Estimation of quantitative genetic parameters

Robin Thompson^{1,2,*}, Sue Brotherstone³ and Ian M. S. White³

¹*Rothamsted Research, Harpenden, Herts AL5 2JQ, UK*

²*Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK*

³*Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Edinburgh EH9 3JT, UK*

This paper gives a short review of the development of genetic parameter estimation over the last 40 years. This shows the development of more statistically and computationally efficient methods that allow the fitting of more biologically appropriate models. Methods have evolved from direct methods based on covariances between relatives to methods based on individual animal models. Maximum-likelihood methods have a natural interpretation in terms of best linear unbiased predictors. Improvements in iterative schemes to give estimates are discussed. As an example, a recent estimation of genetic parameters for a British population of dairy cattle is discussed. The development makes a connection to relevant work by Bill Hill.

Keywords: genetic parameters; variance components; heritability; residual maximum likelihood; milk yield

1. INTRODUCTION

One important strand of Bill Hill's research has been the application of genetics to livestock improvement. This requires the accurate prediction of breeding values, which in turn requires knowledge of the variances and covariances of random genetic and environmental effects. This paper gives a short review of the development of genetic parameter estimation. Again, this is an area to which Bill Hill has made major contributions. Typically, phenotypic covariances between relatives are partitioned into causal components such as variances due to additive effects, and temporary and permanent environmental effects (Falconer & Mackay 1996). In this paper, because of space limitations, general principles are illustrated using an infinitesimal additive model for normally distributed traits. In this model, the estimation of genetic parameters is often equivalent to the estimation of variance and covariance components. The book of Lynch & Walsh (1998) gives a very readable and comprehensive account of most of the statistical and genetic principles in this review. As an example, a recent estimation of genetic parameters using random regression models for a British population of dairy cattle is discussed.

2. ESTIMATION OF GENETIC PARAMETERS

(a) *Direct methods based on covariance between relatives*

At the simplest levels, analysis of variance methods can provide estimates of phenotypic covariances when groups can be constructed with the same degree of genetic relationship for all members of a group. In the

1960s, at the beginning of Bill Hill's career, estimation was commonly based on paternal half-sib groups or groups of parents and offspring. Once phenotypic covariances between relatives had been estimated, these estimates were interpreted to give estimates of genetic parameters. Some designs generated both types of relationship, and it is sometimes appropriate to combine the estimates. For example, if several offspring are raised from each sire, sire-offspring and half-sib covariances could be combined. These estimates are correlated (Hill & Nicholas 1974), and weighted estimates can be derived. However, sib covariance estimates are biased by the selection of parents (e.g. Robertson 1977). Henderson *et al.* (1959) and Curnow (1961), using a sequential decomposition of the likelihood, show how maximum-likelihood (ML) techniques can be used to correct for this bias (provided the base animals are unselected). Optimal designs are given by Hill & Nicholas (1974) for unselected parents, and by Thompson (1976) for assortative mating and the selection of parents in a variety of hierarchical mating structures with measurement on parent and offspring. The general conclusions are that ML gives the most gain (about 20%) in accuracy from combining the estimates at intermediate heritabilities ($h^2 = 0.2$) when there is no selection of parents. At low values of h^2 , most of the information comes from sib covariances, and at high values, from offspring-parent regression. When there is selection of parents, the combination of estimates is more useful at low heritabilities.

In the field it is uncommon to obtain equal sized families, and it is common to need to adjust data for environmental effects. Henderson (1953) described general methods for these situations, based on the analysis of variance idea of constructing sensible partitions of sums of squares and equating these sums of squares to their expectation. This, in general, was not optimal, as it used a fixed effect model to construct

* Author for correspondence (robin.thompson@bbsrc.ac.uk).

One contribution of 16 to a Theme Issue 'Population genetics, quantitative genetics and animal improvement: papers in honour of William (Bill) Hill'.

the sums of squares and then a mixed model to interpret them. In some special cases it was found that incorporating data with small family sizes gave larger sampling variances than when data on the smallest family sizes were removed. This was partly a consequence of inappropriate weighting to the family means based on different family sizes, and Robertson (1962) suggested a way of alleviating this problem by giving optimal weights for different family sizes for a one-way classification.

(b) *Maximum-likelihood methods*

This section is based on a recent review by Thompson & Mäntysaari (2004) who give a more algebraic presentation. Patterson & Thompson (1971) introduced a residual maximum likelihood (REML) method based on maximizing the likelihood of error contrasts, i.e. contrasts that contributed no information on fixed effects. This REML method agreed with analysis of variance methods in balanced cases and effectively eliminated bias in variance estimation due to not knowing the fixed effects. The implementation of REML estimation depends intimately on the mixed model equations (MME) introduced by Henderson (1973) to construct best linear unbiased predictors (BLUP). To allow the accurate prediction of breeding values, it is important to take account of genetic relationships between all animals. Animal models have been introduced to allow the specification of all additive covariances between relatives by incorporating a random effect for the breeding value of each animal. In the formation of the MME it is the inverse of the numerator relationship matrix, $\mathbf{A}$, that is needed. Algorithms have been developed (Henderson 1976; Quaas 1976) to calculate the inverse directly from a list of pedigrees. Since the only non-zero elements of the inverse are those linking parents to offspring and individuals to themselves, $\mathbf{A}^{-1}$ is very sparse. One way of explaining the sparsity is to say that given all ancestors' breeding values, the breeding value of an individual depends only on its parents' breeding values. Before the introduction of the animal model, parameter estimation had been a three-stage procedure: adjusting data for environmental effects, estimating covariance between relatives, and interpreting them in terms of genetic parameters. Now, it consisted of fitting a mixed linear model with specification-of-expectation and variance submodels.

