PRESENT STATUS OF KNOWLEDGE ABOUT STATISTICAL PROCEDURES AND ALGORITHMS TO ESTIMATE VARIANCE AND COVARIANCE COMPONENTS

Karin Meyer

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

SUMMARY

Estimation of variance components for continuous traits is reviewed, concentrating on Restricted Maximum Likelihood and algorithms specific to animal breeding applications. Special emphasis is given to developments since the last World Congress. Problems and areas for future research are discussed.

INTRODUCTION

At each World Congress, problems and methodology to estimate genetic parameters have been reviewed. Hill (1974) considered heritability estimation, concentrating on aspects of experimental design. The question of method of analysis was dealt with by pointing towards the availability of general purpose computer programs, especially Harvey's (1972) least squares program, though the author briefly commented on the usefulness of Maximum Likelihood (ML) methods to combine data from different sources in an optimal way. Since then, rapid developments in computer power and computing facilities available have stimulated extensive research on ML and related procedures to estimate variance and covariance components (henceforth referred to as variance components) for unbalanced, mixed linear models. Harville (1977) presented an comprehensive general review of the earlier work, emphasizing among other points the need for future research to identify the best computing algorithms for particular classes of models. Thompson (1982) discussed the scope of ML for the estimation of genetic parameters, showing that some of the terms involved have a 'natural' interpretation in the animal breeding context. An update was given by Thompson and Cameron (1986).

After a brief review of ML estimation and relevant models of analysis, this paper attempts to present on overview of developments in the estimation of genetic parameters for continuous traits, in particular since the last World Congress. Emphasis is given almost exclusively to ML estimation, concentrating on algorithms and models specific to animal breeding.

MAXIMUM LIKELIHOOD

General interest in ML estimators of variance components has been propelled by their desirable statistical properties : they are consistent, asymptotically normal and efficient (Harville, 1977). Furthermore, the ML framework provides a great deal of flexibility, allowing for designs and models of analysis which cannot be accommodated by analysis of variance (ANOVA) type methods of estimation, and constraints on the parameter space can be imposed readily. A recent review of Henderson's (1953) ANOVA type methods in contrast to ML and related procedures has been given by Searle (1989), and empirical comparisons examining the efficacy of the two approaches are available for a variety of designs (e.g. Hocking and Kutner, 1975; Corbeil and Searle, 1976; Henderson and Quaas, 1977; Lin and McAllister, 1984; Swallow and Monahan, 1984; Shaw, 1987).

In animal breeding, data used to estimate variance components frequently originates from selection experiments or livestock improvement schemes which involve continuous culling of animals on the basis of their performance. In that case, ANOVA estimators which assume that data are randomly sampled tend to be subject to selection bias (Robertson, 1977), while under certain conditions ML will account for selection or at least yield estimates less biased than their ANOVA

counterparts. Initial interest in ML, to estimate both genetic parameters and fixed effects, was stimulated by concern about bias due to selection (Henderson, 1949; Lush and Shrode, 1950; Henderson *et al.*, 1959; Curnow, 1961; Thompson, 1973). A number of simulation studies have illustrated the effect of selection on ANOVA and ML estimators of variance components in uni- and multivariate analyses (e.g. Rothschild *et al.*, 1979; Meyer and Thompson, 1984; Sorensen and Kennedy, 1984; Walter and Mao, 1985).

As emphasized by Harville (1977), there are two major conceptual drawbacks of ML estimators. Firstly, the loss in degrees of freedom due to fixed effects in the model of analysis is ignored. Fortunately, this can be overcome by maximizing only the part of the likelihood (L) which is independent of the fixed effects (Patterson and Thompson, 1971), a procedure which is now generally referred to as Restricted Maximum Likelihood (REML). Secondly, ML estimation requires the distribution of the data to be specified. In estimating variance components, this is usually a multivariate normal distribution. Various authors (e.g. Harville, 1977; Banks *et al.*, 1985; Westfall, 1987), however, have indicated that ML or REML estimators may be an appropriate choice even if normality does not hold.

In spite of 'built-in' optimal properties of (RE)ML estimation, practical applications have lagged considerably behind theoretical developments. This has largely been due to computational requirements involved for all but very simple cases. Over the last decade, extensive research effort has been directed towards the development of specialized and efficient algorithms for particular classes of models.

MODELS OF ANALYSIS

Developments in variance component estimation specific to animal breeding have been closely linked with advances in the genetic evaluation of animals by Best Linear Unbiased Prediction (BLUP); see Hill and Meyer (1988) for a review. Early REML applications were generally limited to models largely equivalent to those in corresponding ANOVA type analyses, considering one random effect only, i.e. genetic variances were commonly estimated from covariances among paternal half-sibs fitting the so-called sire model.

Recently, the conceptually simpler Animal Model (AM) has come to dominate genetic evaluation schemes. Rather than describing an animal's record in terms of sire and dam effects, the AM includes an additive genetic effect for each animal, both animals with records and animals which are parents only. This allows information on all known relationships between animals to be incorporated in the analysis. With the AM other effects, such as maternal genetic, non-additive genetic, cytoplasmic or permanent environmental effects, can be accounted for simply by fitting corresponding additional random effects. Henderson (1988) described a variety of such expanded AMs. Genetic properties of the AM have been discussed by Kennedy *et al.* (1988).

In terms of variance component estimation, the AM has changed thinking, from the expectation of mean squares and the interpretation of observational components of variance in genetic terms, to a more direct approach where we fit a vector of random effects for each component of interest. This requires not only the covariance matrix amongst the levels of each random effect to be specified but, for most applications, also the respective inverse. Efficient algorithms to obtain the inverse of the additive genetic relationship matrix directly from a list of pedigree information have been presented by Henderson (1976) and Quaas (1976). Rules to set up inverse covariance matrices for non-additive genetic effects indirectly have recently been derived by Smith (1984), Chang (1988), Schaeffer *et al.* (1989), Smith and Maki-Tanila (1989), and Hoeschele and Van Raden (1989).

