Sequential Imputation and Multipoint Linkage Analysis

Augustine Kong, Nancy Cox, Mike Frigge, and Mark Irwin

Departments of Statistics (A.K., M.F., M.I.) and Medicine (N.C.), University of Chicago, Chicago, Illinois

A novel Monte Carlo method for linkage analyses involving large pedigrees and many polymorphic loci is introduced. Issues related to the efficiency of the method are discussed. © 1993 Wiley-Liss, Inc.

Key words: disease mapping, marker mapping, importance sampling, Monte Carlo

INTRODUCTION

In both disease mapping and mapping of genetic markers, often many linked loci have to be handled simultaneously. Computationally efficient algorithms for calculating likelihoods are available for large pedigrees with a small number of loci, and small pedigrees with a large number of loci. However, for large pedigrees with a large number of loci, especially those that have substantial missing data, exact evaluation of a single likelihood value can be prohibitive because of the required memory and computing time. In this paper, a method called *sequential imputation* is proposed to handle problems of this type. It is a Monte Carlo approach that applies the traditional technique of importance sampling in a novel fashion. Based on a fixed value of the parameter, missing data are imputed conditioned on the observed data. The loci are processed one (or a few) at a time to reduce the demand on computational resources. The result is a collection of complete data sets with associated weights. Not only can the weights be used to estimate the likelihood of the parameter value used for the imputations, but the complete data sets can also be used to approximate the whole likelihood surface.

METHOD

In multilocus linkage problems, if for each person and each locus it is known exactly what allele is inherited from the father and what allele is inherited from the mother, the likelihood function is trivial to evaluate. We refer to this information which

Address reprint requests to Dr. Augustine Kong, Department of Statistics, University of Chicago, 5734 S. University Avenue, Chicago, IL 60637.

© 1993 Wiley-Liss, Inc.

is desirable, but not available (as least not entirely), as *missing data* and denote it by $\mathbf{z}$. The observed data, denoted by $\mathbf{y}$, include marker genotypes for some members of the pedigree. In the case of disease mapping, $\mathbf{y}$ will also include available disease phenotypes of the members. The combination $\mathbf{y}$, $\mathbf{z}$ is referred to as the *complete data.*

Let θ be the unknown parameter vector so that the likelihood function is $L(\theta) = p_\theta(\mathbf{y})$. In the case of disease mapping, θ is often a scalar that denotes the location of the disease gene relative to a set of markers whose locations are assumed to be known; θ may also incorporate other parameters such as marker allele frequencies and parameters relating the disease genotype and phenotype. In linkage mapping of markers, θ is a vector that denotes the relative locations among a collection of markers.

Let $\{y_1,\dots,y_n\}$ and $\{z_1,\dots,z_n\}$ be some decomposition of $\mathbf{y}$ and $\mathbf{z}$. At this time, assume that there are n loci so that for $t = 1,\dots,n$, y_t and z_t are respectively the observed and missing data on locus t. Other decompositions will be considered later. Note that the labels t, $t = 1,\dots,n$, do not necessarily correspond to the *physical* ordering, assumed or real, of the loci. Given a certain value of θ, sequential imputation [Kong et al., 1991] is a Monte Carlo method that allows us to obtain an unbiased estimate of $L(\theta)$ and generate weighted samples of $\mathbf{z} = \{z_1,\dots,z_n\}$ from the conditional distribution $p_\theta(\mathbf{z} \mid \mathbf{y})$. The method involves first simulating z_1^* from $p_\theta(z_1 \mid y_1)$ and computing $w_1 = p_\theta(y_1)$. Then the following two steps are applied for $t = 2,\dots,n$, in increasing order of t:

(A) Simulate z_t^* from the conditional distribution $p_\theta(z_t \mid y_1, z_1^*,\dots y_{t-1}, z_{t-1}^*, y_t)$. Notice that the z_t^*'s have to be simulated sequentially since each z_t^* is simulated conditioned on the previously imputed missing parts $z_1^*,\dots, z_{t-1}^*$

(B) Sequentially compute the predictive probabilities $p_\theta(y_t \mid y_1, z_1^*,\dots y_{t-1}, z_{t-1}^*)$ and $w_t = w_{t-1}\, p_\theta(y_t \mid y_1, z_1^*,\dots y_{t-1}, z_{t-1}^*)$ and set $w = w_n$.

