PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Monte Carlo Likelihood in the Genetic Mapping of Complex Traits [and Discussion]

E. A. Thompson and C. A. B. Smith

Phil. Trans. R. Soc. Lond. B 1994 **344**, 345-351
doi: 10.1098/rstb.1994.0073

Monte Carlo likelihood in the genetic mapping of complex traits

E. A. THOMPSON

Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195, U.S.A.

SUMMARY

Many of the likelihoods arising in the analysis of complex genetic traits, particularly in linkage analysis, are computationally infeasible. Where exact likelihoods cannot be computed, Monte Carlo estimates of likelihoods may provide a satisfactory alternative. Although simulation on pedigrees is straightforward, simulation conditional upon observed phenotypic data is not. However, recent advances in Markov chain Monte Carlo methods have provided a method well suited to this problem. From realizations of underlying genes, simulated under a genetic model, conditional upon observed data, a Monte Carlo estimate of this likelihood surface can be formed. Various sampler and model modifications are needed to enhance the statistical efficiency of the Monte Carlo estimator; as these methods become increasingly developed, this approach becomes a useful tool in resolving the genes contributing to the phenotypes associated with genetically complex diseases.

1. INTRODUCTION

In statistical analyses of the genetic epidemiology of complex traits a major limitation has been practical and theoretical bounds on computational feasibility of likelihood evaluation. In some quite standard applications, a single run on the LINKAGE program (Lathrop *et al.* 1984) may take several months (Schellenberg *et al.* 1992). For an ongoing study, with continuing data collection, this is not acceptable. In other potential applications, it can be shown that exact evaluation of the likelihood would take millions of years, even if computer speeds continued to multiply at the same rate as over the past four decades. Potential solutions include improvement of the programs, improvement of the computational algorithms, or a radically different approach to likelihood assessment. Recently Cottingham *et al.* (1993) have shown that fairly standard computer science programming procedures can improve performance of the LINKAGE program by an order of magnitude. However, the immediate response of practitioners is to wish to address a tenfold larger problem. With the current algorithms, based on the method of Elston & Stewart (1971), computer times increase exponentially with pedigree complexity, numbers of alleles and numbers of loci modelled; computer science cannot keep up with increases in genetic complexity of models for traits and DNA markers.

Owing to the infeasibility or impracticality of exact likelihood computation, approximate methods are often used. One of the most long-standing is that of Hasstedt (1982) for mixed models; another is that of Hoeschele *et al.* (1987) for ordered categorical data determined by an underlying quantitative liability. Most recently Curtis & Gurling (1993) have proposed an approximation to multilocus linkage likelihoods using a combination of pairwise log-likelihoods. For practical purposes, these methods may be excellent, but without some method of exact computation this is impossible to assess. One approach which provides such an assessment is Monte Carlo likelihood, in which exact evaluation of likelihoods and likelihood ratios is replaced by a Monte Carlo estimate. For data analysis, Monte Carlo methods may not provide a practical solution, but for assessment of alternative approximations they are ideal, as, provided the Monte Carlo experiment is run for long enough, the exact likelihood can be computed to an arbitrary degree of accuracy.

Monte Carlo estimates of integrals or expectations are not new, either in general (Hammersley & Handscomb 1964) or in genetic linkage analysis (Thompson *et al.* 1978), but in Monte Carlo likelihood approaches the problem is complicated by the fact that realizations are required from a probability distribution with unknown normalizing constant, this normalizing constant being precisely the required likelihood. Thus methods of Markov chain Monte Carlo (Hastings 1970) are applicable; their application in genetic analysis raises many interesting statistical questions. Note that these statistical questions addressed here are not those of the likelihood estimation of genetic models from data. Rather they are the statistical properties of the Monte Carlo procedures used to estimate the likelihood function.

2. MONTE CARLO LIKELIHOOD IN MISSING DATA PROBLEMS

The statistical problems involved in fitting genetic linkage models to trait data, $\mathbf{Y}$, may be viewed as latent variable or 'missing data' problems. Were the underlying haplotypes (multilocus genotypes) of all individuals observable, estimation would be trivial, but only the trait data (phenotypes) and single-locus marker genotypes of some individuals are observed. We denote the observed trait and marker data on a set of related individuals by $\mathbf{Y}$, additional latent variables by $\mathbf{X}$, and the genetic parameters (recombination fractions, penetrances, etc.) by θ. The likelihood is

$$L(\theta) = P_\theta(\mathbf{Y}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y},\mathbf{X}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y}|\mathbf{X}) P_\theta(\mathbf{X}). \quad (1)$$