NUMERICAL PROCEDURES

Estimation by ML almost invariably involves the numerical solution of a constrained, non-linear optimization problem in an iterative scheme (Harville, 1977). Procedures to locate the mini-

mum or maximum of a function can be classified according to the amount of information from derivatives which is available or can be utilized (e.g. Gill *et al.*, 1981).

Using Second Derivatives

The so-called Newton methods have generally been found quickest to converge. Newton-Raphson (NR) iteration, which requires second partial derivatives of the likelihood function to be calculated, has been described by Jennrich and Sampson (1976), and improvements have been suggested by Lindstrom and Bates (1988). Replacing the second derivatives by their expected values, which tend to be somewhat easier to compute (Harville, 1977), yields Fisher's Method of Scoring (MSC). Efron and Hinkley (1978) give a comparison of the observed and expected Fisher information. The MSC is equivalent to Anderson's (1973) algorithm, or its REML analogue, and, except for constraints on the parameter space, to 'iterated MIVQUE' (Harville, 1977).

Patterson and Thompson (1971) formulated a MSC algorithm, and most early REML applications in animal breeding used this procedure. Specialized REML algorithms for multivariate analyses of models with one random effect were described, for instance, by Thompson (1973), Schaeffer *et al.* (1978), Meyer (1983, 1985) and Henderson (1984). Extensions to models with more random effects are straightforward in principle (e.g. Searle, 1979) and have been considered in the animal breeding context (e.g. Cue, 1986; Meyer, 1987a), but found little practical use.

In a number of 'Quasi-Newton' (QN) procedures (e.g. Dennis and More, 1977) attempts are made to approximate the matrix of (expected) second derivatives, some requiring first derivatives to be evaluated others not. A problem common to all QN methods is to ensure that the approximate Hessian matrix is positive definite; see standard textbooks such as Gill *et al.* (1981) or Kennedy and Gentle (1980) for further details. Harville (1977) mentioned 'variable-metric' methods as a promising new development. Redner and Walker (1984) considered one of them, a secant algorithm, which approximates the Hessian matrix by successive rank updates, and Robinson (1988) applied such algorithm to a multivariate analysis of dairy cattle data. For data with an identical, independent distribution first derivatives of log L (=scores) are calculated and summed over individual observations. In that case, an 'empirical' information matrix can be calculated from the sums of crossproducts individual scores (Meilijson, 1989). Corresponding orthogonal representations of the score function may be found for other cases. Klassen and Smith (1990) suggested estimating expected values of second derivatives for a given design by simulation. This is achieved by calculating BLUP solutions for random effects and quadratic forms in the vector of solutions for each replicate, and subsequently estimating the second derivatives required as coefficients in a multiple, linear regression of these quadratics on the population parameters.

Using First Derivatives

General methods which require first derivatives of the function to be optimized include various gradient algorithms and QN procedures. A scheme which is appropriate whenever there is missing data is the Expectation-Maximization (EM) algorithm of Dempster *et al.* (1977). In estimating variance components by (RE)ML the EM algorithm yields estimators equivalent to those obtained by setting first derivatives of log L to zero, i.e. can be thought of as exploiting information from first derivatives. These are generally considerably easier to calculate than second derivatives, and the EM algorithm has been used in the majority of REML applications in animal breeding. In contrast to the Newton methods, it is also guarantied to yield estimates within the parameter space without the need to impose constraints explicitly. Moreover, for some models the comparatively simple form of the resulting equations to be solved has facilitated the use of numerical techniques to reduce computational requirements substantially. There is a multitude of publications outlining the estimation of variance components via the EM algorithm for specific models (e.g. Laird and Ware, 1982; Dempster *et al.* 1984; Henderson, 1984; 1985a; 1986a; Taylor and Everett, 1985; Meyer, 1987a; Lin and Lee, 1989; Da *et al.*, 1989).

While algorithms using second derivatives are generally fast to converge, the EM algorithm can be described as a first-order successive approximation method with linear convergence at the

end of iterations (Laird *et al.*, 1987). Thus it can be very slow to converge, in estimating genetic parameters especially if heritabilities are low. There have been a number of attempts to improve convergence. Some yielded QN procedures requiring first derivatives and have been discussed above. Describing how to estimate the observed information matrix when using the EM algorithm, Louis (1982) suggested a modification which is essentially an application of Aitken's acceleration method. Laird *et al.* (1987) described the application of this technique to the estimation of covariance components by ML and REML. Examples given by Laird *et al.* (1987) and Lindstrom and Bates (1988) as well as an application to dairy data (Colleau *et al.*, 1989) showed that the acceleration could substantially reduce the number of iterations required.

Schaeffer's (1979) Common Intercept Approach (CIA) attempts to speed up the EM algorithm from differences between starting values and estimates from the last two rounds of iteration. Misztal and Schaeffer (1986) showed that the CIA is equivalent to assuming a non-linear model to describe the rate of convergence. Simianer (1988) examined the convergence of the EM algorithm for unmodified EM, the CIA, a procedure which, at each iterate, predicted converged values by fitting a secant through estimates for two adjacent starting values, and an iterated version of the CIA. For simulated data, the latter two, in essence approximating a NR algorithm using finite differences, converged after 5 and 4 EM iterations, while the former two had not achieved convergence after 100 rounds.

Thompson and Meyer (1986) showed that a reparameterization from variance components to terms derived from expectations of mean squares in corresponding ANOVAs, which makes the new parameters almost orthogonal, can improve rate of convergence of the EM algorithm dramatically. Harville and Callanan (1990) considered this reparameterization together with two others to 'linearize' the REML equations, demonstrating that it improves convergence for both EM type and NR algorithms. A 'shortcut' to the MSC, i.e. combining MSC estimation of sire covariances with EM estimation of residual covariances, for a multi-trait sire model with missing observations accelerated convergence greatly over the unmodified EM algorithm (Meyer, 1986). For a parameterization as suggested by Thompson and Meyer (1986) though, the EM algorithm performed almost as well as the 'shortcut', i.e. second derivatives, which describe the curvature of the likelihood surface, seem most useful when estimating parameters with high sampling correlations.

Derivative Free Methods

The optimum of a function can also be determined without knowing its derivatives. Louis (1982) commented on derivative free (DF) algorithms as an alternative to EM estimation. These numerical techniques range from direct search procedures, based on mere comparisons of function values, to methods which approximate first or even second derivatives (Gill *et al.*, 1981). The use of a DF approach for REML estimation of variance components has been described by Graser *et al.* (1987) for an AM. For a univariate analysis with animals as the only random effect, log L is maximized with respect to only one parameter, the ratio of error and additive genetic variance, using a quadratic approximation of the log likelihood function. They showed that the residual sum of squares and the determinant of the coefficient matrix in the mixed model equations, both required to evaluate the log L, can be determined in a general way through a series of Gaussian Elimination steps. At convergence, the error variance can be estimated directly from the residual sum of squares. Reducing the dimension of search by one by eliminating the error variance has been referred to by Harville and Callanan (1990) as 'concentrated likelihood' approach.

Derivative free REML estimation has been extended to models with additional random effects (Meyer, 1988 and 1989a) and multivariate analyses (Thompson and Juga, 1989; Meyer, 1989b). Using simulation, Meyer (1989a) compared the convergence of a quadratic approximation of the likelihood, both by least squares and numerical differentiation, a QN procedure approximating first and second derivatives, and a direct search procedure, the so-called Simplex method due to Nelder and Mead (1965). While the quadratic approximation performed consistently best for the single parameter case, it frequently yielded non-positive definite estimates of the Hessian matrix for several parameters. For most cases examined, the Simplex and QN procedures required about equivalent numbers of function evaluations to locate the maximum of log L. The Simplex proce-

dure, however, was easier to use and more robust against starting values far from the estimates. Furthermore, estimates could be constrained simply by assigning a very large (negative) function value to parameter vectors out of bounds (Nelder and Mead, 1965).

Further research is required to examine alternative search strategies. Simianer (1988) showed that it can be easier to locate the maximum of the likelihood function than of its logarithm. Box (1966) found the Simplex method to perform well up to 5 parameters while Powell's (1965) method of 'conjugate directions' performed better for higher dimensions of search. Preliminary experience with Powell's method suggests that it may be advantageous for multivariate analyses, in particular for models with more than one random effect.

SPECIALISED ALGORITHMS

Most general descriptions of (RE)ML procedures for estimating variance components, especially of those utilizing derivatives of the likelihood function, involve terms which require the inverse of the complete covariance matrix of the vector of observations. This makes their application for a large proportion of practical analyses in animal breeding computationally unfeasible. Thompson (1973, 1982) showed that REML estimation can be based on the mixed model equations, as given in Henderson *et al.* (1959) and widely used in the genetic evaluation of animals by BLUP, and Searle (1979) gave an extensive treatment of the matrix equalities involved. Even then, the direct inverse of the coefficient matrix for random effects after absorbing fixed effects, **C**, of order equal to the total number of random effects levels multiplied by the number of traits analyzed, and corresponding matrix products and traces are generally required in each round of iteration. This imposes severe limitations on both the size and model of analysis computationally feasible. Fortunately, there are a number of numerical techniques which can be exploited in conjunction with specific features of the data structure.

Transformation of the Mixed Model Equations

As noted by Patterson and Thompson (1971), the REML equations could be expressed in terms of latent roots of the covariance matrix of residuals. Dempster *et al.* (1984) described the use of a singular value decomposition of the MME when estimating variance components via the EM algorithm for a model with one random effect. For this strategy, the major computational burden imposed is the calculation of eigenvalues and eigenvectors of **C**, with similar restrictions on size as applicable to inversion, but this is required only once for each analysis and iterations are subsequently very fast. Along the same lines, Quaas and Smith (cited Taylor *et al.*, 1985) and Smith and Graser (1986) advocated the use of a Householder transformation to reduce **C** to tridiagonal form. In contrast to the calculation of eigenvalues, which is generally an iterative procedure, this requires a finite number of steps proportional to the number of levels of the random effect. Individual EM iterates can then be performed in linear time. Lin (1988) demonstrated that tridiagonalization and diagonalization of **C** produce the same iterates as direct inversion. Smith and Lin (1989) emphasized that in applying these techniques the eigenvectors do not need to be calculated explicitly, and summarized comparative operation counts and computing times required.

Smith and Graser (1986) showed that the reduction to tridiagonal form can be exploited for a broad range of models. For instance, two random effects can be accommodated by combining an EM type step to estimate variances due to one random effect and residuals, with a direct search for the maximum of the likelihood with respect to the variance due to the second random effect. Thompson and Meyer (1990) extended this application to EM estimation for a Reduced Animal Model (RAM; Quaas and Pollak, 1980) by introducing an auxiliary, imaginary effect with negative variance to make residual variances for parents and non-parents homogeneous. Furthermore, it can be used with an MSC algorithm as the trace of the square of C inverse required in this case, can be calculated from the tridiagonalized coefficient matrix (Colleau *et al.*, 1989; Smith and Lin, 1989).

Equivalent Models

Equivalent models, defined as having equal expectations and moments, can be employed to reduce computational requirements (Henderson, 1985b). A prominent example is the RAM, equivalent to the full AM, used widely in the genetic evaluation of pigs and beef cattle. Describing the estimation of non-additive genetic variances, Henderson (1985a) fitted a model with a total genetic value for each animal rather than additive, dominance and epistasis effects separately. Use of an equivalent model enabled Thompson and Meyer (1990) to exploit tridiagonalization of **C** for a RAM. For estimation under an AM, all animals without records and only single links to other animals via the relationship matrix can be absorbed or, equivalently, treated as if they were unknown. This reduces not only the number of random effects levels to be dealt with, but it has also been found to decrease the number of likelihood evaluations required for a DF algorithm. Henderson (1989) illustrated the equivalence of the AM and a sire-maternal grandsire model when data on dams are missing.

If random effects levels are correlated, i.e. relationships between animals are taken into account, analyses can be performed as if they were uncorrelated under an equivalent model where information on the relationship structure is incorporated into the design matrix for random effects (Quaas, 1984; Harville and Fenech, 1985; Smith and Graser, 1986). This involves a Cholesky decomposition of the numerator relationship matrix and multiplication of the original design matrix with the lower triangular Cholesky matrix **L**, or, equivalently, pre- and postmultiplication of **C** with **L** and **L'** and multiplication of the right hand side in the MME with **L**. **L** can be set up efficiently following rules given by Quaas (1976). Meyer (1987b) outlined a strategy to carry out multiplications of the MME without requiring additional storage, setting up one column of **L** at a time, while Mrode and Thompson (1989) subsequently described an alternative, computationally faster scheme.

Transformation of the Data

In special cases, a transformation applied to the data can reduce computational requirements for multivariate analyses; see Jensen and Mao (1988) for a more detailed review. Consider a model with one random effect and equal design matrices for all traits, i.e. all traits recorded for all animals at the same or strictly corresponding time(s). In this case, a canonical transformation (CT), derived from the residual and random effects covariance matrices, can be applied to yield new variables which are both genetically and phenotypically uncorrelated (e.g. Hill and Thompson, 1978; Hayes and Hill, 1980). This reduces the multivariate analysis to a series of corresponding univariate analyses (Thompson, 1977). Meyer (1985) presented a MSC algorithm for this case, and Colleau *et al.* (1989) outlined both MSC and NR procedures incorporating a tridiagonalization of the coefficient matrix. Corresponding EM algorithms with (tri)diagonalization of **C** have been described by numerous authors (e.g. Taylor *et al.*, 1985; Van Raden and Freeman, 1986; Schaeffer, 1986; Lin, 1987).

Use of the CT for DF algorithms allows log L to be calculated for one trait at a time, summing over traits. Moreover, terms involved are of interest in their own right, providing a reparameterization to heritabilities on the canonical scale and to elements of the transformation matrix, i.e. eigenvalues and pertaining eigenvectors of covariance matrices on the original scale. Thompson and Juga (1989) considered this for a bivariate analysis, employing a grid search with quadratic approximation of log L. More generally, it allows maximization of the likelihood in a nested two-step procedure. For each set of canonical heritabilities examined in the exterior step, a series of computationally demanding Gaussian Elimination steps, as outlined by Graser *et al.* (1987), is required. For given canonical heritabilities, maximization in the interior step with respect to the elements of the transformation matrix, however, involves scalar calculations only and is computationally inexpensive. Effectively, this reduces the dimension of search (in the exterior step) to the number of traits. Using the Simplex procedure, Meyer (1989b) showed that this strategy could dramatically reduce the number of likelihood evaluations required (compared to maximization with respect to the covariance components). For AMs with equal design matrices but additional random effects, the number of iterates or function evaluations can not be decreased. Nevertheless, eliminating genetic and error covariances through the CT reduces the number of non-zero off-diagonal

elements in the coefficient matrix and thus computational requirements in each set of Gaussian Elimination steps substantially.

A MSC algorithm for the case where residual covariances between traits are zero has been described by Schaeffer *et al.* (1978) for a sire model. This is the case, for example, when different traits are measured on different animals. In some instances, this constellation can also be achieved by transforming the data. If traits are recorded sequentially, the (inverse of the) Cholesky decomposition of the residual covariance matrix provides a linear transformation to variables with uncorrelated residuals (Schaeffer, 1986a). Garrick (1988) described an efficient computing strategy for this case using an algorithm requiring first derivatives of the likelihood function, and considered an extension for models with more than one random effect.

OTHER METHODS

Non-iterative Procedures

Two translation invariant methods closely related to REML, minimum norm (MINQUE) and minimum variance (MIVQUE) quadratic unbiased estimation, have been described by Rao (1972). In contrast to REML, they are non-iterative, i.e. estimates depend on the assumed (prior) values for the parameters, and no constraints on the parameter space are imposed. While MIVQUE assumes normality, no specific distribution of the data is required for MINQUE. There are strong links between ANOVA, REML, MINQUE and MIVQUE (Searle, 1979). Disregarding differences in constraints on the parameter space, MIVQUE is equal to one round of REML for a MSC algorithm. Under normality, MINQUE and MIVQUE are equivalent (Harville, 1977). Assuming prior values of zero for all components except residual variances, MINQUE is identical to estimation by Symmetric Differences Squared with weights proportional to the inverse of the residual covariance matrix (Keele and Harvey, 1989).

MIVQUE is locally best, i.e. has minimum variance only if priors are equal to the true variances. Hence, iterated MIVQUE (or MINQUE) and, for its constraints on the parameter space, REML are often preferred as the precision of estimates does not depend on the initial guess for the parameters to be estimated (Searle, 1989). In some situations, however, MIVQUE or MINQUE may be preferable to REML, not only for their smaller computational requirements. A typical example is the estimation of variances specific to a relatively small sample or subset of data, such as within a herd for dairy cattle data, when good estimates of the population values are available.

Approximate REML

A number of attempts have been made to derive approximate REML procedures, hoping to preserve most of its optimal properties but to reduce computational demands. Harville (1977) suggested replacing the REML quadratic forms by others resembling them when construction and direct solution of the MME was not feasible. Henderson's (1980) 'simple method' ignored non-zero off-diagonal elements in **C**, i.e. effectively replaced BLUP solutions of random effects and the corresponding quadratic(s) by their contemporary comparison counterparts. Henderson (1986b) described an approximate multitrait MIVQUE in which **C** and its inverse were approximated by considering diagonal blocks for animals and fixed effects together with the appropriate off-diagonal blocks.

