Genetic Epidemiology 9:73–86 (1992)

Extended Pedigree Patterned Covariance Matrix Mixed Models for Quantitative Phenotype Analysis

Nicholas J. Schork

Division of Hypertension, Department of Medicine and Department of Epidemiology, University of Michigan, Ann Arbor, Michigan

Overt computational constraints in the formation of mixed models for the analysis of large extended-pedigree quantitative trait data which allow one to reliably characterize and partition sources of variation resulting from a variety sources have proven difficult to overcome. The present paper suggests that by combining a restricted patterned covariance matrix approach to modeling and partitioning the variation arising from polygenic and environmental forces with an Elston–Stewart like algorithmic approach to modeling variation resulting from a single genetic locus with large phenotypic effects one can produce a model that is at once intuitively appealing, efficient computationally, and reliable numerically. Extensions and variations of this approach are also discussed, as are some simulation and timing studies carried out in an effort to validate the accuracy and computational efficiency of the proposed methodology. © 1992 Wiley-Liss, Inc.

Key words: mixed models, variance components, Elston–Stewart algorithm, quantitative traits

INTRODUCTION

Despite the plethora of molecular genetic advances made in recent years, human geneticists are still faced with a number of problems which have proven difficult to overcome. Most of these problems derive from the obvious fact that breeding experiments and experimental genetic interventions involving humans are, for the most part, unethical. Human geneticists must therefore rely heavily on statistical approaches to modeling and exploring the natural variation occurring within and among human populations. Because the complexity of the genetic phenomena currently understood through laboratory-based methods has grown dramatically, a gap between basic genetic theory and implementable, statistical model-based tools for studying and assessing human

Received for publication July 1, 1991; revision accepted January 28, 1992.

Address reprint requests to Nicholas J. Schork, Department of Medicine and Department of Epidemiology, University of Michigan, R6592 Kresge I, Ann Arbor, MI 48109-0500.

variation has arisen, as noted by Ott [1990]. This is especially true in the case of quantitative phenotype analysis: theoretically sound methods have been devised for dissecting and exploring the basis of quantitative inheritance in plant and animal populations, and have, in fact, been implemented in studies on these populations with great success, but have not proven to be useful or applicable in human pedigree data contexts. For example, the elegant work of Lander et al. [1989] describes methods for the analysis of quantitative *plant* traits that are theoretically sound, but rely almost entirely on backcross and intercross information.

One particularly troublesome problem in human pedigree quantitative phenotype analysis centers around the development of realistic and efficient models incorporating both the effects of a single locus with large phenotypic effects and polygenic and complex environmental factors—the so called "mixed model"— for moderate to large extended pedigree (i.e., nonnuclear family) data. In their pioneer work, Elston and Stewart [1971] described methods whereby one could ease the computational burden of many pedigree models and suggested that the implementation of a true mixed model would be difficult at best. This suggestion was later elaborated by Boyle and Elston [1979]. Morton and MacLean [1974] devised an implementable mixed model but relied on a numerical method for the relevant function evaluation that made their approach approximate, computationally burdensome, and unflexible (see also Elston [1981]). Ott [1979] incorporated the patterned covariance matrix model first discussed by Lange et al. [1976] to model polygenic variation in a mixed model setting. Though the resulting model's likelihood function was exact, its utility is limited to small pedigrees because its computational load increases exponentially with the pedigree size. In two important papers, Hasstedt [1982, 1991] derived an approximation to certain mixed models which is computationally feasible, but whose current implementations do not allow easy inclusion of arbitrary covariance components, has not been well studied (e.g., through large scale simulation studies) on large extended pedigree data, and relies on a method for approximating polygenic parameters which may produce different results depending on the order in which extended pedigree members are incorporated into the relevant likelihood function evaluation "peeling" procedure. On another plane, Bonney [1984] elaborated a number of computationally feasible and mathematically elegant models for exploring the variation of many genetic phenomena. However, many of these models have formulations for characterizing residual (i.e., nonmajor locus) variation which do not exploit the sound principles of genetic transmission, and hence do not permit the partition of the residual variance into genetic and nongenetic components.

In fairness to the approaches developed by Hasstedt and Bonney, it is important to note that the published, large-scale studies gauging the reliability of each method have relied primarily on nuclear family data [Konigsberg et al., 1989; Demenais and Bonney, 1989; Demenais et al., 1990]. As such, it would be premature to suggest that the sources of covariation sacrificed in each approach, as well as the lack of residual variance partitioning the case of the regressive models, have a dramatic negative effect. In addition, recent simulation work by John Blangero and colleagues suggests that arbitrary covariance components may be implemented reliably within the framework of Hasstedt's later model [Hasstedt, 1991; J. Blangero, personal communication]. Other papers discussing some relevant aspects of mixed model formulation and computation are Bonney [1982] and Lalouel et al. [1983].

In what follows, a formulation of the mixed model is outlined which can be con-

sidered a combination of the approach of Ott and the fundamental algorithm described by Elston and Stewart [1971], which was later extended by Lange and Elston [1975] and Cannings et al. [1978]. The primary advantages of the proposed model are that it allows for the arbitrary partition of the residual variance in extended pedigree contexts, will produce the same likelihood value irrespective of the order in which the relevant parts of the pedigree are incorporated into the likelihood function evaluation, and is compatible with the efficient computational strategies outlined in Schork [1991a] and Schork [1991b]. As an aside, this paper also presents a simple framework which may allow one to assess the effects of sacrificing sources of covariation in mixed model likelihood evaluation schemes. As such, this framework could be used in computational experiments with both regressive models and approaches like Hasstedt's which, like the proposed method, do not use the full covariance structure of extended pedigrees in their constructions.

