Evaluation of likelihood ratios for complex genetic models.

E. A. Thompson,[1,2] and Sun Wei Guo.[2]

1: Department of Statistics, GN-22, University of Washington, Seattle, WA 98195.

2: Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195.

Abstract

Although methods for computing likelihoods for simple genetic models on large and complex pedigrees have been known for some time, and although methods for evaluating likehoods for complex genetic models on small pedigrees have likewise been well known, likelihood evaluation for complex models given data on extended pedigrees has remained an intractable problem. The Gibbs sampler provides a method of Monte Carlo evaluation of likelihood ratios for complex models on extended and/or complex pedigrees. With increasing computer speeds, this approach provides a tractable and efficient approach to many such likelihood evaluation problems in linkage and segregation analysis. In the current paper, however, we restrict attention to two basic building-blocks of the overall process. The first is the sequential computation of Gaussian likelihoods for multiple random-effects models on extended pedigrees. The second is the use of this in the Monte Carlo evaluation of likelihoods for the classical mixed model of segregation analysis. We detail the implementation of the Gibbs sampler on pedigrees that permits this Monte Carlo evaluation. We then present an example, and finally, in the context of this same example, show also how linkage analysis for a quantitative trait falls within this same framework.

Keywords: Monte Carlo integration; Gibbs sampler; extended pedigrees; complex genetic models; likelihood evaluation.

1. Introduction.

A major part of Genetic Epidemiology is concerned with the likelihood analysis of genetic models for phenotypic data observed on related individuals. Many traits have some genetic component. In this paper we will restrict attention to a single quantitative trait and denote the set of observed values by $\mathbf{y}$, with y_j being the value for individual j. A pedigree (or set of related individuals) is most conveniently thought of as a collection of "marriages", each indicating a male and a female parent in the marriage, and a list of the offspring of the marriage. Any given individual may be in several of these nuclear families (sets of parents and offspring). An individual can be an offspring of only one marriage, but could be a parent in many, with different "spouses". It is convenient to require (by the addition of unobserved "dummy" individuals, if necessary), that every marriage specifies two parent individuals. Individuals who are not the offspring of any specified marriage are the "founders" of the pedigree; the remainder are the non-founders. Figure 1 shows the pedigree used in the examples of this paper.

Specification of a genetic model requires specification of the types of genetic effects involved, their pattern of transmission from parents to offspring, and the relationship between the genetic effects present in an individual and the observed value of the trait for that individual. Genetic effects can be modelled either as discrete Mendelian genotypes, or as the combined effects of genes at many loci each making a small contribution to the trait ("polygenic effects"). The "mixed model" is characterised by having both fixed and random effects as components contributing additively to phenotype. Moreover, in genetics, the mixed model refers to the case where components of each type are heritable; "major genes" contribute effects dependent on the discrete Mendelian genotype, while "polygenic background" contributes the random effects. The mixed model in its classical form has not contributed much to the resolution of complex genetic traits. The future of resolving the genetic components of such traits is now generally accepted to lie in linkage analysis. Only by linking the major gene components to a marker locus will the genetic components be resolved (Elston et al., 1989). However, for human familial data, linkage analysis and mapping of quantitative trait loci (Lander and Botstein, 1989) is not straightforward. Thus useful mixed models must include not only the classical components of fixed and random, heritable and non-heritable, effects, but the major-gene components may involve also linked markers. Consequently notation is complex; a summary of the notation used for variables and parameters is given in Table 1.

In likelihood analysis of complex genetic models on extended or complex genealogies, the first requirement is computation of a likelihood. The most successful method of computing likelihoods on large and/or complex pedigrees is that of "peeling". Peeling algorithms originated with Hilden (1970), Elston and Stewart (1971) and Heuch and Li (1972). The approach was generalised to more complex pedigrees and more complex models by Cannings, Thompson and Skolnick (1976, 1978, 1980), and is very closely related to more recently developed methods for the propagation of probabilities in belief networks and probabilistic expert systems (Lauritzen and Spiegelhalter, 1988). The key feature of this class of algorithms lies in the use of the conditional independence that is a feature of genetic models: given the types of the genes carried by a set of individuals who divide the pedigree, phenotypic events observed on the two subdivisions are independent. This conditional independence will be used in a slightly different way in the current paper.

The peeling algorithm has enabled the computation of likelihoods on large and complex pedigrees for very simple genetic models. It permits linkage analysis for simple Mendelian traits on large pedigrees provided these are not too complex, and has been used to compute likelihoods for simple polygenic models on large but simple pedigrees. (For a more detailed description of peeling, and of its use in the computation of probabilities on pedigrees, see Thompson (1986)). These seem to be its limits, although it has also been used to suggest an approximate computation of likelihoods for mixed models (Hasstedt, 1982). However, many current genetic epidemiological questions relate to the inheritance of complex quantitative genetic traits, where there may be several heritable effects, some most appropriately modeled as Mendelian (fixed) effects, and others as polygenic effects. Several programs have been developed for the segregation analysis of such traits in nuclear families; these are reviewed and compared by Konigsberg et al. (1989). However, to ensure sufficient data in circumstances of relative genetic homogeneity it will be necessary to analyse large pedigrees, and many large pedigrees are complex. Additionally, it will be desirable to map any Mendelian effects, requiring linkage analysis to be performed in conjunction with complex segregation analysis on extended pedigrees.

2. Peeling polygenic (random effects) models

The application of (a version of) the peeling algorithm to the computation of likelihoods for the simple additive polygenic model on a large pedigree was first developed by Elston and Stewart (1971). Although the fact that likelihoods for more complex random effects models could be computed by peeling on large pedigrees is well known (Ott, 1979), the only implementations are for additive genetic components, on zero-loop pedigrees (Elston and Stewart, 1971; Hasstedt and Cartwright, 1979). The possibility of inclusion of dominance, common sib environment, maternal line effects, or other heritable random effects is not generally appreciated, The joint peeling of polygenic components of several individuals (Cannings et al., 1980) has also not been implemented. Here we provide a general formulation and implementation of peeling of complex polygenic models on possibly complex pedigrees. This will be a component of our method for the Monte Carlo evaluation of likelihood ratios for more complex genetic models.

The sequential computation of likelihoods for a Gaussian random effects model is based on two essential facts about multivariate Gaussian distributions. The first is that such distributions are completely specified by the mean and by the covariance matrix. The second is that the marginal distribution for any subset of the variables is also Gaussian, and that the distribution formed by conditioning on any subset of the variables is likewise Gaussian. Here it is convenient to specify our Gaussian form as

$$K \exp(-\mathbf{w}^T \mathbf{B} \mathbf{w}/2) \tag{2.1}$$

where $\mathbf{B}$ is a symmetric positive semidefinite matrix (corresponding to the inverse of the covariance matrix of the random variables), and K is a multiplying constant. By including the number 1 as a component of $\mathbf{w}$, linear terms (corresponding to means) and constant terms are incorporated into the exponent of expression (2.1). Strictly speaking, K is therefore unnecessary, but it is convenient to allow a separate constant multiplier.

