Bayesian Inference of Primate Divergence Times

Richard David Wilkinson

Downing College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

September 2007

Bayesian Inference of Primate Divergence Times

Richard David Wilkinson

This thesis is concerned with how to estimate species divergence times using only the fossil record. Although the methods used are applicable to any taxonomic grouping of species, the focus here is solely on the primates.

The work is motivated by the wide discrepancy between the molecular and palaeontological estimates of divergence times (for primates, molecular estimates of the divergence time are generally about 80-90 million years (My) ago, whereas palaeontological estimates are about 60-65 My ago). While it is known that a direct reading of the fossil record can only ever be a lower bound on the age of a group, there is little theory to suggest what size temporal gap is expected between the oldest fossil, and the divergence time. The work in this thesis estimates the posterior distribution of this temporal gap under a Markovian branching process model of evolution. Several variations on the model are explored, including a model of the mass extinction event that killed the dinosaurs 65 My ago.

One of the main developments is a methodology to estimate the joint distribution of two split points. This allows the structure of the data to be fully utilised, and thus more information can be extracted from the limited data available. The conclusions of the work are that molecular estimates of the divergence time are consistent with the fossil record.

There are two main methodological developments contained in this thesis. The first concerns Approximate Bayesian Computation (ABC), which is a likelihood-free Monte Carlo method that only requires the ability to simulate sample data sets from the model. The basic rejection algorithm approach is inefficient due to repeatedly sampling from the prior distribution. I have developed an ABC-MCMC hybrid algorithm that is more accurate and efficient for the models in which it can be applied.

The second development concerns conditioned branching processes. A fully Bayesian method of inference requires that we simulate Galton-Watson trees conditioned on having a branching event at a given time. I prove a result showing that these conditioned trees are size-biased Galton-Watson trees and I describe their fish-bone like structure.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. This work has not been submitted for a degree, diploma or any other qualification at any other university.

Acknowledgements

Huge thanks go to my supervisor Simon Tavaré for his inexhaustible supply of good humour, patience and ideas. His constant enthusiasm has kept me going throughout the last three years. I would also like to thank Robert Martin at the Field Museum in Chicago, and Christophe Soligo at UCL, for their ideas, data and advice. This dissertation was supported financially by the Engineering and Physical Sciences Research Council.

Finally, I want to thank my wife Libby Wilson for putting up with me while writing this thesis and for looking after me when I needed it most.

Contents

List of Figures

Chapter 1

Introduction

Ever since Darwin published his theory of evolution in *The Origin of Species* scientists have been studying the relationships between humans and other species. His remarkable ideas led to a paradigm shift in human thought, with mankind appreciating its transient role in a continuous cycle of evolution and extinction. Since then a huge amount of time has been spent analysing the primates, and while the picture is clearer, it is still far from complete; for example, debate still rages about when primates first evolved. The prevailing view from palaeontology is that primates diverged from other mammals in the Paleocene after the decline of the dinosaurs 65 million years ago. Recent advances in genetic dating methods, however, estimate the divergence time to be way back in the Cretaceous some 90 million years ago.

The standard approach to inferring divergence times amongst palaeontologists was encapsulated by Simpson [110] when he wrote 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". The motivation behind this thesis is that this principle is misleading and leads to the wrong conclusions.

The primate fossil record has been estimated to be less than 10% complete [77], meaning that of all the species that have ever existed, less than 10% of them have been preserved and then discovered as fossils. While it is intuitively obvious and long been recognised

that the age of the oldest fossil can only be a lower bound on the age of the group, few have studied the expected temporal gap between divergence and first appearance in the fossil record. It is this calculation that is the main focus of this thesis.

While the majority of the thesis is devoted to the principle aim stated above, several theoretical developments are made along the way. The mathematical model I use is unable to be explicitly solved and so a simulation-based Monte Carlo technique known as *Approximate Bayesian Compuation* (ABC) is used to perform inference. ABC is a relatively new methodology that is still in its infancy and there are consequently many unknowns about how best to apply it. I look at a few points of contention and give a development that can increase the speed of the algorithms in certain models. Another theoretical development is contained in the penultimate chapter. I give the structure of Markovian branching processes conditioned to have a death followed by a birth at a given point in time. This allows us to model the divergence of primate suborders such as the anthropoids.

1.1 Estimating Divergence Times Using the Fossil Record

Finding the shape of the tree of life is one of the most fundamental and important problems in the biological sciences. Darwin's idea that all living things can be related to one another via a vast family tree is one of the most vivid and romantic scientific ideas of the last two centuries and has had a profound effect on human beings' understanding of their place in the world. The tree of life (or phylogeny) can be considered as two parts: the shape or topology and the length of each lineage. The shape shows the relationships between each taxon and can be determined by either morphology or molecular methods; the time information, shown through the length of each lineage, gives the time of origin of each taxa. It is this second part that I concentrate on in this thesis.

The primary aim is to develop a methodology for the inference of divergence times using the information contained in the fossil record. The *divergence time* of a clade is the point in time at which the last common ancestor (LCA) of that clade diverged into distinct species. A *clade* is a monophyletic group of species, which means that it is a taxonomic group of organisms that consists of a single common ancestor and all of its descendants.

Throughout the thesis I focus solely on the primates, although the methodology could in theory be applied to any monophyletic group. The primates are of particular interest, not only because they provide the zoological context for human evolution, but also because the primate fossil record is relatively poor compared to many other modern orders [121], making a careful analysis vital.

Dating mammalian divergence times is currently a controversial subject. There are two main approaches, one based on a direct reading of the fossil record, and the other on a

molecular DNA-based approach. Benton [14] characterised the debate between the proponents of the two methods as *molecules versus morphology*. The argument is caused by the large disparity between the estimates given when dating mammalian and avian divergence times, with molecular estimates often twice as large as palaeontological estimates [15, 16]. For example, modern orders of placental (eutherian) mammals are dated to have originated in the Paleocene and Eocene 50-65 My[(1)] ago by palaeontological data [13], and much earlier in the mid-to-late Cretaceous 80-100 My ago by molecular estimates [41, 58]. The proponents of both approaches have spent much time criticising the other group's methods, with neither side seemingly willing to admit defeat.

Argument about primate divergence times is based around whether they diverged from other placental mammals in the Cretaceous or the Cenozoic[(2)], and therefore whether they coexisted with the dinosaurs (which went extinct 65 My ago). Palaeontological estimates for the primate divergence time are based on the oldest primate fossil, which is from the Early Eocene some 55 My ago. The mainstream palaeontology consensus is that the primates originated not long before the earliest known fossil [66], with the 95% confidence interval of 55-63 My ago given by Gingerich and Uhen [51] being fairly typical. This date is supported by the belief that the downfall of the dinosaurs 65 My ago opened up an evolutionary niche that was quickly filled by the primates and other mammals [2].

Molecular dates for the primate divergence time, however, tend to be well back into the Cretaceous, meaning that primates and dinosaurs coexisted. Hedges *et al.* [58] and Kumar *et al.* [70] give a date of at least 90 My ago, Arnason *et al.* [5, 4] suggest the divergence was at least 80 My ago, and Bininda-Emonds *et al.* [19] specifically examine the effect of the end-Cretaceous mass extinction event and calculate a primate divergence time of 87.7 ± 2.7 My. I will now examine some of the main criticisms of each method.

Molecular estimates of divergence times arise from comparing the genomes of different taxa. Although DNA does not contain direct information about time, by comparing the distance between two sequences and assuming clock-like behaviour of genetic mutations, estimates can be given for how far back in the past any two species shared a common ancestor. Critics of molecular methods [15] claim that the molecular clock assumption is violated during periods of explosive radiation, such as at the Cretaceous-Cenozoic boundary 65 My ago. During these periods it is claimed that the molecular clock runs at an accelerated rate and that organisms evolve more quickly than at other times.

An important point to note is that molecular methods require calibration of the molecular clock and that this is done by using a divergence time from another part of the tree. Because DNA does not directly contain any time information, at some point a divergence

[(1)] Throughout this thesis time will be measured in units of millions of years (My).

[(2)] Palaeontologists have divided time into eons, eras, periods and epochs with each representing a finer division of time than the proceeding term. The Cretaceous period ended 65 My ago coinciding with the start of the Cenozoic era. A complete list of currently used terminology and a full explanation can be found in Gradstein *et al.* [54].

time must be estimated using fossil evidence. And so even if we prefer to use molecular estimates of divergence times, we must at some point make an inference using the fossil record. The fossil date used should ideally be the divergence time of a well sampled taxon so that all parties can agree on its validity. But this, as Graur *et al.* [55] showed, is not always the case. They documented numerous examples of molecular divergence times being estimated to a supposedly high-degree of precision using fossil calibration points without any acknowledgement of the possible error in these dates. A particularly startling example was a paper where the arthropod-nemotode divergence was dated to 1167 My ago with a 95% confidence interval 350 My wide. When the uncertainty in the fossil calibration point was taken into account the 95% confidence interval grew to a considerable 14.2 billion years.

Palaeontological estimates of divergence times are obtained from the fossil record and are based on the age of the oldest fossil representative of the group. It is common practice to date the divergence time of a group from just before the first known fossil representative [78]. Notice that there will always be a time gap between the origination of the family and the preservation of the first discovered fossil. For this reason, it is recognised that the fossil record can only give a lower bound on the age of a taxon [78]. It is this range extension (or temporal gap) between the age of a clade and the age of the oldest fossil representative that is the cause of much of the debate. The idea of a temporal gap is demonstrated in Figure 1.1.

The sampling rate, by which I mean the proportion of species that are preserved in the fossil record, has a large effect on the accuracy of the palaeontological estimates. For well sampled clades, which have a large number of fossil representatives, we expect the range extension to be smaller than that observed for a clade which has a low sampling rate. So for example many marine invertebrates, which live in areas of net deposition and have hard bodies, are believed to be well sampled as there are many fossils and so we expect only a small range extension. Primates, which are soft bodied, live in areas of net erosion, and prefer tropical environments not conducive to fossil preservation, have a poor fossil record and so we expect a larger range extension.

It is possible that there is a simple explanation for why molecular and fossil based estimates disagree. Each fossil that is discovered is assigned to a group on the basis of its morphology. This is a difficult and often subjective task, with palaeontologists looking for some defining character, such as some aspect of the skull shape (cf. Figure 1.2) or dentition[(3)]. While molecular estimates date the time at which genetic intermixing ceased, the fossil dates, at best, record the time at which diagnostic characters are obtained [32].

[(3)]The difficult and often controversial nature of this task can be seen in the recent hominoid fossil found on the Indonesian island Flores. Brown *et al.* [25] claim this small bodied creature represents a new hominin (the tribe containing humans and chimpanzees) Homo floresiensis which survived until at least 18000 years ago. On the other hand, Martin *et al.* [82] claim it is a microcephalic modern human. Identifying a species from a single skull fragment, as in this case, is not uncommon.

The first lineage to diverge is likely to have looked nothing like modern species and may not have the defining characters required to be recognised. This simple explanation, cannot however, explain away the disagreement over primate divergence times. There are two suborders of primates, known as the haplorhini and the strepsirrhini suborders, which Arnason *et al.* [4] dated as diverging 80 My ago. The defining primate morphological characters, such as opposable thumbs, five fingers with fingernails, a generalised dental-pattern and forward facing binocular vision, are seen in both suborders. As it is much more likely that these characters evolved in a common ancestor, rather than through convergent evolution in the two suborders, we cannot dismiss the difference in the divergence time estimates by saying that character states were slow to evolve.

In this thesis I give an alternative explanation for the discrepancy between molecular and palaeontological estimates. Namely, that current methods of estimating divergence times from fossil evidence massively underestimate the size of the temporal gap that is expected between the divergence time and the oldest fossil. This is especially true for taxa with low sampling rates, such as the primates. Following the novel approach of Tavaré *et al.* [118], I use a model for species diversification from the literature, apply a model for fossil preservation, and then ask what size temporal gap is observed on average between divergence and the first fossil. Figure 1.1 gives an illustration of the basic idea. An important aspect of this approach is that the sampling rate is taken into account, unlike in the method of Gingerich *et al.* [51, 52], which gives the same estimate regardless of the completeness of the fossil record. My approach builds on that of Tavaré *et al.* using a different inference technique, and by removing several of the assumptions needed previously, such as the need to have the sampling fractions fixed into a constant ratio [113]. I also give a more realistic sampling model and investigate the effect of the mass extinction event that wiped out the dinosaurs.

Throughout the thesis I focus on extracting the maximum amount of information possible from the primate fossil record. By taking into account the modern primate diversity (number of extant species) and the structure of the primate phylogeny suggested by the anthropoid subtree, it is possible to get a more accurate picture than would otherwise be the case. Even within the palaeontology community, there is disagreement about the quality of the fossil record and about how much can be learnt from it. Advocates of the fossil record point to the strong correlation between the order in which taxa originate in the molecular and the fossil-based estimates as evidence of the success of the fossil record. Critics, however, highlight the correlation between the number of rock formations and the fossil record [92, 93] and question whether the appearance of mass extinctions and explosive radiations are really an artefact of the fluctuation in the rock record [109]. While there is both a biotic and an abiotic signal in the fossil record, most palaeontologists believe and hope that the biotic signal is strong enough for us to accurately estimate divergence dates.

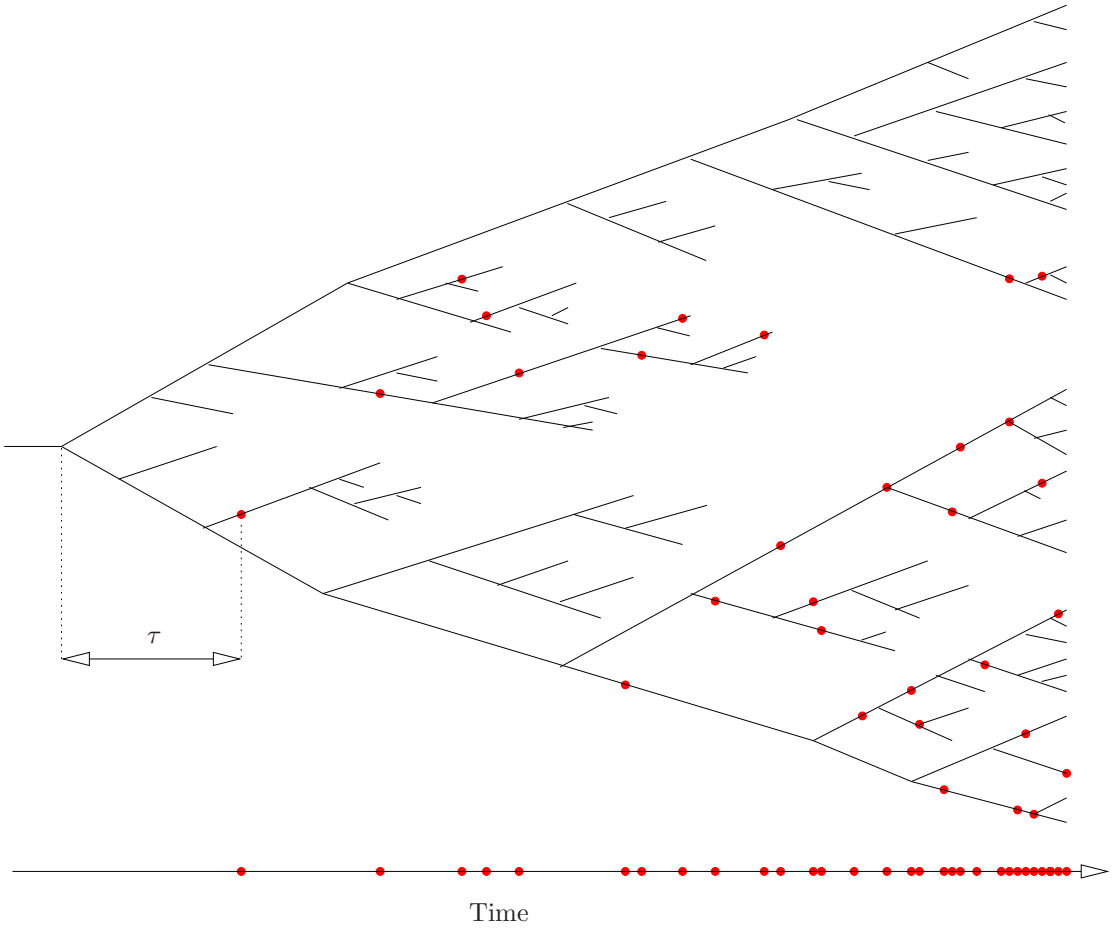

Figure 1.1: *A sample evolutionary tree showing relationships between species. Red dots indicate that the species has been preserved as a fossil. The horizontal line is the information we get from the fossil record and consists of the fossil finds from the tree projected onto the time axis. Parameter τ represents the temporal gap between the divergence time of the LCA and the oldest discovered fossil.*

At several points during this thesis we find that extensions to the model have introduced too many degrees of freedom, and that we are at the limit of what is possible with the data set. Knowing this limitation is important, as otherwise false conclusions can easily be drawn.

Giving a careful analysis of the primate fossil record is particularly important. Firstly, the primate fossil record is known to be particularly poor compared to many other orders [121]. It is far from complete and contains several major holes. For example, the Malagasy lemurs, a primate infraorder, represent more than 20% of modern primate genera and about 16% of modern species; there are at least 85 modern lemur species [86]. Yet to date no lemur fossils have been discovered. The presence of the loris sister group, documented

to be at least 37 My old, shows that the lemur *ghost lineage* must be at least 37 My in length [83], proving that long temporal gaps do occur. The completeness of the primate fossil record has been estimated to be as low as 7% [118]. This is partially due to the fact that primates are soft bodied, arboreal and tend to live in warm wet places that are not conducive to fossil preservation. Indeed, the fact that nearly all modern primates live in tropical or subtropical climates [77] may provide another explanation as to why their fossil record is so poor. A direct reading of the record suggests that primates originated in the Northern Hemisphere and spread to the southern continents[(4)]. However, the first appearance of fossil primates coincides with a worldwide temperature increase of 5-10°C during the Eocene (55 - 37 My ago) [85]. It is believed that the southern continental regions show lower fossil preservation rates compared to the northern continents [78] and we know that Europe and North America are much better sampled by palaeontologists than Africa and Asia. This suggests the alternative theory that primates originated in the southern continents and spread northwards during warmer climatic conditions in the Eocene [20].

Secondly, it is important to accurately date the primate divergence time because of the consequences of either an early or late divergence time on the plausibility of various evolutionary scenarios. For example, primate evolution may be linked to the great angiosperm (flowering plants) radiation that is thought to have occurred in the Mid-Cretaceous about 100 My ago [35, 71]. Conroy [31] suggests that the appearance of large fruit as a food source led to the evolution of the primates from other mammals species. This theory is less plausible, however, if the primates did not evolve until the Cenozoic. Another example of its importance is that the timing of the divergence affects the potential relevance of continental drift. If it is true that these radiations (evolutionary splits) took place only in the Cenozoic, then continental drift is only of limited relevance to primate evolution. Alternatively, however, if mammalian groups were present in the Cretaceous then continental drift may have had a large effect. There is also interest in the placement of the divergence times relative to the Cretaceous-Tertiary (K-T) boundary 65 My ago, which marked the extinction of the dinosaurs. The traditional view is that the Cretaceous era (144 to 65 My ago) is the *age of dinosaurs* whereas the following Cenozoic era is the *age of mammals* and that the rapid evolutionary diversification of the mammals was made possible by the ecological vacuum created by the extinction of the dinosaurs [2].

Finally, a great deal of time and money has been spent collating primate fossils and classifying them into the database shown below. It seems like an act of folly to date the divergence time solely from the age of the oldest fossil. The expense of the data justifies a considerable amount of time and effort being spent on its analysis, with care taken to extract all the information possible. This thesis forms part of that effort.

[(4)] Over 40% of all known fossil primates from the period 55 My ago to 30 My ago come from Wyoming or France [125].

Data

The primate fossil database I use throughout this thesis is described below in Table 1.1. This was compiled by our palaeontologist collaborators Christophe Soligo from University College London and Robert Martin from the Field Museum in Chicago. Each known fossil species has been categorised by the epoch in which it was preserved, by the sub and infraorder to which it belongs, and whether it is a crown or stem species. There is uncertainty in most aspects of these data.

Fossils are only preserved in sedimentary rock, which is formed when soil is compressed over a long period of time. As sedimentary rock cannot easily be dated, palaeontologists rely on nearby igneous rock and the principle of geological superposition (younger rock generally lies above older rock) to date fossils. Consequently, there is often some uncertainty about the age of any given fossil. There is also uncertainty over the number of extant species in each family with several large changes over the previous few years. The known primate diversity has increased from 235 in 2001 to 376 in 2005 [56]. This is due to both the discovery of new species and the division of taxa previously thought to represent a single species into a number of distinct species.

All primates belong to either the strepsirrhine (wet nosed primates) or the haplorhine (dry nosed primates) suborder. The haplorhini are further divided into the tarsiidae (tarsiers) and anthropoidea (monkeys and apes) infraorders. The anthropoids are then divided into platyrrhine (new world monkeys) and catarrhine (old world monkeys) species. Again, there is often uncertainty about the classification of fossils, with some species known from samples containing only a few teeth. Figure 1.2, by Christophe Soligo, shows skulls from each primate family and suggests a possible phylogeny. The oldest known fossil primate is at most 54.8 My old, and the range extension from the oldest fossil to the primate divergence time is based on inferences made in Tavaré *et al.* [118]. The evolutionary relationships shown in the phylogeny are far from certain, with some authors suggesting that the adapiforms and omomyiforms are not necessarily stem strepsirrhine and haplorhine groups respectively. See Chapter 6 for more information.

The concept of *crown* and *stem* groups was suggested by Willi Hennig as a way to classify living organisms relative to extinct ones. The concepts are not necessarily universally accepted [39], as they do not obey the systematics of living organisms. A *crown group* of species is a living monophyletic clade. It contains all living species, their last common ancestor (LCA), and all of its descendants. *Stem groups* lie basally to the LCA of the crown group (i.e., further down the tree of life) and do not contain any living species. The species in the stem group always lack one or more of the diagnostic characters present in the crown group.

