select animals that can satisfy this goal. Moreover, if genetic heterogeneity is present then traditional methods for predicting selection response can be misleading (Hill & Zhang, 2004; Mulder *et al.*, 2007).

Methods have previously been developed to estimate the degree of genetic heterogeneity. SanCristobal-Gaudy *et al.* (1998) have developed a likelihood-based expectation-maximization (EM) algorithm. Sorensen & Waagepetersen (2003) have applied a Bayesian Markov chain Monte Carlo (MCMC) algorithm to estimate the parameters in a similar model, which has the advantage of producing model-checking tools based on posterior predictive distributions and model-selection criteria based on

* Corresponding author: School of Technology and Business Studies, Dalarna University, SE-79188 Falun, Sweden. Tel: +46(0)23 77 82 82. E-mail: mfl@du.se

the Bayes factor and deviances. At the same time, Bayesian methods to fit models with residual hetero-scedasticity for multiple breed evaluations (Cardoso *et al.*, 2005) and generalized linear mixed models allowing for a heterogenetic variance term (Kizilkaya & Tempelman, 2005) have been developed. Wolc *et al.* (2009) have studied a sire model, with random genetic effects included in the residual variance, by fitting squared residuals with a gamma generalized linear mixed model. Mulder *et al.* (2009) fitted a bivariate linear model for the trait and the log-squared residuals, and they estimated the correlation between effects for mean and variance. Hill & Mulder (2010) reviewed the topic of genetic control of environmental variation. Yang *et al.* (2011) showed that inferences under the genetically structured heterogeneous variance model can be misleading when the data are skewed.

Lee & Nelder (2006) developed their framework of double hierarchical generalized linear models (DHGLM), which has been applied to stochastic volatility modelling in finance (del Castillo & Lee, 2008) and also to robust inference against outliers by allowing heavy tailed distributions (Lee & Nelder, 2006). Inference in DHGLM is based on hierarchical likelihood (h-likelihood) theory (Lee & Nelder, 1996) and DHGLM is a direct extension of hierarchical GLM (HGLM) including random effects for the dispersion. Two main computational strategies are available. In the first, the parameters are estimated by iterating a hierarchy of generalized linear models (GLM), where each GLM is estimated by iterative reweighted least squares (IRWLS; see e.g. Rönnegård *et al.*, 2010). The second strategy works by numerically maximizing the h-likelihood (see e.g. Molas & Lesaffre, 2011). The first is computational fast, while the other may produce higher-order estimators (Noh & Lee, 2007). DHGLM give model checking tools based on GLM theory and model-selection criteria are calculated from the h-likelihood (Lee & Nelder, 1996). Both the theory and the fitting algorithm are explained in detail in Lee *et al.* (2006).

Rönnegård *et al.* (2010) used DHGLM to estimate breeding values for mean and dispersion in an animal model, and also their variances, but the correlation between them was not included. The computational demands were decreased compared with Markov chain Monte Carlo estimation. The method by Rönnegård *et al.* (2010) can be used to estimate genetic heterogeneity of the residual variance in animal models with many observations.

Previously correlation between random effects for the same level in DHGLM have been estimated (Lee *et al.*, 2006), i.e. correlation between random effects for the mean, or correlation between random effects for the dispersion. Correlation between random effects for different levels, that is between an effect for the mean model and an effect for the residual variance model, has not previously been reported within the DHGLM framework.

The aim of this paper is to extend DHGLM, and thereby the method from Rönnegård *et al.* (2010), to include correlation between random effects for mean and variance, and moreover to evaluate the performance with regard to bias and precision.

2. Material and methods

In this section, we start by defining the exponential model (Hill & Mulder, 2010) introduced by SanCristobal-Gaudy *et al.* (1998). Thereafter, a bivariate linear mixed model is used as a tool in an IRWLS algorithm for fitting the exponential model. The algorithm is a natural extension of the algorithm by Rönnegård *et al.* (2010) by including a correlation between random effects for mean and variance. It is also quite similar to the one presented by Mulder *et al.* (2009), except that the algorithm below corrects for the fact that estimated, and not true, residuals are used, and that the squared residuals are gamma distributed. In the Appendix, we derive that the IRWLS algorithm can be motivated using the h-likelihood method.

(i) *Model*

The model fitted is the exponential model,

$$E(y|a, a_d) = \mu \tag{1}$$

with a linear predictor

$$\mu = X\beta + Za.$$

The dispersion part of the model is specified as

$$var(y|a, a_d) = \Phi$$

with linear predictor

$$\Phi = \text{diag } \phi, \quad \log \phi = X_d\beta_d + Za_d.$$

By y is denoted a vector of n responses depending on fixed effects β and random effect a (length k), and ϕ is a vector of residual variances depending on fixed effects β_d and random effect a_d. The response vector y given a and a_d is normal distributed. Matrices for design X and X_d, and the incidence matrix Z are known.

The random effects a and a_d are normal distributed additive genetic effects with mean 0 and variance

$$var\begin{pmatrix} a \\ a_d \end{pmatrix} = G \otimes A, \tag{2}$$

$$G = \begin{pmatrix} \sigma_a^2 & \sigma_a \sigma_{a_d} \rho \\ \sigma_a \sigma_{a_d} \rho & \sigma_{a_d}^2 \end{pmatrix},$$

where A is the additive genetic relationship matrix of dimension k for the additive genetic effects, thus $\sigma_a^2 A$ is the covariance of a and $\sigma_{a_d}^2 A$ is the covariance of a_d. The parameter ρ is the correlation between a and a_d as included in the estimation by Sorensen & Waagepetersen (2003).

The model on responses y is referred to as the mean part, and the model on the residual variances ϕ is referred to as the residual variance part. Other un-correlated random effects (permanent environmental effects) can be added to both the mean and the residual variance parts.