Any multivariate Gaussian density can be explicitly integrated with respect to any variable to give an expression of the same form in the remaining components. Specifically, in the notation of (2.1),

$$\int_{-\infty}^{\infty} \exp(-\mathbf{w}^T\mathbf{B}\mathbf{w}/2)\; dw_k = (2\pi/b_{kk})^{\frac{1}{2}} \exp(-\mathbf{w}*^T\,(\mathbf{B}* - \mathbf{b}\mathbf{b}^T/b_{kk})\mathbf{w}*/2) \tag{2.2}$$

where $\mathbf{w}^T = (\mathbf{w}*^T, w_k)$ and $\mathbf{B}*$ is $\mathbf{B}$ with the k^{th} row/column eliminated, $\mathbf{b}$ is that row/column, without the (k,k) element, and b_{kk} is that element i.e.;

$$\mathbf{B} \;=\; \begin{bmatrix} \mathbf{B}* & \mathbf{b} \\ \mathbf{b}^T & b_{kk} \end{bmatrix}$$

Further, multiplication of any current form by another in some or all of the same variables, leads to a combined form:

$$K_1 \exp(-\mathbf{w}^{(1)T}\mathbf{B}_1\mathbf{w}^{(1)}/2) \times K_2 \exp(-\mathbf{w}^{(2)T}\mathbf{B}_2\mathbf{w}^{(2)}/2) \;=\; K_3 \exp(-\mathbf{w}^{(3)T}\mathbf{B}_3\mathbf{w}^{(3)}/2)$$

where $K_3 = K_1K_2$, $\mathbf{w}^{(3)}$ is the union of the component variables of $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ and $\mathbf{B}_3 = \mathbf{B}*_1 + \mathbf{B}*_2$ where $\mathbf{B}*_1$ is the matrix $\mathbf{B}_1$ augmented by columns/rows of zeros corresponding to the variables present in $\mathbf{w}_2$ but not in $\mathbf{w}_1$, and $\mathbf{B}_2$ is similarly defined. Thus repeated multiplication and integration of Gaussian forms can be effected simply by storing (and later retrieving) the set of variables on which the current form is defined, the the multiplicative constant K, and the symmetric matrix $\mathbf{B}$. In the pedigree context, each stored form will normally be defined by a set of individuals (for a zeroloop pedigree, one individual), and the terms $\mathbf{w}$ will correspond to the heritable random effects of those individuals.

As an example, we shall consider the specific case of a simple polygenic model for a normally distributed quantitative trait affected by common sib environment, with observed values $\mathbf{y}$ on a set of related individuals, the model may be written

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{z} + \mathbf{d} + \mathbf{e} \tag{2.3}$$

where $\boldsymbol{\mu}$ is a vector of means (incorporating perhaps some fixed effects), $\mathbf{z}$ is a vector of breeding values distributed as $N(\mathbf{0}, \boldsymbol{\sigma_a}^2\mathbf{G})$, $\mathbf{d}$ is a vector of common sib effects distributed as $N(\mathbf{0}, \boldsymbol{\sigma_d}^2\mathbf{D})$, and $\mathbf{e}$ is a vector of uncorrelated residuals, each of variance $\boldsymbol{\sigma_e}^2$. The matrix $\mathbf{G}$ has components which are twice the kinship coefficients between individuals; for non-inbred individuals $G_{ii} = 1$. Alternatively the vector of additive effects $\mathbf{z}$ can be generated via

$$z_i = (z_m + z_f)/2 + s_i \tag{2.4}$$

where m and f are the parents of i, the segregation effects s_i are independent, mean zero, and variance $\boldsymbol{\sigma_a}^2/2$. The matrix $\mathbf{D}$ does not have full rank; $D_{ij} = 1$ if individuals i and j are sibs, and is 0 otherwise. In the case of a zeroloop pedigree, common-sib effects are indistinguishable from dominance effects; a reparametrisation

$$\boldsymbol{\sigma_e}^2* = \boldsymbol{\sigma_e}^2 - 3\,\boldsymbol{\sigma_d}^2, \qquad \boldsymbol{\sigma_d}^2* = 4\,\boldsymbol{\sigma_d}^2 \tag{2.5}$$

translates one model into the other provided the necessary positivity constraints are met. Where an individual has no sibs in the pedigree, we can consider the single variable $d_i + e_i$ with variance $\boldsymbol{\sigma_d}^2 + \boldsymbol{\sigma_a}^2$.

We shall detail the computation for the model (2.3) only for a zeroloop pedigree, since this is more easily described. The procedure is outlined in figure 1. Many of the extended pedigrees encountered in human genetic epidemiological analyses are zeroloop pedigrees. As for the Mendelian case (Cannings, Thompson and Skolnick, 1978), the approach is essentially identical to that for an arbitrary pedigree, although then effects for several individuals must be handled jointly. First, we define a "peeling sequence" of the marriages in the pedigree. This is simply an ordering of the nuclear families defined by these marriages, ordered in such a way that, when each family comes to be considered, only one individual (known as the "pivot") connects the family to that part of the pedigree not yet considered. For a zeroloop pedigree, clearly this is always possible.

Peeling a nuclear family

Step 1: the initial state Consider a marriage in this sequence, defining a nuclear family, with mother m, father f, and offspring i, $i = 1, \ldots, k$. We may suppose that likelihoods have been computed from parts of the pedigree connected to the marriage node via all save one of the individuals, this individual to be the "pivot" for the current computation. The pivot must be either a parent or an offspring, so, without loss of generality, suppose that partial likelihoods have been computed for the part of pedigree connected to the marriage node via f, and for offspring individuals $1, \ldots, k-1$, and either through k so that the pivot is the remaining parent m, or through m the pivot then being the final offspring k.

Step 2; peeling non-pivot offspring i. Thus we have a vector of effects to consider when peeling the offspring i, $\mathbf{w}^{(1)} = (1, z_m, z_f, d, z_i)^T$. Here d is the shared sib effect in the sibship, and the final component "1" is used to incorporate all linear terms into the quadratic form. Any constant term may be held either in the term $\exp(-B_{00})$ or as a multiplying constant K. From previous peeling to individual i we have say, likelihood

$$K_i \exp(-(1, z_i)^T B_i (1, z_i)/2) \tag{2.6}$$

The segregation from parents m and f to i gives an additional factor

$$(2\pi \boldsymbol{\sigma_a}^2/2)^{-\frac{1}{2}} \exp(-(z_i - 0.5(z_m + z_f))^2/\boldsymbol{\sigma_a}^2) \tag{2.7}$$

and data on individual i (if this individual is observed) provides a factor

$$(2\pi v)^{-\frac{1}{2}} \exp(-(z_i + \mu_i + d - y_i)^2/(2v)) \tag{2.8}$$

where v is the relevant variance of individual-specific effects (normally $\boldsymbol{\sigma_e}^2$; see above). Combining the quadratic forms (2.6)-(2.8), we obtain a form in $\mathbf{w}^{(1)}$. Specifically, segregation adds a term

$$\frac{2}{\boldsymbol{\sigma_a}^2} \, (0, -\tfrac{1}{2}, -\tfrac{1}{2}, 0, 1) \, (0, -\tfrac{1}{2}, -\tfrac{1}{2}, 0, 1)^T$$

into the matrix of the exponent, and the data observation a term

$$\frac{1}{v} \, (\mu_i - y_i, 0, 0, 1, 1) \, (\mu_i - y_i, 0, 0, 1, 1)^T$$

Thus if $\mathbf{B}_i = \begin{pmatrix} b_{00} & b_{0i} \\ b_{0i} & b_{ii} \end{pmatrix}$ then the form $K^{(1)} \exp(-\mathbf{w}^{(1)^T} B^{(1)} \mathbf{w}^{(1)}/2)$ has

$$\mathbf{B}^{(1)} = \begin{pmatrix} b_{00} + (\mu_i - y_i)^2/v & 0 & 0 & (\mu_i - y_i)/v & b_{0i} + (\mu_i - y_i)/v \\ 0 & 1/(2\boldsymbol{\sigma_a}^2) & 1/(2\boldsymbol{\sigma_a}^2) & 0 & -1/\sigma_a{}^2 \\ 0 & 1/(2\boldsymbol{\sigma_a}^2) & 1/(2\boldsymbol{\sigma_a}^2) & 0 & -1/\sigma_a{}^2 \\ (\mu_i - y_i)/v & 0 & 0 & 1/v & 1/v \\ b_{0i} + (\mu_i - y_i)/v & -1/\sigma_a{}^2 & -1/\boldsymbol{\sigma_a}^2 & 1/v & b_{ii} + (2/\boldsymbol{\sigma_a}^2) + (1/v) \end{pmatrix}$$

and

$$K^{(1)} = K_i\,(\pi\boldsymbol{\sigma_a}^2)^{-\frac{1}{2}}\,(2\pi v)^{-\frac{1}{2}}$$

The variable z_i may now be integrated from this exponential quadratic form, using equation (2.2). This provides quadratic form and accompanying constant term $K^{(2)}\exp(-\mathbf{w}^{(2)^T}B^{(2)}\mathbf{w}^{(2)}/2)$ where $\mathbf{w}^{(2)} = (1, z_m, z_f, d)^T$.