METHODS

As with all mixed models, there are two basic components to the proposed mixed model: a component characterizing the effects of the major locus and a component characterizing the polygenic and environmental "background" effects. Each component will be briefly described, though their combination—the crucial aspect of the proposed method—will be considered in greater detail.

Consider the case of a single locus with 2 alleles, A and a, which works to form 3 genotypes AA, Aa, and aa. Each genotype, g, has an associated frequency f_g, $g \in \{AA, Aa, aa\}$, and mean effect μ_g. Associated with all genotypes is a common variance σ^2. The frequencies, f_g, are, for pedigree members whose parents are not in the pedigree, functions of the allele frequencies [e.g., $f_{AA} = p^2, f_{Aa} = 2p(1 - p), f_{aa} = (1 - p)^2$, where p is the frequency of the A allele and Hardy–Weinberg equilibrium is assumed], and are dictated by transmission probabilities consistent with Mendelian theory for those pedigree members whose parents are in the pedigree [Elston, 1981]. A likelihood based model assuming a pedigree with n members then involves consideration of all possible genotypes arrangements for the pedigree members. A likelihood function for such a model can be written as

$$\begin{aligned} L_n(\theta|X) &= \sum_{g_1}\sum_{g_2}\cdots\sum_{g_n} f_{g_1} f_{g_2} \cdots f_{g_n} \\ &\cdot \varnothing(\varkappa_1|\mu_{g_1},\sigma^2)\phi(\varkappa_2|\mu_{g_2},\sigma^2) \cdots \phi(\varkappa_n|\mu_{g_n},\sigma^2) \end{aligned} \tag{1}$$

where θ denotes the parameters p, μ_{AA}, μ_{Aa}, μ_{aa}, σ^2, X denotes the trait values, $\varkappa_1, \ldots, \varkappa_n$, of the n pedigree members, g_i refers to the possible genotypes associated with pedigree member i, f_g is the appropriate allele frequency or transmission probability associated with genotype g_i, and $\varnothing$ is the "penetrance probability," or the probability that i shows $\varkappa_i$ given genotype g_i. Note that $\varnothing$ is typically taken to be the normal density function with mean μ and variance σ^2.

Elston and Stewart [1971], Lange and Elston [1975], Ott [1974], and Cannings et al. [1978] all elaborate and expand on a method whereby the multiple sum given in Eq. (1) can be written as an iterated sum. Since these authors characterize their methods differently and use a terminology that is consistent with subtleties unique to these different characterizations, the characterization of the Elston–Stewart algorithm offered in this paper will be based on the notion of a "partition-set." Basically, "partition-

sets" are comprised of subsets of closely related pedigree members (e.g., nuclear families) within the pedigree. One member in each partition-set who is common to another partition-set, termed the "pivot," is then chosen. The likelihood of the other partition-set members is then computed conditionally on the genotypes of this pivotal member. These conditional likelihoods are then saved and are incorporated into calculations involving other partition-sets of which the pivot is a member. The order in which the partition-sets are dealt with is dictated by the dependence of the conditional likelihood calculations of each partition-set on the other partition sets. To illustrate, consider the pedigree depicted in Figure 1a. The nuclear families defined by $nc_1 = \{1,2,4,5\}$, $nc_2 = \{3,4,7,8\}$, and $nc_3 = \{5,6,9,10\}$ define partition-sets; persons 4 and 5 are pivotal. Starting with nc_2, the genotypes of member 4 are fixed and the three conditional likelihoods, $L_{nc_2}(\theta|x_3,x_7,x_8,g_4 \in \{AA, Aa, aa\})$, are computed and saved. The same is done for nc_3 using member 5 to condition on. These conditional likelihoods are then used to "weight" the genotype assignments for members 4 and 5 when the likelihood of nc_1 is computed. If the conditional likelihoods of relevant pedigree partition-sets are incorporated in the likelihood function evaluation in this iterative fashion, then the likelihood evaluation involving the entire pedigree will be exact. The basic mechanism behind this algorithm has been extended to work with complex pedigrees and pedigrees with loops (see Lange and Elston [1975] and Cannings et al. [1978]).

