2.2 Exact Bayesian inference

Consider a sample of $n = 2$ DNA sequences under the standard coalescent model with infinite-sites mutation. In this setting the number of segregating loci S is equivalent to the full dataset. Suppose now that θ is known, and is here the *total* scaled mutation rate over all loci, or sites, in the dataset. The unknown of interest is the coalescence time t_1. Before observing its value, the coalescent model specifies an exp(1) distribution for t_1, with probability density function (pdf):

$$\pi(t_1{=}t) = \exp(-t). \tag{1}$$

The likelihood function is Poisson(θt):

$$\mathrm{P}(S{=}s|t_1{=}t) = \frac{1}{s!}(\theta t)^s \exp(-\theta t). \tag{2}$$

By Bayes theorem we obtain the posterior pdf of t_1:

$$\pi(t_1{=}t|S{=}s) = c(\theta t)^s \exp(-(\theta{+}1)t), \tag{3}$$

where c is a constant (does not depend on t). The RHS of (3) has the form of the gamma($1{+}s, 1{+}\theta$) pdf, and it follows that

$$\mathrm{E}[t_1|S{=}s] = \frac{1+s}{1+\theta} \qquad \text{and} \qquad \mathrm{V}[t_1|S{=}s] = \frac{1+s}{(1+\theta)^2}, \tag{4}$$

which may be compared with prior moments $\mathrm{E}[t_1] = \mathrm{V}[t_1] = 1$. Noting that $\mathrm{E}[S] = \theta$, we see that if $s < \mathrm{E}[S]$ then $\mathrm{E}[t_1|S{=}s] < \mathrm{E}[t_1]$, and vice-versa. Further, $\mathrm{V}[t_1|S{=}s] < \mathrm{V}[t_1]$ unless $s > 2\mathrm{E}[S] + \mathrm{E}[S]^2$.

The prior density curve and posterior curves for several values of s when $\theta = 1$ are shown in Figure 1. Recall that t_1 is measured in coalescent time units: to convert to generations we need to multiply by the effective population size N.

As this example illustrates, the coalescent model can be thought of as specifying a prior distribution for the genealogical tree, which after observing data can be updated to give the corresponding posterior distribution. Some researchers have treated unknowns such as coalescent times and ages of mutations as classical parameters, and estimated them ignoring the information in the prior distribution, even when working within the coalescent framework. This waste of information is inefficient. For example, a natural estimator of t_1 within the classical framework is the moments estimator S/θ, for which the mean square error (MSE) is

$$\mathrm{MSE}(S/\theta) = \mathrm{E}_{t_1}[\mathrm{E}_{S|t_1}[(S/\theta - t_1)^2]] = 1/\theta,$$

which is uniformly larger than $1/(1{+}\theta)$, the MSE of the posterior mean, given at (4), which might be considered as a point estimator for t_1. However, under the

