

Methods to Estimate Genetic Components of Variance for Quantitative Traits in Family Studies

Mariza de Andrade,* Christopher I. Amos, and Tracy J. Thiel

Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston

The aim of this paper was to compare several methods of estimating the genetic components of a quantitative trait in familial data. The Expectation and Maximization (E-M) algorithm, the Newton-Raphson method, and the scoring method were compared for estimating polygenic and environmental effects on nuclear families. We also compared scoring and quasilikelihood (QL) methods when a linked genetic marker was available to estimate effects from a major gene. Generally, all procedures performed similarly in estimating polygenic and environmental variance components. The E-M algorithm yielded more precise estimators when heritability was low. The scoring method was much faster than the other methods and yielded slightly more precise estimates of mean effects but slightly less precise estimates of the variance components. Estimates of major gene effects were not affected by the number of alleles at the trait locus. For these relatively large sample sizes, QL and scoring had similar precision, but QL took 32 times longer than scoring. Finally, we compared the results of applying these methods to data from the Bogalusa Heart Study. Results showed larger imprecision when the QL method was applied, consistent with earlier studies that showed decreased precision of quasilikelihood compared with maximum likelihood in moderately small sample sizes. *Genet. Epidemiol.* 17:64–76, 1999. © 1999 Wiley-Liss, Inc.

Key words: variance components; iterative estimation methods; major gene effect; linkage

Contract grant sponsor: National Institute of Health; Contract grant numbers: R01-GM52607, P01-CA34936, 5R01HD32194, 5U01HL38844.

*Correspondence to: Dr. Mariza de Andrade, Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, P.O. Box 189, 1515 Holcombe Blvd, Houston TX 77030. E-mail: mandrade@request.mdacc.tmc.edu

Received 30 March 1998; Revised 29 June 1998; Accepted 31 August 1998

© 1999 Wiley-Liss, Inc.

INTRODUCTION

One of the primary goals of human genetics research is to study the familial aggregation of quantitative traits such as systolic blood pressure or high-density lipoproteins. Of particular interest is whether these traits are more affected by genes or by the environment, or whether both genes and environment are involved. If genetic factors influence interindividual variability of a trait, identifying the specific factors involved is a major goal of most current genetic studies. Although some environmental factors can be observed directly and modeled as fixed effects, genetic effects are typically unobservable. To identify the total proportion of variance that can be attributed to genetic factors, mixed linear models, path analysis, and other methodologies have been developed and applied, as discussed by Thompson and Shaw [1990] and Rao et al. [1974]. Although effects from specific alleles at identified genetic loci are usually not available for the study of quantitative traits, a dense map of markers is currently available for human genome wide linkage analysis. These markers usually do not have a direct effect on trait variability, but if they are linked, cosegregation of the linked markers with a trait locus can be used to partition interindividual variability into linked and unlinked components of variance [Amos, 1994].

Some environmental effects can be attributed to known and measurable factors, and a fixed effects model is appropriate to represent these sources of variability. Other sources of variability, such as measurement error, reflect an inherently variable process and are specified by a random effects model. In some cases, unmeasurable environmental effects due to factors such as shared household effect can be modeled as a random effect. Genetic sources of variability can be modeled as either fixed or random components of variance. Observed effects from specific alleles at a locus, which are believed to directly affect trait variability, are modeled as fixed effects. Polygenic effects, which arise from effects of many unlinked loci with unmeasurably small effects, induce a correlation structure among relatives [Fisher, 1918] and are modeled as a random effect. The effect from a major locus, which by definition has an estimably large effect, can be modeled as a fixed or random component. If the number of alleles at a locus is known, the unobservable genetic effects from the locus could be modeled as a fixed effect. Standard models of qualitative traits, such as being affected by a disease, assume a simple biallelic genetic model. For many of the quantitative traits that have been characterized at the molecular level, such as Lp(a), α 1-antitrypsin, and galactosemia, this assumption seems to be invalid, because a large number of variant alleles are typically observed in those cases. Thus, a random effects model may be more appropriate for assessing the effects from a genetic factor linked to a marker. If the genetic effect actually results from a single locus that has two alleles, specifying a random effects model has been shown in simulation studies to lead to unbiased estimates of the variance components provided the sample size is large enough and families are not selected through extreme individuals [Amos et al., 1996].

For the preliminary evaluation of genetic linkage using quantitative traits, numerous strategies have been developed. The Haseman-Elston (H-E) method is the simplest of these [Haseman and Elston, 1972]. In this procedure, the squared difference between trait values is regressed upon the estimated proportion of genes “identical by descent” (IBD) at a marker locus. In the absence of linkage, there is no

