Estimating the primate divergence time using conditioned birth-and-death processes

Richard D. Wilkinson*,[a], Simon Tavaré[b]

[a] *Department of Probability and Statistics, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield S3 7RH, United Kingdom*
[b] *DAMTP, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, United Kingdom*

Abstract

The fossil record provides a lower bound on the primate divergence time of 54.8 million years ago, but does not provide an explicit estimate for the divergence time itself. We show how the pattern of diversification through the Cenozoic can be combined with a model for speciation to give a distribution for the age of the primates. The primate fossil record, the number of extant primate species, and information about the structure of the primate phylogenetic tree are combined to provide an estimate for the joint distribution of the primate and anthropoid divergence times. To take this information into account, we derive the structure of the birth-and-death process conditioned to have a subtree originate at a particular point in time. This process has a size-biased law and has an immortal line running from the root of the tree to the root of the subtree, with species on the spine having modified offspring and length distributions. We conclude that it is not possible, with this model, to rule out a Cretaceous origin for the primates.

Key words: Conditioned birth-and-death process, size-biased trees, primate divergence time, inference from the fossil record

*Corresponding author
Email addresses: r.d.wilkinson@sheffield.ac.uk (Richard D. Wilkinson), st321@cam.ac.uk (Simon Tavaré)

1. Introduction

Sam Karlin made numerous contributions to the theory of stochastic processes. Among the earliest of these is the now-classical Karlin-McGregor integral representation of the transition function of birth-and-death and related processes (Karlin and McGregor, 1958) and its application to coincidence probabilities (Karlin and McGregor, 1959a,b). For applications of the latter to combinatorics, see Karlin (1988). Many of Sam's results arose in the study of evolutionary or population genetics, beginning with a mathematical analysis of Moran's model (Karlin and McGregor, 1962), a continuous-time analogue of the Wright-Fisher model of gene frequency change in a finite population. A more general class of discrete-time models obtained by conditioning branching processes on a fixed total size was described in Karlin and McGregor (1964). This provided motivation for Cannings' exchangeable models (Cannings, 1974) and their recent developments (Möhle and Sagitov, 2001).

Several of Sam's papers exploited compound stochastic process arguments, in particular to study population models in which new populations arise at the points of non-homogeneous Poisson processes (Karlin and McGregor, 1967); we use a similar approach in the present paper. Karlin and McGregor (1972) gave a prescient argument that we now recognize as a "coalescent method" to derive the celebrated Ewens Sampling Formula (Ewens, 1972). Kingman's elegant formulation of the ancestral structure of neutral population genetics models, the coalescent, appeared in 1982 and its uses are now commonplace in population genetics (cf. Tavaré (2004); Hein et al. (2005); Wakeley (2008)).

One of us (ST) was a postdoc of Sam's at the time computational molecular biology was coming into its own. We still worked on stochastic problems in population genetics (such as Karlin and Tavaré (1982)), but DNA sequencing had become a reality (although for a while the data came in *books* – ST remembers typing all 48,502 basepairs of bacteriophage lambda into a text file!) and there were new problems to think about. Sam's interests moved towards statistical issues in sequence analysis, resulting first in Karlin et al. (1983), and remained there for the rest of his life. Nonetheless, he still had time for questions about stochastic processes. With this in mind, we think Sam would have liked the problem (and perhaps even the approach) we describe in our paper, which we dedicate to his memory.

2. Estimating divergence times by using the fossil record

The crown divergence time of a monophyletic group of species is the most recent time at which all the species shared a common ancestor. Informally, one can think of the divergence time as the point at which a single ancestor species first diverged into two or more distinct species. Thinking in terms of phylogenetic trees, estimating divergence times is essentially a problem of how to learn the depth of a tree from an incomplete snapshot of its various parts.

In this paper we show how to date the divergence time of a taxonomic group using the fossil record. This is an important problem as fossil evidence is the only direct source of information about the age of a group of species. Genetic data do not explicitly contain any information about age; dating methods that use DNA rely on one or more dates estimated from the fossil record in order to calibrate the speed of mutation in a dating model (the so-called molecular clock). Fossils, on the other hand, can be dated to provide tangible evidence of the existence of a species at a particular point in time. However, fossil evidence only provides a lower bound on the age of a group, with the divergence time of a taxon bounded above by the age of the oldest fossil. For well-sampled taxa with relatively complete fossil records, such as marine invertebrates, it is likely that this lower bound will be close to the true divergence age (Raup and Sepkoski, 1982). However, for poorly sampled taxa, a category including most terrestrial vertebrates, we intuitively expect the temporal gap between the divergence time and the oldest fossil discovery to be more variable and potentially much longer than for well-sampled taxa.

While fossil data do not explicitly provide an upper bound on the age of a clade, the pattern of fossil finds can provide information about how the diversity (number of species) of the clade varied through time. This signal will often be highly noisy, and using it to infer the true diversity is complicated by not knowing the completeness of the fossil record and by the belief that the fossil sampling and discovery rate varies over the geologic time scale (Raup, 1979). However, by modelling diversification and fossil preservation we can use the fossil record, along with other information such as the modern diversity and the known phylogenetic structure, to estimate the divergence time of a clade. We can then give a probability distribution for the divergence time which represents our remaining uncertainty given the data, giving a credibility interval for the range and estimating the most likely divergence time.

2.1. The primate fossil record

We extend the work of Tavaré et al. (2002), estimating the joint distribution of the primate and anthropoid divergence times. The estimation of these divergence times merits special care and attention because of the debate about the primate divergence time that has taken place in recent years. The argument has been characterized by Benton (1999) as 'molecules versus morphology' and concerns whether the primates coexisted with the dinosaurs during the Cretaceous over 65 million years (My) ago. Direct readings of the fossil record tend to place the divergence time in the Cenozoic (Gingerich and Uhen, 1994; Kay et al., 1997), whereas molecular dates tend to place the divergence time in the Cretaceous (Kumar and Hedges, 1998; Arnason et al., 1996; Hedges et al., 1996; Bininda-Emonds et al., 2007). There are sound reasons for why some disparity is expected between the two dating methods, as genetic dates record when inter-breeding ceased, whereas fossils date when morphological difference arose. However, this cannot account for the magnitude of the difference and there is reason to believe that date estimation from fossil evidence can be improved (Martin, 1993). The completeness of the primate fossil record (the proportion of species preserved as fossils) has been estimated to be less than 10% by Martin (1990) and as noted above, for incomplete taxa the temporal gap between oldest fossil and divergence time will be stochastically large.