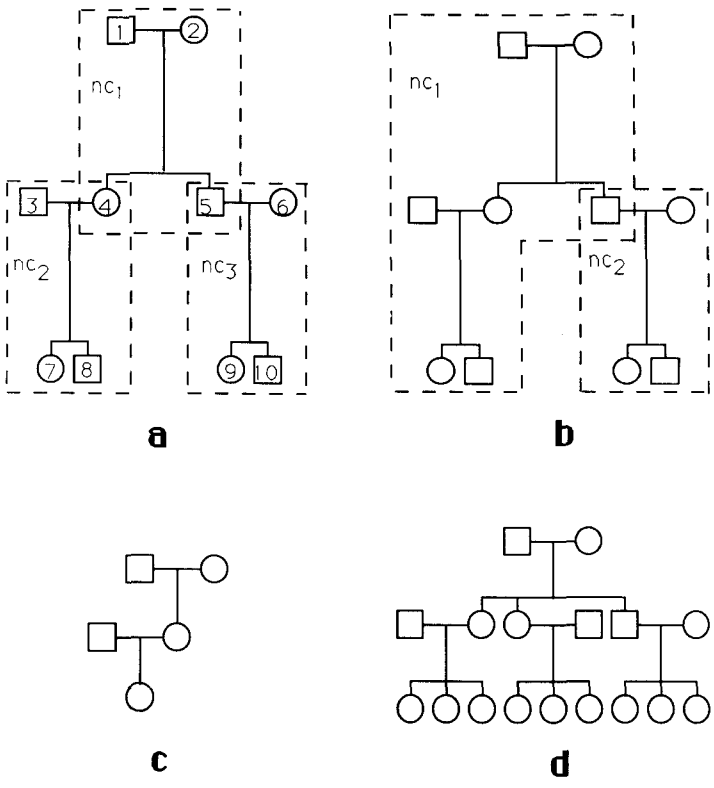

Fig. 1. (**a,b**) Two different "partition-set" partitions for a 10-member, o-parameter of 2, pedigree. (**c,d**) Pedigrees structured with an o-parameter of 1 and 3, respectively.

To model polygenic and environmental factors, the patterned covariance model of Lange et al. [1976] can be adopted. If it is assumed that the variation in the pedigree trait values is controlled by the additive and dominance effects of polygenes, shared household factors, and a random environmental effect, then the covariation among the n pedigree members can be modeled with an $n \times n$ matrix defined by

$$\Omega = 2A\sigma_a^2 + D\sigma_d^2 + H\sigma_h^2 + I\sigma_e^2 \tag{2}$$

where σ_a^2, σ_d^2, σ_h^2, and σ_e^2 are the additive variance, dominance variance, shared household variance, and random environmental variances, respectively. The other terms in Eq. (2) are $n \times n$ matrices that relate the variance terms to the relevant pedigree member pairs. Thus, A is the kinship coefficient matrix, D is Jacquard's delta-7 matrix [Jacquard, 1974], H is a matrix such that its ijth component is 1 if i and j share a household and 0 if they do not, and I is the identity matrix. If a mean, μ, is associated with the trait in question and one assumes multivariate normality of the trait values among the pedigree members [Lange, 1978], then one can compute estimates of μ and the variance components in Ω by maximizing the relevant multivariate normal likelihood function as discussed by Lange et al. [1976].

The mixed model of Ott [1979] is obtained by evaluating a mixture of multivariate normal distributions, where each component in the mixture is associated with a particular genotype arrangement among the pedigree members. The relevant likelihood function is given by

$$L_n(\theta|X) = \sum_{k=1}^{G} F_k \cdot \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} exp[-\tfrac{1}{2}(\varkappa - \mu_k)'\Omega^{-1}(\varkappa - \mu_k)] \tag{3}$$

where $\theta = (p, \mu_{AA}, \mu_{Aa}, \mu_{aa}, \sigma_a^2, \sigma_d^2, \sigma_h^2, \sigma_e^2)$, X is a vector of length n holding the pedigree member trait values, the sum is over all G possible genotype arrangements for the n pedigree members, F_k is the product of all relevant f_g for the kth genotype arrangement, Ω is given in Eq. (2), and μ_k is a vector of a mean genotype effects consistent with the genotype arrangement k. One should consult Ott [1979] or Schork [1991] for further details. As intimated in the introduction, because the sum in Eq. (3) grows exponentially with the size of the pedigree, the model in (3) is impractical for all but the smallest of pedigrees.

The mixed model advocated in this paper involves the use of the Elston–Stewart approach (or, more specifically, the Lange–Elston extension of the Elston–Stewart algorithm) to evaluating the major locus component of the model, but uses the model offered in Eq. (3) to model the variation *within each partition-set*. That is, for each partition-set, nc_i, with m_i members, one computes the probability of the $m_i - 1$ nonpivotal members conditionally on the genotypes of the pivotal member ℓ, which can be written in the form

$$p_{nc_i}(\varkappa^1, \ldots, \varkappa_{m_i}|\theta, g_\ell) = \frac{p(\varkappa_1, \ldots, \varkappa_{m_i} \cap \theta g_\ell)}{p(\varkappa_\ell|\theta, g_\ell)} \tag{4a}$$

where $g_\ell \in \{AA, Aa, aa\}$. This equation can be written in a manner consistent with Eqs. (1) and (3) to produce a likelihood equation which forms the basis for the type of

equation which is iterated in the Elston–Stewart algorithm implementation used in the proposed model:

$$L_{nc_i}(\theta|g_\ell) = \sum_{k^*=1}^{G^*} F_{k^*} \frac{1}{(2\pi)^{n/2}|\Omega^*|^{1/2}} exp[-1/2(\varkappa - \mu_{k^*})'\Omega^{*-1}(\varkappa - \mu_{k^*})]/(\phi(\varkappa_\ell|\mu_{g_\ell},\sigma_t^2), f_{g\ell}) \quad (4b)$$

where nc_i denotes the ith partition-set, θ is given as in Eq. (3), the sum is only over those genotype arrangements involving the relevant m_i partition-set members, F_{k^*} is the product of the relevant f_g, for the members in nc_i, $\varkappa$ is the vector of trait values associated with the members of nc_i, μ^* is a vector of the mean genotype effects consistent with the k^*th genotype arrangement, Ω is an $m_i \times m_i$ covariance matrix of the type assumed in Eq. (2), but whose terms are restricted to and dictated by the m_i partition-set members' relationships, and where σ_t^2 is the total variation $\sigma_t^2 = \sigma_a^2 + \sigma_d^2 + \sigma_h^2 + \sigma_e^2$. As such, the combination and joint derivation of the various conditional likelihoods follow the scheme outlined in Lange and Elston [1975].