Although the summation may be infeasible, the latent variables $\mathbf{X}$ are to be chosen in such a way that each term of the expression is easily computed. In fact, $\mathbf{X}$ may contain both discrete and continuous variables, but for simplicity of notation we restrict to the case of a discrete sum. Now

$$P_\theta(\mathbf{X}|\mathbf{Y}) = \frac{P_\theta(\mathbf{Y},\mathbf{X})}{P_\theta(\mathbf{Y})}, \quad (2)$$

and (Thompson & Guo 1991)

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = E_{\theta_0}\left(\frac{P_\theta(\mathbf{Y},\mathbf{X})}{P_{\theta_0}(\mathbf{Y},\mathbf{X})}\middle|\mathbf{Y}\right). \quad (3)$$

As the probability distribution (2) is known up to the normalizing factor $L(\theta) = P_\theta(\mathbf{Y})$ Markov chain Monte Carlo provides a method for providing N dependent realizations $\mathbf{X}^{(l)}$, $l = 1, \dots, N$, from $P_{\theta_0}(\cdot|\mathbf{Y})$ which may be used in a Monte Carlo estimator of the likelihood ratio (3):

$$\frac{1}{N}\sum_{l=1}^{N}\left(\frac{P_\theta(\mathbf{Y},\mathbf{X}^{(l)})}{P_{\theta_0}(\mathbf{Y},\mathbf{X}^{(l)})}\right). \quad (4)$$

This estimator (4) works well if θ is close to θ_0, but, in comparing alternative genetic models, it is rarely the local characteristics of the likelihood surface that are of interest. Rather geneticists will with to compare best fitting models within each broad class. These models are widely spaced in the hypothesis space in terms of the induced probability distribution on the latent variables; a chain run at either hypothesis will provide a very poor estimate of the likelihood ratio relative to the other.

To overcome this problem, chains may be run at a set of parameter combinations, $\theta_0, \theta_1, \theta_2, \dots, \theta_K$, spanning the range between the two hypotheses of interest, θ_0 and θ_K. Whereas realizations from chain θ_j could simply be used to estimate adjacent log-likelihood differences $L(\theta_{j+1})/L(\theta_j)$ or $L(\theta_j)/L(\theta_{j-1})$, this would be wasteful. An importance sampling approach allows one to use all the samples in a combined estimate of log likelihood differences along the chain $\theta_0, \theta_1, \theta_2, \dots, \theta_K$ (Geyer 1991*a*). An intuitive way to view this approach is as follows. Assume N_j realizations are taken from chain $P_{\theta_j}(\cdot|\mathbf{Y})$. Rather than retaining the chain parameter value corresponding to each realization, the collection of realizations are 'pooled', and the pooled sample is regarded as a sample size $\sum_j N_j$ from the weighted average of the distributions indexed by $\theta_0, \dots, \theta_K$, that is

$$\frac{1}{\sum_j N_j}\sum_{j=0}^{K} N_j P_{\theta_j}(\mathbf{Y},\mathbf{X}) \exp(\nu_j),$$

where $\nu_j = -\log L(\theta_j)$. Then we have an auxiliary likelihood problem in which the unknown 'parameters' ν_j are estimated by maximum likelihood, given the sample from this 'mixture distribution', or equivalently by solving the equations

$$\exp(-\nu_j) = \sum_{\mathbf{X}^*}\left(\frac{P_{\theta_j}(\mathbf{Y},\mathbf{X}^*)}{\sum_{l=0}^{k} N_l P_{\theta_l}(\mathbf{Y},\mathbf{X}^*)\exp(\nu_l)}\right) \quad \text{for} \quad j = 0, \dots, K, \quad (5)$$

where the summation is over the total combined sample of realizations $\mathbf{X}^*$. (These equations determine the log-likelihoods ν_j only up to an additive constant, so log-likelihood differences $\nu_j - \nu_0$, $j = 1, \dots, K$ are estimated.) If this procedure is implemented every realization contributes to the estimate of $L(\theta_j) = \exp(-\nu_j)$ for all j in accordance with the appropriate importance sampling weights (Geyer 1991*a*).