Table 1 shows the available primate fossil data. It consists of a collection of counts of the number of distinct primate species in each of the past 14 geologic epochs, along with the number of extant primate species (reported in Groves (2001) and Groves (2005)). It also gives the number of anthropoid species known from the fossil record. These data are an unpublished updated version of the data given in Tavaré et al. (2002). The anthropoids are an infraorder of the primates consisting of the new and old world monkeys and the apes, and they form a monophyletic subtree in the primate phylogeny (so that $D_i \geq A_i$ for all i). Further information on the primates and along with information about the data is available in Martin et al. (2007). There are two important points to note from the data: no primate fossil predating the Eocene has been found, with the oldest primate fossil being at most 54.8 My old and no anthropoid fossil has been found before the Late-Eocene, with the oldest anthropoid fossil being at most 37 My old. Throughout this paper, we let τ denote the temporal gap between the oldest primate fossil and the primate crown divergence time, so that the primate divergence occurred $54.8 + \tau$ My ago. We similarly define τ^* to be the temporal gap

between the oldest anthropoid fossil and the anthropoid divergence time, so that the anthropoid divergence time was $37+\tau^*$ My ago. Figure 1 is a simple illustration showing this structure.

Table 1: A summary of the number of primate and anthropoid species known from the fossil record. Time during the Cenozoic is divided into 14 geologic epochs, with the dates for each epoch given in the table in millions of years (My). Also given is the modern diversity.

Epoch	k	Time at base of interval k (My)	Primate fossil counts, $\mathcal{D}$	Anthropoid fossil counts, $\mathcal{A}$
Extant	0		376	281
Late-Pleistocene	1	0.15	22	22
Middle-Pleistocene	2	0.9	28	28
Early-Pleistocene	3	1.8	30	30
Late-Pliocene	4	3.6	43	40
Early-Pliocene	5	5.3	12	11
Late-Miocene	6	11.2	38	34
Middle-Miocene	7	16.4	46	43
Early-Miocene	8	23.8	34	28
Late-Oligocene	9	28.5	3	2
Early-Oligocene	10	33.7	22	6
Late-Eocene	11	37.0	30	2
Middle-Eocene	12	49.0	119	0
Early-Eocene	13	54.8	65	0
Pre-Eocene	14		0	0

Aside from the fossil data, there are other sources of information that can be utilized. Firstly, the modern diversity can inform us about fossil sampling rates and the completeness of the record. Secondly, morphological considerations can allow for the identification of some phylogenetic structure. For example, it is known that the anthropoids are a monophyletic subgroup of the primates, so that the anthropoid phylogenetic tree is a subtree of the primate tree, as shown in Figure 1. Knowing this structure can inform our beliefs about the placement of the root and shape of the tree. Molecular evidence can also provide information about the divergence time, although

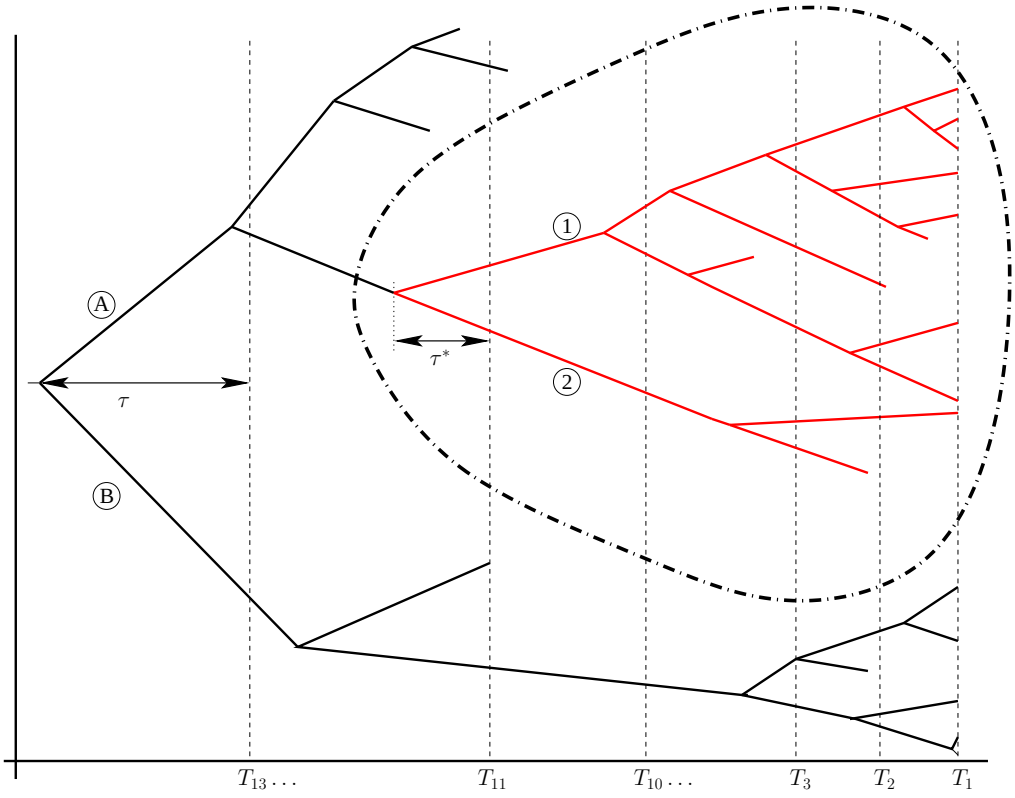

Figure 1: A sample primate speciation tree with anthropoid-subtree highlighted. Here, tree A represents the haplorhini and tree B the strepsirrhini, while subtrees 1 and 2 represent the platyrrhine and catarrhine species. Parameter τ records the distance between the base of the Eocene and the base of the primate tree, whereas τ^* records the distance between the base of the Late-Eocene and the anthropoid subtree.

we do not explore that route here.

The focus in this paper is on combining the information in the fossil record with the modern diversity and on using the known phylogenetic structure to date multiple divergence times simultaneously. By using this structure we hope to date two divergence times with more accuracy than is possible in dating a single divergence time. Also, by giving the joint distribution, it is possible to quantify the joint distribution of the error terms, offering a potential improvement in accuracy if these dates are used as calibration nodes in subsequent molecular analyses. We model both the primate and anthropoid divergence times with the aim of learning how these times can

be constrained given our model and the data. We take a forwards modelling approach, giving a model for speciation and fossil discovery, and then fit the model to the data to learn about the temporal gaps τ and τ^*.