The mixed model scheme just outlined does afford a great deal of flexibility. Partition-sets can be chosen in a variety of ways; the limiting factor being the amount of computation assumed in modeling the variation in the partition-sets through Eq. (4). For instance, Figure 1b depicts alternative partition-sets to those depicted in Figure 1a. Complex pedigrees and pedigrees with loops can still be dealt with in the manner described in Lange and Elston [1975]. Further variance components modeling a variety of genetic and environmental phenomena can be easily added by extending Eq. (2). In addition, covariates, y, can be added to model by adapting the mean vectors appearing in the quadratic form of Eqs. (3) and (4) to, say

$$[\varkappa - \alpha_\mu(y)]'\Omega^{-1}[\varkappa - \alpha_\mu(y)] \quad (5)$$

where $\alpha_\mu(y)$ is a function (e.g., regression, growth curve, etc.) mapping the covariates, y, to the relevant mean genotype effects μ which are dictated by the appropriate genotype arrangement under consideration.

Obviously, the drawback to the proposed scheme is that it sacrifices covariance terms between members of the pedigree by allowing parameterization only of the covariation between members in a partition-set. This may not be a serious drawback, however, given the structure of Eq. (2), since sibs contribute most of the information in the identification of dominance polygenic effects, and the kinship coefficient decreases exponentially as members become separated by either more generations or greater "horizontal" distancing (e.g., cousins to second cousins to third cousins, etc.). A related drawback is that the proposed method can deal only with matrilineal or other long-term effects by using large partition-sets—a strategy that may be too challenging computationally to be useful in certain situations (e.g., in settings in which the relevant partition-sets have extremely large sibships).

SIMULATION STUDIES

Monte Carlo Studies

In order to investigate the reliability of the proposed mixed model, a large simulation study was employed. Because an overriding concern associated with the pro-

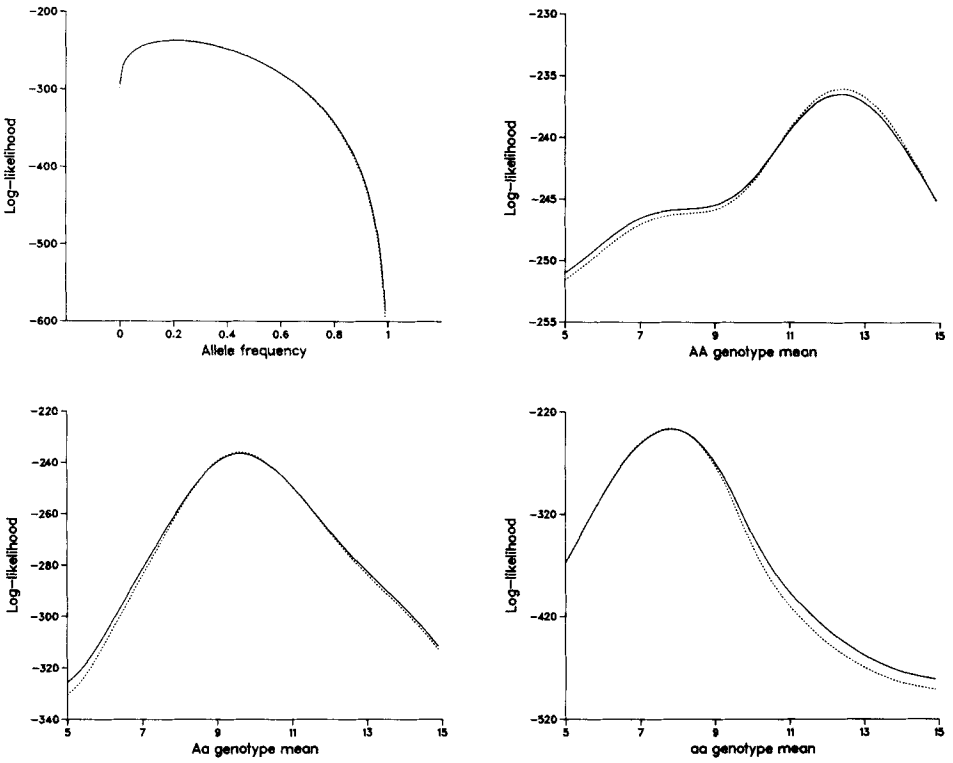

Fig. 2. Profile likelihoods for an exact and approximate mixed model for an allele frequency and mean genotype parameters assuming 25 pedigrees with 10 members each following a segregation pattern with parameters $p = 0.25$, $\mu_{AA} = 12.0$, $\mu_{Aa} = 10.0$, $\mu_{aa} = 8.0$, $\sigma_a^2 = 0.5$, $\sigma_d^2 = 0.5$, and $\sigma_e^2 = 1.0$. Note that the dashed line represents the approximation and the solid line the exact method. Close agreement causes the lines to overlap. A value of 230 was added to likelihoods obtained with the approximate model.

TABLE I. Characteristics of the Pedigree Structures Investigated in a Simulation Study of the Proposed Mixed Model*

o	*nuc*	*mem*	*tp*	*tz*	*tm*	*frac*	$n = 250$
1	2	5	15	2	11	0.85	50 (250)
2	3	10	55	8	28	0.60	25 (250)
3	4	17	153	24	58	0.45	15 (255)
4	5	26	351	56	103	0.35	10 (260)
5	6	37	703	110	166	0.28	7 (259)
6	7	50	1275	192	250	0.23	5 (250)

*o is the offspring parameter; *nuc* the number of nuclear families in the pedigree; *mem* the number of members in the pedigree; *tp*, *tz*, *tm* are defined by Eqs. (6), (7), and (8), respectively; *frac* is the fraction of the total pairs of pedigree members that could potentially assume nonzero values in the covariance matrix built from these pairs used in the proposed model; $n = 250$ is the number of pedigrees needed to produce at least 250 subjects (actual number of subjects given $n = 250$ is in parentheses).