3. SAMPLING GENES ON PEDIGREES

(*a*) *Gene drop and gene lift*

Simulation on pedigrees pre-dates digital computers (Wright & McPhee 1925), but the increasing power of computing makes it a more useful proposition. Simulation of the genes descending a pedigree has been used in many studies, and is easily done. Genes are assigned to the founders of the pedigree, segregation of genes down the pedigree is simulated, and the required statistics relating to the resultant current genes are computed. This approach has been used to provide estimates of structural parameters such as inbreeding coefficients (Edwards 1967), or to investigate gene loss in endangered species (MacCluer *et al.* 1986).

However, simulation in the presence of genetic data on current individuals is far harder. The number of possible genotypic configurations on a pedigree is immense, and the proportion that are compatible with data observed on individuals is minute. Normally data are observed on the final members of a pedigree, the current individuals, and the ancestors are unobserved. Simulating the descent of genes from these ancestors to the current individuals will very seldom produce a genotypic configuration compatible with current data: simple-minded rejection sampling is useless.

An alternative is to attempt 'gene lift', simulating backwards from current data, according to some probabilities for parents conditional upon offspring, and then reweight to the true probabilities under the genetic model, using importance sampling. This is seldom successful in large problems. If there are multiple alleles, incompatibilities will arise and the simulation cannot proceed. Even for the simulation of a rare recessive allele, for example, where there will be no incompatibility problems, gene lift does not work. The failure of the algorithm to be able to look at distant ancestry, and so bring together the multiple descendant copies of a rare allele, means that realisations of 'gene lift' are very far removed from those having non-negligible probability under the genetic model.

(*b*) *Markov chain genotype updating*

Alternatives to 'gene drop' and 'gene lift' are 'gene updating' samplers. We review briefly the Metropolis-Hastings class of algorithms (Hastings 1970) for generating dependent realizations from a probability distribution $P_\theta(\mathbf{X})$ on a space $\mathscr{X}$, where $P_\theta(\cdot)$ may be known only up to a normalizing constant. For each $\mathbf{X}$ in $\mathscr{X}$ a 'proposal distribution' $q(\cdot, \mathbf{X})$ is defined. Then, if the process is now at $\mathbf{X}$ the next value is generated as follows:

1. Generate $\mathbf{X}^*$ from the proposal distribution $q(\cdot, \mathbf{X})$.
2. Compute the Hastings ratio

$$h = \frac{q(\mathbf{X}, \mathbf{X}^*) P_\theta(\mathbf{X}^*)}{q(\mathbf{X}^*, \mathbf{X}) P_\theta(\mathbf{X})}.$$

Note that h depends only on the ratio of densities $P_\theta(\cdot)$, so that any normalizing constant need not be computed.

3. With probability $r = min(1, h)$ the process moves to $\mathbf{X}^*$ and with probability $(1 - r)$ it remains at $\mathbf{X}$.

The algorithm of Metropolis *et al.* (1953) is a special case; if $q(\mathbf{X}^*, \mathbf{X}) = q(\mathbf{X}, \mathbf{X}^*)$ the Hastings ratio reduces to the odds ratio of the proposal state $\mathbf{X}^*$ versus the current state $\mathbf{X}$. The Gibbs sampler (Geman & Geman 1984) is also a special case, in which $\mathbf{X}^*$ differs from $\mathbf{X}$ in only one component X_i say, and X_i has the distribution $P_\theta(X_i|\mathbf{X}_{-i})$ where $\mathbf{X}_{-i}$ denotes the components of $\mathbf{X}$ other than X_i. In this case $r = h = 1$; in the Gibbs sampler there is no rejection step, but steps are necessarily small, with only one component of $\mathbf{X}$ being changed at each step.

Returning to the missing-data likelihood formulation, we require realisations from $P_\theta(\mathbf{X}|\mathbf{Y})$, a distribution known up to the normalizing constant $P_\theta(\mathbf{Y})$. Unfortunately, in genetic examples the constraints on $\mathbf{X}$ imposed by Mendelian segregation mean that any attempted proposal distribution that makes multiple changes to the current value of $\mathbf{X}$ is likely to result in a configuration with zero probability. However, one-variable-at-a-time updating methods, such as the Gibbs sampler, are well suited to this sampling problem, at least in proposing feasible configurations. The genes and heritable effects in an individual are determined by those in his parents, and jointly with those in his spouse, influence those in his offspring. This neighbourhood structure means that the Gibbs sampler is easy to implement; each genetic effect in each individual is successively updated, conditional upon the remainder.

