Estimation of Parameters in Random Effect Models with Incidence Matrix Uncertainty

Xia Shen[1,2] * **and Lars Rönnegård**[2,3]

1 The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden;

2 School of Technology and Business Studies, Dalarna University, Borlänge, Sweden;

3 Department of Animal Breeding & Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.

Abstract

Random effect models have been widely applied in many fields of research. However, models with uncertain design matrices for random effects have been little investigated before. In some applications with such problems, an expectation method has been used for simplicity. This method does not include the extra information of uncertainty in the design matrix is not included. The closed solution for this problem is generally difficult to attain. We therefore propose an two-step algorithm for estimating the parameters, especially the variance components in the model. The implementation is based on Monte Carlo approximation and a Newton-Raphson-based EM algorithm. As an example, a

*Corresponding author. Email: xiashen@me.com

simulated genetics dataset was analyzed. The results showed that the proportion of the total variance explained by the random effects was accurately estimated, which was highly underestimated by the expectation method. By introducing heuristic search and optimization methods, the algorithm can possibly be developed to infer the 'model-based' best design matrix and the corresponding best estimates.

Random effect models have been used by researchers in various fields, e.g. genetics, quality control, and medicine. Generally, explanatory variables are regarded as fixed or random effects, and through design matrices, formulate a *linear predictor* that explains the response variable. Since dispersion parameters can be estimated for each random element in the model, such models are usually referred to as *variance component* (VC) models. Models involving normal-distributed random effects are linear mixed models when the response variable comes from a normal distribution.

Many variables with unobservable information, e.g. genetic markers (small parts of DNA), actually have uncertainty in their design matrices, which cannot be taken into account by simply specifying random effects (GOLDGAR 1990; SCHORK 1993; XU and ATCHLEY 1995; XU 1996; RÖNNEGÅRD, BESNIER and CARLBORG 2008). This means that the random effect design matrix itself is a random variable that has a particular distribution. It is a challenging problem how to estimate parameters and to do inference for such models which we refer to as incidence-matrix-uncertain random effect models (IMURM).

The aim of this paper is to develop a general method for estimating pa-

rameters in IMURM, especially variance components for the random effects with an uncertain design matrix. The rest of the paper is arranged as three parts: We describe the statistical model and the theoretical background in the method section, including a proposed algorithm; A simulated genetics dataset is analyzed using our method as an example; The paper is closed by discussing the relative numerical problems and possible development in the future.

METHODS

Models and Likelihoods:

We consider the normal linear mixed model

$$\mathbf{y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is the vector of dependent variable, $\boldsymbol{\beta}$ is the fixed effect vector, $\boldsymbol{\gamma}$ is the multivariate normal-distributed random effect vector with a zero mean and variance-covariance matrix $Var(\boldsymbol{\gamma}) = \sigma_g^2 \mathbf{I}_q$, and $\boldsymbol{\epsilon}$ the normal-distributed error term with a zero mean and variance $Var(\boldsymbol{\epsilon}) = \sigma_e^2 \mathbf{I}_N$. $\mathbf{X}$ $(N \times p)$ and $\mathbf{Z}$ $(N \times q)$ are the design matrices, where N is the number of observations.

The variance-covariance matrix of $\mathbf{y}$ is given by $\mathbf{V} = \sigma_g^2 \mathbf{A} + \sigma_e^2 \mathbf{I}_N$. The variance components are $\boldsymbol{\theta} = (\sigma_g^2, \sigma_e^2)'$. We have

$$\mathbf{A} \quad = \quad \mathbf{Z}\mathbf{Z}' \tag{2}$$

$\mathbf{A}$ is a semi-positive definite matrix that determines the correlation of the levels of the random effect. When $\mathbf{A}$ is known for a random effect model (1), the design matrix $\mathbf{Z}$ can be Cholesky-decomposed from $\mathbf{A}$. Fitting such models with correlated random effects is straightforward in R (R DEVELOPMENT

CORE TEAM 2009) using for instance, the **hglm** package (RÖNNEGÅRD, SHEN and ALAM 2010).