As an example, look at the crown and stem strepsirrhine fossil counts in Table 1.1 and the strepsirrhine side of the tree in Figure 1.2. The crown strepsirrhini consist of the lemurs,

Fossil data as supplied June 2006											
Epoch	k	Time at the base of interval k, T_k (My)	Strepsirrhini		Haplorhini						
			Stem	Crown	Stem	Crown					
						Tarsiidae	Anthropoidea				
						Stem & Crown	Stem	Crown			
								Platyrrhini		Catarrhini	
								Stem	Crown	Stem	Crown
Extant	0	0	-	88	-	7	-	-	128	-	153
Late-Pleistocene	1	0.15	-	-	-	-	-	-	2	-	20
Middle-Pleistocene	2	0.9	-	-	-	-	-		-	-	28
Early-Pleistocene	3	1.8	-	-	-	-	-	-	-	-	30
Late-Pliocene	4	3.6	-	3	-	-	-	-	-	-	40
Early-Pliocene	5	5.3	-	1	-	-	-	-	-	-	11
Late-Miocene	6	11.2	3	1	-	-	-	-	-	5	29
Middle-Miocene	7	16.4	1	1	-	1	-	-	17	10	16
Early-Miocene	8	23.8	-	6	-	-	-	-	3	22	3
Late-Oligocene	9	28.5	-	-	1	-	-	1	-	1	-
Early-Oligocene	10	33.7	3	1	1	1	10	-	-	6	-
Late-Eocene	11	37.0	14	-	7	-	7	-	-	2	-
Middle-Eocene	12	49.0	47	2	59	2	9	-	-	-	-
Early-Eocene	13	54.8	26	-	37	-	2	-	-	-	-
Pre-Eocene	14		-	-	-	-	-	-	-	-	-

Table 1.1: *Data showing the number of primate species discovered in each geological epoch, broken down into sub and infraorder, and into crown and stem species. This database was kindly provided by Christophe Soligo and Robert Martin. The number of extant species comes from Groves [56].*

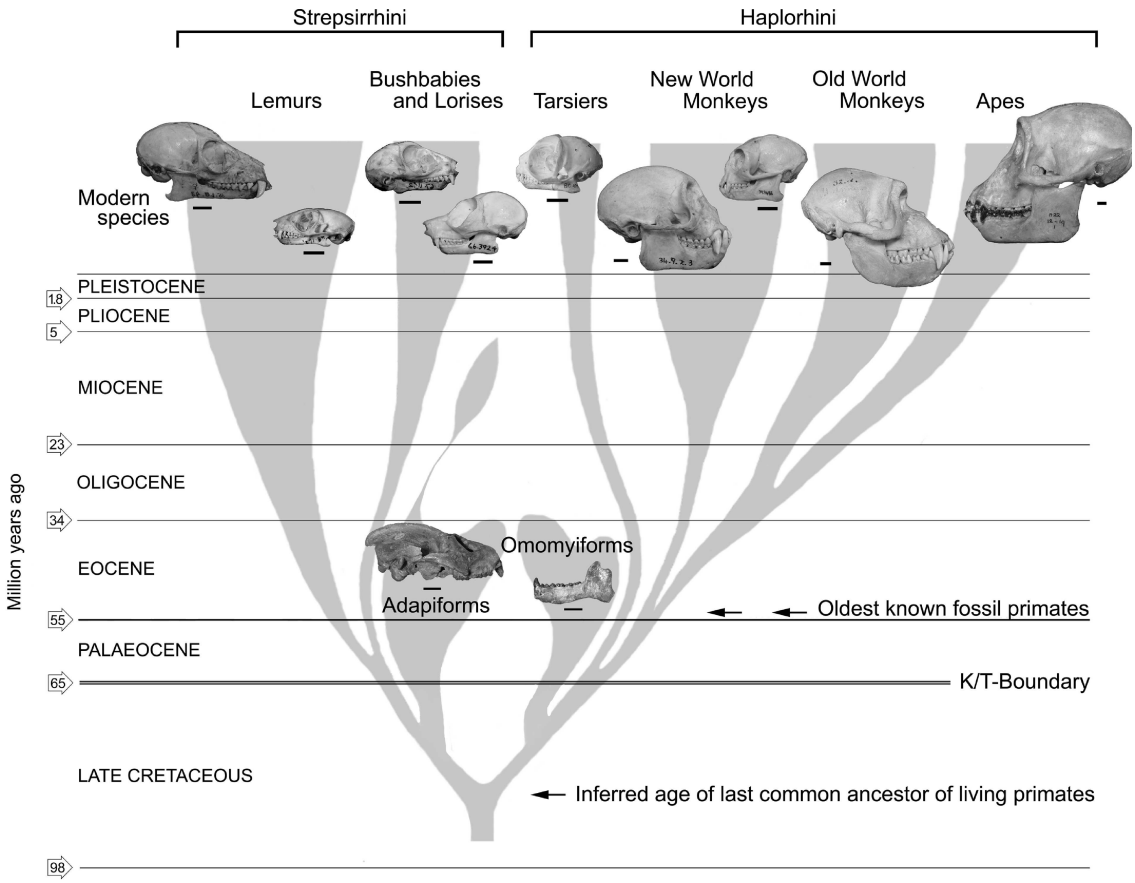

Figure 1.2: *A possible primate phylogeny by Christophe Soligo. The placing of the primate divergence time in the Late Cretaceous is based on inferences made in Tavaré et al. [118]. Notice that the most complete fossil omomyiform consists of only a lower jaw.*

lorises and bushbabies and all their ancestors back to the LCA in the Mid-Palaeocene. The stem strepsirrhine in Figure 1.2 are the adapiformes. This helps to make clear the difference between the crown and stem divergence time. According to this diagram, the strepsirrhine crown divergence time was about 60 My ago, whereas the strepsirrhine stem divergence time was approximately 75 My ago.

The notion of crown and stem groups allows us to be more specific about which divergence time we shall be dating in this thesis. Throughout, the main focus is on the primate crown divergence time. The primate crown group consists of both the stem and crown strepsirrhine and haplorhine species. The primate stem group would include the plesiadapiformes and is not studied here. I shall also look at the crown anthropoid divergence time. The crown anthropoids are the new and old world monkeys and the apes.

Finally, before discussing the statistical approach used, I want to reiterate the importance of a careful analysis. The data in Table 1.1 was expensive to collect and analyse, both in terms of time and money, and it deserves great care and attention to make sure

that a correct and coherent picture of primate evolution is being drawn from it.

1.2 Statistical Inference

The job of a modeller is usually to provide a mathematical description of reality, which says that given a set of conditions, this is the type of data or behaviour we expect to observe. In other words, they provide a forward model. While there are several key model developments in this thesis, model development is not the main aim. Instead, the focus is largely on how to do inference for models previously used in the literature. Where the modeller gives forwards models, the statistician looks at the inverse problem. Namely, given a data set and a model, what can we learn about the state of reality or set of conditions necessary in the model. While many forward models of evolution exist, only a few people have examined the inverse problem of what the fossil record can teach us given a choice of model. This is the main aim of my thesis.

There are two schools of thought on the correct approach to inference, usually known as *Bayesian* and *frequentist* statistics. I subscribe to the Bayesian viewpoint set out by de Finetti [37, 38] in which probability is a measure of uncertainty and where a priori knowledge is taken into account in all inferences. If we consider statistics to be the study of uncertainty [72] (an apt description of our task), then I believe the Bayesian viewpoint is best equipped to deal with this. The frequentist approach, on the other hand, assumes we can embed the process into a sequence of repeated processes, and is a poor setting for the unique evolutionary process studied here.

The Bayesian approach to inference can be described as follows. Suppose that the model is parameterised by θ and that there is uncertainty surrounding its value. This uncertainty is expressed by a prior probability distribution $\pi(\theta)$. Suppose further that we are in possession of data $\mathcal{D}$ which it is assumed come from our model. Our aim is to find the posterior distribution of the parameters given the data:

$$\pi(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\pi(\theta)}{\mathbb{P}(\mathcal{D})}, \tag{1.1}$$

where $\mathbb{P}(\mathcal{D}|\theta)$ represents the probability of the data given that the parameter takes value θ. The normalising constant is chosen so that the posterior distribution, $\pi(\theta|\mathcal{D})$, integrates to one, i.e., $\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)d\theta$. This notation is used throughout the thesis.

The past few decades have seen a growth in the use of Bayesian ideas in many areas of science [62, 96]. This is largely due to the increased availability and lower cost of computer power. This increased power has allowed for the development of a large variety of computationally intensive *Monte Carlo* techniques which are able to numerically integrate the terms in Equation (1.1). Traditional frequentist maximum-likelihood calculations often limited the number of parameters one could use in any given model [10], whereas

techniques such as *Markov Chain Monte Carlo* (MCMC) [119] allow models of greater complexity to be analysed.

Software packages which simplify the use of MCMC methods, such as the BUGS project [115], have proven very popular in the scientific community. For while MCMC is increasingly being used it is not necessarily easy to master; there are numerous time consuming obstacles to overcome such as making sure the chain mixes well and that it converges to equilibrium. Another problem is that MCMC methods require knowledge of the likelihood function, $\mathbb{P}(\mathcal{D}|\theta)$. For complex models, which are increasingly being studied, this quantity is often not known. This is the case for the model considered in this thesis.

Before introducing an alternative to MCMC, I shall discuss why the problem of inferring divergence times is difficult. The main reason inference is difficult is due to the sheer amount of missing information. We are essentially interested in inferring the depth of an evolutionary tree such as that shown in Figure 1.1. However, we only see a one dimensional snapshot of this tree via the fossil evidence that has been deposited in the rock record. Our task is to infer the depth of a complex tree structure using only the one dimensional information contained in the frequency pattern of fossil discoveries. In fact the situation is harder still as the age of a fossil cannot be clearly resolved. The best that can be done is the placement of each fossil in the appropriate epoch.

The second reason inference is difficult is due to the role that conditioning plays. We know that modern primates exist, and so any calculation must be conditional on this fact. At various points throughout the thesis conditioning will play a part, building to the final two chapters where the evolution of the anthropoid suborder is conditioned upon. Notice that the paucity of the primate fossil data does not in itself add any difficulty. It does, however, make the inferences less accurate and increases the need for a rigourous approach.

Hidden tree structures, such as the hidden primate phylogeny, often occur in models used in fields such as population genetics and ecology. They often lead to highly dependent data structures which are difficult to deal with using standard statistical methods. This led to the development of Monte Carlo techniques that can deal with problems where the likelihood function is unknown. This collection of *likelihood-free* techniques are usually known as *Approximate Bayesian Computation* (ABC) methods, and they only require that we can simulate data sets from the model [11, 98, 117]. The secondary aim of this thesis is to add to the catalogue of ABC methods, to investigate their use, and to shed light on their advantages and disadvantages.

While ABC methods are only approximate, as their name suggests, they do have potential advantages over MCMC techniques. Firstly, they are almost trivial to use once you are able to simulate from the model as they do not require the fine tuning that is often needed to ensure convergence in MCMC. Secondly, changes to the model or data structure

can easily be incorporated without any changes to the inference mechanism, whereas in MCMC adding a different type of data may render inference impossible. Thirdly, a natural measure of model fit comes out of the ABC approach giving an approximate model selection technique. Finally, if the likelihood function is unknown, then MCMC methods can not be used. This final point is the main reason to use and develop ABC methods.

1.3 Structure of the Thesis

Chapter 2 contains details of the model that will be used throughout the thesis. I carefully highlight the assumptions that have been made and indicate the more problematic ones. I give calculations of some basic probabilities and show what effect conditioning on non-extinction has on the evolutionary process.

In Chapter 3 I introduce the concept of Approximate Bayesian Computation and give a few examples, before moving on to apply the inference algorithms to the model from Chapter 2. Chapter 4 looks at ways to exploit all of the information contained in the data set, such as the number of extant species. I also give an optimal subtree selection algorithm which allows us to find the joint distribution of two divergence times, and use this to find the distribution of the primate and anthropoid divergences. The final two parts of Chapter 4 discuss an improvement to the basic ABC algorithm, given by Beaumont *et al.* [11], and assess its accuracy with a simulation study.

In Chapter 5 I develop the ABC methodology to produce an ABC-MCMC hybrid algorithm. This allows exact inference to be done on aspects of the model that we can compute and approximate inference on the parts that we cannot. This new algorithm speeds up the inference approach and allows us to improve the model by removing one of the more controversial assumptions made by Tavaré *et al.* [118].

Chapter 6 contains several improvements to the basic model, including a more realistic fossil find model in which the duration a species lives for is taken into account, and a model of the K-T crash that killed the dinosaurs 65 My ago. I also outline a potential model selection approach that is made possible by our ABC algorithm and use it to justify using a logistic growth curve to model primate diversity.

Chapter 7 is an entirely theoretical chapter in which I give the structure of branching processes conditioned to have a birth at a given point in time. These trees have a simple and pleasing fish-bone structure. This then leads to a discussion of size-biased trees and a long proof based on the Palm distribution of random measures. This theory is then used in Chapter 8 to give a fully Bayesian approach to inferring the joint distribution of two divergence times. By simulating speciation trees conditioned to have a birth at a given time, I can choose both the primate and anthropoid divergence times from prior distributions and thus find the joint posterior distribution of both divergence times.

Finally, Chapter 9 contains a short conclusion of my findings followed by a bibliography.

Chapter 2

Modelling Approach

> All models are wrong but some are useful. G. E. P. Box

The aim of this thesis is to develop a method for the inference of species divergence times. In this chapter, I take a modelling approach from the literature, which I then build upon throughout the remainder of the chapter and at various other points in the thesis. Occam's razor suggests that a model should give an economical account of the phenomena under consideration that is both simple and evocative; over-elaboration and over-parameterisation complicates inference. For example, although it is possible to conceive of models in which we keep track of fluctuations in the population size of each species, the interactions between them, and many other factors scientists claim to know something about, it would not be possible to do so without the introduction of assumptions open to criticism by others. For this reason, I aim to keep the model as flexible as possible, and will try to let the data provide the information. A proviso to this is that for sparse data sets, strong modelling assumptions are often required. I will explore what is and is not possible with the available data throughout the thesis.

Following Tavaré *et al.* [118], I break the modelling problem into two parts. In Section 2.1 I give a model for primate evolution, and this is followed in Section 2.2 with a model for fossil finds. Section 2.3 gives the inference approach one might take in an ideal world

and the reason why this is not possible. Along the way, I give a few calculations that will be required later in the thesis.

2.1 Speciation Model

In this section I give a model that can be used to generate speciation trees. These trees, often called phylogenetic trees, show the evolutionary relationships between different species. We consider a branching structure in which species diverge into distinct and different species at random times (see Figure 2.1 for an example). Each branch (or lineage) represents a single species and its length represents the amount of time for which that species existed.

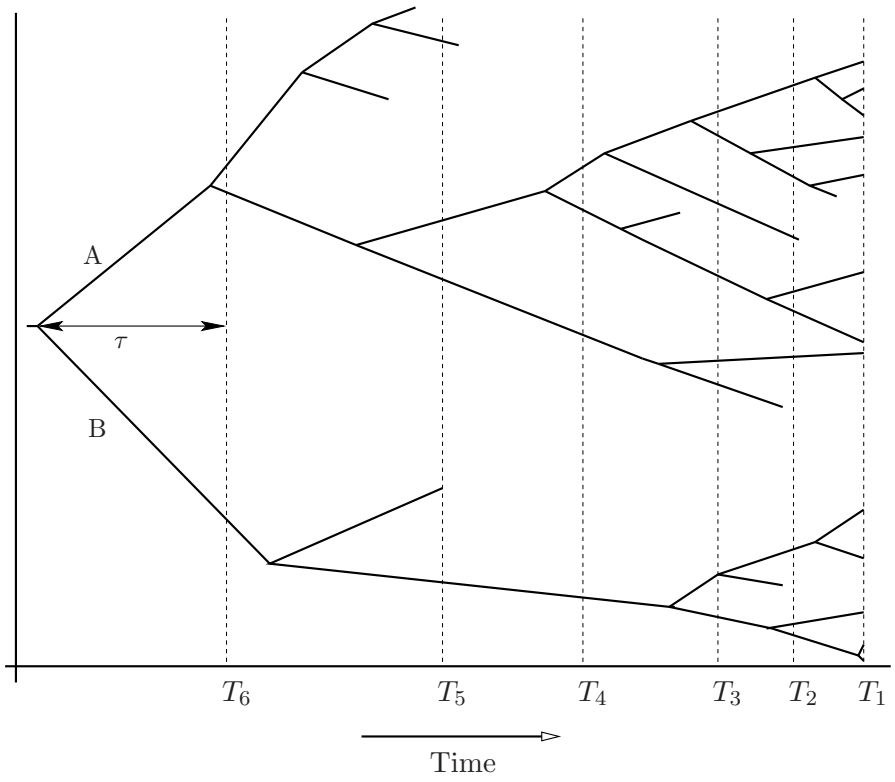

Figure 2.1: *An example evolutionary tree. Side A might represent the haplorhine species and side B the strepsirrhine species. Parameter τ represents the temporal gap between the oldest fossil and the primate divergence time. Time is divided into geological epochs, shown here by the vertical lines.*

The study of branching processes has a long history. In 1873 Francis Galton posed a question in the *Educational Times* about the extinction of aristocratic family names. This led to Galton and Watson's seminal paper [123] in which they described the branching process which now takes their name[(1)]. The discrete time Galton-Watson process assumes

[(1)]I. J. Bienaymé had previously introduced this concept [59] and so some authors refer to the process

that in each generation, each individual gives birth to an independent and identically distributed number of offspring. In the next generation these then go onto give birth to an independent, identically distributed number of offspring, and so on. And so in essence, the study of Galton-Watson processes is the study of sums of independent, identically distributed random variables.

Since 1875, the theory of branching processes has been generalised and developed, and there is now a large and diverse literature on the general theory. Athreya and Vidyashankar [9] give a survey of some of the different areas in the field, and Athreya and Ney [8] give the introductory theory. A more abstract treatment can be found in Asmussen and Hering [6]. Here we follow in the tradition of Raup, Gould, Foote and others (see [102] and [87]), by using branching processes to model speciation.

More specifically, I use a continuous-time, non-homogeneous, binary Galton-Watson process, also known as a generalised birth-death process. We assume that each species lives for an exponential period, with mean $1/\lambda$ million years (My), independently of other species. If a species dies at time t it branches into either zero or two new species, with branching probabilities $p_0(t)$ and $p_2(t)$ respectively ($p_0(t) + p_2(t) = 1$ for all t). A method for determining values of $p_0(t)$ and $p_2(t)$ is given in Section 2.1.1.

We want to estimate the primate divergence time, which is how far back in time we must go down the evolutionary tree until we reach the last common ancestor of the primates. To model this we assume that this ancestor diverged into two new species at time 0, so that $Z(0) = 2$, where $Z(t)$ is the diversity (number of species) at time t[(2)]. $Z(t)$ will be referred to as the *size-process* of the tree. The branching property allows us to consider $Z(t) = X_1(t) + X_2(t)$, where X_1 and X_2 are independent copies of the branching process started from one individual at time 0.

An alternative model I could have used is splitting (or budding) trees [45] where each species produces offspring at the jump times of a Poisson process. Geiger and Kersting [47] show that under a parameter transformation splitting trees are equivalent to Galton-Watson processes. However, when fossil finds are considered the two models are no longer equivalent and may lead to different results. In this thesis I consider only Galton-Watson processes.

Before giving some basic results about the Galton-Watson process, I want to highlight some of the assumptions that have been made in the above approach. Firstly, we are assuming that each species lives for an exponentially distributed period of time. This is necessary so that we can use the memoryless property of the exponential distribution,

---

as the Bienaymé-Galton-Watson process. I will follow the majority and refer to the process as the Galton-Watson process.

[(2)] Care needs to be taken so as not to be confused by time scales. When referring to the real world, the present is time 0 and we measure time back into the past in units of millions of years (My). So for example, the Cretaceous era ended 65 My ago, i.e., at time -65 My. However, in our model I use a different time scale in order to simplify the algebra, where the primate divergence is at time 0.

which ensures that the size process, $Z(t)$, is Markovian. Without this assumption, the past and future of $Z(t)$ would not be independent given the present, which would make the mathematics more difficult. A potential drawback of the memoryless property is that species do not age (see Jagers [60]). However, I consider this a reasonable assumption to make as there are no clear reasons to expect that because a species has already existed for 10 million years it shouldn't continue to exist for another 10 million. A theory of age-dependent branching processes has been developed, see [12], but will not be used here.

Secondly, we assumed that when a species dies, it gives birth to either zero or two new species, i.e., a binary tree structure. As speciation events are rare phenomena, we can assume that the probability of a species diverging into three or more new species 'simultaneously' is zero. We assume $p_1(t) = 0$ as these events do not change the shape of the tree, which can be seen by considering a parameter transformation. See Athreya and Ney [8] (p118) for details.

The important thing to remember here is that all models are wrong. As scientists, we must be aware of what is importantly wrong. Or as Box [21] says, *'It is inappropriate to be concerned about mice when there are tigers abroad'.*

While I feel that the two assumptions stated above are reasonable, there are potentially more serious drawbacks to the branching process approach. The first, highlighted by Jagers [60], is that in the limit, $Z(t) \to 0$ or ∞ with probability one. There does not exist a stable branching process that settles down to some positive finite value (for a proof see page 18 in Asmussen *et al.* [6]). The second problem is that species do not interact. In any given ecosystem, there is an upper bound on the number of related species (those that compete for similar resources) that can maintain a population. This is not taken into account in our model. However, we hope that the key elements of the model are sound and that something useful can be learnt.

2.1.1 Basic Calculations

I now outline the basic calculations needed to find the transition probabilities of the size process $Z(t)$. These were first given by Kendall [67] in a simpler setting, and hence I only give a brief outline here. In general, explicit calculation of the transition functions is not possible for the Galton-Watson process. It is only possible here because we are using the simpler birth-death process. The solution is found via the moment generating function of the process. Let

$$P_{ik}(s,t) = \mathbb{P}(Z(t) = k | Z(s) = i)$$
$$F_j(u,s,t) = \mathbb{E}(u^{Z(t)} | Z(s) = j).$$

The values $P_{ik}(s,t)$ form a stochastic semigroup due to the assumption of exponential lifetime distributions. This ensures that the size-process $Z(t)$ is a Markov process. We denote the generator of the Markov process by the matrix $\boldsymbol{A}(t)$, where

$$A_{ij}(t) = \lim_{\delta \downarrow 0} \frac{1}{\delta}(P_{ij}(t, t+\delta) - I_{ij}) = \begin{cases} i\lambda p_2(t) & \text{if } j = i+1 \\ i\lambda p_0(t) & \text{if } j = i-1 \\ -i\lambda & \text{if } j = i \\ 0 & \text{otherwise,} \end{cases}$$

and where $\boldsymbol{I}$ represents the identity matrix. We can then use Kolmogorov's forward equations (see for example [104])

$$\frac{\partial \boldsymbol{P}}{\partial t}(s,t) = \boldsymbol{P}(s,t)\boldsymbol{A}(t)$$

to write down a system of differential-difference equations for the transition probabilities:

$$\frac{\partial}{\partial t}P_{10}(s,t) = \lambda p_0(t) P_{1,1}(s,t) \tag{2.1}$$

$$\frac{\partial}{\partial t}P_{1n}(s,t) = (n+1)\lambda p_0(t) P_{1,n+1}(s,t) + (n-1)\lambda p_2(t) P_{1,n-1}(s,t) - n\lambda P_{1,n}(s,t),\ n \geq 1. \tag{2.2}$$

Multiplying by u^n and summing gives a partial differential equation for the probability generating function:

$$\frac{\partial F_1}{\partial t}(u,s,t) = \lambda(u-1)(up_2(t) - p_0(t))\frac{\partial F_1}{\partial u}(u,s,t). \tag{2.3}$$

Kendall gives the solution for the moment generating function as

$$F_1(u,s,t) = \frac{\xi + (1-\xi-\eta)u}{1-\eta u} \tag{2.4}$$

and upon expansion we find that

$$P_{1n}(s,t) = \begin{cases} \xi & \text{if } n = 0 \\ (1-\xi)(1-\eta)\eta^{n-1} & \text{if } n > 0, \end{cases} \tag{2.5}$$

where $\xi \equiv \xi(s,t)$ and $\eta \equiv \eta(s,t)$ are functions of s and t. The values of ξ and η are found by substituting back into (2.1) and are equal to

$$\xi(s,t) = 1 - \frac{e^{-\omega(s,t)}}{W(s,t)}, \quad \eta(s,t) = 1 - \frac{1}{W(s,t)}. \tag{2.6}$$

Functions ω and W are defined by

$$\omega(s,t) = \lambda \int_s^t (p_0(\tau) - p_2(\tau))\, d\tau, \tag{2.7}$$

$$W(s,t) = e^{-\omega(s,t)} \left(1 + \lambda \int_s^t e^{\omega(s,u)} p_0(u) \mathrm{d}u\right). \tag{2.8}$$

It then follows from Equation (2.5) that the expected diversity at time t, when starting with one species at time s, is

$$\mathbb{E}(Z(t) \mid Z(s) = 1) = \frac{1-\xi}{1-\eta} = \exp\left(\lambda \int_s^t (m(u) - 1) \mathrm{d}u\right), \tag{2.9}$$

where $m(u)$ is the mean number of offspring produced when a species dies at time u.