4. IMPROVED MODELS AND SAMPLES

(*a*) *Burn-in problems with the Gibbs sampler*

The Gibbs sampler for underlying genotypes works well in simple examples, but can run into serious problems. If phenotypes rather closely define underlying genotypes of sampled individuals, the total space of genotypic configurations on all individuals can be difficult to sample efficiently. An example is given by the mixed model for cholesterol levels on a 232-member pedigree (Thompson *et al.* 1993). The model is that cholesterol levels ($\mathbf{Y}$) depend additively on effects ($\boldsymbol{\mu}$) due to a segregating major gene (genotypes $\mathbf{G}$), on additional polygenic heritable effects ($\mathbf{Z}$), and independent residuals ($\mathbf{e}$). Fixed effects ($\boldsymbol{\gamma}$) due to covariates ($\mathbf{f}$) such as age and sex can also be incorporated as fixed effects. Thus, the model is

$$\mathbf{Y} = \boldsymbol{\gamma}(\mathbf{f}) + \boldsymbol{\mu}(\mathbf{G})\mathbf{Z} + \mathbf{e}. \tag{6}$$

In sampling, the latent variables are taken as the major genotype and the polygenic value of each number of the pedigree.

Provided the initial latent variable configuration $(\mathbf{G}^{(0)}, \mathbf{Z}^{(0)})$, belongs to a set M of configurations in which certain individuals with very high observed cholesterol values do indeed carry the gene for high cholesterol reliable results are obtained. These configurations have a joint log-probability, $\log P_{\theta_0}(\mathbf{Y}, \mathbf{G}, \mathbf{Z})$, between -1802 and -1812. There are also a large number of configurations, B, with $\log P_{\theta_0}(\mathbf{Y}, \mathbf{G}, \mathbf{Z})$ between -1950 and -1970, and these give a quite different picture of the log-likelihood differences between alternative models. Moreover, if the chain is started in B it can take as many as 1 200 000 pedigree scans before, quite suddenly over a period of less than 30 pedigree scans, $\log P_{\theta_0}(\mathbf{Y}, \mathbf{G}, \mathbf{Z})$ will increase by about 150, and $(\mathbf{G}, \mathbf{Z})$ reaches M (figure 1). On the other hand, if started in M, the process has not, in tens of millions of scans, left M. Thus the total weight of B, each realization

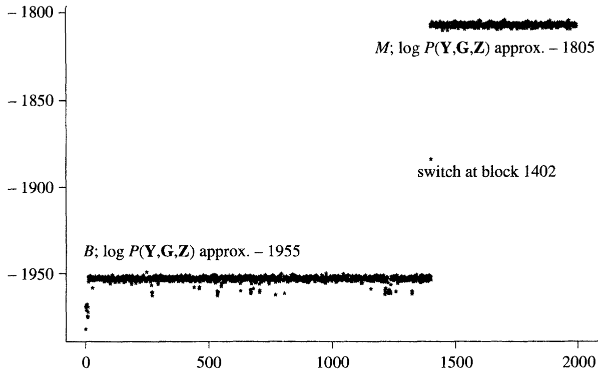

Figure 1. Block means of $\log P_{\theta_0}$ $(\mathbf{Y}, \mathbf{G}, \mathbf{Z})$ under the Gibbs sampler, when the initial configuration is chosen poorly. Each point is a mean over 200 realizations, with realizations taken at intervals of 4 pedigree scans.

contributing only e^{-150} times as much as a single point in M, must be negligible.

In general, the problem is not so easily solved. There are diagnostics for being in a low-density part of the space. For example, one indication is that realizations deriving from chain run at parameter values θ_j have a higher posterior probabilities of deriving from the distribution indexed by some other θ_{j*} than by θ_j. Another is that, as runs get longer, the batch-mean autocorrelation diagnostic statistics become more extreme. However, even if the existence of a region B is diagnosed, it must still be validated that any region such as B, in which the probability density may be very much smaller, does not contain a significant part of the total probability mass.