For known $\mathbf{A}$, using the *restricted maximum likelihood* (REML) adjustment, the likelihood function of $\boldsymbol{\theta}$ *profiling* out the fixed effects is given by (e.g. PAWITAN 2001)

$$\begin{aligned} p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{A},\mathbf{y}) &= f(\mathbf{y}|\boldsymbol{\theta},\mathbf{A}) \\ &= |2\pi\mathbf{V}|^{-1/2} \\ &\quad \cdot \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{A}})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{A}})\right) \\ &\quad \cdot \left|\frac{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}}{2\pi}\right|^{-1/2} \end{aligned} \tag{3}$$

In some applications, the design matrix $\mathbf{Z}$ of the random effect $\boldsymbol{\gamma}$ is uncertain. For instance, for the purpose of ranking animals, breeders estimate breeding values as BLUP, i.e. *best linear unbiased predictor* (ROBINSON 1991), which are the random effect estimates from the REML estimation procedure. When estimating breeding values, a linear mixed model is built for linking phenotypic values (observed trait values) to genetic information. The genetic information comes in through the variance-covariance matrix $\mathbf{A}$ that describes the kinship correlation between each pair of animals. Because the real gene flow is not observable, uncertainty exists in $\mathbf{A}$.

Given the incomplete known information, there exists a probability space $(\Omega,\mathscr{F},\mathscr{P})$ where Ω denotes the sample space of all the possible $\mathbf{A}$ matrices, $\mathscr{F}$ is the σ-algebra of subsets of Ω, and the probability measure $\mathscr{P}$ on $(\Omega,\mathscr{F})$ satisfies $\mathscr{P}(\Omega)=1$. If the sample size equals s, any element $\mathbf{A}_i \in \Omega$, $i=1,\ldots,s$, is a possible variance-covariance matrix, and $E[\mathbf{A}]=\sum_{i=1}^{s}\mathbf{A}_i\mathscr{P}(\mathbf{A}_i)$ is the expectation of $\mathbf{A}$. Geneticists often use an average matrix of $\mathbf{A}$, i.e. the expectation

of $\mathbf{A}$ conditioning on the observable kinship information. This method is referred to as the expectation method.

Taking the uncertainty in $\mathbf{A}$ into account, $\mathbf{A}$ can be regarded as a random variable or a parameter in the model. A joint likelihood of $\boldsymbol{\theta}$ and $\mathbf{A}$ is considered, and inference of $\boldsymbol{\theta}$ should be made from the marginal likelihood of $\boldsymbol{\theta}$ integrating out $\mathbf{A}$. Hence, based on profile likelihood (3), the marginal likelihood of $\boldsymbol{\theta}$ is

$$\begin{aligned} m_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y}) &= \sum_{\mathbf{A}} p_{\boldsymbol{\beta}}(\boldsymbol{\theta},\mathbf{A}|\mathbf{y}) \\ &= \sum_{\mathbf{A}} p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A})\mathscr{P}(\mathbf{A}) \\ &= E_{\mathbf{A}}[f(\mathbf{y}|\boldsymbol{\theta},\mathbf{A})] \end{aligned} \tag{4}$$

Maximizing this marginal likelihood gives the estimate of $\boldsymbol{\theta}$ from IMURM.

Estimating Algorithm:

Since the distribution of $\mathbf{A}$ is rather complicated, marginal likelihood (4), involving an expectation with respect to $\mathbf{A}$, is hardly derivable unless the number of observations is extremely small. Therefore, we propose a Monte Carlo (MC) strategy that approximates the marginal likelihood $m_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y})$, which is

$$\tilde{m}_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y}) \approx \frac{1}{m}\sum_{i=1}^{m} p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i) \tag{5}$$

where m is the number of imputes drawn based on the known information. Each impute $\mathbf{A}_i$ corresponds to an incidence matrix $\mathbf{Z}_i$, and equation (2) holds for them, so that $\mathbf{Z}_i\mathbf{Z}_i' = \mathbf{A}_i$. $\tilde{m}_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y})$ converges to $m_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y})$ as $m \to \infty$.

$\boldsymbol{\theta}$ is identical for all the imputes of $p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i)$, which means that instead of maximizing each likelihood impute, the entire sum $\sum_{i=1}^{m} p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i)$ needs

to be maximized. The first and second derivatives of $\log p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i)$ with respect to $\boldsymbol{\theta}$ have closed solutions (HARVILLE 1977). Let $\ell = \log \tilde{m}_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y})$ and $\ell_i = \log p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i)$. ℓ is the target log-likelihood to maximize. Using the derivatives $\partial \ell_i/\partial \boldsymbol{\theta}$ and $\partial^2 \ell_i/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'$, a Newton-Raphson-based EM algorithm can be used to estimate $\boldsymbol{\theta}$. In the following steps, m is the number of MC imputes, and k is the iteration index in the EM algorithm.

i) Find an initial estimate $\hat{\boldsymbol{\theta}}_0$.

ii) Loop on k until convergence.

$$\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\theta}}_{k-1} - \delta \left(\frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \right)^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{k-1}} \left(\frac{\partial \ell_i}{\partial \boldsymbol{\theta}} \right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{k-1}} \tag{6}$$

where δ is a step size constant and

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{m} w_i \frac{\partial \ell_i}{\partial \boldsymbol{\theta}} \tag{7}$$

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \sum_{i=1}^{m} w_i \left(\left(\frac{\partial \ell_i}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ell_i}{\partial \boldsymbol{\theta}} \right)' + \frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \right) - \left(\frac{\partial \ell}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ell}{\partial \boldsymbol{\theta}} \right)' \tag{8}$$

In equations (7) and (8), the weights are defined as

$$w_i = p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i) / \sum_{i=1}^{m} p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i) \tag{9}$$

iii) Take the converged estimate $\hat{\boldsymbol{\theta}}$ as the variance component estimate of IMURM.

Simulated Example:

In order to test our algorithm, a small animal dataset was simulated. In a three-generation pedigree (Figure 1), the pair of grandparents (animals 1 and

2) were mated to give birth to one male (animal 3) and two female progeny (animals 4 and 5). Animal 3 were mated with animal 4 and produced animals 6, 8, 9 and 10. Animal 3 were also mated with animal 5 and gave birth to animal 7. Genotypes were simulated for each animal, and there are 4 alleles (A, B, C and D) throughout the pedigree, which have genetic effects of $\gamma_A = 3$, $\gamma_B = 6$, $\gamma_C = 9$ and $\gamma_D = 12$, respectively. The phenotypic value for animal i was simulated by

$$y_i = \mu + \gamma_{i1} + \gamma_{i2} + \epsilon_i \tag{10}$$

where $\mu = 50$, $\epsilon_i \sim N(0,1)$, and γ_{i1} and γ_{i2} correspond to the two allele effects of animal i inherited from the father and the mother respectively. The simulated phenotypic values were 64.87, 56.21, 61.28, 69.20 and 62.08 for animals 6-10, respectively. Note that in Figure 1, the kinship information is known but the genotypes are unobservable.

The *(narrow sense) heritability* (LYNCH and WALSH 1997) or *intra-class correlation* of the studied trait is defined by

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \tag{11}$$

which measures how large proportion of the trait is determined by inheritance. The five animals in the last generation are used to estimate h^2. The conventional expectation method estimates h^2 using model (1), where the variance-covariance matrix $\mathbf{A}$ of the genetic random effect is the kinship ma-

trix

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1.250 & 0.625 & 0.750 & 0.750 & 0.750 \\ 0.625 & 1.250 & 0.625 & 0.625 & 0.625 \\ 0.750 & 0.625 & 1.250 & 0.750 & 0.750 \\ 0.750 & 0.625 & 0.750 & 1.250 & 0.750 \\ 0.750 & 0.625 & 0.750 & 0.750 & 1.250 \end{pmatrix} \tag{12}$$

which is an expected relationship matrix with uncertain genotypes. Taking into account such uncertainty, we estimate h^2 by estimating the variance components using IMURM and compare with the expectation method. Assuming known allele effects and residual variance, the simulated value of h^2 is $s^2_{\boldsymbol{\gamma}}/(s^2_{\boldsymbol{\gamma}} + \sigma^2_e) = .9375$, where $\boldsymbol{\gamma} = (\gamma_A, \gamma_B, \gamma_C, \gamma_D)'$ and s represents the standard deviation.