Determining the Branching Probabilities

Equation (2.9) shows that the expected growth of the branching process is determined by the mean of the lifetime distribution, $1/\lambda$, and the mean of the offspring distribution, $m(u)$. As we are using a binary process, we can use knowledge of $m(u)$ to determine the branching probabilities, as $m(u) = 2p_2(u)$. To determine $m(u)$, we assume some parametric form for the expected diversity, $\mathbb{E}Z(t)$. In Section 6.4 I try several different forms for $\mathbb{E}Z(t)$, including linear and exponential forms, and then develop a model selection technique to choose between them. For the majority of this thesis, however, I use a logistic growth form, which was advocated by Raup *et al.* [102] as the most biologically realistic model. The logistic model is also in agreement with the predictions made by competition models used in ecology, such as the Lotka-Voltera model [16, 75]. For constants $\rho \geq 0$ and $0 \leq \gamma \leq 1$, I assume that

$$\mathbb{E}Z(t) = \frac{2}{\gamma + (1-\gamma)e^{-\rho t}}. \tag{2.10}$$

Equating Equations (2.9) and (2.10), substituting $m(u) = 2p_2(u)$, and differentiating, we find that

$$p_2(t) = \frac{\frac{\rho}{2\lambda}(1-\gamma)}{(1-\gamma) + \gamma e^{\rho t}} + \frac{1}{2} \tag{2.11}$$

$$p_0(t) = 1 - p_2(t). \tag{2.12}$$

Notice that as t tends to infinity, $p_0(t)$ and $p_2(t)$ both tend to $\frac{1}{2}$, and so in the limit as $t \to \infty$ the process behaves like a critical binary Galton-Watson process. Using Equations

(2.7) and (2.8), we can observe that

$$\omega(s,t) = \log\left(\frac{\gamma+(1-\gamma)e^{-\rho t}}{\gamma+(1-\gamma)e^{-\rho s}}\right)$$
$$W(s,t) = \frac{1}{\gamma+(1-\gamma)e^{-\rho t}}\left[(\gamma+(1-\gamma)e^{-\rho s})+\frac{\lambda}{2}\left(\gamma(t-s)+(\frac{1}{\lambda}-\frac{1}{\rho})(1-\gamma)(e^{-\rho t}-e^{-\rho s})\right)\right].$$

Substituting these values into Equation (2.6) we find that

$$\eta(s,t) = 1 - \frac{\gamma+(1-\gamma)e^{-\rho t}}{(\gamma+(1-\gamma)e^{-\rho s})+\frac{\lambda}{2}\left(\gamma(t-s)+(\frac{1}{\lambda}-\frac{1}{\rho})(1-\gamma)(e^{-\rho t}-e^{-\rho s})\right)} \tag{2.13}$$

$$\xi(s,t) = 1 - \frac{\gamma+(1-\gamma)e^{-\rho s}}{(\gamma+(1-\gamma)e^{-\rho s})+\frac{\lambda}{2}\left(\gamma(t-s)+(\frac{1}{\lambda}-\frac{1}{\rho})(1-\gamma)(e^{-\rho t}-e^{-\rho s})\right)}. \tag{2.14}$$

Notice that when $s \leq t, \gamma \leq 1$ and $\rho \leq \lambda$, the values of η and ξ are bounded between 0 and 1. This ensures that Equation (2.5) is a proper probability distribution, and is a type of modified geometric distribution.

2.2 Fossil Finds Model

In the previous section I gave a speciation model which allows us to construct sample phylogenetic trees. I now give a model for fossil preservation, enabling us to generate sample data sets of the form given in Chapter 1.

There are many factors that determine whether any given species is found in the fossil record. It must firstly be preserved as a fossil, and secondly, it must be discovered. Many factors affect preservation: hard bodied species are more likely to be preserved than soft; large individuals more likely than small; common species more likely than rare; and marine species more likely than terrestrial.

Since I am dealing with primate species only, the factors above, with the exception of the abundance of each species, are roughly the same for all species. No information is available on species abundance, and so I make the assumption that all species are equally likely to be preserved as a fossil in a given time period.

Many authors have linked fossil discovery rates to the rock record [92, 93]. The age distribution of the exposed rock on the Earth's surface is not uniform, with some epochs having a large number of formations in a variety of locations, whereas others have very few. It is also known that primates are most prevalent in the tropics, and so if the location of the rock formations from a given epoch are not in these regions, the probability of finding primate fossils from this epoch is low. For these reasons, I assume the probability of preservation can change over time, as done by Foote [43].

Initially, I focus on a simple model in which each species is equally likely to be preserved

in any interval in which it lived. Splitting time into the geological epochs specified in Chapter 1, let D_i represent the number of fossil species discovered in interval i, N_i the total number of species that lived during interval i, and α_i the *sampling fraction* for interval i (the proportion of species preserved as fossils). This leads to a simple binomial structure:

$$\mathbb{P}(D_j = d_j | N_j = n_j, \alpha_j) = \binom{n_j}{d_j} \alpha_j^{d_j} (1 - \alpha_j)^{n_j - d_j}. \tag{2.15}$$

I shall refer to this model as the *binomial fossil finds* model.

The elephant in the room is the species lifetime L_i. It can be argued that the fossil record is biased in favour of long-lived species; intuitively, a species that existed for 10 My is more likely to be preserved than one that lived for only 1 My. In Chapter 6 I develop a Poisson model for fossil finds which takes the species lifetime into account. Until then, however, the binomial model will be used.

The most important parameter throughout this thesis is the temporal gap τ. This is the time between the oldest primate fossil and the primate divergence time. The age of any given fossil cannot be determined with any great precision. The best that can be done is assigning each fossil to the epoch in which it was formed. The age of the oldest primate fossil is 54.8 My, which corresponds to the start of the Early-Eocene epoch. Figure 2.1 illustrates the temporal gap τ, and a sample evolutionary tree with epochs overlaid.

2.3 Inference in an Ideal World

In the previous sections I specified a forward model for the phenomena we wish to investigate. By *forward model*, I mean that given values for the temporal gap τ, the speciation model parameters λ, ρ and γ, and the sampling fraction $\boldsymbol{\alpha}$, we can simulate sample data sets. It is the inverse of this procedure, however, that is of interest. We have a set of fossil counts $\mathcal{D}$, and we wish to make inferences about the underlying process by specifying likely values for the parameters.

Previous research has been done on inference in branching processes, and Guttorp [57] is a useful starting point for exploring previously published work. However, most of the effort has been on questions of how to infer birth and death rates, and offspring distributions. There is no literature dealing with inference in the partially observed tree processes specified here.

Inference is usually done, whether in a frequentist or a Bayesian analysis, via the likelihood function, $\mathbb{P}(\mathcal{D}|\theta)$, where $\theta = (\boldsymbol{\lambda} = (\lambda, \gamma, \rho, \tau), \boldsymbol{\alpha})$ is the multidimensional model parameter. To calculate the likelihood, we can write

$$\mathbb{P}(\mathcal{D}|\theta) = \sum_{\mathcal{N}} \mathbb{P}(\mathcal{D}|\mathcal{N}, \boldsymbol{\alpha}) \mathbb{P}(\mathcal{N}|\boldsymbol{\lambda}), \tag{2.16}$$

where $\mathcal{N} = (N_1, \ldots, N_{14})$ are the counts of the total number of primate species in each interval. Equation (2.15) leads to the first term in the sum, and the second term is dealt with in Section 2.3.2.

Equation (2.5) gives the transition probabilities for the diversity jumping from $Z(s) = 1$ to $Z(t) = n$, and it will be useful later to have the transition probabilities for jumps starting from general diversity values. Let $\{X_i(t)\}_{i\geq 1, t\geq s}$ be independent, identically distributed copies of the branching process begun at time s, with $X_i(s) = 1$ for all i. The branching property states that if $Z(s) = k$, then

$$Z(t) = \sum_{i=1}^{k} X_i(t).$$

To calculate the transition probability from k to n, we must calculate the k-fold convolution of k independent branching processes. To simplify the notation, I write $Z(t) = X_1 + \cdots + X_k$, where $X_i \sim X(t)$ independently of other X_j, and I write ξ for $\xi(s,t)$ and η for $\eta(s,t)$. Then, note that

$$\begin{aligned}
\mathbb{P}(Z(t) = n \mid Z(s) = k) &= \mathbb{P}(X_1 + \cdots + X_k = n) \\
&= \sum_{r=1}^{k\wedge n} \mathbb{P}(X_1 + \cdots + X_k = n, R = r) \text{ for } n \geq 1, \\
&\qquad\qquad \text{where } R = \sum_{i=1}^{k} \mathbb{I}_{X_i \neq 0} \\
&= \sum_{r=1}^{k\wedge n} \sum_{\substack{j_1, \ldots, j_k : \\ \sum j_k = n,\, R = r}} \mathbb{P}(X_1 = j_1, \ldots, X_k = j_k, R = r) \\
&= \sum_{r=1}^{k\wedge n} \sum_{\substack{j_1, \ldots, j_k : \\ \sum j_k = n,\, R = r}} \xi^{k-r}(1-\xi)^r(1-\eta)^r\eta^{n-r}
\end{aligned}$$

$$\mathbb{P}(Z(t) = n \mid Z(s) = k) = \begin{cases} \sum_{r=1}^{k\wedge n} \xi^{k-r}(1-\xi)^r(1-\eta)^r\eta^{n-r} \begin{pmatrix} k \\ r \end{pmatrix} \begin{pmatrix} n-1 \\ r-1 \end{pmatrix}, & n > 0 \\ \xi^k, & n = 0. \end{cases} \tag{2.17}$$

The two combinatorial terms above are the number of ways of putting n balls into k boxes in such a way that there are exactly r non-empty boxes. We also wish to be able to calculate the joint distribution of the size process at times $s \leq t_1 \leq \ldots \leq t_k$, which we

can find by observing that the Markov property implies that

$$\mathbb{P}(Z(t_1) = n_1, \ldots, Z(t_k) = n_k) = \mathbb{P}(Z(t_1) = n_1) \prod_{i=1}^{k-1} \mathbb{P}(Z(t_{i+1}) = n_{i+1} \mid Z(t_i) = n_i).$$

2.3.1 Conditioning on Non-Extinction

A mistake commonly made by previous authors [87] is failing to condition calculations on the non-extinction of the process. We observe that there are extant primates, and therefore all analyses should be conditional on this fact. In this section I give calculations for the transition probabilities of the conditioned process. Firstly, consider a branching process started from one species at time s:

$$\begin{aligned}\mathbb{P}(X(t) = j \mid X(t) > 0) &= \frac{\mathbb{P}(X(t) = j, X(t) > 0)}{\mathbb{P}(X(t) > 0)} \\ &= (1-\eta)\eta^{n-1},\end{aligned}$$

where $\eta \equiv \eta(s,t)$. Thus, $X \mid X > 0$ is a standard geometric random variable with parameter $1-\eta$, and thus $\mathbb{E}(X(t) \mid X(t) > 0) = 1/(1-\eta)$.

The process we are interested in starts with two lineages and survives to the present. We also require that both sides of the tree survive, i.e., both lineages present at time $t = 0$ have extant descendants. Otherwise the first lineage does not represent the LCA of the primates, but an ancestor on the primate stem. Our modelling assumption is that one side of the tree represents the haplorhini, the other the strepsirrhini, both of which have extant representatives. Mathematically, the process we are interested in is $\widetilde{Z}(t) = \big(X_1(t) + X_2(t)\big)\, \mathbb{I}_{X_1(t), X_2(t) \geq 1}$. It takes the value 0 if either side of the tree dies out, and equals $Z(t)$ if both sides survive.

$$\begin{aligned}\mathbb{P}(\widetilde{Z}(t) = j) &= \sum_{k=1}^{j-1} \mathbb{P}(X_1(t) = k)\mathbb{P}(X_2(t) = j-k) && \text{for } j \geq 1 \\ &= (1-\xi)^2(1-\eta)^2\eta^{j-2}(j-1) && \text{for } j \geq 1 \\ \mathbb{P}(\widetilde{Z}(t) = 0) &= \mathbb{P}(\{X_1(t) = 0\} \cup \{X_2(t) = 0\}) \\ &= 2\xi - \xi^2\end{aligned}$$

Thus, conditioning this process on non-extinction, we find that

$$\mathbb{P}(\widetilde{Z}(t) = j | \widetilde{Z}(t) > 0) = (j-1)(1-\eta)^2\eta^{j-2} \qquad \text{for } j > 0 \tag{2.18}$$

$$\mathbb{E}(\widetilde{Z}(t) | \widetilde{Z}(t) > 0) = \frac{2}{1-\eta}. \tag{2.19}$$

We are now able to calculate the transition probabilities, if desired.

$$\begin{aligned}\mathbb{P}(\widetilde{Z}(t) = n \mid \widetilde{Z}(s) = k) &= \sum_{r=1}^{n-1} \mathbb{P}(X_1(t) = r,\ X_2(t) = n - r \mid \widetilde{Z}(s) = k) \\ &= \sum_{m=1}^{k-1} \sum_{r=1}^{n-1} \mathbb{P}(X_1(t) = r,\ X_2(t) = n - r,\ X_1(s) = m,\ X_2(s) = k - m \mid \widetilde{Z}(s) = k) \\ &= \sum_{m=1}^{k-1} \sum_{r=1}^{n-1} \mathbb{P}(X_1(t) = r \mid X_1(s) = m)\mathbb{P}(X_2(t) = n - r \mid X_2(s) = k - m) \\ &\qquad \times \mathbb{P}(X_1(s) = m,\ X_2(s) = k - m \mid \widetilde{Z}(s) = k).\end{aligned}$$

All three terms in the above equation are known. A result of interest is the following:

$$\begin{aligned}\mathbb{P}(X_1(t) = m,\ X_2(t) = k - m \mid \widetilde{Z}(t) = k) &= \frac{\mathbb{P}(X_1(t) = m,\ X_2(t) = k - m)}{\mathbb{P}(\widetilde{Z}(t) = k)} \\ &= \frac{\mathbb{P}(X_1(t) = m)\ \mathbb{P}(X_2(t) = k - m)}{\mathbb{P}(\widetilde{Z}(t) = k)} \\ &= \frac{1}{k-1}.\end{aligned}$$

In other words, given the value of $\widetilde{Z}(t)$, the diversity is split uniformly across the two sides of the tree, i.e., $X_1|\widetilde{Z}(t) > 0 \sim U\{1,\ \ldots, \widetilde{Z}(t)\}$.

An oversight was made in previous publications [94, 118] when the parameters γ and ρ, used in the logistic growth equation, were fixed at incorrect values. We can see from Equation (2.10) that as $t \to \infty$, the expected population size tends to $2/\gamma$. As the known diversity at the time of publication was thought to be 235 species, the authors set $\gamma = 2/235 = 0.0085$. The value of ρ was fixed at $\rho = 0.2995$ by assuming that 90% of modern diversity had been reached by the Middle Eocene, 49 My ago. While it is true that the expected final population size will be 235 when using these values, it should be noted that this is not the case if you then condition on non-extinction. The problem arises due to the difference between Equations (2.9) and (2.19).

There are essentially two different ways to condition the process on non-extinction. The classical approach is to look at the population size at time t, conditional on the tree surviving to time t. In this case, the population size no longer tends to a limiting value:

$$\begin{aligned}\mathbb{E}(\widetilde{Z}(t) \mid \widetilde{Z}(t) > 0) &= \frac{2}{1-\eta} \\ &= \frac{\lambda(\gamma t + (\frac{1}{\lambda} - \frac{1}{\rho})(1-\gamma)(e^{-\rho t} - 1))}{\gamma + (1-\gamma)e^{-\rho t}}\end{aligned} \tag{2.20}$$

where I have substituted Equations (2.13) and (2.14) for ξ and η. As $t \to \infty$,

$$\mathbb{E}\left(\frac{Z(t)}{t} \mid Z(t) > 0\right) \to \frac{\lambda}{2}.$$

Under this conditioning scheme we find that $\mathbb{E}(\widetilde{Z}(t)|\widetilde{Z}(t) > 0) \to \infty$ as t goes to infinity. This result was noted by Foote and Raup [44] in the case of constant origination and extinction rates[3], and Asmussen and Hering [6] proved that conditional on non-extinction, the discrete-time Galton-Watson process tends to infinity almost surely. Note that the long-term expected behaviour of $\widetilde{Z}(t)$ now depends on the average species lifetime, $1/\lambda$, whereas λ does not appear in the logisitic equation (2.10).

The other way to condition on non-extinction is to look at the population size at time t conditional on the tree surviving to some future time T. This is the form we are interested in as we know the process survives to the present. Calculating this expectation we find that

$$\mathbb{E}(Z(t) \mid Z(T) > 0) = \frac{(1-\xi(0,t))(1-\xi(t,T))(1-\eta^2(0,t)\xi(t,T))}{(1-\xi(0,T))(1-\eta(0,t))(1-\eta(0,t)\xi(t,T))^2} \text{ for } t < T \qquad (2.21)$$

which is different to Equation (2.10) and Equation (2.20). I have made explicit the dependance of ξ and η on time as the values vary with both t and T. Figure 2.2 shows all three expected growth curves for $\gamma = 0.0085, \rho = 0.2995$ and $\lambda = 0.4$, which are the values used in Tavaré *et al.* [118]. Notice how much faster and larger the population grows when we condition on non-extinction. In Tavaré *et al.* the parameter values were chosen so as to make the expected diversity equal to the observed diversity of 235. However, Equation (2.10) was used for the calculations, rather than one of the two conditional forms, so the parameter choice would not have had the desired effect. Our simulations assume that extinction does not take place, and therefore we should use Equation (2.21) to determine the parameter values. In fact, we do not take this approach, as we wish to allow the parameters (including $1/\lambda$) to vary and to see if there is any signal from the data about their true value.

A note on the simulations

The way we described the model makes clear how to simulate $X(t)$ and hence $Z(t)$. To simulate $\widetilde{Z}(t)|\widetilde{Z}(T) > 0$, we simulate realisations of $Z(t)$, and discard trees in which either side of the tree dies out. This is an inefficient approach, however, as a large proportion of the simulation time is wasted on simulating trees which go extinct. It is not currently known how to efficiently simulate Galton-Watson trees conditioned on non-extinction.

[3] We are using heterogeneous rates $p_0(t)$ and $p_2(t)$.

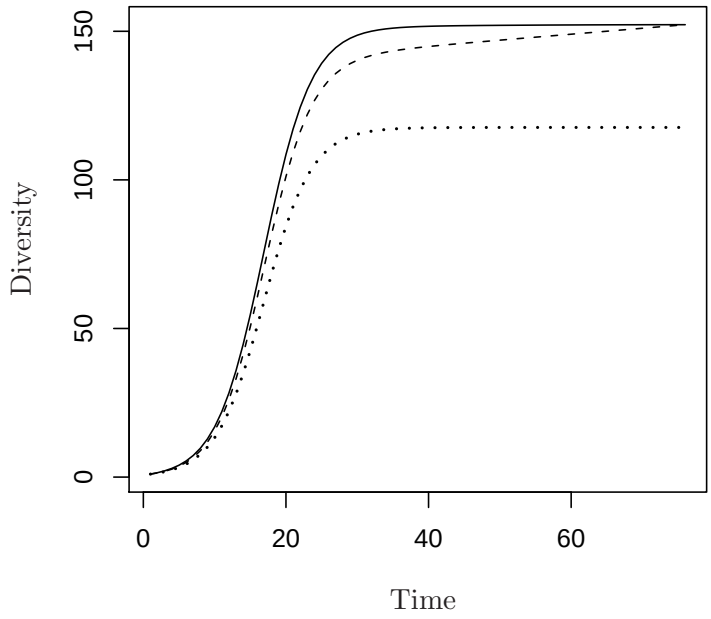

Figure 2.2: *Plots of the expected diversity curves for (i)* $\mathbb{E}(Z(t))$ *(dotted/bottom line); (ii)* $\mathbb{E}(\widetilde{Z}(t) \mid \widetilde{Z}(t) > 0)$ *(dashed/middle line); and (iii)* $\mathbb{E}(\widetilde{Z}(t) \mid \widetilde{Z}(T) > 0)$ *(solid/upper line) when* $\gamma = 0.0085, \rho = 0.2995, \lambda = 0.4$.

2.3.2 The Cumulative Process

In Equation (2.16), the likelihood function is decomposed into two parts. The first is part of the model definition, whereas the second, $\mathbb{P}(\mathcal{N} = (N_1, \ldots, N_k)|\boldsymbol{\lambda})$, we must calculate. In an attempt to calculate this probability, let $N(t)$ be the cumulative process, that is, the number of species that have existed by time t. Note that $N(t)$ is not Markovian. To see this consider two trees, one with $N(t) = Z(t) = n$ and the other with $N(t) = n, Z(t) = 0$. In the former, $N(t)$ will increase with positive probability, whereas in the latter case the population is extinct and so $N(s) = 0$ for all $s > t$.

The two-dimensional process $(N(t), Z(t))$ is, however, a Markov Chain. The chain can make the following transitions:

$$\begin{array}{lcll} (n,z) & \to & (n,z-1) & \text{at rate } z\lambda p_0(t) \quad \text{(a death)} \\ (n,z) & \to & (n+2,z+1) & \text{at rate } z\lambda p_2(t) \quad \text{(a birth)} \\ (n,z) & \to & (n,z) & \text{at rate } 1-z\lambda \quad \text{(no change).} \end{array}$$

Substituting these rates into the Kolmogorov forward equation gives

$$\begin{aligned}\frac{\partial P_{(n,z)}}{\partial t}(t) &= \sum_{(k,l)\in\mathbb{Z}^2} P_{(k,l)}(t)A_{(k,l)(n,z)}(t) \\ &= (z+1)\lambda p_0(t)P_{(n,z+1)}(t) + (z-1)\lambda p_2(t)P_{(n-2,z-1)}(t) - z\lambda P_{(n,z)}(t).\end{aligned}$$

Multiplying by $u^z w^n$, summing over z and n, and letting $G(u,w,t) = \mathbb{E}(u^{Z(t)}w^{N(t)})$ gives the partial differential equation

$$\frac{\partial G}{\partial t} = \left(\lambda p_2(t)u^2w^2 - \lambda u + \lambda p_0(t)\right)\frac{\partial G}{\partial u}, \tag{2.22}$$

with boundary condition $G(u,w,0) = uw$. This is a Riccati equation. No solution is listed in the *Handbook of First Order Partial Differential Equations* by Polyanin *et al.* [97], and Riccati equations do not in general have known solutions. I believe this is the case here.

Kendall [67] approaches the problem by considering the birth count $B(t)$. This leads, in a similar way to above, to an insoluble partial differential equation. He is able, however, to give the asymptotic distribution of the cumulative process for the homogeneous subcritical binary Galton-Watson process (i.e., where $p_0(t) \equiv p_0 > p_2 \equiv p_2(t)$). Puri [99] and Dwass [40] derive related distributions, also, however, in the limit as $t \to \infty$.

The insolubility of Equation (2.22) means that we are unable to find an expression for the likelihood function, $\mathbb{P}(\mathcal{D}|\theta)$. As the likelihood is the medium through which we would usually do inference, we shall need to consider a non-standard approach to inference. Fortunately, all is not lost. For although the likelihood is unknown, we are able to simulate data sets from the model. In the next chapter I show that this is sufficient to enable us to do inference.

Chapter 3

Approximate Bayesian Computation

The past few decades have seen a revolution in the approach mathematicians and scientists take to modelling. The increased availability of cheap computer power has meant that a modeller is now able to solve on his desktop machine equations for which he would have needed an expensive supercomputer as little as a decade or two ago. This has led to the consideration of models of greater complexity than was previously possible. Models with tens or hundreds of parameters are now commonplace, and, as is to be expected, these models are often mathematically intractable.

The main inference methods used for complex models are a collection of algorithms known as Monte Carlo methods, and these are now commonplace in many fields of research. However, most of these methods depend upon knowledge of the likelihood function, and as in the previous chapter this is not always known. The likelihood function, $\mathbb{P}(\mathcal{D}|\theta)$, is of fundamental importance to both the frequentist and the Bayesian schools of statistical inference. The frequentist approach is based on maximum likelihood estimation in which we aim to find $\hat{\theta} = \arg\max_\theta \mathbb{P}(\mathcal{D}|\theta)$, whereas the Bayesian approach is based around finding the posterior distribution $\pi(\theta|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)$. If the likelihood is unknown, both of these approaches may be impossible, and we will need to perform *likelihood-free* inference.

There are a large number of models for which the likelihood function is unknown, but for which it is possible to simulate sample data sets from the model, $\mathbb{P}(\cdot|\theta)$. *Likelihood-free* inference techniques that can be applied in this situation have been developed over the previous decade and are often known as *Approximate Bayesian Computation* (ABC) methods. The first use of ABC ideas was in the field of genetics, as developed by Tavaré *et al.* [117] and Pritchard *et al.* [98]. ABC methods have since proved useful in a variety of scientific application areas, including epidemiology, ecology, and genetics (see [111]).

The basic algorithm is based upon the rejection method, but has since been extended in various ways. Marjoram *et al.* [76] suggested an approximate MCMC algorithm (cf. Chapter 5), Sisson *et al.* [112] a sequential importance sampling approach, and Beaumont, Zhang and Balding [11] improved the accuracy by performing local-linear regression on the output (see Chapter 4 for more information). I give an extension in Chapter 5 by combining MCMC with ABC to produce a hybrid sampler.

ABC methods are, as their name suggests, approximate. Little, however, is known about their accuracy. In this chapter I outline the basic algorithm and illustrate the approach with a few simple examples. I draw special attention to some of the problems and challenges still to be solved, many of which are vital if these methods are to become part of the statisticians tool-kit. In Sections 3.1.4 and 3.1.5 I give some of the advantages and disadvantages of ABC compared with MCMC. Finally, in Section 3.2 I apply these methods to the primate divergence time problem given in Chapter 2.

3.1 A Short Introduction to Approximate Bayesian Computation

Suppose that we have discrete data $\mathcal{D}$ which we assume is from model $\mathcal{M}$, parameterised by parameter θ (possibly multidimensional), and that θ has the prior distribution $\pi(\theta)$. Our principal aim throughout this chapter is to find the posterior distribution of the parameter given the data:

$$\pi(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\pi(\theta)}{\mathbb{P}(\mathcal{D})}.$$

Here, $\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)\mathrm{d}\theta$ is a normalising constant often referred to as the *evidence* for model $\mathcal{M}$, or the *marginal likelihood*. For most models an analytic expression for $\pi(\theta|\mathcal{D})$ can not be found, usually because the marginal likelihood can not be computed. The usual alternative to analytic computation is to use one of the numerous Monte Carlo techniques that have been developed. Our problem goes deeper than an inability to calculate the marginal likelihood, we cannot calculate the likelihood term $\mathbb{P}(\mathcal{D}|\theta)$ either. However we can simulate data sets $\mathcal{D}$ from the model and so all is not lost.

3.1.1 Basic Algorithm

The simplest approximate Bayesian computation algorithm [98] is based upon the rejection algorithm. This was first given by von Neumann [120] in 1951 and is as follows:

Algorithm A: Rejection Algorithm 1

A1 Draw θ from $\pi(\cdot)$,

A2 Accept θ with probability $r = \mathbb{P}(\mathcal{D}|\theta)$.

This algorithm, as it stands, is of no help. For the class of models we are interested in, the likelihood $\mathbb{P}(\mathcal{D}|\theta)$ is unknown. However, there is an alternative version of this algorithm which does not depend on explicit knowledge of the likelihood, but only requires that we can simulate from the model:

Algorithm B: Rejection Algorithm 2

B1 Draw θ from $\pi(\cdot)$,

B2 Simulate data $\mathcal{D}'$ from $\mathbb{P}(\cdot|\theta)$,

B3 Accept θ if $\mathcal{D} = \mathcal{D}'$.

Algorithm B can be thought of as a mechanical version of Bayes Theorem. To see that these two algorithms do indeed give a sample from the posterior distribution, set $I = \mathbb{I}_{\theta \text{ is accepted}}$ to be the indicator of whether θ is accepted. Then

$$\mathbb{P}(I = 1) = \int \mathbb{P}(I = 1|\theta)\pi(\theta)\mathrm{d}\theta = \int \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)\mathrm{d}\theta = \mathbb{P}(\mathcal{D}),$$

which gives that

$$\mathbb{P}(\theta|I = 1) = \frac{\mathbb{P}(I = 1|\theta)\pi(\theta)}{\mathbb{P}(I = 1)} = \pi(\theta|\mathcal{D})$$

as is required.

The acceptance rate of both algorithms is the normalising constant, $\mathbb{P}(I = 1) = \mathbb{P}(\mathcal{D})$. Thus, to get a sample of size n we have to wait a period of time that has a negative binomial distribution, with parameters n and $\mathbb{P}(\mathcal{D})$. The mean number of data sets we shall need to simulate is $n(1 - \mathbb{P}(\mathcal{D}))/\mathbb{P}(\mathcal{D})$, which is large when $\mathbb{P}(\mathcal{D})$ is small[(1)]. This presents us with a problem: for complex problems the probability that a simulated data set, $\mathcal{D}'$, exactly matches the observed data, $\mathcal{D}$, will be tiny, and hence Algorithm B will run extremely slowly.

[(1)] Note that the acceptance rate gives a natural way to estimate the usually tricky to calculate normalising constant $\mathbb{P}(\mathcal{D})$. In Chapter 6 I look at using the acceptance rate to do model selection via the Bayes factor.

A solution to this problem is to relax the requirement of equality in step B3, and to accept θ values when the simulated data is 'close' to the real data. To define 'close' we require a metric ρ on the state space $\mathcal{X}$, with $\rho(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, and a *tolerance* ϵ. The algorithm, our first approximate Bayesian computation algorithm, is then as follows:

Algorithm C: ABC 1

C1 Draw θ from $\pi(\cdot)$,

C2 Simulate data $\mathcal{D}'$ from $\mathbb{P}(\cdot|\theta)$,

C3 Accept θ if $\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$.

Unlike algorithms A and B, Algorithm C is not exact. Accepted θ values do not form a sample from the posterior distribution, but from some distribution that is an approximation to it. The accuracy of the algorithm (measured by some suitable distance measure) depends on ϵ in a non-trivial manner (cf. Section 3.1.2). Let $\pi_\epsilon(\theta)$ denote the distribution of accepted θ values when using tolerance ϵ. The distribution $\pi_\epsilon(\theta)$ also depends on the choice of metric, but this is not made explicit in the notation. If $\epsilon = 0$, then we only accept θ values when the simulated data precisely matches the observed data. In this case Algorithm C is equivalent to Algorithm B and gives a sample from the posterior distribution, i.e., $\pi_0(\theta) = \pi(\theta|\mathcal{D})$ and the algorithm is exact[(2)]. If $\epsilon = \infty$, all θ values are accepted regardless of the simulated data. In this case we will be left with a sample from the prior distribution, i.e., $\pi_\infty(\theta) = \pi(\theta)$.

Note that the approximation in step C3 means that the algorithm can also be used for continuous models. I will continue to use $\mathbb{P}(\mathcal{D}|\theta)$ to denote the model likelihood, even though it may now represent a density function rather than a probability. This abuse of notation should hopefully not cause any confusion. The acceptance rate is

$$\begin{aligned}\mathbb{P}(I = 1) &= \int_\Theta \mathbb{P}(\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon|\theta)\pi(\theta)\mathrm{d}\theta \\ &= \int_\Theta \int_{\mathcal{D}':\rho(\mathcal{D},\mathcal{D}')\leq\epsilon} \mathbb{P}_\theta(\mathcal{D}'|\theta)\pi(\theta)\mathrm{d}\mathcal{D}'\mathrm{d}\theta\end{aligned}$$

which is an increasing function of ϵ, as is expected (Θ denotes the parameter space). If the data are discrete, the integrals over the data can be replaced by sums over the same range.

The tolerance ϵ represents a trade-off between computability and accuracy; large ϵ values will mean more acceptances in step C3 above and will enable us to generate samples

[(2)]Note that I am assuming that $\rho(\cdot,\cdot)$ is a metric, so that $\rho(\mathcal{D}, \mathcal{D}') = 0$ implies $\mathcal{D} = \mathcal{D}'$. In general this will not be true, as it will be necessary to summarise the data if it is high dimensional. In this case ρ will be a distance function, but not a metric. See Section 3.1.3 for more details.

more quickly. However, the distribution obtained, $\pi_\epsilon(\theta)$, may be a poor approximation to $\pi(\theta|\mathcal{D})$. Small ϵ values will mean that our approximation is more accurate, but the acceptance rate in step C3 will be lower and so we will require more computer time to generate a sample of a given size. To choose a value of ϵ, we must decide upon a compromise between the amount of time and computer power available, and the desired accuracy. If we have a large cluster of computers available for use, we will be able to use a smaller value of ϵ than if we only have a single desktop machine. Knowledge of acceptance rates for different values of ϵ is at present a matter of experience. In general we find that

$$\begin{aligned}\pi_\epsilon(\theta) &= \frac{\int_{\mathcal{D}':\rho(\mathcal{D},\mathcal{D}')\leq\epsilon}\mathbb{P}(\mathcal{D}'|\theta)\pi(\theta)\mathrm{d}\mathcal{D}'}{\int_\Theta\int_{\mathcal{D}':\rho(\mathcal{D},\mathcal{D}')\leq\epsilon}\mathbb{P}(\mathcal{D}'|\theta)\pi(\theta)\mathrm{d}\mathcal{D}'\mathrm{d}\theta}\\ &= \frac{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon|\theta)\pi(\theta)}{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon)}.\end{aligned}$$

Output from Algorithm C satisfies

$$\mathbb{E}(\theta|\rho(\mathcal{D},\mathcal{D}')\leq\epsilon) = \frac{\int_A\mathbb{E}(\theta|\mathcal{D}')\mathbb{P}(\mathcal{D}')\mathrm{d}\mathcal{D}'}{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon)}$$

where $A=\{\mathcal{D}':\rho(\mathcal{D},\mathcal{D}')\leq\epsilon\}$ and

$$\mathbb{E}(\theta^2|\rho(\mathcal{D},\mathcal{D}')\leq\epsilon) = \frac{\int_A\mathbb{E}(\theta^2|\mathcal{D})\mathbb{P}(\mathcal{D})\mathrm{d}\mathcal{D}}{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon)} = \frac{\int_A[\mathbb{V}\mathrm{ar}(\theta|\mathcal{D})+\mathbb{E}(\theta|\mathcal{D})^2]\mathbb{P}(\mathcal{D})\mathrm{d}\mathcal{D}}{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon)}.$$

These results follow from an application of Fubini's Theorem:

$$\begin{aligned}\mathbb{E}(\theta|\rho(\mathcal{D},\mathcal{D}')\leq\epsilon) &= \int_\Theta\theta\pi_\epsilon(\theta)\mathrm{d}\theta\\ &= \int_\Theta\frac{\theta\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon|\theta)\pi(\theta)\mathrm{d}\theta}{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon)}\\ &= \frac{\int_\Theta\theta\pi(\theta)\int_A\mathbb{P}(\mathcal{D}|\theta)\mathrm{d}\mathcal{D}\mathrm{d}\theta}{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon)}\\ &= \frac{\int_A\int_\Theta\theta\pi(\theta)\mathbb{P}(\mathcal{D}|\theta)\mathrm{d}\theta\mathrm{d}\mathcal{D}}{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon)} \quad \text{by Fubini's Theorem}\\ &= \frac{\int_A\mathbb{E}(\theta|\mathcal{D})\mathbb{P}(\mathcal{D})\mathrm{d}\mathcal{D}}{\mathbb{P}(\rho(\mathcal{D},\mathcal{D}')\leq\epsilon)}\end{aligned}$$

A similar calculation works for the second moment. Experience suggests that under suitable conditions on the choice of metric, the variance of the approximation $\pi_\epsilon(\mu)$ is always greater than the true posterior variance. If this is true, the approximation would always be over-dispersed compared with the true distribution and thus the approximation is a conservative one. That is, predicted posterior credibility intervals will be larger than the true posterior credibility interval.

3.1.2 An Illustrative example

I now illustrate the basic idea with a simple normal example for which we can compute everything analytically. This will allow us to examine more closely the effects of the ABC algorithm. Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ are an independent sample from the normal distribution with known variance σ^2. For simplicity suppose that the mean μ has a uniform prior on $[a, b]$. The posterior distribution of μ, given the data, can be found as follows:

$$\begin{aligned} \pi(\mu|\mathcal{D}) &\propto \mathbb{I}_{a\leq\mu\leq b} \exp\left(-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right) \\ &\propto \mathbb{I}_{a\leq\mu\leq b} \exp\left(-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right). \end{aligned}$$

So the posterior distribution of μ is a truncated normal distribution (truncated outside of $[a, b]$) with mean $\bar{x}$ and variance σ^2/n. If the prior distribution for μ is the uninformative improper prior $\pi(\mu) \propto 1$, $\mu \in \mathbb{R}$, then the posterior distribution becomes $\mu|\mathcal{D} \sim \text{Normal}(\bar{x}, \frac{\sigma^2}{n})$.

The mean of the data is sufficient for μ (cf. Section 3.1.3 for details of summarising data) and so we can compare data sets by comparing their means. I shall assume without loss of generality that the data has a mean value of 0. Using the absolute distance metric $\rho(\mathbf{x}, \mathbf{x}') = |\bar{\mathbf{x}} - \bar{\mathbf{x}}'|$ gives the following algorithm:

- Pick μ from the prior distribution,
- Simulate $X'_1, \ldots, X'_n$ from $N(\mu, \sigma^2)$,
- Accept μ if $|\bar{x}' - 0| \leq \epsilon$.

Figure 3.1 shows plots of the sample distribution obtained from this algorithm, along with the true posterior distributions for various different values of ϵ. The dashed lines are the densities obtained from using kernel estimates on 1000 successful draws from the above algorithm. Notice that for small ϵ the approximation is excellent, but for larger ϵ we have an over dispersed approximation, as expected.

For the normal distribution with uniform priors, calculation of $\pi_\epsilon(\theta)$ is possible. We can write down the approximate density in a semi-explicit form:

$$\begin{aligned} \pi_\epsilon(\mu) &= \frac{\int_{-\epsilon}^{\epsilon} \sqrt{\frac{n}{2\sigma^2}} e^{-\frac{n}{2\sigma^2}(y-\mu)^2} \mathrm{d}y}{\int_{-\infty}^{\infty}\int_{-\epsilon}^{\epsilon} \sqrt{\frac{n}{2\sigma^2}} e^{-\frac{n}{2\sigma^2}(y-\mu)^2} \mathrm{d}y\mathrm{d}\mu} \\ &= \frac{\Phi\left(\frac{\epsilon-\mu}{\sqrt{\sigma^2/n}}\right) - \Phi\left(\frac{-\epsilon-\mu}{\sqrt{\sigma^2/n}}\right)}{2\epsilon}. \end{aligned} \tag{3.1}$$

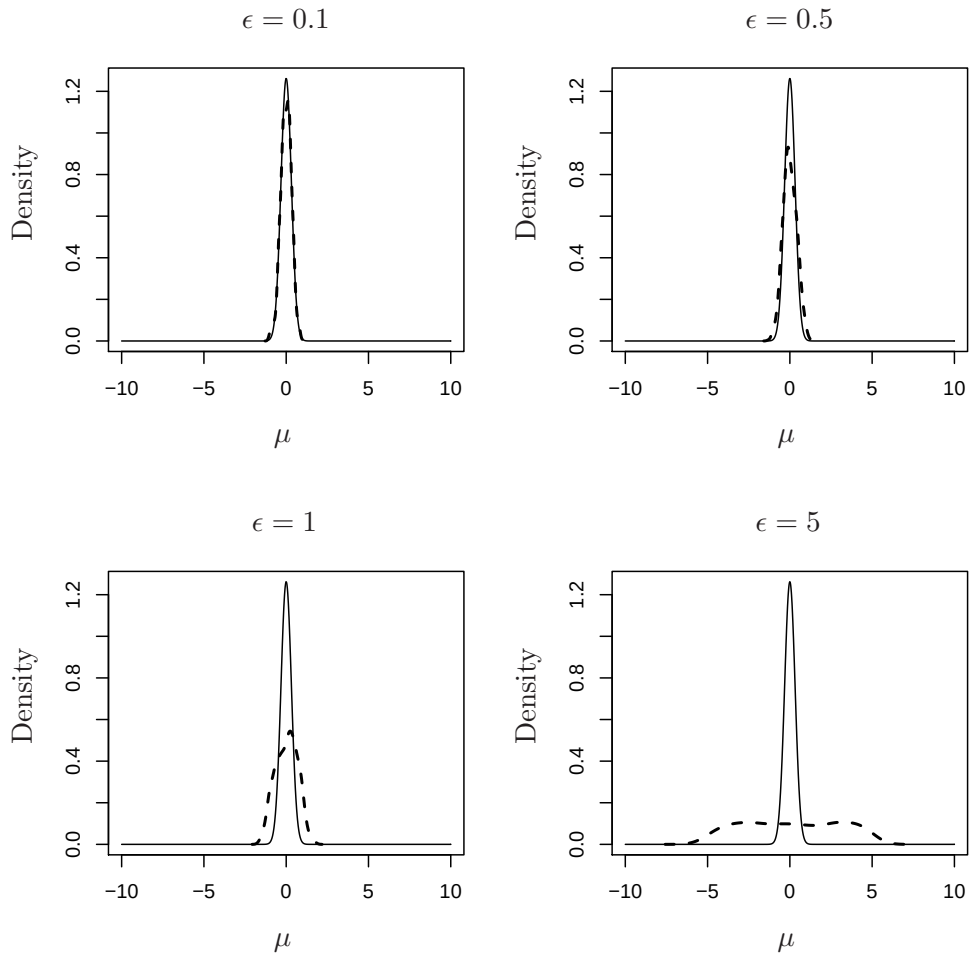

Figure 3.1: *Plots of the true posterior distribution for μ (solid line), and the ABC estimate using 1000 samples (dashed line). A value of $\sigma^2 = 1$ was used for all four plots.*

where Φ is the standard normal distribution function. In a similar way to above, we can calculate the first two moments of the approximate distribution:

$$\begin{aligned}
\mathbb{E}(\mu||\bar{x}| \leq \epsilon) &= \int_{-\infty}^{\infty} \frac{\mu \mathbb{P}(|\bar{x}| \leq \epsilon|\mu)\pi(\mu)}{\mathbb{P}(|\bar{x}| \leq \epsilon)} \mathrm{d}\mu \\
&= \frac{\int_{-\infty}^{\infty} \mu \int_{-\epsilon}^{\epsilon} \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n}{2\sigma^2}(y-\mu)^2} \mathrm{d}y \mathrm{d}\mu}{\int_{-\infty}^{\infty} \int_{-\epsilon}^{\epsilon} \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n}{2\sigma^2}(y-\mu)^2} \mathrm{d}y \mathrm{d}\mu} \\
&= \frac{\int_{-\epsilon}^{\epsilon} \int_{-\infty}^{\infty} \mu \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n}{2\sigma^2}(y-\mu)^2} \mathrm{d}\mu \mathrm{d}y}{\int_{-\epsilon}^{\epsilon} \int_{-\infty}^{\infty} \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n}{2\sigma^2}(y-\mu)^2} \mathrm{d}\mu \mathrm{d}y} \text{ by Fubini's Theorem} \\
&= 0
\end{aligned}$$

$$\mathbb{V}\mathrm{ar}(\mu|\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon) = \frac{\int_{-\infty}^{\infty} \mu^2 \int_{-\epsilon}^{\epsilon} \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n}{2\sigma^2}(y-\mu)^2} \mathrm{d}y \mathrm{d}\mu}{\int_{-\infty}^{\infty} \int_{-\epsilon}^{\epsilon} \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n}{2\sigma^2}(y-\mu)^2} \mathrm{d}y \mathrm{d}\mu} = \frac{\sigma^2}{n} + \frac{1}{3}\epsilon^2.$$

This shows that the estimate of the posterior mean is unbiased in this case. We can also see that the approximate posterior variance increases quadratically in ϵ. This is a precise mathematical description of the over-dispersion of $\pi_\epsilon(\mu)$ observed above.

We can examine the error between the approximation and the true posterior distribution by performing a Taylor expansion on Equation (3.1). Letting $x = -\mu/\sqrt{\sigma^2/n}$ and $h = \epsilon/\sqrt{\sigma^2/n}$ and Taylor expanding the Gaussian distribution function about x for small h gives:

$$\Phi\left(\frac{\epsilon-\mu}{\sqrt{\sigma^2/n}}\right) = \Phi(x+h) = \Phi(x) + h\Phi'(x) + \frac{h^2\Phi''(x)}{2!} + \frac{h^3\Phi'''(x)}{3!} + o(h^3)$$

so that

$$\begin{aligned}\pi_\epsilon(\mu) &= \frac{\Phi(x+h)-\Phi(x-h)}{2\epsilon} \\ &= \frac{1}{\epsilon}\left(h\Phi'(x) + \frac{h^3}{6}\Phi'''(x)\right) + o(h^2)\end{aligned}$$

Because $\Phi'(x) = \phi(x)$, i.e., the standard Gaussian probability density function, we can see that to first order the approximation is exact:

$$\pi_\epsilon(\mu) = \pi(\mu|\mathcal{D}) + o(\epsilon).$$

To second order we find that

$$\pi_\epsilon(\mu) = \left(1 - \frac{n\epsilon^2}{\sigma^2} + \frac{n^2\epsilon^2\mu^2}{\sigma^4}\right)\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\Big(-\frac{\mu^2 n}{2\sigma^2}\Big) + o(\epsilon^2).$$

The total variation distance between the posterior distribution and the approximation is defined to be

$$d_{TV}(\pi_\epsilon(\mu), \pi(\mu|\mathcal{D})) = \frac{1}{2}\int |\pi(\mu|\mathcal{D}) - \pi_\epsilon(\mu)|\,\mathrm{d}\mu.$$

We can calculate this as follows:

$$\begin{aligned}d_{TV}(\pi_\epsilon(\mu), \pi(\mu|\mathcal{D})) &= \frac{1}{2}\int_{-\infty}^{\infty}\left|\frac{n^2\epsilon^2\mu^2}{\sigma^4} - \frac{n\epsilon^2}{\sigma^2}\right|.\frac{\exp(-\frac{\mu^2 n}{2\sigma^2})}{\sqrt{2\pi\sigma^2/n}}\,\mathrm{d}\mu \\ &= \frac{n\epsilon^2}{2\sigma^2}\int_{-\infty}^{\infty}|x^2-1|\frac{\exp(-x^2/2)}{\sqrt{2\pi}}\mathrm{d}x \\ &= \frac{n\epsilon^2}{2\sigma^2}\left(\int_{-1}^{1}(1-x^2)\frac{\exp(-x^2/2)}{\sqrt{2\pi}}\mathrm{d}x + 2\int_1^{\infty}(x^2-1)\frac{\exp(-x^2/2)}{\sqrt{2\pi}}\mathrm{d}x\right) \\ &= \sqrt{\frac{2}{\pi}}\exp(-1/2)\frac{\epsilon^2}{\sigma^2/n}\end{aligned}$$

Hence, the total variation distance between the two distributions is approximately

$$\frac{cn\epsilon^2}{\sigma^2} + o(\epsilon^2)$$

where $c = \sqrt{2/\pi}\exp(-1/2) \approx 1/2$. Thus, for a given size of error, the tolerance value we are required to use depends on the size of the posterior variance σ^2/n. For small posterior variances we shall need to use a smaller value of ϵ, whereas for large variances we can use a larger values of ϵ.

Another simple example involves data $X \sim \text{Poisson}(\theta)$, with a conjugate $\text{Gamma}(\alpha, \beta)$ prior distribution for θ. The posterior distribution of θ is then a $\text{Gamma}(x + \alpha, 1 + \beta)$ distribution. Using the metric $\rho(x, x') = |x - x'|$, where x is the observed data, we find that the approximate posterior distribution can be computed exactly. By writing

$$\mathbb{P}(\rho(X, x) \leq \epsilon|\theta) = \sum_{y=x-\epsilon}^{y=x+\epsilon} \frac{e^{-\theta}\theta^y}{y!}$$

we find that for general ϵ, the approximate posterior distribution is

$$\pi_\epsilon(\theta) = \frac{\theta^{\alpha+x+\epsilon-1}e^{-\theta(1+\beta)}}{Z}\left(\sum_{n=0}^{2\epsilon}\frac{\theta^n(x-\epsilon)!}{(x-\epsilon+n)!}\right)$$

where the normalising constant Z, is

$$Z = \frac{1}{(1+\beta)^{\alpha+x-\epsilon}}\sum_{n=0}^{2\epsilon}\frac{(x-\epsilon)!}{(x-\epsilon+n)!}\frac{\Gamma(x-\epsilon+\alpha+n)}{(1+\beta)^n}.$$

The left-hand plot of Figure 3.2 shows how $\pi_\epsilon(\theta)$ changes as ϵ takes values from 0 to 4. The red curve with the highest peak (and smallest variance) is the true posterior ($\epsilon = 0$), and as ϵ decreases we can see that the dispersion increases. The green curve furthest to the right is the prior and we can see this is close to $\pi_4(\theta)$. The right hand plot shows how the total variation distance between the two distributions changes as ϵ increases for given parameter values.