Given the decompositions described above, for each t, (A) and (B) are done simultaneously and involve a single locus peel. Kong [1991] provides details on how to simulate missing data for one locus conditioned on the imputed missing data of other loci. Steps (A) and (B) are done repeatedly and independently m times. The choice of m, the number of imputations, is discussed later. Let the results be denoted $\mathbf{z}^*(1),\dots, \mathbf{z}^*(m)$ and $w(1),\dots,w(m)$, where $\mathbf{z}^*(j) = \{z_t^*(j),\dots,z_n^*(j)\}$ for $j = 1,\dots,m$. Note that $\mathbf{z}^*(j)$ is simulated from the distribution

$$
\begin{aligned}
p_\theta^*(\mathbf{z}^*(j) \mid \mathbf{y}) &= p_\theta(z_1^*(j) \mid y_1) \prod_{t=2}^{n} p_\theta(z_t^*(j) \mid y_1, z_1^*(j),\dots,y_{t-1}, z_{t-1}^*(j), y_t) \\
&= \frac{p_\theta(z_1^*(j), y_1)}{p_\theta(y_1)} \prod_{t=2}^{n} \frac{p_\theta(y_1,\dots,y_t, z_1^*(j),\dots,z_t^*(j))}{p_\theta(y_1,\dots,y_t, z_1^*(j),\dots,z_{t-1}^*(j))} \\
&= \frac{p_\theta(y_1,\dots,y_n, z_1^*(j),\dots,z_n^*(j))}{p_\theta(y_1)} \prod_{t=2}^{n} \frac{p_\theta(y_1,\dots,y_{t-1}, z_1^*(j),\dots,z_{t-1}^*(j))}{p_\theta(y_1,\dots,y_t, z_1^*(j),\dots,z_{t-1}^*(j))}
\end{aligned}
\qquad (1)
$$

$$=p_\theta(\mathbf{y},\mathbf{z}^*(j))\frac{1}{p_\theta(y_1)\prod_{t=2}^{n}p_\theta(y_t|y_1,z_1^*(j),\ldots,y_{t-1},z_{t-1}^*(j))}$$
$$=\frac{p_\theta(\mathbf{y},\mathbf{z}^*(j))}{w(j)}.$$

Hence $$E_{p^*}[w(j)]=\sum_{\mathbf{z}^*}w(j)p_\theta^*(\mathbf{z}^*(j)\,|\,\mathbf{y})=\sum_{\mathbf{z}^*}w(j)\frac{p_\theta(\mathbf{y},\mathbf{z}^*(j))}{w(j)}=\sum_{\mathbf{z}^*}p_\theta(\mathbf{y},\mathbf{z}^*(j))=p_\theta(\mathbf{y}).\quad(2)$$

As a consequence, $\bar{w}=m^{-1}\sum_{j=1}^{m}w(j)$ is an unbiased estimate of $L(\theta)$.

Besides getting an unbiased estimate of the likelihood, the samples $\mathbf{z}^*(j)$, $j=1,\ldots,m$, generated by sequential imputation can be treated as weighted samples (weight $\propto w(j)$) taken from the conditional distribution $p_\theta(\mathbf{z}|\mathbf{y})$. These samples can be used to estimate the likelihoods of other parameter values. Following Thompson and Guo [1991], let

$$h(\mathbf{y},\mathbf{z})=p_{\theta_1}(\mathbf{y},\mathbf{z})\,/\,p_{\theta_0}(\mathbf{y},\mathbf{z})$$

be the complete data likelihood ratio, where θ_0 and θ_1 are two values of θ. Note that both and $p_{\theta_0}(\mathbf{y},\mathbf{z})$ and $p_{\theta_1}(\mathbf{y},\mathbf{z})$, the complete data likelihoods, usually can be easily evaluated given any $\mathbf{z}$. If the sequential imputation is applied based on a parameter value θ_0, then