RESULTS

Convergence rate of the IMURM algorithm depends on the problem size (number of observations), complexity of the random effect (rank of the design matrix) and the number of MC imputes used in the computation. Generally, more MC imputes are required when the problem size and the random effect complexity increases. For the simulated example, the genetic variance estimate converged after about 50 MC imputes, which means $m = 50$ is sufficient for this small dataset (Figure 2 (a)). Although for one run of the algorithm, convergence becomes slower if m increases, the algorithm does not require many iterations since it uses Newton-Raphson optimization. With 10000 MC imputes of $\mathbf{Z}_i$, the algorithm converged after around 5 iterations (Figure 2 (b)).

Using the generated 10000 MC imputes, we obtained an empirical distribution of the possible genotypes, namely, the $\mathbf{Z}$ matrices (Figure 2 (c)). The true $\mathbf{Z}$ matrix, constructed according to the true genotypes, was an extreme value of the distribution, which led to a variance component estimate of the random effect falling on the tail of the variance component distribution (Figure 2 (d)) and a maximum extreme value of the profile likelihood (Figure 2 (e)).

Our target was to estimate the heritability h^2 by estimating the variance components σ_g^2 and σ_e^2. If the true $\mathbf{Z}$ matrix was used, it turned out with an estimate of .9785 and the bias of .0410 (Table 1). The estimate by IMURM converged to .9087 using all the 10000 imputes, and the bias was -.0288. The expectation method using variance-covariance matrix 12 gave an estimate of .4485 and the bias of -.4890. The IMURM method was therefore more reliable on estimating the proportion of total variance explained by the random effect than the expectation method.

DISCUSSION

We have employed three terminologies: a Monte Carlo method, a Newton-Raphson algorithm, and an EM algorithm. Note that the entire estimation method for IMURM is a two-step method. In the first step, the Monte Carlo method is used to construct an empirical distribution of $\mathbf{A}$, which enables approximating the marginal likelihood of $\boldsymbol{\theta}$. In the second step, the EM algorithm is used to estimate $\boldsymbol{\theta}$, which is based on a Newton-Raphson framework, that is, the weights are updated as the 'E' step, and the estimates are updated using Newton-Raphson iterations as the 'M' step.

It is a reasonable question how difficult the IMURM estimation method can be extended for more flexible models. One extension of normal linear mixed models is the *generalized linear mixed models* (GLMM) (BRESLOW and CLAYTON 1993), where the mean of the non-normal response variable is explained by the linear predictor through a *link function*. The distribution of random effects can also be relaxed to non-normal, which was presented by LEE and NELDER (1996) who called such models the *hierarchical generalized linear models* (HGLM). HGLM can be estimated based on the *hierarchical likelihood* (h-likelihood), and modeling dispersion parameters, i.e. double HGLM (LEE and NELDER 2006), is also straightforward via the h-likelihood algorithm (LEE, NELDER and PAWITAN 2006). For HGLMs, the IMURM method can be easily extended when the family of response and the family of random effects are a conjugate pair, that is, closed solutions for the necessary derivatives can be attained. For other random effects models such as GLMMs, derivatives of the profile likelihood function can be used, however, bias should be corrected using REML adjustment (LEE, NELDER and PAWITAN 2006).