Step 3; Case 1; m is pivot; peeling d and z_f

If all the offspring have been peeled, and the pivot is m, the effect d may next be eliminated after multiplying by the population probability density for d,

$$(2\pi\boldsymbol{\sigma_d}^2)^{-\frac{1}{2}}\exp(-d^2/2\boldsymbol{\sigma_d}^2), \tag{2.9}$$

and integrating with respect to d.

Any previously computed quadratic form in z_f, which will have already incorporated data observations and/or sibship effects relating to that parent, are next incorporated. For a founder parent f not previously considered, a form incorporating the data observation (if any) and the population probability density for z_f is incorporated:

$$(2\pi v^*)^{-\frac{1}{2}}\ exp(-(z_f + \mu_f - y_f)^2/(2v^*)) \text{ and } (2\pi\boldsymbol{\sigma_a}^2)^{-\frac{1}{2}}\exp(-z_f{}^2/(2\boldsymbol{\sigma_a}^2)), \tag{2.10}$$

where $v* = \boldsymbol{\sigma_d}^2 + \boldsymbol{\sigma_e}^2$. Then, z_f is eliminated by integration, leaving finally a quadratic form in $\mathbf{w}^{(3)} = (1, z_m)^T$.

Data on m will be incorporated at the point at which m is peeled as an offspring, or, if m is a founder, must be incorporated before m is peeled as the parent of another sibship. In the latter case a data observation y_m on m gives rise to a contribution

$$(v^*)^{\frac{1}{2}}\ exp(-(z_m + \mu_m - y_m)^2/(2v^*))$$

as for f in (2.10).

Step 3; Case 2; pivot is child k

In the event that the pivot of the nuclear family is the unpeeled offspring individual k, previously computed likelihoods (or data and population priors on additive genetic values) for both parents m and f are incorporated, and the integration w.r.t. z_m and z_f successively performed, leaving a quadratic from in $\mathbf{w}^{(4)} = (1, d, z_k)^T$. If k is observed, the data for this individual must now be incorporated, since it involves d;

$$v^{-\frac{1}{2}}\exp(-(z_i + \mu_i + d - y_i)^2/(2\boldsymbol{\sigma_e}^2) \tag{2.11}$$

as in equation (2.8) above. and then the prior density for d can be incorporated as in (2.9)

$$(2\pi\boldsymbol{\sigma_d}^2)^{-\frac{1}{2}}\exp(-d^2/2\boldsymbol{\sigma_d}^2).$$

Then d is integrated out, leaving a final quadratic form in $\mathbf{w}^{(5)} = (1, z_k)^T$, to be carried forward to computations on families in which k contributes as a parent.

Peeling of the nuclear family is now complete

The process is repeated successively for each family of the pedigree, incorporating previous contributions until a final likelihood is obtained.

Note that any individual is an offspring either once or zero times, but can be a parent in any number of sibships. It is important to ensure that data on each observed individual gets incorporated once and once only. This is straightforward on a zeroloop pedigree, but requires care in complex pedigrees. The dominance effect for a sibship should be incorporated only when it is to be peeled, which can be done only after the data on the last sibship member is incorporated. Thus, data for an offspring pivot should be incorporated in the computation relating to that family, and not deferred until he is considered as a parent. Data on founder individuals should be incorporated when they are to be peeled. Introducing each quadratic form contribution at the last possible time in the sequence of computations not only simplifies the contributions held at any time, but also ensures all effects are incorporated once and once only. This is algorithmically a little different to peeling of qualitative traits, where it can be more efficient to introduce penetrance probabilities earlier in the process, to avoid consideration of zero terms; this does not apply to quantitative traits.

We have described the above process in detail, because of its parallels to the usual peeling for likelihoods for simple Mendelian traits which will be familiar to some readers. However, it should be clear that the overall validity and feasibility of the procedure does not depend on the particular random effects model chosen or on the particular form of pedigree. As for Mendelian models, if a pedigree is looped, it will be necessary to carry joint quadratic forms on the breeding values of several individuals. We note that this is in fact computationally simpler than in the Mendelian case. For k individuals in a cutset, a 3-genotype discrete Mendelian model requires up to 3^k numbers to be held, one corresponding to each genotype combination on the cutset. For a matrix of coefficients of a quadratic form on k breeding values, the storage required is of order k^2.

In our detailed description, we have included only the simplest possible effects. Other random effects, both heritable and non-heritable, can be easily incorporated, provided only all of these are Gaussian and contribute additively to an observed phenotype. It is also possible to peel the case ${\boldsymbol{\sigma}_e}^2 = 0$, in which heritability is 1. This can be a useful model for comparison of likelihoods. With regard to additional effects, a maternal line (mitochondrial) effect may be of interest, for example. In this case the design matrix $\mathbf{M}$ has $M_{ij} = 1$ if individuals i and j are in the same maternal line, and 0 otherwise. In terms of the peeling, each individual contains an additive effect u_i equal for all members of the line. The quadratic form will carry terms in the current maternal effects until all members of the line have been peeled, at which point it can be integrated out, after incorporation of a prior $(2\pi{\boldsymbol{\sigma}_m}^2)^{-\frac{1}{2}} \exp(-u^2/(2{\boldsymbol{\sigma}_m}^2))$. Maternal line effects are used to illustrate the Monte-Carlo EM approach to maximisation of likelihoods for variance component models on pedigrees by Guo and Thompson (1991a).

3. The mixed model likelihood

The mixed model incorporates both a polygenic contribution to the trait and the effects of a major genotype. The simplest mixed model can be written, in the notation of Table 1 and of the previous section

$$\mathbf{y} = \mathbf{g} + \mathbf{z} + \mathbf{e}$$

where g_j is the contribution to the trait value due to the major genotype G_j of individual j. In general, there might be more than one Mendelian locus whose genotype contributes to the trait, and the random effects components of the model might include more than the additive genetic value z and residual e. The mixed model has played a large role in human genetic epidemiological endeavour (Ott, 1979; Lalouel et al, 1983). While the main objective has been the detection of major genes contributing to a quantitative trait, the modelling of additional heritable variation via polygenic contributions to the trait value has been the usual procedure. The two longest established program packages capable of complex segregation analysis are PAP (Hasstedt and Cartwright, 1979) and POINTER (Morton et al., 1983). The theoretical backgrounds to the methods in these two packages are given by Hasstedt (1982) and Lalouel et. al, (1983), respectively. Both include mixed-model computations. Konigsberg et al. (1989), in comparing the performance of these and other program packages for the analysis of nuclear family data, used the mixed model as one of their key standards for comparison.

The likelihood under a mixed model, with a contribution to the quantitative trait from a major genes as well as a polygenic component, is a mixture of likelihood contributions of polygenic form, one corresponding to each possible major-genotype combination over the entire pedigree; this is a very large number! Since the overall form is neither Gaussian nor discrete, the mixed model is intrinsically "unpeelable". Hasstedt (1982) investigated, on small pedigrees, the approximation of replacing the extant Gaussian components at each stage of peeling by a set of three components. Each of these is peeled at the next stage, proliferating more components in accordance with the major genotype combinations on the current nuclear family, and the reduction/approximation process repeated. This is the method used by PAP (Hasstedt and Cartwright, 1979) to compute mixed-model likelihoods, but the properties and performance of the approximation have not been generally investigated.