relationship between IBD at the marker locus and the squared pair differences. In the presence of linkage, pairs of sibs who share two genes IBD are concordant for the genetic factor that influences variation in trait levels. Therefore, they show a smaller squared pair difference than pairs that share no alleles IBD. Thus, a simple test for linkage can be developed by testing for a negative regression of squared pair differences onto IBD. The H-E method is easy to apply, has been shown to detect linkage in the presence of nonnormally distributed residual nongenetic variance [Blackwelder and Elston, 1985], and has been extended to provide a test for linkage using multivariate data [Amos et al., 1990]. This method has also been shown, however, to have less power than methods that jointly model the distribution of either sib pairs or members of sibships [Amos et al., 1996; Wright, 1997]. Additionally, modeling covariate effects is difficult with the H-E method [Pugh et al., 1997; Amos, 1994]. To provide a simpler framework for modeling data from sibships and extended families and for inclusion of covariates, Hopper and Mathews [1982] and Amos [1994] suggested the use of a components of variance approach. The obvious approach of simply applying standard likelihood theory and assuming a multivariate normal distribution to model the residual distribution of the phenotypes (after conditioning on fixed effects such as measurable covariates) in families may not be appropriate if unobservable major genes are segregating, because this segregation introduces platykurtosis and may skew the distribution of trait values. Extensive simulations by Amos et al. [1996] failed to document significant bias attributable to the moderate kurtosis introduced by major gene effects, but robust estimation procedures were preferable when the nongenetic variation was markedly nonnormal.

Comparisons of the effectiveness of various estimation procedures of components of variance models in genetic studies are rather limited in the literature. Thompson and Shaw [1990] applied the Expectation and Maximization (E-M) algorithm to the polygenic model in families. Maximum likelihood (ML) methods have been applied by Lange et al. [1976], Hopper and Mathews [1982], and Lange and Boehnke [1983] to the polygenic model as well. Schork [1991] compared various methods of computation of the mixed model using a variety of purely numerical approaches. Results indicated that a variety of methods using variable metric adaptations of N-R methods yielded similar likelihood values across several replicates, suggesting that these methods provided similar estimates. Finally, Amos et al. [1996] applied quasilielihood (QL) methods to model genetic parameters in nuclear families. For a simple model involving only polygenic and nongenetic components of variance, the QL method reduces to ML estimation. The estimation procedures we used for Fisher scoring and for QL estimation are similar, so we do not compare this case. Modeling the major gene component, however, requires evaluating a nonpatterned variance-covariance matrix, so that QL estimation is no longer equivalent to ML estimation in this context. Amos et al. [1996] compared application of the scoring algorithm for QL estimators to Newton-Raphson (N-R) and direct search methods under ML estimation. Because the numerical optimization of QL and ML estimating procedures was not identical, results of the Amos group's studies [1996] provide a limited evaluation of the efficiency of the estimation methods per se. In particular, because numerical optimization methods were used for ML estimation, the precision of the ML procedure could have been underestimated.

Our main aim in this paper is to overview the iterative methods commonly ap-

plied in variance components (VC) models and to evaluate how they behave when they are used to estimate the genetic components of variance. Our comparative studies evaluate the precision of various methods for variance components estimation as well as the computing time required by the methods. It is important to note that the structure of VC in familial data is slightly different from that of the usual VC using repeated measures or longitudinal data. In familial data, each family is a cluster of individuals who share common genes. Therefore, the maximum likelihood expression cannot usually be written in terms of design matrices [Searle et al., 1992]. In Methods, we describe the VC models of familial data and how the iterative procedures behave. In Results, we present findings from simulation studies and an application to data from the Bogalusa Heart Study.

METHODS

Background

The model we describe here represents the observed quantitative trait of an individual as a function of fixed and random effects. The random effects can be represented by the polygenic and major gene components. For the polygenic component, as is customary in genetics, we assume that a large number of loci are involved in influencing the trait. We assume no interaction among the polygenic loci (no epistasis) and random mating. Based on this assumption, the dominance effect is minimized for the polygenic effect [Morton et al., 1968; Morton, 1974], and the polygenic component is assumed to be adequately represented by the additive genetic effect.

Consider the case in which a particular trait, such as systolic blood pressure, is observed for families (or clusters of related individuals). Under a polygenic model the observed values of the trait for the members of the i th family can be represented by

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{a}_i + \boldsymbol{\varepsilon}_i, \quad (1)$$

where \mathbf{y}_i is the vector of the observed values of the trait (or phenotype) for the i th family, $\boldsymbol{\mu}$ is the vector of overall mean, \mathbf{a}_i is the unobservable vector of the additive random genetic effects for the i th family, \mathbf{X}_i is the matrix of observable covariates, $\boldsymbol{\beta}$ is the vector of fixed covariate effects uncorrelated with the additive genetic effects and environmental effects, and $\boldsymbol{\varepsilon}_i$ is the vector of environmental effects for the i th family, $\forall i, i = 1, 2, \dots, k$.

We assume that the additive genetic effect for the i th family is $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{G})$, where \mathbf{G} is the matrix of coefficients of relationship between pairs of related individuals, for example, $1/2$ for full sibs and parent-offspring; the environmental effect for the i th family is $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$, where \mathbf{I} is the identity matrix; and the additive genetic effect is uncorrelated with the environmental effect. Then $\mathbf{y}_i \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{X}_i\boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{V} = \sigma^2\mathbf{G} + \tau^2\mathbf{I}$.