3.1.3 Data Summaries

For problems with large amounts of high-dimensional data, Algorithm C will be impractical, as the simulated data will never closely match the observed data. The standard approach to reducing the number of dimensions is to use a (possibly multidimensional) summary statistic, $S(\mathcal{D})$ say, which should summarise the important parts of the data. We then adapt Algorithm C so that parameter values are accepted if the summary of the simulated data is 'close' to the summary of the real data.

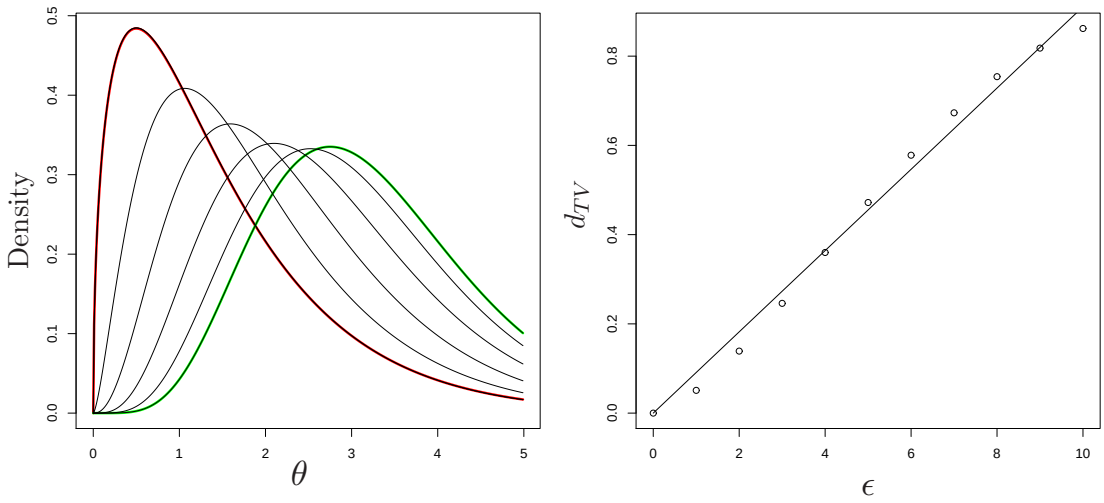

Figure 3.2: *The plot on the left shows $\pi_\epsilon(\theta)$ as ϵ varies from 0 (red left hand curve) to 4. The green curve furthest to the right is the prior density. The right hand plot shows how variation distance between the approximation and the posterior changes with ϵ. The following values were used: $\alpha = 1.5, \beta = 1, x = 10$.*

Algorithm D: ABC 2

D1 Draw θ from $\pi(\cdot)$,

D2 Simulate data $\mathcal{D}'$ from $\mathbb{P}(\cdot|\theta)$,

D3 Accept θ if $\rho(S(\mathcal{D}), S(\mathcal{D}')) \leq \epsilon$.

The statistic $S(\mathcal{D})$ is called a *sufficient statistic* for θ if and only if $\pi(\theta|\mathcal{D}) = \pi(\theta|S(\mathcal{D}))$, i.e., if the conditional distribution of θ given the summary equals the posterior distribution[(3)]. The idea is that if $S(\mathcal{D})$ is known, then the full data set, $\mathcal{D}$, can not provide any extra information about θ.

If a sufficient statistic is available, then Algorithm D is essentially the same as Algorithm C (in the normal distribution example given above, we really used Algorithm D). However, for problems where both the posterior distribution and the likelihood function are unknown, it will not in general be possible to determine whether a statistic is sufficient. Instead, it is left to the intuition and experience of the practitioner to find a summary statistic that works well and captures the key aspects of the data. For difficult problems this may be a case of trial and error, rather than following a particular strategy.

[(3)]This is the Bayesian definition of sufficiency. An equivalent definition is that S is sufficient if and only if the conditional distribution of $\mathcal{D}$ given $S(\mathcal{D})$ does not depend upon θ.

3.1.4 Advantages of ABC

The approximate Bayesian computation techniques listed here and elsewhere have potential advantages over many other Monte Carlo techniques. In this section I list some of these advantages. However, it should be noted that these are all subject to the provisos of the following section.

The first thing to note, is that if the likelihood function is unavailable or prohibitively expensive to compute we will be unable to use traditional Monte Carlo approaches. However, even if the likelihood function is known, ABC methods can be useful as a potential first-pass inference technique.

Many Monte Carlo techniques can be difficult to code and suffer from tuning problems that can take the statistician much time to solve. Markov chain Monte Carlo (MCMC) algorithms, for example, have several associated difficulties: the Markov chains may not converge to equilibrium, there may be many parameters to fine-tune to control the mixing, and the output received is dependent and so the output may need to be thinned. The need to 'tune' so many parameters means that using MCMC is often time consuming as each time the model is changed, the mixing parameters must be retuned in order to get satisfactory results. On the other hand, ABC methods need no readjustment upon changes to the model, which means they could be used in the development stage when the model is likely to undergo several iterations[(4)]. Another advantage of ABC over MCMC, is that ABC can be run in parallel on different machines.

ABC methods also allow different types of data to be used, even those with a complex dependency structure. In many fields such as epidemiology, population genetics and evolutionary biology, there is often an underlying unobserved tree structure, which leads to complex dependencies in the data. In Section 4.2, I show that combining different data types is simple when using ABC and is done by changing the metric that is used.

The acceptance rate of these algorithms may provide a natural measure of model fit that can be used to estimate Bayes factors. These can then be used in model selection. This idea is briefly explored in Chapter 6.

Finally, there is a large appetite in the scientific community for any method that allows for easier use of Bayesian methods. For example, the BUGS project [115] provides free and easy to use software that simplifies the setting up of Bayesian hierarchical models and then helps to find the posterior distributions using Gibbs sampling. ABC, unlike MCMC, is almost trivial to code once you are able to simulate observations from the model.

[(4)] The ABC algorithms do require some tuning as we must choose a metric and a value for ϵ. However, this can be done after the simulations have been run. If we save all the output from our simulations, we can then post-process this data by trying various different values for ϵ and ρ. This allows us to choose optimal values with a minimal amount of computation.

3.1.5 Disadvantages and Problems

ABC methods are still in their infancy. There are many technical questions which remain to be answered, and until this time it is likely that these techniques will be used as a method of last resort only. For example, currently it is not known how accurate the approximation $\pi_\epsilon(\theta)$ is, or how the accuracy depends on the choice of metric and ϵ. It is shown in Section 4.5 that the choice of metric does matter.

Another problem is that if data summaries need to be used, the intuition of the practitioner is relied upon when choosing a good summary statistic. Ideally, some notion of approximate sufficiency is needed (cf. Le Cam [26]) along with a methodical way of finding summaries which are nearly sufficient and which therefore capture the pertinent parts of the data.

However, even if these two technical issues are addressed satisfactorily, some disadvantages of ABC will remain. The most serious of these, as pointed out by Sisson [111], is that due to sampling from the prior distribution each time, these algorithms often run slowly when compared to MCMC techniques, especially as the number of parameters increases. Marjoram *et al.* [76] tried to address this with an approximate MCMC algorithm, and in Chapter 5, I give a method that allows exact MCMC and ABC to be combined for problems in which some computation is possible.

The remainder of the chapter contains the details of how to apply these methods to the problem outlined in Chapter 2.

3.2 Inference of a Single Divergence Time

We are now able to perform the first inferences for the primate evolution model discussed in Chapter 2. In this chapter I focus only on estimating the primate crown-divergence time. This problem been tackled previously in Tavaré *et al.* [118] using the same model, but a different non-Bayesian inference technique based on minimising the sum of square differences between observed and expected data sets. This problem was also addressed in Plagnol *et al.* [94] using a similar approach to ours, but with a couple of important differences. Since these publications, the data set has changed due to the discovery of new fossils, and the reclassification of some others, and so for this reason, and to introduce a couple of changes and developments from [94, 118], I give another analysis of the data here. This analysis will also be important in order to have an understanding of the developments that follow. The data set used is given below in Table 3.1 and is obtained from Table 1.1 by finding the row totals. Note that the strepsirrhine-haplorhine stem-divergence time is the same as the primate crown-divergence time, as the strepsirrhini and haplorhini are the only two extant suborders of the primates.

Before implementing the inference algorithms discussed in the previous section, we must

Table 3.1: *The strepsirrhine and haplorhine stem fossil counts*

Epoch	T_k (My)	No. of Primate Fossils
Extant	0	376
Late-Pleistocene	0.15	22
Middle-Pleistocene	0.9	28
Early-Pleistocene	1.8	30
Late-Pliocene	3.6	43
Early-Pliocene	5.3	12
Late-Miocene	11.2	38
Middle-Miocene	16.4	46
Early-Miocene	23.8	34
Late-Oligocene	28.5	3
Early-Oligocene	33.7	22
Late-Eocene	37.0	30
Middle-Eocene	49.0	119
Early-Eocene	54.8	65
Pre-Eocene		0

decide on a set of prior distributions to use. This is done in the next section.

3.2.1 Prior Distributions

There is a large and diverse literature concerning the selection of prior distributions. Argument rages about whether one should take an objective Bayesian viewpoint, as promoted in Berger [17], or a subjective viewpoint, as argued in Goldstein [53]. An objective Bayesian approach involves choosing prior distributions according to some rule or convention (see Kass *et al.* [65]) in the hope that this apparent objectivity will silence the critics of the Bayesian paradigm. The subjective approach is to say that the very notion of randomness is inherently subjective, as is the scientific method, and that it only makes sense to have personal prior distributions. While this is the viewpoint to which I subscribe, it needs to be born in mind that scientists often turn to statistics for an objective validation of their work. It should also be remembered that in the majority of cases the choice of model has a far greater effect on the outcome of any analysis than the choice of prior distribution.

Our approach here is a compromise between the two schools of thought. Due to the complex nature of our problem, the calculation and simulation of reference priors, such as a Jeffreys's prior [61] or a Bernardo's reference prior [18], is infeasible or impossible. However, flat prior distributions are usually seen as more objective than other distributions and so we shall use uniform prior distributions on all of the parameters. The ranges are

decided upon after taking the advice of our scientific collaborators, although we shall vary them throughout the thesis, depending on the model we are examining.

It is important when choosing prior distributions to relate theory to practice. In other words, when choosing the prior distributions for the logistic growth parameters we keep in mind that ρ represents the speed at which diversity reaches its maximum, and $2/\gamma$ determines the expected long-term population size, etc. In later chapters, special attention will be given to the sampling fractions $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{14})$, but for this chapter and the next, I follow Tavaré *et al.* and make the parsimonious assumption that the α_k have fixed ratios. We write

$$\alpha_k = \alpha p_k, \quad \text{for } k = 1, 2, \ldots, 14,$$

where $\boldsymbol{p} = (p_1, \ldots, p_{14})$ is a constant vector, and $\alpha \in \mathbb{R}^+$ is allowed to vary. The main reason for making this assumption is that the ABC algorithm runs slowly when there are many parameters. I shall use the same fixed ratio as used in [118], with the values reproduced in Table 3.2. This approach is essentially a subjective Bayesian approach; the authors are saying that this is their considered opinion, here are the conclusions it leads to. Others are free to try their own values as they please. In Chapter 5 I question the validity of this assumption and introduce new methodology which allows us to replace it with independent priors for each α_k.

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14
p_k	1.0	1.0	1.0	1.0	0.5	0.5	1.0	0.5	0.1	0.5	1.0	1.0	1.0	0.1

Table 3.2: *Sampling ratios p_k, taken from [118].*

Initially, I use the following set of prior distributions:

$$\begin{array}{lcll} \tau & \sim & U[0, 100] & \text{(temporal gap)} \\ \alpha & \sim & U[0, 0.3] & \text{(sampling fraction)} \\ \gamma & \sim & U[0.005, 0.015] & \text{(growth parameter)} \\ \rho & \sim & U[0, 0.5] & \text{(growth parameter)} \\ 1/\lambda & \sim & U[2, 3]. & \text{(mean lifetime)} \end{array} \tag{3.2}$$

3.2.2 Choice of metric

One crucial aspect of approximate Bayesian computation is the choice of ρ in algorithm D. As already mentioned, there is no theory to help guide our choice, or to help assess the quality of any given choice. I use several different metrics throughout this thesis, with the choice depending on what aspect of the data I wish to concentrate on. However, the metric used most of the time, suggested in [94], is what I will refer to as the *standard*

metric and label ρ_s. Let D_+ represent the total number of fossils found,

$$D_+ = D_1 + \cdots + D_{14}$$

and let $\boldsymbol{Q}$ be a vector of proportions

$$\boldsymbol{Q} = (Q_1, \ldots, Q_{14}) := \left(\frac{D_1}{D_+}, \ldots, \frac{D_{14}}{D_+} \right).$$

Measure the distance between the real data $\mathcal{D}$, and a simulated data set $\mathcal{D}'$, by

$$\rho_s(\mathcal{D}, \mathcal{D}') = \left| \frac{D'_+}{D_+} - 1 \right| + \frac{1}{2} \sum_{j=1}^{14} |Q_j - Q'_j|. \tag{3.3}$$

The first term on the right measures the difference in the total number of fossils found in the simulated and real data sets, while the second term is the total variation distance between the two vectors of proportions.

Another metric, which is perhaps more intuitive, is the *sum of squares*, or *Euclidean* metric

$$\rho_e(\mathcal{D}, \mathcal{D}') = \sum_{i=1}^{14} (D_i - D'_i)^2. \tag{3.4}$$

Figure 3.3 shows scatter plots of ρ_e and ρ_s values for different simulated data sets. Plot A shows that from a distance both metrics give similar rankings to the data sets. However, when we zoom in and look more closely in Plot B, we can see that the metrics do rank data sets differently and so using different metrics will lead to different results. In Section 4.5 I give details from a simulation study comparing the effects of the two metrics.

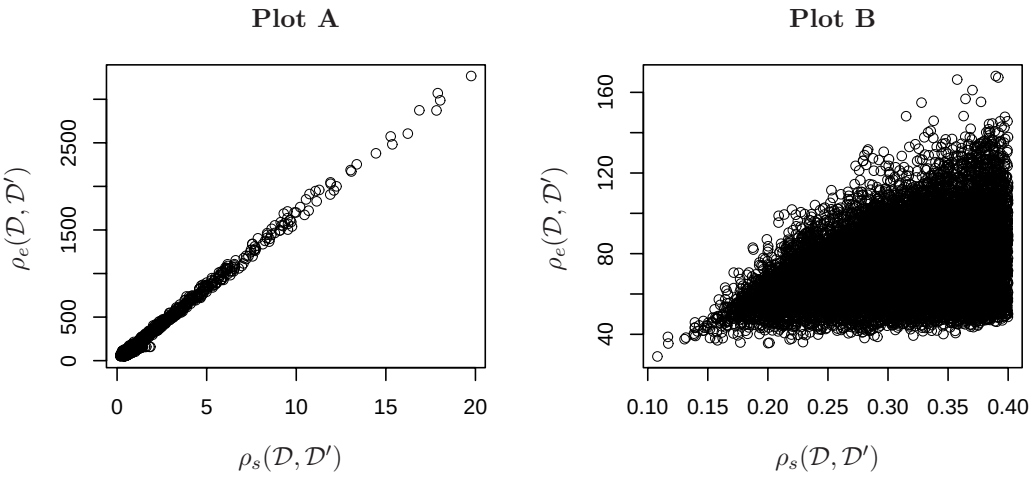

Figure 3.3: *Scatter plots of values of ρ_p and ρ_e on a variety of simulated data sets.*

3.2.3 Inference Algorithm

We are now in a position to write down our simulation approach. Firstly, I describe how to simulate continuous-time Galton-Watson trees as described in Chapter 2:

1. Begin with two species at time $t = -54.8 - \tau$ My[5] and let Z keep track of the number of extant species at time t. Initially $Z = 2$.

2. Simulate a realisation of $L \sim \text{Exp}(\lambda Z)$ (i.e., the time to the minimum of Z independent $\text{Exp}(\lambda)$ random variables) and set the new time to be $t = t + L$.

3. Randomly select one of the Z extant species and set its death time to be t. With probability $p_2(t)$ give the terminated species two offspring with birth time t and set $Z = Z + 1$. Otherwise set $Z = Z - 1$. Return to step 2.

Beginning the simulations with two species allows us to control the initial divergence time. In future I shall refer to the two sides of the tree, each evolving from one of the initial species. One of the sides represents the haplorhine species, and the other the strepsirrhine species. We require that both sides of the tree survive to time 0, as otherwise the initial divergence time does not represent the primate crown-divergence time, but would represent the stem-divergence time instead.

Recall that we wish to sample from the posterior distribution $\mathbb{P}(\mathcal{D}|\theta)$, and that $\mathcal{N}$ represents the number of species in each interval. Inference in this model is slightly more complicated than in Section 3.1 due to the hidden variable $\mathcal{N}$. By writing $\pi(\theta, \mathcal{N}) = \mathbb{P}(\mathcal{N}|\theta)\pi(\theta)$ for the joint distribution of θ and $\mathcal{N}$, we can see that the posterior distribution of $(\theta, \mathcal{N})$ is found via

$$\pi(\theta, \mathcal{N} \mid \mathcal{D}) \propto \mathbb{P}(\mathcal{D} \mid \theta, \mathcal{N})\pi(\theta, \mathcal{N}).$$

We can then use the algorithms stated previously to get draws from the posterior distribution $\pi(\theta, \mathcal{N}|\mathcal{D})$ as follows:

Algorithm E: Inferring Divergence Times

E1 Draw θ from $\pi(\theta)$.

E2 Simulate a tree and fossil finds using parameter θ. Count the number of simulated fossils in each interval, $\mathcal{D}' = (D'_1, \ldots, D'_{14})$.

E3 Accept parameter θ if $\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$. Return to step E1.

By ignoring accepted values of $\mathcal{N}$, we are left with values of θ drawn from an approximation to the marginal posterior, $\pi(\theta|\mathcal{D})$, as required. The results from this algorithm are presented in Section 3.2.4.

[5] Recall that the oldest primate fossil is at most 54.8 My old.

A note about the simulations

The simulations in this thesis are computationally intensive. The majority of the results are obtained by using a high-performance computing facility where it is possible to use many processors simultaneously. One of the advantages of ABC methods are that the algorithms are extremely easy to run in parallel with no issues about collating results from different nodes.

In later chapters, it will be important to know the complete structure of the tree, rather than just the population size at time t. Consequently, we shall need to keep track of the relationships between species. This can be done easily in the C programming language by storing the tree as a sequence of nested pointers. We declare `treenode` to be a structure which contains two floating point numbers representing that species birth and death time, and three pointers to other `treenode`s, one of which is its parent (allowing us to search backwards in time) and the other two are its offspring (allowing us to search forwards in time). If the species has no offspring the pointers are set to `NULL`. We also declare the structure to contain an array of integers 14 numbers long. Each number is set to 0 or 1 according to whether a fossil of that species is discovered in the corresponding interval. If a species is not extant during a given interval, obviously it cannot be preserved as a fossil in that interval.

Once we have a sample tree we can simulate fossil finds by drawing a Bernoulli(α_i) random variable for each species that lived in interval i to determine whether the species was found as a fossil in that interval. The nested tree structure used allows us to use recursive functions making operations on the tree computationally simple. A simple function, such as

```
gen_fos(struct treenode *tree)
 {
   generatefossils(tree) // generates fossils for that species only
   if (tree→left != NULL) gen_fos(tree→left)
   if (tree→right != NULL) gen_fos(tree→right)
}
```

will then generate fossils on the entire tree.

3.2.4 Results

Our initial simulations were done using the prior distributions given by Equation (3.2), the standard metric ρ_s, and $\epsilon = 0.1$. This value of ϵ was chosen by initially saving all of the output from the simulations, and then reducing the value of ϵ until the number of accepted results became too low. After approximately 50 hours of computer time, we were left with 7076 results, which is sufficient to get a clear picture of the marginal posterior distributions. If we use a smaller value ϵ we do not get a large enough sample.

For example, with $\epsilon = 0.05$ we are left with only 76 accepted results. By the *acceptance rate* I mean the reciprocal of the number of trees successfully simulated (i.e., at time 0 both sides of the tree have extant representatives) per accepted tree in step E3 of the inference algorithm. Lower acceptance rates will require that we simulate for longer. The acceptance rate for this simulation was 2560 successfully simulated trees per simulated tree. Note, however, that on average only 1 out of every 2.5 attempts at simulating a tree is successful (i.e., 60% of all simulations go extinct on one side or the other before time t). Table 3.3 contains a summary of the posterior distributions obtained, and Figure 3.4 contains plots of the marginal distributions. I have also included summaries of N_0 values, the present day diversity, for reasons that will become clear in later chapters.

Table 3.3: *A summary of the posterior distributions found when dating the primate divergence time. The results were obtained using ρ_s with $\epsilon = 0.1$ (n=7076). LQ denotes the lower quartile and UQ the upper quartile of the posterior distributions.*

	Min.	LQ	Median	Mean	UQ	Max.
N_0	50	112	167	215	262	2136
τ	3.9	16.0	23.0	25.8	32.4	93.4
α	0.010	0.073	0.118	0.130	0.178	0.300
ρ	0.020	0.294	0.373	0.359	0.440	0.500
γ	0.0050	0.0077	0.0101	0.0101	0.0126	0.0150
$1/\lambda$	2.00	2.34	2.59	2.57	2.81	3.00

The first thing to notice about these results is that the posterior distribution for τ is bounded away from the origin. Recall from the introduction that palaeontologists interpreting the fossil record have concluded that the primates did not coexist with the dinosaurs (the dinosaurs died out 65 My ago at the end of the Cretaceous era). These results suggest that conclusion can not be made using this data. The posterior probability of a Cretaceous origin for the primates is $\mathbb{P}(\tau > 10.2|\mathcal{D}) \approx 0.95$, where 10.2 is significant as it is the minimum temporal gap size required for a primate divergence time in the Cretaceous. This clearly shows that under this model, the primate fossil data is insufficient to constrain primate origins to the Cenozoic.

The marginal posterior distributions of the growth parameters ρ, γ and λ suggest that we should perhaps extend the range of their prior distributions. If we do this, however, we find that we must extend the ranges a large amount to find the complete range of values supported. For example, values of $1/\lambda$ in the range 0 to 30 all have positive posterior probability. The effect on the posterior of τ is to lose information, with the posterior much more uniformly spread across the interval $[0, 100]$. I do not extend the prior ranges of ρ, γ and λ for future calculations as these ranges represent our a priori beliefs, and can be defended on those grounds.