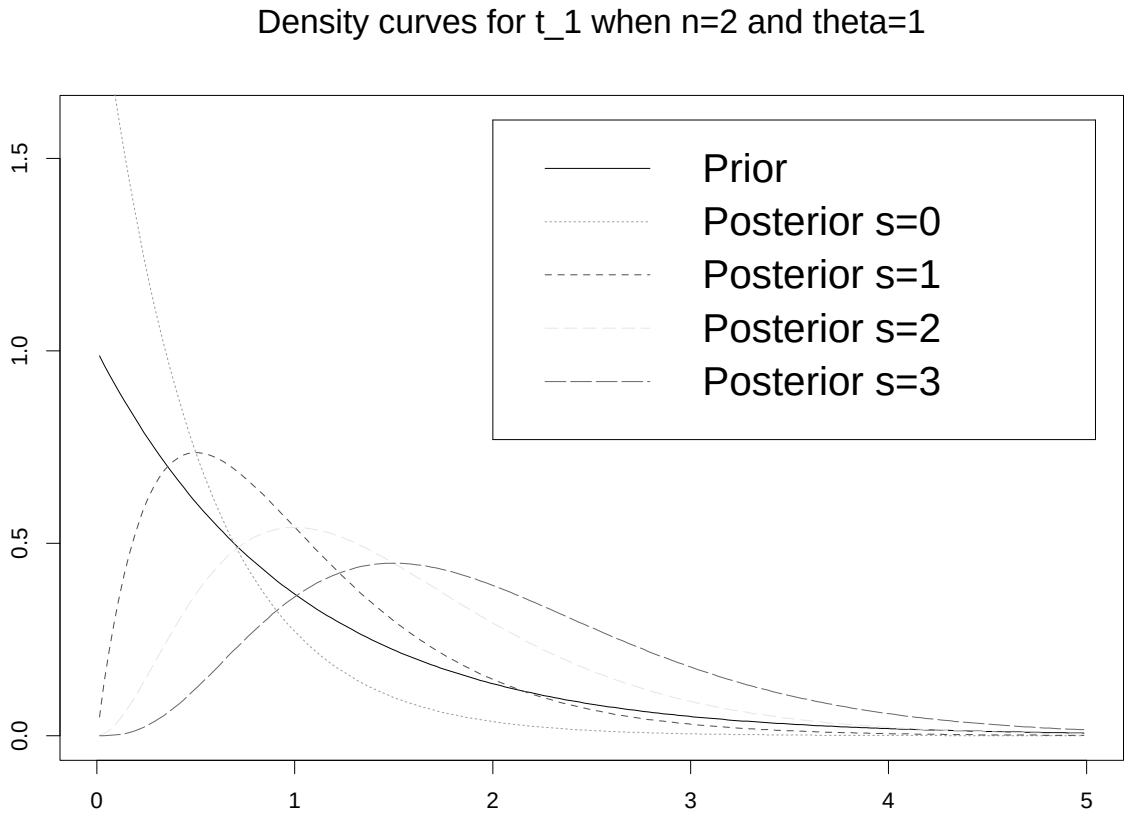

Figure 1: Prior is gamma$(1,1) \equiv$ exp(1); posteriors are gamma$(1{+}s, 2)$

Bayesian paradigm point estimators are often avoided as they do not reflect the full information in the posterior density curves, such as those in Figure 1.

For Bayesian inferences about θ, the unknown coalescence times and tree topology need to be integrated out with respect to their coalescent prior distribution in order to obtain the marginal likelihood for θ. Continuing the $n = 2$ example, the marginal likelihood is

$$\mathrm{P}(S|\theta) = \int_0^\infty \mathrm{P}(S|t_1{=}t, \theta)\pi(t_1{=}t)\mathrm{d}t,$$

which on substituting from (1) and (2) can be evaluated exactly:

$$\mathrm{P}(S|\theta) = \int_0^\infty \frac{(\theta t)^s}{s!}\exp(-(\theta{+}1)t)\mathrm{d}t = \frac{\theta^s}{(1+\theta)^{s+1}}. \tag{5}$$

On maximising (5) wrt θ we obtain the MLE $\hat{\theta} = S$.

When $n > 2$, a marginal likelihood for θ can be evaluated exactly in the special case that no variation is observed in the sample, and it is assumed that these data

reflect no mutations in the underlying genealogy. Then the marginal likelihood is the product over i of the probability that no mutation occurs when the genealogy has exactly i lineages:

$$\mathrm{P}(S{=}0|\theta) = \prod_{i=1}^{n-1} \frac{i}{i+\theta}. \tag{6}$$

The MLE is $\hat{\theta} = 0$, which is usually implausible *a priori*.