(*b*) *Choice of latent variables*

There are two approaches to obtaining more precise Monte Carlo likelihood ratio estimates. One involves better samples, the other involves changing the model framework. It seems natural to take all individual genotypes and heritable effects as the latent variables in a complex genetic model (Guo & Thompson 1992). However, this may not be statistically effective. In some cases, the joint probability of data $\mathbf{Y}$ and latent variables may be computable for a subset of variables $\mathbf{X}$, say $\mathbf{X}_{(1)}$, where $\mathbf{X} = (\mathbf{X}_{(1)}, \mathbf{X}_{(2)})$. That is, it may be possible to integrate or sum analytically over the variables $\mathbf{X}_{(2)}$. In this case, although a sampler generating the full $\mathbf{X}$ values from a chain with equilibrium distribution $P_\theta(\mathbf{X}|\mathbf{Y})$ may be employed, it is possible to use only the $\mathbf{X}_{(1)}$ values in the Monte Carlo estimate of likelihood (4) or from (5). Examples of this discussed in Thompson (1994*a*); note, however, that as the generated $\mathbf{X}$ values are dependent, there is no guarantee that the exact integration of some variables reduces the Monte Carlo variance of the estimated likelihood surface. An alternative approach is to change the latent variables; for example, Thompson (1994*b*) uses as latent variables only the indicators of grandparental origins of genes, for each individual at each locus. In fact, Lange & Matthysse (1989) use both genotypes and grandparental indicators in their Metropolis formulation; thus use only of genotypes as in Guo & Thompson (1992) could be regarded as having integrated over grandparental gene origins, while use only of gene origins as in Thompson (1994*b*) requires integration over genotypes.

(*c*) *Improved samplers for genetic problems*

An alternative approach to more efficient Monte Carlo likelihood estimation is to construct samplers which sample the space of genotypic configurations more effectively than does the Gibbs sampler. The constraints of Mendelian genetics limit the proposal distributions that can be usefully employed in the Metropolis-Hastings algorithm. Further, constraints on the feasible genotypic configurations can lead to failure of irreducibility of the Gibbs sampler. Even where irreducible, the Gibbs sampler may be impractical, as the Markov chain can be very poorly mixing. Sheehan & Thomas (1993) address the reducibility problem by modifying zeros in either the genotype transmission probabilities or the genotype–phenotype correspondence (penetrances) and using importance weighting (in fact with zero/one weights) to obtain realizations from the correct conditional distribution of genotypes given the data. Lin (1993) showed how penetrance modifications could be limited and made individual-specific. Earlier, Geyer (1991*b*) had proposed Metropolis-coupled samplers, and Lin (1993) proposed coupling a Gibbs sampler for the true genetic model to one for the penetrance-modified model, providing an irreducible sampler with the correct equilibrium distribution, without any reweighting being required. Although resolving problems of reduciblity, these samples remain impractical on large pedigrees.

Lin (1993) also proposed a class of proposal probabilities proportional to $1/T$ powers of the local conditional probabilities used in the Gibbs sampler, where T is a 'temperature' parameter. Lin *et al.* (1993) use Metropolis coupled samplers at varying temperatures and with modified penetrance matrices, coupled with a Gibbs sampler for the true model. Using the three ideas of coupled samplers, individual-specific and genotype-specific penetrance modifications, and 'high temperature' Metropolis-Hastings proposal distributions, the speed of sampling the space of genotypic configurations is greatly enhanced, particularly where genetic marker loci have several alleles.

Geyer & Thompson (1993) use a different form of penetrance modification and coupling of samplers to sample the genotypic configurations underlying the cystic fibrosis trait on a 2024 member pedigree. Instead of 'swapping' configurations between samplers with the appropriate Metropolis-Hastings acceptance probabilities, the single chain of realizations 'jumps' from sampler to sampler, in a procedure akin to the simulated tempering procedure proposed by Marinari & Parisi (1992). The distributions of each sampler form a sequence from a degenerate sampler providing regeneration points to the Gibbs sampler for the true model. Intermediate samplers were defined by varying the penetrance probabilities, the influence of the observed data increasing from none at the degenerate 'hot' chain to complete at the true model.

These ideas of Geyer (1991*b*), Lin (1993) and Geyer & Thompson (1993) provide samplers that are far more effective than the simple Gibbs sampler for problems of sampling genotypic configurations on large pedigrees with data observed on some individuals. Their ideas have not been implemented in the examples described in this paper, but many of them could be combined with alternative choices of latent variables, or with the model framework described below, to improve further the performance of Monte Carlo likelihood ratio estimators.

(*d*) *A class of fractional penetrance models*

An alternative modification of genetic models lies in the idea of fractional contributions of genes to traits.

This also is a form of penetrance modification, related to those proposed by Lin *et al.* (1993) and by Geyer & Thompson (1993) to obtain valid and effective samplers for otherwise intractable genetic models on large and complex pedigrees. However, here it relates rather to defining a sequence of models to use for likelihood ratio estimation between disparate end-points as in (5). Consider again the simple additive mixed model for a quantitative genetic trait (6). Having estimated the parameters within this class of models, we may then wish to compare with a model without major-gene effects, or without polygenic effects.

