Likelihood calculations to evaluate experimental designs to estimate genetic variances

Karin Meyer

Animal Genetics and Breeding Unit[1], University of New England, Armidale NSW 2351, Australia

E-mail : kmeyer@didgeridoo.une.edu.au
Phone : +61 2 6773 3331
Fax : +61 2 6773 3266

Abstract

Mixed model analyses via restricted maximum likelihood, fitting the so-called animal model, have become standard methodology for the estimation of genetic variances. Models involving multiple genetic variance components, due to different modes of gene action, are readily fitted. It is shown that likelihood based calculations may provide insight into the quality of the resulting parameter estimates, and are immediately applicable to the validation of experimental designs. This is illustrated for the example of a design suggested recently to estimate X-linked genetic variances. In particular, large sample variances and sampling correlations are demonstrated to provide an indication of 'problem' scenarios. Using simulation, it is shown that the profile likelihood function provides more appropriate estimates of confidence intervals than large sample variances. Examination of the likelihood function and its derivatives are recommended as part of the design stage of quantitative genetic experiments.

1 Introduction

Estimation of genetic parameters or, equivalently, (co)variance components in modern quantitative genetic analyses is by and large performed by fitting a linear mixed model and using maxi-

[1] AGBU is a joint venture between the NSW Department of Primary Industries and the University of New England

mum likelihood or Bayesian inference. In particular, residual or restricted maximum likelihood (REML), due to Patterson and Thompson (1971), is routinely employed. Early applications fitted models where observations were attributed to the genotype of parents, e.g. a 'sire' or 'sire and dam' model, mimicking corresponding analyses of variance (ANOVA) where the resulting estimates of variance components were equated to their expectations (Shaw, 1987; Fry, 1992). Today, however, we generally fit a model which relates observations to the additive genetic effects of the individuals recorded, accounting for covariances with relatives through the so-called numerator relationship matrix (whose elements are twice the coefficient of co-ancestry). This directly yields estimates of the genetic variance components. Impetus for this has largely come from applications to data from livestock improvement schemes. Hence the model is commonly referred to as the 'animal model' or even 'individual animal model' (Meyer, 1989). More recently, animal model type analyses have seen increasing uptake in other areas, especially for quantitative genetic analyses of data from wild animal populations and evolutionary biology; see, Kruuk (2004) for a review.

Apart from desirable statistical properties of REML estimators (Harville, 1977), a key attraction of the animal model is that it can utilise information from all types of covariances between relatives in complex pedigrees simultaneously and combine them in an optimal way. If covariances with appropriate expectations are available, it readily allows additional random effects – genetic or environmental – to be fitted and the corresponding covariance components to be estimated. Common examples are maternal additive genetic and permanent environmental effects. These are routinely considered, for instance, in the analysis of data on early growth of meat producing livestock, but are equally relevant in other areas; see Kruuk and Hadfield (2007) or Postma and Charmantier (2007) for recent exposés. Less often considered are other additive and non-additive genetic sources of variation, such as dominance and epistasis, cytoplasmic effects, imprinting or sex-linked effects.

Mixed model methodology together with the availability of REML software tailored towards quantitative genetic analyses and modern computers make complex analyses with numerous fixed effects, covariables, unbalanced data and several variance components technically straightforward. However, reliable estimates require a data and pedigree structure which provides the necessary information to separate the observed variation into its causal components. This holds especially, when multiple random effects are fitted and the corresponding variance components are

to be separated. Prior to maximum likelihood estimation and the use of mixed models, analyses to estimate multiple variance components typically required determining a number of different types of covariances between relatives with sufficiently different expectations that, by equating covariances to their expected values, all components of interest could be uniquely determined; see Bondari *et al.* (1978) for an example. For instance, to estimate direct and maternal additive genetic variances, the direct-maternal genetic covariance and maternal permanent environmental variance, a minimum set of covariances between relatives would comprise the covariance among paternal half sibs, the covariance among full sibs, the covariance between sires and their off-spring and the covariance between dams and their offspring. This requirement remains unchanged with modern, animal model type analyses.

Hence, experimental designs suggested to estimate multiple, genetic variance usually aim at generating as many close types of relatives as possible. However, this does not provide a guarantee for estimability of all components of interest. Early literature on experimental design for the estimation of genetic parameters generally considered the expectation of mean squares in ANOVA, e.g. Robertson (1959). Later, Hill and Nicholas (1974) and Thompson (1976a) applied maximum likelihood estimation to combine information from parent-offspring regression and the covariance among sibs, and considered the resulting, optimal family structure for heritability estimation. In a maximum likelihood framework of inference, estimates of sampling variances and confidence intervals are commonly derived from the likelihood function and its derivatives. This can be utilised to assess the scope of a design to estimate all the parameters fitted and, moreover, obtain a measure of the accuracy which might be achieved. Applying REML to the designs suggested by Eisen (1967) and Bondari *et al.* (1978) to estimate maternal genetic variances, Thompson (1976b) demonstrated that the derivatives of the covariance matrix of the vector of observations with respect to some of the variance components were collinear, and that these components could thus not be estimated separately.

Recently, Fairbairn and Roff (2006) proposed a design to estimate genetic variance components due to autosomal and sex-linked genetic effects. The authors suggested an extended model, fitting maternal genetic and environmental as well as dominance effects in addition, and strongly advocated analysis via REML fitting an animal model. While they examined the expectation of covariances between relatives and showed that their design provided a 'sufficient set', no attempt was made to assess the sampling properties of resulting parameter estimates. After a brief review

of pertinent principles, this paper illustrates the use of likelihood based calculations to examine the quality of estimates that might be obtained, using the design of Fairbairn and Roff (2006) as an example.

2 Likelihood principles

Let $\boldsymbol{\theta}$, with elements θ_i, denote the vector of p parameters to be estimated and $\log \mathcal{L}(\boldsymbol{\theta})$ the logarithm of the corresponding REML likelihood function, given the model of analyis and vector of observations $\mathbf{y}$, assumed to have multivariate normal distribution. Under large sample theory, i.e. asymptotically, the REML estimate of $\boldsymbol{\theta}$ then has a normal distribution with covariance matrix given by the inverse of the Fisher (or expected) information matrix of $\boldsymbol{\theta}$ (denoted by $\mathbf{H}$ here rather than $\mathbf{I}$ to avoid confusion with an identity matrix)

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \mathbf{H}^{-1} \tag{1}$$

evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. The elements of $\mathbf{H}$ are the negatives of the expected values of the partial second derivatives of $\log \mathcal{L}(\boldsymbol{\theta})$ with respect to the elements of $\boldsymbol{\theta}$.

Hence, the diagonal elements of $\mathbf{H}^{-1}$ provide estimates of the (asymptotic) sampling variances of the $\hat{\theta}_i$. These can be used to construct confidence intervals, based on the normal distribution, or carry out hypotheses tests. For instance, based on the normal approximation, the $(1 - \alpha)$ confidence interval for parameter θ_i is simply $[\hat{\theta}_i - z_{\alpha/2}\sqrt{h^{ii}}, \hat{\theta}_i + z_{\alpha/2}\sqrt{h^{ii}}]$, where h^{ii} is the $i-$th diagonal element of $\mathbf{H}^{-1}$, α denotes the error probability and z_α is the truncation point of the standard normal distribution corresponding to α.

Partition $\boldsymbol{\theta}$ into sub-vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, with lengths k and $p - k$, respectively. The *Wald test* criterion for the hypothesis that $\hat{\boldsymbol{\theta}}_1$ is equal to a value $\mathbf{t}$ is

$$\Omega = (\hat{\boldsymbol{\theta}}_1 - \mathbf{t})'(\mathbf{H}^{-1})_{11}^{-1}(\hat{\boldsymbol{\theta}}_1 - \mathbf{t}) \tag{2}$$

which is assumed to have a χ^2 distribution with k degrees of freedom, where $(\mathbf{H}^{-1})_{11}$ the sub-matrix of $\mathbf{H}^{-1}$ corresponding to $\boldsymbol{\theta}_1$.

In practice, however, the shape of the likelihood surface can deviate substantially from normality and use of large sample results can give misleading estimates of sampling errors and confidence intervals (Sprott, 1973). This can be due to limited sample sizes as well as the parameterisation chosen. Corrections for skewness of the likelihood functions have been suggested (Bartlett, 1953a,b), but it is often preferred to base inference directly on the likelihood function, in particular for problems with more than one parameter. This tends to be computationally more demanding, but with modern computing facilities, that is generally not an issue.

The *likelihood ratio test* criterion for $\hat{\boldsymbol{\theta}}_1$ equal to $\mathbf{t}$ is

$$\Lambda = -2\log\left(\mathcal{L}(\boldsymbol{\theta}_1 = \mathbf{t}|\hat{\boldsymbol{\theta}}_2)/\mathcal{L}(\hat{\boldsymbol{\theta}})\right) = -2\left(\log\mathcal{L}(\boldsymbol{\theta}_1 = \mathbf{t}|\hat{\boldsymbol{\theta}}_2) - \log\mathcal{L}(\hat{\boldsymbol{\theta}})\right) \tag{3}$$

where $\log\mathcal{L}(\boldsymbol{\theta}_1 = \mathbf{t}|\tilde{\boldsymbol{\theta}}_2)$ is the value of the log likelihood obtained by fixing $\boldsymbol{\theta}_1$ at $\mathbf{t}$ and maximising the likelihood function with the respect to the elements of $\boldsymbol{\theta}_2$. Except for cases where $\mathbf{t}$ comprises values at the boundary of the parameter space, Λ again has an asymptotic χ^2 distribution. Hence, the likelihood ratio and Wald test are asymptotically equivalent. However, when both are available, the former is generally preferred as it is less affected by deviations of the likelihood function from normality than the Wald test and as it is invariant to parameter transformations (Meeker and Escobar, 1995; Pawitan, 2000). If the hypothesis test involves parameters at the boundary, Λ is distributed according a mixture of χ^2 distributions (Self and Liang, 1987; Stram and Lee, 1994). Only one component of this mixture has degrees of freedom equal to the number of parameters tested. The other components have smaller degrees of freedom, i.e. if the boundary conditions are ignored, the resulting test tends to be too conservative; see Dominicus *et al.* (2006) or Visscher (2006) for recent treatments in a genetic context.

The value determined for the likelihood ratio test, $\log\mathcal{L}(\boldsymbol{\theta}_1 = \mathbf{t}|\tilde{\boldsymbol{\theta}}_2)$, represents a point on the *profile* likelihood surface for $\boldsymbol{\theta}_1$. The profile likelihood can be thought off as being obtained by projecting the likelihood surface for all p parameters on the sub-space for the k parameters in $\boldsymbol{\theta}_1$. This is illustrated in Figure 1. Shown are the contour lines of the likelihood surface for $p = 2$ parameters, $\theta_1 = 40$ and $\theta_2 = 60$ which represent the additive genetic and environmental variance in a paternal half-sib design comprising 450 families of size 10. Viewing this surface from the direction of either co-ordinate axis then yields the profile likelihood as the outline or 'silhouette' of the likelihood surface for both parameters projected onto the axis for each parameter. An

important property of this projection is that the curvature of the surface is preserved (Patefield, 1977). In other words, partial second derivatives of the profile likelihood function are the same as those of the likelihood function for all parameters. This implies that we can obtain estimates of sampling variances and confidence regions for a subset of parameters from the profile likelihood function.

Let $\log \mathcal{L}_P(\theta_i)$ denote the relative profile likelihood for a single parameter θ_i, i.e. the profile likelihood deviated from its maximum or, equivalently, the maximum of $\log \mathcal{L}(\boldsymbol{\theta})$. Ignoring any boundary conditions and assuming a χ^2 distribution with one degree of freedom, the confidence limits for θ_i are then obtained as the roots of $\log \mathcal{L}_P(\theta_i) - \frac{1}{2}\chi^2_{\alpha,1}$, with $\chi^2_{\alpha,1}$ the critical value for α from the χ^2 distribution with one degree of freedom. Graphically, determining the confidence interval for θ_i is as simple as drawing a line at $-\frac{1}{2}\chi^2_{\alpha,1}$ parallel to the axis for θ_i and finding the intercepts with $\log \mathcal{L}_P(\theta_i)$. This is shown in Figure 1 for $\alpha = 5\%$, i.e. $-\frac{1}{2}\chi^2_{\alpha,1} = -1.92$. A number of strategies have been suggested to find the confidence limits numerically. The simplest techniques involve evaluation of $\log \mathcal{L}_P(\theta_i)$ for a segment of interest or a grid search, but can be computationally demanding. Meyer and Hill (1992) employed a quadratic approximation of $\log \mathcal{L}_P(\theta_i)$ to obtain sampling errors of REML estimates when using a derivative-free algorithm, and a cubic or quartic approximation to determine corresponding, non-symmetric confidence intervals. Venzon and Moolgavkar (1988) described a Newton-Raphson type algorithm which is widely used, with subsequent improvements or modifications for special cases (e.g. Minkin and Venzon, 1990; Neale and Miller, 1997; Gimenez *et al.*, 2005; Virtanen and Uusipaikka, 2008).

3 Material and Methods

3.1 Experimental design

Fairbairn and Roff (2006) suggested a design comprising a number of unrelated families of the same structure, where each family involves 3 generations of animals to be recorded. An implicit assumption of the design is that the litter size and sex ratios in each litter can be chosen as required. Briefly, for each family generation 1 consists of 8 individuals (4 males and 4 females) assumed to be unrelated, which form 4 pairs of grandparents. Each pair is assumed to have 4 off-spring, resulting in 16 animals in generation 2. For pairs 1 and 2, offspring consist of 2 males

and 2 females, while all 4 offspring for litters 3 and 4 are female. Each of the 2 full-sib males (from pairs 1 and 2) is then mated to the unrelated females in the other 3 families (3 matings per sire, 12 in total). Figure 2 gives a representation of the mating scheme. This generates sets of full- and half-sibs and single- and double-first cousins in generation 3. In addition the design provides parent-offspring and grandparent-offspring as well as various types of uncle/aunt-nephew/niece type covariances.

No details for the size and composition of litter in generation 3 are postulated in Fairbairn and Roff (2006). In the following, we assume that there are 4 offspring per family in generation 3, 2 of either sex. This gives 48 animals in generation 3 and a total of 72 animals with records in each family.

3.2 Model

Let **y** denote the vector of observations for a trait of interest, with records continuous and normally distributed. Similarly, define **a**, **s** and **m** as the corresponding vectors of autosomal, sex-linked and maternal, additive genetic effects. In addition, allow for both autosomal and sex-linked dominance effects, represented by vectors **da** and **ds**, respectively. Further, let the vector of maternal, permanent environmental effects be given by **c**, and let **e** denote the vector of residuals. The resulting 'animal' model of analysis then is

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}_1(\mathbf{a} + \mathbf{s} + \mathbf{da}) + \mathbf{Z}_2\mathbf{ds} + \mathbf{W}(\mathbf{m} + \mathbf{c}) + \mathbf{e} \tag{4}$$

with $\boldsymbol{\beta}$ a vector of fixed effects, and $\mathbf{X}$ the corresponding design matrix. $\mathbf{Z}_1$ is the design matrix for direct genetic effects other than X-linked dominance effects. As there are no such effects for males which have only one copy of the gene, $\mathbf{Z}_2$, the design matrix for **ds**, is obtained by replacing all elements in rows of $\mathbf{Z}_1$ pertaining to records on males by zero. Finally, $\mathbf{W}$ is the design matrix for maternal effects.

Assume that **a** and **m** are distributed proportionally to the numerator relationship matrix $\mathbf{A}$, and let the corresponding matrix of additive genetic relationships amongst the levels of **s** be denoted by $\mathbf{S}$. Similarly, define $\mathbf{D}_\mathbf{A}$ and $\mathbf{D}_\mathbf{S}$ as the matrices of relationships between the levels of **da** and **ds**. Assume further that the random effects fitted are uncorrelated, except for the direct and maternal

genetic effects, **a** and **m**. This gives covariance of matrix, **V**, of **y**

$$\mathbf{V} = \mathbf{Z}_1 \left(\sigma^2_A \mathbf{A} + \sigma^2_S \mathbf{S} + \sigma^2_{DA} \mathbf{D_A}\right) \mathbf{Z}'_1 + \sigma^2_{DS}\, \mathbf{Z}_2 \mathbf{D_S} \mathbf{Z}'_2 + \sigma_{AM} \left(\mathbf{Z}_1 \mathbf{A} \mathbf{W}' + \mathbf{W} \mathbf{A} \mathbf{Z}'_1\right) + \mathbf{W} \left(\sigma^2_M \mathbf{A} + \sigma^2_C \mathbf{I}\right) \mathbf{W}' + \sigma^2_E \mathbf{I}$$

with **I** an identity matrix, σ^2_A, σ^2_S, σ^2_M, σ^2_C, σ^2_{DA}, σ^2_{DS} and σ^2_E the variance components due to **a**, **s**, **m**, **c**, **da**, **ds** and **e**, respectively, and σ_{AM} the direct-maternal genetic covariance.

3.3 Maximum likelihood estimation

The REML log likelihood pertaining to (Eq. 4) is (e.g. Harville, 1977)

$$\log \mathcal{L}(\boldsymbol{\theta}) = const. - 1/2 \left(\log |\mathbf{V}| - \mathrm{tr}(\mathbf{X}'_+ \mathbf{V}^{-1} \mathbf{X}_+) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \quad (5)$$

where $\mathbf{X}_+$ denotes a full rank submatrix of **X**. In some cases, **V** can be partitioned into a number of independent matrices of sums of squares and cross-products. We can then evaluate $\log \mathcal{L}(\boldsymbol{\theta})$ by summing over the independent matrices, and obtain REML estimates by maximizing the likelihood of these matrices (Thompson, 1976b). If families are unrelated and have the same structure, **V** is block-diagonal with identical blocks $\mathbf{V}_0$ of size $N \times N$, where N is the number of observations per family (assuming **y** is ordered according to individuals within families). This gives $\mathbf{V} = \mathbf{I}_F \otimes \mathbf{V}_0$, with F denoting the number of families and $\otimes$ the direct (Kronecker) matrix product (Harville, 1997). In the absence of fixed effects, other than an overall mean, (Eq. 5) can be rewritten as

$$\log \mathcal{L}(\boldsymbol{\theta}) = const. - 1/2\, d \left(\log |\mathbf{V}_0| + \mathrm{tr}(\mathbf{V}_0^{-1} \mathbf{M})\right) \quad (6)$$

Here $d = F - 1$ denotes the degrees of freedom and **M** is the $N \times N$ matrix of mean squares and cross-products amongst records of family members, accumulated over families. The expected information matrix is

$$\mathbf{H} = -1/2\, d \left\{\mathrm{tr}(\mathbf{V}_0^{-1} (\partial \mathbf{V}_0 / \partial \theta_i) \mathbf{V}_0^{-1} (\partial \mathbf{V}_0 / \partial \theta_j))\right\} \quad (7)$$

with θ_i the $i-$th parameter to be estimated. If $\mathbf{V}$ is linear in the parameters to be estimated (i.e. $\mathbf{V} = \sum_i (\partial \mathbf{V}/\partial \theta_i)\, \theta_i$), maximum likelihood estimates of $\boldsymbol{\theta}$ can be obtained as iterative solutions to

$$\mathbf{H}\,\hat{\boldsymbol{\theta}} = \mathbf{g} \qquad \text{with} \quad \mathbf{g} = -1/2\, d\, \left\{ \mathrm{tr}(\mathbf{V}_0^{-1}(\partial \mathbf{V}_0/\partial \theta_i)\mathbf{V}_0^{-1}\mathbf{M}) \right\} \tag{8}$$

3.4 Calculations

Sampling properties of variance component estimates for the design described above were examined for population values of $\sigma^2_A = 400$, $\sigma^2_S = 100$, $\sigma^2_M = 120$, $\sigma_{AM} = -30$, $\sigma^2_C = 150$, $\sigma^2_{DA} = 60$, $\sigma^2_{DS} = 20$ and $\sigma^2_E = 600$. Values imply moderate heritabilities for autosomal additive genetic effects (0.30 for males and 0.28 for females). Maternal effects are assumed to explain approximately 10% of the phenotypic variation, and dominance effects are assumed to be small, amounting to 4.2% of the total variation in females for autosomal and 1.4% for X-linked effects. Sizes of experiment considered were $F = 50, 100, 200$ and 1000 independent families.

Matrices of additive genetic relationships for a family, $\mathbf{A}_0$ and $\mathbf{S}_0$, were set up directly from the pedigree generated by the mating scheme, as described by Henderson (1976) and Fernando and Grossman (1990), respectively. Elements of $\mathbf{D}_{\mathbf{A}\,0}$, i.e. the coefficients of fraternity, were determined as $\delta_{ij} = 0.25(a_{km}a_{ln} + a_{kn}a_{lm})$ for individual i with parents k and l and individual j with parents m and n and a_{ij} the $ij-$th element of $\mathbf{A}_0$ (Lynch and Walsh, 1998). Finally, the non-zero elements of $\mathbf{D}_{\mathbf{S}\,0}$ were obtained from the expectations of covariances amongst female relatives given by Fairbairn and Roff (2006, Table 5).

For each scenario, asymptotic sampling errors, 95% confidence intervals and sampling correlations were obtained from the information matrix. The profile log likelihood for each variance component was evaluated for a range of values in steps of 1.0. This allowed the corresponding 95% confidence interval to be determined simply by finding the closest pairs of values (either side of the maximum of $\log \mathcal{L}_P(\theta_i)$) which encompassed the value of $-1/2\chi^2 = -1.92$, and, if necessary, interpolating between the members of the pair. In addition, empirical means, standard deviations and confidence intervals were obtained from simulation. For $F = 50$, observations for each family were sampled from a multivariate normal distribution $\mathrm{N}(\mathbf{0}, \mathbf{V}_0)$, accumulating means and the matrix of sums of squares and crossproducts among family members to calculate the matrix of mean squares and products $\mathbf{M}$. For F larger than the dimension of $\mathbf{V}_0$, $\mathbf{M}$ was sampled

directly from a central Wishart distribution as described by Odell and Feiveson (1966). REML estimates of variance components were obtained constraining variance components to an operational zero of 10^{-8} and the direct-maternal genetic correlation to have absolute value of $1 - 10^{-8}$. Estimation was carried out employing a Method of Scoring algorithm as described above (Eq. 8). If this did not yield a higher $\log \mathcal{L}(\boldsymbol{\theta})$ in an iterate, the change in $\boldsymbol{\theta}$ given by (Eq. 8) was successively halved until $\log \mathcal{L}(\boldsymbol{\theta})$ was increased or the scale factor for the step size was less than 0.01. In addition, derivative-free optimisation steps were employed to deal with any problem cases and to check for convergence. A total of 50 000 replicates were performed for each case. After sorting estimates obtained in ascending numerical order, limits of empirical 95% confidence regions for each component were determined as midpoints between the values separating the top 2.5% and bottom 2.5% of estimates from the remainder. For example, the lower bound for a parameter was the average of the 1250−th and the 1251−th estimate.

The full model with 8 (co)variance components was contrasted to reduced models assuming either maternal ($\sigma^2_M = \sigma_{AM} = \sigma^2_C = 0$) or dominance ($\sigma^2_{DA} = \sigma^2_{DS} = 0$) were absent. In addition, a 'wrong' reduced model ignoring dominance effects when true values were non-zero was examined. Furthermore, we investigated the scope of improving estimates for the full model through embryo transfer by splitting full-sib families in generations 2 and 3, so that each litter had as many 'rearing' dams as possible. For each family in generation 2, only one member of each litter was assumed to be raised by its genetic dam, with the remaining 3 individuals exchanged for a member of each of the other 3 families. In generation 3, 8 families were similarly split, while the remaining 4 families were assumed to have a pair-wise exchange (see Figure 3).

4 Results

4.1 Means and standard deviations

Means and standard deviations (SDEV) across replicates from the simulation are summarised in Table 1. Constraining estimates of variance components to the parameter space yields biassed estimates. Biasses decrease as the number of families increases, i.e. as sampling variances decrease and fewer samples need to be constrained. In particular, estimates for the numerically small components σ^2_{DA} and σ^2_{DS} are biassed upwards due to enforcing non-negative values. In turn, this

causes other estimates to be biassed downwards, as the total amount of variation to be partitioned in the analysis is limited. These 'trade-offs' depend on sampling correlations. For the full model, biasses for the design of Fairbairn and Roff (2006) are substantial, even for a large experiment. Specifically, estimates of σ^2_{DA} are biassed upwards, while σ^2_C and σ^2_E are correspondingly biassed downwards.

As shown above (Eq. 7), expected SDEVs derived from the information matrix, are inverse proportional to $\sqrt{F-1}$. Hence only values for $F = 1000$ are shown in Table 1 - values for other F can be obtained by appropriate scaling. Empirical SDEVs are given as relative deviations from the theoretical values. For instance, for the full model and $F = 50$, the empirical SDEV for σ^2_C is 68.04. The theoretical value for $F = 50$ is $24.8 \times \sqrt{999/49} = 112.0$. This gives a relative deviation of $(112.0 - 68.0)/112.0 = -0.393$ or -39.3% as shown in Table 1. Variance components estimated with bias also show substantially lower empirical than theoretical SDEVs. As for empirical means, deviations decrease with increasing numbers of families and are less prominent for analyses with reduced numbers of parameters, i.e. again reflect the effect of constraints imposed on the parameter space.

4.2 Sampling correlations

Theoretical values for sampling correlations between parameter estimates are independent of the size of the experiment, i.e. depend only on the family relationship structure or, more specifically, the differences between the matrices of derivatives $\partial \mathbf{V}_0/\partial\theta_i$. Table 2 contrasts selected empirical and theoretical sampling correlations, omitting any pairs of parameters for which the absolute value of the correlation is less than 0.2 in all instances. Overall there is good agreement between theoretical and empirical values. Differences between them are again most pronounced for small family sizes and components with large theoretical standard deviations, indicating that constraining estimates modified sampling correlations for these parameters. In most cases, this involves a reduction in strength of correlation (absolute value). As observed in previous studies, there are moderate to strong sampling correlations between the direct and maternal genetic covariance components σ^2_A, σ^2_M and σ_{AM}, which are substantially reduced with embryo transfer (Meyer, 1992).

More strikingly, for the full model of Fairbairn and Roff (2006), there are very strong sampling

correlations between σ^2_C, σ^2_{DA} and σ^2_E. A negative correlation of -0.90 (theoretical) value between σ^2_C and σ^2_{DA} implies that the sum of the two components can be estimated accurately, but that is is difficult to disentangle the individual components. With σ^2_{DA} close to the boundary of the parameter space and subject to large sampling variation, constraints on σ^2_{DA} then give rise to the substantial downwards bias observed for σ^2_C. Similar arguments explain the bias in σ^2_E, which has theoretical sampling correlations of 0.90 with σ^2_C and -0.99 with σ^2_{DA}. Consequently, $\mathbf{H}$ has two eigenvalues amounting to less than 1% of the sum of all eigenvalues. With the smallest eigenvalue 0.03% of total, the information matrix is close to not being positive definite.

Such strong sampling correlations are not surprising. Figure 3 gives the coefficients in the expectation of the covariance matrix among the members of a family for σ^2_{DA}, σ^2_{DS} and σ^2_C. For our scenario, the design matrix ($\mathbf{Z}_1$) for **da** and **c** is an identity matrix, i.e. these also represent the corresponding matrices of partial derivatives of $\mathbf{V}_0$. As shown, only full sibs contribute to the estimation of σ^2_C. Similarly, with a coefficient of $1/4$ in the expectation of the covariance between full-sibs, most information to estimate σ^2_{DA} comes from this source. While covariances between double first cousins contribute, the corresponding coefficient is only $1/16$. 'Embryo transfer' breaks up this close link between σ^2_C and σ^2_{DA} (c.f. Figure 3) and, for our example with as much embryo transfer as possible, reduces the sampling correlations between σ^2_C and σ^2_{DA} or σ^2_E to values close to zero. Though estimation of σ^2_{DS} again relies solely on the covariances between full-sibs and double first cousins, this component is only expressed for females. Hence, with a substantial number of records on males, derivatives $\partial\mathbf{V}_0/\partial\sigma^2_{DS}$ and $\partial\mathbf{V}_0/\partial\sigma^2_{DA}$ or $\partial\mathbf{V}_0/\partial\sigma^2_C$ are sufficiently different for sampling correlations and SDEVs for σ^2_{DS} to be much lower than for σ^2_{DA}.

4.3 Profile likelihood

Profile log likelihood functions for individual components, summarised in Figure 4 identify the parameters for which accurate estimation in the full model is problematic. Profiles for σ^2_C, σ^2_{DA} and σ^2_E are non-symmetric, very flat in parts and clearly non-normal, emphasizing that large sample theory may not be applicable in this case. In the absence of maternal effects, $\log \mathcal{L}_P(\theta_i)$ for σ^2_{DA} is approximately normal and the corresponding confidence interval is narrow enough to suggest that σ^2_{DA} may be estimated with reasonable accuracy. Conversely, in the absence of dominance effects, estimability of σ^2_C improves dramatically. Either scenario ensures an accurate estimate of

σ^2_E, but not estimating dominance variances only slightly improves the accuracy of estimation for the other components.

Confidence intervals derived from the profile log likelihood together with their counterparts from the normal approximation and simulation are shown in Figure 5. 'Means' for the likelihood derived intervals are the maximum likelihood estimates obtained when constructing the matrix of mean squares (**M**) for the population values of the variance components, i.e. do not reflect the effects of constraining estimates which are outside the parameter space due to sampling. Truncating theoretical, large sample intervals at zero there is good agreement between all three measures for σ^2_{DA}, σ^2_S, σ^2_M, σ_{AM} and σ^2_{DS}. For the dominance variances, σ^2_{DA} and σ^2_{DS}, with relatively low population values, the design does not allow estimates significantly different from zero (at $\alpha = 5\%$) to be identified when fitting the full model. Fitting the wrong model by assuming dominance variances to be zero, estimates of σ^2_E are biassed upwards by almost 10%, while confidence intervals are dramatically reduced. Hence, the square root of the mean square error (RMSE) for σ^2_E from the simulation dropped from 124.4 in the full analysis to 24.0 for the 'wrong' model. Similarly, the estimate of σ^2_C was biassed upwards, but the RMSE was reduced from 43.5 to 24.1, while values for the other components were little affected.

5 Discussion

Likelihood based calculations have been shown to be useful in evaluating experimental designs. Even measures derived from the information matrix which invoke large sample theory can identify potential areas of weakness: If confidence intervals derived from the diagonal elements of the information matrix exceed the boundaries of the parameter space, the corresponding parameter are likely to be affected by constraints applied during estimation and biassed. Eigenvalues of the information matrix amounting to a very small proportion of their total indicate overparameterisation of the model and very strong sampling correlations.

The impact on sampling distribution of individual parameters can be assessed by inspecting the corresponding profile likelihood functions. Shapes deviating from the normal distribution signal scenarios for which large sample theory is not applicable (Sprott, 1980). For these cases, confidence intervals tend to be markedly asymmetrical. As illustrated for σ^2_C and σ^2_E (Figure 5),

confidence regions derived from the profile likelihood function then show good agreement with empirical results.

Ignoring any fixed effects and assuming no missing observations, calculations shown clearly represent a 'best possible' scenario. However, this should suffice for the purpose of comparing designs. Calculations shown are straightforward and computationally undemanding, as they involve manipulation of matrices of size equal to the number of records in a family only. They are readily performed for a range of hypothetical population values for the parameters to be estimated, and can be recommended as part of the design phase of any experiment to facilitate estimation of genetic parameters. The methodology presented is readily extensible to address other questions, which are often examined by simulation only, e.g. the impact of pedigree errors or sources of (co)variation which have been ignored. The 'animal model' and availability of corresponding software, in particular for REML estimation, make estimation of multiple genetic (co)variance components conceptually easy. However, one of the undesirable properties of REML is that it can provide 'estimates' for parameters for which little information is available. Inspection of the information matrix and profile likelihoods may help identifying such instances, which are less readily found by inspection of the expectation of covariances between different types of relatives alone.

Results show that the scheme of Fairbairn and Roff (2006) is excellent for the prime purpose it has been designed for, i.e. to disentangle additive genetic variances due to X-linked and autosomal effects. In the absence of maternal effects, it is also well suited to the estimation of the corresponding variances due to dominance effects, though a relatively large experiment may be required to ascertain estimates different from zero.

Estimation of maternal effects and the corresponding (co)variance components has long been recognised as being inherently problematic (Willham, 1980), with the impact of the data structure on estimates from animal model analyses investigated repeatedly (e.g. Gerstmayr, 1992; Clèment *et al.*, 2001; Maniatis and Pollott, 2003). If estimation of both (autosomal) dominance and maternal permanent environmental variances is necessary, special care must be taken to ensure the data and pedigree structure allows these components to be separated. What is feasible depends on the species concerned. As illustrated above, 'embryo transfer' has the potential to reduce problematic sampling correlations, but is a technically demanding option. A simpler measure would be to augment the experimental design so as to include maternal half-sibs. Alternatively, covariances

between relatives in different generation with a non-zero component of σ^2_{DA} could be generated through assortative mating or parent/offspring type matings to produce inbred animals. However, this might necessitate the estimation of additional parameters such as the genetic correlation between mates (Lynch and Walsh, 1998) or to accommodate the effects of inbreeding on dominance (co)variances (Smith and Mäki-Tanila, 1990).

6 Conclusions

Separating variances between relatives into causal components due to different modes of gene action is inherently difficult. Examination of the information matrix in REML analyses and the profile likelihood functions for individual parameters can be illuminating and aid in the validation of experimental design. The calculations involved are straightforward and computationally undemanding. Examination of the likelihood function and its derivatives can thus be recommended as part of the design stage of quantitative genetic experiments.

Acknowledgments

This work was supported by grant BFGEN.100 B of Meat and Livestock Australia Ltd (MLA).

References

Bartlett M S (1953a). Approximate confidence intervals. *Biometrika* **40**: 12–19.

Bartlett M S (1953b). Approximate confidence intervals. II. More than one unknown parameter. *Biometrika* **40**: 306–317.

Bondari K, Willham R L, Freeman A E (1978). Estimates of direct and maternal genetic correlations for pupa weight and family size of Tribolium. *J Anim Sci* **47**: 358–365.

Clèment V, Bibé B, Verrier E, Elsen J M, Manfredi E, Bouix J, , Hanocq E (2001). Simulation analysis to test the influence of model adequacy and data structure on the estimation of genetic parameters for traits with direct and maternal effects. *Genet Select Evol* **33**: 369–395. doi: DOI: 10.1051/gse:2001123.

Dominicus A, Skrondal A, Gjessing H, Pedersen N, Palmgren J (2006). Likelihood ratio tests in behavioral genetics: Problems and solutions. *Behav Genet* **36**: 331–340.

Eisen E J (1967). Mating designs for estimating direct and maternal genetic variances and direct-maternal covariances. *Can J Genet Cytol* **9**: 13–22.

Fairbairn D J, Roff D A (2006). The quantitative genetics of sexual dimorphism : assessing the importance of sex-linkage. *Heredity* **97**: 319–328. doi: 10.1038/sj.hdy.6800895.

Fernando R L, Grossman M (1990). Genetic evaluation with autosomal and X-chromosomal inheritance. *Theor Appl Genet* **80**: 75–80. doi: 10.1007/BF00224018.

Fry J D (1992). The mixed-model analysis of variance applied to quantitative genetics: biological meaning of the parameters. *Evolution* **46**: 540–550.

Gerstmayr S (1992). Impact of the data structure on the reliability of the estimated genetic parameters in an animal model with maternal effects. *J Anim Breed Genet* **109**: 321–336.

Gimenez O, Choquet R, Lamor L, Scofield P, Fletcher D, Lebreton J D, Pradel R (2005). Efficient profile–likelihood confidence intervals for capture–recapture models. *J Agric Biol Env Stat* **10**: 184–196. doi: 10.1198/108571105X46462.

Harville D A (1977). Maximum likelihood approaches to variance component estimation and related problems. *J Amer Stat Ass* **72**: 320–338.

Harville D A (1997). *Matrix Algebra from a Statistician's Perspective*. Springer Verlag.

Henderson C R (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**: 69–83.

Hill W G, Nicholas F W (1974). Estimation of heritability by both regression of offspring on parent and intra-class correlation of sibs in one experiment. *Biometrics* **30**: 447–468.

Kruuk L E B (2004). Estimating genetic parameters in natural populations using the 'animal model'. *Phil Trans R Soc B* **359**: 873–890.

Kruuk L E B, Hadfield J D (2007). How to separate genetic and environmental causes of similarity between relatives. *J Evol Biol* **20**: 1890–1903. doi: 10.1111/j.1420-9101.2007.01377.x.

Lynch M, Walsh J B (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Assoc.s, Inc., Sunderland, MA.

Maniatis N, Pollott G E (2003). The impact of data structure on genetic (co)variance components of early growth in sheep, estimated using an animal model with maternal effects. *J Anim Sci* **81**: 101–108.

Meeker W Q, Escobar L A (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *Amer Stat* **49**: 48–53.