The problem with the MLE mentioned above can be overcome by reporting, for example, a posterior 95% highest-density interval, using either an improper uniform prior for θ or a proper, informative prior. An additional advantage to a Bayesian interpretation is that it becomes possible to report inferences about N and μ separately. Recall that the scaled mutation rate θ is equal to $2N\mu$. The reason for working with θ rather than N and μ separately is that marginal likelihoods such as (5) and (6) depend on these two parameters only through their product. However, there is usually background information about N and μ to formulate informative prior distributions, in which case it is possible to obtain a useful joint posterior distribution for them. This is of great practical importance because the effective population size N is the key to interpreting times under the coalescent model. It will be explored further below using approximate methods of Bayesian inference. Note here that inferences about N and μ are always sensitive to the prior assumptions whereas, in the presence of substantial data, inferences about θ are robust to the prior.

3 Bayesian Inference via Rejection Sampling

3.1 Explicit likelihood computations

Let us return to the general setting of a sample of size n sequences under any genealogical and mutation models. Often it is reasonable to assume *a priori* that the unknowns of interest are mutually independent, and drawn from a family of distributions such as the gamma or lognormal, for which accurate simulation algorithms are available. Such simulation algorithms are not directly useful, because we need to simulate from the posterior, not the prior. However, if the likelihood $\mathrm{P}(\mathrm{data}|\theta)$ can be computed it is possible to generate a posterior random sample for some parameter of interest, say θ, as follows.

Rejection Algorithm 1:

1. simulate θ from its prior, with density $\pi(\theta)$;
2. if $u > c\mathrm{P}(\mathrm{data}|\theta)$ then accept θ, otherwise reject. Here, u is a standard uniform random variable, and c is any constant such that the RHS of the

inequality does not exceed unity. To maximise acceptances c should be as large as possible subject to this constraint.

Rejection Algorithm 1 is a sort of "mechanical" version of Bayes Theorem: values in the resulting sample must first be simulated and must then be accepted; the probability of this occurring is proportional to $\pi(\theta) \times \mathrm{P}(\mathrm{data}|\theta)$, i.e. prior times likelihood.

This algorithm isn't generally useful, because the (marginal) likelihood $\mathrm{P}(\mathrm{data}|\theta)$ isn't usually tractable. However, we can bring it closer to practical usefulness by also simulating the entire genealogical tree G under the assumed model. Then, we accept or reject the pair (G, θ) according to the conditional likelihood $\mathrm{P}(\mathrm{data}|G, \theta)$. The accepted pairs form a random sample from the joint posterior distribution of G and θ. To obtain a posterior random sample from a marginal of interest it suffices merely to ignore the other simulated values. For example, to obtain a marginal posterior sample for θ, the values of G are ignored.

The conditional likelihood $\mathrm{P}(\mathrm{data}|G, \theta)$ can often be evaluated using some version of the so-called "peeling algorithm", which requires successively summing over the possible haplotypes at each internal node. Although feasible, this is usually computationally expensive, except in the case of the infinite-sites mutation model with known ancestral haplotype. In this case, given G the data determine uniquely the branches on which the mutations occur, and the conditional likelihood for a single locus then takes the form:

$$\mathrm{P}(\mathrm{data}|G, \theta) = \prod_{i \notin M} \exp(-\theta l_i/2) \prod_{j \in M} (1 - \exp(-\theta l_j/2)) \,,$$

where M denotes the set of branches on which a mutation arose.

3.2 Likelihood approximation via summary statistics

A further simplification arises if we replace the full data with the summary statistic S. As noted above, this replacement entails some loss of information, but the resulting approximate algorithm is very easy to code and quick to run. It combines some of the advantages of a full Bayesian approach with the computational convenience of methods based on summary statistics.

We present a version of the algorithm in which θ is replaced with N and μ.

Rejection Algorithm 2:

1. simulate μ and N from their joint prior;

2. simulate G from the chosen model (e.g. standard coalescent);

3. calculate L, the total branch length of G;

4. accept (G, N, μ) if

$$u < \frac{\text{Pois}(s, N\mu L)}{\text{Pois}(s, s)}, \tag{7}$$

where s is the observed value of S and $\text{Pois}(m, \lambda)$ denotes the probability that a Poisson(λ) random variable takes value m.