Numerical Difficulties Implementation of equations (7) and (8) requires consideration of numerical aspects. One important part is to calculate the weights accurately. Since w_i is defined as a fraction where the terms involved in the numerator and denominator are all profile likelihood values, each term $p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i)$ is an extremely small probability especially when $\boldsymbol{\theta}$ is multidimensional. Such small probabilities are zeros on a computer, so that the computation loses numerical accuracy and in all likelihood, generates zero-weights or even an error. Fortunately, log-likelihood values do not suffer from such a

numerical problem, simple transformation might be adopted on the log scale and maintains the equality. We propose an easy solution which calculates the weights as

$$\begin{aligned} w_i &= p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i)/\sum_{i=1}^{m} p_{\boldsymbol{\beta}}(\boldsymbol{\theta}|\mathbf{y},\mathbf{A}_i) \\ &= \frac{\exp \ell_i}{\sum_{i=1}^{m} \exp \ell_i} \\ &= \frac{\exp(\ell_i - \overline{\ell})}{\sum_{i=1}^{m} \exp(\ell_i - \overline{\ell})} \end{aligned} \tag{13}$$

where $\overline{\ell} = \frac{1}{m}\sum_{i=1}^{m} \ell_i$. After centering towards the mean, the log-likelihood values are close to zero, and the corresponding profile likelihood values can be accurately computed.

Future Development It was found that in the sample space of $\mathbf{Z}$, the 'model-based' best design matrix was an extreme value that led to the maximum of the profile likelihood values. This implies, after profiling out the fixed effects, like equation (3), that the 'model-based' best design matrix and its corresponding $\hat{\boldsymbol{\theta}}$ maximize the joint likelihood $p_{\boldsymbol{\beta}}(\boldsymbol{\theta},\mathbf{A}|\mathbf{y})$. By maximizing the joint likelihood with respect to $\boldsymbol{\theta}$ and $\mathbf{Z}$ simultaneously, it is possible to infer the best random effect design matrix and the best variance component estimates. Nevertheless, our IMURM algorithm is not implemented for maximizing the joint likelihood but the marginal. More advanced methods, e.g. those used in phylogenetic tree estimation (FELSENSTEIN 2004), using the optimality criterion of maximum likelihood, often under a Bayesian framework, might be utilized to estimate IMURM jointly. Identifying the optimal design matrix is NP-hard using these methods (FELSENSTEIN 2004), so heuristic search and optimization methods might be used to identify a reasonably good

design matrix that fits the data. In genetics application using Bayesian methods, Gibbs sampling can be used to estimate the incidence matrix, however, it is not guaranteed that the chain is irreducible in large complex problems, and even if the chain is irreducible, mixing can be quite slow (SORENSEN and GIANOLA 2002). We may introduce Bayesian computation when it is possible and reasonable to joint-estimate the model, otherwise, inference should focus on the parameters.

AUTHORS CONTRIBUTIONS

XS and LR derived the main algorithm for estimating covariate-uncertain random effect models together. XS implemented the main algorithm, did the simulation and drafted the paper. XS and LR worked on the revision together and both approved the final version.

ACKNOWLEDGEMENTS

XS is funded by a Future Research Leaders grant from the Swedish Foundation for Strategic Research (SSF) to professor Örjan Carlborg at Uppsala University and Swedish University of Agricultural Sciences. LR is funded by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS).

LITERATURE CITED

BRESLOW, N. E., and D. G. CLAYTON, 1993 Approximate inference in generalized linear mixed models. Journal of the American Statistical Association

.

FELSENSTEIN, J., 2004 *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

GOLDGAR, D., 1990 Multipoint analysis of human quantitative genetic variation. Am. J. Hum. Genet. **47**: 957–967.

HARVILLE, D. A., 1977 Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association **72**: 320–338.

LEE, Y., and J. A. NELDER, 1996 Hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society. Series B (Methodological) **58**: 619–678.

LEE, Y., and J. A. NELDER, 2006 Double hierarchical generalized linear models. Applied Statistics **55**: 139–185.

LEE, Y., J. A. NELDER, and Y. PAWITAN, 2006 *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall/CRC.

LYNCH, M., and B. WALSH, 1997 *Genetics and Analysis of Quantitative Traits*. Sinauer Assoc.

PAWITAN, Y., 2001 *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications.

R DEVELOPMENT CORE TEAM, 2009 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

ROBINSON, G. K., 1991 That BLUP is a good thing: The estimation of random effects. Statistical Science **6**: 15–32.

