Proc. Assoc. Advmt. Anim. Breed. Genet. 20:523-526

SAMPLING BASED APPROXIMATION OF CONFIDENCE INTERVALS FOR FUNCTIONS OF GENETIC COVARIANCE MATRICES

Karin Meyer[1] **and David Houle**[2]

[1]Animal Genetics and Breeding Unit*, University of New England, Armidale, NSW 2351
[2]Department of Biological Science, Florida State University, Tallahassee, FL 32306-4295

SUMMARY

Approximate lower bound sampling errors of maximum likelihood estimates of covariance components and their linear functions can be obtained from the inverse of the information matrix. For non-linear functions, sampling variances are commonly determined as the variance of their first order Taylor series expansions. This is used to obtain sampling errors for estimates of heritabilities and correlations, and these quantities can be computed with most software performing such analyses. In other instances, however, more complicated functions are of interest or the linear approximation is difficult or inadequate. A pragmatic alternative then is to evaluate sampling characteristics by repeated sampling of parameters from their asymptotic, multivariate normal distribution, calculating the function(s) of interest for each sample and inspecting the distribution across replicates. This paper demonstrates the use of this approach and examines the quality of approximation obtained.

INTRODUCTION

Maximum likelihood (ML) theory indicates that ML estimates asymptotically have a multivariate normal (MVN) distribution with covariance matrix given by the inverse of the information matrix, i.e. the inverse of the matrix of second, partial derivatives of the likelihood function. Hence lower bound sampling errors of ML estimates are usually obtained from the diagonal elements of this matrix, and pertaining confidence limits are determined multiplying these values with the appropriate intercepts of a standard normal distribution. Corresponding statistics for linear functions of the parameters estimated are readily derived. For a non-linear function, the standard procedure is to replace the function with its first order Taylor series expansion and to calculate the variance of this linear approximation, a strategy sometimes referred to as the Delta method (e.g. Oehlert 1992). In genetic parameter estimation, this is used to approximate sampling errors of variance ratios and correlations, and is implemented in most restricted maximum likelihood (REML) software available.

In some cases, however, more complicated functions and their sampling distribution are of interest, which may not be approximated closely enough by a linear expansion. Others may involve variables afflicted by constraints on the parameter space or may simply not be accommodated by the facilities to calculate approximate variances of 'user-defined' functions of covariances available in software such as ASReml (Gilmour *et al.* 2009) and WOMBAT (Meyer 2007). A simple alternative then is to evaluate asymptotic sampling characteristics for such functions by repeated sampling of parameter estimates from their asymptotic, MVN distribution, calculating the function(s) of interest for each sample and inspecting their distribution(s) across replicates. This paper describes a suitable sampling strategy and examines the quality of approximation of sampling distributions obtained.

SAMPLING STRATEGY

Newton-Raphson type algorithms to maximise the REML (log) likelihood ($\log \mathcal{L}$) function utilize second derivatives of $\log \mathcal{L}$ and are well established as the most efficient methods available, especially the so-called average information variant (Gilmour *et al.* 1995). However, these involve an unconstrained optimization. Hence, estimation of covariance components is generally performed

*AGBU is a joint venture of NSW Department of Department of Primary Industries and the University of New England employing a re-parameterisation to functions which do not require constraints to ensure positive (semi-) definite estimates of covariance matrices. A common choice for covariance matrices is to estimate the elements of their Cholesky factors, transforming diagonal elements to logarithmic scale (Meyer and Smith 1996). Furthermore, performing the factorization with pivoting on the largest diagonal readily facilitates reduced rank analyses (Meyer and Kirkpatrick 2005).