$$E_{p^*}[h(\mathbf{y},\mathbf{z}^*(j))w(j)\,|\mathbf{y}]=\sum_{\mathbf{z}^*}\frac{p_{\theta_1}(\mathbf{y},\mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y},\mathbf{z}^*(j))}w(j)p_{\theta_0}^*(\mathbf{z}^*(j)\,|\mathbf{y})$$
$$=\sum_{\mathbf{z}^*}\frac{p_{\theta_1}(\mathbf{y},\mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y},\mathbf{z}^*(j))}\frac{w(j)p_{\theta_0}(\mathbf{y},\mathbf{z}^*(j))}{w(j)}$$
$$=\sum_{\mathbf{z}^*}p_{\theta_1}(\mathbf{y},\mathbf{z}^*(j))$$
$$=p_{\theta_1}(\mathbf{y}),$$

which implies that $m^{-1}\sum_{j=1}^{m}h(\mathbf{y},\mathbf{z}^*(j))w(j)$ is an unbiased estimate of $L(\theta_1)$. Hence, by applying sequential imputation based on a parameter value θ_0, an unbiased estimate of the likelihood for any other parameter value can be obtained.

EFFICIENCY AND SAMPLE SIZE

The coefficient of variation of $\bar{w}$, $C[\bar{w}]$, measures the *relative* standard error of $\bar{w}$, as an estimate of $p_\theta(\mathbf{y})$. We have

$$C[\bar{w}]=\frac{1}{\sqrt{m}}C[w(j)]=\frac{1}{\sqrt{m}}\frac{\sqrt{Var_{p^*}[w(j)]}}{E_{p^*}[w(j)]}=\frac{1}{\sqrt{m}}\frac{\sqrt{Var_{p^*}[w(j)]}}{p_\theta(\mathbf{y})}$$

Its sample estimate is $\hat{C}[\bar{w}]=\frac{1}{\sqrt{m}}\hat{C}[w(j)]=\frac{1}{\sqrt{m}}\frac{s_w}{\bar{w}}$, where s_w denotes the sample standard deviation of the $w(j)$'s. For $C[\bar{w}]$ to be some desirable value δ, m has to be equal to $\delta^{-2}\times(C[\bar{w}(j)])^2$. For example, suppose we want $C[\bar{w}]$ to be around 0.1. Based on the samples, this implies that the number of imputations needs to be about $100 \times s_w^2/\bar{w}^2$. Note that $C[\bar{w}]$ is approximately the standard error of $\log_e \bar{w}$ as an estimate of $\log_e L(\theta)$. In the scale of the lod score, $C[\bar{w}] = 0.1$ corresponds to a standard error of 0.1 x $\log_{10} e \approx 0.043$.

As demonstrated, the efficiency of the method is inversely proportional to $(C[w(j)])^2$. Note that $C[w(j)]$ depends on the distribution from which the $\mathbf{z}^*$'s are simulated, which in turn depends on the decompositions of **y** and **z** used for sequential imputations. From **(1)** we have $w(j) = p_\theta(\mathbf{y})p_\theta(\mathbf{z}^*(j) \mid \mathbf{y})/p^*_\theta(\mathbf{z}^*(j) \mid \mathbf{y})$. In importance sampling, p^* is referred to as the *trial distribution*, and the ratio $p_\theta(\mathbf{z}^*(j) \mid \mathbf{y})/p^*_\theta(\mathbf{z}^*(j) \mid \mathbf{y})$ is called the importance sampling weight, so that $w(j)$ is the importance sampling weight multiplied by the unknown constant $p_\theta(\mathbf{y})$. This implies that $(C[w(j)])^2 = Var_{p^*}\left[p_\theta(\mathbf{z}^*(j)|\mathbf{y}) / p^*_\theta(\mathbf{z}^*(j)|\mathbf{y})\right]$ can be considered as a measure of *distance* between the actual conditional distribution of **z**, $p_\theta(\mathbf{z} \mid \mathbf{y})$, and the trial distribution $p^*_\theta(\mathbf{z} \mid \mathbf{y})$. To keep this distance small, it is desirable to have $p^*_\theta(\cdot \mid \mathbf{y})$ as close to $p_\theta(\cdot \mid \mathbf{y})$ as possible.