Schaeffer (1986b) outlined a 'pseudo-REML' procedure which replaced quadratic forms in the vector of BLUP solutions for random effects, **u**, by bilinear forms in **u** and the corresponding vector of right hand sides, taking expectations of the latter pretending the true variances were known. A related method, referred to as the 'tilde-hat' approach was described by Van Raden and Jung (1988). Simulating unselected data, Schaeffer (1986b) reported similar means over replicates and an empirical correlation close to unity between REML and pseudo-REML estimates with a slightly higher sampling variance for the latter. However, Ouweltjes *et al.* (1988) demonstrated that approximate REML procedures are biased by selection when REML is not.

Bayesian Estimation

An extensive review of Bayesian methods in animal breeding, including their application in estimating genetic parameters, has been given by Gianola and Fernando (1986). The most distinctive feature of the Bayesian approach is that it allows prior information to be incorporated in the analysis, basing inferences and decisions on posterior distributions. It is comforting to realize that the statistical procedures widely used at present, BLUP and (RE)ML estimation, have a Bayesian interpretation.

Assuming flat priors for fixed effects and variance components and 'integrating out' random effects, the mode of the joint posterior density of fixed effects and variances is their ML estimate. Integrating out fixed effects in addition, i.e. maximizing the marginal density, then yields REML estimates of the variance components. Predicting random effects when variances are unknown, Gianola *et al.* (1986) showed that an iteratve scheme estimating the variances by REML yielded Empirical Bayes estimators. Alternative procedures, using non-informative priors, may provide estimators of variance components with smaller mean square errors although these may be biased (Harville, 1977; Gianola and Fernando, 1986). Foulley (1990) discussed some of the alternatives to REML and BLUP. In the animal breeding context, Bayesian estimation has so far primarily been considered for categorical data. To date, no explicit guide-lines are available to indicate when Bayesian procedures may be preferable over REML in the analysis of quantitative traits.

PROBLEMS AND FUTURE RESEARCH

Selection

As discussed above, REML can account for selection and yield estimates of genetic parameters in the unselected population. In essence, all information which has contributed to selection decisions must be available and utilized in the analysis. In some circumstances this may not be sufficient, e.g. translation invariance of the selection criteria may be required (Schaeffer, 1987). Gianola *et al.* (1988) discussed the 'ignorability' of selection in the prediction context, and Im *et al.* (1988 and 1989) considered likelihood inferences under selection. Yet, as emphasized by Foulley (1990), the effects of selection on estimates of genetic parameters are still under dispute.

Mixed model methodology, in particular REML estimation under an AM, has been advocated for the optimal analysis of data from selection experiments (e.g. Kennedy, 1990). Simulation demonstrated that the AM with complete relationships would account for inbreeding and selection and give estimates of the base population genetic parameters, even if some data from earlier generations was omitted (Sorensen and Kennedy, 1984; Jensen and Mao, 1989). Further work, however, showed that the latter does not hold, i.e. all data has to be included, although the observed selection bias was considerably less than expected from normal theory (Van der Werf, 1990). For analyses of dairy data under a sire model, proven bulls without information on their first batch of daughters have commonly been treated as fixed in order to avoid a biased estimate of the variance between sires (Van Vleck, 1985). Similarly, Graser *et al.* (1987) suggested treating selected base animals in the AM as fixed and described a simple way to do so. Simulating a multi-generation selection experiment with missing records for earlier generations, however, identified problems with this approach. Provided inbreeding was correctly taken into account, the estimate of the variance within full-sib families was unbiased, while the additive genetic variance was biased downwards and the error variance accordingly biased upwards (Van der Werf, 1990). Clearly, the mechanisms involved are not fully understood yet, especially how information within and across families and generations is combined.

Shape of the Likelihood Surface

There are a number of problems and questions relating to the geography of the likelihood surface. As discussed above, the performance of a maximization algorithm for a specific model can be affected considerably by the parameterization chosen. If there is a large negative sampling correlation between two parameters, the likelihood surface has a ridge, i.e. log L changes very little

as long as the sum of the two components remains constant, and its maximum is difficult to locate. Parameterizations which reduce correlations between parameters have been described for sire models (Thompson and Meyer, 1986; Harville and Callanan, 1990), and corresponding representations need to be determined for AMs and multivariate analyses. For instance, for an AM with litters as an additional, permanent environmental random effect, it appears to be advantageous to maximize log L with respect to the within family variance, or the sum of (half) additive genetic and litter variance, rather than the litter variance directly. For multivariate problems, maximization with respect to the Cholesky decomposition of the covariance matrices has been recommended to improve speed of convergence (Lindstrom and Bates, 1988).

This issue is closely related to the question of experimental design for REML estimation. Criteria, similar to the determinant of the coefficient matrix in least-squares analyses, need to be identified to assess the amount of information available in a particular analysis readily. As interest in more detailed models increases, e.g. AMs including maternal or non-additive genetic effects, guide-lines need to be developed to ensure a data structure which will allow all components to be estimated. The linear model framework, in particular estimation under an AM using a DF algorithm, makes it easy to overparameterize the model of analysis. Thompson (1986) outlined the information contributed by different types of relatives in the ML estimation of maternal genetic variances. For these models especially, little is known about the number of maxima of the likelihood function, the chances of convergence to a local rather than the global maximum and which algorithms are best when several maxima exist. Hoeschele (1989) indicated that the REML likelihood always has a single maximum over the permissible parameter space for a model with one random effect, estimating two variance components. Unless there was some source of bias, simulation studies for the simpler models generally yielded mean REML estimates equal to the population parameter, suggesting that convergence to a local maximum was rare. Problems have been reported, however, for multivariate analyses using a DF algorithm (Kovac and Groeneveld, 1989).

