African Journal of Biotechnology Vol. 9(49), pp. 8304-8307, 6 December, 2010
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB10.1478
ISSN 1684–5315 © 2010 Academic Journals

Full Length Research Paper

Hypothesis testing in genetic linkage analysis via Gibbs sampling

Gholamreza Jandaghi

Qom College, University of Tehran, Iran. E-mail: jandaghi@ut.ac.ir.

Accepted 12 November, 2010

Genetic linkage analysis involves estimating parameters in a genetic model in which a genetic trait is regressed on some factors such as polygenic values and environmental effects. Since only phenotypes are observed, hypothesis testing in such cases needs calculation of likelihood function in which one needs to consider all compatible configurations of genotypes. The number of these configurations increases as the size of a pedigree and the number of loci involved increase, Monte Carlo methods play an important role. The existing theory assumes an asymptotic normality for score statistics which is violated on boundary values which is the case in genetic linkage analysis. In this paper, a Markov Chain Monte Carlo approach is proposed to overcome this problem.

Key words: Gibbs sampling, pedigree, linkage analysis, likelihood.

INTRODUCTION

Geman and Geman (1984) proposed an iterative procedure called Gibbs sampler, for drawing multiple dependent realizations from a distribution known only to be proportionally constant. Sheehan et al. (1989) show that Markov chain Gibbs sampler produces a chain which is irreducible. Some difficulties are encountered when using Gibbs sampler as the method of resampling. These are: Slow convergence, the problem of multimodality of the distribution from which we are resampling and the problem of setting initial values. These problems have been addressed by Sheehan and Thomas (1993) and Jandaghi (1994). The Gibbs sampling method has been used for statistical inference in other biological researches during the past decade (Rezhetsky and Morozov, 2001). Since hypothesis testing is mainly required in genetic models, most researchers use the asymptotic normality property. In linkage analysis, testing hypothesis on recombination fraction usually involves testing either $H_0:\theta=0$ or $H_0:\theta=0.5$.

Asymptotic theory states that under some regularity conditions on the likelihood function, the unrestricted maximum likelihood estimate is asymptotically normal (Cox and Hinkley, 1974). One of these regularity conditions is that θ_0 is in the interior of the parameter space Ω. Since the usual hypothesis test in linkage is either $\theta=0$ or $\theta=0.5$, the regularity condition is violated, and we need to take account of the truncation of the limiting normal distribution (Cox and Hinkley, 1974). As a result, when dealing with the restricted maximum likelihood estimate and θ_0 is on the boundary, which is the case in linkage analysis, the likelihood ratio and the score statistic are not guaranteed to be asymptotically chi-squared distributed. In this paper, a two level Gibbs sampling is proposed to find the true shape of their distributions.

TERMINOLOGY AND NOTATION

To consider a pedigree with n individuals, let $g=\left(g_1,g_{2,},...,g_{_n}\right)$ be the vector of genotypes of the individuals in the pedigree, where g_i is the genotype of the i-th individual. Let g_{-i} = $\left(g_i \mid g_1,...,g_{i-1},g_{i+1},...,g_n\right)$, let $x=\left(x_1,x_2,...,x_n\right)$ be the observed data, where x_i is the observed phenotype of the i-th individual, $P\left(x\mid g\right)$ is the penetrance

Abbreviation: MSC2010: 62C10; 62F03; 62F40.

probability, that is, the probability that an individual with genotype g has phenotype x, and $P\left(g_k \mid g_f, g_m\right)$ is the transmission probability, that is, the probability that an individual has genotype g_k given the parental genotypes g_f and g_m. Let x_j and g_j be the phenotype and the genotype of individual j, respectively. Let $g_{f_j}, g_{m_j}, g_{s_j}$ and $g_{s_{jl}}$ be the phenotype of father, mother, spouse and the offspring of individual j.