Meyer K (1989). Estimation of genetic parameters. In: Hill W G, McKay T F M (Eds.) *Evolution and Animal Breeding.*, Reviews on Molecular and Quantitative Approaches in Honour of A. Robertson, CAB International. pp. 161–167.

Meyer K (1992). Bias and sampling covariances of estimates of variance components due to maternal effects. *Genet Select Evol* **24**: 487–509.

Meyer K, Hill W G (1992). Approximation of sampling variances and confidence intervals for maximum likelihood estimates of variance components. *J Anim Breed Genet* **109**: 264–280.

Minkin S, Venzon D J (1990). Profile-likelihood-based confidence intervals. Reply. *Appl Stat* **39**: 125–127.

Neale M C, Miller M B (1997). The use of likelihood-based confidence intervals in genetic models. *Behav Genet* **27**: 113–120. doi: 10.1023/A:1025681223921.

Odell P L, Feiveson A H (1966). A numerical procedure to generate a sample covariance matrix. *J Amer Stat Ass* **61**: 199–203.

Patefield W M (1977). On the maximized likelihood function. *Sankhyā Ser B* **39**: 92–96.

Patterson L D, Thompson R (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545–554.

Pawitan Y (2000). A reminder of the fallibility of the Wald statistic: Likelihood explanation. *Amer Stat* **54**: 54–56.

Postma E, Charmantier A (2007). What 'animal models' can and cannot tell ornithologists about the genetics of wild populations. *J Ornith* **140**: 633–642. doi: 10.1007/s10336-007-0191-8.

Robertson A (1959). Experimental design in the evaluation of genetic parameters. *Biometrics* **15**: 219–226.

Self S G, Liang K Y (1987). Asymptotic properties of the maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Amer Stat Ass* **82**: 605–610.

Shaw R G (1987). Maximum-likelihood approaches applied to quantitative genetics of natural populations. *Evolution* **41**: 812–826.

Smith S P, Mäki-Tanila A (1990). Genotypic covariance matrices and their inverses for models allowing dominance and inbreeding. *Genet Select Evol* **22**: 65–91. doi: 10.1051/gse:19900106.

Sprott D A (1973). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika* **60**: 457–465. doi: 10.1093/biomet/60.3.457.

Sprott D A (1980). Maximum likelihood in small samples: Estimation in the presence of nuisance parameters. *Biometrika* **67**: 515–523. doi: 10.1093/biomet/67.3.515.

Stram D A, Lee J W (1994). Variance component testing in the longitudinal model. *Biometrics* **50**: 1171–1177.

Thompson R (1976a). Design of experiments to estimate heritability when observations are available on parents and offspring. *Biometrics* **32**: 283–304.

Thompson R (1976b). The estimation of maternal genetic variance. *Biometrics* **32**: 903–917.

Venzon D J, Moolgavkar S H (1988). A method for computing profile-likelihood-based confidence intervals. *Appl Stat* **37**: 87–94.

Virtanen A, Uusipaikka E (2008). Computation of profile likelihood-based confidence intervals for reference limits with covariates. *Stat Med* **00**: 000–000 (in press). doi: 10.1002/sim.3000.

Visscher P M (2006). A note on the asymptotic distribution of likelihood ratio tests to test variance components. *Twin Res Hum Genet* **9**: 490–495.

Willham R L (1980). Problems in estimating maternal effects. *Livest Prod Sci* **7**: 405–418.

Table 1. Means (as deviation from population values) and empirical standard deviations (as relative deviation from theoretical values, ×100) for estimates of variance components from simulation, for experiments comprising 50, 100, 200 or 1000 families.

	Full model								No maternal				
	σ^2_A	σ^2_S	σ^2_M	σ_{AM}	σ^2_C	σ^2_{DA}	σ^2_{DS}	σ^2_E	σ^2_A	σ^2_S	σ^2_{DA}	σ^2_{DS}	σ^2_E
						Means							
50	0.2	-3.0	-3.7	2.3	-30.1	125.1	12.3	-98.0	1.3	-2.1	2.3	7.5	-5.9
100	-0.0	-1.8	-3.5	2.0	-21.7	93.0	6.6	-71.8	0.9	-1.2	-0.2	3.8	-1.8
200	-0.1	-0.7	-2.5	1.2	-14.2	62.1	3.0	-47.4	0.4	-0.5	-0.6	1.8	-0.3
1000	-0.1	0.0	-0.6	0.2	-3.9	17.3	0.1	-13.0	0.0	-0.0	0.0	0.0	-0.0
						Standard deviations							
50	-0.5	-3.4	-7.2	-4.0	-39.3	-47.2	-29.9	-45.8	-1.0	-2.4	-18.6	-27.9	-11.1
100	0.3	-1.8	-2.2	-0.5	-30.8	-38.2	-25.2	-37.0	-0.5	-1.5	-11.2	-21.9	-6.2
200	0.1	-1.2	-1.5	-0.8	-26.0	-32.8	-19.6	-31.8	-0.4	-0.4	-4.6	-14.6	-2.0
1000	-0.0	-0.3	-0.9	-0.3	-17.9	-22.7	-4.2	-22.1	0.6	0.1	-0.1	-1.2	0.1
E(SD)[a]	19.3	11.1	16.6	12.4	24.8	101.0	12.0	76.0	12.2	10.0	15.2	9.3	14.3

	'Embryo transfer'								No dominance					
	σ^2_A	σ^2_S	σ^2_M	σ_{AM}	σ^2_C	σ^2_{DA}	σ^2_{DS}	σ^2_E	σ^2_A	σ^2_S	σ^2_M	σ_{AM}	σ^2_C	σ^2_E
						Means								
50	0.5	-2.0	0.0	-0.0	-0.2	4.4	9.0	-8.1	0.8	-0.1	1.1	-0.7	-0.7	-0.2
100	0.9	-1.2	-0.1	0.2	0.0	0.5	5.1	-3.1	0.1	-0.1	0.0	0.0	-0.0	-0.0
200	0.3	-0.7	0.0	0.0	-0.0	-0.9	2.5	-0.4	0.1	-0.1	0.3	0.0	-0.2	0.0
1000	0.1	-0.1	-0.0	0.0	-0.0	-0.2	0.1	0.1	-0.1	-0.1	-0.1	0.1	0.1	0.1
						Standard deviations								
50	-1.0	-2.3	0.4	0.7	0.1	-21.8	-30.7	-12.8	0.1	0.1	-2.7	-2.1	-1.7	-0.3
100	-0.6	-1.2	0.4	0.5	0.1	-14.5	-24.3	-7.7	0.1	-0.1	-0.7	-0.4	-0.6	-0.0
200	-0.1	-0.8	-0.1	-0.2	0.1	-7.1	-17.1	-3.7	0.1	-0.3	0.2	0.3	0.3	0.2
1000	-0.1	-0.7	0.2	-0.2	0.4	-0.5	-3.1	-0.2	-0.7	1.2	-0.2	-1.0	-1.3	-0.9
E(SD)	14.0	10.7	8.6	6.2	8.2	17.6	10.7	16.0	18.5	10.1	15.5	11.6	10.1	10.1

[a]Expected value of standard deviation (from information matrix) for 1000 families

Table 2. Selected empirical sampling correlations (×100) between estimates of variance components for experiments comprising 50, 100, 200 or 1000 families, together with their expected values (EXP) from the information matrix.

