Problems with computational methods in population genetics

Matthew Stephens
Department of Statistics
University of Oxford
1, South Parks Road, Oxford, UK
stephens@stats.ox.ac.uk

1. Introduction

Advances in computational power, statistical methodology, and population genetics theory have recently had an impact on the way in which information is extracted from molecular genetic data. Several computationally intensive schemes have already been developed for analysis of such data under simple models (see for example Griffiths and Tavaré, 1994; Kuhner et al., 1995; Wilson and Balding, 1998), and their popularity will no-doubt increase as they become able to incorporate more biologically-realistic assumptions. It is important to recognize that these methods are being used to tackle difficult problems which are extremely challenging, even using the sophisticated theory and technology now available. Since these methods are likely to be widely applied by people who are not experts in computational statistics, it is particularly important to develop methods which are easily used to produce reliable results. This piece aims to describe one sense in which some current approaches are more likely to provide reliable results than others.

2. Background

Analyses of molecular genetic data typically wish to answer questions based on the genetic types A_n of a sample of n chromosomes from a population. Coalescent theory (see Donnelly and Tavaré, 1997, for a review) provides a way of modelling the distribution of the (random) genealogical tree $\mathcal{G}$ relating the sampled chromosomes, prior to observing their genetic types. This model will depend on the sampling scheme used, and on the size and mating behaviour of the population. For example, the standard coalescent introduced by Kingman (1982) is appropriate for a random sample from a constant-sized randomly-mating population, and generalisations of this process have been developed for more complex settings (see for example Herbots, 1997).

Given the genealogical tree $\mathcal{G}$, the distribution of genetic types is obtained by specifying a model for the mutation mechanism, and a distribution for the genetic type of the individual at the top of the tree (often the stationary distribution of the mutation mechanism). The effects of mutation can then be traced down the branches of $\mathcal{G}$, assuming mutations occur independently on the branches as independent Poisson processes of some rate $\theta/2$, giving a distribution for the genetic types of the sampled chromosomes at the tips of the tree. More background is given by Donnelly and Tavaré (1997). Questions of interest may relate to the genealogy $\mathcal{G}$, the mutation mechanism, and the population demography. We focus here on the problem of performing inference for the parameter θ.

3. Methods

Computational methods for performing inference for θ typically proceed by considering some aspect, $\mathcal{H}$, of the ancestry of the sample to be "missing data". The exact details of which data are considered to be missing vary from method to method, but they all share the property that

$\mathsf{P}_\theta(A_n \mid \mathcal{H})$ is (relatively) easy to calculate. A naive estimator of the likelihood for θ is then given by

$$
\begin{aligned}
L(\theta) &= \mathsf{P}_\theta(A_n) \\
&= \int \mathsf{P}_\theta(A_n \mid \mathcal{H}) \mathsf{P}_\theta(\mathcal{H})\, d\mathcal{H} \approx \frac{1}{M} \sum_{i=1}^{M} \mathsf{P}_\theta(A_n \mid \mathcal{H}^{(i)})
\end{aligned} \tag{1}
$$

where $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots, \mathcal{H}^{(M)}$ are independent samples from $\mathsf{P}_\theta(\mathcal{H})$. This estimator typically suffers from having a very large variance, and the values of M required to get reliable results are typically too large for this method to be practicable. Furthermore the estimator requires a different set of simulations to be performed for each value of θ at which the likelihood is to be estimated, making estimation of a likelihood surface even more computationally demanding. We divide the methods which have been used to tackle this problem into three types:

Importance Sampling

Importance sampling (see Ripley, 1987, for an introduction) is a standard method of reducing the variance of estimators such as (1). For any probability distribution $Q(\mathcal{H})$ whose support contains the set $\{\mathcal{H} : \mathsf{P}(A_n \mid \mathcal{H})\mathsf{P}(\mathcal{H}) > 0\}$, we have

$$
\begin{aligned}
\mathsf{P}_\theta(A_n) &= \int \mathsf{P}_\theta(A_n \mid \mathcal{H}) \mathsf{P}_\theta(\mathcal{H})\, d\mathcal{H} \\
&= \int \frac{\mathsf{P}_\theta(A_n \mid \mathcal{H}) \mathsf{P}_\theta(\mathcal{H})}{Q(\mathcal{H})} Q(\mathcal{H})\, d\mathcal{H} \\
&\approx \frac{1}{M} \sum_{i=1}^{M} \frac{\mathsf{P}_\theta(A_n \mid \mathcal{H}^{(i)}) \mathsf{P}_\theta(\mathcal{H}^{(i)})}{Q(\mathcal{H}^{(i)})}
\end{aligned} \tag{2}
$$