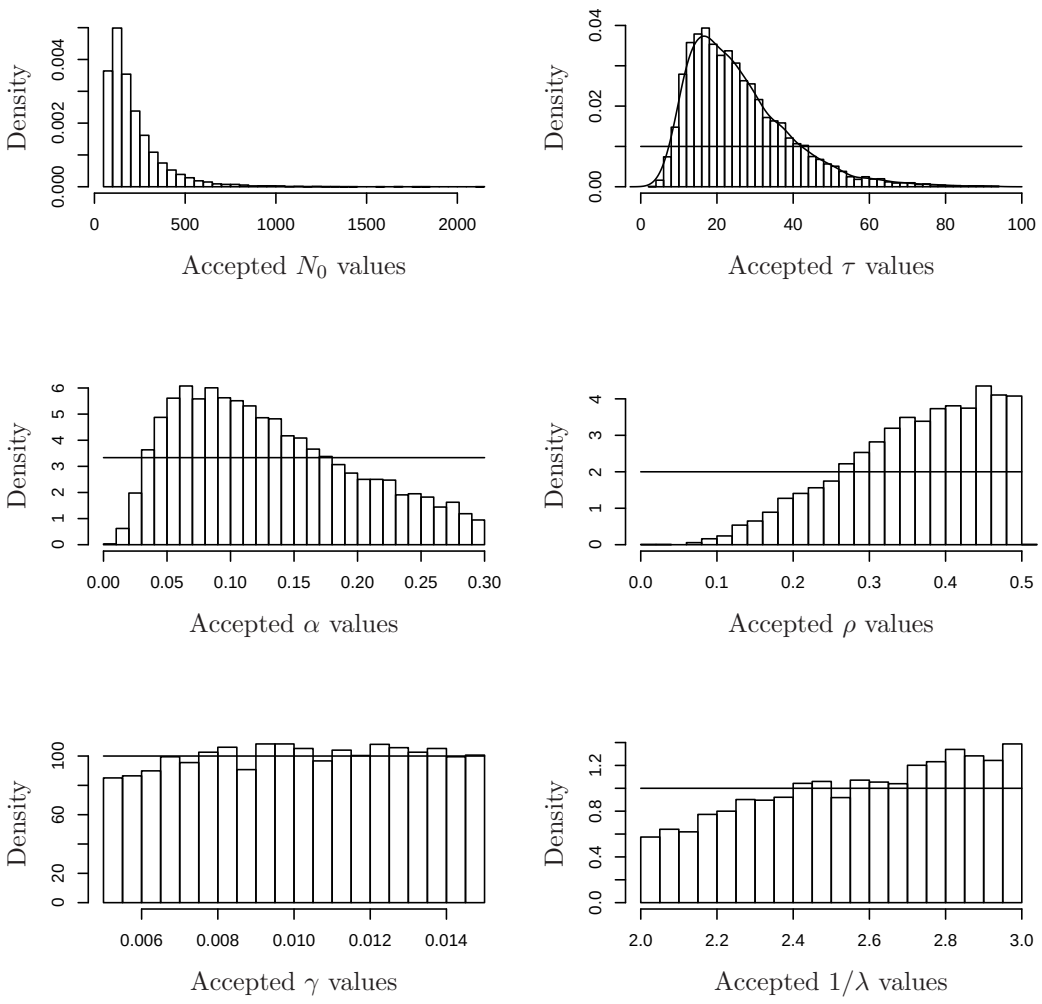

Figure 3.4: *Marginal posterior distributions obtained when dating the primate divergence time. The prior distributions are overlaid. The results were obtained using ρ_s with $\epsilon = 0.1$.*

The final point to note is that the majority of the mass in the distribution of N_0 is concentrated on values of 250 and less. Recall that we know that there are 376 extant primate species. In the next chapter I show how to take this data into account in our simulations.

3.2.5 Summary

In this chapter I have introduced Approximate Bayesian Computation and shown how it can be used to perform inference in the speciation model given in Chapter 2. The main conclusion is that the primate fossil record is not sufficient to constrain the primate divergence time to the Cenozoic, as claimed by many palaeontologists.

The analysis in this chapter has only made use of part of the information contained in

the data set. We ignored the fact that we know the number of extant primate species, and the fact that we have information about the internal structure of the tree. The anthropoids are a monophyletic clade which diverged from other primates at least 37 My ago, and we can use the anthropoid fossil counts to inform the structure of the simulated trees. In the next chapter, I exploit these data to get more information out of the limited data that is available.

Chapter 4

Using more of the Available Information

The inference technique described in the previous chapter is an important first step towards a more rigourous approach to inferring divergence times from the fossil record. Yet a quick comparison between Table 1.1 containing the complete data set, and Table 3.1 containing the data used for that analysis, shows that we have not used most of the rich data structure given in Table 1.1. Every known primate species has been carefully classified into one of two suborders, the haplorhini and the strepsirrhini. The strepsirrhine species have wet noses and the suborder contains the lemur and loris families. The haplorhini are the dry-nosed primates and contains the tarsier (long feet) and anthropoid (the higher primates) families. The anthropoids are in turn divided into two parvorders: the platyrrhini and catarrhini. The platyrrhini are the new world (American) monkeys and tend to be small, arboreal (live in trees) and nocturnal. The catarrhini are distinguished from the platyrrhini by their downwards pointing noses, and includes the old world monkeys and the ape family. They are generally diurnal (active during the day) and are consequently better understood than the platyrrhini.

The paucity of the primate fossil record, and the limited opportunites available for its improvement, make it vital that we extract the maximum amount of information possible from the data we have. Our aim in this chapter is to set out an approach that will take into account some of the structure in the data so that we are able to simultaneously estimate

two different divergence times. In the next section I give a method that will allow the joint distribution of the primate and anthropoid divergence times to be inferred. By inferring the joint distribution we hope to obtain a more accurate picture than would be possible by doing two independent inferences.

The modern diversity (number of extant species) of each taxonomic grouping is also known and can be used as an additional piece of data. The values shown in Table 1.1 come from Groves [56] and are continually being updated. In 2002 the known modern diversity was 235 primate species [118], but by 2005 Groves gives 376 species. Since this publication there have been several new discoveries. These include the Bemaraha Woolly Lemur (Avahi cleesei), named after lemur enthusiast and actor John Cleese, and the GoldenPalace.com monkey (Callicebus aureipalati), whose name was sold for $650,000 at auction. In Section 4.2 I include this information about the diversity in the analysis and explain its importance to our inferences. In Section 4.3 I allow the haplorhini and strepsirrhini, the two different sides of the evolutionary tree, to evolve independently. In Section 4.4 I introduce the idea of using local-linear regression on the output, before giving a short simulation study to indicate the accuracy of our methods.

4.1 Dating Two Divergence Times

Our aim in this section is to develop a methodology that takes into account the internal tree structure and allows us to infer the joint distribution of two divergence times.

Table 4.1 shows the fossil counts for two different taxonomic groupings. The first column of integers gives the number of crown-primate species discovered, $\mathcal{D} = (D_1, \ldots, D_{14})$, say. The second column gives the number of crown-anthropoid species, $\mathcal{A} = (A_1, \ldots, A_{14})$. Notice that $D_k \geq A_k$ for all k, as the antropoids are a subset of the primates. No primate fossils predating the Eocene epoch have been found, with the oldest primate fossil being at most 54.8 My old. No anthropoid fossils have been found before the Late-Eocene epoch, with the oldest anthropoid fossil being at most 37 My old.

As in Chapter 3, let τ denote the temporal gap between the oldest primate fossil and the last common ancestor (LCA) of the extant primates. Analogously, let τ^* denote the temporal gap between the oldest anthropoid (platyrrhine-catarrhine) fossil and the LCA of the anthropoids. See Figure 4.1 for clarification.

We could choose to date the anthropoid divergence time using the single-split approach given in the previous chapter. Using just the anthropoid fossil counts in Table 4.1, the prior distributions (3.2), and the sampling ratios given in Table 3.2 (but with $p_{12} = p_{13} = 0.1$), we find the posterior distributions for τ^* and α shown in Figure 4.2. The range extension observed for τ^* is small, with a median value of 2.6 My, and lower and upper quartiles of 1.1 My and 5.6 My respectively. This is much shorter than the values of τ observed in the

Table 4.1: *Strepsirrhini/haplorhini (primate-crown) and platyrrhini/catarrhini (anthropoid crown) stem fossil counts*

Epoch	T_k (My)	Primate Crown Fossil Counts, $\mathcal{D}$	Anthropoid Crown Fossil Counts, $\mathcal{A}$
Extant		376	281
Late-Pleistocene	0.15	22	22
Middle-Pleistocene	0.9	28	28
Early-Pleistocene	1.8	30	30
Late-Pliocene	3.6	43	40
Early-Pliocene	5.3	12	11
Late-Miocene	11.2	38	34
Middle-Miocene	16.4	46	43
Early-Miocene	23.8	34	28
Late-Oligocene	28.5	3	2
Early-Oligocene	33.7	22	6
Late-Eocene	37.0	30	2
Middle-Eocene	49.0	119	0
Early-Eocene	54.8	65	0
Pre-Eocene		0	0

previous chapter.

The negative aspect of this approach is that it involves eliminating a large part of our data set. We made no use of the primate fossil counts $\mathcal{D}$, and consequently ignored the unusual structure shown by the data. Note that in the most recent epochs, nearly all fossil finds have been of anthropoid species. We must go back at least 33 My before we find a significant number of non-anthropoid fossils, giving the data an unusual structure. The anthropoid subtree that originates some time before the start of the Late-Eocene soon grows to dominate the tree in terms of the number of fossil species found. We know that there are at least 95 species of primates which are not anthropoids. However, for some reason these species were not preserved as fossils.

The unusual tree structure may be of some importance, and taking it into account may help to give a clearer picture of evolutionary events. It is only by doing joint inference of τ and τ^* that we will see the significance, or otherwise, of this structure. Our approach is based on finding the subtree with fossil counts that most closely match the anthropoid fossil counts. We extend Algorithm E in Chapter 3, so that once we have a simulated tree with fossil counts close to $\mathcal{D}$, we perform an exhaustive search for the subtree that has fossil counts closest to $\mathcal{A}$. If the distance between simulated and real anthropoid fossil counts is less than our tolerance, ϵ_2, we measure the temporal gap from the base of this subtree to the start of the Late-Eocene interval. This is then our estimate for τ^*.

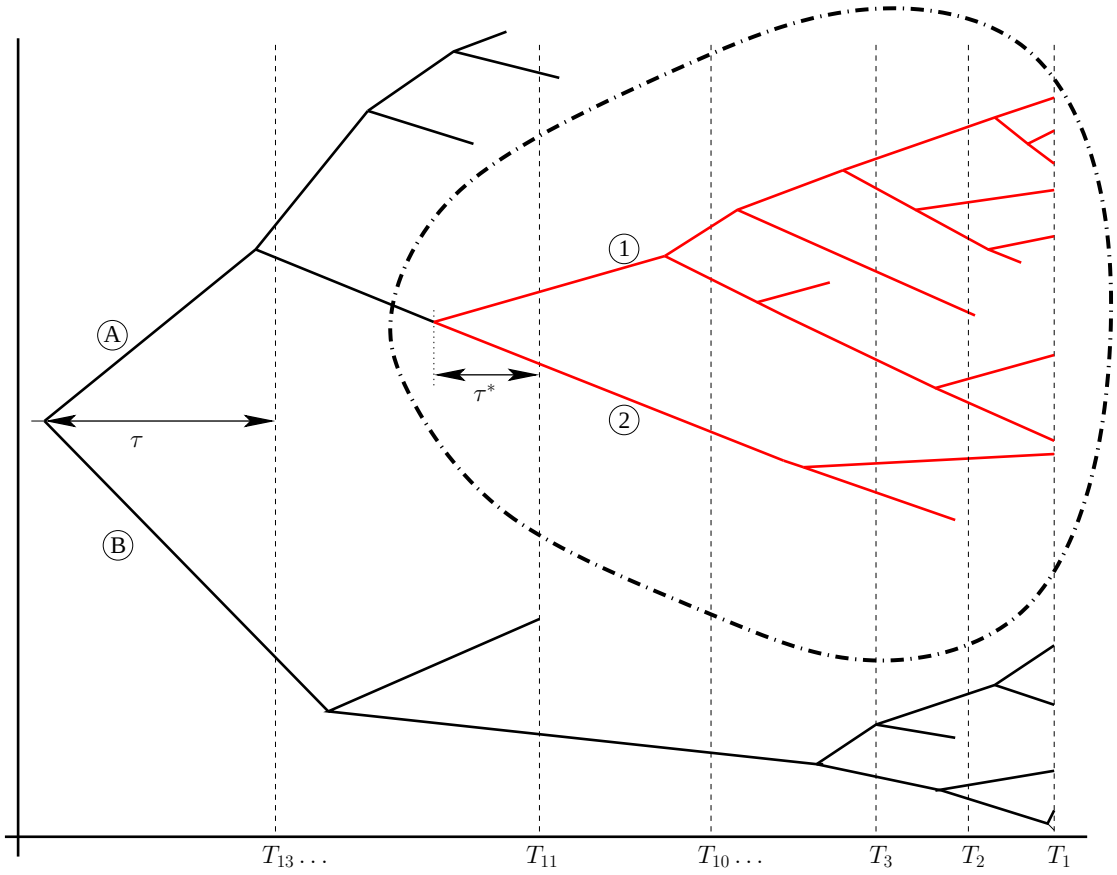

Figure 4.1: *A sample primate speciation tree with anthropoid-subtree highlighted. Here, tree A represents the haplorhini and tree B the strepsirrhini, while subtrees 1 and 2 represent the platyrrhine and catarrhine species. Parameter τ records the distance between the base of the Eocene and the base of the tree, whereas τ^* records the distance between the base of the Late-Eocene and the anthropoid subtree.*

There are a couple of technical points we must check. Firstly, we require that the root of the subtree (the death time of the LCA) occurs earlier than the beginning of the Late-Eocene epoch 37 My ago. Secondly, both branches leading from the root must have extant descendants. This is required because the subtree represents the anthropoids. One side of the subtree represents the platyrrhini and the other the catarrhini, and both of these groups have modern representatives. For the remainder of the thesis, I refer to this approach as *Optimal Subtree Selection (OSS)*.

Algorithm F: Optimal Subtree Selection (OSS)

F1 Draw parameters $\boldsymbol{\theta} = (\tau, \alpha, \gamma, \rho)$ from $\pi(\cdot)$.

F2 Simulate a tree and fossil finds using parameter $\boldsymbol{\theta}$. Count the number of simulated fossils in each interval, $\mathcal{D}' = (D'_1, \ldots, D'_{14})$.

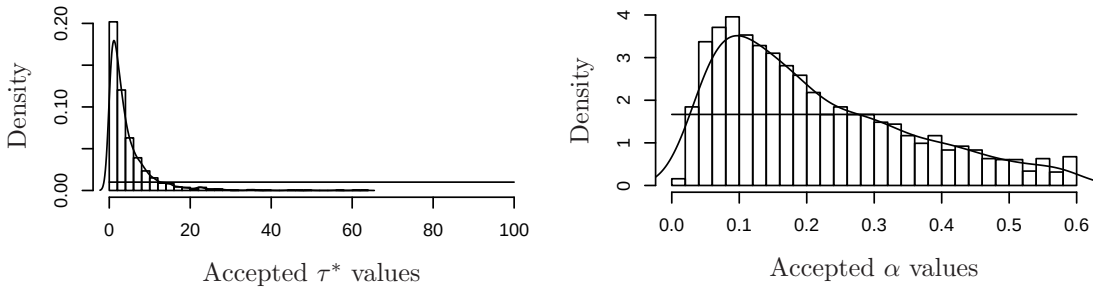

Figure 4.2: *Posterior distributions when using the approach of Chapter 3 on the anthropoid data $\mathcal{A}$. Results obtained using ρ_s with $\epsilon = 0.1$ ($n = 2224$).*

F3 Calculate the value of the metric $\rho(\mathcal{D}', \mathcal{D})$. If

$$\rho(\mathcal{D}', \mathcal{D}) \quad \begin{array}{ll} \leq \epsilon_1, & \text{go to step F4.} \\ > \epsilon_1, & \text{reject } \boldsymbol{\theta} \text{ and go to step F1.} \end{array}$$

F4 Perform an exhaustive search of all possible subtrees. For each subtree check that the root of the subtree occurs at least 37 My ago, and that both offspring of this root produce trees which survive to the present. If both conditions are met, count the number of fossils on the subtree, $\mathcal{A}' = (A_1, \ldots, A_{14})$.

F5 Determine which subtree has the smallest value of $\rho(\mathcal{A}, \mathcal{A}')$, i.e., has fossil counts closest to the anthropoid fossil set. This is the *optimal subtree*. If

$$\rho(\mathcal{A}, \mathcal{A}') \quad \begin{array}{ll} \leq \epsilon_2, & \text{accept } \boldsymbol{\theta} \text{ and measure } \tau^*. \\ > \epsilon_2, & \text{reject } \boldsymbol{\theta} \text{ and return to step F1.} \end{array}$$

4.1.1 OSS Results

We are now able to run the simulations to find the joint distribution of two divergence times. Using the standard metric for steps F3 and F4 in the optimal subtree search with a tolerance of $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2) = (0.2, 0.2)$, leads to the posterior distributions summarised in Table 4.2 and Figure 4.3[(1)]. Notice that Algorithm F uses a two-step metric. We could, however, rewrite this as a single metric to be applied after the optimal subtree has been found. However, splitting up the acceptance step by using the metric on two different occasions allows us to reject trees with primate fossil counts unlike our data before we go

[(1)] I do not interpret these results here, as there are still problems to be solved. In Chapter 5 I give a full description of the results and give three-dimensional plots of the joint posterior distributions of τ and τ^*.

through the expense of finding the optimal subtree. Note also that using a different metric for the subtree search will, in general, lead to a different subtree being found as optimal and hence a different value of τ^*. A larger value of ϵ than in Chapter 3 is necessary as we now have two filtering steps, which essentially squares the acceptance rate.

Table 4.2: *A summary of the posterior distributions from an optimal subtree search; obtained using ρ_s with $\boldsymbol{\epsilon} = (0.2, 0.2)$ (n=3210).*

	Min.	LQ	Median	Mean	UQ	Max.
N_0	43	87	103	112	127	369
τ	0.2	14.4	27.4	32.4	46.5	99.6
τ^*	0.0	12.4	17.8	20.2	25.3	90.5
α	0.060	0.193	0.235	0.227	0.271	0.300
ρ	0.004	0.297	0.386	0.361	0.446	0.500
γ	0.0050	0.0088	0.0114	0.0110	0.0132	0.0150
$1/\lambda$	2.00	2.23	2.45	2.37	2.71	3.00

Notice the contrast between this approach to inference for τ^*, which takes into consideration the structure, and the approach from Chapter 3 where the anthropoid data is considered in isolation (Figure 4.2). The new approach leads to a marginal posterior distribution for τ^* which supports much larger values. Also, the posterior is bounded away from the origin, unlike in Figure 4.2 where values are clustered around it.

The other point to note is that the distribution for N_0 (the modern diversity) is unrealistic; we know modern primate diversity is about 376, whereas our observations are centred around values of about 100. This is dealt with in the following section.

Finally, note that Algorithm E could easily be extended to do inference for three or more split points. The limiting factor, however, is computer power. It is not feasible to date three split points as the unusual data structure and associated poor model fit would cause the acceptance rate to become too small and we would be forced to use large ϵ values, which would lead to a low degree of accuracy.

4.2 Using the Modern Diversity

One of the more obvious problems with the analysis presented in the previous section is that observed values of N_0 are too small. Groves [56] gives the modern primate diversity to be 376 species, whereas our simulations are predicting average values of around 110 species. In this section we incorporate this knowledge into our data and calculate the posterior distribution of the parameters given the fossil counts $\mathcal{D}$ and the fact that modern diversity is 376.

Augmenting $\mathcal{D}$ and $\{N_0 = 376\}$ differs from our previous approach, which was to treat

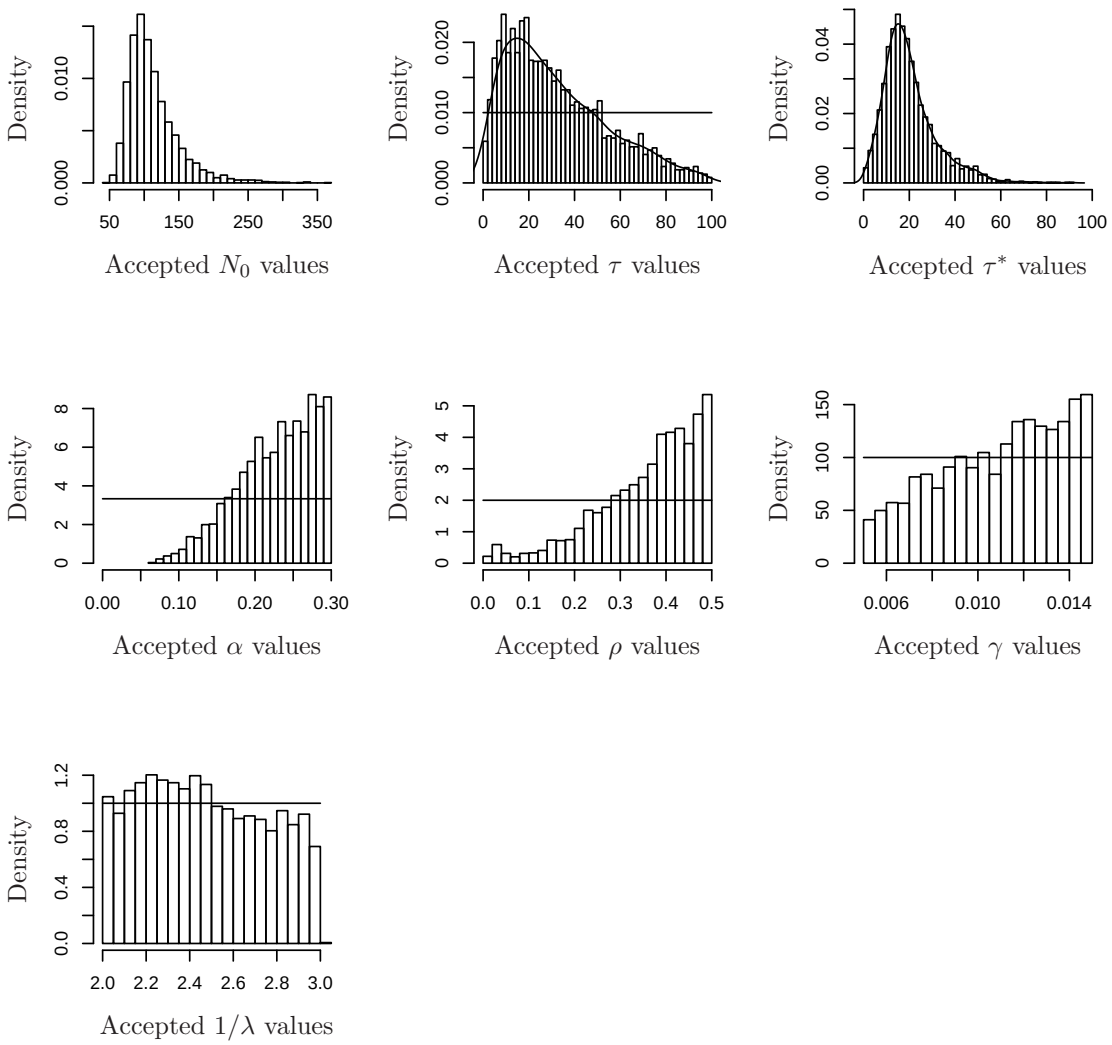

Figure 4.3: *Marginal posterior distributions obtained when dating the joint distribution of the primate and anthropoid divergence times using the optimal subtree selection approach. Obtained using ρ_s and $\boldsymbol{\epsilon} = (0.2, 0.2)$.*

the diversity as a priori knowledge. The prior ranges for ρ and γ were chosen so as to make N_0 take values of about 376 species. Instead, we aim to find the posterior distribution of the parameters given the fossil data and the observed number of modern species, i.e., $\pi(\cdot|\mathcal{D}, N_0 = 376)$. In an ideal world, this is easy to do. One would simulate trees conditioned to have 376 modern species, i.e., from

$$\mathbb{P}(Z_t = \cdot \mid Z_T = 376) \text{ for } t \leq T. \tag{4.1}$$

Unfortunately, although we can write down the probabilities in Equation (4.1), we do not know how to efficiently simulate trees with this size-process. A crude approach would be to use the rejection method: simulate unconditioned trees as before, discarding those that do

not have $N_0 = 376$. Although this does give observations from the required distribution, it is inefficient and unacceptably slow for our purposes.