The additive genetic effect a_d for the residual variance $\log \phi$ is the genetic control of environmental variation, which is a measure on the uniformity of trait expression. The correlation parameter ρ indicates how the uniformity varies with the breeding value of the animal. A numerically high value of ρ would mean that selection of high breeding values for the mean would change the environmental variance.

(ii) *A bivariate linear mixed model and a fitting algorithm*

Consider the bivariate linear mixed model

$$\begin{pmatrix} y \\ z_d \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & X_d \end{pmatrix} \begin{pmatrix} \beta \\ \beta_d \end{pmatrix} + \begin{pmatrix} Z & 0 \\ 0 & Z \end{pmatrix} \begin{pmatrix} a \\ a_d \end{pmatrix} + \begin{pmatrix} e \\ e_d \end{pmatrix}, \tag{3}$$

with working weight matrix and residual variance

$$\Sigma^{-1} = \begin{pmatrix} \Phi^{-1} & 0 \\ 0 & \mathrm{diag}\left(\frac{1-q}{2}\right) \end{pmatrix},$$

$$\mathrm{var}\begin{pmatrix} e \\ e_d \end{pmatrix} = \Sigma \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & \sigma_d^2 I_n \end{pmatrix}.$$

Here the residual variances σ^2 and σ_d^2 are both 1 (but defined for later use in the fitting algorithm below), z_d is the vector of linearized working variables (McCullagh & Nelder 1989)

$$z_{d,i} = \log \phi_i + \phi_i^{-1} \left(\frac{\hat{e}_i^2}{1-q_i} - \phi_i \right), \tag{4}$$

and q is the vector of hat values defined in Appendix (i)(b).

This gives estimates similar to h-likelihood estimates, as shown in the Appendix. We call the bivariate REML method used to fit the linear mixed model above an IRWLS approximation of the h-likelihood.

Fitting algorithm for the IRWLS approximation

The above linear mixed model (3) is fitted by the following IRWLS algorithm.

1. Initialize weights $\Sigma^{-1} = \begin{pmatrix} \Phi^{-1} & 0 \\ 0 & \mathrm{diag}\left(\frac{1-q}{2}\right) \end{pmatrix}$ and working variables z_d.
2. Fit (3) with correlation structure $G \otimes A$ between a and a_d, but y and z_d conditional uncorrelated, that is, $\mathrm{var}(y|a, a_d) = \sigma^2 \Phi$ and $\mathrm{var}(z_d|a, a_d) = \sigma_d^2 \mathrm{diag}(2/(1-q))$.
3. Update $\hat{\sigma}^2$, $\hat{e}$ and q, and thereby $z_d = \log(\hat{\sigma}^2 \phi) + \frac{1}{\hat{\sigma}^2} \Phi^{-1}(\hat{e}^2/(1-q) - \hat{\sigma}^2 \phi)$ and $\mathrm{diag}((1-q)/2)$ in Σ^{-1}.
4. Repeat step 2.
5. Update $\Phi^{-1} = \mathrm{diag}(\exp(\hat{z}_d))^{-1}$ and update Σ^{-1}.
6. Iterate steps 2–5 until convergence ($\hat{\sigma}^2 = 1$).

(iii) *Simulations for validating the h-likelihood method*

Estimation of correlation between random effects for mean and dispersion is new within the h-likelihood framework and there are potential applications in other research areas than genetics. A relatively simple simulation structure was therefore considered. Using these simulations the h-likelihood method and the IRWLS approximation were compared. The h-likelihood was implemented in the R software using a Newton–Raphson algorithm on the score functions. The score functions are given in the Appendix for the effects for mean and variance, together with the adjusted profile likelihood for the variance components. The variance components were estimated using transformation $\xi = (\log \sigma_a^2, \log \sigma_{a_d}^2, \log \frac{1+\rho}{1-\rho})$ to avoid boundary problems.

We simulated 10 000 observations and a random group effect. The number of groups was either 10 or 100. An observation for individual j with covariate x belonging to group k was simulated as $y_{jk} = 1 \cdot 0x + a_k + e_{jk}$, where the random group effects were iid with $a_k \sim N(0, \sigma_a^2)$, and the residual effect e_{jk} was sampled from $N(0, \mathrm{var}(e_{jk}))$ with $\mathrm{var}(e_{jk}) = \exp(0 \cdot 2x + a_{d,k})$. The covariate x was simulated binary to resemble sex effect. Furthermore, $a_{d,k} \sim N(0, \sigma_{a_d}^2)$ with $\mathrm{cov}(a_k, a_{d,k}) = \rho \sigma_a \sigma_{a_d}$. The simulated variance components were $\sigma_a^2 = 1$ and $\sigma_{a_d}^2 = 0 \cdot 1$ or $0 \cdot 5$, whereas the correlation ρ was either $-0 \cdot 5$ or $0 \cdot 95$. The value of $\sigma_{a_d}^2 = 0 \cdot 1$ gives a substantial variation in the simulated elements of a_d, where one standard deviation difference between two values $a_{d,l}$ and $a_{d,m}$ increases the residual variance $1 \cdot 37$ times. We replicated the simulation 100 times and obtained estimates of variance components using the h-likelihood method and the IRWLS algorithm.

The computational demands of the h-likelihood method, implemented in R, limited analyses of more sophisticated unbalanced data scenarios

and were therefore not used in the following analyses.

(iv) *Pig litter size data and model*

Pig litter size was previously analysed by Sorensen & Waagepetersen (2003) using the Bayesian method, and the data are described therein. The data includes 10 060 records from 4149 sows and a total number of 6437 pigs in the pedigree. Hence, repeated measurements on sows were present and a permanent environmental effect of each sow was included in the model. The maximum number of parities is nine. The data include the following class variables: herd (82 classes), season (four classes), type of insemination (two classes) and parity (nine classes). The data are highly unbalanced; 13 herds contribute five observations or less, and the ninth parity includes nine observations.