The likelihood approach allows for hypothesis testing through the likelihood ratio test (LRT; Kendall and Stuart, 1973). This requires evaluation of log L under both the null and alternative hypothesis. Minus twice the difference in log L then has an asymptotic Chi-Squared distribution with degrees of freedom corresponding to the number of parameters tested and can be contrasted to table values. Foulley *et al.* (1990) advocated a LRT based on the marginal rather than full likelihood, to test for sources of heterogeneity in a model with heterogeneous variances. Properties of the LRT as applied to REML estimates of variances are largely unexplored. For small samples of simulated data, Shaw (1987) found the power of the LRT to be considerable less than expected when testing for an estimate of the genetic variance different from zero.

The asymptotic covariance matrix of (RE)ML estimates is given by the inverse of the pertaining information matrix, i.e. the matrix of expected values of second derivatives of log L. Some techniques to approximate this matrix have been discussed above. Meyer (1989a) found good agreement between sampling variances approximated by numerical differentiation of the likelihood function and empirical variances for a simple AM, while the approximation failed for AMs including a maternal genetic effect. In cases where the shape of the likelihood surface does not allow reliable estimates of sampling variances to be obtained, a grid of likelihood values around the point estimates may provide a good guide to their precision. Confidence intervals for REML estimates have been considered by Harville and Callanan (1990).

Computational Aspects

Even with highly specialized and efficient algorithms and modern day computers, REML estimation of genetic parameters is often limited by computational requirements. Judicious implementation and exploitation of specific features of hardware and software available can often considerably extend the range of analyses feasible. Graser *et al.* (1987) showed that the use of sparse matrix storage techniques with a DF algorithm allowed AMs with thousands of animals to be handled. Misztal (1989) considered the use of sparse matrix library routines for an EM type REML algorithm. Tier and Smith (1989) illustrated the use of so-called 'linked lists' in mixed model applications. In using sparse matrix techniques, the ordering of equations is of crucial importance

in minimizing 'fill-in', i.e. the number of additional non-zero elements created, and thus computations required. A number of computing systems allow code to be 'vectorized', i.e. operations to be performed for all elements of a vector simultaneously, or almost so. If programs can be adapted to utilize this feature, computing time required may be reduced substantially. Some machines offer parallel processing facilities. Relative performance of different algorithms then depends heavily on the extent to which they can make use of the parallel environment.

CONCLUSIONS

Accurate estimation of genetic parameters is one of the fundamental tasks of applied quantitative genetics. Since the first World Congress, we have seen a transgression from least squares type estimation, interpreting covariances between relatives for a single type of relationship, to ML estimation exploiting information from all relatives simultaneously. To date, REML estimation fitting an AM appears to represent the 'state of art' though developments and practical applications are still quite immature. Derivative free algorithms are highly flexible, accommodating a wide range of models of interest in the analysis of animal breeding data. Future advances are required to make routine applications for AMs with additional random effects and multivariate analyses feasible. There remains considerable scope for further improvement through an optimal combination of parameterization, maximization technique and numerical transformations for specific types of analysis.

Conditions under which alternative, e.g. Bayesian, estimators are better than REML need to be established. We need to gain a better understanding of the relative importance of different sources of information for AM analyses, and guide-lines for experimental design for REML estimation need to be developed. For models fitting parental or non-additive genetic effects in particular, it is of crucial importance to ensure that there is sufficient information on all parameters to be estimated; even with the best method of analysis, our estimates can only be as good as the data they are based on.

REFERENCES

ANDERSON, T.W. 1973. *Ann. Statist.* **1** : 135-141.
BANKS, B.D., MAO, I.L. and WALTER, J.P. 1985 *J. Dairy Sci.* **68** : 1785-1792.
BOX, M.J. 1966. *Computer J.* **9** : 67-77.
CHANG, A. 1988. Ph.D. Thesis, University of Illinois, Urbana-Champaign.
COLLEAU, J.J., BEAUMONT, C. and REGALDO, D. 1989. *Livest. Prod. Sci.* **23** : 47-66.
CORBEIL, R.R. and SEARLE, S.R. 1976. *Biometrics* **32** : 779-792.
CUE, R.I. 1986. *J. Anim. Breed. Genet.* **103** : 334-341.
CURNOW, R.N. 1961. *Biometrics* **17** : 553-566.
DA, Y., GROSSMAN, M. and MISZTAL, I. 1989. *J. Dairy Sci.* **72** : 2125-2135.
DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. 1977. *J. Roy. Statist. Soc. B* **39** : 1-38.
DEMPSTER, A.P., SELWYN, M.R., PATEL, C.M. and ROTH, A.J. 1984. *Appl. Stat.* **33** : 203-214.
DENNIS, J.E. and MORE, J.J. 1977. *SIAM Rev.* **19** : 46-89.
EFRON, B. and HINKLEY, D.V. 1978. *Biometrika* **65** : 457-488.
FOULLEY, J.L. 1990. In *Proc. 4th World Congr. Genet. Appl. Livest. Prod., Edinburgh,* (in press).
FOULLEY, J.L., GIANOLA, D., SAN CRISTOBAL, M. and IM, S. 1990. *J. Dairy Sci.* (in press).
GARRICK, D.J. 1988. Ph.D. Thesis, Cornell University, Ithaca, NY.
GIANOLA, D. and FERNANDO, R.L. 1986. *J. Anim. Sci.* **63** : 217-244.
GIANOLA, D., FOULLEY, J.L. and FERNANDO, R.L. 1986. *Genet. Sel. Evol.* **18** : 485-498.
GIANOLA, D., IM, S. and FERNANDO, R.L. 1988. *J. Dairy Sci.* **71** : 2790-2798.
GILL, P.E., MURRAY, W. and WRIGHT, M.H. 1981. Academic Press, New York.
GRASER, H.-U., SMITH, S.P. and TIER, B. 1987. *J. Anim. Sci.* **64** : 1362-1370.
HARVEY, W.R. 1972. Mimeograph, Ohio State University.
HARVILLE, D.A. 1977. *J. Amer. Stat. Ass.* **72**: 320-340.
HARVILLE, D.A. and FENECH, A.P. 1985. *Biometrics* **41** : 137-152.