TABLE II. Results of the Simulation Study Using 100 Independent Replications of the Pedigree Types and Numbers Given in Table I for Two Different Segregation Parameters Settings*

	o					
	1	2	3	4	5	6
p	0.27 (0.01)	0.30 (0.01)	0.27 (0.01)	0.27 (0.01)	0.27 (0.01)	0.27 (0.01)
$m\text{-}_{AA}$	11.9 (0.11)	11.7 (0.09)	11.9 (0.10)	12.0 (0.12)	11.8 (0.10)	11.9 (0.10)
$m\text{-}_{Aa}$	9.99 (0.07)	9.88 (0.07)	9.96 (0.06)	9.97 (0.07)	9.89 (0.08)	9.98 (0.08)
$m\text{-}_{aa}$	7.94 (0.05)	7.84 (0.06)	7.91 (0.04)	7.95 (0.05)	8.00 (0.05)	7.97 (0.05)
$v\text{-}_{a}$	1.00 (0.08)	0.88 (0.06)	0.79 (0.06)	0.89 (0.07)	1.03 (0.08)	0.78 (0.06)
$v\text{-}_{d}$	0.18 (0.01)	0.28 (0.04)	0.16 (0.03)	0.25 (0.04)	0.19 (0.04)	0.23 (0.04)
$v\text{-}_{e}$	0.71 (0.03)	0.67 (0.04)	0.92 (0.04)	0.75 (0.05)	0.79 (0.05)	0.85 (0.05)
p	0.28 (0.01)	0.27 (0.01)	0.25 (0.01)	0.28 (0.01)	0.25 (0.01)	0.25 (0.01)
$m\text{-}_{AA}$	11.8 (0.09)	11.9 (0.10)	11.9 (0.10)	11.7 (0.11)	11.8 (0.08)	12.0 (0.12)
$m\text{-}_{Aa}$	9.95 (0.06)	9.95 (0.07)	10.1 (0.08)	9.93 (0.08)	9.94 (0.06)	10.1 (0.07)
$m\text{-}_{aa}$	7.92 (0.04)	7.91 (0.04)	7.95 (0.04)	7.91 (0.04)	7.98 (0.04)	8.09 (0.05)
$v\text{-}_{a}$	0.54 (0.06)	0.39 (0.04)	0.47 (0.05)	0.38 (0.04)	0.50 (0.05)	0.43 (0.05)
$v\text{-}_{d}$	0.51 (0.08)	0.67 (0.07)	0.62 (0.06)	0.49 (0.06)	0.48 (0.05)	0.59 (0.06)
$v\text{-}_{e}$	0.84 (0.03)	0.77 (0.06)	0.83 (0.06)	1.04 (0.06)	0.97 (0.06)	0.94 (0.06)

*Results shown are the mean and standard deviation of the parameter estimates gleaned from the 100 replications. The upper panel reports the results from the first parameter settings, the lower panel the second (see text). p is the frequency of the A allele: m_{AA}, $m\text{-}_{Aa}$, $m\text{-}_{aa}$ are the mean genotype effects; $v\text{-}_{a}$, $v\text{-}_{d}$, and $v\text{-}_{e}$ are the additive, dominance, and environmental variance, respectively.

posed method is its sacrificing of covariance terms, a special type of pedigree was investigated in the simulation study. This pedigree type has three generations, but its size is dictated by a "number of offspring" parameter o. This parameter gives the number of offspring in the second and third generations for each set of parents in the first and second generations. Figure 1a and b depicts a pedigree of this type with $o = 2$. Figure 1a and b depicts pedigrees with $o = 1$ and $o = 3$, respectively. Such pedigrees allow easy characterization of their covariance and related terms. The number of nuclear families in such pedigrees is simply $o + 1$; the number of pedigree members, m, is simply $m = 2 + 2o + o^2$. The number of possible pedigree member "pairs," tp, whose covariation *could* be parameterized is, because of symmetry

$$tp = m \cdot (m + 1)/2. \tag{6}$$

Because only additive variance, dominance variance, and random environmental variance terms were assumed in the models studied, many of the tp covariance terms in Eq. (6) would naturally be 0 (e.g., excluding inbreeding, spouses should not share genes, neither should inlaws, etc.). The number of such terms, tz, given a pedigree of the type discussed above with o offspring parameter is

$$tz = 2 \cdot o + o^2(o - 1). \tag{7}$$

Since the partition-sets used in the simulation study were defined by the $o + 1$ nuclear families, the number of pedigree members pairs, tm, whose covariation is parameterized in the proposed mixed model is given by

$$tm = (o + 1)\left\{\frac{(o + 2)(o + 3)}{2}\right\} - o. \tag{8}$$

Six different pedigree types defined by the parameter o were investigated. A total of 250 simulated subjects were targeted for inclusion in any one simulation study. Because this number was not possible to achieve given certain of the pedigree sizes investigated, an effort was made to get as close to this number as possible. Table I displays some of the characteristics of the pedigree settings investigated. As noted in Table I, the pedigree types investigated (i.e., 3-generation, o parameter dictated) allow easy quantification of the sacrificed covariance terms.