We may write the likelihood under a mixed model in the generic form

$$L(\theta) = \sum_{\mathbf{G}} \int_{\mathbf{z}} f_\theta(\mathbf{y} \mid \mathbf{z}, \mathbf{G})\ P_\theta(\mathbf{G})\ f_\theta(\mathbf{z})\ d\mathbf{z} \tag{3.1}$$

where $\mathbf{G}$ incorporates the major genes, $\mathbf{z}$ incorporates all random effects, and θ denotes the set of parameters of the model. (For simplicity, one may think of $\mathbf{z}$ being the vector of breeding values, and $\mathbf{G}$ the vector of genotypes at a single locus, but the model would normally be more complex than this.)

Although the mixed model likelihood is intrinsically unpeelable, for given $\mathbf{z}$,

$$f_\theta(\mathbf{y} \mid \mathbf{z}) = \sum_{\mathbf{G}} f_\theta(\mathbf{y} \mid \mathbf{z}, \mathbf{G})\ P_\theta(\mathbf{G}) \tag{3.2}$$

is computable by Mendelian peeling (Cannings et al., 1978), and for given $\mathbf{G}$

$$f_\theta(\mathbf{y} \mid \mathbf{G}) = \int_{\mathbf{z}} f_\theta(\mathbf{y} \mid \mathbf{z}, \mathbf{G})\ f_\theta(\mathbf{z})\ d\mathbf{z} \tag{3.3}$$

is computable by polygenic peeling, as described in the previous section. Then

$$L(\theta) = \sum_{\mathbf{G}} f_\theta(\mathbf{y} \mid \mathbf{G})\, P_\theta(\mathbf{G}) \tag{3.4}$$

$$= \int_{\mathbf{z}} f_\theta(\mathbf{y} \mid \mathbf{z})\, f_\theta(\mathbf{z})\, d\mathbf{z} \tag{3.5}$$

Thus the problem is particularly suited to Monte Carlo evaluation, where either $\mathbf{z}$ or $\mathbf{G}$ is sampled from an appropriate distribution. For ease of presentation we shall focus on the form (3.4) viewing the likelihood as a mixture of polygenic likelihoods.

The simplest way to use Monte Carlo methods to evaluate the sum (3.4) would be simply to generate realisations of the major genotype configuration $\mathbf{G}$ under P_θ. Such "gene-drop" (MacCluer et al., 1986) realisations can be very easily obtained. Given a sample $\{\mathbf{G}^{(1)}, \cdots, \mathbf{G}^{(N)}\}$ of such realisations.

$$\hat{L}(\theta)^{(1)} = \frac{1}{N} \sum_{i=1}^{N} f_\theta(\mathbf{y} \mid \mathbf{G}^{(i)})$$

provides an estimate of $L(\theta)$. However, on a large pedigree such an estimation method is ineffectual. Most of the realisations $\mathbf{G}^{(i)}$ will provide only miniscule contribution to the likelihood, and those few genotypic configurations that provide substantial contribution have miniscule probability of being realised even in a large sample. Instead, some form of importance sampling is needed. One form of such sampling was suggested by Lange in Ott(1979). Instead of sampling the $\mathbf{G}^{(i)}$ according to the distribution indexed by θ, it was suggested that sampling according to a base-point P_{θ_0} might be more effective, providing an estimator

$$\hat{L}(\theta)^{(2)} = \frac{1}{N} \sum_{i=1}^{N} f_\theta(\mathbf{y} \mid \mathbf{G}^{(i)})\, \frac{P_\theta(\mathbf{G}^{(i)})}{P_{\theta_0}(\mathbf{G}^{(i)})}$$

This proposal has two advantages. First it may be possible to choose θ_0 so that realisations are easier to generate. Second, one set of realisations according to the distribution P_{θ_0} will provide estimates for a number of other θ-values. However, it remains ineffectual on pedigrees even of moderate size, for the same reason as before: the realisations bear no relation to the observed data $\mathbf{y}$ and therefore the vast majority make miniscule contribution to the likelihood.

Clearly what is required is a method of sampling that takes the data into account. In the usual importance-sampling framework (Hammersley and Handscombe, 1964), we need a sampling distribution which mimics in shape the summand $f_\theta(\mathbf{y} \mid \mathbf{G})P_\theta(\mathbf{G})$ which, as a function of $\mathbf{G}$, is proportional to the conditional distribution $P_\theta(\mathbf{G} \mid \mathbf{y})$. Specifically, suppose we could sample from the global conditional distribution of genotypes given the phenotypes on the entire pedigree, under some set of parameter values θ_0, where $P_{\theta_0}(.)$ is a similar distribution to $P_\theta(.)$ (That is, the parameter values θ and θ_0 should be broadly similar.) This form of sampling (if feasible) would be particularly effective in focusing the sample on those values of $\mathbf{G}$ and/or $\mathbf{z}$ which are *a posteriori* probable, given the data $\mathbf{y}$ and the parameter value θ_0 (or θ). It is thus far more efficient than any "gene-drop" sampling of the latent genotypic variables. We defer showing how this sampling is accomplished, and show first how it would be used.

To implement this sampling scheme, and obtain valid estimates of $L(\theta)$ we would have to incorporate the appropriate importance-sampling weighting:

$$L(\theta) = \sum_{\mathbf{G}} \frac{f_\theta(\mathbf{y} \mid \mathbf{G}) P_\theta(\mathbf{G})}{P_{\theta_0}(\mathbf{G} \mid \mathbf{y})} P_{\theta_0}(\mathbf{G} \mid \mathbf{y})$$

giving estimate

$$\hat{L}(\theta)^{(3)} = \frac{1}{N} \sum_{i=1}^{N} \frac{f_\theta(\mathbf{y} \mid \mathbf{G}^{(i)})\ P_\theta(\mathbf{G}^{(i)})}{P_{\theta_0}(\mathbf{G}^{(i)} \mid \mathbf{y})} \tag{3.6}$$

where now the $\{\mathbf{G}^{(i)}: i = 1, \cdots, N\}$ are sampled from the global posterior distribution of genotype given phenotype under θ_0. At first sight this is unpromising, for $P_{\theta_0}(\mathbf{G} \mid \mathbf{Y})$ is not explicitly known; if it were known there would be no problem in direct likelihood evaluation! Suppose, however, that instead of considering $L(\theta)$, we consider the likelihood ratio $l(\theta: \theta_0) = L(\theta)/L(\theta_0)$ In likelihood analysis of models we are, in any case, only interested in likelihood ratios; the absolute value of $L(\theta)$ is not useful. Then

$$l(\theta: \theta_0) = \sum_{\mathbf{G}} \frac{f_\theta(\mathbf{y} \mid \mathbf{G}) P_\theta(\mathbf{G})}{f_{\theta_0}(\mathbf{y}) P_{\theta_0}(\mathbf{G} \mid \mathbf{y})} P_{\theta_0}(\mathbf{G} \mid \mathbf{y})$$

since, by definition, $L(\theta_0) = f_{\theta_0}(\mathbf{y})$. But now we may rewrite the denominator containing the product of two probabilities that are not explicitly known, as the product of two known probabilities;

$$l(\theta: \theta_0) = \sum_{\mathbf{G}} \frac{f_\theta(\mathbf{y} \mid \mathbf{G})\ P_\theta(\mathbf{G})}{f_{\theta_0}(\mathbf{y} \mid \mathbf{G})\ P_{\theta_0}(\mathbf{G})} P_{\theta_0}(\mathbf{G} \mid \mathbf{y})$$

(An additional computational advantage of considering likelihood ratios is that underflow and overflow can generally then be avoided, whereas for the likelihoods themselves, contributions are often infinitesimally small.)