where $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots, \mathcal{H}^{(M)}$ are independent samples from the *importance sampling function* Q. For appropriate choice of Q this estimator will have much smaller variance than (1). Indeed, if Q is the posterior distribution $\mathsf{P}_\theta(\mathcal{H} \mid A_n)$ then the estimator (2) will have zero variance, but unfortunately this posterior distribution is not known for most cases of interest. In any case, it is desirable to use a single Q to estimate the likelihood at several values of θ, and no single Q is optimal for all values of θ. Felsenstein et al. (1998) point out that the method of Griffiths and Tavaré (1994) can be interpreted as an importance sampling scheme, and its performance is investigated and compared with other possible importance sampling schemes by Stephens and Donnelly (1999).

MCMC methods (fixed θ)

An alternative method of estimating (relative) likelihood surfaces is provided by Markov chain Monte Carlo (MCMC) methods. Standard MCMC methods can be used to construct a Markov chain with the posterior distribution of the missing data for some fixed θ_0, $\mathsf{P}_{\theta_0}(\mathcal{H} \mid A_n)$, as its stationary distribution. A relative likelihood surface for θ may then be estimated using

$$
\begin{aligned}
\frac{\mathsf{P}_\theta(A_n)}{\mathsf{P}_{\theta_0}(A_n)} &= \int \frac{\mathsf{P}_\theta(A_n \mid \mathcal{H}) \mathsf{P}_\theta(\mathcal{H})}{\mathsf{P}_{\theta_0}(A_n)}\, d\mathcal{H} \\
&= \int \frac{\mathsf{P}_\theta(A_n \mid \mathcal{H}) \mathsf{P}_\theta(\mathcal{H})}{\mathsf{P}_{\theta_0}(A_n) \mathsf{P}_{\theta_0}(\mathcal{H} \mid A_n)} \mathsf{P}_{\theta_0}(\mathcal{H} \mid A_n)\, d\mathcal{H} \\
&\approx \frac{1}{M} \sum_{i=1}^{M} \frac{\mathsf{P}_\theta(\mathcal{H}^{(i)}) \mathsf{P}_\theta(A_n \mid \mathcal{H}^{(i)})}{\mathsf{P}_{\theta_0}(\mathcal{H}^{(i)}) \mathsf{P}_{\theta_0}(A_n \mid \mathcal{H}^{(i)})} = \frac{1}{M} \sum_{i=1}^{M} w_i \text{ say,}
\end{aligned} \tag{3}
$$

where $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots, \mathcal{H}^{(M)}$ are samples from the Markov chain at stationarity. This method, which is used by Kuhner et al. (1995), may be viewed as a version of importance sampling using $\mathsf{P}_{\theta_0}(\mathcal{H} \mid A_n)$ as the importance sampling function. As noted above, this is the optimal importance sampling function for $\theta = \theta_0$, but since in this case the relative likelihood is trivially one, this efficiency is rather illusory.

MCMC methods (variable θ)

A third approach, used for example by Wilson and Balding (1998), is to place a prior on θ and create a Markov chain with the joint posterior distribution of $(\theta, \mathcal{H})$ as its stationary distribution. This may naturally be used to report a posterior distribution for θ given the data, but may also be used to estimate a relative likelihood surface for θ, for example by using non-parametric methods to estimate the posterior density from the MCMC sample, and dividing this density estimate by the prior density to give an estimate of the relative likelihood.

4. Problems

The problem with the first two methods described above is essentially that choosing a suitable importance sampling function $Q(\cdot)$ which is efficient for a wide range of values of θ seems to be extremely difficult. In particular, although the posterior distribution $\mathsf{P}_{\theta_0}(\mathcal{H} \mid A_n)$ is the optimal importance sampling function for estimating the likelihood at $\theta = \theta_0$, the variance of the estimator (3) may become extremely large (or even infinite) for θ away from θ_0.

