The Use of Linear Mixed Models to Estimate Variance Components from Data on Twin Pairs by Maximum Likelihood

Peter M. Visscher, Beben Benyamin, and Ian White

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Scotland, United Kingdom

It is shown that maximum likelihood estimation of variance components from twin data can be parameterized in the framework of linear mixed models. Standard statistical packages can be used to analyze univariate or multivariate data for simple models such as the ACE and CE models. Furthermore, specialized variance component estimation software that can handle pedigree data and user-defined covariance structures can be used to analyze multivariate data for simple and complex models, including those where dominance and/or QTL effects are fitted. The linear mixed model framework is particularly useful for analyzing multiple traits in extended (twin) families with a large number of random effects.

Estimation of variance components in the classical twin design have been performed using least squares (Jinks & Fulker, 1970), weighted least squares (e.g., Kendler et al., 1994) and maximum likelihood (e.g., Martin & Eaves, 1977). Maximum likelihood has become the method of choice, because of its desirable asymptotic properties and the availability of versatile and powerful computer programs such as Fisher (Hopper, 1988; Lange et al., 1976) and Mx (Neale, 1997; Neale & Maes, 2004). These programs are typically parameterized as covariance models, so that the variance–covariance structure of pairs of twins needs to be specified.

In this note we show that standard linear mixed models can be used to estimate variance components, and in particular that pedigree packages that are designed to estimate variance components in large general pedigrees can be exploited to efficiently estimate parameters in univariate and multivariate models.

Methods

In a mixed linear model, the latent variables (random effects) are specified in the model for the means, and the covariance structure(s) are specified by linking phenotypes with levels of random effects. Consider the commonly used ACE model for individual *i*,

$$y_i = \mu + A_i + C_i + E_i \qquad [1]$$

with μ the overall mean and A, C and E random additive genetic, common environmental and residual effects, respectively. The phenotypic variance in the population is partitioned as

$$\mathrm{var}(y_i) = \mathrm{var}(A) + \mathrm{var}(C) + \mathrm{var}(E)$$

and the covariance between individuals *i* and *j*, who belong to the same family, is

$$\mathrm{cov}(y_i,y_j) = a_{ij}\mathrm{var}(A) + \delta_{ij}\mathrm{var}(C)$$

with a_{ij} the coefficient of relationship (1 for monozygotic [MZ] and 1/2 for dizygotic [DZ]) and δ_{ij} an indicator variable which is 1 if *i* and *j* are twins raised together and 0 otherwise. The additive genetic value (A) for an individual can be partitioned into the effects inherited from the parents and the deviation (M_i) from the parental average (A_{pa}):

$$A_i = 1/2\,A_{dad} + 1/2\,A_{mum} + M_i = A_{pa} + M_i$$

(e.g., Falconer & Mackay, 1996). In a noninbred randomly mating population with homogeneous variances in males and females, the total additive genetic variance is partitioned as

$$\mathrm{var}(A) = 1/4\mathrm{var}(A_{dad}) + 1/4\mathrm{var}(A_{mum}) + 1/2\,\mathrm{var}(A)$$

The term A_{ap} is the average additive genetic value of the parents, with variance $1/2\,\mathrm{var}(A)$. M_i is sometimes called the Mendelian sampling term, and its variance, sometimes called the segregation variance, is the within-family additive genetic variance ($= 1/2\,\mathrm{var}(A)$). Hence, an equivalent model to [1] is

$$y_i = \mu + A_{pa(i)} + C_i + M_i + E_i = \mu + \mathrm{Pair}_i + M_i + E_i \qquad [2]$$

The random effect *Pair* is common to a pair of twins, whether MZ or DZ. The random effect *M* is shared by a pair of MZ twins but not by a pair of DZ twins. Hence, the (co)variances of the new random effects are

Received 3 August, 2004; accepted 28 September, 2004.

Address for correspondence: Peter M. Visscher, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, Scotland, UK. E-mail: peter.visscher@ed.ac.uk

https://doi.org/10.1375/twin.7.6.670 Published online by Cambridge University Press

var(Pair) = ½var(A) + var(C), var(M) = ½var(A), and cov(y_i,y_j) = m_{ij}var(M) + δ_{ij}var(Pair).

The indicator variable m_{ij} is 1 for an MZ pair and 0 otherwise. In slightly different notation, with *i* indicating a pair and *j* (= 1,2) the individual within a pair

MZ: $y_{ij} = \mu + Pair_i + M_i + E_{ij}$

DZ: $y_{ij} = \mu + Pair_i + M_{ij} + E_{ij}$

The parameterization in [2] lends itself to an analysis with any statistical packages that can fit a mixed linear model with uncorrelated random effects, for example, SAS, SPSS, Genstat, Splus and R. It does not require a statistical package specifically designed for genetic analysis. An example of the required coding of the data file is given in Table 1. A mixed linear model analysis using an appropriate standard statistical package, specifying *M* and *Pair* as random effects, would produce maximum likelihood estimates of random effects which are linear functions of the parameters of interest. Hence,

$\hat{\sigma}^2_a = 2\hat{\sigma}^2_m$ and $\hat{\sigma}^2_c = \hat{\sigma}^2_{pair} - \hat{\sigma}^2_m$

A disadvantage of this approach is that negative estimates of the common environmental variance can occur when $\hat{\sigma}^2_{pair} < \hat{\sigma}^2_m$, unless a constraint can be specified. This problem is likely to be more severe for multivariate models. Fitting a CE model in this way is straightforward, but fitting an AE or ADE model is not, because there is no linear combination of two (three) uncorrelated random effects that correspond to the underlying AE (ADE) effects. To allow for more complex models and for implicit constraints on the estimates of the causal variance components, statistical packages can be used that allow for arbitrary pedigrees files and/or user-defined covariance structures.

Variance component estimation software that can handle pedigree data and multiple random effects with user-defined covariance structures can also be used for twin analyses, and provides a flexible way to fit commonly used models (with the usual constraints) and to fit single or multiple QTL effects. For such software, for example, ASREML (Gilmour et al., 1995; Gilmour et al., 2002) and VCE (Neumaier & Groeneveld, 1998), typically two input files are specified: a data file containing rows for each individual with a measured phenotype and codes for fixed effects, covariates and levels of random effects, and a pedigree file which specifies the genetic relationship between individuals with phenotypes. To fit data from DZ and MZ twins in this framework, the pedigree file needs to be coded so that DZ are recognized as 'normal' full-sibs, by specifying common parents, whereas MZ individuals appear only once in the pedigree file as a single entity (genotype) with unknown parents. In the phenotype file each MZ individual has an entry, with the same code for the random effects as its co-twin. With this parameterization, the effects A, C and E are modeled directly. An example of the coding of a pedigree and data file is given in Tables 1 and 2. In the model specification for the analysis, the effects of factors *Pair* and *M* are random and the covariance structure of M is determined by the pedigree file. Note that although the coding of the *Pair* and *M* factors in the data file is the same for nonpedigree and pedigree packages, the variance of the effects are different. In the nonpedigree packages the variance associated with *Pair* and *M* is (½var(A) + var(C)) and ½var(A), respectively, whereas in the pedigree package the variances are var(C) and var(A), respectively.

The extension to multivariate analysis is straightforward. In Appendix A an example of an ASREML command file for an ACE model with three traits is given.

If the option of user-defined covariance matrices is supported, then the mixed linear model approach can be used for QTL analysis or fitting addition twin models. In ASREML, for each additional random effect with an arbitrary covariance structure, an additional file is supplied that contains the inverse of the covariance matrix. In Appendix B, how to code a dominance effect (D) when fitting an ADE model is shown.

Maximum likelihood estimation of variance components in effect assumes that the fixed effects are known without error, which leads to biased estimates of the variance components. In the simplest case of $y_i = \mu + e_i$ and *n* observations, the maximum likeli-

Table 1

Example Data Input File for Variance Component Analysis[1]

Individual	Twin pair	MZ/DZ (1/2)	Level of Pair	Level of M	Trait
1	1	1	P1	M1	Y1
2	1	1	P1	M1	Y2
3	2	2	P2	M2	Y3
4	2	2	P2	M3	Y4
5	3	2	P3	M4	Y5
6	3	2	P3	M5	Y6
7	4	1	P4	M6	Y7
8	4	1	P4	M6	Y8

Note: [1]The first 3 columns are only given for clarification, but are not needed.

Table 2

Example Pedigree File for the Same Individuals as in Table 1[1]

Individual	Twin pair	MZ/DZ (1/2)	ID	ID dad	ID mum
1 & 2	1	1	ID1	0	0
3	2	2	ID2	DAD2	MUM2
4	2	2	ID3	DAD2	MUM2
5	3	2	ID4	DAD4	MUM4
6	3	2	ID5	DAD4	MUM4
7 & 8	4	1	ID6	0	0

Note: [1]The first 3 columns are only given for clarification, but are not needed.

https://doi.org/10.1375/twin.7.6.670 Published online by Cambridge University Press

hood estimate of σ_e^2 is $\sum(y - \bar{y})^2/n$, which is biased by a factor of $(n - 1)/n$. When the number of observations is large relative to the number of fixed effects or covariates to be estimated, this bias is small. In residual (or restricted) maximum likelihood (REML, Patterson & Thompson, 1971), only the part of the likelihood which is independent of fixed effects is maximised, by taking into the account the loss in degrees of freedom by estimating fixed effects. In balanced designs, REML estimates are identical to ANOVA estimates of variance components. For the analysis of samples from human populations, the number of covariates is usually small relative to the number of observations, so that the use of either ML or REML is likely to lead to the same statistical inference. Pedigree software written for large populations, such as ASREML (Gilmour et al., 1995) and VCE (Neumaier & Groeneveld, 1998) are based on residual maximum likelihood estimation. These programs were designed for large complex pedigrees, and a multivariate analysis of, say, tens of thousands of twin pairs would be feasible and computationally efficient.

Discussion and Conclusions

In this note the equivalence between a twin covariance model and a mixed linear model has been shown. Standard statistical software can be used for the simplest of models, for example, for CE and ACE models.

Guo and Wang (2002) showed how many complex models used in behavior genetic analysis could be fitted with mixed or 'multilevel' models and presented SAS codes for maximum likelihood (or REML) analysis for a number of models. Guo and Wang (2002) group the type of relationships that are available in the data into clusters, and perform between- and within-cluster analysis of variance. For example, if the data consists of observations on MZ pairs, DZ pairs and pairs of halfsibs, then three between- and three within-cluster variances are estimated. Essentially, their proposed method is a (maximum likelihood) generalization of analysis of variance, in which the 'observable' rather than causal effects are modeled. However, the proposed method does not generate maximum likelihood estimates of causal components (or their ratios) for nonsaturated models, because the underlying (hypothesized) model is not fitted directly. For example, in the case of 'clusters' of MZ and DZ twin pairs and the ACE model, the estimate of the heritability from the Guo and Wang (2002) mixed model is twice the difference in the ML estimate of the MZ and DZ intraclass correlation coefficient (= between-pair variance/[between-pair variance + within-pair variance]), which is the same as the least-squares estimator. The assumption of the ACE model that the total phenotypic variance is the same for MZ and DZ pairs is not explicitly taken into account because four variance components are estimated. In contrast, our mixed model approach to twin data generates maximum likelihood estimates of the components of the model. To fit and test parsimonious submodels using the Guo and Wang (2002) mixed model approach is not obvious, as acknowledged by the authors.

Pedigree-based variance component estimation software can be used to fit more complex models, including additional random effects such as dominance and (multiple) QTL and easily allows more complex pedigree structures such as extended twin families and arbitrary deep pedigrees (e.g., George et al., 2000). Thus, the same pedigree-based software can in principle be used to analyze (multivariate) data from a wide variety of pedigree structures and complex linear models, from analysis of the classical twin design to multivariate QTL mapping in complex pedigrees.

Acknowledgments

PMV thanks the UK Biotechnology and Biological Research Council for financial support. We thank Karoline Schousboe for discussions (in Melbourne) that led to this study and Kirsten Ohm-Kyvik and Thorkild Sorensen for continued support. We thank Robin Thompson, Arthur Gilmour and Mike Goddard for helpful discussions and suggestions on how to parameterize 'clones' in variance component estimation software. We thank Shaun Purcell for useful comments.

References

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed.). Harlow, UK: Longman.

George, A. W., Visscher, P. M., & Haley, C.S. (2000). Mapping quantitative trait loci in complex pedigrees: A two step variance component approach. *Genetics, 156*, 2081–2092.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics, 52*, 1440–1450.

Gilmour, A. R., Cullis, B. R., Welham, S.I., & Thompson, R. (2002). *ASReml Reference Manual* (2nd ed.) [Release 1.0]. NSW Agriculture Biometrical Bulletin 3, NSW Agriculture, Australia.

Guo, G., & Wang, J. (2002). The mixed or multilevel model for behavior genetic analysis. *Behavior Genetics, 32*, 37–49.

Hopper, J. L. (1988). Review of FISHER. *Genetic Epidemiology, 5*, 473–476.

Jinks, J. L., & Fulker, D. W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin, 73*, 311–349.

Kendler, K. S., Neale, M. C., Heath, A. C., Kessler, R. C., & Eaves, L. J. (1994). A twin-family study of alcoholism in women. *American Journal of Psychiatry, 151*, 707–715.

Lange, K., Westlake, J., & Spence, M. A. (1976). Extensions to pedigree analysis. III. Variance components by the scoring method. *Annals of Human Genetics, 39*, 485–491

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity, 38*, 79–95.

Neale, M. C. (1997). *Mx: Statistical modeling* (3rd ed.). Richmond, VA: Department of Psychiatry, Medical College of Virginia.

Neale, M. C., & Maes, H. H. M. (2004). *Methodology for genetics studies of twins and families*. Dordrecht, the Netherlands: Kluwer Academic.

Neumaier, A., & Groeneveld, E. (1998). Restricted maximum likelihood estimation of covariance in sparse linear model. *Genetics Selection Evolution, 30*, 3–26.

Patterson, H. D., & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika, 58*, 545–554.

APPENDIX A

Multivariate ASREML Script for 3-trait ACE Model, Fitting Sex and Age as Covariates

```
Text following # is for clarification
Trivariate Analysis
  factorM !P      # Random effect with pedigree file
  pair !I         # Unstructured random effect
  sex 2           # The factor sex has 2 levels
  age
  y1              # }
  y2              # } Three traits
  y3              # }
twins.ped         # Pedigree file for random effect factorM
twins.dat         # Data file
y1 y2 y3 ~ Trait Trait.sex Trait.age !r Trait.pair Trait.factorM
# Model statement. Sex is fitted as a fixed effect and age as a covariate
# for all traits. Pair and factorM are fitted as random effects.
1 2 2 # 1 = no. sites; 2= error dimension; 2 = no. G structures
0 # default {ASREML Syntax}
Trait 0 US !GP # Trait = no. columns (= 2); 0 = default
1       # E 3x3 covariance matrix, starting values
0 1
0 0 1
Trait.pair 2 # 1st G-structure & dimension
Trait 0 US !GP
1       # C 3x3 covariance-matrix, starting values
0 1
0 0 1
pair 0 ID # no structure on pair
Trait.genotype 2 # 2nd G-structure
Trait 0 US !GP
1       # A 3x3 covariance matrix, starting values
0 1
0 0 1
factorM 0 AINV # FactorM follows standard relationship rules
```

Example first few lines of `twins.dat`:

```
M1 P1 male 39.0 1.0 0.9 0.5
M1 P1 male 39.0 1.1 0.8 0.6
M2 P2 female 35.0 1.8 0.4 1.3
M3 P2 male 35.0 1.2 0.7 2.5
```

Example first few lines of twins.ped:

```
M1 0 0
M2 dad23 mum23
M3 dad23 mum23
```

For more details on ASREML, see http://www.vsn-intl.com/ASReml/index.htm

https://doi.org/10.1375/twin.7.6.670 Published online by Cambridge University Press

APPENDIX B

Performing an ADE Analysis in a Mixed Linear Model

We first demonstrate how to fit an ADE model with user-defined covariance structures, without the use of a pedigree file. A random factor ID is created which is the same as the pair identifier for MZ twins but takes a different value for DZ twins. For example, if pairs 1 and 2 are MZ and pairs 3 and 4 are DZ, we have,

PAIR	TWIN	ID
1	1	1
1	2	1
2	1	2
2	2	2
3	1	3
3	2	4
4	1	5
4	2	6

In the model specification, ID is fitted as a random effect and fitted twice, with two user-defined covariance structure, one pertaining to additive genetic (A) effects, and the other one to dominance (D) effects. In ASREML the model statement is

Y ~ mu !r giv(ID,1) giv(ID,2)

In programs such as ASREML, the inverse of the covariance structures of all levels of ID need to be specified in separate files. The first random effects corresponds to additive genetic effects. The covariance structure of DZ twins for A is

$$\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

with inverse

$$\begin{bmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{bmatrix}$$

MZ twins are specified as only one level of the factor ID with variance (and inverse) of 1. The second random term corresponds to the dominance effects. For the dominance effect, the covariance structure of DZ twins is

$$\begin{bmatrix} 1 & 1/4 \\ 1/4 & 1 \end{bmatrix}$$

with inverse

$$\begin{bmatrix} 16/15 & -4/15 \\ -4/15 & 16/15 \end{bmatrix}$$

Again, for MZ the same rules apply as for the additive effects. Hence, for the above example, the nonzero entries for the first random term (A) in the model statement are:

1 1 1
2 2 1
3 3 1.3333
4 3 –0.6667
4 4 1.3333
5 5 1.3333
6 5 –0.6667
6 6 1.3333

and the file associated with the second term (D) is similar but with 1.3333 (4/3) replaced by 1.0667 (16/15) and –0.6667 (–2/3) replaced by –0.2666 (–4/15).

This parameterization requires the creation of user-defined covariance matrices but the script files remain simple. The same model could also be fitted using a standard pedigree file and a single user-defined covariance structure corresponding to dominance effects.

https://doi.org/10.1375/twin.7.6.670 Published online by Cambridge University Press