More generally, considering likelihood ratios, (3.1), (3.4) and (3.5) become

$$l(\theta: \theta_0) = \sum_{\mathbf{G}} \int_{\mathbf{z}} \frac{f_\theta(\mathbf{z})}{f_{\theta_0}(\mathbf{z})}\ \frac{P_\theta(\mathbf{G})}{P_{\theta_0}(\mathbf{G})}\ \frac{f_\theta(\mathbf{y} \mid \mathbf{z}, \mathbf{G})}{f_{\theta_0}(\mathbf{y} \mid \mathbf{z}, \mathbf{G})}\ dF_{\theta_0}(\mathbf{z}, \mathbf{G} \mid \mathbf{y}) \tag{3.7}$$

$$= \sum_{\mathbf{G}} \frac{P_\theta(\mathbf{G})}{P_{\theta_0}(\mathbf{G})}\ \frac{f_\theta(\mathbf{y} \mid \mathbf{G})}{f_{\theta_0}(\mathbf{y} \mid \mathbf{G})}\ P_{\theta_0}(\mathbf{G} \mid \mathbf{y}) \tag{3.8}$$

$$= \int_{\mathbf{z}} \frac{f_\theta(\mathbf{z})}{f_{\theta_0}(\mathbf{z})}\ \frac{f_\theta(\mathbf{y} \mid \mathbf{z})}{f_{\theta_0}(\mathbf{y} \mid \mathbf{z})}\ dF_{\theta_0}(\mathbf{z} \mid \mathbf{y}) \tag{3.9}$$

Recall again that, in general, $\mathbf{G}$ denotes any underlying major genotype contribution (which could include linked markers) and $\mathbf{z}$ the polygenic contribution (which may include also common environmental effects, or other heritable Gaussian random effects). The estimates of likelihood ratio, $\hat{l}\,(\theta: \theta_0)$, corresponding to each of the above forms are as follows;

$$\frac{1}{N} \sum_{i=1}^{N} \frac{f_\theta(\mathbf{z}^{(i)})}{f_{\theta_0}(\mathbf{z}^{(i)})}\ \frac{P_\theta(\mathbf{G}^{(i)})}{P_{\theta_0}(\mathbf{G}^{(i)})}\ \frac{f_\theta(\mathbf{y} \mid \mathbf{z}^{(i)}, \mathbf{G}^{(i)})}{f_{\theta_0}(\mathbf{y} \mid \mathbf{z}^{(i)}, \mathbf{G}^{(i)})} \tag{3.10}$$

$$\frac{1}{N} \sum_{i=1}^{N} \frac{P_\theta(\mathbf{G}^{(i)})}{P_{\theta_0}(\mathbf{G}^{(i)})} \; \frac{f_\theta(\mathbf{y} \mid \mathbf{G}^{(i)})}{f_{\theta_0}(\mathbf{y} \mid \mathbf{G}^{(i)})} \tag{3.11}$$

$$\text{and } \frac{1}{N} \sum_{i=1}^{N} \frac{f_\theta(\mathbf{z}^{(i)})}{f_{\theta_0}(\mathbf{z}^{(i)})} \; \frac{f_\theta(\mathbf{y} \mid \mathbf{z}^{(i)})}{f_{\theta_0}(\mathbf{y} \mid \mathbf{z}^{(i)})} \tag{3.12}$$

where the samples $\mathbf{G}^{(i)}$ and/or $\mathbf{z}^{(i)}$ are from the (joint) conditional distribution of $\mathbf{G}$ and $\mathbf{z}$ given the data $\mathbf{y}$. Specifically, in (3.10), Monte Carlo sampling would be effected with respect to (w.r.t) to both $\mathbf{G}$ and $\mathbf{z}$. In (3.11), sampling would be w.r.t $\mathbf{G}$, with polygenic likelihoods $f_\theta(\mathbf{y} \mid \mathbf{G})$ for given $\mathbf{G}$ computed by polygenic peeling. In form (3.12), sampling would be w.r.t $\mathbf{z}$, with major gene likelihoods $f_\theta(\mathbf{y} \mid \mathbf{z})$ computed by Mendelian peeling. In every case, the sampling is conditional on the data $\mathbf{y}$ and (as will be shown below) can be achieved via the Gibbs sampler. Note also that a single set of realisations, generated according to P_{θ_0} can be used to provide estimates of likelihood ratios for any other θ relative to that base point. Estimates at different θ-values are, of course, highly correlated. This ensures that $\hat{l}\,(\theta; \theta_0)$ is, like $l(\theta; \theta_0)$, a smooth function of θ. There is, however, a caution. The estimation will be most effective where θ and θ_0 provide broadly similar distributions; that is, for θ close to θ_0. For a more general discussion of these points see Geyer (1990), and Geyer and Thompson (1991).

Focusing again on the forms (3.4), (3.8) and the estimate (3.11), we see that if the difference between θ and θ_0 is only in the transmission of the major genotype, then $f_\theta(\mathbf{y} \mid \mathbf{G}) = f_{\theta_0}(\mathbf{y} \mid \mathbf{G})$ and only the major genotype configuration enters explicitly into the estimate. Computation of the polygenic likelihood is avoided. Thus in comparing alternative mixed models, comparison within any class for which the random effects (polygenic) component has the same parameter values can be made much more efficiently. Calibration of the likelihood ratios between these subclasses will require many evaluations of polygenic likelihoods, but the number of such inter-class calibrations can be kept to a minimum. This point is also of great importance in the linkage analysis of complex traits (Guo and Thompson, 1991c), where again the main parameter difference of interest is in the transmission of Mendelian genotypes (in that case, their transmission in relation to than of a Mendelian marker, parametrised by a recombination fraction r).

Traditionally, the mixed model likelihood is thought of as a mixture of polygenic likelihoods, but it is equally a multinormal mixture of major locus likelihoods. In this paper, we address the mixed model by sampling the major genotype (and using the estimate (3.11)), but it is equally possible to sample the polygenic values (and use the estimate (3.12)). It is also possible to sample both $\mathbf{G}$ and $\mathbf{z}$, and avoid all consideration of peeling by using the estimate (3.10). This last is presumably the least effective, but there is no universal preference between the other two; the relative efficiencies of the two alternative Monte Carlo views will depend on the relative trait contributions of the two components. Many other aspects of this simulation approach also require further investigation -- for example, the best prescriptions for simulation values of θ_0 in order to obtain an overview of the likelihood surface $L(\theta)$.

4. The Gibbs sampler on pedigrees

An essential component of a Monte-Carlo approach to the likelihood analysis of complex genetic models is the Gibbs sampler, which provides a simple and effective method for sampling from the global conditional probability distribution of genotypes given phenotypes (Sheehan, Possolo, and Thompson, 1989). As shown in the previous section, this ability to sample from this conditional distribution opens the way for a Monte Carlo approach to the computational problems of fitting complex genetic models to data on extended pedigrees. The Gibbs sampler (Hastings, 1970; Geman and Geman, 1984) is becoming widely used in several statistical settings (Gelfand and Smith, 1990).

The idea of the Gibbs sampler relies on the same conditional independence structure of genetic models that is the basis of other methods of computation (see section 1). Specifically, the conditional probability of an individual's genotype, given phenotypic information on any subset of the pedigree, and on the genotypes of all other individuals, depends only on his own phenotype (if observed) and the genotypes of his parents (if present in the genealogy), his offspring (if any) and his offspring's other parents (his "spouses"). The Gibbs sampler on pedigrees is thus specified as follows:

a) Find any genotypic configuration over the pedigree that is consistent with phenotypic observations.

b) Take a random permutation of all individuals in the pedigree.

c) For each individual, in the order indicated by (b), update the genotypic configuration by replacing the genotype of the individual by a realisation from the local conditional distribution given his own phenotype, and current genotypic estimates for his parents, spouses and offspring.

d) Repeat (b) and (c) for as long as required, storing and/or using the configuration achieved (for example) after the 400th. scan, and every subsequent 20 th. scan of the pedigree.