To estimate the VCs, σ^2 and τ^2 , we can use the E-M algorithm, N-R, and scoring methods. These methods are well described in the literature for the case in which a design matrix can be specified for each VC [Searle et al., 1992]. For most genetic problems in humans, however, no design matrix can be specified a priori because the structure of the matrix will vary from family to family and markers are usually less than fully informative, and few reports concern the estimation of VC in such a case.

The E-M algorithm is based on sufficient statistics of the complete data, which consist of the observed and unobserved data. Observed data are the observed quantitative traits or phenotypes, and unobserved data are the random components, in our case, the genetic and environmental components. The N-R and scoring methods are based on the derivatives of the log likelihood function and each method has its own advantages and disadvantages. The E-M algorithm generally takes longer to converge but does not produce negative estimates of the variance components, whereas the N-R and scoring methods converge faster, but can produce negative estimates unless boundary constraints are imposed. The E-M algorithm can yield the breeding values, an important measure in animal breeding but it does not provide an information matrix of the estimates. To produce this matrix, we need to apply an additional step [Louis, 1982; Meng and Rubin, 1991]. Conversely, the scoring and N-R methods automatically provide this information matrix.

Expectation-Maximization Algorithm

The basic form of E-M equations for the above model are well known [Laird, 1982; Dempster et al., 1977]. The natural sufficient statistics for the “complete data” (\mathbf{y} , \mathbf{a}) when there are no fixed effects except for the overall mean $\boldsymbol{\mu}$, are $k^{-1}\mathbf{1}'(\mathbf{y} - \mathbf{a})$, $k^{-1}\mathbf{a}'\mathbf{G}^{-1}\mathbf{a}$, and $k^{-1}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, whose unconditioned expectations are $\boldsymbol{\mu}$, σ^2 , and τ^2 , respectively. Thus, the iterative equations are formed by setting new values for these parameters equal to the conditional expectations of the statistics taken at current parameter values

$$\boldsymbol{\mu}^{(+)} = k^{-1}\mathbf{E}(\mathbf{1}'(\mathbf{y} - \mathbf{a}) \mid \mathbf{x}, \boldsymbol{\mu}, \sigma^2, \tau^2) = k^{-1}\mathbf{1}'(\mathbf{y} - \boldsymbol{\eta}), \quad (2)$$

$$\sigma^{2(+)} = k^{-1}\mathbf{E}(\mathbf{a}'\mathbf{G}^{-1}\mathbf{a} \mid \mathbf{y}, \boldsymbol{\mu}, \sigma^2, \tau^2) = k^{-1}[\text{tr}(\mathbf{G}^{-1}\boldsymbol{\Sigma}) + \boldsymbol{\eta}'\mathbf{G}^{-1}\boldsymbol{\eta}], \quad (3)$$

$$\tau^{2(+)} = k^{-1}\mathbf{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \mid \boldsymbol{\mu}, \sigma^2, \tau^2) = k^{-1}[\text{tr}(\boldsymbol{\Sigma}) + \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}], \quad (4)$$

where

$$\boldsymbol{\eta} = \mathbf{E}(\mathbf{a} \mid \mathbf{x}, \boldsymbol{\mu}, \sigma^2, \tau^2) = \mathbf{y} - \boldsymbol{\mu} - \boldsymbol{\varepsilon},$$

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{G}(\mathbf{I} - \sigma^2\mathbf{V}^{-1}\mathbf{G}) = \sigma^2\tau^2\mathbf{G}\mathbf{V}^{-1},$$

$$\text{tr}(\mathbf{G}^{-1}\boldsymbol{\Sigma}) = \sigma^2\tau^2\text{tr}(\mathbf{V}^{-1}),$$

$$\text{tr}(\boldsymbol{\Sigma}) = \sigma^2\tau^2\text{tr}(\mathbf{G}\mathbf{V}^{-1}) = \tau^2\text{tr}(\mathbf{I} - \tau^2\mathbf{V}^{-1}).$$

The problem with these E-M equations is that they all require computation of \mathbf{V}^{-1} at each iteration. For large pedigrees, the computation of \mathbf{V}^{-1} at each iteration can become computationally intensive. To deal with this issue, Thompson and Shaw [1990] solved the above E-M equations without determining \mathbf{V}^{-1} . Instead, only the eigenvalues of \mathbf{G} needed to be determined, which provide the eigenvalues of \mathbf{V} , hence, \mathbf{V}^{-1} . Thus the trace terms of equations (3) and (4) are easily computed. In our study we analyzed only small pedigrees and therefore inverted \mathbf{V} to solve iterative equations.

Newton-Raphson Method

In the VC estimation problem, we can estimate the set of parameters denoted by $\boldsymbol{\theta}$, which consists of the fixed effects coefficients and the VC parameters, by the N-R iterations

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - (\mathbf{H}^{(m)})^{-1} \nabla \mathbf{L}^{(m)}, \quad (5)$$

where \mathbf{L} indicates the log likelihood function, $\mathbf{H}^{(m)}$ the Hessian matrix (second-derivative matrix), and $\nabla \mathbf{L}^{(m)}$ the gradient vector, with $\boldsymbol{\theta}$ replaced by $\boldsymbol{\theta}^{(m)}$. In our studies, we used a numerical method to perform N-R iterations as implemented by MAXFUN [Sorant and Elston, 1994].