The easiest way to maximize the likelihood is to find the zeros of the first differential. In the simplest case, this is equivalent to equating a residual sum of squares and a sum of squares of predicted values to their expectation. These expectations are functions of the inverse of the coefficient matrix, $\mathbf{C}^{-1}$, in the MME. For a one-way classification, the method is equivalent to Robertson's (1962) suggestion. In most cases explicit estimates are not available and iterative schemes are needed to locate the maximum of the likelihood. A Fisher scoring scheme based on the expected value of the second differentials, again a function of $\mathbf{C}^{-1}$, was used by Patterson & Thompson (1971).

An alternative 'expectation-maximization' (EM) algorithm based on thinking of the random effects as 'missing' was suggested by Dempster *et al.* (1977). This algorithm gives an iterative scheme for updating

the variance parameters that can be found from the manipulation of the equations through equating the first differentials of the likelihood to zero. Two advantages of this EM algorithm are that variance parameters stay positive, and there is a guaranteed increase in likelihood at each M-step. However the method can be slow to converge, especially if the heritabilities are small or if there are missing data in multivariate cases. Indeed, this method is said to be the most widely used in animal breeding studies in terms of numbers of iterations. It also requires the inversion of $\mathbf{C}$ in each iteration.

An important development was the introduction by Smith & Graser (1986) of an alternative form for the likelihood that naturally leads to a sequential formation of the likelihood. The likelihood can be thought of as a log determinant term (D) and a residual sum of squares (S). The terms D and S can be formed sequentially by using terms associated with eliminating the fixed and random effects from the MME, or equivalently, in terms of the factorization of $\mathbf{C}$. An advantage of this is that $\mathbf{C}^{-1}$ is not calculated and, because of the large number of zero elements, computation can be substantially reduced by using sparse matrix methods. One disadvantage was that the differentials of the likelihood were not easily available. To maximize the likelihood with one parameter, Smith & Graser (1986) suggested using a quadratic approximation. With more than one parameter, methods that avoid calculating derivatives became a popular, flexible alternative. These methods were used for animal and reduced animal models, and for both univariate and multivariate data (Meyer 1989; Thompson & Meyer 1990; Meyer 1991). The basic framework was extended to include more biologically appropriate models with genetic components, including maternal models with both Wilham and Falconer terms (Koerhuis & Thompson 1997), and models with mutation terms (Wray 1990). A major disadvantage of derivative-free methods is that the computational effort increases dramatically as the number of variance parameters increases.

An important advance in derivative-based methods was the rediscovery (Misztal & Perez-Enriso 1993) of an algorithm (Takahashi *et al.* 1973) allowing calculation of the terms in the inverse of $\mathbf{C}$ that are required for forming the first differentials (a 'sparse' inverse of $\mathbf{C}$) without calculating all the elements of the inverse. The sequential formation of this sparse inverse of $\mathbf{C}$ can be described in a way that parallels the sequential formation of the likelihood (Thompson *et al.* 1994). Meyer & Smith (1996) introduced an alternative way of calculating these first differentials by performing the 'automatic' differentiation of the Cholesky decomposition of $\mathbf{C}$. These techniques for forming first derivatives both require twice the computational effort (at least in terms of the number of multiplications) compared with forming the likelihood.

This result allowed the implementation of EM algorithms that could estimate variance parameters (Misztal 1994). These were an improvement on derivative-free methods, but could still be slow to converge. It is possible to calculate second differentials using automatic differentiation (Smith 1995), but the computation of each second differential requires six

times as many multiplications as those involved in a single likelihood calculation (Smith 1995), and this becomes more costly as the number of parameters increases. There are various suggestions on approximating the second differential. Mäntysaari & Van Vleck (1989) suggest accelerating the EM algorithm based on the observed geometric rate of convergence. A quasi-Newton scheme using first differential values to build up an approximate second differential was suggested by Neumaier & Groeneveld (1998). A third method, based on manipulation of the alternative information matrices, was suggested by Thompson and co-workers (Gilmour *et al.* 1995; Johnson & Thompson 1995; Jensen *et al.* 1997). They show that the average of the observed and expected information (AI) is relatively easy to calculate. The AI matrix is based on the differential of the sum of squares S , and can be written in a similar way to S itself, using 'working' variables based on estimates of the fixed effects and predictors of the random effects.

A synthesis of comparisons of the different algorithms used to compute REML estimates was carried out by Hofer (1998), and was updated by Thompson & Mäntysaari (2004). These comparisons show the expected improvement of EM methods over derivative-free methods. The comparisons also show that most second differential methods converge in relatively few iterations.