In addition, such parameterisation directly allows sampling of estimates of covariance matrices which are guaranteed to be in the parameter space, mimicking the constraints imposed in REML estimation. Let $\hat{\boldsymbol{\theta}}$, of length p, denote the vector of parameter estimates and $\mathbf{H} = \text{Var}(\hat{\boldsymbol{\theta}})$ the corresponding inverse of the information matrix at convergence. Samples of parameters from $N(\hat{\boldsymbol{\theta}}, \mathbf{H})$ are obtained as $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \mathbf{L}_H\mathbf{d}$ with $\mathbf{L}_H$ the Cholesky factor of $\mathbf{H}$ and $\mathbf{d}$ a vector of standard normal deviates, $d_i \sim N(0,1)$. Samples of covariance matrices can then be constructed from $\tilde{\boldsymbol{\theta}}$ by reversing the transformation.

APPLICATION

Data for 6 traits recorded on 4000 individuals in 500 independent families of size 8 were simulated for the design of Bondari *et al.* (1978). Population parameters assumed all residual correlations were equal to 0.3. Heritabilities were 0.2, 0.3 and 0.4 for two traits each, and all phenotypic variances were equal to 100. For Case I, all genetic correlations were assumed to be equal to 0.5, while for case II values for traits i and j were $0.7^{|i-j|}$.

REML estimates of genetic and residual covariances matrices were obtained fitting an animal model, using an average information algorithm. Three estimates of sampling (co)variances for covariance components and functions thereof were contrasted:

A) Values from the REML analysis, obtained from $\mathbf{H}$ using the Delta method. Let σ_{ij} denote the elements of a covariance matrix $\boldsymbol{\Sigma} = \mathbf{LL}'$, with $\mathbf{L} = \{l_{ij}\}$ its Cholesky factor. For $\text{Cov}(\hat{l}_{ij}, \hat{l}_{km})$ given by the corresponding element of $\mathbf{H}$, $\text{Cov}(\hat{\sigma}_{ij}, \hat{\sigma}_{kl})$ is approximated as

$$\sum_{t=1}^{f(i,j)} \sum_{s=1}^{f(k,m)} \left[\hat{l}_{jt}\hat{l}_{ms}\,\text{Cov}\left(\hat{l}_{it}, \hat{l}_{ks}\right) + \hat{l}_{jt}\hat{l}_{ks}\,\text{Cov}\left(\hat{l}_{it}, \hat{l}_{ms}\right) + \hat{l}_{it}\hat{l}_{ms}\,\text{Cov}\left(\hat{l}_{jt}, \hat{l}_{ks}\right) + \hat{l}_{it}\hat{l}_{ks}\,\text{Cov}\left(\hat{l}_{jt}, \hat{l}_{ms}\right) \right]$$

with $f(i,j) = \min(i,j,r)$, and r the rank at which $\boldsymbol{\Sigma}$ is estimated. Similar formulations apply when diagonal elements l_{ii} are transformed to logarithmic scale or for covariances among components belonging to matrices $\boldsymbol{\Sigma}$ pertaining to different sources of variation.

B) Empirical values obtained by repeatedly sampling data for the given structure from appropriate normal distributions with population values equal to the estimates of covariances, and carrying out a REML analysis for each sample. A total of 10,000 analyses were performed, and sampling variances determined as the variances across replicates.

C) Approximate values obtained as covariances across 200,000 samples drawn from a MVN distribution as described above.

For both empirical and MVN samples, 95% confidence intervals were obtained after sorting samples in numerical order as the mid-points between the 2.5% top and bottom samples and the remainder. REML estimation and sampling from the MVN distribution were carried out using WOMBAT.

Results. Estimates of sampling covariances among the distinct elements of the genetic covariance matrix ($\hat{\boldsymbol{\Sigma}}_G$) for case I are contrasted in Figure 1, showing excellent agreement between all three values [■ depicting variances $\text{Var}(\hat{\sigma}_{G\,ij})$ and • covariances $\text{Cov}(\hat{\sigma}_{G\,ij}, \hat{\sigma}_{G\,kl})$]. For case II, the estimate of the smallest genetic eigenvalue was not significantly different from zero, i.e. a full rank estimate of $\boldsymbol{\Sigma}_G$ represented an over-parameterised model. As illustrated in Figure 2, this resulted in an overestimate of $\text{Var}(\hat{\sigma}_{G\,ij})$ obtained from the MVN approximation. The component affected pertained to the variance of the trait considered last in the Cholesky decomposition of $\boldsymbol{\Sigma}_G$, i.e. the overestimate reflected