The general theoretical justification for this procedure was developed by Hastings (1970) and Geman and Geman (1984). It has been shown to produce realisations from the global posterior distribution of genotype given phenotype on a pedigree (Sheehan, 1989). Unlike the classical direct simulation approaches, the Gibbs sampler, does not, of course, produce independent realisations.

Sheehan (1989,1990) has investigated the determination of initial configurations, the advantages of random scans versus fixed scans or completely random choice, and other aspects of the procedure. For the current class of problems relating to quantitative traits, determination of a feasible initial configuration presents no problems. All major-genotype combinations are consistent with the data, and all additive genetic values are consistent with each other. In fact, the major-gene allele frequency and founder phenotypes were used to assign founder genotypes, and genes were then "dropped" down the pedigree, with appropriate weighting according to the phenotypic values of observed individuals.

With regard to (d), note that the procedure must be run "long enough" to "forget" the initial configuration before the configuration will be a reasonable realisation from the posterior distribution, and that adjacent realisations will be highly dependent. However, convergence of the distribution occurs quickly. In simple models, dependence between realisations often decays rapidly, although for more complex models with multiple effects there are high cross-correlations between the different variables, as well as high autocorrelation in some cases. In a Monte Carlo estimate such as (3.11), there is a "trade-off" between the level of dependence between successive $\mathbf{G}^{(i)}$ and the amount of computation that must be done using each realisation. Where one is simply "counting" realisations of genotypes, use of every scan is not unreasonable (e.g. Sheehan, 1990). Where substantial computation is involved in the use of each $\mathbf{G}^{(i)}$ as in (3.11), it is inefficient to use very highly dependent realisations, although the dependence does not invalidate the estimate. For a more general discussion see Geyer (1990); for the current example we have used a "dememorisation period of" 400 scans, and thereafter used every 20 th. scan.

5. Application in the evaluation of likelihood ratios for the mixed model.

In this section we consider the combination of the Gibbs sampler (section 4), the Monte Carlo estimation of likelihood ratios (section 3), and the polygenic exact integration ("peeling") algorithm (section 2), in providing a mixed-model analysis of a quantitative trait observed for members of a large pedigree.

The data are simulated. The pedigree of 230 individuals is shown in Figure 2. The true parameter values are given in Table 2. This example is also the one analysed via Monte-Carlo EM procedures by Guo and Thompson (1991b); details of how maximum likelihood estimates are found are given in that paper. The maximum likelihood estimates are also given in Table 2. It is not the purpose of this paper to compare true and MLE parameter values, or to discuss the power of such data to resolve major genes, but rather to show how the methods of previous sections permit the likelihood surface to be explored.

In general there are two aims of likelihood evaluation. One is to monitor the Monte Carlo EM procedure to find MLE's. In that case, the Gibbs samples at current parameter estimates have already been drawn for the EM procedure, and likelihood evaluation can use these samples at no additional sampling cost. These samples will of necessity be small (in our examples, usually 200) since many EM steps will be necessary. However, to explore the likelihood surface in the neighbourhood of the maximum, or of some hypothesis of interest, it is more convenient to choose a fixed θ_0 and generate a substantial sample of realisations to be used in several likelihood ratio evaluations. Other things being equal, the Monte Carlo standard error of course decreases as $N^{-\frac{1}{2}}$ where N is the number of relisations used. In this example, it was found that Monte Carlo samples of size 4000 provided reliable estimates of likelihood ratios. Log-likelihood ratios against various alternative models were evaluated, using the methods of sections 3 and 4: these are also shown in Table 2. We see, not only from the parameter estimates but also from these likelihood ratios, that the presence of a major gene is clear; the log likelihood ratio against the polygenic model is 141.94. Further, the codominant expression of this major gene is also clear -- the log likelihood ratios against a mixed model in which the expression is either dominant or recessive are substantial. The heritable nature of the polygenic background is also apparent; the log likelihood ratio against a model with no additive genetic effects is 19.31. The only component whose presence cannot be clearly inferred is the residual environmental "error"; the true σ_ϵ^2 is small, the maximum likelihood estimate smaller, and the log-likelihood difference against an alternative which does not include this component only 4.35.

Additionally the method was used to explore the likelihood surface for the two variance components, in the neighbourhood of the maximum likelihood estimate. Some results are given in Table 3. Two Gibbs sampler runs were made, each providing 4000 samples, one at the true parameter values and the other at the MLEs, but in each case using the samples to estimate the likelihood ratio relative to θ_0 at the same 9 points -- the presumed MLE and eight neighbours. Since the Monte Carlo estimate will be most accurate for points in the neighbourhood of θ_0, the lower left-hand set of values represents our best estimate of the log likelihood ratios. We see that the point estimate which had been found by the Monte Carlo EM algorithm may not be the MLE -- with this much longer run, there is an indication that smaller values of ${\boldsymbol\sigma_e}^2$ may give slightly higher likelihoods. The values also show the negative relationship between estimates of ${\boldsymbol\sigma_a}^2$ and ${\boldsymbol\sigma_e}^2$; their sum is quite precisely estimated. Also shown in Table 3 (right-hand side) are the log likelihood ratios for the same nine points relative to the truth, from a Gibbs sampler sample of size 4000 run at the true simulation parameter values. All these points have a log-likelihood about 5 greater than the true parameter values, but more importantly the relative likelihoods of these 9 points is reasonably well estimated even by this Gibbs sampler run at parameter values quite different from the MLE. The lower right-hand panel, obtained by taking the values of the upper right-hand panel and subtracting 5.108 (the estimate log likelihood ratio between the MLE and the truth) from each term gives values similar to those from the Gibbs sampler run at the MLE. This provides an indication both of how large a Monte-Carlo sample may be needed to estimate likelihood ratios to a given degree of precision, and of how far from the Gibbs sampler parameter value (θ_0) one can expect to be able to estimate differences in log-likelihood.

6. Linkage analysis for complex genetic traits

So far we have considered only the quantitative trait, but in many cases it will be necessary also to consider linked DNA markers in order to resolve complex genetic traits. One great strength of the current approach is that the likelihood ratio estimation method of section 3 applies equally to any genetic model. To consider linkage analysis, we need only subdivide the major genotype $\mathbf{G}$ into two components $\mathbf{T}$ (for the locus affecting the trait phenotype), and $\mathbf{M}$ (for the DNA marker). Again, a Monte Carlo EM approach will permit likelihood analysis and maximisation of likelihoods (Guo and Thompson, 1991c). Here we focus only on the estimation of the likelihood ratios.

Since the phenotypes $\mathbf{y}$ depend only on the trait genes $\mathbf{T}$, and not on the marker types $\mathbf{M}$, equation (3.11) becomes;

$$\hat{l}\,(\theta\colon\theta_0) = \frac{1}{N}\sum_{i=1}^{N}\frac{P_\theta(\mathbf{T}^{(i)})}{P_{\theta_0}(\mathbf{T}^{(i)})}\,\frac{P_\theta(\mathbf{M}^{(i)}\mid\mathbf{T}^{(i)})}{P_{\theta_0}(\mathbf{M}^{(i)}\mid\mathbf{T}^{(i)})}\,\frac{f_\theta(\mathbf{y}\mid\mathbf{T}^{(i)})}{f_{\theta_0}(\mathbf{y}\mid\mathbf{T}^{(i)})} \tag{6.1}$$

Suppose now we wish to compare models P_θ and P_{θ_0} that differ only in the degree of linkage to the marker. Then equation (6.1) becomes

$$\hat{l}\,(\theta\colon\theta_0) = \frac{1}{N}\sum_{i=1}^{N}\frac{P_\theta(\mathbf{M}^{(i)}\mid\mathbf{T}^{(i)})}{P_{\theta_0}(\mathbf{M}^{(i)}\mid\mathbf{T}^{(i)})} \tag{6.2}$$

Although $(\mathbf{T}^{(i)},\mathbf{M}^{(i)})$ must be generated using the Gibbs sampler for the full set of genetic effects under the model P_{θ_0}, not only are the polygenic values $\mathbf{z}$ not used in the likelihood ratio estimate, but also it is no longer necessary to compute the polygenic likelihoods $f_\theta(\mathbf{y}\mid\mathbf{T})$. and $f_{\theta_0}(\mathbf{y}\mid\mathbf{T})$. Thus the estimate here is particularly easily evaluated.