(c) Other parameterizations and iterative schemes

In some cases, transformations can aid in estimation. If we have multivariate data with two ($p \times p$) variance matrices to estimate, say $\mathbf{G}$ and $\mathbf{R}$, then a canonical transformation (Hayes & Hill 1980; Meyer 1985, 1997) can help in reducing one $p \times p$ estimation into p independent analyses.

A related problem is that often we require $\mathbf{G}$ and $\mathbf{R}$ to be positive definite, and schemes based on second differentials do not necessarily lead to positive definite matrices. One suggestion is to use transformed parameters. In the univariate case ($\mathbf{R} = \sigma^2 \mathbf{I}$), σ or $\log \sigma$ have been suggested instead of σ^2 . Multivariate analogues such as Cholesky transformations have been suggested by Lindstrom & Bates (1988), and Groeneveld (1994). Recent work on EM algorithms (Foulley & Quaas 1995; Meng & Van Dyk 1998) has suggested that this Cholesky, or linear, parameterization has a natural interpretation and can lead to faster convergence.

For example, Foulley & Quaas (1995) use a model with residual variance σ_E^2 and genetic variance σ_G^2 , and the breeding values $\mathbf{u}^*$ scaled by σ_G . Estimation of σ_G follows a two stage procedure of first predicting $\mathbf{u}^*$ with mixed model equations. Then, regression of data on breeding values (taking into account the uncertainty of the breeding values) gives a natural way of estimating σ_G and keeps σ_G^2 within the parameter space. For a balanced sire model, Foulley & Quaas (1995) note that this algorithm has a geometric rate of convergence with the distance of a parameter from its final value approximately reducing by a factor, r , in each iteration. This factor r is a function of $\sigma = n/(n + \sigma)$, with $\sigma = \sigma_E^2/\sigma_G^2$. For $\sigma = 0.2, 0.5$ and 0.8 , estimation of σ_G using an EM algorithm gives $r = 0.27, 0.45$ and 0.31

respectively, compared with $0.03, 0.25$ and 0.63 for a scheme based on updating σ_G^2 . This shows the advantage of the σ_G parameterization for small values of σ . The first scheme estimates a linear parameter σ_G and the second scheme estimates a quadratic parameter σ_G^2 .

A more recent development is the parameter extension or PX-EM algorithm suggested by Liu *et al.* (1998). In our case it involves estimating $\sigma_G^2 = (\sigma_{G1})^2 \sigma_{G2}^2$ and σ_{G1} by the first linear parameterization scheme, and σ_{G2}^2 by the second quadratic parameterization scheme. This scheme is at first sight counter-intuitive in that σ_{G1} and σ_{G2}^2 look confounded, but each is individually updated and σ_G^2 is estimable. The scheme has a rate of convergence that again depends on σ , but is faster than the two previous schemes, with rates of convergence of $0.30, 0.60$ and 0.80 for $\sigma = 0.2, 0.5$ and 0.8 for the balanced sire case (Thompson & Mäntysaari 2004).

There has been interest recently in parameterizing variance matrices in different ways to lead to more parsimonious models. One way that has similarities with PX-EM methods is to write the p breeding values on one animal as $\mathbf{u} = \mathbf{S}\mathbf{f} + \mathbf{e}$ with $\text{var}(\mathbf{f}) = \mathbf{U}$ and $\text{var}(\mathbf{e}) = \mathbf{D}$, a diagonal matrix. The matrix $\mathbf{S}$ ($p \times f$) could be thought of as representing a set of f factors, and the parameterization can represent a reduced rank or latent regression parameterization. Then an estimation of $\mathbf{S}$ can be found from the prediction of $\mathbf{f}$ and then regressing data on $\mathbf{f}$, taking into account the uncertainty in $\mathbf{f}$ (Thompson *et al.* 2003). This provides insight into the PX-EM schemes and might suggest alternative parameterizations or hybrid schemes that have better iterative convergence properties.

We think that AI iterative schemes are attractive in that they usually only need a small number of iterations. The two drawbacks are that they do not always improve the likelihood, and that they can lead to estimates outside the parameter space. The first difficulty reduces as parameters get nearer to a maximum value of the likelihood. One hybrid scheme that is attractive is the suggestion (Cullis *et al.* 2004) of starting with a PX-EM scheme and switching to an AI scheme when the iterates are 'close' to a REML solution.

(d) Sampling approaches

Thus far we have concentrated on exact methods of analysis. A recent, excellent book (Sorensen & Gianola 2002) has discussed Bayesian and Markov chain Monte Carlo (MCMC) methods. One reason often advocated for using a Bayesian approach is that it allows the natural use of 'prior' information on genetic parameters. This prior evidence is not often well explained. For example, in the recent 2002 World Congress on Genetics Applied to Livestock Production, over thirty papers presented Bayesian analyses of quantitative genetic parameters. Over half the papers did not quantify prior knowledge, and a third used vague or flat prior knowledge for the genetic parameters. In some papers at least it seems that only lip service was paid to the Bayesian paradigm.

In a sense there is a direct analogy between direct and iterative estimation in linear estimation, and exact and sampling-based methods in quadratic estimation.