CALCULATION OF LIKELIHOOD

In any genetic model, the trait of interest is modeled against some genotypic values or parameters and by using statistical methods, these model parameters are estimated. When performing pedigree analysis, the basic statistical tool used is the likelihood function. We can use the conditional independence imposed by Mendel's laws to express the likelihood function L as a product of transmission probabilities $P\left(g_k \mid g_f, g_m\right)$, penetrance probabilities $P\left(x_i \mid g_i\right)$ and population gene frequencies $P\left(g\right)$, and sum over all genotype possibilities for all pedigree members to calculate L as outlined below. Using the terminology and notations presented in the previous section, the likelihood for the pedigree is:

$$L(\theta) = P(x,\theta) = \sum_{\textit{feasible genes}} P(x \mid g) P(g,\theta)$$

Where,

$$P(x \mid g) = \prod_{j=1}^{n} P(x_i \mid g_i)$$

and

$$P(g,\theta) = \prod_{j \notin F} P_\theta\left(g_j \mid g_{f_j}, g_{m_j}\right) \prod_{j \in F} P\left(g_j\right)$$

or equivalently,

$$L(\theta) = \sum P(x_1 \mid g_1) P_\theta(g_1 \mid .) \ldots \sum P(x_n \mid g_n) P_\theta(g_n \mid .)$$

CALCULATION OF THE SCORE AND LIKELIHOOD RATIO STATISTICS

One of the statistics often used for testing the null hypothesis of no linkage is the score test. To do such a test, we need to estimate the first and second derivatives of the likelihood. Knowing that:

$$L(\theta) = \sum P(x,\theta) = \sum \frac{P(x,\theta)}{P(x,\theta_0)} P(x,\theta_0)$$

and its approximation

$$L(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{P\left(g^i,\theta\right)}{P\left(g^i,\theta_0\right)}$$

Where, g^i is the i-th Gibbs sample, we can estimate the first derivative of the logarithm of the likelihood function

$$S = \frac{d}{d\theta} \log\left(L(\theta)\right) = \frac{L^{'}(\theta)}{L(\theta)}$$

by

$$S \approx \frac{\displaystyle\sum_{i=1}^{N} \frac{P^{'}\left(g^i,\theta\right)}{P\left(g^i,\theta_0\right)}}{\displaystyle\sum_{i=1}^{N} \frac{P\left(g^i,\theta\right)}{P\left(g^i,\theta_0\right)}} \qquad (1)$$

Since $P(g,\theta) = \prod_{j=1}^{n} P_j\left(g_j,\theta\right)$, its derivative will be

$$\frac{d}{d\theta} P(g,\theta) = P(g,\theta) \sum_{j=1}^{n} \frac{P_j^{'}\left(g_j,\theta\right)}{P_j\left(g_j,\theta\right)} \qquad (2)$$

Substituting (2) in (1), we will have

$$S = \frac{\displaystyle\sum_{i=1}^{N} \left[\frac{P\left(g^i,\theta\right)}{P\left(g^i,\theta_0\right)} \sum \frac{P_j^{'}\left(g_j^i,\theta\right)}{P_j\left(g_j^i,\theta_0\right)} \right]}{\displaystyle\sum_{i=1}^{N} \frac{P\left(g^i,\theta\right)}{P\left(g^i,\theta_0\right)}} \qquad (3)$$

The estimate of the variance of the score statistic is calculated by taking the second derivative of the logarithm of the likelihood with respect to θ, such that

$$\frac{d^2}{d\theta^2} \log\left(L(\theta)\right) = \frac{d}{d\theta} \left[\frac{L^{'}(\theta)}{L(\theta)} \right]$$

which is equal to

$$\frac{L''(\theta)}{L(\theta)} - \left[\frac{L'(\theta)}{L(\theta)}\right]^2 \tag{4}$$

Where

$$L''(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\frac{P''\left(g^i,\theta\right)}{P\left(g^i,\theta_0\right)}$$