Two different segregation parameter settings were studied. The first assumed allele frequencies of 0.25 and 0.75, codominance, with mean genotype effects of 12.0, 10.0, and 8.0, an additive variance of 1.0, a dominance variance of 0.0, and a random environmental variance of 1.0. The second setting was similar to the first except that the additive variance and dominance variance were both set to 0.5. Pedigree data conforming to each of these parameter configurations for each of the pedigree types outlined in Table I were generated. Pedigree data were generated with the covariance matrix defining genetic parameters [i.e., Eq. (2)] *complete* and intact to produce the polygenic and environmental effects. Major locus genotypes and effects were assigned to each pedigree member by the simple "gene dropping" method applied to the founders of each pedigree [MacCluer et al., 1986]. One hundred replications for each configuration were run. For each replicate run, estimates of the segregation parameters

were obtained by maximizing a likelihood function consistent with the scheme outlined above and Eq. (4) with the NPSOL optimization package [Gill et al., 1984]. The object of the study was to see if the proposed mixed model scheme could recover the generating parameters *despite* the loss or sacrifice of covariance terms assumed in the model. FORTRAN subroutines written by the author to implement the relevant likelihood function for the pedigree settings discussed are available at no charge from the author through e-mail (Nicholas-Schork@um.cc.umich.edu).

The results of the study are displayed in Table II. As can be seen from Table II, the generating parameters were recovered well, although there does appear to be a consistent underestimation of the environmental variance. The settings involving a true dominance variance of 0.0 did, of course, force some bias in the estimation of the variance terms in those settings. Remarkably, there does not appear to be any relationship between the amount of covariation *not* parameterized in the proposed model and the reliability of the parameter estimation, which suggests that the scheme may work in other settings and have some general utility in identifying major loci in the presence of polygenic and environmental "background" effects while allowing for the isolation and separation of these other genetic and environmental sources of variation.

Profile Likelihood Evaluation

In order to better assess the accuracy of the proposed methodology, a comparison of likelihoods evaluated *exactly* and those obtained using the approximation discussed in this paper was pursued. Three data sets with 25 pedigrees, each with 10 members and 3 generations (i.e., an o-parameter of 2), were generated that assumed different segregation patterns. Each data set assumed an allele frequency, p_a, of 0.25 and mean genotype effects of $\mu_{AA} = 12.0$, $\mu_{Aa} = 10.0$, and $\mu_{aa} = 8.0$. However, the first data set assumed $\sigma^2_a = 0.5$, $\sigma^2_d = 0.5$, and $\sigma^2_e = 1.0$, the second, $\sigma^2_a = 1.0$, and $\sigma^2_e = 1.0$, and the third, $\sigma^2_a = 2.0$, and $\sigma^2_d = 0.0$, and $\sigma^2_e = 1.0$. MLEs were obtained for the exact model by maximizing the complete multivariate normal likelihood functions over all the pedigrees' 10 members (i.e., no partition-sets or conditioning were used) using a complete covariance structure (i.e., no sacrficing of covariance terms was used). This function maximization required consideration of all 3,607 possible Mendelian genotype arrangements for each pedigree's 10 members at each iteration. Maximization was carried out using the NPSOL package. MLEs using the approximation discussed in this paper were also obtained from these data sets. After MLEs were obtained, each parameter was allowed to vary within a reasonable range while the others were held fixed at their maximum values so that an investigation and comparison of the effects this variation had on the likelihood of the parameter estimates for each data set could be made. It should be emphasized that the approximate model did yield likelihoods that were uniformly lower than those produced by the exact method. This is to be expected because of the sacrifice of covariance terms. However, by adding a

Fig. 3. Profile likelihoods for an exact and approximate mixed model for variance component parameters assuming 25 pedigrees with 10 members each following segregation patterns with parameters $p = 0.25$, $\mu_{AA} = 12.0$, $\mu_{Aa} = 10.0$, $\mu_{aa} = 8.0$, $\sigma^2_a = 0.5$, $\sigma^2_d = 0.5$, and $\sigma^2_e = 1.0$ (upper panel) $p = 0.25$, $\mu_{AA} = 12.0$, $\mu_{Aa} = 10.0$, $\mu_{aa} = 8.0$, $\mu^2_a = 1.0$, $\sigma^2_d = 1.0$, and $\sigma^2_e = 1.0$. The dashed line represents the approximation and the solid line the exact method. A value of 230 was added to likelihoods obtained with the approximate model.