Rejection Algorithm 2 has been implemented in the following S-plus code. Rather than output the entire tree G, here we output only its height, $\sum_{i=1}^{n-1} w_i$, which is the (scaled) time since the MRCA of the sample.

```
qrej <- function(nacc=10000, nblk=round(nacc/2), nsamp=6, s=3)
{
  ns1 <- nsamp-1
  rate <- (ns1:1)*(nsamp:2)/2
  acc <- matrix(0,1,3)
  while(nrow(acc)<nacc+1)
  {
    w <- matrix(rexp(ns1*nblk,rate),ns1,nblk)
    TMRCA <- apply(w,2,sum)
    L <- apply((nsamp:2)*w,2,sum)
    u <- runif(nblk)
    N <- rgamma(nblk,5,10^{-3})
    mu <- rgamma(nblk,2,10^4)
    ind <- u<dpois(s,L*N*mu)/dpois(s,s)
    acc <- matrix(c(acc[,1],N[ind],acc[,2],mu[ind],acc[,3],TMRCA[ind]),,3)
  }
  acc[2:(nacc+1),]
}
```

One small difficulty with rejection algorithms is that the number of acceptances is random, whereas the user usually wants a pre-determined size for the posterior sample. This is overcome in the above algorithm by the use of `nblk`: batches of size `nblk` are processed until at least `nacc` acceptances have been achieved.

Some results of a run of `qrej` using the default settings are shown Figure 2. Results are included both for N and μ separately, and for twice their product, θ. The density curves have been obtained using Gaussian kernel density estimation, implemented via the S-plus `density` command, except for the prior density curves for N and μ which were created using the S-plus `dgamma` command.

The top left panel of Figure 2 gives a scatter plot of 100 accepted (N, μ) values, indicating their negative posterior correlation (they are independent *a priori*). A