Since

$$P''\left(g,\theta\right) = \frac{d}{d\theta}P'\left(g,\theta\right),$$

$$P''\left(g,\theta\right) = P\left(g,\theta\right)\left(\left[\sum_{j=1}^{n}\frac{P_j'\left(g^i,\theta\right)}{P_j\left(g^i,\theta\right)}\right]^2 + \sum_{j=1}^{n}\left[\frac{P_j''\left(g_j,\theta\right)}{P_j\left(g_j,\theta\right)} - \left[\frac{P_j'\left(g_j,\theta\right)}{P_j\left(g_j,\theta\right)}\right]^2\right]\right).$$

(5)

So the variance of the score statistic is estimated by:

$$\widehat{\mathrm{var}}\left(\hat{S}\right) = \frac{\sum_{i=1}^{N}\left\{\frac{P\left(g^i,\theta\right)}{P\left(g^i,\theta_0\right)}\left[\left[\sum_{j=1}^{n}\frac{P_j'\left(g_j^i,\theta\right)}{P_j\left(g_j^i,\theta\right)}\right]^2 + \sum_{j=1}^{n}\left[\frac{P_j''\left(g_j^i,\theta\right)}{P_j\left(g_j^i,\theta\right)} - \left[\frac{P_j'\left(g_j^i,\theta\right)}{P_j\left(g_j^i,\theta\right)}\right]^2\right]\right]\right\}}{\sum_{i=1}^{N}\frac{P\left(g^i,\theta\right)}{P\left(g^i,\theta_0\right)}} - S^2.$$

(6)

Therefore, Equations (3) and (6) can be used for estimation of the score statistic and its variance and testing of hypothesis is straightforward. The derivatives of the likelihood function are easily computed, because the probabilities $P_j\left(g_j,\theta\right)$ involve the terms θ and/or $\left(1-\theta\right)$. We can assign an array corresponding to $P_j\left(g_j,\theta\right)$ and every time a probability is calculated, we can specify the term corresponding to it and thus the derivatives of the likelihood function are analytically calculated.

Similar to the calculation of the score statistic, we can use Monte Carlo simulation to calculate the likelihood ratio:

$$\ell\left(\theta,\theta_0\right) = \frac{L(\theta)}{L\left(\theta_0\right)}.$$

Using the notation of Guo and Thompson (1991), the likelihood ratio can be written as

$$\ell\left(\theta,\theta_0\right) = \sum_{g}\frac{f_\theta\left(y \mid g\right)P_\theta\left(g\right)}{f_{\theta_0}\left(y \mid g\right)P_{\theta_0}\left(g \mid y\right)}P_{\theta_0}\left(g \mid y\right) \tag{7}$$

Where, g is the genotype, y is the phenotype of the pedigree and θ is the parameter. Since by definition

$$L\left(\theta_0\right) = f_{\theta_0}\left(y\right)$$

Equation (7) can be rewritten as:

$$\ell\left(\theta,\theta_0\right) = \sum_{g}\frac{f_\theta\left(y \mid g\right)P_\theta\left(g\right)}{f_{\theta_0}\left(y \mid g\right)P_{\theta_0}\left(g\right)}P_{\theta_0}\left(g \mid y\right) \tag{8}$$

which can be estimated by the average

$$\ell\left(\theta,\theta_0\right) \square \frac{1}{N}\sum_{g}\frac{P_\theta\left(g^i\right)f_\theta\left(y \mid g^i\right)}{P_{\theta_0}\left(g^i\right)f_{\theta_0}\left(y \mid g^i\right)} \tag{9}$$

Therefore, based on N Gibbs realization of genotypes for the pedigree, we can use Equation (9) to calculate the estimate of the likelihood ratio.