3. Modelling speciation

In order to combine the fossil record, the number of extant species, and the known phylogenetic structure to estimate divergence times, we need a model which incorporates all three aspects. We take a forwards modelling approach, explicitly modelling speciation using a simple stochastic birth-and-death process. Although it is easy to criticise the model, it should be borne in mind that this is an advance over previous approaches to dating using the fossil record, which tend to have been statistical approaches relying on correlations, rather than process models. It is also unclear, due to the limited data available, whether a more complex modelling approach is feasible. Our model can then be used to asses the range of uncertainty one can expect for the temporal gap between the oldest fossil and the divergence time.

We now describe the basic model, which is then conditioned to account for the known phylogenetic structure. The notation and development follow that given in Harris (1963). We consider the birth-and-death process to be an evolving tree process, with each lineage in the tree representing a different species. In order to describe the dynamics of the process, it will be useful to have the following definition of an exponential distribution with time-varying rate.

Definition. Let $b(\cdot)$ denote a positive integrable function with $\int_s^\infty b(t)\mathrm{d}t = \infty$ for all s. We say the random variable X has an inhomogeneous exponential distribution begun at time s, and write $X \sim \mathrm{Exp}_s(b(\cdot))$, if X has the probability density function

$$\pi_s(x) = b(s+x)\exp\left(-\int_s^{s+x} b(t)\mathrm{d}t\right), \quad x > 0.$$

We consider the inhomogeneous birth-and-death process. Each lineage lives for an inhomogeneous exponential period of time with variable rate $b(t) = \lambda(t) + \mu(t)$. Upon the death of a species at time t, it is replaced with either zero (a death) or two (a birth) new species with probabilities

$$p_0(t) = \frac{\mu(t)}{\lambda(t)+\mu(t)} \text{ and } p_2(t) = \frac{\lambda(t)}{\lambda(t)+\mu(t)} \tag{1}$$

respectively. If we denote the number of species alive at time t by $Z(t)$, the process can be described in terms of the infinitesimal change equations

$$Z(t+h) = \begin{cases} Z(t)+1 & \text{w.p. } Z(t)\lambda(t)h + o(h) \\ Z(t)-1 & \text{w.p. } Z(t)\mu(t)h + o(h) \\ Z(t) & \text{w.p. } 1 - Z(t)(\lambda(t)+\mu(t))h + o(h) \end{cases}$$

completing the description of the basic model.

In order to date two or more divergence times simultaneously, we must be able to include any known phylogenetic structure into the model. The type of information that is typically available is that a subgroup of the species form a subtree within the main tree. One approach to using this information is to find post-hoc within the tree, the most likely subtree; we can simulate a sample tree for the complete phylogeny, then exhaustively search all subtrees to find the subtree that most closely matches the data for the subgroup. We then measure the divergence time from this optimal subtree. The problem with this approach is that it is difficult to interpret the results. The optimal subtree is in some sense the closest match to the data, but is not interpretable as a posterior distribution or in any other standard way.

A more satisfactory approach to modelling subtree origination is to condition the birth-and-death process to have a subtree originate at a given point in time. We can then draw divergence times from their prior distributions, and simulate a sample tree rooted at those divergence times.

3.1. Modelling subtree origination

We now consider the effect that conditioning a birth-and-death process on subtree origination has on the size distribution of the process, where by the size distribution is meant $\{\mathbb{P}(Z(t) = k), k = 0, 1, \ldots\}$. Values for the unconditioned process were given by Kendall (1948). Initially, consider a process which begins from one individual at time zero, $Z(0) = 1$. We can then derive the distribution of $Z(t)$ conditional on the event that a birth occurs at time $y > 0$. Extension of the result to processes which begin with n lineages at time 0 follows from an application of Lemma 2.

Conditioning a birth-and-death process to have a subtree originate at time y is equivalent to conditioning the process to have a branch die at time y and give birth to at least two new lineages. Let $B(y)$ denote the event that one of the branches of $Z(\cdot)$ dies at time y and gives birth to at least two new lineages. The distribution of the conditioned process is then given

by $\mathbb{P}(Z(t) = k|B(y))$ for $k = 0, 1, 2, \ldots$. As we are considering a continuous-time birth-and-death process $\mathbb{P}(B(y)) = 0$, so to calculate the conditional probability $\mathbb{P}(Z(t) = k|B(y))$ we consider the limit

$$\lim_{h \downarrow 0} \mathbb{P}(Z(t) = k|B(y, y + h))$$

where $B(y, y + h)$ is the event that a birth occurs during the time interval $[y, y + h]$. To describe the distribution $\mathbb{P}(Z(t) = k|B(y))$, we make use of the following definition.

Definition [Size-biased random variables]. Let X be a discrete positive random variable with $\mathbb{P}(X = j) = p_j$ and $\mathbb{E}(X) < \infty$. Then the size-biased version, $\widehat{X}$ say, has distribution

$$\mathbb{P}(\widehat{X} = j) = \frac{jp_j}{\mathbb{E}X}, \quad j = 1, 2, \ldots.$$

The following lemma, whose proof is given in the appendix, gives the size distribution for the conditioned process.

Lemma 1. *The continuous-time birth-and-death process $Z(t)$ starting from $Z(0) = 1$ and conditioned to have a subtree originate at time y has the size distribution of a size-biased process up to time y, and a standard process from time y onwards.*

$$\begin{aligned}
&\mathbb{P}(Z(t) = j|B(y)) \\
&= \frac{j\mathbb{P}(Z(t) = j)}{\mathbb{E}Z(t)} \quad \textit{for} \quad 0 \le t \le y, &(2)\\
&= \sum_{k=1}^{\infty} \mathbb{P}(Z(t) = j|Z(y) = k + 1) \frac{k\mathbb{P}(Z(y) = k)}{\mathbb{E}Z(y)} \ \textit{for} \quad t > y. &(3)
\end{aligned}$$

Equation (2) says that up until the conditioned subtree origination at time y the conditioned process has a size-distribution which is the size-biased version of the original distribution. Equation (3) says that after the conditioned birth, the process evolves as a standard birth-and-death process begun from $k + 1$ lineages. Note that if we write

$$F_i(s, u, t) = \sum_{j \ge 0} \mathbb{P}(Z(t) = j|Z(u) = i)\, s^j, \tag{4}$$

then the distribution on the right of (2) has probability generating function (pgf)

$$F^{\mathrm{SB}}(s,t) = \frac{s\frac{\mathrm{d}}{\mathrm{d}s}F_1(s,0,t)}{\mathbb{E}Z(t)}, \quad t < y. \tag{5}$$

The calculations above are for the birth-and-death process starting from a single individual at time 0. The size-process of trees which start with more than one individual can be found using the following lemma, proved in Brown (2000).