One can think of Gibbs sampling methods as adding noise at every step of a simplified exact analysis. For instance, estimate fixed effects and add noise, predict random effects and add noise, form sums of squares for breeding values and add noise to give an estimate of σ_G^2 . One does not need to be Bayesian to use MCMC methods: Guo & Thompson (1994) use the above paradigm, with the estimation of σ_G^2 given by an EM step. In a sense the difficulties of calculating prediction error variances are replaced by sampling them. Thompson (1994) and Groeneveld & García-Cortés, (1998) have pointed out that the sampling error can be reduced when updating σ_G^2 by taking account of the variance of the noise added to random effects, although this is simpler to do for uncorrelated effects. One can also get nearer to exact methods by using block updating, but this leads to more complicated variance correction formula. It is not always clear which computational scheme, exact, Gibbs sampling or intermediate, will minimize computational effort.

A recent suggestion by Clayton & Rasbash (1999) for imputation can also reduce the computational effort. In a simple application, their idea suggests fitting two models. In the first model the data are adjusted for predicted random effects, and the fixed effects are estimated; and in the second model the data are adjusted for fixed effects, and breeding values are predicted and genetic variance and residual variance are estimated. The procedure is repeated many times, at each stage adding noise to the estimates of the fixed and random effects. The iterative estimates have some sampling variation and this can be reduced by averaging a sequence of estimates after 'burn-in'. This scheme is similar to Gibbs sampling but should have less sampling variation. Thompson (1994), and more recently García-Cortés & Sørensen (2001) and Harville (2004), have discussed related ideas.

(e) *Validation of models*

The purpose of a large amount of parameter estimation is to allow the efficient prediction of breeding values and efficient selection procedures. In some ways it is more important to know if we are selecting the best animals, rather than whether we are fitting the best model. In some ways, using Falconer's realized heritability, relating selection response to selection differential could be thought of as more important than estimates of heritability based on covariances between relatives in unselected base populations. Hill (1971) put the estimation of realized heritability from the comparison of divergent lines, or a selected line and a control line, on a sound and rigorous scientific and statistical basis, in particular deriving the appropriate sampling variance matrix for the generation-line means. Sheridan (1988) reviewed experiments in all animal species in which base population and realized heritability estimates were available, and concluded that there was not a strong agreement between the realized genetic and base genetic parameters. James (1990) suggested that a major cause of this discrepancy appeared to be sampling errors.

One concern with the validation of animal models is whether different types of relationship provide consistent information. Thompson & Atkins (1994) developed methods to allow within- and between-line

information on genetic variation to be disentangled. This led to an empirical verification of Hill's analytical justification for the sampling variances of the generation-line means. Thompson & Atkins (1994), found there was substantial agreement between the two types of estimates for an extensive sheep-selection experiment. Another attempt to disentangle the contribution of different components of the individual animal model to heritability estimation was due to Visscher & Thompson (1992), who fitted a model with different male and female heritabilities in the analysis of a British dairy cattle population. They showed that approximately two-thirds of the information came from female (mostly dam-offspring) relationships.

Another way of partially validating a model was the method R introduced by Reverter (Reverter *et al.* 1994a, b). The initial suggestion was to check whether a model is appropriate by comparing predictions based on 'early' data with predictions based on 'recent' data. Reverter *et al.* (1994a) show that the regression of 'recent' prediction on 'early' predictions is 1. Informally, this statistic is asking the question: does recent data change the prediction of early animals? In a sense this is looking backwards. By contrast, the analysis of selection experiments and the partitioning of likelihood often look forward and ask: does response agree with prediction for 'recent' animals? Reverter *et al.* (1994b) have suggested that the method could be extended to estimate genetic parameters, essentially choosing an estimated heritability to make the regression coefficient 1. One problem with ML techniques is the dependence of the procedures on having estimates of the prediction error variances. Method R introduced an ingenious algorithm that avoids calculation of prediction error variances. This method has been used in several large genetic situations (e.g. Misztal 1997; Van Tassell *et al.* 1999), but the method is not completely understood. A concern is whether or not the method can be used with selected data. Recent simulation evidence (Cantet *et al.* 2000; Schenkel & Schaeffer 2000) and theoretical evidence (Thompson 2001) has suggested that method R estimates are biased in selected populations.

The introduction of multiple-trait across country evaluation (MACE; Schaeffer 1994) has also led to interest in validating models. One concern is that if the genetic trend is not well estimated it might influence predictions of breeding values and their comparison. Boichard *et al.* (1995) introduced three methods for checking on the genetic trend. In the first, predictions from first parity records are compared with prediction from all records using a repeatability model, on the basis that the first parity model is less affected by parity effect biases. In the second method, daughter yield deviations are analysed within sire by calving-year to determine if these remain stable over time or if there is evidence of selection biases. The third method, an extension of that of Reverter *et al.* (1994b), compares recent with early predictions, but includes terms to measure differences between the two predictions and a term to allow the estimation of genetic trend. The methods have been influential in validating and improving predictions of data used by INTERBULL, although the results of the tests have not been published (U. Emanuelson, personal communication).

Figure 1. Genetic correlation matrix for the Legendre polynomial coefficients, ordered as intercept, slope and curvature, then by lactation, then by the traits milk, fat and protein yield.

3. GENETIC PARAMETERS FOR THE UK TEST DAY MODEL

Current software programs such as Asreml (Gilmour *et al.* 2002) for fitting the mixed model incorporate many of the developments described above. This allows problems involving a very large number of variance components to be tackled which would have been impossible even a few years ago. As an example, we consider work in preparation by I. M. S. White, S. Brotherstone and R. Thompson concerning the estimation of about 800 variances and covariances for the UK test day model.

Data were available on 22 343 daughters of 823 bulls. Milk (M), fat (F) and protein (P) yield in each of the first three lactations were analysed as nine traits. The mean lactation curve was fitted as a quartic polynomial function of days in milk (DIM). Sire and cow (permanent environmental) deviations from the mean curve were fitted as Legendre polynomials (L_0 , L_1 and L_2) with random coefficients. Covariance matrices were therefore 27×27 (3 variables $\times$ 3 lactations $\times$ 3 coefficients). The advantages of such random regression models are that they allow more appropriate modelling of environmental effects, more flexible modelling of the lactation curves, and the prediction

of functions of these curves, including persistency. Space does not allow us to adequately discuss these random regression models, so we refer the reader to Meyer & Hill (1997) and Meyer & Kirkpatrick (2005) for a more extensive discussion.

The traits were analysed in sets of three, with milk in first lactation always included to allow for selection. There were thus 28 runs required to analyse all 9 traits, each run estimating a 9×9 submatrix of the complete 27×27 between- and within-sire matrices. The contemporary group (CG) effect was allowed to depend linearly on DIM, with an intercept and slope for each of milk, fat and protein (independent of lactation). Variation in intercept and slope was treated as fixed between herds and random within herds. This required the estimation of a 6×6 within-herd covariance matrix (intercept and slope for each of milk, fat and protein). Complete matrices were reconstituted from overlapping submatrices so that they were positive definite. This was achieved by using the Cholesky factors of submatrices as 'pseudo-data' for the full set of traits. This computationally attractive procedure is similar to the 'iterative summing of expanded part matrices' of Mäntysaari (1999). The idea can be illustrated with a simple example of 3 traits (X , Y , Z) with covariance matrix V

Table 1. Heritabilities at various days in milk.

		days in milk					
lactation	trait	50	100	150	200	250	300
1	milk	0.445	0.484	0.507	0.502	0.478	0.432
1	fat	0.298	0.327	0.358	0.356	0.368	0.361
1	protein	0.380	0.404	0.417	0.405	0.390	0.378
2	milk	0.413	0.477	0.497	0.482	0.452	0.401
2	fat	0.237	0.254	0.280	0.293	0.340	0.336
2	protein	0.304	0.358	0.386	0.385	0.369	0.326
3	milk	0.467	0.499	0.493	0.480	0.481	0.434
3	fat	0.201	0.209	0.222	0.244	0.323	0.339
3	protein	0.338	0.347	0.329	0.319	0.343	0.343

Table 2. Estimates of heritability (along diagonal) and genetic correlations (below diagonal) at 100 days in milk.

		1	1	1	2	2	2	3	3	3
lactation	trait	milk	fat	protein	milk	fat	protein	milk	fat	protein
1	milk	0.48								
1	fat	0.71	0.33							
1	protein	0.92	0.80	0.40						
2	milk	0.73	0.53	0.64	0.48					
2	fat	0.47	0.88	0.59	0.59	0.25				
2	protein	0.76	0.68	0.78	0.89	0.68	0.36			
3	milk	0.62	0.31	0.52	0.94	0.43	0.78	0.21		
3	fat	0.46	0.60	0.63	0.59	0.77	0.69	0.58	0.21	
3	protein	0.68	0.48	0.67	0.87	0.52	0.86	0.86	0.65	0.35

(3×3). Suppose the traits are analysed in pairs, and the analysis of (X, Y), (X, Z) and (Y, Z) produces estimated covariance matrices (2×2) with Cholesky roots **A**, **B** and **C**. The pseudo-data take the form

$$\begin{array}{ccc} A_{11} & A_{12} & * \\ 0 & A_{22} & * \\ B_{11} & * & B_{12} \\ 0 & * & B_{22} \\ * & C_{11} & C_{12} \\ * & 0 & C_{22} \end{array}$$

Here * represents a missing value. This pseudo-data is presented to a REML program as multivariate normal data with zero mean and covariance matrix **V**. For the test day model, between- and within-sire matrices were dealt with simultaneously, with constraints to ensure that both genetic (4×between-sire) and permanent environment (within-sire – 3×between-sire) components were positive definite. Residual variances and covariances were estimated separately for each stage of each lactation (i.e. there were twelve 3×3 residual covariance matrices for milk, fat and protein).

Each of the 28 runs estimated up to 135 variance-covariance components and involved up to 836 500 mixed model equations. Each iterative round took approximately 5 min on a Pentium 3, 850 MHz personal computer. Approximately 20% of the time taken in each iteration was concerned with the likelihood calculation and approximately 40% was concerned with the calculation of the sparse inverse of **C**, in agreement with the theoretical suggestions. The other

major proportion of time, 30%, was associated with the computation of the first differentials of the likelihood. The number of iterations varied from run to run and ranged from 10 to over 30 iterations. The increased iterations occurred when the estimated matrices were nearly singular.