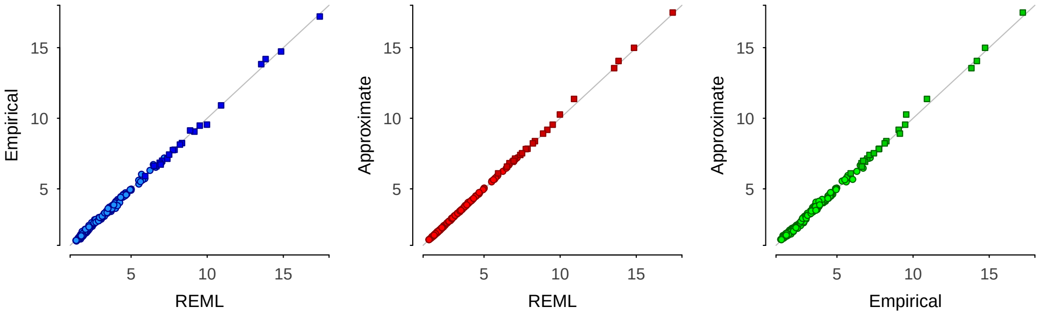

Figure 1. REML, empirical and approximate sampling (co)variances for case I

accumulation of errors for a redundant parameter. Reducing the number of parameters by estimating $\mathbf{\Sigma}_G$ at reduced rank again yielded very good agreement between empirical and approximated values.

Empirical and approximated sampling distributions for selected functions of covariances are compared in Figure 3, with left and right solid vertical bars marking the 95% confidence limits obtained as truncation points between the top and bottom 2.5% of samples and dashed bars showing their 'standard' counterparts, 1.96 standard deviations either side of the mean. Again, there was close agreement between empirical results obtained by re-sampling data and the MVN approximation. For functions at the boundary of the parameter space, such as the genetic correlation between traits 1 and 2, sampling distributions tend to be skewed and confidence intervals derived directly from the distribution tend to be more appropriate than those calculated from sampling errors and normal intercepts. Estimates of genetic eigenvalues are generally reported without any measure of their precision. Similarly, canonical eigenvalues and the number of effective dimensions, $\sum_i \lambda_i/\lambda_1$ (with λ_i the eigenvalues of the matrix of (co)heritabilities and λ_1 the largest value; Kirkpatrick (2009)) are functions of both genetic and phenotypic covariance matrices, and calculation of sampling variances using the Delta method would be, at the least, tedious while it is straightforward using MVN sampling.

DISCUSSION

By definition, REML estimation of covariance components involves the solution of a constrained optimisation problem. Fortunately, this task can be made easier by a transformation to parameters which do not require constraints to yield valid estimates of covariance matrices. Sampling from the asymptotic distribution of these parameters has been shown to yield numerical estimates of sampling covariances, distributions and confidence intervals in close agreement with those obtained by resampling data. It has to be emphasized though that for this to hold, large sample properties need to apply, i.e. the inverse of the information matrix has to provide an adequate description of