Several models were analysed by Sorensen & Waagepetersen (2003) with an increasing level of complexity in the model for the residual variance, and with the model for the mean $y = X\beta + Wp + Za + e$. Here y is the litter size (vector of length 10 060), β is a vector including the fixed effects of herd, season, type of insemination and parity, and X is the corresponding design matrix (10 060 × 94), p is the random permanent environmental effect (vector of length 4149), W is the corresponding incidence matrix (10 060 × 4149) and $\mathrm{var}(p) = \sigma_p^2 I$, a is the random additive genetic effect, Z is the corresponding incidence matrix (10 060 × 6437) and $\mathrm{var}(a) = \sigma_a^2 A$, where A is the additive relationship matrix. Hence, the LHS of the mixed model equations is of size 10 680 × 10 680.

The residual variance $\mathrm{var}(e)$ was modelled as follows, where Model III and Model IV are model names from Sorensen & Waagepetersen (2003):

Model III: Random additive genetic effect and fixed effects for the linear predictor for the residual variance

$$\mathrm{var}(e_i) = \exp(X_{d,i}\beta_d + Z_i a_d),$$

where β_d is a parameter vector including an intercept β_0 and effects of parity and type of insemination, $X_{d,i}$ is the ith row in the design matrix X_d, Z_i is the ith row of Z, and a_d is the random additive genetic effect.

Model IV: Permanent environmental effect, additive genetic effect, and fixed effects for the linear predictor for the residual variance

$$\mathrm{var}(e_i) = \exp(X_{d,i}\beta_d + W_i p_d + Z_i a_d),$$

where W_i is the ith row of W and p_d is a random permanent environmental effect with $p_d \sim N(0, \sigma_{p_d}^2 I)$.

(v) *Simulations for validating the IRWLS approximation*

To test whether the IRWLS algorithm gives unbiased variance components on realistic examples for animal breeding, we simulated observations using the pedigree from the pigs litter size data. The number of sows with records was fixed as in the original dataset. The total number of observations was either kept ($n = 10\,060$), or increased by changing the number of repeated records per sow (parities) to 4 ($n = 4 \times 4149 = 16\,596$) or 9 ($n = 9 \times 4149 = 37\,341$).

On the mean level the observation for animal j, parity k and insemination type x was $y_{jk} = 11 \cdot 16 + \beta_{\mathrm{parity},k} + 0 \cdot 45x + p_j + a_j + e_{jk}$. The residual effect was sampled from $N(0, \mathrm{var}(e_{jk}))$ with var $(e_{jk}) = \exp(1 \cdot 77 + \beta_{d,\mathrm{parity},k} - 0 \cdot 17x + p_{d,j} + a_{d,j})$. Additive genetic effects were sampled from (2) with $\sigma_a^2 = 1 \cdot 62$, $\sigma_{a_d}^2 = 0 \cdot 09$ and $\rho = -0 \cdot 62$. Permanent environmental effects were sampled from $p_j \sim N(0, 0 \cdot 60)$ and $p_{d,j} \sim N(0, 0 \cdot 06)$ (Model IV) or $p_{d,j} = 0$ (Model III). These values came from Sorensen & Waagepetersen (2003). The simulations were replicated 100 times and analysed using the IRWLS algorithm.

3. Results

(i) *Simulations for validating the h-likelihood method*

For the simulations with balanced data, the two methods h-likelihood and IRWLS gave identical results (Table 1) up to the fourth decimal. When the number of groups was small (10 groups), there was a large variation in the estimates, because only ten values of a and a_d were sampled in each replicate. In that case there was a tendency for ρ to be overestimated (estimate $-0 \cdot 42$, 16 % overestimated) because of the parameter boundary $\rho \geqslant -1$. An alternative way to estimate ρ is to take mean of $\log((1 + \rho)/(1 - \rho))$. By doing that we avoid boundary problems and obtained the value $-0 \cdot 48$ (4 % overestimated).

The algorithms performed well for estimating variance components for models with correlation between random effects for mean and residual variance.

(ii) *Data on pigs litter sizes*

For the pigs litter size data, the results for the IRWLS algorithm (Table 2) were similar to the MCMC estimates of Sorensen & Waagepetersen (2003) for most parameters. The MCMC estimates of ρ were slightly more negative ($-0 \cdot 57$ for Model III and $-0 \cdot 62$ for Model IV) than the IRWLS estimates ($-0 \cdot 49$

Table 1. *Mean (standard errors) for 100 replicates of 10 000 balanced observations in k groups using the h-likelihood estimator.* (*The IRWLS algorithm gave identical results*)

True values				Estimates		
k	σ_a^2	$\sigma_{a_d}^2$	ρ	σ_a^2	$\sigma_{a_d}^2$	ρ
100	1·0	0·1	−0·5	1·02 (0·014)	0·10 (0·002)	−0·50 (0·008)
10	1·0	0·1	−0·5	0·93 (0·044)	0·10 (0·004)	−0·42 (0·033)
100	1·0	0·5	−0·5	1·02 (0·015)	0·50 (0·007)	−0·50 (0·007)
100	1·0	0·1	0·95*	1·00 (0·021)	0·11 (0·002)	0·94 (0·003)

*Results in the fourth row are calculated from 58 replicates that converged for both methods (97 replicates converged for the h-likelihood method and 60 replicates converged for the IRWLS algorithm).

for Model III and −0·52 for Model IV). Similarity of estimates for the two methods was found for the variance components σ_a^2 and $\sigma_{p_d}^2$, and for all of the fixed effects for the residual variance, β_d. The permanent environmental effect variance σ_p^2 in the mean model was about half the MCMC estimate, and the additive genetic variance $\sigma_{a_d}^2$ in the residual variance was considerably larger, compared with Sorensen & Waagepetersen (2003).