(a) Legendre polynomial coefficients

Figure 1 shows a colour image of the genetic correlation matrix for the Legendre polynomial coefficients, ordered as intercept, slope and curvature, then by lactation, then by the traits milk, fat and protein yield. Within each type of coefficient (intercept, slope or curvature) there are some strong positive correlations between the traits (M1–P3). Correlations between intercepts or curvatures and slopes tend to be small, while correlations between intercepts and curvature coefficients are generally substantial and negative, reflecting a tendency for deviation curves above the mean curve to be concave, and *vice versa*. The structure of the 27×27 matrices is approximately that of a Kronecker product of a 3×3 lactations matrix and a 9×9 traits matrix. Visscher (1991) and Veerkamp & Goddard (1998) have found similar multiplicative models useful in the analysis of multivariate dairy-cattle data. To a first approximation, the correlation between parameters in different lactations (e.g. intercept for milk in lactation 1 and slope for fat in lactation 2) is the product of two correlations, one depending only on the parameters chosen (milk intercept and fat slope), and one depending only on the two lactations (e.g. 1 and 2).

(b) Heritabilities

Heritabilities calculated from the variance components are shown below. For milk, the average heritability in the first two lactations is about 0.45, with the maximum about day 150 (table 1). In lactation 3 the heritability is slightly higher on average and more nearly constant over the lactation. For protein, heritability drops from about 0.4 in lactation 1 to about 0.35 in lactation 3. Heritabilities for fat show a rising trend within lactations, but average values drop from about 0.35 in lactation 1 to about 0.25 in lactation 3.

(c) Genetic correlations

Table 2 gives estimated genetic correlations between the traits at DIM=100 days.

Numbers on the diagonal are heritabilities. In each lactation there are strong correlations between protein and the other traits, and a weaker correlation between milk and fat. Cross-correlations between traits in different lactations follow the same pattern, but are reduced by a factor of approximately 0.8 for adjacent lactations and 0.6 for lactations 1 and 3. Lagged correlations (not shown) show a similar pattern, but are reduced by a factor depending on the separation of the time points.

4. CONCLUSIONS

We have shown that the area of genetic parameter estimation has advanced tremendously over the last 40 years, allowing more appropriate models to be fitted to larger datasets.

What do we expect to happen in the next 10 years? We expect to see even more complex models fitted using molecular information. The infinitesimal model is often a useful first step in understanding genetic data and remains the basis of practical improvement schemes. Even when more precise molecular models are fitted, it is often useful to include infinitesimal model terms to sweep up unexplained genetic variation.

Fernando & Grossman (1989) have shown how to incorporate marker information into prediction, essentially extending the Quaas (1976) results on the inverse of the 'average' relationship matrix to the case when there is information on markers in pedigrees. Meuwissen & Goddard 2000 show how to incorporate this work into fine scale mapping using both linkage and linkage disequilibrium information.

In some ways, work on estimation of variance parameters, validation of models and prediction of breeding values are thought of as separate exercises, partly because of the difficulty of estimating variance parameters on complete populations. Faster genetic progress might be possible if we could integrate the methods, for example through a synthesis of sampling and analytic methods. Work is needed to reconcile the sometimes discrepant results obtained from sampling variances of parameters, for example those calculated with the average information method, and confidence intervals obtained from profile likelihoods. This work may highlight cases where it is difficult to disentangle parameter estimates. Sales & Hill (1976) suggested formulae for genetic progress taking account of

uncertainty in parameters in simple cases. We expect this to be extended to more complicated models, allowing alternative genetic models to be compared based on measures of genetic progress, rather than the statistical model selection criteria (e.g. Akaike's information criterion; Akaike 1973) that correct the model likelihood for the estimation of parameters.

R.T. would like to thank Bill Hill for his multifaceted and unstinting support throughout his career; this has ranged from wise counsel to the Animal Breeding Research Organization and its successors and involvement in the joint supervision of over fifteen Ph.D. students (when R.T. learnt more than the students), to encouragement to learn at the feet of Oscar Kempthorne. The authors would like to acknowledge the comments and patience of Peter Visscher and the two referees.

REFERENCES

- Akaike, H. 1973 Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (ed. B. N. Petrov & F. Csaki), pp. 267–281. Budapest: Akademiai Kiado.
- Boichard, D., Bonati, B., Barbat, A. & Mattila, S. 1995 Three methods to validate the estimation of genetic trend for dairy cattle. *J. Dairy Sci.* **78**, 431–437.
- Cantet, R. J. C., Birchmeier, A. N., Santos-Crista, M. G. & de Avila, V. S. 2000 Comparison of residual maximum likelihood and method R for estimating heritability and predicting breeding value under selection. *J. Anim. Sci.* **78**, 2554–2560.
- Clayton, D. & Rasbash, J. 1999 Estimation in large crossed random effects models by data augmentation. *J. R. Stat. Soc. A* **162**, 425–436.
- Cullis, B. R., Smith, A. B. & Thompson, R. 2004 Perspectives of ANOVA, REML and a general linear mixed model. In *Methods and models in statistics—in honour of Professor John Nelder FRS* (ed. N. M. Adams, M. J. Crowder, D. J. Hand & D. A. Stephens).
- Curnow, R. N. 1961 The estimation of repeatability and heritability from records subject to culling. *Biometrics* **17**, 553–566.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38.
- Falconer, D. S. & Mackay, T. F. C. 1996 *Introduction to quantitative genetics*. Harlow: Longman.
- Foulley, J. L. & Quaas, R. L. 1995 Heterogeneous variances in Gaussian linear mixed models. *Genet. Sel. Evol.* **27**, 211–228.
- Garcia-Cortes, L. A. & Sorensen, D. 2001 Alternative implementations of Monte Carlo EM algorithms for likelihood inferences. *Genet. Sel. Evol.* **33**, 443–452.
- Gilmour, A. R., Thompson, R. & Cullis, B. R. 1995 Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450.
- Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J. & Thompson, R. 2002 *ASREML user guide, Release 1.0*. Hemel Hempstead: VSN International.
- Groeneveld, E. 1994 A reparameterization to improve numerical optimization in multivariate REML (co)variance component estimation. *Genet. Sel. Evol.* **26**, 537–545.
- Groeneveld, E. & Garcia-Cortés, A. 1998 VCE4.0, a (co)variance component package for frequentists and Bayesians. In *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production, Armidale* **27**, 455–456. International Committee of World Congresses on Genetics applied to Livestock Production.