An alternative and approximate approach is to accept trees with $N_0 \in S$, where S is a set of acceptable N_0 values, such as $\{n : n \geq 376\}$ or $\{n : |n - 376| \leq \delta\}$ for some choice of δ. Using $|N_0 - 376| \leq \delta$ is probably the better choice as the value of 376 has some uncertainty surrounding it. As noted in the introduction, the estimated diversity has increased from 235 to 376 in the past decade and seems set to change again as new species are discovered and others are identified as being the same.

A different approach to solving this problem lies in adapting the metric used in the ABC algorithm. We can sample (approximately) from the conditional distribution $\pi(\theta|\mathcal{D}, N_0 = 376)$, by changing the metric ρ so that trees with $N_0 \neq 376$ are penalised in proportion to how far they are from reality. A metric I found to work well is

$$\rho_p(\mathcal{D}, \mathcal{D}') = \sum_{i=1}^{k} \left| \frac{D_i}{D_+} - \frac{D'_i}{D'_+} \right| + \frac{1}{2} \left| \frac{D'_+}{D_+} - 1 \right| + \frac{1}{2} \left| \frac{N'_0}{N_0} - 1 \right| \tag{4.2}$$

where N'_0 is the simulated diversity. I will refer to this as the *population-adjusted metric* or ρ_p, in contrast to the standard metric, $\rho_s(\cdot, \cdot)$, which has been used thus far. Note that

$$\rho_p = \rho_s + \frac{1}{2} \left| \frac{N'_0}{N_0} - 1 \right|$$

so that $\rho_p \geq \rho_s$ for all data sets. Consequently, for the same value of ϵ we expect the acceptance rate to be lower when using the population-adjusted metric.

Changing the metric illustrates one of the advantages of the ABC approach. It was noted in Chapter 3 that when using ABC it is simple to combine different types of data. Data $\mathcal{D}$ and N_0 are dependent in some complicated manner. To do a traditional analysis we would need to calculate the likelihood $\mathbb{P}(\mathcal{D}, N_0 = 376|\boldsymbol{\theta})$, which, due to this dependence, would likely be a non-trivial calculation, even if we did know $\mathbb{P}(\mathcal{D}|\theta)$. However, with the ABC methodology we can simply add a term to our metric, and the posterior distribution is automatically altered.

Using ρ_p with $\boldsymbol{\epsilon} = (0.4, 0.4)$ and prior distributions (3.2), gives the results shown in Table 4.3 and Figure 4.4. The acceptance rate was approximately 25900 simulated trees per accepted tree, leaving 8931 results.

Notice that this has solved the problem of observing unrealistic values for the modern diversity. The distribution of N_0 is now nicely clustered around the true value of 376. Notice also that a large value of ϵ was needed in order to make the simulations run, and that the acceptance rate is low even using these large ϵ values. This is potentially due to the fact that $\rho_p \geq \rho_s$, but also because the model is still not producing N_0 values that are close to the data. Finally, note that the posterior distribution for α is now much more

Table 4.3: *A summary of the posterior distributions found when taking the modern primate diversity into account by using the population-adjusted metric. Obtained using ρ_p with $\boldsymbol{\epsilon} = (0.4, 0.4)$ (n=8931)*

	Min.	LQ	Median	Mean	UQ	Max.
N_0	208	377	408	415	449	597
τ	0.0	6.8	16.0	23.6	33.8	100.0
τ^*	2.0	18.6	25.5	32.6	39.6	115.6
α	0.022	0.049	0.063	0.066	0.080	0.140
ρ	0.0123	0.118	0.205	0.226	0.329	0.500
γ	0.0050	0.0072	0.0095	0.0096	0.0119	0.0150
$1/\lambda$	2.00	2.27	2.53	2.52	2.76	3.00

.

closely defined than in previous analyses. A rough calculation for α, based on the number of fossils in the Late-Pleistocene (most recent) epoch, suggests a value of $\alpha = \frac{19}{376} = 0.051$, which fits with the value obtained here.

The posterior distributions of γ, ρ and $1/\lambda$ contain very little information about these parameters, with the prior and posterior distributions being almost identical. Rather than giving these parameters prior distributions we could estimate them and use the point estimates in the simulations. Estimating these parameters by their posterior mode and repeating the simulations of this section, we find the results shown in Table 4.4 and Figure 4.5. This increases the strength of the information about the parameters of interest, τ and τ^*, but at the expense of taking account of less of the uncertainty. By fixing the values of the growth parameters we are no longer taking into account the uncertainty surrounding the growth rate of the primate diversity, assuming instead that it took place at a specified rate.

Table 4.4: *A summary of the posterior distributions found when taking the modern primate diversity into account by using the population-adjusted metric. Obtained using ρ_p with $\boldsymbol{\epsilon} = (0.2, 0.2)$ (n=2721) and fixing the growth parameters to be equal to their posterior mode; $\rho = 0.12, \gamma = 0.0095, 1/\lambda = 2.53$.*

	Min.	LQ	Median	Mean	UQ	Max.
N_0	289	358	377	378	397	469
τ	0.0	2.5	4.8	5.4	7.6	23.4
τ^*	2.5	14.7	17.7	17.8	20.9	36.2
α	0.036	0.056	0.061	0.062	0.067	0.090

.

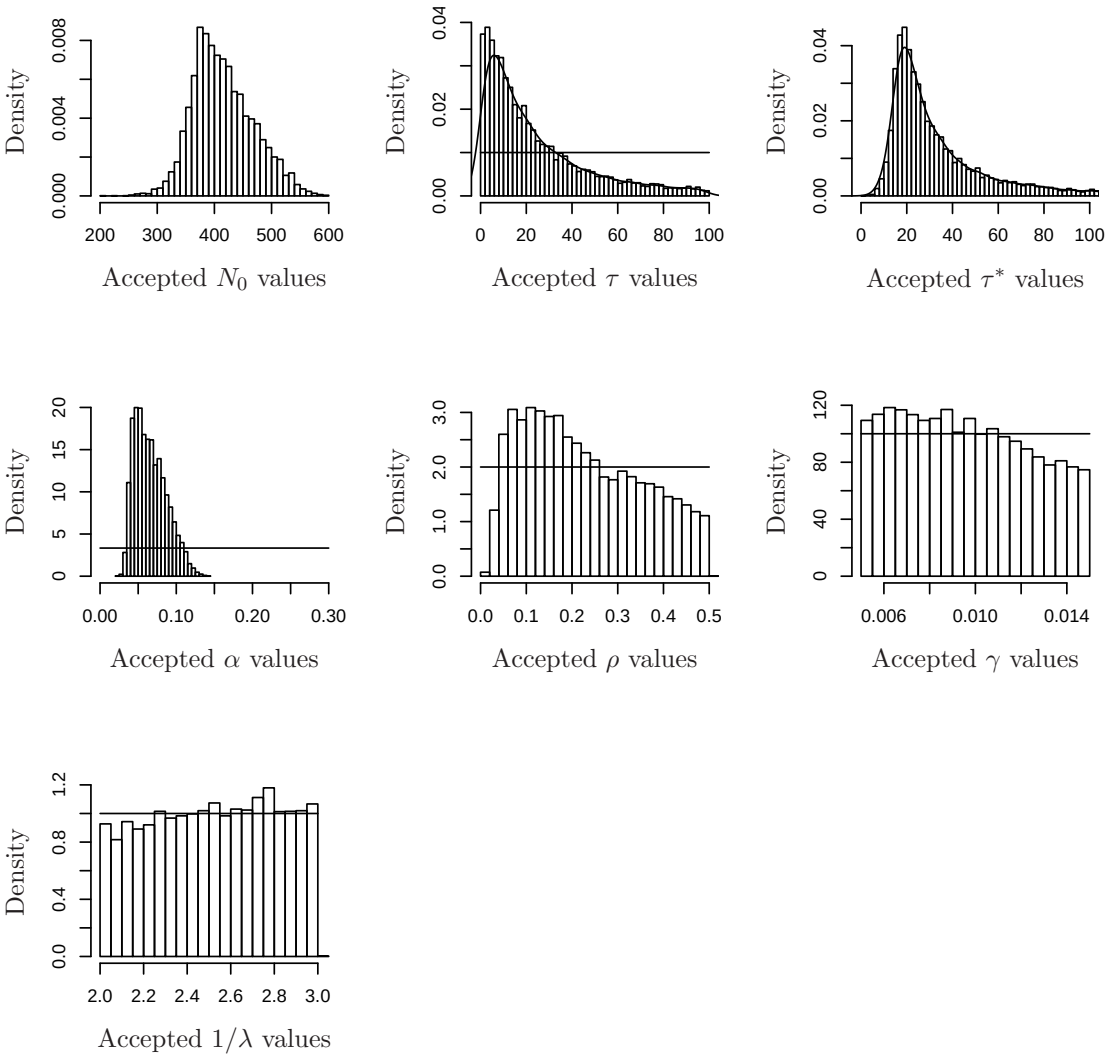

Figure 4.4: *Marginal posterior distributions obtained when dating the joint distribution of the primate and anthropoid divergence times and conditioning on* $N_0 = 376$. *Obtained using* ρ_p *and* $\boldsymbol{\epsilon} = (0.4, 0.4)$.

4.3 Modelling with Two Trees

In Section 4.1 some of the data structure was taken into account when we considered methods for finding the joint distribution of two split points. We noted then that the data has an unusual structure due to the quick rise to dominance by the anthropoids. Here we break the data down in a different way and consider the haplorhine and strepsirrhine suborders separately. Table 4.5 gives data broken down by suborder.

Notice that there are only a few fossil representatives of the strepsirrhini in recent epochs, despite a relatively large modern diversity. This suggests that the strepsirrhine and haplorhine suborders are subject to different sampling rates (and possibly different growth parameters too). We test this idea by allowing ρ, γ and $\boldsymbol{\alpha}$ to take different values

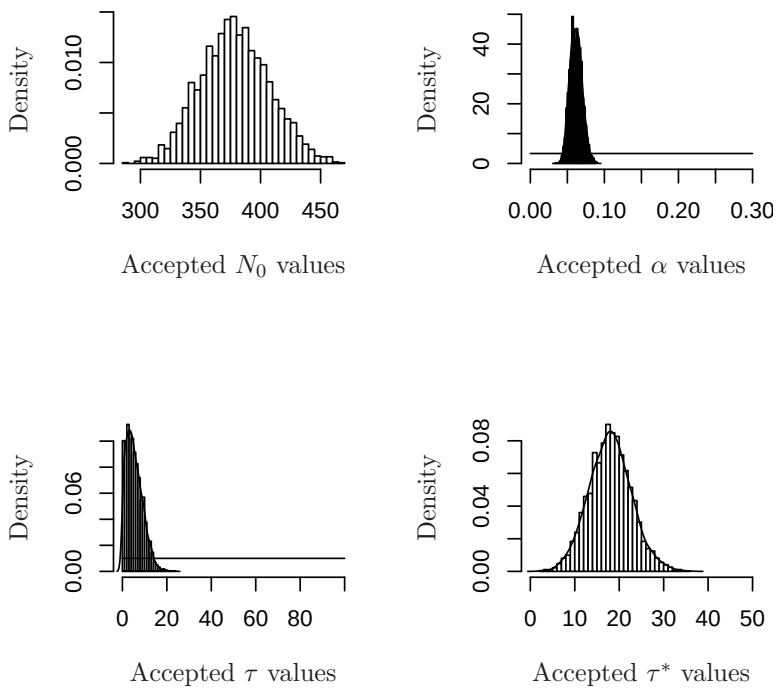

Figure 4.5: *Marginal posterior distributions obtained when dating the joint distribution of the primate and anthropoid divergence times and conditioning on* $N_0 = 376$. *Obtained using* ρ_p *and* $\boldsymbol{\epsilon} = (0.2, 0.2)$ *and fixing the growth parameters to be equal to their posterior mode;* $\rho = 0.12, \gamma = 0.0095, 1/\lambda = 2.53$.

on the different sides of the tree. Let α^S be the sampling fraction for the strepsirrhini and α^h the sampling fraction for the haplorhini, with a similar convention for ρ^H, γ^H, etc. We then simulate each side of the tree separately and decide whether to accept the parameters according to whether the data they generate satisfies $\rho(\mathcal{H}, \mathcal{H}') \leq \epsilon_1$ and $\rho(\mathcal{S}, \mathcal{S}') \leq \epsilon_2$, where $\mathcal{H}'$ are the simulated haplorhine fossil counts and $\mathcal{S}'$ the simulated strepsirrhine counts.

When this is done with the standard metric, we get accepted values of N_0^S, modern strepsirrhine diversity, that have a mean value of 9 species, whereas the average value of N_0^H is 105 species. Both of these values fall a long way short of the observed diversity. Using the population-adjusted metric solves this problem, and with $\boldsymbol{\epsilon} = (0.4, 0.4)$ we find the posterior distributions shown in Figure 4.6 and Table 4.6. The acceptance rate was low with an average of 310090 simulated trees per accepted tree.

Notice that the marginal posterior distributions for the two sampling fractions, α^H and α^S, are different. The distribution of the strepsirrhine sampling fraction takes values that are smaller than those taken by the haplorhine sampling fraction. The growth parameters ρ and γ do not show much difference between the two sides, however.

We shall have to wait until after Chapter 5 to perform this analysis properly, as at

Table 4.5: *Primate fossil counts, sorted by suborder.*

Epoch	T_k	Number of strepsirrhine fossils, $\mathcal{S}$	Number of haplorhine fossils , $\mathcal{H}$
Extant	0	88	288
Late-Pleistocene	0.15	0	22
Middle-Pleistocene	0.9	0	28
Early-Pleistocene	1.8	0	30
Late-Pliocene	3.6	3	40
Early-Pliocene	5.3	1	11
Late-Miocene	11.2	4	34
Middle-Miocene	16.4	2	44
Early-Miocene	23.8	6	28
Late-Oligocene	28.5	0	3
Early-Oligocene	33.7	4	18
Late-Eocene	37.0	14	16
Middle-Eocene	49.0	49	70
Early-Eocene	54.8	26	39
Pre-Eocene		0	0

present both sides of the tree use the same fixed sampling proportions $\boldsymbol{p}$, as given in Table 3.2. The next chapter gives a method of removing this assumption, and contains the results from performing a similar analysis to the above, but without the assumption that $\alpha_i = \alpha p_i$.

Finally, it is possible to model each side of the tree separately and to use the optimal subtree selection approach at the same time. Currently, however, the acceptance rate is too low and we do not have the computer power required to successfully perform such a simulation.

4.4 Regression Analysis

One of several extensions to the basic ABC algorithm was given by Beaumont, Zhang and Balding in 2002 [11]. Their idea is that rather than using a hard cut-off, whereby we reject or accept θ according to whether $\rho(\mathcal{D}, \mathcal{D}')$ is greater or less than some tolerance ϵ, we use a soft cut-off and weight each observation according to its accuracy. The simulated data sets that are closest to $\mathcal{D}$ will have the largest weights and will be given a greater importance when doing inference than those which are far from $\mathcal{D}$. They then recommend fitting a local-linear regression of simulated parameter values on the data and using this to predict the true parameter values by substituting the true data back into the regression equation.

Table 4.6: *A summary of the posterior distributions found when dating the primate divergence time by modelling the haplorhine and strepsirrhine suborders separately. Obtained using ρ_p with $\boldsymbol{\epsilon} = (0.4, 0.4)$ (n=1090).*

	Min.	LQ	Median	Mean	UQ	Max.
N_0^H	111	205	251	252	291	419
N_0^S	53	81	87	86	91	118
τ	1.9	23.6	37.0	41.7	57.6	99.0
α^H	0.027	0.052	0.066	0.072	0.087	0.176
α^S	0.018	0.032	0.038	0.038	0.044	0.073
ρ^H	0.022	0.182	0.284	0.281	0.388	0.500
ρ^S	0.018	0.249	0.341	0.331	0.423	0.500
γ^H	0.0050	0.0069	0.0092	0.0095	0.0119	0.0150
γ^S	0.0050	0.0073	0.0096	0.0098	0.0122	0.0150
$1/\lambda^H$	2.00	2.30	2.55	2.53	2.77	3.00
$1/\lambda^S$	2.00	2.22	2.46	2.48	2.71	3.00

Formally, they aim to fit a model of the form

$$\boldsymbol{\Theta} = \boldsymbol{DB} + \boldsymbol{e}$$

where for our case the response variables $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i)_{i\in\{1,\ldots,n\}}$ are the parameters, the regression variables $\boldsymbol{D} = (d_{ij})_{i\in\{1,\ldots,n\},j\in\{1,\ldots,p\}}$ are the simulated fossil counts[(2)], and $\boldsymbol{e}$ is a matrix of errors. It is assumed that $\mathbb{E}(\boldsymbol{e}) = \boldsymbol{0}$, and that the errors are uncorrelated but have different magnitudes, so that $\mathbb{C}\text{ov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$ for some diagonal matrix $\boldsymbol{V}$. The variance of each error is assigned according to the weight of each observation, so that if $\boldsymbol{W}$ is the diagonal matrix of weights, we set $\boldsymbol{V} = \boldsymbol{W}^{-1}$. Then observations with large weights (a small value of $\rho(\mathcal{D}, \mathcal{D}')$) are assumed to be more accurate and so the error term has a smaller variance. On the other hand, observations with small weights (large $\rho(\mathcal{D}, \mathcal{D}')$) are assumed to be less accurate and consequently $\mathbb{V}\text{ar}(e_i)$ is larger.

The solution is found by minimising the weighted sum of squares

$$\sum_{i=1}^{n} (\boldsymbol{\theta}_i - \boldsymbol{D}_{i.}\boldsymbol{B})^2 W_i$$

with respect to $\boldsymbol{B}$. Here $\boldsymbol{D}_i$ represents row i of matrix $\mathcal{D}$. The optimal choice for $\boldsymbol{B}$ is given by

$$\hat{\boldsymbol{B}}^* = (\boldsymbol{D}^t\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^t\boldsymbol{W}\boldsymbol{\Theta}$$

Substituting the real fossil counts gives parameter predictions $\hat{\boldsymbol{\theta}} = \mathcal{D}\hat{\boldsymbol{B}}$. To determine the

[(2)]The model can be given an intercept term by prepending a column of $(1, \ldots, 1)^t$ to the matrix $\boldsymbol{D}$.

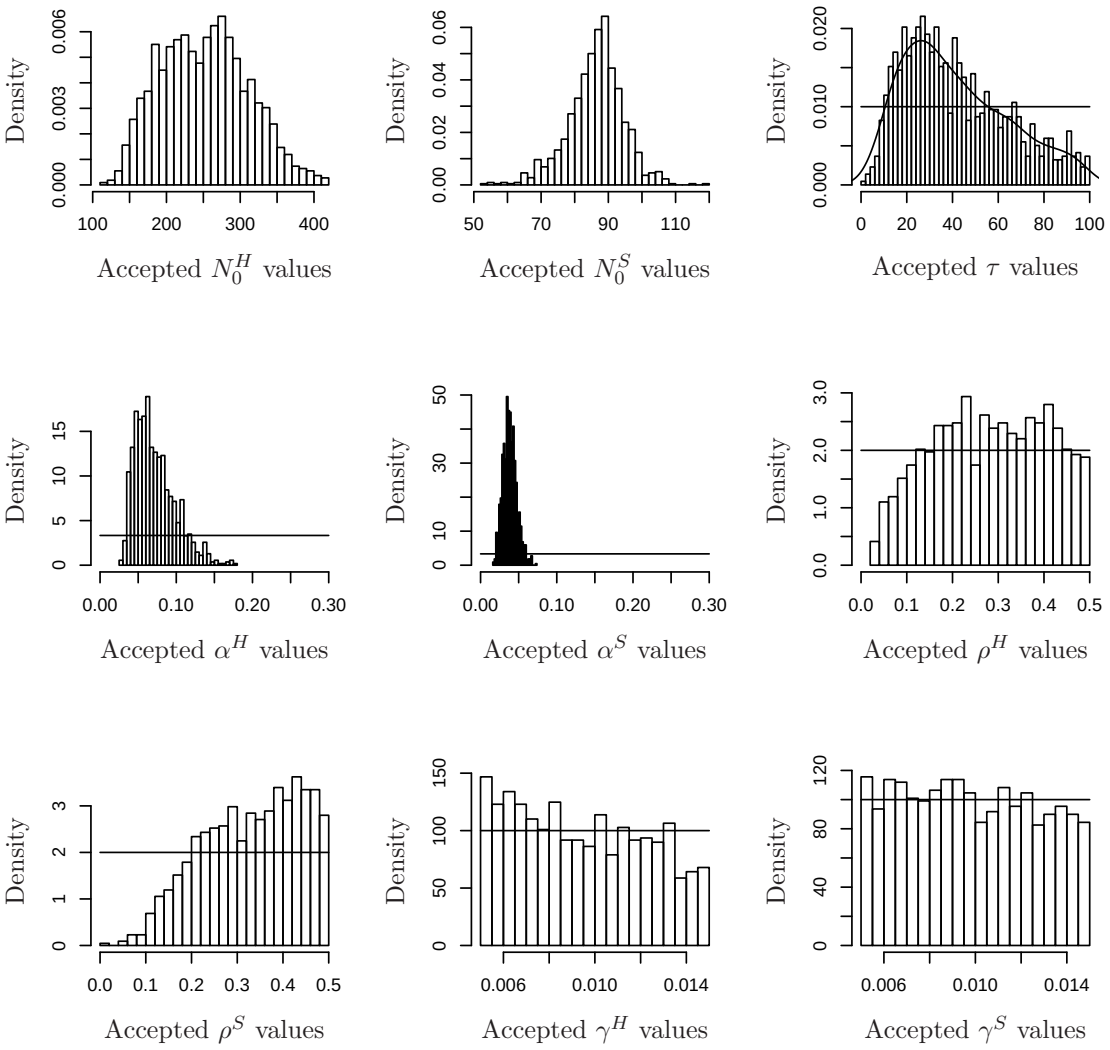

Figure 4.6: *Marginal posterior distributions obtained when dating the primate divergence time by modelling the haplorhine and strepsirrhine suborders separately. Obtained using ρ_p and $\boldsymbol{\epsilon} = (0.4, 0.4)$.*

weight of an observation that has metric value $t = \rho(\mathcal{D}, \mathcal{D}')$, Beaumont *et al.* recommend the use of the Epanechnikov kernel

$$K_\delta(t) = \begin{cases} c\delta^{-1}\left(1 - \left(\frac{t}{\delta}\right)^2\right), & t \leq \delta \\ 0, & t > \delta \end{cases}$$

Here, c is a normalising constant and δ is a tolerance value which determines the proportion of the results that are given non-zero weights. It is claimed that the method is insensitive to the choice of δ. To do inference for $\boldsymbol{\theta} = (\tau, \tau^*, \alpha)$, we must combine the two metric scores into a single value. Initially, I use $t = \rho(\mathcal{D}, \mathcal{D}') + \rho(\mathcal{A}, \mathcal{A}')$. I then use the following algorithm:

Algorithm G: Local-Linear Regression

G1 Generate the parameter values and data sets as before, but keeping all of the runs, regardless of the value of $\rho(\mathcal{D}, \mathcal{D}')$ or $\rho(\mathcal{A}, \mathcal{A}')$.

G2 Pick out the runs with $t = \rho(\mathcal{D}, \mathcal{D}') + \rho(\mathcal{A}, \mathcal{A}')$ in the bottom $\delta\%$ of all t values.

G3 Set the weights $W_i = 1 - \frac{t^2}{\delta^2}$.

G4 Set $\hat{\boldsymbol{B}} = (\boldsymbol{D}^t \boldsymbol{W} \boldsymbol{D})^{-1} \boldsymbol{D}^t \boldsymbol{W} \boldsymbol{\Theta}$, where each row of $\boldsymbol{D}$ contains one set of simulated primate fossil counts $\mathcal{D}'$ and anthropoid counts $\mathcal{A}'$, and each row of $\boldsymbol{\Theta}$ the corresponding $\boldsymbol{\theta}$ values.

G5 Substitute the real data into $\boldsymbol{\theta} = \mathcal{D}\hat{\boldsymbol{B}}$ to find the point estimate, or substitute each of the simulated data sets to find the adjusted posterior distributions.