The procedure for calculating the distribution needs two levels of Gibbs sampling. In other words, different quantiles of the distribution correspond to different sets of phenotypes and to generate different sets of phenotypes, one needs one step of Gibbs sampling which we refer to as Outer-Gibbs sampling. After generating a collection of different sets of phenotypes, the Gibbs sampler is run to compute the score statistic and its variance for each set of phenotypes. We call this step Inner-Gibbs sampling. Therefore, we can first generate n_O outer-Gibbs samples and then for each one of n_O different sets of phenotypes, we generate n_I inner-Gibbs samples to calculate the score statistic and its variance. The stepwise procedure is as follows:

Step 1. Generate n_O outer-Gibbs samples from the distribution of pedigree genotypes unconditional on their phenotypes, that is, generate the vector $g = \left(g_1, g_2, ..., g_n\right)$ from the following distribution:

$$\begin{aligned} P_\theta\left(g_j \mid g^{-j}\right) &= P_\theta\left(g_j \mid \left\{g_{s_j}\right\}, \left\{g_{s_{jl}}\right\}, g_{f_j}, g_{m_j}\right) \\ &\propto \prod_{j,l} P_\theta\left(g_{s_{jl}} \mid g_j, g_{s_j}\right) P_\theta\left(g_j \mid g_{f_j}, g_{m_j}\right) \end{aligned} \tag{10}$$

Step 2. Assign the phenotypes consistent with each set of genotypes generated in step 1, so, we have n_O different sets of phenotypes for the pedigree.

Step 3. For each set of phenotypes produced in step 2, they generate n_I inner-Gibbs samples from the conditional distribution of genotypes on phenotypes:

$$P_\theta\left(g_j \mid g^{-j}, x_j\right) \propto \prod_{j,l} P_\theta\left(g_{s_{jl}} \mid g_j, g_{s_j}\right) P_\theta\left(g_j \mid g_{f_j}, g_{m_j}\right) P_\theta\left(x_j \mid g_j\right) \tag{11}$$

Step 4. Using equations (3) and (6) to calculate estimates of score statistic and their variances, we have n_O different values of score statistic on which we can build up the distribution of S .

Same procedure can be followed to calculate the distribution of likelihood ratio. We can also use these approximate distributions or their moments for power calculation and testing hypotheses of the parameter θ .

CONCLUSION

In genetic linkage, usually the focus is to test some hypotheses on genetic model parameters. Since estimation of the parameters involves the calculation of likelihood, the complexity of pedigree likelihood forces the researchers to do a huge amount of calculation. Therefore one needs to use Monte Carlo methods for estimation. Furthermore, testing the boundary value of parameters does not allow the researcher to assume the asymptotic normality regulations in Cox and Hinkley (1974). The proposed approach allows accessing the exact distribution of likelihood statistics and the use of its quantiles for confidence interval building and hypothesis testing purposes. Although, there are some difficulties with using Gibbs sampling, especially its slow convergence and setting the initial values, it still remains an efficient method of dealing with computational problems in linkage analysis.

REFERENCES

Cox DR, Hinkley DV (1974). Theoretical Statistics, Chapman and Hall, London.

Geman S, Geman D (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6: 721-741.

Guo SW, Thompson EA (1991). Monte Carlo Estimation of Variance Component Models for Large Complex Pedigrees, IMA J. Math. Appl. Med. Biol. 8: 171-189.

Jandaghi G (1994). Monte Carlo Estimation of the Pedigree likelihood and Its Statistics using Gibbs Sampler, Ph.D. Thesis, Department of Preventive Medicine and Biostatistics, University of Toronto, Canada.

Rezhetsky A, Morozov P (2001). Markov Chain Monte Carlo Computation of Confidence Intervals for Substitution-Rate Variation in Proteins, Pacific Symposium in Biocomputing, 6: 203-214.

Sheehan NA, Possolo A, Thompson EA (1989). Image Processing Procedures Applied to the Estimation of Genotypes on Pedigrees, Am. J. Hum. Genet. 45(Suppl.): A248.

Sheehan NA, Thomas A (1993). On the Irreducibility of a Markov Chain Defined on a Space of Genotype Configurations by a Sampling Scheme, Biometrics, 49: 163-175.