(iii) Simulations for validating the IRWLS approximation

The general trend showed that bias and standard error (SE) decreased when the number of parities was increased for each sow (Table 3). There were exceptions to this trend, including the intercept term β_{d0} of Model IV, which could be due to the fact that the number of parities for some sows was actually smaller in the setting with four parities for all sows compared with the original data structure. For the scenario with nine parities per sow, both bias and SE were very small or negligible for all parameters.

The estimates of $\sigma_{a_d}^2$ and ρ were biased for Model III, and Model IV seems to be more appropriate to use for situations with repeated observations. For Model IV, the magnitude in percentage bias was less than 10 % for all parameters except the permanent environmental effects for the mean and residual variance models (i.e. σ_p^2 and $\sigma_{p_d}^2$) (Table 3 (b)). The genetic parameters σ_a^2, $\sigma_{a_d}^2$ and ρ were estimated with small or no bias, for all scenarios and both models, indicating that the method will give good estimates for these parameters in a very general setting.

4. Discussion

We have extended the DHGLM framework to include correlation between random effects for the mean and the residual variance, for a normal response and normal distributed random effects. We have approximated the h-likelihood by an IRWLS algorithm that in summary works by iteratively updating and fitting a bivariate linear mixed model simultaneously on mean and residual variance. The IRWLS approximation of DHGLM is a fast and easily implemented algorithm for genetic heterogeneity estimation building on REML tools. The additional functions required are implemented in the developmental version of ASReml (Gilmour, 2010) to be released as ASReml 4 and an example code is available on request from the authors.

Significant bias was found for data structures having few repeated observations per individual, where the bias decreased quickly as the number of repeated observations increased. For data structures having few repeated observations, the largest bias was detected for σ_p^2. This is perhaps not surprising, because for an individual having a single observation the permanent environmental effect p_i and the residual e_i are indistinguishable, and part of the σ_p^2 will therefore be picked up by the residual variance.

It is possible that more accurate estimates could be obtained by implementing a computationally efficient algorithm (using sparse matrix techniques) for the h-likelihood estimation method presented in our paper, instead of using the IRWLS algorithm. This would, however, require further research. Furthermore, the second-order h-likelihood estimation method is known to eliminate bias for binary outcomes in HGLM (Lee *et al.*, 2006), but can be more demanding to compute.

Similar bias patterns were found when comparing the simulation study with the difference between the estimates obtained by Sorensen & Waagepetersen (2003) and the IRWLS estimates (Tables 2 and 3). The variance components for the mean model, the fixed effects for the residual variance and the correlation showed a similar pattern in differences. This might give an indication of a problem with the IRWLS estimates being biased in certain directions, when the data structure does not contain

Table 2. *Estimates and 95% confidence intervals of chosen parameters for pigs litter size data in Model III (first section) and Model IV (second section) used by Sorensen & Waagepetersen (2003). Results obtained by Sorensen & Waagepetersen (2003) (first row in each section), by Rönnegård* et al. *(2010) (second row) and using IRWLS (third row)*

| | Mean model | | Residual variance model | | | | | Cor ρ |
| | Variances | | Fixed effects* | | | Variances | | |
	σ_a^2	σ_p^2	β_{d_0}	$\beta_{d_{ins}}$	$\beta_{d_{par}}$	$\sigma_{a_d}^2$	$\sigma_{p_d}^2$	
Sorensen & Waagepetersen (2003) III	1·58	0·60	1·78	−0·16	0·34	0·11		−0·57
	1·13, 2·00	0·31, 0·96	1·65, 1·90	−0·24, −0·09	0·25, 0·43	0·08, 0·15		−0·72, −0·41
Rönnegård et al. (2010)	1·35	0·53	1·73	−0·17	0·32	0·13		
	0·99, 1·71	0·25, 0·81	1·61, 1·85	−0·23, −0·11	0·26, 0·39	0·09, 0·16		
IRWLS	1·61	0·34	1·70	−0·17	0·32	0·18		−0·49
	1·24, 1·97	0·08, 0·61	1·57, 1·82	−0·23, −0·11	0·26, 0·39	0·14, 0·22		−0·63, −0·36
Sorensen & Waagepetersen (2003) IV	1·62	0·60	1·77	−0·17	0·35	0·09	0·06	−0·62
	1·20, 2·05	0·30, 0·92	1·65, 1·89	−0·25, −0·09	0·26, 0·44	0·06, 0·13	0·05, 0·09	−0·80, −0·43
Rönnegård et al. (2010)	1·35	0·44	1·72	−0·17	0·32	0·09	0·06	
	1·00, 1·70	0·17, 0·71	1·62, 1·83	−0·23, −0·11	0·26, 0·39	0·05, 0·14	0·02, 0·11	
IRWLS	1·61	0·28	1·69	−0·17	0·32	0·15	0·05	−0·52
	1·25, 1·96	0·02, 0·54	1·57, 1·81	−0·23, −0·11	0·26, 0·39	0·10, 0·20	0·00, 0·09	−0·66, −0·37

*β_{d_0} is the intercept term in the model for the residual variance, $\beta_{d_{ins}}$ is the fixed effect for insemination and $\beta_{d_{par}}$ is the fixed effect for the difference in first and second parity.