To do so we need a set of models at which to sample, connecting the estimated mixed model to either of these two nested extremes. One way to do this is to change the parameters underlying the probability distributions of **G** and/or **Z**. However, this leads to severe problems in sampling and in importance weighting estimates, as **G** and **Z** values generated under one model may have infinitesimal probability under an adjacent one, particularly at the endpoints. For example, one can eliminate polygenic effects by setting the prior variance of the **Z** values to zero, but then any non-zero **Z** values generated under another model are impossible under this endpoint. Likewise, one way to eliminate major gene effects, is to give all individuals the same genotype with probability one, by setting one allele frequency to one, but then any realizations in which other alleles are present have zero probability.

Fractional contributions of genes provides a more successful solution. Rather than modifying the prior distributions of **G** and **Z** the relation to phenotype is modified to

$$\mathbf{Y} = \boldsymbol{\mu}(\mathbf{f}) + \lambda_1\boldsymbol{\mu}(\mathbf{G}) + \lambda_2\mathbf{Z} + \mathbf{e},$$

where λ_1 and λ_2 range from 0 to 1. Thus a sequence of models is defined by varying λ_1 and/or λ_2. The only parameter of the genetic model that is altered is the variance of the residuals **e** which is modified to maintain a constant marginal variance of each component of **Y** as the λ-values change. Only $P_{\theta_j}(\mathbf{Y}|\mathbf{G},\mathbf{Z})$ contributes to the estimation of likelihood ratios since the realized **G** and **Z** have equal prior probability under all models in the set.

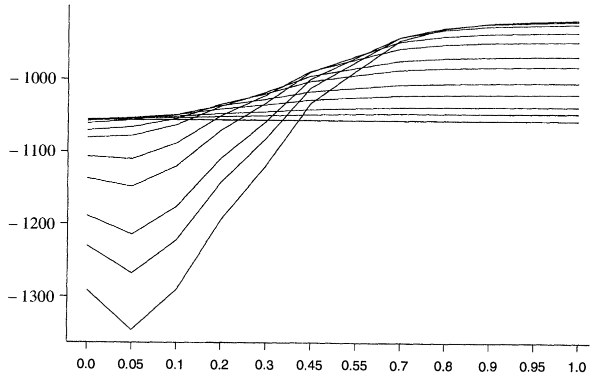

Figure 2. Mean values of $\log P_{\theta_j}(\mathbf{Y}|\mathbf{G})$, $j = 0, \ldots, 11$ for samples **G** obtained using the Gibbs sampler and the fractional penetrance model with $\lambda_2 = 1$ and the 12 x-axis values of λ_1. (Latent variables **Z** are integrated analytically in this example.)

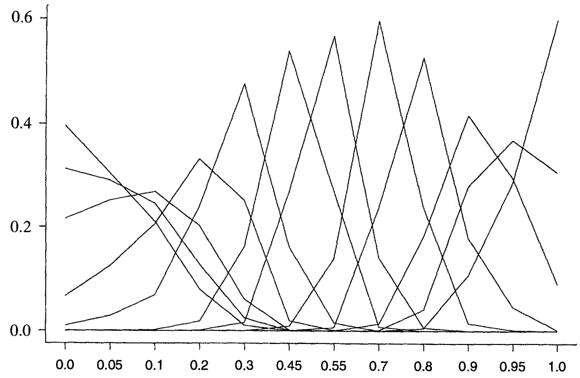

Figure 3. For the same samples as for figure 2, mean posterior probabilities that realizations **G** obtained under the Gibbs sampler for $P_{\theta_j}(\cdot|\mathbf{Y}), j = 0, \ldots, 11$ derive from $P_{\theta_{j^*}}(\cdot|\mathbf{Y}), j^* = 0, \ldots, 11$. The example shows good probability overlap between successive samplers.