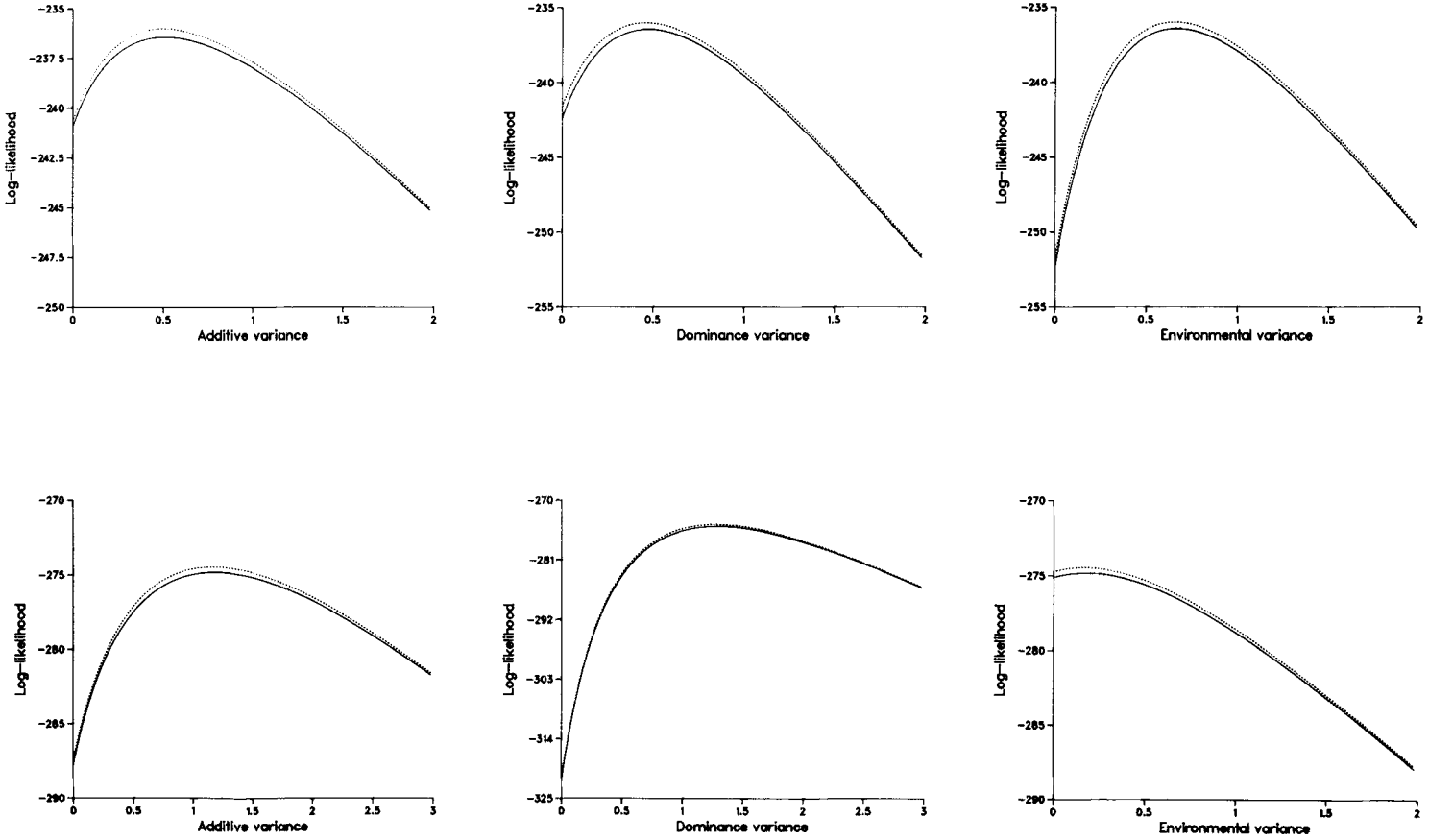
Log-likelihood
Additive variance
Dominance variance
Environmental variance

Figure 3.

TABLE III. Time in Seconds on Various Computers Needed to Compute One Iteration of the Proposed Mixed Model for Pedigrees of Various Sizes*

	o					
Computer	1	2	3	4	5	6
IBM XT (6 mhz)	4.390	13.840	43.280	138.570	455,930	1533.680
Apollo DN3000	0.100	0.251	0.830	2.723	9.000	30.101
Apollo DN3500	0.033	0.117	0.400	1.267	4.167	13.867
Apollo DSP4000	0.033	0.100	0.367	1.267	4.083	13.800
IBM PS/2 80 (16 mhz)	0.052	0.109	0.432	1.260	3.960	12.960
Sun SPARCstation 1	0.016	0.012	0.023	0.081	0.256	0.817
Dec 3100	0.008	0.008	0.019	0.055	0.178	0.676
Dec 5000	0.004	0.004	0.008	0.031	0.113	0.402
IBM 3090-600E (scalar)	0.001	0.002	0.005	0.016	0.052	0.174
IBM RS/6000	0.001	0.001	0.012	0.023	0.045	0.150
IBM 3090-600E (vector)	0.001	0.002	0.003	0.009	0.024	0.074

*o corresponds to the o-parameter used to characterize the pedigree size and shape (see text).

constant of 230 to all the likelihoods produced by the approximate model, agreement between the likelihoods obtained by both methods was found. This is encouraging since it suggests that the approximate model yields likelihoods that may only differ by a constant amount from the exact method in certain settings.

Figure 2 displays the "profile" likelihoods for the allele frequency and mean genotype effect parameters for the first data set. The general agreement between the exact and approximate (plus a constant of 230) likelihoods, especially around the maximum, was also found using the other data sets. Figure 3 displays the profile likelihoods for the variance component parameters obtained with the first two data sets. There is marked agreement between the two methods, even with respect to the underestimation of σ_e^2 for the second data set. Excellent agreement was also seen between the two methods in the consideration of the variance component parameters for the third data set.

Timing Studies

In order to assess the efficiency with which parameters could be estimated using the proposed scheme, some timing studies were performed. Table III reports the time in seconds needed to compute one function evaluation of the proposed method for various pedigree sizes and structures on some commonly used computers. Schork [1991a] suggested that it generally takes 150–200 function evaluations to maximize mixed model likelihoods for nuclear family data using NPSOL (note: this includes function evaluations used to compute finite-difference approximations of derivatives). This range of numbers can be used with Table III to gauge the length of time one might need to obtain estimates of segregation parameters using the proposed scheme. Further comparisons involving other machines can be obtained by using Table III and the relative speeds of certain machines described by Dongarra [1991], since the program used to carry out the studies described in this paper was written to exploit the matrix/vector characterization of the likelihood functions given in Eqs. (3) and (4).