Scoring Method

This method uses an iteration scheme similar to N-R, replacing the Hessian matrix with the information matrix, which is the negative expectation of the Hessian matrix. By doing so, the information matrix need only be calculated once, thus avoiding the computational burden of iteratively calculating the Hessian as required by N-R. Then the scoring method uses the following iteration scheme:

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + (\mathbf{I}^{(m)})^{-1} \nabla \mathbf{L}^{(m)}, \quad (6)$$

where $\mathbf{I}^{(m)}$ is the information matrix. We wrote scientific code to solve iterative equations for E-M and scoring using the C programming language. For all methods we used the identical convergence criterion that the change in likelihood between iterations must be less than 1×10^{-5} .

Major Gene Models

Advances in molecular biology enable us to evaluate the association and genetic linkage of markers with quantitative traits. Let us assume that in addition to the polygenic additive effect, a major gene is responsible for the trait and a marker locus is linked with this major gene. Then equation (1) can be rewritten as

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{a}_i + \mathbf{g}_i + \boldsymbol{\varepsilon}_i, \quad (7)$$

where \mathbf{g}_i is the unobservable major gene effect for the i th family. Amos [1994] showed that $\mathbf{g}_i \sim (\mathbf{0}, \sigma_g^2 \mathbf{F}_i)$, where $\mathbf{F}_i = (f(\theta, \pi_{ijl}))$, θ is the recombination fraction between the major locus, and π_{ijl} is the IBD sharing assessed by marker typings for pairs of individuals (j, l) at the i th family. The values for $f(\theta, \pi_{ijl})$ for several relationships are given by Amos [1994].

We assume that the dominance component of variance for the major gene effect is negligible. Amos [1988] showed that, except for unusual situations, the additive component of variance usually dominates the dominance variance. For most linkage testing situations, the additive assumption is reasonable. The dominance variance can become appreciable for recessive traits. When the recessive disease allele frequency is less than 1/3, the dominance variance is greater than the additive variance. Consequently, here σ_g^2 refers to the additive major gene component of variance. The value of π_{ijl} can assume only 3 values: 0, 1/2, or 1. However, because data from markers are often incomplete, estimates of π_{ijl} are routinely used as described by Haseman and Elston [1972]. This leads to a less discrete formulation of π values.

Under tight linkage, we assume that $\theta = 0$. Then $f(\theta, \pi_{ijl}) = \pi_{ijl}$, and consequently, $\mathbf{g}_i \sim (\mathbf{0}, \sigma_g^2 \mathbf{Z}_i)$, where $\mathbf{Z}_i = (\pi_{ijl})$. In this case, the variance-covariance matrix for each family \mathbf{y}_i is $\mathbf{V} = \sigma^2 \mathbf{G} + \sigma_g^2 \mathbf{Z}_i + \tau^2 \mathbf{I}$. Here we assume that $\mathbf{y}_i \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{X}_i \boldsymbol{\beta},$

V), despite the fact this assumption may not be correct because the major gene effect does not necessarily follow a normal distribution.

To estimate σ_g^2 , we applied the three methods mentioned above and the QL approach proposed by Amos [1994]. The maximization step of the E-M algorithm is given by

$$\hat{\sigma}_g^2 = E[\mathbf{g}'_i \mathbf{Z}_i^{-1} \mathbf{g}_i | \mathbf{y}_i] = (k_i)^{-1} [\delta'_i \mathbf{Z}_i^{-1} \delta_i + \text{tr}(\mathbf{Z}_i^{-1} V_g)], \quad (8)$$

where k_i is the family size,

$$\delta_i = E[\mathbf{g}_i | \mathbf{y}_i] = \sigma_g^2 \mathbf{Z}_i V^{-1} (\mathbf{y}_i - \boldsymbol{\mu}),$$

$$V_g = \text{var}[\mathbf{g}_i | \mathbf{y}_i] = \sigma_g^2 \mathbf{Z}_i [\mathbf{I} - V^{-1} \sigma_g^2 \mathbf{Z}_i].$$

In this case, we are required to invert \mathbf{Z}_i for each family i . For some families, this matrix may be singular. In particular, singular matrices occur for any family in which two or more sibs share all alleles IBD or in sibships of size 4 or larger, where two sibpairs share no alleles IBD, for example, in a family of 4 siblings whose parents have alleles 1,2 and 3,4, and offspring have alleles 1,3; 2,4; 1,2; and 3,4, respectively. The generalized inverse could be applied to obtain estimators, but usual forms of generalized inverses [Searle, 1982] essentially work with the nonsingular parts of the matrix, deleting rows and columns that result in singularities. In the case covered by equation (8), this would result in incomplete use of phenotype information.