In tests, this fractional penetrance model has proven quite successful. Returning again to the cholesterol pedigree of Thompson *et al.* (1993), varying λ_2 with $\lambda_1 = 1$ has very efficiently confirmed previous results on the log-likelihood difference between the mixed model maximum likelihood estimate and major gene models. Moreover, varying λ_1 with $\lambda_2 = 1$ has provided accurate estimates of the much larger log-likelihood difference between the mixed model and polygenic models. For this case, figure 2 shows the mean contributions $\log P_{\theta_{j^*}}(\mathbf{Y}|\mathbf{G})$ for samples **G** obtained from each sampler $P_{\theta_j}(\cdot|\mathbf{Y})$ while figure 3 shows, for the same sets of realizations, the mean posterior probabilities that realizations **G** obtained from each sampler $P_{\theta_j}(\cdot|\mathbf{Y})$ derive from the sampler $P_{\theta_{j^*}}(\cdot|\mathbf{Y})$. (In this example, $N_j = 2000$ for each of the 12 samplers, and the chosen values of λ_1 label the 12 sets of realizations in figures 2 and 3.) Finally, in cases where the pedigree is too complex to obtain base-point log-likelihoods of either major-gene or polygenic models and where no latent variables can be integrated analytically, varying $\lambda_1 = \lambda_2$ will provide a Monte Carlo estimate of the log-likelihood relative to a pure environmental model.

5. CONCLUSION

This paper has demonstrated the scope for Monte Carlo likelihood in the mapping of complex traits, where exact likelihood computation is often infeasible. Much remains to be explored concerning the statistical properties of Monte Carlo likelihood surfaces, but some practical concerns and solutions are presented here. Methods to improve performance include the alternative specification of latent variables, better samplers, and alternative definition of intermediate models. These approaches are not mutually exclusive. For example, a fractional penetrance sequence of models can be easily combined with the simulated tempering approach of Geyer & Thompson (1993). Using methods in combination, it seems that effective methods can be found for previously intractable problems.

I am grateful to Charles Geyer for many helpful discussions on Monte Carlo likelihood functions and on 'gene lift'. Research supported in part by NIH grant GM-046255.

REFERENCES

Cottingham, R.W., Idury, R.M. & Schaffer, A.A. 1993 Faster sequential genetic linkage computations. *Am. J. Hum. Genet.* **53**, 252–263.

Curtis, D. & Gurling, H. 1993 A procedure for combining two-point lod scores into a summary multipoint map. *Hum. Hered.* **43**, 173–185.

Edwards, A.W.F. 1967 Automatic construction of genealogies from phenotypic information (AUTOKIN). *Bull. Eur. Soc. Hum. Genet.* **1**, 42–43.

Elston, R.C. & Stewart, J. 1971 A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542.

Geman, S. & Geman, D. 1984 Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Analysis Mach. Intell.* **6**, 721–741.

Geyer, C.J. 1991*a* Reweighting Monte Carlo Mixtures. Technical Report No. 568, School of Statistics, University of Minnesota.

Geyer, C.J. 1991*b* Markov chain Monte Carlo maximum likelihood. In *Computing science and statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156–163. Interface Foundation of North America.

Geyer, C.J. & Thompson, E.A. 1993 Annealing Markov chain Monte Carlo with applications to pedigree analysis. Technical Report No. 589, School of Statistics, University of Minnesota.

Guo, S.W. & Thompson, E.A. 1992 A Monte Carlo method for combined segregation and linkage analysis. *Am. J. Hum. Genet.* **51**, 1111-1126.

Hammersley, J.M. & Handscomb, D.C. 1964 *Monte Carlo methods*. Methuen & Co., London.

Hasstedt, S.J. 1982 A mixed-model likelihood approximation on large pedigrees. *Comput. Biomed. Res.* **15**, 295–307.

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

Hoeschele, I., Gianola, D. & Foulley, J.L. 1987 Estimation of variance components with quasi-continuous data using Bayesian methods. *J. Anim. Breed. Genet.* **104**, 334.

Lange, K. & Matthysse, S. 1989 Simulation of pedigree genotypes by random walks. *Am. J. Hum. Genet.* **45**, 959–970.

Lathrop, G.M., Lalouel, J.-M., Julier, C. & Ott, J. 1984 Strategies for multilocus linkage analysis in humans. *Proc. natn. Acad. Sci. U.S.A.* **81**, 3443–3446.

Lin, S. 1993 Markov chain Monte Carlo estimates of probabilities on complex structures. Ph.D. thesis, University of Washington.

Lin, S., Thompson, E.A. & Wijsman, E.M. 1993 Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. *IMAJ Math. appl. Med. Biol.* **10**, 1–17.

MacCluer, J.W., VandeBerg, J.L., Read, B. & Ryder, O.A. 1986 Pedigree Analysis by Computer Simulation. *Zoo Biol.* **5**, 147–160.