DISCUSSION

As noted in the introduction, the construction of reliable and practical mixed models for human quantitative phenotypes has been hindered by computational difficulties. The methodology outlined previously responds to these difficulties. The method is both intuitive and flexible. Though more detailed investigations involving a variety of possible segregation parameter settings are called for, especially in the area of hypothesis testing, the results of the simulation study described previously suggest that the method may also be a reliable one. Evaluation of the necessary likelihood functions can be facilitated through the vectorization strategy outlined in Schork [1991a] (see Table III, last row) for computing over patterned covariance mixed models involving partition-set members, and can be further facilitated through the parallel strategy discussed in Schork [1991b] for large, possibly complex, zero-loop pedigrees.

The methodology described in this paper invites comparison with the "regressive" approach advocated by Bonney [1984]. Bonney [1984] also suggested that modeling variation occurring in the "background" of a major locus could be restricted to partition-sets within the pedigree. However, Bonney explicitly restricted his partition-sets to nuclear families within the pedigree and parameterized his "background" effects in terms of correlations that were not wedded to Mendelian theory. Though such parameterization does afford some flexibility, it does not allow the isolation and distinction of variation attributable to genetic and nongenetic sources. In addition, although Demenais and Bonney [1989] showed that one parameterization of a regressive model was equivalent to a certain mixed model parameterization for nuclear family data, they later considered the use of less complicated models in an effort to avoid computational burdens [Demenais et al., 1990]. The method described in this paper is relatively efficient computationally (Table III) and can be made more so, as noted above. In addition, the covariance matrix used in Eq. (4) for each partition-set *could* be parameterized in terms of correlations compatible with regressive models, if it was desired to do so.

Obviously, greater focus on, and research interest in, both previously proposed mixed models and the method outlined in this paper will force a more precise characterization and determination of the practical and theoretical merits of various mixed model strategies—something sorely needed if the investigation of the genetic basis of human quantitative traits is to keep pace with developments in laboratory technologies and plant and animal genetic analysis techniques.

REFERENCES

Bonney GE (1982): Maximum likelihood methods for genetic analysis of multivariate pedigree data. Ph.D. dissertation, Department of Biostatistics, University of North Carolina.

Bonney GE (1984): On the statistical determination of major gene mechanisms in continuous human traits: regressive models. Am J Med Genet 35:816–826.

Boyle CR, Elston RC (1979): Multifactorial genetic models for quantitative traits humans. Biometrics 35:55–68.

Cannings CA, Thompson EA, Skolnick M (1978): Probability functions on complex pedigrees. Adv Appl Prob 10:26–61.

Demenais FM, Bonney GE (1989): Equivalence of the mixed and regressive models for genetic analysis. I. Continuous traits. Genet Epidemiol 6:597–617.

Demenais FM, Murigande C, Bonney GE (1990): Search for faster methods of fitting the regressive models to quantitative traits. Genet Epidemiol 7:319–334.

Dongarra JJ (1991): Performance of various computers using standard linear equations software. Argonne National Laboratory Report MSCO-TM-23m.

Elston RC (1981): Segregation analysis. In Harris H, Hirshhorn K (eds): "Advances in Human Genetics." New York: Plenum, 63–120.

Elston RC, Stewart J (1971): A general model for the analysis of pedigree data. Hum Hered 21:323–342.

Gill PE, Murray W, Saunders MA, Wright MH (1984): User's guide for SOL/NPSOL: A FORTRAN package for non-linear programming. Stanford University, Department of Operations Research, Report SOL 83-1.

Hasstedt SJ (1982): A mixed-model likelihood approximation on large pedigrees. Comput. Biomed. Res. 15:295–307.

Hasstedt SJ (1991): A variance components/major locus likelihood approximation on quantitative data. Genet Epidemiol 8:113–126.

Jacquard A: (1974) "The Genetic Structure of Populations." New York: Springer.

Konigsberg LW, Kammerer CM, MacCluer JW (1989) Segregation analysis of quantitative traits in nuclear families: comparison of three program packages. Genet Epidemiol 6:713–726.

Lalouel JM, Rao DC, Morton NE, Elston RC (1983): A unified model for complex segregation analysis. Am J Hum Genet 35:816–826.

Lander ES, Botstein D (1989): Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.

Lange K (1978): Central limit theorems for pedigrees. J Math Biol 6:59–66.

Lange K, Elston RC (1975): Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. Hum Hered 25:95–105.

Lange K, Westlake J, Spence MA (1976): Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet 39:485–491.

MacCluer JW, VandeBerg JL, Read B, Ryder DA (1986): Pedigree analysis by computer simulation. Zoo Biol 5:147–160.

Morton NE, MacLean CJ (1974): Analysis of family resemblance. III. Complex segregation analysis of complex traits. Am J Hum Genet 26:489–503.

Ott J (1974): Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588–597.

Ott J (1979): Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigree analysis. Am J Hum Genet 31:161–175.

Ott J (1990): Cutting a Gordian knot in the linkage analysis of complex human traits. Am J Hum Genet 46:219–221.

Schork NJ (1991a): Efficient computation of patterned covariance matrix mixed models in quantitative segregation analysis. Genet Epidemiol 8:29–46.

Schork NJ (1991b): The parallel computation of pedigree likelihoods. In Keramidas EM (ed): Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Fairfax Station, VA: Interface Foundation of North America, Inc., pp 262–265.

Edited by G.P. Vogler