- Guo, S. W. & Thompson, E. A. 1994 Monte-Carlo estimation of mixed models for large complex pedigrees. *Biometrics* **50**, 417–432.
- Fernando, R. L. & Grossman, M. 1989 Marker-assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.* **21**, 246–477.
- Harville, D. A. 2004 Making REML computationally feasible for large data-sets: use of the Gibbs sampler. *J. Stat. Comput. Sim.* **74**, 135–154.
- Hayes, J. F. & Hill, W. G. 1980 A reparameterization of a genetic selection index to locate its sampling properties. *Biometrics* **36**, 237–248.
- Henderson, C. R. 1953 Estimation of variance and covariance components. *Biometrics* **9**, 226–252.
- Henderson, C. R. 1973 Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. Jay L. Lush*, pp. 10–41. Champaign, Illinois: American Society of Animal Science and American Dairy Science Association.
- Henderson, C. R. 1976 A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**, 69–82.
- Henderson, C. R., Kempthorne, O., Searle, S. R. & von Krosigk, C. M. 1959 The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192–218.
- Hill, W. G. 1971 Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics* **27**, 293–311.
- Hill, W. G. 1974 Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics* **30**, 447–468.
- Hofer, A. 1998 Variance component estimation in animal breeding: a review. *J. Anim. Breed. Genet.* **115**, 247–265.
- James, J. W. 1990 Selection theory versus selection results—a comparison. In *Proceedings of the fourth World Congress on Genetics Applied to Livestock Production, Edinburgh*, vol. 13 pp. 195–204. International Committee of World Congresses on Genetics applied to Livestock Production.
- Jensen, J., Mäntysaari, E. A., Madsen, P. & Thompson, R. 1997 Residual maximum likelihood estimation of (co) variance components in multivariate mixed linear models using average information. *J. Ind. Soc. Agric. Stat.* **49**, 215–236.
- Johnson, D. L. & Thompson, R. 1995 Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* **78**, 449–456.
- Koerhuis, A. N. M. & Thompson, R. 1997 Models to estimate maternal effects for juvenile body weight in broiler chickens. *Genet. Sel. Evol.* **29**, 225–299.
- Lindstrom, M. J. & Bates, D. M. 1988 Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Am. Stat. Assoc.* **83**, 1014–1022.
- Liu, C., Rubin, D. B. & Wu, Y. N. 1998 Parameter expansion to accelerate EM: The PX–EM algorithm. *Biometrika* **85**, 755–770.
- Lynch, M. & Walsh, B. 1998 *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer Associates.
- Mäntysaari, E. 1999 Derivation of multiple trait reduced rank random regression (RR) model for the first lactation test day records of milk, protein and fat. In *Proceedings of the 50th Annual Meeting EAAP, Zurich, Switzerland, August 22–26 1999*, p. 26. European Association for Animal Production.
- Mäntysaari, E. & Van Vleck, L. D. 1989 Restricted maximum likelihood estimates of variance components from multi-trait sire models with large number of fixed effects. *J. Anim. Breed. Genet.* **106**, 409–422.
- Meng, X. L. & Van Dyk, D. A. 1998 Fast EM-type implementation for mixed effects models. *J. R. Stat. Soc. B* **60**, 559–578.
- Meuwissen, T. H. E. & Goddard, M. E. 2002 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**, 421–430.
- Meyer, K. 1985 Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics* **41**, 153–165.
- Meyer, K. 1989 Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genet. Sel. Evol.* **21**, 317–340.
- Meyer, K. 1991 Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genet. Sel. Evol.* **23**, 67–83.
- Meyer, K. 1997 An ‘average information’ restricted maximum likelihood algorithm for estimating reduced rank genetic covariance matrices or covariance functions for animal models with equal design matrices. *Genet. Sel. Evol.* **29**, 97–116.
- Meyer, K. & Hill, W. G. 1997 Estimation of genetic and phenotypic covariance functions for longitudinal data by restricted maximum likelihood. *Livest. Prod. Sci.* **47**, 185–200.
- Meyer, K. & Kirkpatrick, M. 2005 Up hill, down dale: quantitative genetics of curvaceous traits. *Phil. Trans. R. Soc. B* **360**, 1443–1455. (doi:10.1098/rstb.2005.1681.)
- Meyer, K. & Smith, S. P. 1996 Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. *Genet. Sel. Evol.* **28**, 23–49.
- Misztal, I. 1994 Comparison of computing properties of derivative and derivative-free algorithms in variance component estimation by REML. *J. Anim. Breed. Genet.* **111**, 346–355.
- Misztal, I. 1997 Estimation of variance components with large scale dominance models. *J. Dairy Sci.* **80**, 965–974.
- Misztal, I. & Perez-Enciso, M. 1993 Sparse matrix inversion for restricted maximum likelihood estimation of variance components by expectation–maximization. *J. Dairy Sci.* **76**, 1479–1483.
- Neumaier, A. & Groeneveld, E. 1998 Restricted maximum likelihood estimation of covariances in sparse linear models. *Genet. Sel. Evol.* **30**, 3–26.
- Patterson, H. D. & Thompson, R. 1971 Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.
- Quaas, R. L. 1976 Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* **32**, 949–953.
- Reverter, A., Golden, B. L., Bourdon, R. M. & Brinks, J. S. 1994a Detection of bias in genetic predictions. *J. Anim. Sci.* **72**, 34–37.
- Reverter, A., Golden, B. L., Bourdon, R. M. & Brinks, J. S. 1994b Method R variance components procedure application on the single breeding value model. *J. Anim. Sci.* **72**, 2247–2253.
- Robertson, A. 1962 Weighting in the estimation of variance components in the unbalanced single classification. *Biometrics* **18**, 413–417.
- Robertson, A. 1977 The effect of selection on the estimation of genetic parameters. *Z. Tierzücht.* **94**, 131–135.
- Sales, J. & Hill, W. G. 1976 Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. *Anim. Prod.* **23**, 1–14.
- Schaeffer, L. R. 1994 Multiple-country comparisons of dairy sires. *J. Dairy Sci.* **77**, 2671–2678.