HARVILLE, D.A. and CALLANAN, T.P. 1990. In *Advances in Statistical Methods for Genetic Improvement of Livestock,* D. Gianola and K. Hammond, eds, Springer Verlag (in press).
HAYES, J.F. and HILL, W.G. 1980. *Biometrics* **36** : 237-248.
HENDERSON, C.R. 1949. *J. Dairy Sci.* **32** : 706 (Abstr.).
HENDERSON, C.R. 1953. *Biometrics* **9** : 226-252.
HENDERSON, C.R. 1976. *Biometrics* **32** : 69-83.
HENDERSON, C.R. 1980. *J. Dairy Sci.* **58** Suppl. 1 : 119 (Abstr.)
HENDERSON, C.R. 1984. *J. Dairy Sci.* **67** : 1581-1589.
HENDERSON, C.R. 1985a. *J. Anim. Sci.* **61** : 113-121.
HENDERSON, C.R. 1985b. *J. Dairy Sci.* **68** : 2267-2277.
HENDERSON, C.R. 1986a. *J. Dairy Sci.* **69** : 1394-1402.
HENDERSON, C.R. 1986b. *J. Anim. Sci.* **63** : 208-216.
HENDERSON, C.R. 1988. *J. Dairy Sci.* **71** Suppl. 2 : 1-16.
HENDERSON, C.R. 1989. *J. Dairy Sci.* **72** : 2592-2605.
HENDERSON, C.R., KEMPTHORNE, O., SEARLE, S.R. and v. KROSIGK, C.M. 1959. *Biometrics* **15** : 192-218.
HENDERSON, C.R. and QUAAS, R.L. 1977. *J. Anim. Sci.* Suppl.1 : 22-23 (Abstr.).
HILL, W.G. 1974. In *Proc. 1st World Congr. Genet. Appl. Livest. Prod., Madrid,* Vol. **IA** : 343-351.
HILL, W.G. and THOMPSON, R. 1978. *Biometrics* **34** : 429-439.
HILL, W.G. and MEYER, K. 1988. In *Animal Breeding Opportunities,* G. Wiener, ed., Brit. Soc. Anim. Prod. and Poul. Breed. Roundtable, Durham, U.K.
HOCKING, R.R. and KUTNER, M.H. 1975. *Biometrics* **31** : 19-28.
HOESCHELE, I. 1989. *J. Statist. Comput. Simul.* **33** : 149-160.
HOESCHELE, I. and VAN RADEN, P.M. 1989. *J. Anim. Sci.* **67** / *J. Dairy Sci.* **72** Suppl.1 : 30 (Abstr.).
IM, S., FERNANDO, R.L. and GIANOLA, D. 1988. *J. Dairy Sci.* **71** Suppl. 1 : 262 (Abstr.).
IM, S., FERNANDO, R.L. and GIANOLA, D. 1989. *Genet. Sel. Evol.* **21** : 399-414.
JENNRICH, R.T. and SAMPSON, P.F. 1976. *Technometrics* **18** : 11-17.
JENSEN, J. and MAO, I.L. 1988. *J. Dairy Sci.* **66** : 2750-2761.
JENSEN, J. and MAO, I.L. 1989. *J. Anim. Sci.* **67** / *J. Dairy Sci.* **72** Suppl. 1 : 33 (Abstr.).
KEELE, J.W. and HARVEY, W.R. 1989. *J. Anim. Sci.* **67** : 348-356.
KENDALL, M.G. and STUART, A. 1973. Macmillan, NY.
KENNEDY, W.J. and GENTLE, J.E. 1980. Marcel Dekker Inc., New York.
KENNEDY, B.W. 1990. In *Advances in Statistical Methods for Genetic Improvement of Livestock,* D. Gianola and K. Hammond, eds, Springer Verlag (in press).
KENNEDY, B.W., SCHAEFFER, L.R. and SORENSEN, D.A. 1988. *J. Dairy Sci.* **71** Suppl. 2 : 17-26.
KLASSEN, D. J. and SMITH, S.P. 1990. In *Proc. 4th World Congr. Genet. Appl. Livest. Prod., Edinburgh,* (in press).
KOVAC, M. and GROENEVELD, E. 1989. *J. Anim. Sci.* **67** / *J. Dairy Sci.* **72** Suppl.1 : 33-34 (Abstr.).
LAIRD, N.M. and WARE, J.H. 1982. *Biometrics* **38** : 963-974.
LAIRD, N.M., LANGE, N. and STRAM, D.O. 1987. *J. Amer. Stat. Ass.* **82** : 97-105.
LIN, C.Y. 1987. *J. Dairy Sci.* **70** : 2680-2684.
LIN, C.Y. 1988. *J. Anim. Sci.* **66** : 1627-1635.
LIN, C.Y. and McALLISTER, A.J. 1984. *J. Dairy Sci.* **67** : 2389-2398.
LIN, C.Y. and LEE, A.J. 1989. Can. *J. Anim. Sci.* **69** : 61-68.
LINDSTROM, M.J. and BATES, D.M. 1988. *J. Amer. Stat. Ass.* **83** : 1014-1022.
LOUIS, T.A. 1982. *J. Roy. Statist. Soc. B* **44** : 226-233.
LUSH, J.L. and SHRODE, R.R. 1950. *J. Dairy Sci.* **33** : 338-347.
MEILIJSON, I. 1989. *J. Roy. Statist. Soc. B* **51** : 127-138.
MEYER, K. 1983. *J. Dairy Sci.* **66** : 1988-1997.
MEYER, K. 1985. *Biometrics* **41** : 153-166.
MEYER, K. 1986. *J. Dairy Sci.* **69** : 1904-1916.
MEYER, K. 1987a. *Genet. Sel. Evol.* **19** : 49-68.
MEYER, K. 1987b. *J. Anim. Breed. Genet.* **104** : 163-168.