Table 3. *Mean (a) and standard errors (b) of estimates of parameters for simulated data over Model III (first section) and Model IV (second section). The left hand column contains the simulated data structure*

| | Mean model | | Residual variance model | | | | | |
| | Variances | | Fixed effects | | | Variances | | Cor ρ |
	σ_a^2	σ_p^2	β_{d0}	$\beta_{d_{ins}}$	$\beta_{d_{par}}$	$\sigma_{a_d}^2$	$\sigma_{p_d}^2$	
(a)								
True values (Model III)	1·62	0·60	1·77	−0·17	0·35	0·09		−0·62
Original parity distribution	1·58	0·45	1·74	−0·17	0·33	0·11		−0·52
Four parities	1·59	0·61	1·73	−0·17	0·34	0·12		−0·56
Nine parities	1·63	0·60	1·74	−0·17	0·35	0·11		−0·60
True values (Model IV)	1·62	0·60	1·77	−0·17	0·35	0·09	0·06	−0·62
Original parity distribution*	1·56	0·24	1·73	−0·16	0·31	0·08	0·13	−0·61
Four parities	1·65	0·51	1·66	−0·16	0·34	0·09	0·15	−0·64
Nine parities	1·62	0·60	1·71	−0·17	0·35	0·09	0·09	−0·64
(b)								
(Model III)								
Original parity distribution	0·018	0·016	0·007	0·003	0·004	0·002		0·008
Four parities	0·015	0·012	0·005	0·003	0·004	0·002		0·007
Nine parities	0·012	0·008	0·005	0·001	0·003	0·001		0·005
(Model IV)								
Original parity distribution*	0·017	0·016	0·006	0·004	0·004	0·003	0·004	0·012
Four parities	0·017	0·012	0·007	0·003	0·004	0·002	0·003	0·008
Nine parities	0·013	0·008	0·005	0·002	0·003	0·001	0·001	0·005

*Twenty-seven out of 100 replicates did not converge. Estimates are for all replicates (with minor differences in results if these 27 replicates were included or not).

enough repeated observations to give good estimates. However, variance components for the residual variance did not follow the same pattern. A simulation study using Bayesian MCMC methods would be desirable but is not within the scope of the current paper.

The IRWLS algorithm is distinct from the algorithm used by Mulder *et al.* (2009) because it uses h-likelihood theory and the algorithm fits corrected squared residuals as a gamma distributed response, whereas the log-squared residuals were fitted as a normal response in Mulder *et al.* (2009). Compared with Bayesian methods we expect the IRWLS algorithm to be faster. The IRWLS computation on pigs litter size data took less than 5 min on a Linux server.

In a previous paper, Rönnegård *et al.* (2010), the variance part was fitted using a gamma generalized linear mixed model. In the presented IRWLS algorithm, we could have fitted a bivariate normal-gamma model, but chose to fit a corresponding bivariate normal–normal model that gives the same estimates at convergence. The bivariate normal–normal model resulted in a user-friendly code and is straightforward to compare with the method in Mulder *et al.* (2009). However, the bivariate normal–gamma model can be implemented by iteratively calling ASReml using some external code (in R or Fortran for instance).

This approach may be required if convergence problems for the IRWLS algorithm occur, but perhaps more importantly the user should assess whether the available data are sufficient to be able to fit the model. When the number of repeated observations is too small (or with a single record per animal) there might not be enough information to fit a model with random additive genetic effects both for the mean and residual variance models.

Inferences under the genetically structured heterogeneous variance model can be misleading when the data are skewed (Yang *et al.*, 2011). In analysis of data using DHGLM, scaling effects should be tested and a possibility would be to fit the IRWLS algorithm for different transformations of *y* combined with model selection tools to find an optimal transformation. This would, however, require further theoretical developments of the DHGLM approach.

The IRWLS algorithm provides a simple implementation of genetic heterogeneity models in existing REML estimation software useful in applied animal breeding and can deal with large datasets. We have focused on applications in animal breeding but the DHGLM framework is applicable in many other fields of research as well (including finance and industrial quality control (Lee *et al.*, 2006; Lee & Nelder, 2006)), and we expect the interest in our

Table A.1 *Parameters and the functions from which they are estimated by using h-likelihood and the IRWLS algorithm*

Parameter	h-likelihood	IRWLS algorithm: iterated REML on bivariate model
Effects for mean: $\tau_M = (\beta, a)$	h	l
Effects for variance: $\tau_D = (\beta_d, a_d)$	$p_{\tau_M}(h)$	l
Variance components: $\xi = (\log \sigma_a^2, \log \sigma_{a_d}^2, \log \frac{1+\rho}{1-\rho})$	$p_\tau(h)$, $\tau = (\tau_M, \tau_D)$	l_{REML}

theoretical development reaches beyond animal breeding.

5. Declaration of interests

The authors declare that they have no competing interests.

Appendix. IRWLS approximation of h-likelihood

We state the h-likelihood for the exponential model (1) and show that the IRWLS fitting algorithm for the bivariate linear mixed model (3) can be derived from it using an approximation. The h-likelihood will be maximized to estimate fixed and random effects for the mean part, an adjusted profile likelihood will be used to estimate fixed and random effects for the variance part, and from an additional profiling the estimates of variance components will be found. Adjusted profile likelihood corresponds to restricted maximum likelihood.

(i) *The h-likelihood for the exponential genetic heterogeneity model*

From Lee & Nelder (1996), the h-likelihood (ignoring constant terms) is

$$h = \log f(y|a, a_d) + \log f(a, a_d)$$
$$= -\frac{1}{2} \log \det \Phi - \frac{1}{2}(y - \mu)^T \Phi^{-1}(y - \mu)$$
$$- \frac{1}{2} \log \det G \otimes A - \frac{1}{2} \begin{pmatrix} a \\ a_d \end{pmatrix}^T G^{-1} \otimes A^{-1} \begin{pmatrix} a \\ a_d \end{pmatrix}.$$

This can be simplified as follows (ignoring a constant involving $\log \det A$):

$$h = -\frac{1}{2} \sum_i \log \phi_i - \frac{1}{2} \sum_i (y_i - \mu_i)^2 / \phi_i$$
$$- \frac{k}{2} \{ \log(\sigma_a^2) + \log(\sigma_{a_d}^2) + \log(1 - \rho^2) \}$$
$$- \frac{1}{2(1 - \rho^2)} \left(\frac{1}{\sigma_a^2} a^T A^{-1} a + \frac{1}{\sigma_{a_d}^2} a_d^T A^{-1} a_d \right)$$
$$+ \frac{\rho}{(1 - \rho^2)\sqrt{\sigma_a^2 \sigma_{a_d}^2}} a^T A^{-1} a_d.$$

