A REVIEW OF GENETIC PARAMETER ESTIMATION

R. Thompson

IACR-Rothamsted Experimental Station, Harpenden, Herts AL5 2JQ, England
and Roslin Institute(Edinburgh), Roslin, Midlothian EH25 9PS, Scotland

INTRODUCTION

We review genetic variance parameter estimation. We concentrate on non-Bayesian estimation. Bayesian methods will be covered in a companion paper by Professor Gianola. My prior belief is that his paper will be a *tour-de-force*. I am more interested in concentrating on obtaining appropriate data and fitting appropriate models rather than using a paradigm that insists on using prior beliefs however (in)substantial. Given the context it is interesting to review the progress in terms of previous Congresses.

Madrid 1982. The comparisons between parameter estimation by fitting different models and equating sums of squares to expectation with estimation by Maximum Likelihood(ML) or Residual Maximum Likelihood (REML) was emphasised. The ML methods were more efficient. The convenient algorithms for the inverse of the Additive Relationship Matrix (Henderson, 1976) allowed in principle use of all covariances between relatives. In some circumstances ML methods took account of selection. There was a *cri du coeur* for help in understanding the L'y selection of Henderson (1975).

Lincoln 1986. ML methods because of the unbalanced nature of the data normally require iterative methods to maximize likelihoods. An important development was the introduction by Smith and Graser (1985) of an alternative form for the likelihood that naturally leads to sequential formation of the likelihood that required much less computation than existing methods at the time. To maximize the likelihood with one parameter Smith and Graser (1985) suggested using a quadratic approximation.

Edinburgh 1990. With more than one parameter, simplex methods become a popular flexible alternative as they avoid calculating derivatives and again existing methods were computationally expensive. The methods were used for Animal, Reduced Animal Models as well as Sire Models, both for univariate and multivariate data (Meyer, 1989). Later more biological appropriate models with genetic components naturally fitted into their framework including maternal models and models with mutation terms (Wray, 1990).

Guelph 1994. This congress saw interesting work presented on maximisation of likelihoods. Numerical methods for maximisation were popular but this became more difficult with more parameters. We consider a linear model $y = Xb + Zu + e$

The residual log-likelihood (REML) is of the form :

$$L \quad \alpha \quad (y - X\hat{b})'V^{-1}(y - X\hat{b}) - \text{logdet}(V) - \text{logdet}(X'V^{-1}X)$$

This is different from the usual likelihood form in that it is a function of error contrasts – contrasts that do not tell us about fixed effects. This difference has two consequences, the use of the weighted least squares estimate $\hat{b}$ of *b*, given by $X'V^{-1}X\hat{b} = X'V^{-1}y$, and a term in $\det(X'V^{-1}X)$ that is sometimes thought of as a penalty function because the fixed effects are

not known. Mixed model equations (Henderson, 1973) pay an important part in the analysis process. These are of the form

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Terms derived from these include prediction error variances found from writing the mixed model equations as $Cs = R$ so that $\text{var}\begin{bmatrix} \hat{b} \\ \hat{u} - u \end{bmatrix} = C^{-1}$

It is often useful to express relevant quantities in terms of the projection matrix

$$P = V^{-1} - X(X'V^{-1}X)^{-1}X'V^{-1}$$

$$L \quad \alpha \quad y'Py - \text{logdet}(V) - \log(X'V^{-1}X)$$

Estimation of a variance parameter θ_i involves setting to zero the first derivatives

$$\partial L/\partial\theta_i = y'P(\partial V/\partial\theta_i)Py - \text{tr}\left[P(\partial V/\partial\theta_i)\right]$$