Lemma 2. *Let $X_1, X_2, \ldots, X_k$ be discrete independent positive random variables with $\mathbb{E}X_j = \lambda_j < \infty$, and let $\widehat{\ }$ denote the size-biasing operator. The size-biased version of the sum $S = X_1 + \cdots + X_k$ has the following distribution:*

$$\widehat{S} =_d X_1 + \cdots + X_{J-1} + \widehat{X}_J + X_{J+1} + \cdots + X_k$$

where the random variable J is independent of the X_j and has distribution

$$\mathbb{P}(J = j) = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_k}, \quad j = 1, 2, \ldots, k.$$

The branching property says that a birth-and-death process with $Z(0) = n$ can be considered as the sum of n independent, identically distributed birth-and-death processes, $Z_1(\cdot), \ldots, Z_n(\cdot)$, with $Z_i(0) = 1$ for each i. Lemmas 1 and 2 then imply that a birth-and-death process with $Z(0) = n$ conditioned to having a subtree originate at time y, can be considered as follows:

- For $t < y$, the process is the sum of $n - 1$ independent standard birth-and-death processes starting from a single branch at time 0, and a size-biased birth-and-death process begun from a single lineage.
- For $t > y$, the process evolves as a standard birth-and-death process.

The law of the size process of the conditioned tree, $\mathbb{P}(Z(t) = j|B(y))$, is of limited value by itself, as it does not describe the structure of the process. As will be described below, we need to be able to simulate observations from the model in order to do inference, and the previous results do not provide a constructive description allowing this to be done easily. We need a more detailed result to describe the structure of the conditioned tree process, which we now provide.

It is possible to show that when we condition on a birth or death at some future time y in the birth-and-death process, we create an immortal line, or distinguished spine, from the root of the tree to time y. Along the immortal spine, the rate of births and deaths is different from the rates along the lineages not on the spine, and the offspring distribution is also modified.

Theorem 1. *Conditioning a birth-and-death process, rooted at time s, to have a death (or birth) occur at time $y > s$, modifies the process as follows:*

1. *There is an immortal line from the root of the tree to the death (or birth) at time y. Any lineage not part of this distinguished spine behaves as in the unconditioned process, independently of all other lineages.*
2. *Lineages on the spine have a modified length distribution, living for a period distributed as an inhomogeneous exponential distribution with rate $2\lambda(t)$ truncated at y, i.e., the length of a lineage born at time x on the spine has density*

$$\pi_x(l) = 2\lambda(x+l)e^{-\int_x^{x+l} 2\lambda(t)\mathrm{d}t}.\mathbb{I}_{x+l<y} + e^{-\int_x^y 2\lambda(t)\mathrm{d}t}.\delta_{y-x}(l).$$

3. *Lineages on the spine have a modified offspring distribution with only births occurring along the spine (i.e. no deaths so the spine cannot die out before time y). One of the two offspring is chosen at random as distinguished. The other offspring evolves as a standard lineage.*

At time y, there is a death on the spine. If conditioning on birth, there is a birth at time y on the spine. For $t > y$, all lineages evolve as standard lineages.

The theorem says that the conditioned process has an spine running from the root to time y. Deaths occur along this spine as the points of a Poisson process with rate $2\lambda(t)$. Deaths on the distinguished spine always lead to the birth of two new species, one of which is distinguished. The undistinguished species evolves as an unconditioned birth-and-death process independently of the spine and other species. This structure has been observed previously in size-biased trees, but under different types of conditioning (Chauvin et al., 1991; Lyons et al., 1995).

We do not prove the theorem here, as its proof is long and complex. But we state a corollary which follows from the theorem by using a compounding argument akin to that described in Karlin and McGregor (1967) and Karlin and Taylor (1981, p406ff).

Corollary 1. *For $t < y$, the pgf of the number of species alive at time t is given by*

$$F^{\mathrm{FB}}(s,t) = s \exp\left\{\int_0^t 2\lambda(u)[F_1(s,u,t) - 1]\,\mathrm{d}u\right\}. \tag{6}$$

In Lemma 3 in the appendix we show the equivalence of the pgfs in (5) and (6) – arguably not obvious at first look! It is possible to generalize Theorem 1 and its corollary to inhomogeneous conditioned Markov branching processes; see Wilkinson and Tavaré (2009) for details.

3.2. Modelling fossil finds

Above we described a model for speciation which allows us to simulate sample family trees that have a subtree originating at a specified point in time. In order to assimilate the fossil data given in Table 1, we need a model for fossil preservation and discovery. This model must act on the simulated trees to produce discrete data that are directly comparable to the recorded data. We do this through a discrete sampling of the branches of the tree. Various models are possible, but we focus on what is perhaps the most natural model and use a Poisson point process to superimpose fossil finds on the branches of the tree. Each species can only be counted once within any epoch, so multiple finds are not counted, but the same species can be discovered in multiple epochs. If the length of time a species lives for in epoch i is l, then under the Poisson sampling scheme the probability that it is preserved and discovered is $1 - \exp(-\beta_i l)$, where β_i is the sampling rate for epoch i. There are reasons to believe that fossil preservation rates vary through time, with fossils from more recent epochs being more likely to be discovered than those from more distant epochs (the "pull of the recent" (Raup, 1979)), and so the sampling rates are allowed to differ between each of the 14 geologic epochs. We treat $\{\beta_i, i = 1, \ldots, 14\}$ as unknown parameters and estimate them along with the other parameters in the next section.

4. Bayesian inference for branching process models

In the previous section we specified a forwards model which can be viewed as a stochastic map $\eta(\cdot)$ from a parameter θ to sample data sets. Our aim is to use the data in Table 1 to learn about θ, finding values which best fit the model and data. In other words, we must solve the inverse problem: given $\eta(\cdot)$ and $\mathcal{D}$, which values of θ are most likely? We take a Bayesian approach

to inference, describing prior distributions for unknown parameters, then computationally inverting the model to find posterior distributions.

Inference for the model described in Section 3 is complicated by the fact that the likelihood function $\pi(\mathcal{D}|\theta)$ is not known explicitly. In Kendall (1948) it was shown that the distribution of the cumulative process of a birth-and-death process is intractable. Without this distribution it is not possible to derive the likelihood of the fossil data, $\pi(\mathcal{D}|\theta)$, under the model. As inference is usually performed using the likelihood function (e.g., in MCMC and maximum likelihood estimation), we are forced here to use a non-standard inference approach. We use a direct inference approach (Diggle and Gratton, 1984) that only requires the ability to simulate from the model.