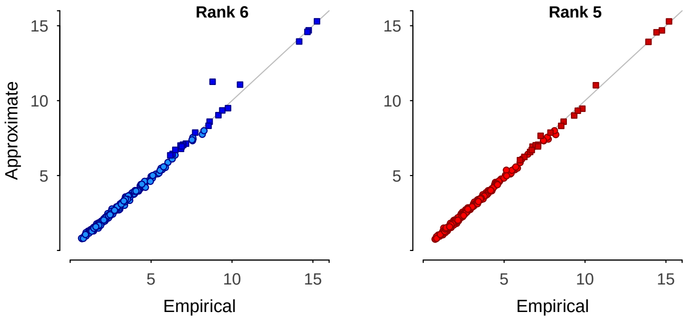

Figure 2. Approximate vs. empirical sampling (co)variances for case II

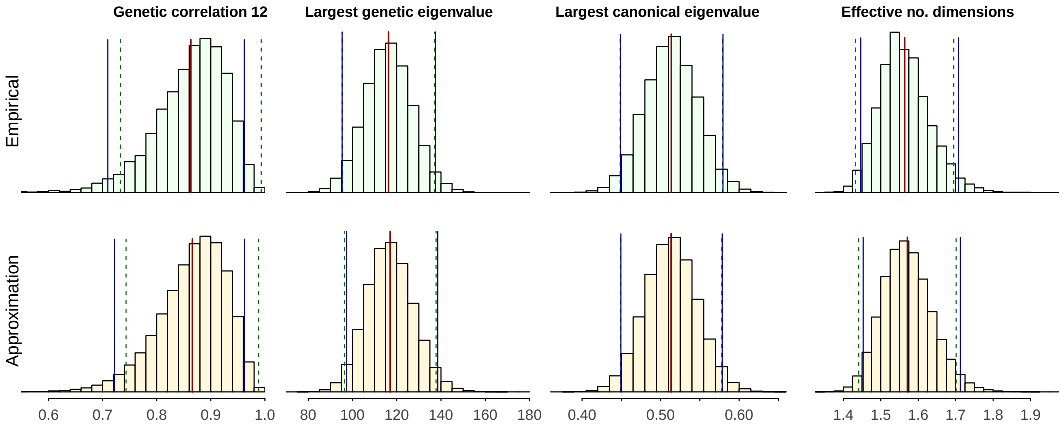

Figure 3. Sampling distributions and confidence intervals for selected functions (Case II)

sampling covariances among the parameters estimated. If this is not the case, estimates of confidence limits derived from the profile likelihood for individual parameters may be more preferable though computationally considerably more demanding (Meyer 2008). In addition, the sampling procedure was found to be sensitive to an overparameterised model, yielding overestimates of sampling variances for redundant parameters, and care needs to be taken for multivariate analyses of more than a few traits to estimate covariance matrices at the appropriate rank.

To facilitate use of the approach described, an option to invoke sampling of parameters from their asymptotic distribution together with the transformation to estimates of covariance matrices has been implemented in our REML package WOMBAT (Meyer 2007) as a post-estimation step. This yields a file with samples of covariance matrices suitable for input to a package such as R (R Core Team 2012) to evaluate the functions of interest and compute summary statistics.

CONCLUSIONS

Sampling of REML estimates from their asymptotic MVN distribution, specified by the inverse of the information matrix, offers a straightforward and computationally undemanding way to derive sampling distributions and confidence intervals for estimates of covariance components and 'non-standard' functions thereof numerically. It is a small but useful addition to our toolkit for estimation.

REFERENCES

Bondari K., Willham R.L. and Freeman A.E. (1978) *J. Anim. Sci.* **47**:358.

Gilmour A.R., Thompson R. and Cullis B.R. (1995) *Biometrics* **51**:1440.

Gilmour A.R., Gogel B.J., Cullis B.R. and Thompson R. (2009) *ASReml User Guide Release 3.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Kirkpatrick M. (2009) *Genetica* **136**:271.

Meyer K. (2007) *J. Zhejiang Uni. SCIENCE B* **8**:815.

Meyer K. (2008) *Heredity* **101**:212.

Meyer K. and Kirkpatrick M. (2005) *Genet. Sel. Evol.* **37**:1.

Meyer K. and Smith S.P. (1996) *Genet. Sel. Evol.* **28**:23.

Oehlert G.W. (1992) *Amer. Stat.* **46**:27.

R Core Team (2012) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

This work was funded by Meat and Livestock Australia under grant B.BFG.0050. Support from the National Institute for Mathematical and Biological Synthesis, University of Tennessee, for working group meetings stimulated the collaboration.