- Schenkel, F. S. & Schaeffer, L. R. 2000 Effects of non-random mating parental selection on variance components. *J. Anim. Genet.* **117**, 225–239.
- Sheridan, A. K. 1988 Agreement between estimated and realised genetic parameters. *Anim. Breed. Abstr.* **56**, 877–889.
- Smith, S. P. 1995 Differentiation of the Cholesky algorithm. *J. Comput. Graph Stat.* **4**, 134–147.
- Smith, S. P. & Graser, H.-U. 1986 Estimating variance components in a class of mixed models by restricted maximum likelihood. *J. Dairy Sci.* **69**, 1156–1165.
- Sorensen, D. & Gianola, D. 2002 *Likelihood, Bayesian and MCMC methods in quantitative genetics*. New York: Springer.
- Takahashi, K., Fagan, J. & Chin, M. S. 1973 Formation of a sparse bus impedance matrix and its application to short circuit study. In *Proceedings of the Eighth Institutional PICA Conference, Minneapolis*, p. 63. IEEE Power Engineering Society.
- Thompson, R. 1976 Design of experiments to estimate heritability when observations are available on parents and offspring. *Biometrics* **32**, 903–917.
- Thompson, R. 1994 Integrating best linear unbiased prediction and maximum likelihood estimation. *Fifth World Congress on Genetics Applied to Livestock Production, Guelph*, vol. 18, pp. 337–340. International Committee of World Congresses on Genetics applied to Livestock Production.
- Thompson, R. 2001 Statistical validation of genetic models. *Livest. Prod. Sci.* **72**, 129–134.
- Thompson, R. & Atkins, K. D. 1994 Sources of information for estimating heritability from selection experiments. *Genet. Res. Camb.* **63**, 49–55.
- Thompson, R. & Mäntysaari, E. 2004 Prospects for statistical methods in animal breeding. *J. Ind. Soc. Agric. Stat.* **57**, 15–25.
- Thompson, R. & Meyer, K. 1990 Estimating genetic parameters using an imaginary animal model. *Genet. Sel. Evol.* **22**, 133–147.
- Thompson, R., Cullis, B., Smith, A. & Gilmour, A. 2003 A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Aust. NZ J. Stat.* **45**, 445–459.
- Thompson, R., Wray, N. R. & Crump, R. E. 1994 Calculation of prediction error variances using sparse matrix methods. *J. Anim. Breed. Genet.* **111**, 102–109.
- Van Tassell, C. P., Wiggans, G. R. & Norman, H. D. 1999 Method R estimates of heritability for milk, fat and protein yields of United States dairy cattle. *J. Dairy Sci.* **82**, 2231–2237.
- Veerkamp, R. F. & Goddard, M. E. 1998 Covariance functions across herd production levels for test day records on milk, fat, and protein yields. *J. Dairy Sci.* **81**, 1690–1701.
- Visscher, P. M. 1991 Estimation of genetic parameters in dairy cattle using an animal model and implications for genetic improvement. Ph.D. thesis, University of Edinburgh.
- Visscher, P. M. & Thompson, R. 1992 Comparisons between genetic variances estimated from different types of relatives in dairy cattle. *Anim. Prod.* **55**, 315–320.
- Wray, N. R. 1990 Accounting for mutation effects in the additive genetic variance covariance matrix and its inverse. *Biometrics* **46**, 197–186.