Approximate Bayesian computation (ABC) methods (Beaumont et al., 2002; Marjoram et al., 2003) are a group of Monte Carlo algorithms used for posterior inference which do not require explicit knowledge of the likelihood function. They use realizations $\eta(\theta)$ from the model and compare these with the data $\mathcal{D}$ to decide whether parameter θ belongs in the posterior sample. Here, we use an ABC algorithm based on the rejection algorithm. To get an approximate sample from the posterior distribution $\pi(\theta|\mathcal{D}) \propto \pi(\mathcal{D}|\theta)\pi(\theta)$ without using evaluations of $\pi(\mathcal{D}|\theta)$, the ABC algorithm can be used as follows:

1. Draw a value of the parameter from its prior $\theta \sim \pi(\cdot)$
2. Simulate data X from the model using parameter θ, $X \sim \eta(\theta)$
3. Accept θ if $\rho(\mathcal{D}, X) \leq \epsilon$.

Here, $\rho(\cdot, \cdot)$ is a distance measure on the output space and ϵ is a tolerance level. If $\epsilon = 0$ then this algorithm is exact, and accepted values of θ are draws from the posterior distribution, whereas if $\epsilon \to \infty$ the algorithm gives draws from the prior distribution. To control the accuracy of the approximation, we take ϵ to be as small as possible. The difficulty arises due to the acceptance rate, with smaller values of ϵ leading to less acceptances in step 3, so that more computation will be required to get a sample of a given size.

The focus of our analysis is on learning the primate and anthropoid divergence times. In the model, these quantities are represented by the two temporal gap parameters τ and τ^*, with the primate divergence time assumed to be $54.8 + \tau$ My ago and the anthropoid divergence time $37 + \tau^*$ My ago. By rooting the tree and subtree at these times in the computer simulation it is possible to simulate sample fossil data generated from phylogenies with these divergence times. To represent prior uncertainty about both variables,

we use flat proper prior distributions, with ranges guided by a combination of pragmatism and expert judgement (R. D. Martin, personnel communication). The priors used were

$$\tau \sim U[0, 50], \qquad \tau^* \sim U[0, 30]$$

and simulation results suggest that the analysis is robust to the range of these priors. The range suggested for τ was $[0, 100]$ My, but experimentation showed that the range $[50, 100]$ contained almost zero posterior mass, and the only effect of using $[0, 100]$ rather than $[0, 50]$ was to double the computational burden. To ensure that τ does represent the primate temporal gap we root the process with two species, and require that in the simulation these species both have modern descendants. Similarly, for the subtree originated at $37+\tau^*$ My ago, we required that both sides of the simulated family tree have extant representatives.

Aside from the two parameters of interest, τ and τ^*, there are numerous other parameters that are required for the model, but which are not of specific interest themselves. Firstly, it is necessary to specify the birth-and-death rates, $\lambda(t)$ and $\mu(t)$, for the speciation model. This can be done by assuming a parametric form for the expected number of species. We set the event rate $b = \lambda(t) + \mu(t)$ to be constant, and then assume a logistic growth model for the diversity through time. Using Equation (8) with the branching property $\mathbb{E}(Z(t)|Z(0) = i) = i\mathbb{E}(Z(t)|Z(0) = 1)$, we equate the expected growth with the logistic growth function, obtaining

$$2\exp\left(b\int_0^t (\lambda(u) - \mu(u))\mathrm{d}u\right) = \frac{2}{\gamma + (1-\gamma)\mathrm{e}^{-\rho t}}$$

for unknown parameters γ and ρ. Using the assumption that $\lambda(t) + \mu(t)$ is constant, this equation can be solved for $\lambda(t)$ and $\mu(t)$ to give expressions for the birth and death rates in terms of γ, ρ and b. We take an empirical Bayes approach for the unknown parameters (γ, ρ, b) and hold them constant at their maximum likelihood values, which were estimated in a previous analysis. We allow different birth-and-death rates on the anthropoid subtree and the encompassing primate tree, so as to allow for differential rates of diversification and the explosion of the anthropoid species sometime in the mid-Cenozoic. For the primate tree we set $\rho = 0.72$ and $\gamma = 0.02$, and for the anthropoid tree we set $\rho = 0.265$ and $\gamma = 0.0065$, and $b = 3$ for both

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$10p_i$	25	26	22	12	5.0	2.7	4.7	3.3	1.6	6.5	18	25	41	3.3

Table 2: Values of the fixed ratio in which the β_i are held, i.e., $\beta_i = \beta p_i$.

trees.

For the sampling rates β_i, we follow Tavaré et al. (2002) and hold the β_i in a fixed ratio, so that $\beta_i = \beta p_i$ for some fixed set of known constants p_i, and unknown parameter β. The values used for p are given in Table 2 and were estimated in a previous analysis. We treat the multiplier β as unknown, and give it a $U[0,1]$ prior distribution and infer its posterior distribution along with the divergence times. Details of how the point estimates for all of these parameters were obtained, along with an alternative ABC approach to estimating posterior distributions of all parameters, are given in Wilkinson (2007).

ABC methods require a distance measure that can be used to compare the model output with the observations. We used the following metric

$$\rho(\mathcal{D}, X) = \sum_{i=1}^{14} \left| \frac{D_i}{D_+} - \frac{X_i}{X_+} \right| + \frac{1}{2} \left| \frac{X_+}{D_+} - 1 \right| + \frac{1}{2} \left| \frac{X_0}{N_0} - 1 \right|.$$

Recall that $\mathcal{D} = (D_1, \ldots, D_{14})$ are the fossil counts from Table 1 and that $X = (X_1, \ldots, X_{14})$ are simulated values of these counts. D_0 and X_0 are the extant number of species observed in the model, $D_+ = \sum D_i$ and $X_+ = \sum X_i$. The first term in the metric is proportional to the total variation distance between the two vectors of proportions $\{D_i/D_+\}$ and $\{X_i/X_+\}$. The $|X_+/D_+ - 1|$ term tries ensure that the total number of fossils found in the simulations is correct, and the $|X_0/D_0 - 1|$ term ensures that the modern diversity is included in the conditioning. Use of this metric gives the posterior distribution of the parameter given the fossil data and the modern diversity, $\pi(\theta|\mathcal{D}, D_0)$. If we were to use the same metric but without the $|X_0/D_0 - 1|$ term then we would find the posterior distribution given the fossil counts only, $\pi(\theta|\mathcal{D})$. The flexibility of ABC algorithms means this is a simple change to make, and it is possible to use the same set of model runs to find both posteriors. To calculate the two different posteriors using a full likelihood calculation would be more difficult (if it were possible), as we would need to calculate the joint distribution of two highly correlated quantities D_0 and $\mathcal{D}$.

The complete inference algorithm used is as follows:

1. Draw a sample value of $\theta = (\tau, \tau^*, \beta)$ from its prior distribution.
2. Simulate a conditioned birth-and-death process starting from two lineages rooted $54.8 + \tau$ My ago, with a subtree originating $37 + \tau^*$ My ago.
3. Check both sides of the tree, and both sides of the subtree, survive to the present. If not, return to step 1.
4. Simulate fossil finds for each epoch using sampling rates $\beta_i = \beta p_i$.
5. Count the fossil finds X on the complete tree. If $\rho(\mathcal{D}, X) \leq \epsilon_1$, go to step 6. Otherwise reject θ and return to step 1.
6. Count fossil finds X' on the subtree. If $\rho(\mathcal{A}, X') \leq \epsilon_2$, where $\mathcal{A}$ are the anthropoid fossil counts, then accept θ into the posterior sample. Otherwise reject θ. Return to step 1.

Step 3 is included in the algorithm so as to condition on non-extinction and to ensure that the specified divergence times do indeed represent the crown divergence time in the simulated phylogeny. We choose tolerance values ϵ_1 and ϵ_2 by taking them as small as is feasible given the computing resources available. Using a cluster of 50 parallel processors, we found $\epsilon = (0.4, 0.4)$ gave a decent acceptance rate for a simulation period of about 10 hours (about 3 weeks of computing time in total) giving a posterior sample of 2,185 values.

The results from the simulations are shown in Figure 2. The posterior 95% credibility interval for the anthropoid divergence time is $[98.9, 54.8]$ My ago, and the posterior 95% credibility interval for the anthropoid divergence time is $[53.2, 37]$ My ago. A Cretaceous origin for the primates corresponds to a divergence time 65 My ago or earlier. This is equivalent to the temporal gap between the oldest fossil discovery and the divergence time, τ, being greater than 10.2 My. Calculating the posterior probability of a Cretaceous origin from the posterior sample, we find $\mathbb{P}(\tau > 10.2 \mid \mathcal{D}) = 0.46$. In other words, we find that a Cretaceous origin for the primates is almost as likely as not. Similarly, we find $\mathbb{P}(\tau^* > 12 \mid \mathcal{D}) = 0.21$ for the posterior probability of an Early-Eocene origin for the anthropoids.

5. Discussion

The answer to the question concerning whether primates existed during the Cretaceous, and hence whether they coexisted with the dinosaurs, has

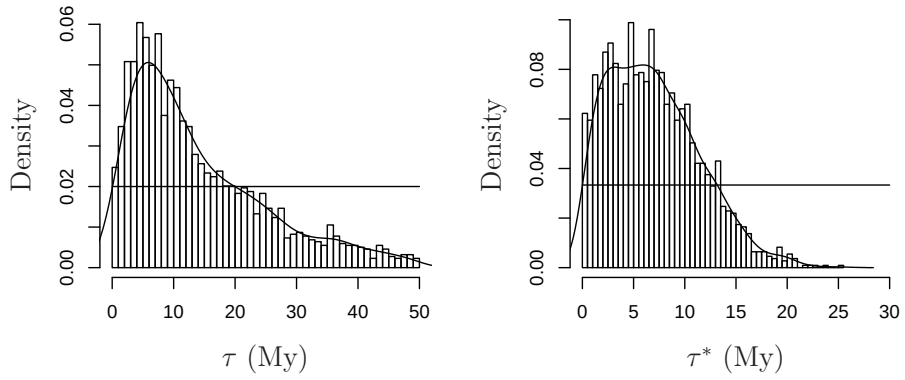

Figure 2: Marginal posterior distributions of the primate and anthropoid divergence time. The histograms are of the raw accepted ABC output, with kernel density estimates overlaid. The horizontal lines are the prior distributions used.

tended to depend on the source of the information held to be of primary importance. Those answering the question who focus on molecular methods using DNA from modern primates have tended to conclude that primates originated in the Cretaceous (Kumar and Hedges, 1998; Arnason et al., 1996; Hedges et al., 1996; Bininda-Emonds et al., 2007). Conversely, those relying on fossil evidence have tended to conclude Cenozoic origins are more likely. The view that the first appearance of a group of species in the fossil record is "... accepted as more nearly objective and basic than opinions as to the time when the group really originated" (Simpson, 1965) still holds weight. What this work has shown is that the primate fossil evidence is not sufficient by itself to make this conclusion. We have aimed to utilize the fossil evidence as fully as possible, extracting structural information and combining it with the modern diversity, and have found Cretaceous origins have significant posterior probability. We used uninformative flat prior distributions so as to clearly observe the signal in the data. Strong prior distributions based on expert opinion (from secondary data sources and accumulated knowledge) may still lead to strong posterior beliefs that a Cenozoic divergence time is much more likely than a Cretaceous divergence. We have simply shown that the data alone are insufficient for this task.

Many other models are possible, and we tried several variants on those

reported here, modelling the K-T crash 65 My ago, trying different beliefs for the expected diversification, and different fossil sampling models. All combinations tried retained significant posterior probability of Cretaceous primate origins (Wilkinson, 2007). It should also be noted that the estimates reported here were found by taking an empirical Bayes approach and fixing many of the unknown parameters at estimated values. This reduces the amount of posterior uncertainty, as we have not fully accounted for all of the uncertainty in the model. If this uncertainty is accounted for, then the posteriors are more diffuse than those reported here.

References

Arnason, U., Gullberg, A., Janke, A., Xu, X. F., 1996. Pattern and timing of evolutionary divergences among hominids based on analyses of complete mtDNAs. J. Mol. Evol. 43, 650–661.

Beaumont, M. A., Zhang, W., Balding, D. J., 2002. Approximate Bayesian computatation in population genetics. Genetics 162, 2025–2035.

Benton, M. J., 1999. Early origins of modern birds and mammals: molecules vs. morphology. BioEssays 21, 1043–1051.

Bininda-Emonds, O. R. P., Cardillo, M., Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., Price, S. A., Vos, R. A., Gittleman, J. L., Purvis, A., 2007. The delayed rise of present-day mammals. Nature 446, 507–512.

Brown, M., 2000. Exploiting the waiting time paradox. Probab. Engrg. Inform. Sci. 20, 195–230.

Cannings, C., 1974. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. Adv. Appl. Probab. 6, 260–290.

Chauvin, B., Rouault, A., Wakolbinger, A., 1991. Growing conditioned trees. Stochastic Process. Appl. 39, 117–130.

Diggle, P. J., Gratton, R. J., 1984. Monte Carlo methods of inference for implicit statistical models. J. Roy. Statist. Soc. Ser. B 46, 193–227.

Ewens, W., 1972. The sampling theory of selectively neutral alleles. Theoret. Popn. Biol. 3, 87–112.

Gingerich, P. D., Uhen, M. D., 1994. Time of origin of primates. J. Hum. Evol. 27, 443–445.

Groves, C. P., 2001. Primate Taxonomy. Smithsonian Institution Press, Washington.

Groves, C. P., 2005. Order primates. In: Wilson, D. E., Reeder, D. M. (Eds.), Mammal Species of the World: A Taxonomic and Geographic Reference, 3rd Edition. Vol. 1. John Hopkins University Press, Baltimore, pp. 111–184.

Harris, T. E., 1963. The Theory of Branching Processes. Springer-Verlag, Berlin.

Hedges, S. B., Parker, P. H., Sibley, C. G., Kumar, S., 1996. Continental breakup and the ordinal diversification of birds and mammals. Nature 381, 226–229.

Hein, J., Schierup, M. H., Wiuf, C., 2005. Gene Genealogies, Variation, and Evolution. Oxford University Press.

Karlin, S., 1988. Coincident probabilities and applications in combinatorics. J. Appl. Probab. 25A, 185–200.

Karlin, S., Ghandour, G., Ost, F., Tavaré, S., Korn, L. J., 1983. New approaches for computer analysis of nucleic acid sequences. Proc. Natl. Acad. Sci. USA 80, 5660–5664.

Karlin, S., McGregor, J., 1958. Linear growth birth and death processes. J. Math. Mech. 7, 643–662.

Karlin, S., McGregor, J., 1959a. Coincidence probabilities. Pacific J. Math. 9, 1141–1164.

Karlin, S., McGregor, J., 1959b. Coincidence properties of birth and death processes. Pacific J. Math. 9, 1109–1140.

Karlin, S., McGregor, J., 1962. On a genetics model of Moran. Proc. Cambridge Philos. Soc. 58, 299–311.

Karlin, S., McGregor, J., 1964. Direct product branching processes and related Markov chains. Proc. Nat. Acad. Sci. USA 51, 598–602.

Karlin, S., McGregor, J., 1967. The number of mutant forms maintained in a population. In: Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66). Univ. California Press, Berkeley, Calif., pp. 415–438.

Karlin, S., McGregor, J., 1972. Addendum to a paper of W. Ewens. Theoret. Popn. Biol. 3, 113–116.

Karlin, S., Tavaré, S., 1982. Linear birth and death processes with killing. J. Appl. Probab. 19, 477–487.

Karlin, S., Taylor, H. M., 1981. A Second Course in Stochastic Processes. Academic Press.

Kay, R. F., Ross, C., Williams, B. A., 1997. Anthropoid origins. Science 275, 797–804.

Kendall, D. G., 1948. On the generalized birth-and-death process. Ann. Math. Stat. 19, 1–15.

Kingman, J., 1982. On the genealogy of large populations. J. Appl. Probab. 19A, 27–43.

Kumar, S., Hedges, S. B., 1998. A molecular timescale for vertebrate evolution. Nature 392, 917–920.

Lyons, R., Pemantle, R., Peres, Y., 1995. Conceptual proofs of $L \log L$ criteria for mean behaviour of branching processes. Ann. Probab. 23, 1125–1138.

Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S., 2003. Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA 100, 15324–15328.

Martin, R. D., 1990. Primate Origins and Evolution: a Phylogenetic Reconstruction. Princeton University Press, New York.

Martin, R. D., 1993. Primate origins: plugging the gaps. Nature 363, 233–234.

Martin, R. D., Soligo, C., Tavaré, S., 2007. Primate origins: Implications of a cretaceous ancestry. Folia Primatol. 78, 277–296.

Möhle, M., Sagitov, S., 2001. A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29, 1547–1562.

Raup, D. M., 1979. Biases in the fossil record of species and genera. Bull. Carnegie Mus. Nat. Hist. 13, 85–91.

Raup, D. M., Sepkoski, J. J., 1982. Mass extinctions in the marine fossil record. Science 215, 1501–1503.

Simpson, G. G., 1965. The Geography of Evolution. Chilton Books, Philadelphia.

Tavaré, S., 2004. Ancestral inference in population genetics. In: Picard, J. (Ed.), Lectures on Probability Theory and Statistics. Ecole d'Etés de Probabilité de Saint-Flour XXXI – 2001. Vol. 1837 of Lecture Notes in Mathematics. Springer Verlag, New York, pp. 1–188.

Tavaré, S., Marshall, C. R., Will, O., Soligo, C., Martin, R. D., 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416, 726–729.

Wakeley, J., 2008. Coalescent Theory. Roberts & Company Publishers.

Wilkinson, R. D., 2007. Bayesian inference of primate divergence times. Ph.D. thesis, University of Cambridge.

Wilkinson, R. D., Tavaré, S., 2009. Conditioning on subtree origination in the Markov branching process, in preparation.

Appendix

Proof of Lemma 1. We assume that $Z(0) = 1$ in what follows. First, consider the growth of the tree before the conditioned subtree origination,

i.e., for $t < y$. By conditioning on the value $Z(y)$ and using Bayes' Theorem we can see that

$$\mathbb{P}(Z(t) = j|B(y)) = \\ \mathbb{P}(Z(t) = j)\lim_{h\downarrow 0}\frac{\sum_{k=1}^{\infty}\mathbb{P}(B(y, y+h)|Z(y)=k)\mathbb{P}(Z(y) = k|Z(t) = j)}{\sum_{k=1}^{\infty}\mathbb{P}(B(y, y+h)|Z(y) = k)\mathbb{P}(Z(y) = k)}$$

Recall that for a population of size k the time to the next birth or death has an exponential distribution with rate $kb(t)$, where $b(t) = \lambda(t) + \mu(t)$, and so $khb(t) + o(h)$ is the infinitesimal probability of a death of one of the k lineages in interval $(t, t + h)$. This gives

$$\mathbb{P}(Z(t) = j|B(y)) = \\ \mathbb{P}(Z(t) = j)\lim_{h\downarrow 0}\frac{\sum(kb(t)h + o(h))\, p_2(y)\,\mathbb{P}(Z(y) = k|Z(t) = j)}{\sum(kb(t)h + o(h))\, p_2(y)\,\mathbb{P}(Z(y) = k)}$$

and dividing through by h, we find that

$$\mathbb{P}(Z(t) = j|B(y)) = \mathbb{P}(Z(t) = j)\frac{\mathbb{E}(Z(y)|Z(t) = j)}{\mathbb{E}Z(y)}.$$

The branching property implies that $\mathbb{E}(Z(y)|Z(t) = j) = j\mathbb{E}(Z(y)|Z(t) = 1)$ and combining this with the tower property of expectation, we find that

$$\begin{aligned}\mathbb{E}Z(y) &= \mathbb{E}[\mathbb{E}(Z(y)|Z(t))] \\ &= \mathbb{E}[Z(t)\mathbb{E}(Z(y)|Z(t) = 1)] \\ &= \mathbb{E}Z(t)\mathbb{E}(Z(y)|Z(t) = 1).\end{aligned}$$

This gives the required size-biased distribution

$$\mathbb{P}(Z(t) = j|B(y)) = \frac{j\mathbb{P}(Z(t) = j)}{\mathbb{E}Z(t)}.$$

Now consider the growth of the process after the conditioned split point, i.e., for $t > y$. We let y_+ denotes the time just after the birth $B(y)$ and y_-

represents the time just before. Then

$$\begin{aligned}\mathbb{P}(Z(t)=j|B(y)) &= \sum_{k=1}^{\infty}\mathbb{P}(Z(t)=j, Z(y_+)=k+1|B(y)) \\ &= \sum_{k=1}^{\infty}\mathbb{P}(Z(t)=j|Z(y_+)=k+1)\mathbb{P}(Z(y_-)=k|B(y)) \\ &= \sum_{k=1}^{\infty}\mathbb{P}(Z(t)=j|Z(y)=k+1)\frac{k\mathbb{P}(Z(y)=k)}{\mathbb{E}Z(y)}.\end{aligned}$$

completing the proof. □

We now establish that the size-biassed distribution given in (2) and the corresponding distribution implicit in Theorem 1 are indeed equivalent. Our approach is to show that the generating functions $F^{\mathrm{FB}}(s,t)$ in (6) and $F^{\mathrm{SB}}(s,t)$ in (5) satisfy the same partial differential equation, and the same boundary conditions, and therefore are equal. First, some standard results for birth-and-death processes: The pgf $F_1(s,u,t)$ satisfies the forward pde

$$\frac{\partial F_1(s,\tau,t)}{\partial t} = (s-1)(\lambda(t)s-\mu(t))\,\frac{\partial F_1(s,\tau,t)}{\partial s}, \quad F_1(s,\tau,\tau+)=s, \tag{7}$$

and so

$$m_1(\tau,t) \equiv \mathbb{E}(Z(t)|Z(\tau)=1) = \exp\left\{\int_\tau^t (\lambda(t)-\mu(t))\,\mathrm{d}u\right\}. \tag{8}$$

Lemma 3. $F^{\mathrm{FB}}(s,t) = F^{\mathrm{SB}}(s,t), \quad 0\le s\le 1, 0\le t\le y.$

Proof. First we derive a pde for

$$G_i^*(s,\tau,t) = \frac{\frac{\mathrm{d}}{\mathrm{d}s}F_i(s,\tau,t)}{\mathbb{E}(Z(t)|Z(\tau)=i)}.$$

Using (8), note that

$$\frac{\partial \mathbb{E}(Z(t)|Z(\tau)=i)}{\partial t} = (\lambda(t)-\mu(t))\,\mathbb{E}(Z(t)|Z(\tau)=i).$$

It follows that

$$\frac{\partial G_i^*}{\partial t} = \frac{1}{\mathbb{E}(Z(t)|Z(\tau)=i)} \left\{ \frac{\partial^2 F_i}{\partial s \partial t} - (\lambda(t) - \mu(t)) \frac{\partial F_i}{\partial s} \right\}.$$

But from (7) we have

$$\frac{\partial^2 F_i}{\partial s \partial t} = (2\lambda(t)s - \lambda(t) - \mu(t)) \frac{\partial F_i}{\partial s} + (s-1)(\lambda(t)s - \mu(t)) \frac{\partial^2 F_i}{\partial s^2}.$$

Substituting, we get

$$\begin{aligned}
\frac{\partial G_i^*}{\partial t} &= \frac{1}{\mathbb{E}(Z(t)|Z(\tau)=i)} \left\{ (2\lambda(t)s - \lambda(t) - \mu(t)) \frac{\partial F_i}{\partial s} \right. \\
&\qquad \left. + (s-1)(\lambda(t)s - \mu(t)) \frac{\partial^2 F_i}{\partial s^2} - (\lambda(t) - \mu(t)) \frac{\partial F_i}{\partial s} \right\} \\
&= \frac{1}{\mathbb{E}(Z(t)|Z(\tau)=i)} \left\{ 2\lambda(t)(s-1) \frac{\partial F_i}{\partial s} + (s-1)(\lambda(t)s - \mu(t)) \frac{\partial^2 F_i}{\partial s^2} \right\} \\
&= (s-1) \left\{ 2\lambda(t) G_i^* + (\lambda(t)s - \mu(t)) \frac{\partial G_i^*}{\partial s} \right\}. \qquad (9)
\end{aligned}$$

The next step is to show that $\hat{F}(s,t) \equiv s^{-1} F^{\mathrm{FB}}(s,t)$ satisfies the pde (9) with $i = 1$. To verify this, write

$$w(s,t) = \int_0^t 2\lambda(u)[F_1(s,u,t) - 1]\mathrm{d}u,$$

and note that

$$\frac{\partial \hat{F}}{\partial t} = \frac{\partial w(s,t)}{\partial t} \hat{F} \text{ and } \frac{\partial \hat{F}}{\partial s} = \frac{\partial w(s,t)}{\partial s} \hat{F}.$$

It follows that we need to show that

$$\frac{\partial w(s,t)}{\partial t} = (s-1) \left\{ 2\lambda(t) + (\lambda(t)s - \mu(t)) \frac{\partial w(s,t)}{\partial s} \right\}. \qquad (10)$$

Now

$$\frac{\partial w(s,t)}{\partial s} = \int_0^t 2\lambda(u) \frac{\partial F_1(s,u,t)}{\partial s} \mathrm{d}u.$$

Use the forward equation (7) to see that the right-hand side of (10) is

$$(s-1)2\lambda(t) + \int_0^t 2\lambda(u)\frac{\partial F_1(s,u,t)}{\partial t}\,\mathrm{d}u.$$

It remains to calculate the left-hand side of (10). But this is

$$\begin{aligned}\frac{\partial w(s,t)}{\partial t} &= \int_0^t \frac{\partial}{\partial t}\,2\lambda(u)(F_1(s,u,t)-1)\mathrm{d}u + 2\lambda(t)(F_1(s,t,t)-1)\\ &= \int_0^t 2\lambda(u)\frac{\partial F_1(s,u,t)}{\partial t} + 2\lambda(t)(s-1),\end{aligned}$$

the last following from the boundary condition in (7). This completes the proof. □