		Full model					No maternal eff.		No dominance eff.		'Embryo transfer'	
		50	100	200	1000	EXP	100	EXP	100	EXP	100	EXP
σ^2_A	σ^2_S	-28.1	-27.2	-28.1	-27.7	-27.9	-55.4	-56.1	-28.9	-28.6	-53.6	-54.2
σ^2_A	σ^2_M	25.1	28.8	27.2	27.8	28.2	–	–	30.0	28.4	0.6	-1.0
σ^2_A	σ_{AM}	-61.7	-63.3	-62.4	-62.4	-62.5	–	–	-64.8	-63.5	-20.8	-19.8
σ^2_A	σ^2_C	-9.3	-11.1	-9.6	-8.5	-5.9	–	–	-22.6	-19.7	-1.5	-0.4
σ^2_A	σ^2_E	-17.8	-15.7	-14.3	-11.9	-8.1	-22.2	-18.3	-79.0	-79.5	-22.4	-17.1
σ^2_S	σ^2_{DS}	-21.0	-21.7	-25.2	-28.9	-30.2	-20.1	-25.7	–	–	-20.6	-26.0
σ^2_M	σ_{AM}	-62.3	-65.0	-65.3	-65.2	-65.4	–	–	-67.3	-65.0	-23.0	-22.7
σ^2_M	σ^2_C	-38.9	-35.8	-33.3	-27.5	-13.8	–	–	-80.8	-80.5	-73.0	-72.8
σ^2_M	σ^2_{DA}	-16.7	-16.1	-15.1	-16.5	-21.8	–	–	–	–	4.4	4.0
σ^2_M	σ^2_E	10.5	10.0	10.0	11.9	18.2	–	–	-27.1	-25.8	-3.5	-3.2
σ_{AM}	σ^2_C	16.1	16.7	15.4	13.2	6.5	–	–	42.7	39.1	10.2	9.4
σ_{AM}	σ^2_E	7.3	5.3	4.0	2.7	-2.4	–	–	60.8	60.5	9.2	10.1
σ^2_C	σ^2_{DA}	-73.7	-78.5	-81.2	-84.6	-90.0	–	–	–	–	-7.3	-8.8
σ^2_C	σ^2_E	72.2	78.0	80.4	84.2	89.8	–	–	9.3	7.5	-2.1	-0.5
σ^2_{DA}	σ^2_{DS}	-3.7	-4.0	-4.4	-7.3	-8.8	-23.4	-34.6	–	–	-21.9	-36.0
σ^2_{DA}	σ^2_E	-96.4	-97.2	-97.7	-98.2	-98.9	-82.4	-82.6	–	–	-79.1	-80.9

Figure 1. Contour levels of the log likelihood function for two parameters (θ_1 and θ_2), together with the profile log likelihood functions for each parameter

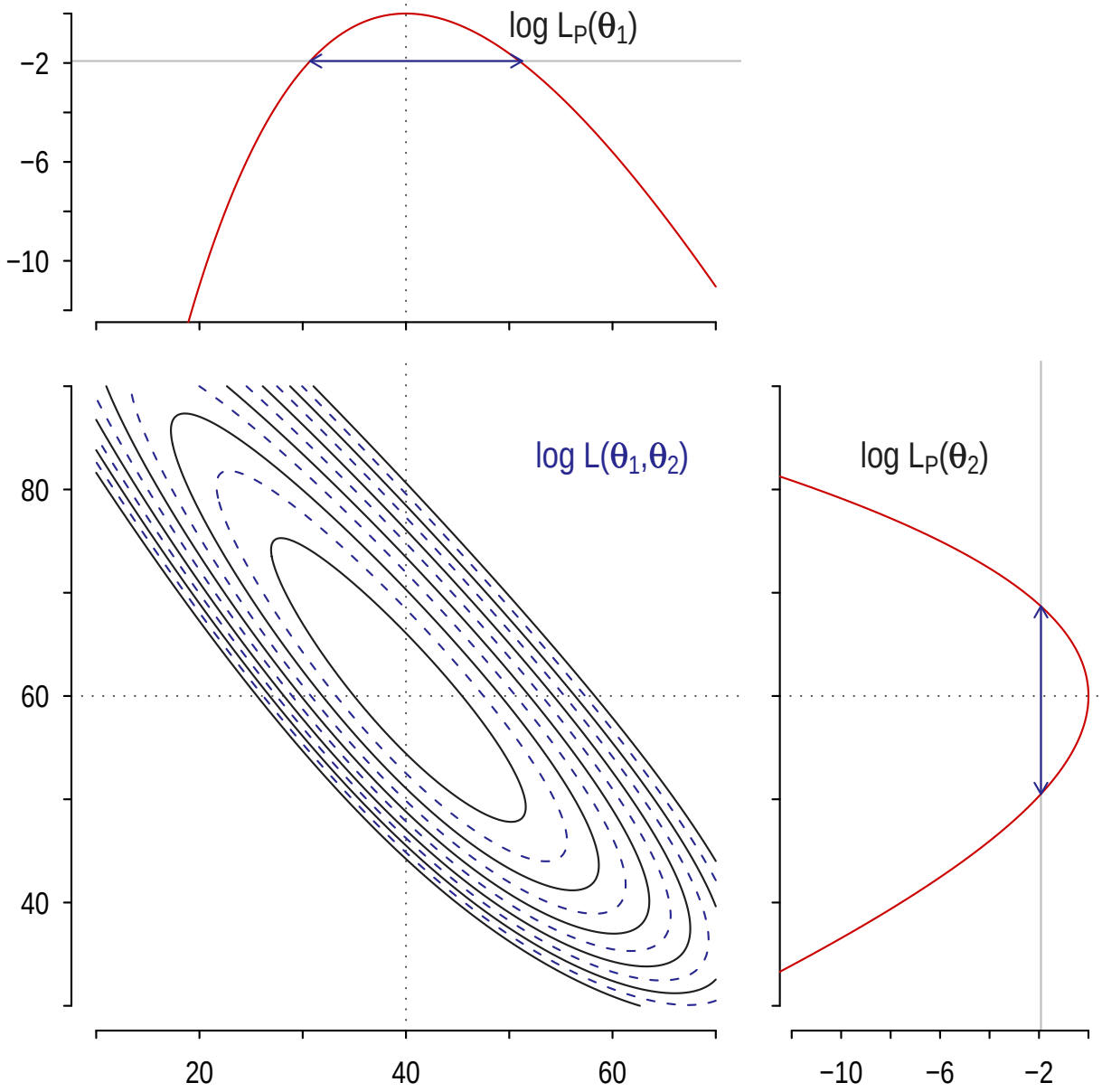

Figure 2. Schematic representation of experimental design (GS: grand-sire, GD: grand-dam, S: sire, D: dam, M: mating to produce generation 3; M4 to M11 omitted)

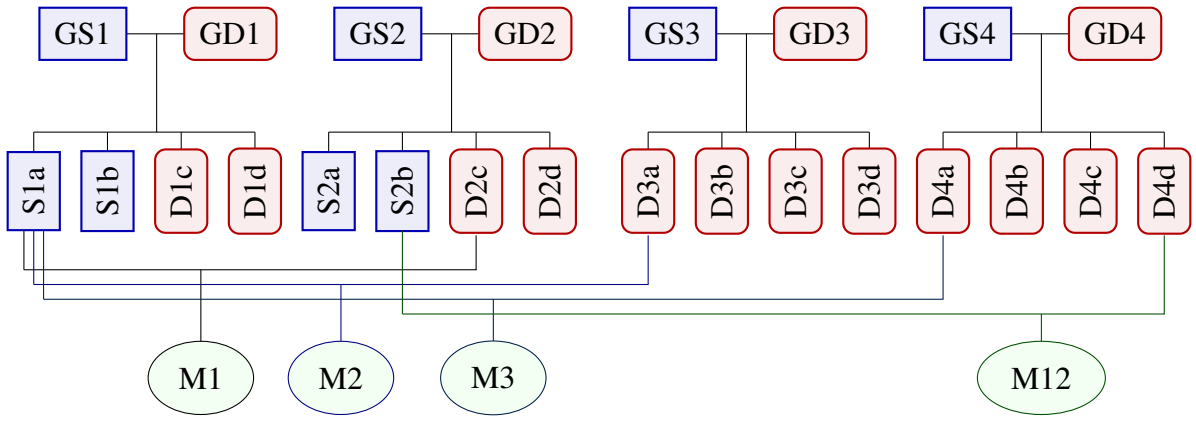

Figure 3. Coefficients of variance components in the expectation of covariances between members of a family: (a) autosomal dominance variance, (b) sex-linked dominance variance, and (c) and (d) maternal, permanent environmental variance with (c) the structure for the 'standard design' and (d) the modified design with 'embryo transfer'