As previously, choice of θ_0 (and additionally, now, the recombination fraction r_0 used in the Gibbs sampler) is open. In the case of linkage analysis it may be particularly convenient to take θ_0 to be absence of linkage ($r_0 = 0.5$), giving directly the usual LOD score based on the log likelihood ratio for linkage. This widely used linkage LOD score is

$$\log_{10}(L(\theta,r)/L(\theta,\tfrac{1}{2})) = \log_{10} l((\theta,r)\colon(\theta,\tfrac{1}{2})),$$

where r is the recombination fraction between trait locus and marker locus, and θ now includes all other parameters of the genetic model. If there is, in fact, tight linkage, so that r is much less than 0.5, running the Gibbs sampler at $r = 0.5$ will not be useful. Instead, it is then more efficient to use a θ_0 close to the MLE, to investigate the likelihood surface in this range, and to calibrate one such value relative to the base-point of absence of linkage.

Using a simulated data set with the same genetic model as above, we show the LOD-score for various recombination fractions, for the same data, where now a codominant diallelic linked marker is simulated at recombination $r = 0.1$. In fact, the MLE is $r = 0.04$, and the trait model is again well estimated (Table 2). We show two estimated linkage LOD score curves (Figure 3), each obtained from a single run of the Gibbs sampler. Each is based on 4000 Gibbs samples, from a realised chain sampled every 30 scans. One chain is run at the MLE parameter values. The second is run at the same other parameter values, but at the true recombination fraction 0.1. In these LOD score estimates, other genetic parameters are held at their MLE values, and not varied to obtain MLEs corresponding to each value of the recombination fraction. That is, equation (6.2) is used. In this example, the linkage is, of course, very clear. Again the point is not to show that the true linkage is well estimated, but rather that a single Markov Chain Monte Carlo sample provides an entire LOD score curve, and that alternative Markov Chains provide curves that are in close agreement, although not identical.

7. Conclusion

The performance, characteristics, and operational guidelines for the pedigree Gibbs sampler remain to be more fully explored. However, the procedures have been implemented, and shown to provide a feasible method of obtaining estimates of parameters and of likelihoods in the analysis of complex genetic traits. This paper has focussed on the evaluation of likelihood ratios: the papers of Guo and Thompson (1991a,b,c) use realisations from the pedigree Gibbs sampler to obtain MLEs. Previous methods for the computation of mixed-model likelihoods on large pedigrees involve approximations whose effect it has not been possible to assess. The current approach provides an alternative, for this and for other models. It enables us to analyse complex genetic traits on large pedigrees, using appropriately complex models. These models may include a linked marker, making possible the mapping of loci having a major effect on a complex trait.

Acknowledgement

The authors are grateful to Charles J. Geyer for many helpful discussions on the Monte Carlo evaluation of likelihood ratios, and for comments on drafts of this paper. Also to Nuala A. Sheehan for access to her programs which first implemented the pedigree Gibbs sampler for Mendelian loci, and to Qing Yang for his work on the approximation of Normal mixtures. This research was supported in part by NSF grants BSR-8619760 and BSR-8921839, USDA contract 88-37151-3958 (EAT) and NIH grant NHLBI-HL3-0086 (SWG).

Literature cited.

Cannings, C., Thompson, E.A. and Skolnick, M.H. 1976. The recursive derivation of likelihoods on complex pedigrees. *Advances in Applied Probability* **8:** 622-625.

Cannings, C., Thompson, E.A., and Skolnick, M.H. 1978. Probability functions on complex pedigrees. *Advances in Applied Probability* **10,** 26-61.

Cannings, C., Thompson, E.A., and Skolnick, M.H. 1980. Pedigree analysis of complex models. In *Current Developments in Anthropological Genetics* (J. Mielke and M. Crawford, eds.). Plenum Press, New York.

Elston, R.C. and Stewart, J. 1971. A general model for the genetic analysis of pedigree data. *Human Heredity* **21,** 523-542.

Elston, R. C., J. W. MacCluer, S. E. Hodge, M. A. Spence, and R. H. King. 1989. Genetic Analysis Workshop 6: Linkage analysis based on affected pedigree members. In *Multipoint Mapping and Linkage Based on Affected Pedigree Members.* R. C. Elston, M. A. Spence, S. E. Hodge and J. W. MacCluer (eds.). Alan R. Liss, New York. Pp. 93-103.

Gelfand, A. E. and Smith, A. F. M. 1990 Sampling-based approaches to calculating marginal densities. *JASA* **85:** 398-409.

Geman S. and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans on Pattern Analysis and Machine Intelligence* **6:** 721-741.

Geyer, C. J. 1990. *Likelihood and Exponential Families.* Ph. D. Thesis, University of Washington.

Geyer, C. J. and Thompson, E. A. 1991. Constrained Monte Carlo maximum likelihood for dependent data. *J R S S (B)* In press.

Guo, S-W and Thompson, E. A. 1991a. Monte-Carlo estimation of variance component models on large and complex pedigrees. *I M A J Math Appl in Med and Biol,* to appear.

Guo, S-W and Thompson, E. A. 1991b. Monte-Carlo estimation of mixed models on large and complex pedigrees. Submitted.

Guo, S-W and Thompson, E. A. 1991c. Monte-Carlo methods for the linkage analysis of complex genetic traits. In preparation.

Hammersley, J. M. and D. C. Handscomb. 1964. *Monte Carlo Methods.* Methuen and Co.: London.

Hasstedt, S.J. and Cartwright, P. (1979). PAP-Pedigree Analysis Package Technical report No. 13, Department of Medical Biophysics and Computing, University of Utah.

Hasstedt, S.J. (1982). A mixed-model likelihood approximation on large pedigrees. *Computers in Biomedical Research,* **15,** 295-307.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications *Biometrika* **57:** 97-109.

Heuch I. and F. M. F. Li. 1972. PEDIG -- A computer program for calculation of genotype probabilities using phenotypic information. *Clin Genet* **3:** 501-504.

Hilden, J. 1970. GENEX -- An algebraic approach to pedigree probability calculus. *Clin Genet* **1:** 319-348.

Konigsberg, L.W., C. M. Kammerer and J.W. MacCluer. 1989. Segregation analysis of quantitative traits in nuclear families: comparison of three program packages. *Genetic Epidemiology* **6:** 713-726.

Lalouel, J.-M., D. C. Rao, N. E. Morton and R. C. Elston. 1983. A unified model for complex segregation analysis. *Amer J Hum Genet* **35:** 816-826.

Lander, E. S. and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121:** 185-199.

Lauritzen, S.L. and D. J. Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems *(with Discussion). J R S S (B)* **50,** 157-224.

MacCluer, J. W., J. L. VandeBerg, B. Read, and O. A. Ryder. 1986. Pedigree analysis by computer simulation. *Zoo Biology,* **5:** 147-160.

Morton, N.E., Rao, D.C. and Lalouel J-M (1983). Methods in genetic epidemiology. In *Contributions to epidemiology and biostatistics, vol 4.* M. A. Klingbuerg ed., pp 52-102. Karger, Basel.

Ott, J. 1979. Maximum likelihood likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *Amer J Hum Genet,* **31:** 161-175.

Sheehan, N. A., A. Possolo and E. A. Thompson. 1989. Image processing procedures applied to the estimation of genotypes on pedigrees. *Amer J Hum Genet* **45** (Suppl.): A248.