Lee & Nelder (2001) considered the adjusted profile (log)- likelihood, which is defined by the generic function:

$$p_\alpha(l) = [l - \tfrac{1}{2} \log \det \{ D(l, \alpha)/(2\pi) \}]|_{\alpha = \tilde{\alpha}},$$

where l is a likelihood (either a log marginal-likelihood or an h-likelihood) with nuisance effects α, $D(l, \alpha) = -\partial^2 l/\partial \alpha^2$ and $\tilde{\alpha}$ solves $\partial l/\partial \alpha = 0$. When profiling out random effects, it is an extension of the REML likelihood and used for estimating variance components. Furthermore, from a classical likelihood point of view, profiling out random effects is a Laplace approximation of an integrated (marginal) likelihood (Molas & Lesaffre, 2011).

A summarization of parameters and the functions from which they are estimated using h-likelihood is found in Table A.1. For the estimation one uses h for $\tau_M = (\beta, a)$, $p_{\tau_M}(h)$ for $\tau_D = (\beta_d, a_d)$ and $p_\tau(h) = p_{\beta_a, \beta_d, a, a_d}(h)$ for $\xi = (\log \sigma_a^2, \log \sigma_{a_d}^2, \log \frac{1+\rho}{1-\rho})$. The fixed and the random effects are found by solving the score equations:

$$\begin{pmatrix} \frac{\partial h}{\partial \tau_M} \\ \frac{\partial p_{\tau_M}(h)}{\partial \tau_D} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{A.1}$$

and the variance components by

$$\frac{\partial p_\tau(h)}{\partial \xi} = 0. \tag{A.2}$$

(a) *Estimation of $\tau_M = (\beta, a)$*

The score function for $\tau_M = (\beta, a)$ is

$$
\begin{aligned}
S(\tau_M) &= \frac{\partial h}{\partial \tau_M} \\
&= \begin{pmatrix} X^{\mathrm{T}} \Phi^{-1}(y - \mu) \\ Z^{\mathrm{T}} \Phi^{-1}(y - \mu) - \frac{1}{(1-\rho^2)\sigma_a^2} A^{-1} a + \frac{\rho}{(1-\rho^2)\sqrt{\sigma_a^2 \sigma_{a_d}^2}} A^{-1} a_d \end{pmatrix},
\end{aligned}
$$

and zero is given by the mixed model equation

$$
\begin{pmatrix} X^{\mathrm{T}}\Phi^{-1}X & X^{\mathrm{T}}\Phi^{-1}Z \\ Z^{\mathrm{T}}\Phi^{-1}X & Z^{\mathrm{T}}\Phi^{-1}Z + \frac{1}{(1-\rho^2)\sigma_a^2}A^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ a \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}\Phi^{-1}y \\ Z^{\mathrm{T}}\Phi^{-1}y + \frac{\rho}{(1-\rho^2)\sqrt{\sigma_a^2 \sigma_{a_d}^2}}A^{-1}a_d \end{pmatrix}. \tag{A.3}
$$

(b) *Estimation of $\tau_D = (\beta_d, a_d)$*

For estimating $\tau_D = (\beta_d, a_d)$, we use the adjusted profile likelihood

$$
p_{\tau_M}(h) = [h - \tfrac{1}{2}\log\det\{D(h, \tau_M)/2\pi\}]\big|_{\tau_M = \hat{\tau}_M},
$$

where

$$
\begin{aligned}
D(h, \tau_M) &= -\frac{\partial^2 h}{\partial \tau_M \partial \tau_M^{\mathrm{T}}} \\
&= H = \begin{pmatrix} X^{\mathrm{T}}\Phi^{-1}X & X^{\mathrm{T}}\Phi^{-1}Z \\ Z^{\mathrm{T}}\Phi^{-1}X & Z^{\mathrm{T}}\Phi^{-1}Z + \frac{1}{(1-\rho^2)\sigma_a^2}A^{-1} \end{pmatrix}.
\end{aligned}
$$

With q being the column consisting of the diagonal elements of the hat matrix $[X\ Z]H^{-1}[X\ Z]^{\mathrm{T}}\Phi^{-1}$ (Hoaglin & Welsch, 1978),

$$
\begin{aligned}
S(\tau_D) &= \frac{\partial p_{\tau_M}(h)}{\partial \tau_D} = \left[\frac{\partial h}{\partial \tau_D} - \frac{1}{2}(\operatorname{vec} H^{-1})^T \frac{\partial \operatorname{vec} H}{\partial \tau_D}\right]\Big|_{\tau_M = \hat{\tau}_M} \\
&= \begin{pmatrix} -\frac{1}{2}X_d^{\mathrm{T}}1 + \frac{1}{2}X_d^{\mathrm{T}}\Phi^{-1}(y-\mu)^2 + \frac{1}{2}X_d^{\mathrm{T}}q \\ -\frac{1}{2}Z^{\mathrm{T}}1 + \frac{1}{2}Z^{\mathrm{T}}\Phi^{-1}(y-\mu)^2 + \frac{1}{2}Z^{\mathrm{T}}q - \frac{1}{(1-\rho^2)\sigma_{a_d}^2}A^{-1}a_d + \frac{\rho}{(1-\rho^2)\sqrt{\sigma_a^2 \sigma_{a_d}^2}}A^{-1}a, \end{pmatrix},
\end{aligned}
$$

where the function vec of a matrix stacks all columns into a single column (Magnus & Neudecker, 1999). The score function is solved by

$$
\begin{pmatrix} X_d^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})X_d & X_d^{T}\operatorname{diag}(\frac{1-q}{2})Z \\ Z^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})X_d & Z^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})Z + \frac{1}{(1-\rho^2)\sigma_{a_d}^2}A^{-1} \end{pmatrix} \begin{pmatrix} \beta_d \\ a_d \end{pmatrix} = \begin{pmatrix} X_d^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})z_d \\ Z^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})z_d + \frac{\rho}{(1-\rho^2)\sqrt{\sigma_a^2 \sigma_{a_d}^2}}A^{-1}a \end{pmatrix}, \tag{A.4}
$$

where z_d was defined in (4) by

$$
z_{d,i} = \log\phi_i + \phi_i^{-1}\left(\frac{e_i^2}{1-q_i} - \phi_i\right).
$$