MEYER, K. 1988. *J. Dairy Sci.* **71** Suppl.2 : 33-34 (Abstr.).
MEYER, K. 1989a. *Genet. Sel. Evol.* **21** :317-340.
MEYER, K. 1989b. *Genet. Sel. Evol.* (submitted).
MEYER, K. and THOMPSON, R. 1984. *J. Anim. Breed. Genet.* **101** : 33-50.
MISZTAL, I. 1989. *J. Anim. Sci.* **67** / *J. Dairy Sci.* **72** Suppl.1 : 30 (Abstr.).
MISZTAL, I. and SCHAEFFER, L.R. 1986. *J. Dairy Sci.* **69** : 2209-2213.
MRODE, R. and THOMPSON, R. 1989. *J. Anim. Breed. Genet.* **106** : 89-95.
NELDER, J.A. and MEAD, R. 1965. *Computer J.* **7** : 147-151.
PATTERSON, L.D. and THOMPSON, R. 1971. *Biometrika* **58** : 545-559.
OUWELTJES, W., SCHAEFFER, L.R. and KENNEDY, B.W. 1988. *J. Dairy Sci.* **71** : 773-779.
POWELL, M.J.D. 1965. *Computer J.* **7** : 155-162.
QUAAS, R.L. 1976. *Biometrics* **32** : 949-953.
QUAAS, R.L. 1984. In *BLUP School Handbook,* A.G.B.U., Univ. New England, Armidale, NSW., p.1-76.
QUAAS, R.L. and POLLAK, E.J. 1980. *J. Anim. Sci.* **51** : 1277-1287.
RAO, C.R. 1972. *J. Amer. Statist. Ass.* **67** : 112-115.
REDNER, R.A. and WALKER, H.F. 1984. *SIAM Rev.* **26** : 195-239.
ROBERTSON, A. 1977. *J. Anim. Breed. Genet.* **94** : 131-135.
ROBINSON, A. 1988. Ph.D. Thesis, Cornell University, Ithaca, NY.
ROTHSCHILD, M.F., HENDERSON, C.R. and QUAAS, R.L. 1979. *J. Dairy Sci.* **62** : 996-1002.
SCHAEFFER, L.R. 1979. In *Proc. Conf. in Honor of C.R. Henderson on Variance Components in Animal Breeding,* p.123-137.
SCHAEFFER, L.R. 1986a. *J. Dairy Sci.* **69** : 187-194.
SCHAEFFER, L.R. 1986b. *J. Dairy Sci.* **69** : 2884-2889.
SCHAEFFER, L.R. 1987. *J. Dairy Sci.* **70** : 661-671.
SCHAEFFER, L.R., WILTON, J.W. and THOMPSON, R. 1978. *Biometrics* **34** : 199-208.
SCHAEFFER, L.R., KENNEDY, B.W. and GIBSON, J.P. 1989. *J. Dairy Sci.* **72** : 1266-1272.
SEARLE, S.R. 179. Paper BU-673-M, Biometrics Unit, Cornell University, NY.
SEARLE, S.R. 1989. *J. Anim. Breed. Genet.* **106** : 1-29.
SHAW, R.G. 1987. *Evolution* **41** : 812-826.
SIMIANER, H. 1988. *J. Anim. Breed. Genet.* **104** : 334-339.
SMITH, S.P. 1984. Mimeograph, Ohio State University (unpublished).
SMITH, S.P. and GRASER, H.-U. 1986. *J. Dairy Sci.* **69** : 1156-1165.
SMITH, S.P. and LIN, C.Y. 1989. *J. Dairy Sci.* (submitted).
SMITH, S.P. and MAKI-TANILA, A. 1989. *Genet. Sel. Evol.* (in press).
SORENSEN, D.A. and KENNEDY, B.W. 1984. *J. Anim. Sci.* **59** : 1213-1223.
SWALLOW, W.H. and MONAHAN, J.F. 1984. *Technometrics* **26** : 47-57.
TAYLOR, J.F., BEAN, B., MARSHALL, C.E. and SULLIVAN, J.J. 1985. *J. Dairy Sci.* **68** : 2703-2722.
TAYLOR, J.F. and EVERETT, R.W. 1985. *J. Dairy Sci.* **68** : 2948-2953.
THOMPSON, R. 1973. *Biometrics* **29** : 527-550.
THOMPSON, R. 1976. *Biometrics* **32** : 903-918.
THOMPSON, R. 1977. In *Proc. Int. Conf. Quant. Genet., Ames, Iowa, p.*639-657.
THOMPSON, R. 1982. In *Proc. 2nd World Congr. Genet. Appl. Livest. Prod., Madrid,* Vol. **V** : 95-103.
THOMPSON, R. and CAMERON, N.D. 1986. In *Proc. 3rd World Congr. Genet. Appl. Livest. Prod., Lincoln, NE,* Vol. **XII** : 371-381.
THOMPSON, R. and MEYER, K. 1986. *J. Statist. Comp. Simul.* **24** : 215-230.
THOMPSON, R. and JUGA, J. 1989. *Acta Agric. Scand.* (submitted).
THOMPSON, R. and MEYER, K. 1990. *Genet. Sel. Evol.* **22** : (in press).
TIER, B. and SMITH, S.P. 1989. *Genet. Sel. Evol.* : (in press).
VAN DER WERF, J.H.J. 1990. Ph.D. Thesis, University of Wageningen, NL.
VAN RADEN, P.M. and FREEMAN, A.E. 1986. *J. Dairy Sci.* **69** Suppl. 1 : 208 (Abstr.).
VAN RADEN, P.M. and JUNG, Y.C. 1988. *J. Dairy Sci.* **71** : 187-194.
VAN VLECK, L.D. 1985. *J. Dairy Sci.* **68** : 2396-2402.
WALTER, J.P. and MAO, I.L. 1985. *J. Dairy Sci.* **68** : 91-98.
WESTFALL, P.H. 1987. *J. Amer. Stat. Ass.* **82**: 866-874.