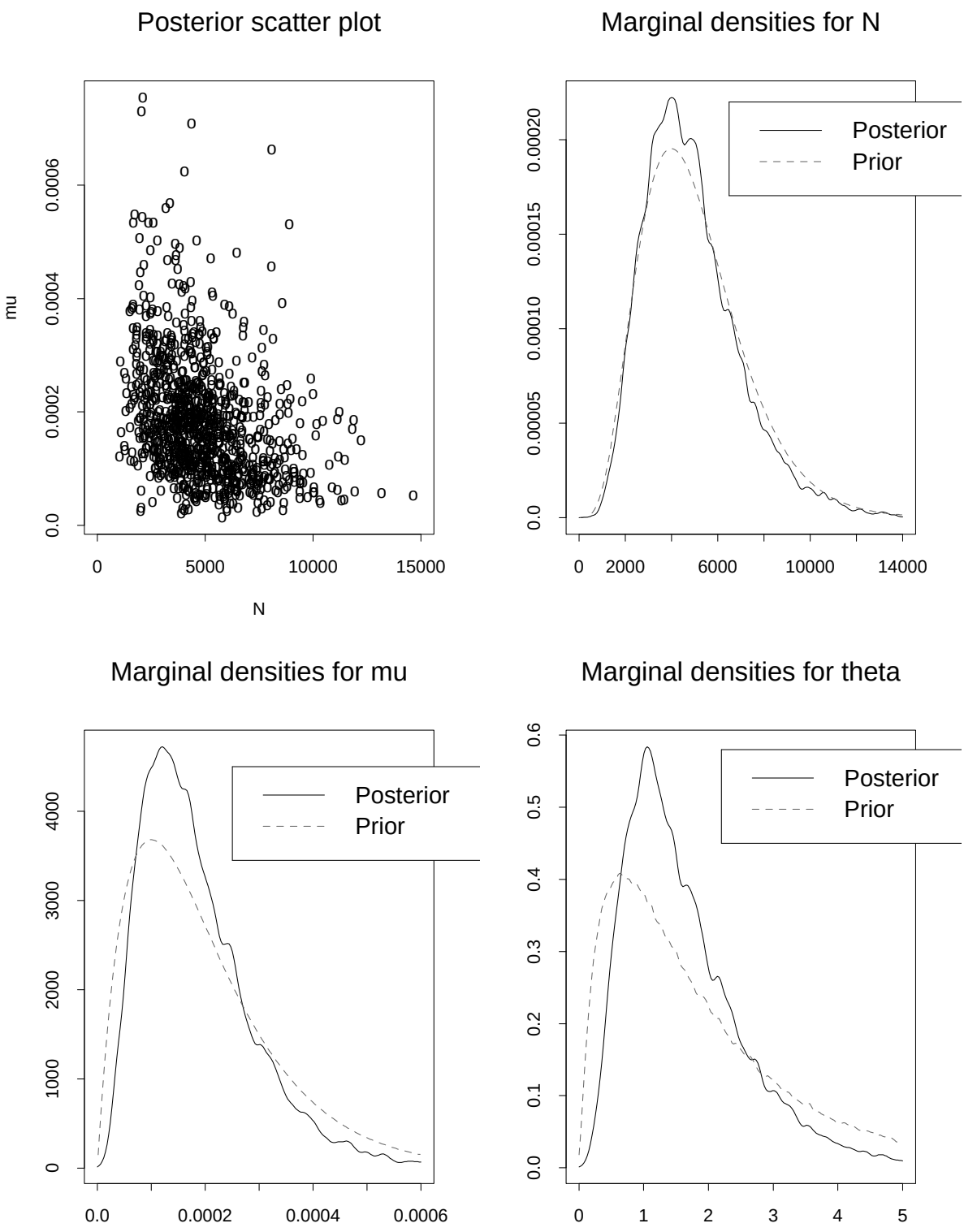

Figure 2: Some results from a run of `qrej` using the default settings

striking feature of the marginal posterior densities shown in the other three panels is that the posteriors appear little different from the priors, indicating that a single observation of $s = 3$ segregating loci gives little information about the parameters of interest here. The most information is available about θ. Its prior expectation of 2 is changed by the observation $s = 3 < \mathrm{E}[S] = 4{\cdot}57$, to a posterior expectation of 1·66. Moreover, the prior standard deviation of 1·8 is reduced to a posterior value of 1·0.