In section 2, we considered a special decomposition of the observed data **y** and the missing data **z**, i.e., y_t and z_t denote respectively the observed and missing data of a single locus t. To improve efficiency, it is necessary to consider other decompositions. Two criteria for choosing the appropriate decompositions are: (I) steps (A) and (B) can be performed inexpensively, in terms of both computing time and memory requirement; (II) the coefficient of variation $C[w(j)]$ is kept small. Note that (I) and (II) are often conflicting criteria. For example, under the trivial decomposition $\mathbf{y} = \{y_1\}$ and $\mathbf{z} = \{z_1\}$, $p_\theta(\mathbf{z} \mid \mathbf{y}) = p^*_\theta(\mathbf{z}^* \mid \mathbf{y})$ and $w(j) = p_\theta(\mathbf{y})$ with zero variation. But doing this requires peeling all the loci jointly, which is exactly what we are trying to avoid. We now present a few modifications of the basic procedure proposed in section 2 which will help to reduce the variation of $w(j)$ *without* increasing difficulties in computation.

Note that $p(\mathbf{z} \mid \mathbf{y})$ can be written as $p(z_1|\mathbf{y})\prod_{t=2}^{n} p(z_t|\mathbf{y},z_1,\ldots,z_{t-1})$. So simulating z^*_1 from $p(z_1 \mid \mathbf{y})$ is obviously preferable to simulating z^*_1 from $p(z_1 \mid y_1)$ *if* the former can be done cheaply. This suggests that, when simulating z^*_1, as much information as possible should be conditioned on as long as it does not increase computational cost. For each locus, each person, and each of their parents, define an identity-by-descent (IBD) variable as the indicator of whether the allele inherited by the person came from the grandfather or the grandmother. Often, some of the IBD variables can be deduced from the observed data **y**. Here we redefine y_1 to include the observed data on the first locus processed **plus** the IBD variables of other loci that can be deduced from the observed data. This change is easy to do and has virtually no effect on the amount of computation needed, but can reduce the variation of the weights substantially.

Further gains can be made by incorporating more than one locus into y_1. Note that the first step of sequential imputation involves computing $p_\theta(y_1)$ and simulating $z^*_1(j)$, $j = 1,\ldots,m$, from $p_\theta(z_1 \mid y_1)$. This requires peeling the loci incorporated in y_1 jointly, but note that only a single peel of these loci is needed for all m imputations. As long as the amount of computing time and memory required to perform this first peel are within acceptable limits, we should incorporate as many loci into y_1 as possible. This will decrease the variation of the weights and, as a consequence, possibly reduce the overall computing time.

The z_i's are imputed only because that helps to simplify computations. In section 2, **z** included every locus and every member in the pedigree. In some cases, some members of the pedigree are typed for some, but not all, of the loci. For a particular person and locus, we call the missing data *ignorable* if neither the person nor any of his/her descendants is typed for that locus. Since there is absolutely no information on these ignorable data, imputing them will only add noise and inflate the variation of the weights. Hence, for each t, $t = 1,\ldots,n$, we redefine z_t to include only data that are not ignorable. This redefinition does not make steps (A) and (B) more difficult and can reduce the computation time. But more important, doing this can drastically reduce the variance of the weights.

The order the loci are processed affects the trial distribution p^* and hence the variance, but not the mean, of the weights $w(j)$. An optimal order is one that minimizes the weight variance. An important guideline for choosing an optimal, or close to optimal, processing order is to start with loci that have the least amount of missing information among the nonignorable data. For two loci that have the same typed individuals, the one with more alleles, and hence usually more informative, should be processed first. Usually it is not too difficult to rank marker loci based on informativeness.