Sheehan, N. A. *Genetic Restoration on Complex Pedigrees.* Ph.D. Thesis, University of Washington.

Thompson, E.A. 1986. *Pedigree Analysis in Human Genetics.* Johns Hopkins University Press, Baltimore, MD.

Tables and Figures.

Table 1; Summary of notation

Table 2; Simulation parameter values and maximum likelihood estimates.

The parameter estimates and likelihood ratios under the heading MLE_1 are those for the simulated data set discussed in section 5. The estimates MLE_2 relate to the linkage example of section 6. The genetic parameters of the trait locus used to obtain these two simulated data sets were the same, and are referred to as the "True" values.

Table 3; Likelihood values in a neighbourhood of the maximum.

Figure captions

Figure 1; The 230-member test pedigree.

The three categories (grey, black and white) denote three subsets of the individuals used in various test examples.

Figure 2; Flow diagram of the zeroloop random-effects peeling program.

Figure 3; LOD score curves for linkage, each estimated from a single Markov Chain Gibbs sampler realisation.

The LOD score is $\log_{10}(L*(r)/L*(0.5))$, where $L*(.)$ is the likelihood function for the recombination fraction r when other parameter values are held fixed at their MLE values. The solid curve is for a Gibbs sampler run at the MLE value 0.04, and the broken curve results from the Gibbs sampler run at the true simulation value for this data set (0.1).

Table 1; Notation summary.

Description of variables or parameter	Notation
a) Variables on the pedigree	
(The subscripted value denotes the value for individual j,	
and the boldface value denotes the set of values over the pedigree.)	
Values of quantitative data observations on individuals	$\mathbf{y}$ and y_j
Genotypes at Mendelian loci relating to the trait	$\mathbf{G}$ and G_j
Contribution of such genotypes to quantitative trait	$\mathbf{g}$ and g_j
Genotypes at a Mendelian locus contributing to the trait	$\mathbf{T}$ and T_j
Genotypes at a linked marker locus	$\mathbf{M}$ and M_j
Polygenic values contributing to the trait	$\mathbf{z}$ and z_j
Dominance or common environmental random effects	$\mathbf{d}$ and d_j
Residual individual environmental random effects	$\mathbf{e}$ and e_j
Fixed (non-heritable) trait contributions	$\boldsymbol{\mu}$ and μ_j
b) Parameters of the genetic models	
Additive genetic variance ($var(z_j)$)	${\sigma_a}^2$
Variance for another heritable/shared random effect ($var(d_j)$)	
(dominance/common sib)	${\sigma_d}^2$
Variance of a maternal line (mitochondrial) effect	$\boldsymbol{\sigma_m}^{\mathbf{2}}$
Variance for residual effects ($var(e_j)$)	${\sigma_e}^2$
Recombination fraction	r
Generic set of all genetic parameters	θ
An alternative set of values for comparison	θ_0
c) Other notation	
Discrete probability distribution at parameter values θ	$P_\theta(.)$
Probability density corresponding to parameter values θ	$f_\theta(.)$
Cumulative probability corresponding to parameter values θ	$F_\theta(.)$
Generic Gaussian quadratic form	$K \exp(-\mathbf{w}^T\mathbf{B}\mathbf{w}/2)$
with vector of variables $\mathbf{w}$, for example	$\mathbf{w}^T = (1, z_m, z_f)$
The number of samples used in Monte Carlo evaluation	N
The values of $\mathbf{G}$ and $\mathbf{z}$ at sample i	$\mathbf{G}^{(i)}$ and $\mathbf{z}^{(i)}$

Table 2; Simulation parameter values and their maximum likelihood estimates.

Parameter		True	MLE_1	MLE_2
Additive variance	σ_a^2	0.6	0.459	0.659
Residual variance	σ_ϵ^2	0.2	0.124	0.145
Major gene frequency	p	0.3	0.292	0.340
Major gene effects:				
homozygote 1	μ_1	2	2.23	2.32
heterozygote	μ_2	0	0.20	0.16
homozygote 2	μ_3	-2	-2.08	-2.11
Marker allele frequency		0.5	---	0.463
Recombination fraction		0.1	---	0.038
Lod score				10.30
log likelihood ratios of MLE_1:				
vs. polygenic; additive + error			141.94	
vs. error component only			489.29	
vs. additive component only			379.16	
vs. dominant; mixed model			72.72	
vs. recessive; mixed model			214.76	
vs. major gene + error			19.31	
vs. major gene + additive			4.35	

Table 3; Likelihood values in a neighbourhood of the maximum.

Gibbs sampler run at		a) putative MLEs			b) "true" simulation values		
σ_a^2		0.412	0.459	0.505	0.412	0.459	0.505
σ_e^2	0.137				4.862	4.976	4.765
	0.124				4.861	5.108	4.962
	0.112				4.772	5.138	5.109
σ_e^2	0.137	-0.249	-0.141	-0.424	-0.246	-0.132	-0.343
	0.124	-0.265	0	-0.167	-0.247	0	-0.146
	0.112	-0.400	0.052	0.022	-0.336	0.030	0.001

Table 4; LOD scores for linkage

Recombination fraction for Gibbs sampler run at	LOD score = $\log_{10}(L*(r)/L*(0.5))$ $r = 0.04$ (MLE)	$r = 0.1$ (truth)
0.001	9.60	8.75
0.010	10.21	9.67
0.020	10.33	9.87
0.030	10.33	9.93
0.040	10.28	9.94
0.050	10.21	9.90
0.100	9.64	9.46
0.200	8.21	7.96
0.300	6.29	5.98
0.400	3.58	3.38

Note: $L*(.)$ here denotes the profile likelihood for the recombination fraction r when other parameter values are held fixed at their MLE values.

Figure 2: Flow diagram for simple peeling of zeroloop pedigree with dominance or common-sib effects.

Set parameter values; $(\mu, \sigma^2, \tau^2, \epsilon^2)$
Read pedigree ; size, (individual, mother, father, data observation) for each individual.
Initialise storage and data indicators for each individual.
FOR each sibship of the pedigree in turn;
    Initialise the matrix $\mathbf{B} = \mathbf{0}$ of quadratic form coefficients
    Initialise constant multiplier $K = 1$
    Read the (mum, dad, number of kids, pivot individual) for the current sibship
    Read the names (identifiers) of each offspring in turn
    Set the main vector of effects $(1, a_m, a_f, d)$ for this sibship.
    For each non-pivot offspring;
        Add component to current vector for this offspring's additive effect.
        Incorporate segregation for this offspring (equation 2.7) updating $\mathbf{B}$ and K
        Check for any previously stored likelihood on this offspring individual
            If any, incorporate it into $\mathbf{B}$ and K.
        Incorporate data if offspring is observed; equation 2.8
        Peel this offspring (i.e. integrate out his additive effect)
    IF all offspring are peeled (i.e. pivot is a parent)
        Incorporate prior Gausssian distribution for d
        Integrate out the common-sib effect.
        Incorporate non-pivot parent:
            Incorporate previous stored likelihood, if any
            If no previous likelihood, incorporate prior distribution for additive effect
                and data, if individual observed (note this contribution incorporates
                    common-sib variance for a founder parent)
        Integrate out additive effect for non-pivot parent.
    OR IF pivot is remaining unpeeled child
        Incorporate data, previous/founder likelihoods on parents
        Integrate out parental additive effects.
        Incorporate data on offspring pivot, if any
        Incorporate prior on shared common-sib effects
        Integrate out common-sib effect.
        IF new pivot has previously stored likelihood, incorporate it.
        (Re)store likelihood for the new pivot, in the form of a 2x2 matrix $\mathbf{B}$ and constant K.
    OR IF no pivot; This was final sibship.
End of FOR loop; (Now all marriages are peeled; all additive effects are integrated out.)
Check all stored likelihoods have been used once and once only.
Print final likelihood.
STOP