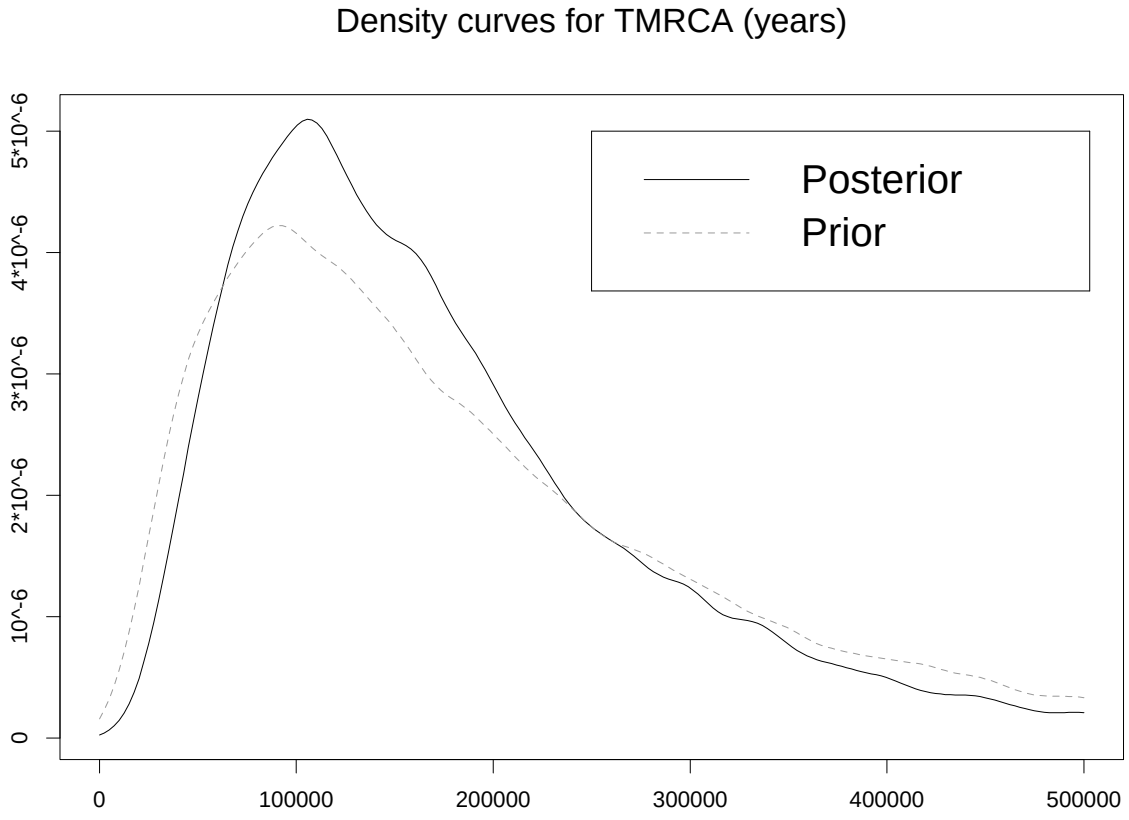

Figure 3: Results for TMRCA from a run of `qrej` with default settings. Generation time is 25 years.

We turn now to posterior inferences about the TMRCA. Since we now have posterior values for N, it becomes possible to report the distribution of the TMRCA in generations. If we choose a generation time, say 25 years, this can be converted to a distribution for the TMRCA in years. Prior and posterior density curves for the TMRCA are shown in Figure 3. Again, there appears to be little information in an observation of $s = 3$ so that the posterior does not seem substantially altered from the prior. However, the mean has been reduced to 180 ky (kilo-years) from

its prior value of 210 ky, and the posterior standard deviation is 130 ky compared with the prior value of 170 ky.

3.3 Likelihood approximation via simulation

The method of approximating the conditional likelihood based on summary statistics, introduced above in section 3.2, has given us our first opportunity for realistic statistical inference. However, it remains unsatisfactory because of the information loss in replacing the full data with a single summary statistic and because it is limited to situations in which the likelihood based on the summary statistic can be readily evaluated.

The likelihood calculation can often be replaced by a further simulation as follows.

Rejection Algorithm 3:

1. simulate all relevant unknowns, both parameters of interest and nuisance parameters, from the prior and genealogical model;

2. given the values generated at step 1, simulate a dataset under the mutation model;

3. accept the parameters of interest if the simulated dataset matches the observed dataset, otherwise reject.

Rejection Algorithm 3 is essentially exact, since the probability of accepting the simulated parameter values is equal to the probability of generating the data given the unknowns, i.e. the likelihood. However, the resulting posterior approximation will usually be poor because acceptances are extremely rare.

To improve the acceptance rate, we weaken the requirement for an exact match of simulated with observed datasets, and require only that they are "similar" in some appropriate sense. The natural way to formalise "similar" is in terms of summary statistics once again, but now that likelihood computation isn't required we can use a vector $\mathbf{S}$ of any desired summary statistics. We accept the parameters of interest if

$$\|\mathbf{S}^* - \mathbf{s}\| \leq \delta,$$

where δ is some pre-assigned tolerance, $\|\cdot\|$ is an appropriate metric, and $\mathbf{S}^*$ and $\mathbf{s}$ denote simulated and observed values of $\mathbf{S}$.

Returning to the example dataset introduced in section 1.2, we resume the framework of section 3.2: the standard coalescent model with infinite-sites mutation. We regard N and μ as parameters of interest, with G a nuisance parameter. As vector of summary statistics $\mathbf{S}$, we choose to use: