Likelihood Function of Time-Dependent Coalescent Models

Emmanuel Paradis

March 3, 2015

Coalescent models describe the distribution of ancestry in a population under some assumptions on the variation in the parameter $\Theta = 2N\nu$, with N the number of alleles in the population and ν the neutral mutation rate. The present document gives the likelihood function, and some computational details, for several models with Θ varying through time. These models are available in coalescentMCMC as R functions (see below).

The general mathematical framework has been given by Griffiths & Tavaré [1]. If Θ is constant, the probability of observing the coalescent times $t_1, \dots, t_n$ is:

$$\prod_{i=2}^{n} \binom{i}{2} \frac{1}{\Theta} \exp\left[-2\binom{i}{2}\frac{t_i - t_{i-1}}{\Theta}\right]$$

where $t_1 = 0$ is the present time and $t_n = T_{\mathsf{MRCA}}$.[1] The general formula for $\Theta(t)$ varying through time is:

$$\prod_{i=2}^{n} \binom{i}{2} \frac{1}{\Theta(t_i)} \exp\left[-2\binom{i}{2}\int_{t_{i-1}}^{t_i} \frac{1}{\Theta(u)} \mathrm{d}u\right] \tag{1}$$

Four specific temporal models are considered below. The time to the most recent ancestor, denoted as T_{MRCA}, is assumed to be known—this assumption could be relaxed though this is not considered here.

1 Models

The *exponential growth model* assumes $\Theta(t) = \Theta_0 e^{\rho t}$, with Θ_0 is the value of Θ at present and ρ is the population growth rate [2]. Because of the exponential function, Θ may reach very high (or low) values. To avoid this, the *linear model* formulated as $\Theta(t) = \Theta_0 + t(\Theta_{T_{\mathsf{MRCA}}} - \Theta_0)/T_{\mathsf{MRCA}}$. This model, like the previous one, has two free parameters: Θ_0 and $\Theta_{T_{\mathsf{MRCA}}}$.

The third model (*step model*) assumes two constant values of Θ before and after a point in time denoted as τ:

$$\Theta(t) = \begin{cases} \Theta_0 & t \le \tau \\ \Theta_1 & t > \tau \end{cases}$$

[1] Note that $2\binom{i}{2} = i(i-1)$ which is easier to calculate.

The last model (*exponential double growth model*) assumes that the population experienced two different phases of exponential growth:

$$\Theta(t) = \begin{cases} \Theta_0 e^{\rho_1 t} & t \le \tau \\ \Theta(\tau) e^{\rho_2 (t-\tau)} = \Theta_0 e^{\rho_2 t + (\rho_1 - \rho_2)\tau} & t \ge \tau \end{cases}$$

which reduces to the first model if $\rho_1 = \rho_2$. These two last models have three free parameters.

1.1 Constant-Θ Model

The log-likelihood is:

$$\ln L = \sum_{i=2}^{n} \ln \binom{i}{2} - \ln \Theta - 2\binom{i}{2} \frac{t_i - t_{i-1}}{\Theta}.$$

Its partial derivative with respect to Θ is:

$$\frac{\partial \ln L}{\partial \Theta} = \sum_{i=2}^{n} -\frac{1}{\Theta} + 2\binom{i}{2} \frac{t_i - t_{i-1}}{\Theta^2},$$

which, after setting $\partial \ln L / \partial \Theta = 0$ can be solved to find the maximum likelihood estimator (MLE):

$$\widehat{\Theta} = \frac{2}{n-1} \sum_{i=2}^{n} \binom{i}{2} (t_i - t_{i-1}).$$

Under the normal approximation of the likelihood function, the variance of $\hat{\Theta}$ is calculated through the second derivative of $\ln L$:

$$\frac{\partial^2 \ln L}{\partial \Theta^2} = \sum_{i=2}^{n} \frac{1}{\Theta^2} - 4 \times \binom{i}{2} \frac{t_i - t_{i-1}}{\Theta^3},$$

and:

$$\widehat{\mathrm{var}}(\widehat{\Theta}) = -\left[\frac{n-1}{\widehat{\Theta}^2} - \frac{4}{\widehat{\Theta}^3} \sum_{i=2}^{n} \binom{i}{2} (t_i - t_{i-1}) \right]^{-1}.$$

This estimator is implemented in `pegas` with the function `theta.tree`.

1.2 Exponential Growth Model

The integral in equation (1) is:

$$\int_{t_{i-1}}^{t_i} \frac{1}{\Theta(u)} \mathrm{d}u = -\frac{1}{\rho \Theta_0} (e^{-\rho t_i} - e^{-\rho t_{i-1}}),$$

leading to the log-likelihood:

$$\ln L = \sum_{i=2}^{n} \ln \binom{i}{2} - \ln \Theta_0 - \rho t_i + 2\binom{i}{2} \frac{1}{\rho \Theta_0} (e^{-\rho t_i} - e^{-\rho t_{i-1}}),$$

with its first partial derivatives being:

$$\begin{aligned}\frac{\partial \ln L}{\partial \Theta_0} &= \sum_{i=2}^{n} -\frac{1}{\Theta_0} - 2\binom{i}{2}\frac{1}{\rho\Theta_0^2}(e^{-\rho t_i} - e^{-\rho t_{i-1}}),\\ \frac{\partial \ln L}{\partial \rho} &= \sum_{i=2}^{n} -t_i + 2\binom{i}{2}\frac{1}{\Theta_0}\left[-\frac{1}{\rho^2}(e^{-\rho t_i} - e^{-\rho t_{i-1}}) + \frac{1}{\rho}(-t_i e^{-\rho t_i} + t_{i-1}e^{-\rho t_{i-1}})\right].\end{aligned}$$

These cannot be solved analytically to find the MLEs $\widehat{\Theta}_0$ and $\hat{\rho}$ but they may be used to speed-up an optimization procedure with analytical gradients.

1.3 Linear Growth Model

Let $\kappa = (\Theta_{T_{\mathsf{MRCA}}} - \Theta_0)/T_{\mathsf{MRCA}}$, so $\Theta(t) = \Theta_0 + \kappa t$. The integral in equation (1) is:

$$\begin{aligned}\int_{t_{i-1}}^{t_i} \frac{1}{\Theta(u)}\mathrm{d}u &= \frac{\ln(\Theta_0 + \kappa t_i)}{\kappa} - \frac{\ln(\Theta_0 + \kappa t_{i-1})}{\kappa}\\ &= \frac{1}{\kappa}\ln\frac{\Theta_0 + \kappa t_i}{\Theta_0 + \kappa t_{i-1}}.\end{aligned}$$

The log-likelihood is thus:

$$\ln L = \sum_{i=2}^{n} \ln\binom{i}{2} - \ln(\Theta_0 + \kappa t_i) - 2\binom{i}{2}\frac{1}{\kappa}\ln\frac{\Theta_0 + \kappa t_i}{\Theta_0 + \kappa t_{i-1}}.$$

The partial derivatives can be calculated analytically.

1.4 Step Model

It is easier to calculate the integral in equation 1 with the difference:

$$\int_{t_{i-1}}^{t_i} \frac{1}{\Theta(u)}\mathrm{d}u = \int_{0}^{t_i} \frac{1}{\Theta(u)}\mathrm{d}u - \int_{0}^{t_{i-1}} \frac{1}{\Theta(u)}\mathrm{d}u. \tag{2}$$

The integral from the origin is:

$$\int_0^t \frac{1}{\Theta(u)}\mathrm{d}u = \begin{cases} \dfrac{t}{\Theta_0} & t \leq \tau \\ \dfrac{\tau}{\Theta_0} + \dfrac{t-\tau}{\Theta_1} & t > \tau. \end{cases}$$

This is then plugged into equation 1 with a simple Dirac delta function.

1.5 Exponential Double Growth Model

In this model the inverse of $\Theta(t)$ is:

$$\frac{1}{\Theta(t)} = \begin{cases} \dfrac{e^{-\rho_1 t}}{\Theta_0} & t \leq \tau \\ \dfrac{e^{-\rho_2 t - (\rho_1 - \rho_2)\tau}}{\Theta_0} & t \geq \tau \end{cases}$$

Again, it is easier to calculate the integral in equation (1) with equation (2). The integral from the origin is:

$$\int_0^t \frac{1}{\Theta(u)}\mathrm{d}u = \begin{cases} -\frac{1}{\rho_1\Theta_0}(e^{-\rho_1 t}-1) & t \le \tau \\ -\frac{1}{\rho_1\Theta_0}(e^{-\rho_1\tau}-1) - \frac{1}{\rho_2\Theta_0}[e^{-\rho_2 t-(\rho_1-\rho_2)\tau} - e^{-\rho_1\tau}] & t \ge \tau \end{cases}$$

This is then plugged into equation (1) with a simple Dirac delta function.

2 Simulation of Coalescent Times

It is generally possible to simulate coalescent times from a time-dependent model by rescaling a set of coalescent times simulated with constant Θ, denoted as t, with:

$$t' = \frac{\int_0^t \Theta(u)\mathrm{d}u}{\Theta(0)}.$$

This gives for the exponential growth model [2]:

$$t' = \frac{e^{\rho t}-1}{\rho},$$

for the linear growth model:

$$t' = t + t^2(\Theta_{T_{\mathsf{MRCA}}}/\Theta_0 - 1)/T_{\mathsf{MRCA}},$$

for the step model:

$$t' = \tau + (t-\tau)\Theta_1/\Theta_0 \qquad \text{if } t > \tau,$$

and for the exponential double growth model:

$$t' = \begin{cases} \frac{e^{\rho_1 t}-1}{\rho_1} & t \le \tau \\ \frac{e^{\rho_1\tau}-1}{\rho_1} + \frac{e^{\rho_2 t+(\rho_1-\rho_2)\tau} - e^{\rho_1\tau}}{\rho_2} & t \ge \tau \end{cases}$$

3 Implementation in coalescentMCMC

Five functions are available in coalescentMCMC which compute the likelihood of the constant-Θ model as well as the four above ones:

```
dcoal(phy, theta, log = FALSE)
dcoal.time(phy, theta0, rho, log = FALSE)
dcoal.linear(phy, theta0, thetaT, TMRCA, log = FALSE)
dcoal.step(phy, theta0, theta1, tau, log = FALSE)
dcoal.time2(phy, theta0, rho1, rho2, tau, log = FALSE)
```

The two arguments common to all functions are:

`phy`: a tree as an object of class `"phylo"`;

`log`: a logical value, if `TRUE` the values are returned log-transformed which is recommended for computing log-likelihoods.

The other arguments are the parameters of the models.

References

[1] R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences*, 344:403–410, 1994.

[2] M. K. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149:429–434, 1998.