To estimate the VC of the major gene effect, one extra parameter is added to the N-R and scoring equations (5) and (6), respectively. These equations can produce negative estimates of the VC parameters, which is an inadmissible solution [Searle et al., 1992]. We applied the step-halving method suggested by Jennrich and Schluchter [1986] to avoid this problem whenever a parameter was estimated outside the admissible region. We also compare the genetic VC results from the scoring method with QL method. For the QL method, we followed approaches provided by Amos et al. [1996].

RESULTS

Simulation Studies

We performed several simulations. First, we compared E-M algorithm, N-R, and scoring methods for polygenic and environmental VC. The E-M algorithm and scoring method are calculated analytically, whereas the N-R method is calculated numerically. The results are shown in Table I, with respective average CPU time requirements for analysis of one simulation run. Although all methods yielded estimates very close to the generating values, in general, the estimates slightly underestimated these values. Because the distribution of VC estimates is usually slightly skewed, the means are even closer to generating values than the medians (results not shown). For low heritability ($h^2 = 0.1$), E-M was substantially more accurate than N-R or scoring. For other models, scoring provided generally more accurate estimates of main effects and slightly less accurate estimates of random effects. Compared to scoring on average, however, the N-R method required roughly 12-fold more time, and the E-M model took 20 times longer than scoring.

TABLE I. Median Values of 1,000 Simulations From 100 Nuclear Families (2 Parents and 4 Offspring) Using 3 Different Approaches (Values in Parentheses Are the Median Squared Error)*

Exp	Method	h^2	Overall mean (μ)	Polygenic VC (σ^2)	Error VC (τ^2)	CPU time
1	E-M	0.1	-0.000145 (0.0101)	1.008 (0.0948)	8.900 (0.2014)	137.2
2	SCOR	0.1	0.002033 (0.0082)	0.9445 (0.1926)	9.027 (0.2522)	8.2
3	N-R	0.1	2699e-05 (0.0101)	0.9567 (0.1670)	8.988 (0.2505)	84.1
4	E-M	0.3	0.001705 (0.0131)	2.949 (0.2934)	6.993 (0.2042)	135.9
5	SCOR	0.3	-0.000130 (0.0119)	2.917 (0.2980)	7.031 (0.2341)	6.3
6	N-R	0.3	0.001709 (0.0131)	2.950 (0.2930)	6.992 (0.2044)	84.4
7	E-M	0.5	0.009510 (0.0168)	4.919 (0.3586)	4.997 (0.1784)	134.9
8	SCOR	0.5	0.001680 (0.0154)	4.919 (0.4143)	5.027 (0.2001)	6.6
9	N-R	0.5	0.009508 (0.0168)	4.919 (0.3600)	4.997 (0.1784)	79.5
10	E-M	0.7	0.010380 (0.0206)	6.957 (0.4062)	2.990 (0.1227)	137.8
11	SCOR	0.7	0.000138 (0.0191)	6.931 (0.4517)	3.025 (0.1495)	6.9
12	N-R	0.7	0.010380 (0.0206)	6.956 (0.4062)	2.990 (0.1222)	83.1
13	E-M	0.9	0.009015 (0.0242)	8.935 (0.3320)	1.003 (0.0595)	135.9
14	SCOR	0.9	0.003434 (0.0223)	8.934 (0.4446)	0.9953 (0.0811)	6.7
15	N-R	0.9	0.008759 (0.0251)	8.966 (0.3789)	0.9927 (0.0712)	82.9

*CPU in seconds per run on average on a SUN Sparcstation 20. For experiments 1-3, $\sigma^2 = 1$; $\tau^2 = 9$, for experiments 4-6; $\sigma^2 = 3$, $\tau^2 = 7$; for experiments 7-9, $\sigma^2 = 5$, $\tau^2 = 5$; for experiments 10-12, $\sigma^2 = 7$, $\tau^2 = 3$; for experiments 13-15, $\sigma^2 = 9$, $\tau^2 = 1$. E-M, expectation and maximization algorithm; SCOR, scoring method; N-R, Newton-Raphson method.

In the second set of simulations, we compared scoring under two different models: one when the quantitative trait locus (QTL) had 2 alleles and the marker locus had 4 equally probable alleles, and the other when the QTL and the marker locus both had 4 alleles. The results are shown in Table II. The accuracy of all variance components tended to improve with increasing major gene and polygenic heritability. Under both models, the scoring method estimates were similar to the true value.

Lastly, we compared our results with those of quasilielihood analysis. We assumed two QTL, each one with 2 alleles. We further assumed that one QTL was linked to a marker locus and the other QTL was not linked. We expected that the

TABLE II. Median Values of 1,000 Simulations From 100 Nuclear Families (2 Parents and 4 Offspring) Using Scoring Method (Values in Parentheses Are the Median Squared Error)*

Exp	True value		2 alleles			4 alleles		
	h^2	h_g^2	σ^2	σ_g^2	τ^2	σ^2	σ_g^2	τ^2
1	0.1	0.3	0.9189 (1.000)	2.949 (0.5711)	5.967 (0.1988)	0.9060 (1.000)	2.909 (0.5449)	5.962 (0.1895)
2	0.3	0.3	2.884 (0.9196)	2.994 (0.6695)	4.006 (0.1764)	3.005 (1.018)	2.988 (0.6991)	3.984 (0.1573)
3	0.5	0.3	4.980 (0.7590)	2.909 (0.4956)	2.017 (0.1002)	5.033 (0.7877)	2.927 (0.5117)	1.998 (0.0988)
4	0.1	0.5	0.9574 (0.9304)	4.918 (0.4753)	3.970 (0.1432)	1.005 (1.000)	4.946 (0.4800)	3.953 (0.1363)
5	0.3	0.5	3.004 (0.6743)	4.915 (0.4347)	2.010 (0.0879)	3.050 (0.7234)	4.915 (0.4747)	1.991 (0.0885)

*Polygenic heritability (h^2), major gene heritability (h_g^2), polygenic VC (σ^2), major gene VC (σ_g^2), error VC (τ^2). For experiment 1, $\sigma^2 = 1$, $\sigma_g^2 = 3$, $\tau^2 = 6$; for experiment 2, $\sigma^2 = 3$, $\sigma_g^2 = 1$, $\sigma_g^2 = 3$, $\tau^2 = 4$; for experiment 3; $\sigma^2 = 5$, $\sigma_g^2 = 3$, $\tau^2 = 2$; for experiment 4; $\sigma^2 = 1$; $\sigma_g^2 = 5$, $\tau^2 = 4$; for experiment 5, $\sigma^2 = 3$, $\sigma_g^2 = 5$, $\tau^2 = 2$.

variance of the unlinked QTL locus would be modeled by the polygenic VC without error, and that is what we observed. We concluded that the results (Table III) were similar, except that the scoring method gave us smaller median squared error when heritability (h^2) was low. The QL method was extremely computationally intensive, requiring about 32 times longer than the scoring method.

Data from the Bogalusa Heart Study

As a part of the Bogalusa Heart study, a large family was selected based on a proband with a heart murmur and a self-reported history of prevalent heart disease on both sides of the family. The data included samples from 196 individuals representing six generations. High-density lipoprotein cholesterol (HDL-C) and serum apolipoprotein B (ApoB) levels were quantitated following standardized protocols [Rosenbaum et al., 1986; Heiba et al., 1993]. The ApoB polymorphism was evaluated using Southern blot analysis of an XbaI site (Heiba et al., 1993; Ma et al., 1987); data from 60 sibling pairs were available. The HDL-C and ApoB levels were log-transformed before the analysis. Following the analyses reported here and in previous publications [Laing et al., 1994; Amos et al., 1987], effects of age, age squared, sex, oral contraceptive use, alcohol consumption, and smoking behavior were statistically removed by a regression analysis and obtaining the residuals. The residuals were subsequently standardized to have unit variance. We previously identified evidence by segregation analysis, for a major gene effect on ApoB levels in this pedigree, with 42% of the variance attributed to a putative major locus. Using model-dependent methods of linkage analysis [Ott, 1991], we also found evidence of linkage between quantitative levels of serum ApoB and the APOB structural gene [Laing et al., 1994], but we did not find statistically significant evidence concerning ApoB levels using the Haseman-Elston method [Heiba et al., 1993]. For the current analysis, we first divided the extended family into nuclear families and then used either (1) the VC method assuming multivariate normality (scoring), or (2) the VC approach and QL. For the ML method, we used the Fisher information matrix to obtain standard errors, while for the QL method, we used a robust estimate of the variance for the estimated parameters [Amos et al., 1996]. Because the family was

TABLE III. Median Values of 1,000 Simulations From 100 Nuclear Families (2 Parents and 4 Offspring) Using Scoring and Quasilikelihood (QL) Methods (Values in Parentheses Are the Median Squared Error)*

	True value		Scoring				QL			
	h^2	h_g^2	σ^2	σ_g^2	τ^2	CPU	σ^2	σ_g^2	τ^2	CPU
1	0.1	0.3	0.919 (1.000)	2.949 (0.5711)	5.967 (0.1988)	24.8	0.919 (1.00)	3.006 (0.7518)	5.907 (0.1915)	413.8
2	0.3	0.3	2.884 (0.9196)	2.994 (0.6695)	4.006 (0.1764)	13.9	2.884 (0.9195)	2.997 (0.6696)	4.005 (0.1747)	677.3
3	0.5	0.3	4.980 (0.7590)	2.909 (0.4956)	2.017 (0.1002)	12.5	4.980 (0.7589)	2.909 (0.4956)	2.017 (0.1002)	678.7
4	0.1	0.7	0.964 (0.5205)	6.919 (0.2888)	1.989 (0.0767)	22.4	0.975 (0.5055)	6.950 (0.3288)	1.981 (0.0723)	610.4

*CPU in seconds per run on average on a SUN Sparcstation 20. Polygenic VC (σ^2), major gene VC (σ_g^2), error VC (τ^2). For experiment 1, $\sigma^2 = 1$, $\sigma_g^2 = 3$, $\tau^2 = 6$; for experiment 2, $\sigma^2 = 3$, $\sigma_g^2 = 3$, $\tau^2 = 4$; for experiment 3, $\sigma^2 = 5$, $\sigma_g^2 = 3$, $\tau^2 = 2$; for experiment 4, $\sigma^2 = 1$, $\sigma_g^2 = 7$, $\tau^2 = 2$.

ascertained on the basis of only one proband, we felt the effect of ascertainment correction would have minimal effect, and therefore, we did not make an ascertainment correction [Majumder, 1985].

Table IV shows estimates based on the analyses by either the ML or QL methods. For this pedigree we find that 69% of the variance was attributed to a major gene when the ML method was used, and 73% of the variance was attributed to a major gene when the QL method was used. Similarly, neither method provided evidence of a major genetic effect by APOB on HDL, although the QL estimates were considerably higher than the total variance of HDL should have been (it was standardized to unit variance before the analysis). Although ML provided more accurate results than QL, it is important to note the very small sample size available for study, with only 60 total sib pairs. For a larger sample size, we would expect QL to provide more accurate results than it did. Previously, Amos et al. [1996] found the ML method to be more accurate for moderately small sample sizes than QL. In addition, application of a robust variance estimate for the QL approach may have led to larger standard errors than those obtained with the scoring method and subsequent inversion of the information matrix.

DISCUSSION

The use of these iterative methods to estimate genetic components of variance has been reported by several groups [Thompson and Shaw, 1990; Henderson, 1986; Lange and Boehnke, 1983; Hopper and Mathews, 1982]. Xu and Atchley [1995] reparameterized the genetic components of variance in terms of heritabilities. However, methods of estimation have not been compared in the literature except by Jennrich and Schluchter [1986], who gave us a statistical overview of these methods but no specific application to quantitative traits.

We compared iterative methods and showed their utility in estimating the genetic variance components. For the analyses presented here, we assumed that the likelihood model postulates a normal distribution of the quantitative trait. The segregation of the major gene violates this assumption, however. Our simulation studies

TABLE IV. Estimated Values of Variance Components Parameters From the Bogalusa Heart Study Using Scoring and QL Methods, With Serum ApoB Levels and High-Density Lipoprotein Cholesterol (HDL) as Quantitative Traits and the APOB Gene as the Major Gene (Values in Parentheses Are the Standard Errors)*

Traits	Method	σ^2	σ_g^2	τ^2
ApoB	Scoring	0.0000 ^a	0.7558 (0.1658)	0.3383 (0.1111)
	QL	0.0000 (0.4231)	0.9449 (0.4262)	0.3488 (0.1078)
HDL	Scoring	0.6752 (0.2229)	0.0000 ^a	0.7887 (0.1861)
	QL	1.2764 (0.7918)	0.0000 (0.7999)	0.7221 (0.2151)

*Polygenic VC (σ^2), major gene VC (σ_g^2), error VC (τ^2).

^aThe estimated value is fixed at the boundary. No standard error value is provided. Elsewhere, standard error is in parentheses.

showed that ML methods of estimating these genetic components of variance are consistent and slightly more efficient than QL methods. We attempted to use E-M methods that included a major gene component but encountered difficulties. First we attempted to apply E-M methods directly but obtained singular matrices for common IBD relationships. We also attempted a two-step procedure in which the polygenic components were first fitted by the E-M algorithm and then removed, and the residual was analyzed by scoring. However, this method underestimated the major gene component (further results available by request). A possible remedy for this problem would be to cycle between the two steps: as the major gene VC is estimated, new residuals are obtained by subtracting off the major gene effect and the new update is then put back into step 1. When we analyzed ApoB and HDL levels in a large pedigree, results of QL and ML methods were similar, but estimates based on the ML method were more precise.

Generally, compared with the other methods, we found scoring to perform well in terms of computational requirements. The computational efficiency of this method is balanced, however, by greater difficulty in providing the information matrix analytically. N-R estimation and scoring provided similar results, and the N-R method may be preferred in situations in which complex genetic models need to be fitted. When computational speed is not a major factor in choosing an algorithm, the N-R method is most appealing because it is easier to implement using available numerical optimization routines. For simulation studies, the N-R approach is excessively computationally intensive.

We have also extended the scoring method for multivariate traits to simultaneously adjust for covariates and to permit the inclusion of an ascertainment correction [de Andrade et al., 1997]. In addition, we included a common environmental effect, and can allow for effects from multiple unlinked loci. In future work, gene-covariate interactions and methods to handle classes of distributions from the exponential family will be developed. All scientific code used for this manuscript is available on request to mandrade@request.mdacc.tmc.edu and our website <http://request.mdacc.tmc.edu/> contains software for most of this work.

ACKNOWLEDGMENTS

We thank our computer programmers Liping Yu and Dakai Zhu for their help. We also thank Dr. Duncan Thomas and an anonymous reviewer for their valuable comments.

REFERENCES

- Amos CI. 1988. Robust methods for detection of genetic linkage for data from extended families and pedigrees. Ph.D. dissertation, Louisiana State University, New Orleans.
- Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 54:535-543.
- Amos CI, Elston RC, Srinivasan SR, Wilson AF, Cresanta JL, Ward LJ, Berenson GS. 1987. Linkage and segregation analyses of apolipoproteins AII and B, and lipoprotein cholesterol levels in a large pedigree with excess coronary disease. The Bogalusa Heart Study. *Genet Epidemiol* 4:115-128.
- Amos CI, Elston RC, Bonney GE, Keats BJB, Berenson GS. 1990. A multivariate method for detecting

- genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. *Am J Hum Genet* 47:247–254.
- Amos CI, Zhu D, Boerwinkle E. 1996. Assessing genetic linkage and association with robust components of variance approaches. *Ann Hum Genet* 60:142–160.
- Blackwelder WC, Elston RC. 1985. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol* 2:85–97.
- de Andrade M, Thiel TJ, Yu L, Amos CI. 1997. Assessing linkage in chromosome 5 using components of variance approach: univariate versus multivariate. *Genet Epidemiol* 14:773–778.
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *JRSS B* 39:1–38.
- Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R S Edinburgh* 52:399–433.
- Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2:3–19.
- Henderson CR. 1986. Recent developments in variance and covariance estimation. *J Anim Sci* 63:208–216.
- Heiba IM, DeMeester CA, Xia YR, Diep A, George VT, Amos CI, Srinivasan SR, Berenson GS, Elston RC, Lusia AJ. 1993. Genetic contributions to quantitative lipoprotein traits associated with coronary artery disease: analysis of a large pedigree from the Bogalusa Heart Study. *Am J Med Genet* 47:875–883.
- Hopper JL, Mathews, JD. 1982. Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* 46:373–383.
- Jennrich RI, Schluchter MD. 1986. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 42:802–820.
- Laing AE, Amos CI, DeMeester C, Diep A, Xia YR, Elston RC, Srinivasan SR, Berenson GS, Lusia AJ. 1994. Linkage between the APOB gene and serum ApoB levels in a large pedigree from the Bogalusa Heart Study. *Genet Epidemiol* 11:29–40.
- Laird NM. 1982. Computation of variance components using the EM algorithm. *J Stat Comp Simulation* 14:295–303.
- Lange K, Boehnke M. 1983. Extensions to pedigree analysis. IV. Covariance components for multivariate traits. *Am J Med Genet* 14:513–524.
- Lange K, Westlake J, Spence MA. 1976. Extensions to pedigree analysis. III. Variance components by the scoring method. *Ann Hum Genet* 39:485–491.
- Louis AL. 1982. Finding the observed information matrix when using the EM algorithm. *JRSS B* 44:226–233.
- Ma Y, Schumaker VN, Butler R, Sparkes RS. 1987. Two DNA restriction fragment length polymorphisms associated with Ag (t/z) and Ag (g/c) antigenic sites of human apolipoprotein B. *Arteriosclerosis* 7:301–305.
- Majumder PP. 1985. Comparison of ascertainment-bias correction schemes for pedigrees ascertained through multiple probands. *Stat Med* 4:163–173.
- Meng X-L, Rubin DB. 1991. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J Am Stat Assoc* 86:899–909.
- Morton NE. 1974. Analysis of family resemblance. I. Introduction. *Am J Hum Genet* 26:318–330.
- Morton NE, Miki C, Yee S. 1968. Bioassay of population structure under isolation by distance. *Am J Hum Genet* 20:411–419.
- Ott J. 1991. Analysis of human genetic linkage. Baltimore: The Johns Hopkins University Press.
- Pugh EW, Jaquish CE, Sorant AJM, Doetsch JP, Bailey-Wilson JE, Wilson AF. 1997. Comparison of sib-pair and variance components methods for genomic screening. *Genet Epidemiol* 14:867–872.
- Rao DC, Morton NE, Yee S. 1974. Analysis of family resemblance. II. A linear model for familial correlation. *Am J Hum Genet* 26:331–359.
- Rosenbaum PA, Amos CI, Shear CL, Elston RC, Sellers TA, Srinivasan SR, Berenson GS. 1986. Description of a large pedigree with an adverse lipoprotein cholesterol phenotype. The Bogalusa Heart Study. *Genet Epidemiol* 3:241–254.
- Schork NJ. 1991. Efficient computation of patterned covariance matrix mixed models in quantitative segregation analysis. *Genet Epidemiol* 8:29–46.

- Searle SR. 1982. Matrix algebra useful for statistics. New York: John Wiley & Sons.
- Searle SR, Casella G, McCulloch CE. 1992. Variance Components. New York: John Wiley & Sons.
- Sorant AJM, Elston RC. 1994. A subroutine package for function maximization (a user's guide to MAXFUN version 6.0). Part of the S.A.G.E. documentation, Department of Biometry and Genetics, Louisiana State University Medical Center, New Orleans.
- Thompson EA, Shaw RG. 1990. Pedigree analysis for quantitative traits: Variance components without matrix inversion. *Biometrics* 46:399–413.
- Wright FA. 1997. The phenotypic difference discards sib-pair QTL linkage information. *Am J Hum Genet* 60:740–742.
- Xu S, Atchley WR. 1995. A random model approach to interval mapping of quantitative trait loci. *Genetics* 141:1189–1197.