These could be thought of equating a function of the data to its expectation. Normally finding a maximum of the likelihood requires an iterative scheme. One suggested by Patterson and Thompson (1971) is based on the expected value of the second differential. Using the first and second differentials we can update θ using the rate that all the terms from solution of MME and C^{-1} for example

$$\hat{\theta} = \theta + \text{EInf}^{-1}(\partial L/\partial\theta).$$

An alternative algorithm was suggested by Dempster, Laird and Rubin (1977). This EM algorithm is based on thinking of the random effects as `missing' .The estimation is based on using $s\hat{\sigma}_g^2 = u'u + \text{PEV}(u)$ writing this as

$$s\hat{\sigma}_g^2 = y'V^{-1}(\partial G/\partial\theta_i)V^{-1}y + s\sigma_g^2 - \text{tr}[V^{-1}(\partial G/\partial\theta_i)],$$

we see this as a manipulation of equating the first differential to zero. It can be also written as $\hat{\theta} = \theta + \text{Inf}^{-1}(\partial L/\partial\theta)$ with Inf representing the information on the complete data. One advantage of this method is that σ_g^2 stays in the parameter space $\sigma_g^2 \geq 0$.

Another advantage is that there is an increase in likelihood in each iteration. Disadvantages are that the method can be slow to converge (indeed this method is said to be the most widely used in terms of numbers of iterations) and it requires the inversion of C in each iteration. An important advance was the rediscovery (Misztal and Perez-Enriso, 1993) of an algorithm (Takahashi *et al*., 1973) that allowed the calculation of the `relevant' terms in the inverse of C required for forming the first differentials without calculating all the elements of the inverse. Meyer and Smith (1996) introduced an alternative way of calculating these first differentials by performing the `automatic' differentiation of the Choleski decomposition of C. These techniques both requiring twice the computational effort of forming the likelihood were derived using properties of Choleski decompositions. An alternative derivation in terms of

sequential formation of C^{-1} parallels the sequential formation of the likelihood (Thompson *et al*., 1994).This result allowed the implementation of EM algorithms to estimate variance parameters, (Misztal, 1994) for bigger problems. These were an improvement on derivative free methods but could still be slow to converge.

It is possible to calculate second differentials using the automatic differentiation ideas of Smith (1995) but the calculation of each second differential requires the computation of the order of six likelihood calculations (Smith, 1995). There are various suggestions on approximating the second differential. Mantysaari and Van Vleck (1989) suggest accelerating the EM algorithm based on the observed geometric rate of convergence. Neumaier and Groeneveld (1998) suggest quasi-Newton scheme using first differential values to build up an approximate second differential. A third suggestion by Thompson and co-workers (Johnson and Thompson, 1995 ; Gilmour *et al*., 1995 ; Jensen *et al*., 1997) is based on manipulation of the alternative information matrices. The second differentials of C with respect to θ_i and θ_j. are

$$(\partial^2 L/\partial\theta_i\partial\theta_j) = (1/2)\mathrm{tr}[P(\partial V/\partial\theta_i)P(\partial V/\partial\theta_j)] - y'P(\partial V/\partial\theta_i)P(\partial V/\partial\theta_j)Py$$

and $$E[(\partial^2 L/\partial\theta_i\partial\theta_j)] = -(1/2)\mathrm{tr}[P(\partial V/\partial\theta_i)P(\partial V/\partial\theta_j)]$$

Both these terms often called observed and expected information are difficult to calculate but the average $AI[(\partial^2 L/\partial\theta_i\partial\theta_j)] = -(1/2)y'P(\partial V/\partial\theta_i)P(\partial V/\partial\theta_j)Py$ can be calculated by using $(\partial V/\partial\theta_i)Py$ and $(\partial V/\partial\theta_j)Py$ as working variables and obtaining the residual cross-product between these working variables. This calculation is much simpler than calculating either the observed and expected information.