(c) *Joint estimation of τ_M and τ_D*

The two MMEs (A.3) for the mean part and (A.4) for the variance part can be expressed as a single MME, namely

$$
C \begin{pmatrix} \beta \\ \beta_d \\ a \\ a_d \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}\Phi^{-1}y \\ X_d^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})z_d \\ Z^{\mathrm{T}}\Phi^{-1}y \\ Z^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})z_d \end{pmatrix}, \tag{A.5}
$$

with

$$
C = \begin{pmatrix} X^{\mathrm{T}}\Phi^{-1}X & 0 & X^{\mathrm{T}}\Phi^{-1}Z & 0 \\ 0 & X_d^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})X_d & 0 & X_d^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})Z \\ Z^{\mathrm{T}}\Phi^{-1}X & 0 & Z^{\mathrm{T}}\Phi^{-1}Z + \frac{1}{(1-\rho^2)\sigma_a^2}A^{-1} & -\frac{\rho}{(1-\rho^2)\sqrt{\sigma_a^2 \sigma_{a_d}^2}}A^{-1} \\ 0 & Z^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})X_d & -\frac{\rho}{(1-\rho^2)\sqrt{\sigma_a^2 \sigma_{a_d}^2}}A^{-1} & Z^{\mathrm{T}}\operatorname{diag}(\frac{1-q}{2})Z + \frac{1}{(1-\rho^2)\sigma_{a_d}^2}A^{-1} \end{pmatrix}. \tag{A.6}
$$

Note that the added terms in the lower right corner of C are simply the elements of $G^{-1} \otimes A^{-1}$.

In a DHGLM setting one would rather state the score function for τ_M and τ_D instead of the MME, that is, for $\tau = (\beta, \beta_d, a, a_d)$,

$$
S(\tau) = \begin{pmatrix} X & 0 & Z & 0 \\ 0 & X_d & 0 & Z \\ 0 & & I_{2q} & \end{pmatrix}^{\mathrm{T}} W^{-1} \begin{pmatrix} y & - & \mu \\ z_d & - & \log\phi \\ & - & a \\ & - & a_d \end{pmatrix},
$$

where

$$
W^{-1} = \begin{pmatrix} \Phi^{-1} & 0 & 0 \\ 0 & \operatorname{diag}(\frac{1-q}{2}) & 0 \\ 0 & 0 & G^{-1} \otimes A^{-1} \end{pmatrix}.
$$

(d) *Estimation of $\xi = (\log \sigma_a^2, \log \sigma_{a_d}^2, \log \frac{1+\rho}{1-\rho})$*

For estimation of $\xi = (\log \sigma_a^2, \log \sigma_{a_d}^2, \log \frac{1+\rho}{1-\rho})$, we use the adjusted profile likelihood

$$p_\tau(h) = [h - \tfrac{1}{2}\log \det\{D/2\pi\}]|_{\tau = \hat\tau},$$

where

$$D = D(h, \tau) = -\frac{\partial^2 h}{\partial \tau\, \partial \tau^T}$$

$$= \begin{pmatrix} X^T\Phi^{-1}X & X^T B_2 X_d & X^T\Phi^{-1}Z & X^T B_2 Z \\ X_d^T B_2 X & X_d^T B_1 X_d & X_d^T B_2 Z & X_d^T B_1 Z \\ Z^T\Phi^{-1}X & Z^T B_2 X_d & Z^T\Phi^{-1}Z + \frac{1}{(1-\rho^2)\sigma_a^2}A^{-1} & Z^T B_2 Z - \frac{\rho}{(1-\rho^2)\sqrt{\sigma_a^2\sigma_{a_d}^2}}A^{-1} \\ Z^T B_2 X & Z^T B_1 X_d & Z^T B_2 Z - \frac{\rho}{(1-\rho^2)\sqrt{\sigma_a^2\sigma_{a_d}^2}}A^{-1} & Z^T B_1 Z + \frac{1}{(1-\rho^2)\sigma_{a_d}^2}A^{-1} \end{pmatrix},$$

$B_1 = \mathrm{diag}(\tfrac{1}{2}(y_i - \mu_i)^2/\phi_i)$ and $B_2 = \mathrm{diag}((y_i - \mu_i)/\phi_i)$.

(ii) *Estimates from a bivariate linear mixed model gives approximate h-likelihood estimates*

In this section, we show that REML estimation from the bivariate linear mixed model (3) can be used to obtain approximate h-likelihood estimates.