RÖNNEGÅRD, L., F. BESNIER, and Ö. CARLBORG, 2008 An improved method for quantitative trait loci detection and identification of within-line segre-

gation in f2 intercross designs. Genetics **178**: 2315–2326.

RÖNNEGÅRD, L., X. SHEN, and M. ALAM, 2010 hglm: a package for fitting hierarchical generalized linear models. Submitted .

SCHORK, N. J., 1993 Extended multipoint identity-by-descent analysis of human quantitative traits: Efficiency, power, and modeling considerations. Am. J. Hum. Genet. **53**: 1306–1319.

SORENSEN, D., and D. GIANOLA, 2002 *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer.

XU, S., 1996 Computation of the full likelihood function for estimating variance at a quantitative trait locus. Genetics **144**: 1951–1960.

XU, S., and W. R. ATCHLEY, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics **141**: 1189–1197.

TABLE AND FIGURE LEGENDS

Table 1 Estimated heritability values using different methods. The variance components of IMURM was estimated using the proposed algorithm with different number of MC imputes. Compared to the simulated true value, the estimated h^2 from IMURM had much less bias than that from the expectation method. For the model using true genotypes, information about the random effect design matrix was complete.

Figure 1 A simulated animal pedigree. There are three generations of animals with five offspring, three parents and two grandparents (founders). Squares and circles denote male and female animals, respectively, with animal indices inside. Bubbles with arrows pointing to each animal indicate the

unobservable true genotypes.

Figure 2 Convergence and estimation behaviors of IMURM. (a) Convergence of the estimated genetic variance with respect to the number of MC imputes, where each point was calculated by one run of the algorithm. (b) Convergence of the genetic variance in one run of the algorithm with 10000 MC imputes. (c) Empirical distribution of 10000 generated sets of genotypes, which was measured as the 'distance to the true genotypes', defined as $\sum_{elements} \text{abs}(\mathbf{Z}_i - \mathbf{Z}_{true})$. (d) Empirical density function of the genetic variance estimates from 10000 generated sets of genotypes. (e) Empirical density function of the (log-) profile likelihood values from 10000 generated sets of genotypes. Results on the true genotypes are pointed in subfigures (c), (d) and (e).

Table 1: **Estimated heritability values using different methods.** (More in TABLE AND FIGURE LEGENDS)

Method	Estimated h^2*	Bias
IMURM - number of MC imputes: 100	.8462	-.0913
200	.8669	-.0706
500	.8829	-.0546
1000	.8955	-.0420
2000	.9037	-.0338
5000	.9086	-.0289
10000	.9087	-.0288
Mixed model using true genotypes	.9785	.0410
Mixed model by expectation method	.4485	-.4890
Simulated value	.9375	-

* Defined as $\hat{\sigma}_g^2/(\hat{\sigma}_g^2 + \hat{\sigma}_e^2)$.

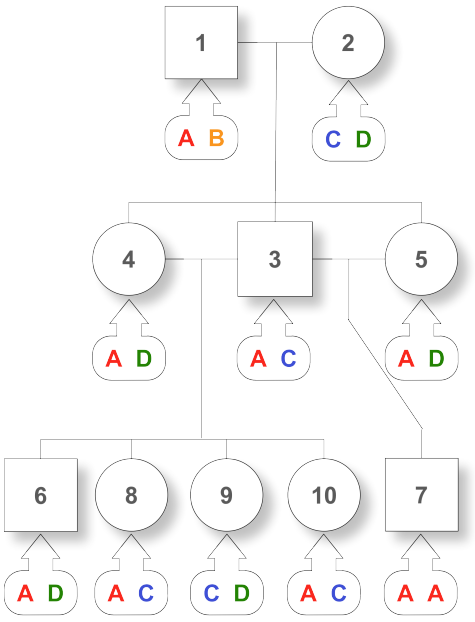

Figure 1: **A simulated animal pedigree.** (More in TABLE AND FIGURE LEGENDS)