For example, consider the method of Kuhner et al. (1995), which takes the missing data $\mathcal{H}$ to be the *scaled genealogical tree*; that is, the genealogical tree $\mathcal{G}$ with its branch lengths scaled by a factor of $\theta/2$, so that mutations may be assumed to occur as a Poisson process of unit rate along the branches. The probability of the data given $\mathcal{H}$, $\mathsf{P}_\theta(A_n \mid \mathcal{H})$, is then independent of θ, and the importance weights w_i in (3) become $w_i = \mathsf{P}_\theta(\mathcal{H}^{(i)})/\mathsf{P}_{\theta_0}(\mathcal{H}^{(i)})$. Consider the variance of the estimator (3), given by $(\mathsf{E}(w^2) - \mathsf{E}(w)^2)/M^2$, in the simple case where A_n is a random sample of size $n = 2$ from a randomly-mating constant-sized population. In this case the standard coalescent may be used to model the genealogy $\mathcal{G}$ relating the two individuals, and the rescaled genealogy $\mathcal{H}$ will be completely determined by its height T, which has exponential distribution with mean $\theta/2$:

$$\mathsf{P}_\theta(t) = (2/\theta)\exp(-2t/\theta). \tag{4}$$

The mean of the square of a w_i is then given by

$$\begin{aligned}
\mathsf{E}(w^2) &= \int \left(\frac{\mathsf{P}_\theta(t)}{\mathsf{P}_{\theta_0}(t)}\right)^2 \mathsf{P}_{\theta_0}(t \mid A_n)\, dt \\
&= \int \frac{\mathsf{P}_\theta(t)^2}{\mathsf{P}_{\theta_0}(t)} \frac{\mathsf{P}_{\theta_0}(A_n \mid t)}{\mathsf{P}_{\theta_0}(A_n)}\, dt \qquad (5) \\
&= \frac{1}{\mathsf{P}_{\theta_0}(A_n)} \frac{2\theta_0}{\theta^2} \int \exp\left\{-\left(\frac{2}{\theta} - \frac{1}{\theta_0}\right) 2t\right\} \mathsf{P}_{\theta_0}(A_n \mid t)\, dt.
\end{aligned}$$

Now as $t \to \infty$, the distribution of A_n given t becomes the same as an independent sample from the stationary distribution of the mutation process. Thus (for many mutation mechanisms) $\mathsf{P}_{\theta_0}(A_n \mid t)$ tends to some limit > 0 as $t \to \infty$, and so the integral (5) is infinite if $2/\theta < 1/\theta_0$. Thus for $\theta > 2\theta_0$ the variance of the estimator (3) is infinite, and estimates of the relative likelihood surface for these values of θ will be extremely unreliable. In particular, the relative likelihood will tend to be severely underestimated, since with high probability only small importance weights will be observed. We note that this result holds however well the chain is

mixing, and would hold even if we were able to obtain independent samples from the posterior distribution of $\mathcal{H}$.

Curiously, this problem appears (on preliminary investigation) to be less severe if we instead work with the unscaled genealogy $\mathcal{G}$ as the missing data, although the general principal that the estimator (3) will have a larger variance, and tend to underestimate the relative likelihood, for θ away from θ_0, will apply whatever is chosen as the missing data. For this reason we believe methods which allow θ to vary to be preferable, even if we do not wish to take a Bayesian approach to inference for θ. That is, introducing a prior distribution for θ and treating it as a random variable is a useful computational device, even if the ultimate goal is to treat θ as a parameter and to estimate it by maximising the likelihood. Alternatively, the importance sampling schemes could be improved by the application of more sophisticated methods such as bridge sampling or path sampling (Gelman and Meng, 1998).

REFERENCES

Donnelly, P. and Tavaré, S., editors (1997). *Progress in Population Genetics and Human Evolution.* Springer, New York.

Felsenstein, J., Kuhner, M. K., Yamato, J., and Beerli, P. (1998). Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. Submitted.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185.

Griffiths, R. C. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology*, 46:131–159.

Herbots, H. (1997). The structured coalescent. In Donnelly, P. and Tavaré, S., editors, *Progress in Population Genetics*. Springer, New York.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13:235–248.

Kuhner, M. K., Yamato, J., and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics*, 140:1421–1430.

Ripley, B. D. (1987). *Stochastic Simulation.* Wiley, New York.

Stephens, M. and Donnelly, P. (1999). Ancestral inference and the coalescent. In preparation.

Wilson, I. J. and Balding, D. J. (1998). Genealogical inference from microsatellite data. *Genetics*, 150:499–510.

RÉSUMÉ

Computationally intensive statistical methods are likely to play an increasing important role in the analysis of molecular genetic data. Typically these data present problems which are extremely challenging, even using the sophisticated computational theory and technology now available. Here we suggest that these problems are easily underestimated, and describe one sense in which some current approaches are more likely to provide reliable results than others.