Location scores for a disease gene relative to a number of marker loci with known locations can be estimated by a simple strategy. Set y_n to be the observed disease data and process the markers first based on the above criteria. The average of the weights before processing the disease data, $\bar{w}_{n-1} = m^{-1}\sum_{j=1}^{m} w_{n-1}(j)$, is an unbiased estimate of $p(y_1,\ldots y_{n-1})$. Hence $\bar{w}_{n-1} \times p(y_n)$ is an unbiased estimate of the likelihood for the scenario that the disease locus is unlinked to the markers. Then process the disease locus at various locations linked to the marker loci. This strategy has the advantage that one set of marker imputations can be used to compute the likelihoods of all locations [Lange and Sobel, 1991]. Note that the estimates of the likelihoods for different locations all have easily computable standard errors. To increase efficiency for those that have very large standard errors, we can contemplate processing the disease locus first, maybe jointly with one or two markers close by for that particular assumed location of the disease locus. Finally, we note that it is usually enough to apply sequential imputation to a single location, probably in the middle, between two physically adjacent markers. Likelihoods for other locations in the interval can be estimated by the procedure discussed at the end of section 2. An example of this procedure, involving a pedigree of 155 individuals, 8 loci, and 32,256 haplotypes, is given in Kong et al. [1992]. By performing $m = 10{,}000$ impuations, which took a total of 20 CPU hours on a SUN SPARC I workstation, location scores were obtained for the whole region. Standard errors were smaller than 0.05 on the lod score scale for most of these estimates.

DISCUSSION

Because of the inherent limitations of existing computer programs and algorithms that do exact computations of likelihoods, investigators often have to reduce the number of loci and the number of alleles per locus in their analyses. This leads to loss of information and can create bias. In addition, because of the inefficiency of computing likelihoods point by point, sensitivity analyses for diagnostic schemes or marker allele frequencies may be prohibitively time consuming. The method of sequential imputation introduced in this paper can drastically reduce the burden for multipoint computations and can encourage more complete analyses.

Sequential imputation and other Monte Carlo methods [Guo and Thompson, 1992; Lange and Sobel, 1991] are most useful for problems in disease mapping where analyses involving large pedigrees with a substantial amount of missing data are unavoidable. For mapping markers, because small nuclear families with little missing data can be concentrated on, exact computations of likelihoods and the implementation of the EM algorithm can be very fast even with a large number of loci, using packages based on the algorithm given in Lander and Green [1987]. Hence, the need for Monte Carlo methods is not as apparent. However, we feel that sequential imputation can still be useful, although less crucial, for marker mapping. The algorithm in Lander and Green [1987] depends critically on the assumption of lack of genetic interference. In comparison, sequential imputation can computationally incorporate interference without any additional cost. While we share the beliefs of many investigators that the effect of interference is likely to be small in most situations, the capability to compute probabilities and likelihoods under models that incorporate interference can only help to refine analyses. A related issue is that estimated genetic distances between markers are often published without associated standard errors. That can be explained partly by the technical difficulties in obtaining standard errors when the data are not complete. Sequential imputation can help to resolve this problem by using the complete data sets to approximate the information matrix at the maximum likelihood estimate. However, when the maximum likelihood estimate of any of the recombination probabilities is too close to zero, the standard errors obtained by inverting the information matrix may not be appropriate. In these situations, an alternative approach is Bayesian analysis. The posterior distributions for the parameters based on carefully selected prior distributions can be approximated using the multiple complete data sets with suitable reweighting [Kong et al., 1991]. Further investigations in these directions are warranted.

ACKNOWLEDGMENTS

This research was facilitated by NSF grant DMS 89-02667 and NIH grant R01-GM46800-01A1. This manuscript was prepared using computer facilities supported in part by NSF grants DMS 89-05292, DMS 87-03942, and DMS 86-01732 to the Department of Statistics, University of Chicago, and by the University of Chicago Block Fund.

REFERENCES

Guo SW, Thompson E (1992): A Monte Carlo method for combined segregation and linkage analysis. Am J Hum Genet 51:1111-1126.

Kong A (1991): Efficient methods for computing linkage likelihoods of recessive diseases in inbred pedigrees. Genet Epidemiol 8:81-103.

Kong A, Liu J, Wong W (1991): Sequential Imputations and Bayesian Missing Data Problems. Technical report No. 321, Department of Statistics, University of Chicago.

Kong A, Irwin M, Cox N, Frigge M (1992): Multi-Locus Problems and the Method of Sequential Imputations. Technical report No. 351, Department of Statistics, University of Chicago.

Lange K, Sobel E (1991): A random walk method for computing genetic location scores. Am J Hum Genet 49:1320-1334.

Thompson E, Guo SW (1991): Evaluation of likelihood ratios for complex genetic models. IMA J Math Appl Med Biol 8:149-169.