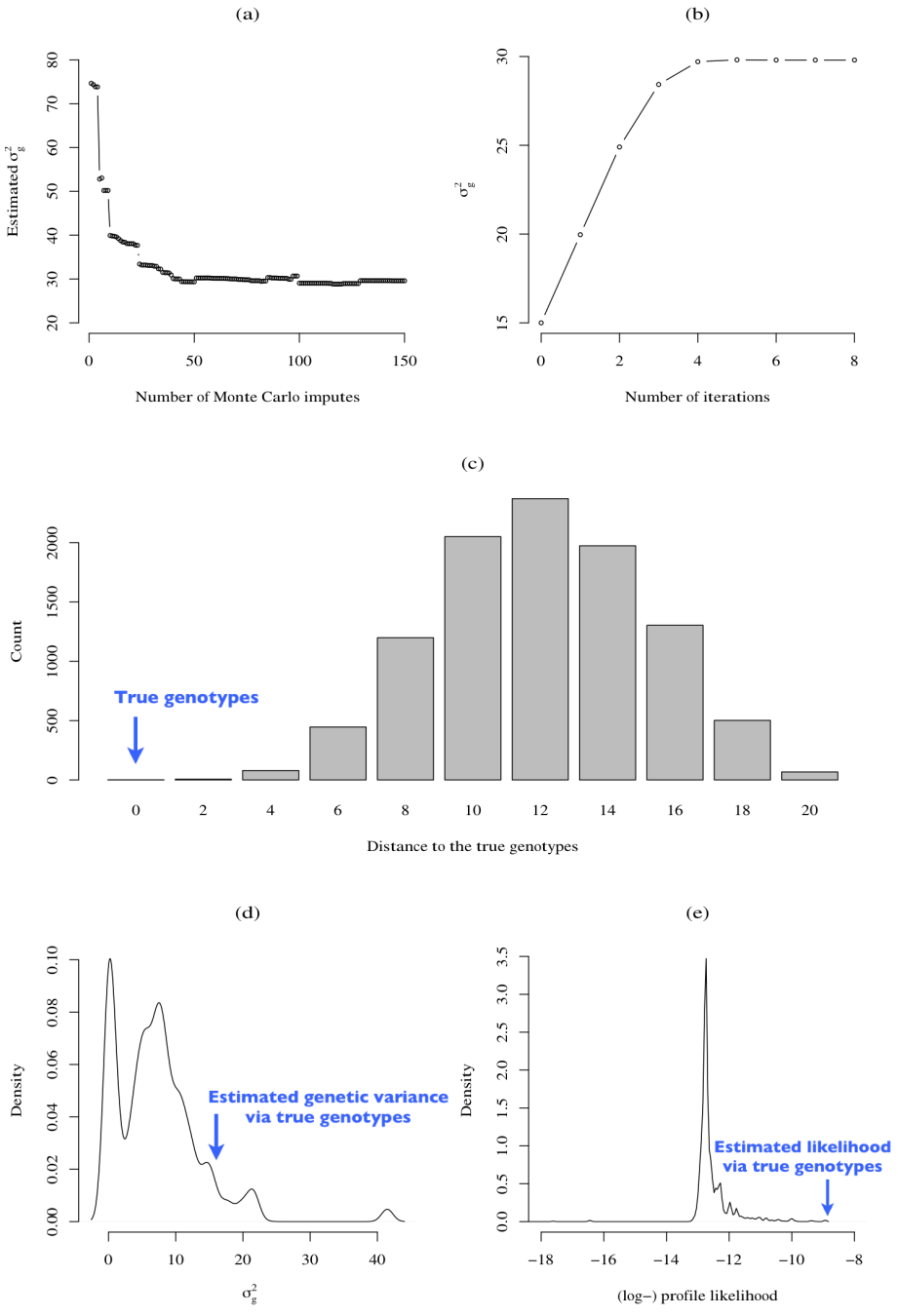

Figure 2: **Convergence and estimation behaviors of IMURM.** (More in TABLE AND FIGURE LEGENDS)