Marinari, E. & Parisi, G. 1992 Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19**, 451–458.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. 1953 Equations of State Calculations by Fast Computing Machines. *J. chem. Phys.* **21**, 1087–1092.

Schellenberg, G.D., Bird, T.D., Wijsman, E.M. *et al.* 1992 Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science, Wash.* **258**, 668–671.

Sheehan, N.A. & Thomas, A.W. 1993 On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49**, 163–175.

Thompson, E.A., Kravitz, K., Hill, J. & Skolnick, M.H. 1978 Linkage and the power of a pedigree structure. In *Genetic epidemiology* (ed. N. E. Morton), pp. 247–253. New York: Academic Press.

Thompson, E.A. & Guo, S.W. 1991 Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. appl. Med. Biol.* **8**, 149–169.

Thompson, E.A., Lin, S., Olshen, A.B. & Wijsman, E.M. 1993 Monte Carlo analysis on a large pedigree. In *Genetic Analysis Workshop 8. Genetic Epidemiology*. **10,** 677–682.

Thompson, E.A. 1994*a* Monte Carlo likelihood in genetic analysis. In *Probability, statistics, optimization* (ed. F. P. Kelly). New York: Wiley. (In the press.)

Thompson, E.A. 1994*b* Monte Carlo likelihood in linkage analysis. *Stat. Sci.* (In the press.)

Wright, S. & McPhee, H.C. 1925 An approximate methods of calculating coefficients of inbreeding and relationship from livestock pedigrees. *J. agric. Res.* **31**, 377–383.

Discussion

C. A. B. Smith (*University College London, U.K.*). Professor Adrian Smith uses a Monte Carlo method with Gibbs sampler for multilocus linkages estimation. Is his method related to that used by Professor Thompson?

E. A. Thompson. Professor C. A. B. Smith and others raised the question of the relationship of this paper to the Monte Carlo approaches to Pedigree Analysis of Ott (1989), Stephens & Smith (1993), and Kong *et al.* (1992). Ott (1989) and Ploughman & Boehnke (1989) both developed methods for simulating marker data conditionally upon trait data, but only for trait models for which exact probabilities can be computed on the pedigree. Their objective was estimation of the power of a potential linkage study, conditional upon observed trait data. That also was the focus of Lange & Matthyse (1989), whose Metropolis Markov chain Monte Carlo approach is much more closely related to that presented here. Kong *et al.* (1992) is closer in objective to this paper, and presents an alternative approach to Monte Carlo estimation of a likelihood ratio function based on independent realizations and importance sampling. Conversely, the Gibbs sampler methods of Stephens & Smith (1993) is an MCMC approach but on a different state space. There the sample space of the MCMC contains also the parameters of the genetic model, and a marginal posterior probability distribution for genetic linkage parameters is estimated.

Another question raised in discussion concerns the interpretation of the latent variables **Z** in the genetic model of this paper. Although the model used was that of the classical additive genetic (polygenic) effects, it is worth noting that any random effects of known correlation structure could be used in place of this specific polygenic form of **Z**. For example, a common family environment effect could be incorporated. Thus, except in the specific example, **Z** should be viewed as a generic random effect component. In application to real data there is usually neither interest in distinguishing, nor sufficient information to distinguish, between alternative forms of familial effects superposed on the effects of a segregating major gene. A polygenic component is simply a convenient way of modelling such additional familial covariation.

Finally, Professor Bodmer raised the question (not printed) of application of MCMC methods of likelihood estimation in 'real' studies of complex traits. The example of this paper is the only real data set for which performance of the methods has been thoroughly investigated, and hence has become our preferred example for comparative studies of alternative methods. However, we are now using these methods in conjunction with other approaches in the analysis of data in a current study of Apolipoprotein B levels and other possible quantitative indicators of heart disease.

References

Kong, K., Irwin, M., Cox, N. & Frigge, M. 1992 Multiloci problems and the method of sequential imputation. Technical Report No. 351, Department of Statistics, University of Chicago.

Ott, J. 1989 Computer simulation methods in linkage analysis. *Proc. natn. Acad. Sci. U.S.A.* **86**, 4175–4178.

Ploughman, L.M. & Boehnke, M. 1989 Estimating the power of a proposed linkage study for a complex genetic trait. *Am. J. Hum. Genet.* **44**, 543–551.

Stephens, D.A. & Smith, A.F.M. 1993 Bayesian inference in multipoint gene mapping. *Ann. Hum. Genet.* **57,** 65–82.