Armidale 1998. The major extensions I saw at Armidale was extension of models more into the area of longitudinal data especially with random regression methods. There was also discussion of the Method R introduced by Reverter (Reverter *et al.*, 1994a ; 1994b) One problem with the ML techniques is the dependence of the procedures on having estimates of the prediction error variances. Method R introduced an ingenious suggestion that avoids calculation of prediction error variances. The initial suggestion was for a way of checking 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 the recent data change the prediction of early animals .In a sense this is looking backwards. By contrast the analysis of selection experiments and partitioning of likelihood often are, in one sense, looking forward and asking 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 1. This method has been used in several large genetic situations (for example : Misztal, 1997 ; van Tassell *et al*., 1999 ; Duagjinda *et al*., 2001a) but the method is not completely understood. Reverter *et al.* (1994b) suggest constructing predictions from all the data (`recent') with predictions on random sub-

samples (`early'). They suggest using sub-samples of 50 % from empirical evidence. Analytical consideration of half-sib data suggests that the sampling variance will have an asymptotic term and a term dependent on the number of samples.

A concern is whether or not the method can be used with selected data. Cantet and Birchmeier (1998) suggested that this was not so, but their verbal presentation contradicted this assertion. More recently simulation evidence (Cantet *et al*., 2000 ; Schenkel and Schaeffer, 2000 ; Duagjinda *et al*., 2001b) have suggested that method R estimates are biased in selected populations when samples based on 50 % of the data. Whilst I accept that the regression of `early' on `recent' is one, however, I doubt this will be the case if `recent' is a random sub-sample. For example, with dam-daughter pairs and selection in the parental generation random samples of dam-daughter data might choose (a) dam alone, (b) daughter alone, (c) dam and daughter together. In case (a) regression is a function of regression of daughter on dam and is unaffected by selection. However, in case (b) the regression is a function of regression of dam on daughter and is affected by selection. Case (c) is essentially uninformative on estimation of heritability. One presumably can avoid the difficulty with case (c) by using several sub-samples. To circumvent problems with selection, one can presumably use a more sequential approach similar to one suggested for discrete data. There is the question about the efficiency of the two suggested schemes and the optimal weighting of information from animals born in different time periods.

MORE RECENTLY

A synthesis of comparisons of these iterative methods was carried out by Hofer (1998) and is updated in table 1. These show the expected improvement of EM methods over derivative free methods. They also show that most second differential methods converge in relatively small number of iterations.

In some cases transformations can aid in estimation. If we have multivariate data with two (pxp) variance matrices to estimate, say G and R, then a canonical transformation (Meyer, 1986 ; 1997) can help in reducing one pxp estimation into p independent analyses. They are modifications using the EM algorithm that allow the same techniques to be used with missing values (Ducrocq and Besbes, 1993) and with unequal design matrices (Ducrocq and Chapuis, 1997).

A related problem is that often we require G and 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 for example σ or log σ instead of σ^2, or multivariate analogues such as Choleski transformations (Lindstrom and Bates, 1988 ; Groenevald, 1994).

Recent work on EM algorithms (Foulley and Quaas, 1995 ; Meng and Van Dyk, 1998) have suggested that this Choleski or linear parameterization has a natural interpretation and can lead to faster convergence.

For example, Foulley and Quaas (1995) use a model $y = X\alpha + \sigma_G Zu^* + e$ and given σ predict *u* with natural mixed model equations. Regression of *y* on σ_G and Zu* (taking into account uncertainty of *u*) gives a natural way of updating σ_G (keeping σ^2_G within the parameter space). For a balanced sire model Foulley and Quaas (1995) note that the rate of convergence depends

on (n / (n + α)) with $\alpha = \sigma^2_E / \sigma^2_G$. For (n / (n + α)) = 0.2, 0.5, 0.8 the rates of convergence for σ_G using an EM algorithm are 0.27, 0.45, and 0.31, compared with 0.03, 0.25, and 0.63 for a scheme based on updating σ^2_G, showing the advantage of the σ_G parameterization for small values of (n / (n + α)).

Table 1. Results of empirical comparison of REML algorithms with regards to rounds of iteration (function evaluations for DF) and total time (h) to convergence [A]

Ref [B]	MME [C]	Par [D]	DF		EM		NR/AI [F]	
			F.Eval	Time	Rounds	Time	Rounds	Time
1	4895	3	26	0.01	24	0.05		
	9790	9	238	0.31	33	0.26		
	14685	18	583	1.77	45	1.02		
2	6192	9	699	1.27			6	0.45
	10230	12	1236	2.33			8	0.90
	14274	18	4751	11.10			18	3.33
3	5731	5	169	0.34			6	0.07
4	8765	6	927	70.60	109	4.91	7	1.86
5	5073	2	39	0.02			5	0.02
	10146	6	472	0.52			9	0.09
6[E]	233796	55	37021	20830			185	40.10
7	46581	12	1435	15.20	1006	88.60	6	0.58
	55410	19	5813	30.60			6	1.00

A Updated from Hofer (1998).

B References 1 Misztal (1999) ; 2 Meyer and Smith (1996) ; 3 Johnson and Thompson (1995) ; 4 Gilmour *et al.* (1995) ; 5 Madsen *et al.* (1994) ; 6 Neumaier and Groeneveld (1998) ; 7 Jansen *et al.* (1997).

C Dimension of mixed model equations (MME).

D Number of (co)variance components.

E `DF' = quasi Newton using finite differences.

F `NR/AI' = quasi-Newton using computed analytic differences.

A more recent development is the suggestion of Lui *et al*. (1998) who suggest a parameter extension or PX-EM algorithm. In our case it involves estimating $\sigma^2_G = (\sigma_{G1})^2\ \sigma^2_{G2}$ and σ_{G1} estimated by the linear scheme and σ^2_{G2} by the quadratic scheme. This scheme at first sight counter-initiative in that σ^2_{G2} are confounded, has a rate of convergence that again depending on n / (n + α) but is faster than the two previous schemes, with rates of convergence of 0.30, 0.60, 0.80 for (n / (n + α)) = 0.2, 0.5 and 0.8. In one sense the extra parameters help to reduce the 'missingness' that slows convergence.

I have found the following argument in trying to understand some of these improved EM schemes. Consider the case when we have N moment matrices M_i (i = 1,…,N) with expectation $G+R_i$. This might arise in considering a p multivariate problem with 'equal designs' with 2 p x p multivariate components and we use a spectral decomposition to construct N independent sets of sums of squares and cross products. We consider the case when R_i is known and we are interested in estimating G. We let $G = SUS^{'}$ that allows a wide range of possible models. If U =

I and S lower triangular we have a Choleski parameterization. The matrix S could be thought representing a set of factors and if S is of size p x f we have V factors, and so S represents a reduced rank or latent regression parameterization. As G is a symmetric matrix there are $px(p+1) / 2$ parameters. Obviously care needs to be taken with S and V as these have $px(2p^2+p+1)/2$ parameters. An estimation procedure based on differentiating the likelihood can be informally thought of as thinking of M_I as $y_i y_I'$ with $E(y_i) = u_i = Sf_i$ estimation of the terms of S can be thought of as predicting f_i from y_i and regressing y_i on the prediction of f_i taking into account the uncertainty in f_i. Estimation of U follows the recipe involving the prediction of f_i and the prediction error variance of f_i. Note that formally y_i does not need forming as all required terms can be constructed from M_i. A similar algorithm can be constructed from PX-EM arguments (B.R. Cullis and A. Smith, *pers. comm*.). I have found this argument useful in (a) understanding the PX-BM methodology, (b) estimation in reduced rank or latent factor models, (c) as a way of constructing hybrid iterative schemes.

I 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 ,but this difficulty reduces as the parameters get nearer to a maximum value of the likelihood and can lead to estimates outside the parameter space. One suggestion is motivated by Lee and Nelder (2001) who base estimation of variance parameters in hierarchical models of pseudo data based on sums of squares of predicted values and their prediction error variances. This suggests when AI algorithms are having problems using (PX)-EM schemes based on constructing pseudo moment matrices and expectations from relevant predictions and prediction error matrices for difficult parameters. A maximization of this likelihood of this pseudo-data could be perhaps used. This updating should get parameters nearer the maximum computationally faster than updating all parameters after each (PX)-EM iterate.

We have concentrated on exact methods of analysis because Professor Gianola will discuss Bayesian and Markov Chain Monte Carlo (MCMC) methods. 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. I tend to think of Gibbs sampling methods as adding noise at every step of a simplified exact analysis. For instance estimate *b* and add noise, estimate *u* and add noise, form sums of squares for *u* and add noise to give an estimate of σ_G^2. One does not need to Bayesian to use MCMC methods and Guo and Thompson (1994) use the above paradigm with the estimation of σ^2_G given by an EM step. In a sense the difficulties of calculating prediction error variances is replaced by sampling them. Thompson (1994) and Garcia-Cortes and Sorensen (2001) have pointed out that the sampling error can be reduced when updating σ^2_G taking account of the variance of the noise added to u 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 and Rasbash (1999) for imputation can also reduce computational effort. In our model, their idea suggests fitting two models

$$y - Z\tilde{u} = Xb + e \qquad (1) \qquad \text{and} \qquad y - X\tilde{b} = Zu + e \qquad (2)$$

In (1) we fit $\hat{b}$ and construct $\tilde{b}$ as $\hat{b}$ plus noise. In (2) we adjust *y* for $\tilde{b}$, estimate $\sigma^2_{G \text{ and }} \sigma^2_E$ and fit $\hat{u}$ and add noise to $\tilde{u}$. Then *y* is adjusted for $Z\tilde{u}$ and the procedure repeated. After burn in averages of σ^2_G and σ^2_E provide estimates of σ^2_G and σ^2_E in the spirit of Gibbs sampling but avoiding some of the noise in $\tilde{u}$ when σ^2_G and σ^2_E are estimated.

The L'y selection idea of Henderson (1975) is still not been completely understood. There has been related work on survival analysis but until recently it has concentrated on univariate sire models. Only recently have there been attempts to integrate the non-linear survival analyses with other related traits in animal models (Ducrocq, 1990 ; 2001). There has been more interest in this area in medical studies under considerations of random, non-random and informative dropout (for example, Diggle and Kenward, 1994)

CONCLUSIONS

We have shown that the area of genetic parameter estimation has advanced tremendously over the last thirty years allowing more appropriate models to be fitted to larger data sets. There are still challenging problems to be solved that we think will build on existing knowledge.

REFERENCES

Cantet, R. and Birchmeier, A.N. (1998) *Proc. 6th WCGALP* **25** : 529-532.
Cantet, R.J.C., Birchmeier, A.N. ,Santos-Cristal, M.G. and de Avila, V.S. (2000) *J.Anim. Sci*. **78** : 2554-2560.
Clayton, D. and Rasbash, J. (1999) *J. R. Stat. Soc. A*. **162** : 425–436.
Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) *J. Royal Stat. Soc., B* **39 :** 1-38.
Diggle, P. and Kenward, M.G. (1994) *Applied Statistics* **43** : 49-93.
Duagjinda, M., Bertrand, J.K. and Misztal, I. (2001) *J. Anim. Sci*. **79** : 2997-3001.
Duagjinda, M., Misztal, I., Bertrand, J.K. and Tsurata, S. (2001) *J. Anim. Sci*. **79** : 2991-2996.
Ducrocq, V. (1990) *Proc. 4th WCGALP* **13** : 419-448.
Ducrocq, V. (2001) *Interbull Bulletin* **27** : 147-152.
Ducrocq, V. and Besbes, B. (1993) *J. Anim. Breed. Genet*. **110** : 81-92.
Ducrocq, V. and Chapuis, H. (1997) *Genet. Sel. Evol*. **29** : 205-224.
Foulley, J.L. and Quaas, R.L. (1995) *Genet. Sel. Evol*. **27** : 211-225.
Garcia-Cortes, L.A. and Sorensen, D. (2001) *Genet. Sel .Evol*. **33** : 443-455.
Gilmour, A.R., Thompson, R. and Cullis, B.R. (1995) *Biometrics* **51** : 1440-1450.
Groeneveld, E. (1994) *Genet. Sel. Evol*. **26** : 537-545.
Guo, S.W. and Thompson, E.A. (1994) *Biometrics* **50** : 7-432.
Henderson, C.R, (1975) *Biometrics* **31** : 3-447.
Henderson, C.R. (1973) In "Proc. Animal Breeding and Genetics Symposium in Honour of Dr. Jay L. Lush" p. 10-41, Champaign, Illinois.
Henderson, C.R. (1976) *Biometrics* **32** : 69-83.
Hofer, A. (1998) *J. Anim. Breed. Genet*. **115** : 247-265.
Jensen, J., Mantysaari, E., Madsen, P. and Thompson, R. (1997) *J. Indian Soc. Agric. Sci*. **49** : 215-236.

Johnson, D.L. and Thompson, R. (1995) *J. Dairy Sci*. **78** : 449-456.
Lee, Y. and Nelder, J.A. (2001) *Biometrika* **88** : 987-1006.
Lindstrom, M.J. and Bates, D.M. (1988) *J. Am. Stat. Assoc*. **83** : 1014-1022.
Lui, C., Rubin, D.B. and Wu, Y.N. (1997) *Biometrika* **85** : 755-770.
Madsen, P., Jensen, J. and Thompson, R. (1994) *Proc. 5th WCGALP* **22** : 19-22.
Mantysaari, E and Van Vleck, L.D. (1989) *J. Anim. Breed. Genet*. **106** : 409-422.
Meng, X.L. and Van Dyk, D.A. (1998) *J.R. Stat. Soc. B*. **60** : 559-578.
Meyer, K. (1985) *Biometrics* **41** : 153-165.
Meyer, K. (1989) *Genet. Sel. Evol*. **21** : 317-340.
Meyer, K. (1997) *Genet. Sel. Evol*. **29** : 97-116.
Meyer, K. and Smith, S.P. (1996) *Genet. Sel. Evol*. **28** : 23-41.
Misztal, I. (1994) *J. Anim. Breed. Genet*. **111** : 346-355.
Misztal, I. (1997) *Journal of Dairy Science* **80** : 965-974.
Misztal, I. and Perez-Enciso, M. (1993) *J. Dairy Sci*. **76** : 1479-1483.
Neumaier, A. and Groeneveld, E. (1998) *Genet. Sel. Evol*. **30** : 3-26.
Patterson, H.D. and Thompson, R. (1971) *Biometrika* **58** : 545-554.
Reverter, A., Golden, B.L., Bourdon, R.M. and Brinks, J.S. (1994a) *J. Anim. Sci*. **72** : 34-37.
Reverter, A., Golden, B.L., Bourdon, R.M. and Brinks, J.S. (1994b) *J. Anim. Sci*. **72** : 2247-2253.
Schenkel, F.S. and Schaeffer, L.R. (2000) *J. Anim. Breed. Genet*. **117** : 225-239.
Smith, S.P. (1995) *J. Comp. Graph Stat*. **4** : 134-147.
Smith, S.P. and Graser, H.-U. (1986) *J. Dairy Sci*. **69** : 1156-1165.
Takahashi, K., Fagan, J. and Chin, M.S. (1973) *Proc. 8th Inst. PICA Conf.Minneapolis* : 63.
Thompson, R. (1994) *Proc. 5th WCGALP* **18** : 337-340.
Thompson, R., Wray, N.R. and Crump, R.E. (1994) *J. Anim. Breed. Genet*. **111** : 102-109.
Wray, N.R. (1990) *Biometrics* **46** : 197-186.