At convergence of the IRWLS algorithm, the relationship between the h-likelihood and the joint log likelihood for the bivariate model (3) is

$$l = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\begin{pmatrix} y-\mu \\ z_d - \log\phi \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} y-\mu \\ z_d - \log\phi \end{pmatrix}$$
$$+ \log f(a, a_d)$$
$$= -\frac{1}{2}\sum_i \log\phi_i - \frac{1}{2}\sum_i \log\left(\frac{2}{1-q_i}\right) - \frac{1}{2}\sum_i (y_i - \mu_i)^2/\phi_i$$
$$- \frac{1}{2}\sum_i (z_{d,i} - \log\phi_i)^2 \frac{1-q_i}{2} + \log f(a, a_d)$$
$$= h - \frac{1}{2}\sum_i \log\left(\frac{2}{1-q_i}\right) - \frac{1}{2}\sum_i (z_{d,i} - \log\phi_i)^2 \frac{1-q_i}{2}.$$

The variance components for the bivariate model are estimated using the REML likelihood

$$l_{\mathrm{REML}} = l - \frac{1}{2}\log \det C, \tag{A.7}$$

where C was previously defined (A.6).

Comparison between functions from which parameters are estimated using h-likelihood and IRWLS is found in Table A.1. Note that the MME (A.5) solves l, h and $p_{\tau_m}(h)$, so if ξ was fixed, the estimates of fixed and random effects (β, β_d, a, a_d) would be the same for h-likelihood and IRWLS. The IRWLS algorithm

gives estimates of ξ from (A.7), but

$$\frac{\partial p_\tau(h)}{\partial \xi} = \frac{\partial h}{\partial \xi} - \frac{1}{2}(\mathrm{vec}\, D^{-1})^T \frac{\partial\, \mathrm{vec}\, D}{\partial \xi}$$
$$= \frac{\partial l}{\partial \xi} - \frac{1}{2}(\mathrm{vec}\, D^{-1})^T \frac{\partial\, \mathrm{vec}\, C}{\partial \xi}$$
$$\approx \frac{\partial l}{\partial \xi} - \frac{1}{2}(\mathrm{vec}\, C^{-1})^T \frac{\partial\, \mathrm{vec}\, C}{\partial \xi}$$
$$= \frac{\partial l_{\mathrm{REML}}}{\partial \xi},$$

so the approximation is only through multiplication by C^{-1} instead of D^{-1}. Moreover C and D are connected through $C = E(D)$. This is the only approximation done to the h-likelihood method by using the IRWLS algorithm.

References

Cardoso, F. F., Rosa, G. J. M. & Tempelman, R. J. (2005). Multiple-breed genetic inference using heavy-tailed structural models for heterogeneous residual variances. *Journal of Animal Science* **83**, 1766–1779.

del Castillo, J. & Lee, Y. (2008). Glm-methods for volatility models. *Statistical Modelling* **8**, 263–283.

Gilmour, A. R. (2010). ASReml 3.1 alpha version. Available at http://www.mmontap.org/downloads (accessed 8 December 2010).

Hill, W. G. (1984). On selection among groups with heterogeneous variance. *Animal Production* **39**, 473–477.

Hill, W. G. & Mulder, H. A. (2010). Genetic analysis of environmental variation. *Genetics Research* **92**, 381–395.

Hill, W. G. & Zhang, X. S. (2004). Effects on phenotypic variability of directional selection arising through genetic differences in residual variability. *Genetics Research* **83**, 121–132.

Hoaglin, D. C. & Welsch, R. E. (1978). The hat matrix in regression and anova. *American Statistician* **32**, 17–22.

Kizilkaya, K. & Tempelman, R. J. (2005). A general approach to mixed effects modeling of residual variances in generalized linear mixed models. *Genetics Selection Evolution* **37**, 31–56.

Lee, Y. & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**, 619–678.

Lee, Y. & Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika* **88**, 987–1006.

Lee, Y. & Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion). *Journal of the*

Royal Statistical Society : Series C (Applied Statistics) **55**, 139–185.

Lee, Y., Nelder, J. A. & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects : Unified Analysis via H-likelihood*. Boca Raton, FL : Chapman & Hall/CRC.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, rev. edn. Chichester : John Wiley & Sons.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn. London : Chapman & Hall.

Meuwissen, T. H. E., de Jong, G. & Engel, B. (1996). Joint estimation of breeding values and heterogeneous variances of large data files. *Journal of Dairy Science* **79**, 310–316.

Molas, M. & Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. *Journal of Statistical Software* **39**(13), 1–20.

Mulder, H. A., Bijma, P. & Hill, W. G. (2007). Prediction of breeding values and selection response with genetic heterogeneity of environmental variance. *Genetics* **175**, 1895–1910.

Mulder, H. A., Hill, W. G., Vereijken, A. & Veerkamp, R. F. (2009). Estimation of genetic variation in residual variance in female and male broiler chickens. *Animal* **3**, 1673–1680.

Noh, M. & Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis* **98**, 896–915.

Rönnegård, L., Felleki, M., Fikse, F., Mulder, H. A. & Strandberg, E. (2010). Genetic heterogeneity of residual variance – estimation of variance components using double hierarchical generalized linear models. *Genetics Selection Evolution* **42**(8), 1–10.

Rönnegård, L., Shen, X. & Alam, M. (2010). HGLM: a package for fitting hierarchical generalized linear models. *The R Journal* **2**, 20–28.

SanCristobal-Gaudy, M., Elsen, J. M., Bodin, L. & Chevalet, C. (1998). Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. *Genetics Selection Evolution* **30**, 423–451.

Sorensen, D. & Waagepetersen, R. (2003). Normal linear models with genetically structured residual variance heterogeneity: a case study. *Genetics Research* **82**, 207–222.

Wolc, A., White, I. M. S., Avendano, S. & Hill, W. G. (2009). Genetic variability in residual variation of body weight and conformation scores in broiler chickens. *Poultry Science* **88**, 1156–1161.

Yang, Y., Christensen, O. F. & Sorensen, D. (2011). Analysis of a genetically structured variance heterogeneity model using the Box–Cox transformation. *Genetics Research* **93**, 33–46.

Yang, Y., Schön, C.-C. & Sorensen, D. (2012). The genetics of environmental variation of dry matter grain yield in maize. *Genetics Research* **94**, 113–119.