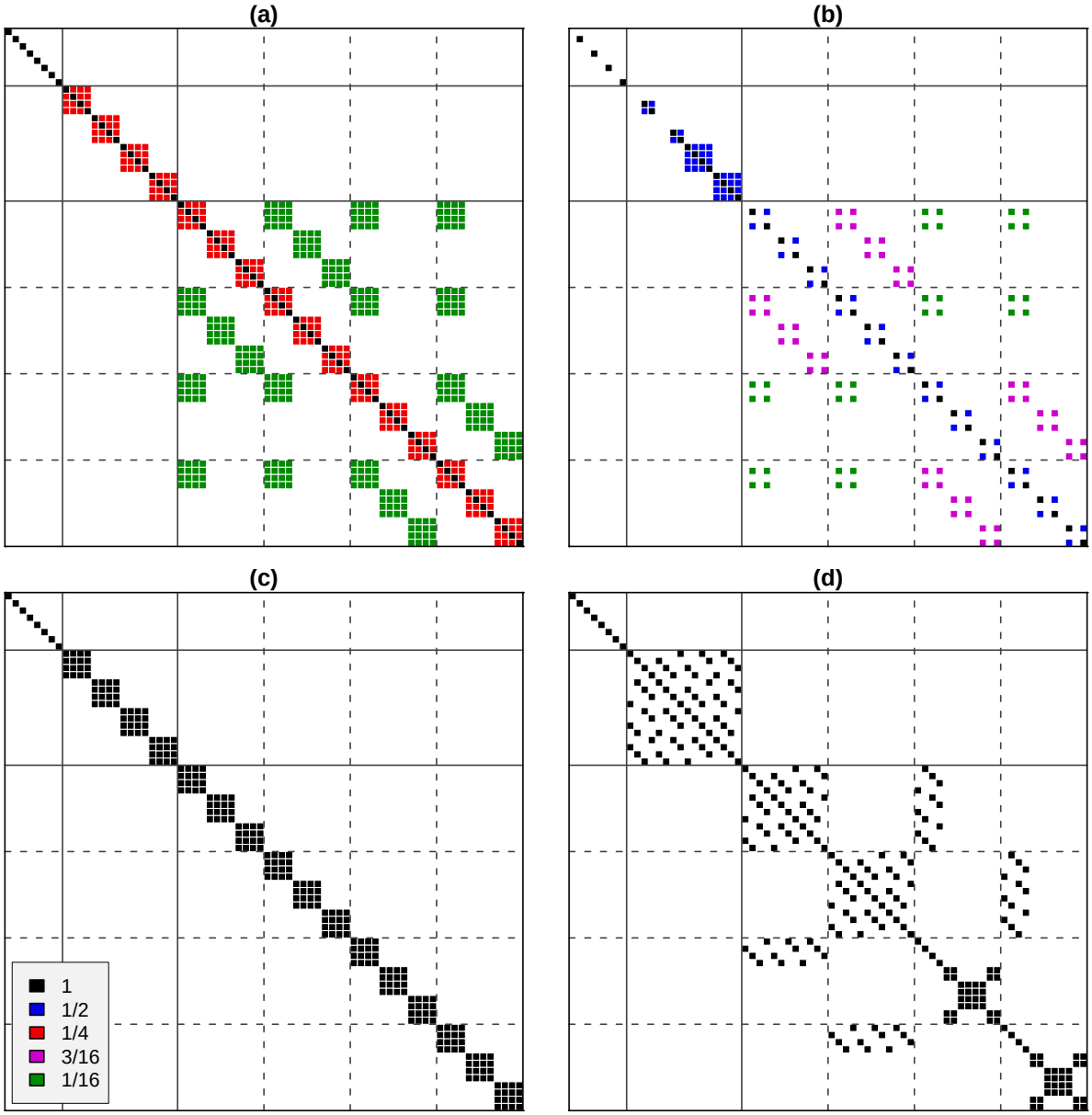

Figure 4. Relative profile log likelihood for individual variance components for an experiment with 200 families
(—— full model, – – – no maternal effects and –·–· no dominance effects)

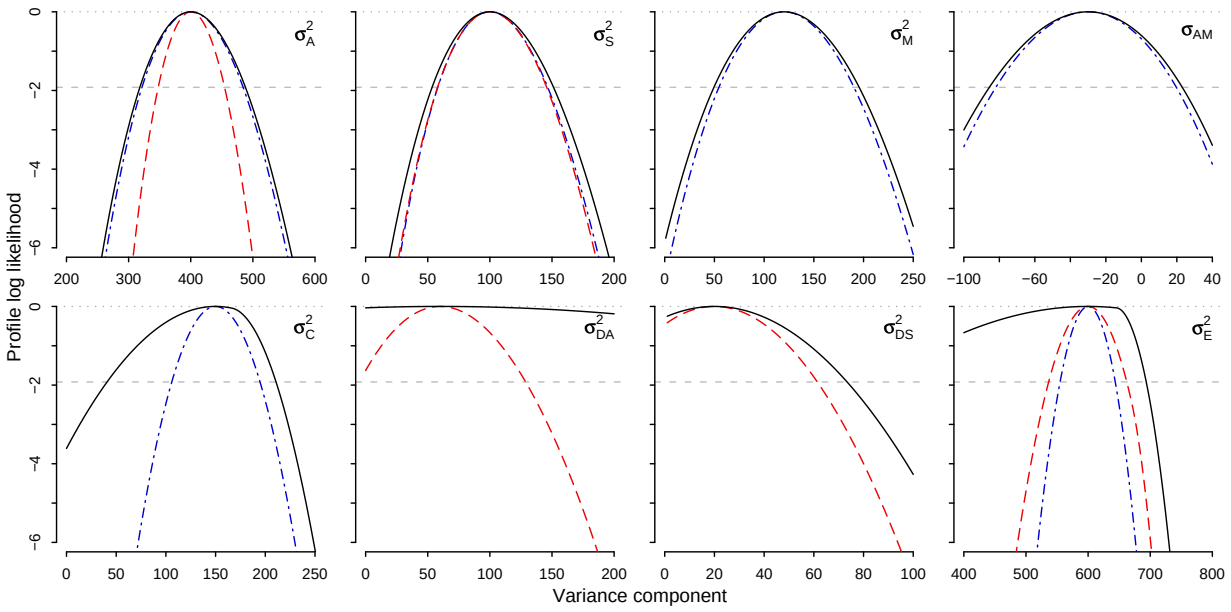

Figure 5. 'Means' and confidence intervals for individual variance components from the information matrix (○), the profile likelihood (△) and simulation(▽) for the full model (open symbols) and a reduced model (filled symbols) ignoring the non-zero dominance components (200 families)