Step G2 performs a similar role to steps 3 and 5 in Algorithm F, in that it outright rejects any tree which has fossil counts (primate or anthropoid) that are not close to the data. It is not feasible to use this algorithm, as it is written here, for the problem under consideration. This is because step G2 requires that we retain all of the output until all simulation has been completed. The memory requirement necessary to do this is is too high. An alternative is to fix a value of ϵ and replace step G1 by

G1′ Generate the parameter values and data sets as before, rejecting any runs which have $\rho(\mathcal{D}, \mathcal{D}') > \epsilon_1$ or $\rho(\mathcal{A}, \mathcal{A}') > \epsilon_2$.

We can then go on to apply the remainder of Algorithm G, however we will not require a small value of δ (such as that used by Beaumont *et al.* [11]) as we have already discarded a large proportion of the simulated data sets.

4.4.1 Results

We are now able to perform local-linear regression on the output given previously. I reanalysed the output from Section 4.2 using the modified version of Algorithm G. I set $\boldsymbol{\epsilon} = (0.5, 0.5)$ in the initial simulations and then used a tolerance of 50% ($\delta = 0.5$) in the Epanechnikov weighting kernel[(3)]. I am unable to use larger ϵ values as the file sizes quickly become too large for the statistical package R to cope with. Using a 50% tolerance means that 94119 results are taken into account in the analysis, although these will be treated with varying importance. The summary of the regression adjusted posteriors is shown in Table 4.7 and Figure 4.7.

[(3)]The use of a rejection step in G1′ has made step G2 superfluous to a certain extent. If the analysis is repeated with $\delta = 1$ the results obtained are almost identical.

	Min.	LQ	Median	Mean	UQ	Max.
τ	0.0	10.9	15.6	18.5	23.0	93.1
τ^*	0.0	13.7	17.4	19.1	22.6	83.8
α	0.042	0.074	0.080	0.081	0.087	0.167

Table 4.7: *Summary of the posterior distributions when using local-linear regression on the output of the ABC algorithm. Obtained using ρ_p with $\epsilon = (0.5, 0.5)$ and $\delta = 0.5$.*

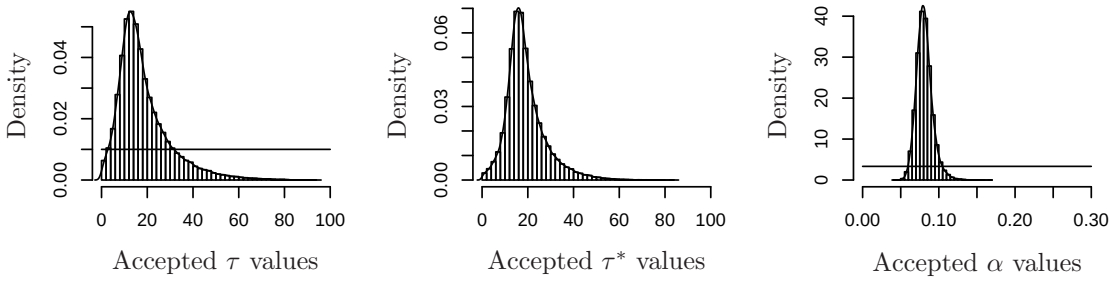

Figure 4.7: *Posterior distributions when using local-linear regression on the output of the ABC algorithm, with a tolerance of 50%.*

Comparing these results with those obtained before applying the regression technique (Figure 4.4 and Table 4.3) we can see some differences. The posterior distribution of τ is now bounded away from 0 and has a much shorter tail than seen previously. The median value of τ is about 16 My in both cases. The posterior distributions of τ^* and α show a larger difference between the two sets of results. The posterior median value of τ^* has decreased from 25.5 My when using standard ABC, to 17.4 My when applying local-linear regression. Its posterior is also more sharply defined now, with a much shorter tail. The posterior distribution of α is also more sharply defined than before, with the interquartile range now only 40% as large as that seen when using standard ABC.

At present, it is not known whether the regression-adjusted output is more or less accurate than the standard ABC output. I did a simulation study to test this, the results from which are shown in the next section.

4.5 Simulation Study

I now give a selection of results from a detailed simulation study which was performed to assess the accuracy of our inference mechanisms. These are then used to estimate the accuracy of the ABC approach. This is done by choosing parameter values and simulating a data set using these values, before using the inference algorithms above to find the posterior distributions. By examining the accuracy of the estimates we hope to show that our inference technique is performing sufficiently well.

The basic idea is as follows: we choose values of τ and α, and then generate a target tree and fossil counts using these parameters. A target subtree is then selected at random and we count the number of fossils and measure its temporal gap τ^*. We then use either Algorithm F or G to estimate the posterior distribution of these parameters in the usual manner. This is then repeated for many different target trees, recording each time the values of the parameters used to generate the tree, and our estimates of those parameters after a short simulation.

We choose target values of τ and α from their prior distributions[(4)]. A target value of τ^* is chosen by randomly picking a split point in the interval $[37, 54.8 + \tau]$, and then finding the first subtree on the path back to the root for which both offspring have extant descendants.

The accuracy of the algorithms is measured by finding the probability that the point estimate (usually the posterior mean or median) is within a factor p of the true value. We aim to find

$$\Phi_\tau(p) := \mathbb{P}\left(\frac{\tau}{p} \leq \hat{\tau} \leq p\tau\right),$$

where $\hat{\tau}$ is the estimate from the algorithm, and τ is the target (true) value. $\Phi_{\tau^*}(p)$ and $\Phi_\alpha(p)$ are defined similarly. An alternative way of measuring the error would be to consider the mean deviation of our estimate from the true value, or the mean integrated square error of our estimates. This would assess the accuracy of the posterior distribution found rather than just the accuracy of a point estimate. With hindsight this may have been the better choice. However, I believe that Φ is a more readily understood measure of the accuracy of the inferences. An approximate value for $\Phi_\cdot(\cdot)$ is found by looking at the proportion of target trees for which the ABC algorithm correctly places the parameter estimates in the required interval. Knowledge of Φ then allows the relative accuracy of different stochastic inference techniques to be compared.

Initially, all the simulation attempts are kept for each run regardless of the size of $\rho(\mathcal{D}, \mathcal{D}')$. This then allows the output files to be post processed and filtered with different levels of ϵ and the effects studied. Because of the computationally heavy burden of the inference algorithms, larger values of ϵ are necessary than might otherwise be used.

4.5.1 Findings

In total 247 different target trees were looked at. One of the problems faced was that there was a high degree of variability of the acceptance rates for different target trees. This made choosing a value of ϵ difficult as some target trees would produce lots of close approximations, whereas others would produce relatively few. So that our impression of

[(4)] Values for ρ, γ and $1/\lambda$ are also chosen from the prior distributions, although no results are shown for these nuisance parameters.

the accuracy was not distorted by runs where very few target trees are generated, we ignore any targets that have less than 1,000 approximation attempts remaining after filtering.

Table 4.8 shows how the accuracy $\Phi(p)$ varies with p varies when using Algorithms F and G. The results for Algorithm F were obtained by using the standard metric with $\boldsymbol{\epsilon} = (0.2, 0.2)$. The results for Algorithm G were obtained by selecting the 1000 results with the smallest $\rho(\mathcal{D}, \mathcal{D}') + \rho(\mathcal{A}, \mathcal{A}')$ values and then using $\delta = 1$.

Table 4.8: *The relative accuracy of Algorithms F and G. Accuracy of Algorithm F when using ρ_s with $\boldsymbol{\epsilon} = (0.2, 0.2)$, and Algorithm G applied to the 1000 results with the smallest $\rho_s(\mathcal{D}, \mathcal{D}') + \rho_s(\mathcal{A}, \mathcal{A}')$ values ($\delta = 1$).*

Algorithm F:

p	2.0	1.8	1.6	1.4	1.2
$\Phi_\tau(p)$	0.82	0.76	0.70	0.60	0.36
$\Phi_{\tau^*}(p)$	0.90	0.88	0.78	0.68	0.46
$\Phi_\beta(p)$	0.91	0.86	0.81	0.63	0.35

Algorithm G:

p	2.0	1.8	1.6	1.4	1.2
$\Phi_\tau(p)$	0.92	0.89	0.85	0.74	0.54
$\Phi_{\tau^*}(p)$	0.91	0.87	0.82	0.76	0.48
$\Phi_\beta(p)$	0.92	0.89	0.85	0.70	0.48

These results show that Algorithm G is more accurate than Algorithm F for this scenario. In other words, using local-linear regression on the output from Algorithm G does improve the accuracy, although not by a great amount. Different values of $\boldsymbol{\epsilon}$ and δ can be used, and these do have a small effect on the value of Φ. In general though, the local-linear regression appears to always improve the accuracy, but by varying degrees.

If we use the Euclidean metric given in the previous chapter (Equation (3.4)), both algorithms perform less well. Table 4.9 shows the results from applying Algorithm G to the 1000 results with the smallest values of $\rho_e(\mathcal{D}, \mathcal{D}') + \rho_e(\mathcal{A}, \mathcal{A}')$. I showed before that there is a strong correlation between the metrics, but this clearly shows that the standard metric ρ_s is better than the Euclidean metric at picking out the important aspects of the data. As before, these results have been confirmed in other settings when using different tolerance and $\boldsymbol{\epsilon}$ values.

Table 4.9: *Accuracy of Algorithm G when using the Euclidean metric ρ_e after choosing the 1000 results with the lowest $\rho_e(\mathcal{D}, \mathcal{D}') + \rho_e(\mathcal{A}, \mathcal{A}')$ values.*

p	2.0	1.8	1.6	1.4	1.2
$\Phi_\tau(p)$	0.87	0.85	0.74	0.58	0.41
$\Phi_{\tau^*}(p)$	0.87	0.76	0.65	0.53	0.34
$\Phi_\beta(p)$	0.72	0.66	0.61	0.46	0.25

One final point to note is the importance of the acceptance rate. For an unknown

reason, the acceptance rate varies widely between different target trees. For some targets we find that after 10 hours of simulation with $\epsilon = 0.1$ there are thousands of accepted simulations, whereas for other targets we find there are less than 100 accepted results. Table 4.10 shows the accuracy when we only look at target trees that gave rise to at least 5000 results after simulating for 10 hours. We can see that the accuracy of our predictions is better when looking at these selected targets. Plots of how Φ changes with p are shown in Figure 4.8.

p	2.0	1.8	1.6	1.4	1.2
$\Phi_\tau(p)$	1.00	1.00	1.00	0.90	0.64
$\Phi_{\tau^*}(p)$	0.92	0.92	0.92	0.82	0.56
$\Phi_\beta(\delta)$	0.98	0.98	0.90	0.84	0.60

Table 4.10: *Accuracy of Algorithm G when looking at only those target trees which lead to high acceptance rates. Filtering is done with $\epsilon_1 = 0.1$ and then only targets with at least 5000 results are considered. Local-linear regression is used on the output with $\delta = 0.2$.*

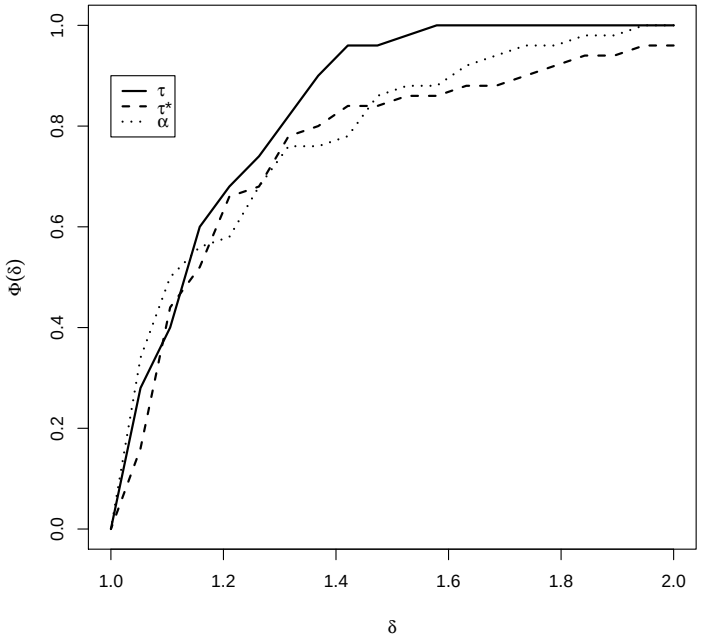

Figure 4.8: *Shows how $\Phi(p)$ changes with p for the data in Table 4.10.*

I also examined how best to weight each result. For the results given above, each simulation was weighted by the sum of the two metrics, $t = \rho_s(\mathcal{D}, \mathcal{D}') + \rho_s(\mathcal{A}, \mathcal{A}')$. However, we can use other weightings to try and emphasise different aspects of the data. I found that none of the other weightings I examined improved the accuracy beyond that found when using the simple weighting first suggested in Algorithm G.

4.5.2 Limitations

There are several limitations as to how realistic these assessments of the accuracy of our simulations really are. Firstly, the target trees we are trying to approximate are generated from our model. So, at best, this simulation study tells us how good our inference methods are, given that the model is the *true* model. Unfortunately, we do not have such confidence in our model.

Secondly, we know that the accuracy of these algorithms is dependent upon the value of ϵ that is used. For this simulation study it was not feasible to simulate each target tree for a length of time comparable to that used for the analysis of the real data in previous sections. On average in this study, each target tree was given six hours on a single node. This is less than 5% of the computing time that was devoted to the analysis of the real data set in Section 4.1. Consequently, the results given above are only really of use in suggesting the relative accuracy of different algorithms.

Finally, the aim of inference is to find the posterior distribution of the parameters. Here, I looked at the accuracy of a point estimate rather than the accuracy of the distributions found. It may be that if we looked at the tails of the distributions (using a measure such as mean integrated square error) we would find a different level of accuracy. Also, even if we were accurately inferring the posterior distribution, we would not expect a point estimate to always be within a factor of p of the true value of the parameter.

In conclusion, the results here should not be taken as a measure of the absolute accuracy of our algorithms, but more as a guide when choosing between methods. The results may also serve to put our mind at ease that the ABC algorithms are not failing in an obvious manner.

Chapter 5

Combining Approximate Bayesian Computation and Markov Chain Monte Carlo

In Chapter 3 I noted that Approximate Bayesian Computation is based upon the rejection algorithm and as a consequence can be inefficient; a large amount of time is spent simulating data with parameter values that lie in the tails of the posterior distribution. The idea behind Markov Chain Monte Carlo methods (MCMC) is that by correlating observations, more time is spent in regions of higher likelihood. The tails of the distribution are still visited, but less time is spent there.

Unlike the approximate Bayesian computation methods studied so far, MCMC methods are exact: if done properly (and this is a potentially big if), they produce draws from the true posterior distribution. Approximate Bayesian computation does not give samples from the posterior distribution, but from some approximation to it.

Motivated by these two thoughts, I use some of the ideas behind MCMC to produce a more efficient and more accurate approximate inference technique that can be used in problems where the likelihood function is unknown.

This will allow the introduction of more parameters into the species divergence time problem than was previously possible. Up until this point of the thesis, I have been assuming that the sampling fractions are held in a fixed ratio to each other, so that

$\alpha_i = \alpha p_i$. The new methodology presented here will allow this assumption to be removed.

In the next section I give a brief overview of previous methods, before going on to give a new development. In Section 5.2 I give details of how to apply these ideas to the species divergence problem, removing the need for the simplifying assumption made about the sampling fractions in Chapter 3. Results and discussion follow in Sections 5.3 and 5.4.

5.1 A Hybrid ABC-MCMC Sampler

In this section I give a brief introduction to Markov Chain Monte Carlo methods. Brooks [22] gives an excellent review of MCMC methods while Tierney [119] gives proofs of the main theoretical results.

The two most popular Markov Chain Monte Carlo (MCMC) algorithms are known as the Metropolis-Hastings sampler and the Gibbs sampler. The latter is a special case of the former, but the distinction is commonly made and serves our purpose here. Although MCMC methods are not inherently Bayesian, MCMC samplers have long been used in Bayesian applications. As this is how they will be used here, I consider the problem of sampling from the posterior distribution $\pi(\theta|\mathcal{D})$. The *Metropolis-Hastings* algorithm for generating dependent observations from $\pi(\theta|\mathcal{D})$ is as follows:

Algorithm H: Metropolis-Hastings

H1 If currently at θ, propose a move to θ' according to transition density $q(\theta, \theta')$.

H2 Accept the move to θ' with *Metropolis-Hastings* acceptance probability

$$h(\theta, \theta') = \min\left(1, \frac{\pi(\theta')\mathbb{P}(\mathcal{D}|\theta')q(\theta', \theta)}{\pi(\theta)\mathbb{P}(\mathcal{D}|\theta)q(\theta, \theta')}\right). \tag{5.1}$$

If rejected, stay at θ. Go to step H1.

Other forms of acceptance probability can be used, but Peskun [91] showed that this form is optimal in the sense that it rejects suitable candidate moves least often. This maximises the statistical efficiency of any Monte Carlo estimate made with output from the algorithm. Subject to the condition that the resulting chain is irreducible and aperiodic, any choice of transition density $q(\cdot, \cdot)$ can be used. If $q(\theta, \theta') \equiv q(\theta')$, so that candidate observations are drawn independently of the current state, then the sampler is known as an *independence sampler*; if $q(\theta, \theta') \equiv q(\theta' - \theta)$, then the sampler is known as a *random-walk sampler*. Another common choice is the Gibbs sampler, which will be explained later. If the parameter θ is multidimensional, then we can partition it into components, and update each component in turn.

Note that knowledge of the normalising constant, $\mathbb{P}(\mathcal{D})$, is not required in the Metropolis-Hastings acceptance probability, but that knowledge of $\mathbb{P}(\mathcal{D}|\theta)$ is necessary. For the problems considered in this thesis, the likelihood function, $\mathbb{P}(\mathcal{D}|\theta)$, is not usually computable, and so using MCMC is not an option. Marjoram *et al.* [76] suggested the following likelihood-free MCMC algorithm:

Algorithm I: Likelihood-Free Metropolis-Hastings

I1 If currently at θ, propose a move to θ' from transition density $q(\theta, \theta')$.

I2 Simulate $\mathcal{D}'$ from the model $\mathbb{P}(\cdot|\theta')$.

I3 If $\mathcal{D} = \mathcal{D}'$, calculate

$$h(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right).$$

Else return to step I1.

I4 Accept the move to θ' with probability $h(\theta, \theta')$. If rejected, stay at θ and return to step I1.

This algorithm is exact, with the difference between algorithms H and I analgous to the difference between algorithms A and B in Chapter 3. No knowledge of the likelihood function is required, but in problems of any complexity we will rarely, if ever, observe simulated data that exactly matches the real data. The approximate version of this algorithm, proposed by Marjoram *et al.*, replaces step I3 in the algorithm above with

I3′ If $\rho(\mathcal{D}, \mathcal{D}') < \epsilon$ calculate $h(\theta, \theta')$. Else return to step I1.

Poor choices for transition kernel q can make the efficiency of the algorithm worse than that of the rejection-based ABC algorithms. In practice, this chain often mixes extremely slowly; each time $\rho(\mathcal{D}, \mathcal{D}') \geq \epsilon$ the chain stays in its current location. Because consecutive results are correlated, very long runs are often required to achieve sufficient output to cover most of the parameter space.

The *Gibbs sampler* (see [27, 50]) is a special type of MCMC algorithm. The parameter is split into a number of components, and at each iteration one of these components is updated by conditioning on the values of the other components. In other words, the new parameter block is sampled from its posterior conditional density given all the other components and the data. An advantage of this type of update is that the Metropolis-Hastings probability, h, is always equal to 1. Consequently, every move that is proposed is accepted.

More formally, if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, then block i is updated from density $\pi(\theta_i \mid \boldsymbol{\theta}_{[-i]}, \mathcal{D})$ where $\boldsymbol{\theta}_{[-i]} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$. If this is done for each $i = 1, \ldots, k$ in turn, then this is known as the *systematic-scan* Gibbs algorithm. If coordinate i is chosen at random

from $\{1, \ldots, k\}$ then it is known as the *random-scan* Gibbs sampler (see [73]). Blocks θ_i can be multidimensional.

Combining Kernels

A Markov Chain Monte Carlo algorithm does not have to always use the same update strategy. In practice, it is often convenient to combine a number of different update strategies. For example, we could randomly choose to use independence samplers, Gibbs samplers or any other Metropolis-Hastings step in any order we wish, as long as some basic conditions are satisfied. Propositions 3 and 4 from Tierney [119] show that under weak conditions on the updates, a hybrid algorithm of mixtures or cycles of these kernels can be combined. A common hybrid algorithm is the Metropolis-within-Gibbs sampler, in which we use Gibbs update steps whenever the full posterior conditionals are known and use Metropolis-Hastings acceptance steps when the conditional distribution is not available.

Here I combine Gibbs sampling steps with rejection or ABC-rejection steps. For ease of exposition, suppose the parameter is split into just two components, $\boldsymbol{\theta} = (\theta_1, \theta_2)$, and that $\pi(\theta_1|\mathcal{D}, \theta_2)$ is known and that $\pi(\theta_2|\mathcal{D}, \theta_1)$ is unknown. Then the following algorithm samples from the posterior distribution:

Algorithm J: Approximate-Gibbs Sampler

J1 If currently at $\boldsymbol{\theta} = (\theta_1, \theta_2)$, draw θ_1' from $\pi(\theta_1|\mathcal{D}, \theta_2)$ and set $\boldsymbol{\theta} = (\theta_1', \theta_2)$.

J2 Draw θ_2' from $\pi(\theta_2)$ and simulate data $\mathcal{D}'$ using parameter $\boldsymbol{\theta} = (\theta_1', \theta_2')$.

J3 If $\mathcal{D} = \mathcal{D}'$, set $\boldsymbol{\theta} = (\theta_1', \theta_2')$ and return to step J1. Otherwise stay at $\boldsymbol{\theta} = (\theta_1', \theta_2)$ and return to step J2.

Steps J2 and J3 above are the mechanical version of the rejection algorithm which gives samples from $\pi(\theta_2|\mathcal{D}, \theta_1)$. By replacing step J3 with

J3′ If $\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$, set $\boldsymbol{\theta} = (\theta_1', \theta_2')$ and return to step J1. Otherwise stay at $\boldsymbol{\theta} = (\theta_1', \theta_2)$ and return to step J2.

we can generate approximate draws from $\pi(\theta_2|\mathcal{D}, \theta_1)$. Alternatively, we could use a likelihood-free MCMC step as follows:

Algorithm K: Approximate Metropolis-within-Gibbs Sampler

K1 If currently at $\boldsymbol{\theta} = (\theta_1, \theta_2)$, draw θ_1 from $\pi(\theta_1|\mathcal{D}, \theta_2)$ and set $\boldsymbol{\theta} = (\theta_1', \theta_2)$.

K2 Draw θ_2' from $q(\theta_2, \theta_2')$ and simulate data $\mathcal{D}'$ using parameter $\boldsymbol{\theta} = (\theta_1', \theta_2')$.

K3 If $\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$, calculate

$$h(\theta_2, \theta_2') = \min\left(1, \frac{q(\theta_2' \to \theta_2)\pi(\theta_2')}{q(\theta_2 \to \theta_2')\pi(\theta_2)}\right).$$

Otherwise return to step K2.

K4 Accept the move to θ_2' with probability h and set $\boldsymbol{\theta} = (\theta_1', \theta_2')$. Otherwise stay at $\boldsymbol{\theta} = (\theta_1', \theta_2)$. Return to step K1.

These algorithms can easily be extended to deal with cases where the parameter is divided into more than two components. Equally, the first step of algorithms J and K can be replaced with any other MCMC update step that preserves detailed balance.

A rule of thumb?

Liu [73] (p27) suggests the following rule of thumb for Monte Carlo computation:

> One should carry out analytical computation as much as possible.

The algorithms above are based on an analogous idea. For the type of intractable problems considered in this thesis (i.e., where we can not compute the likelihood) the following rule of thumb might apply:

> One should carry out exact Monte Carlo inference as much as possible.

5.2 Independent Sampling Fractions

The sampling fraction for interval i, α_i, is defined to be the proportion of species that lived in interval i that are preserved as fossils. In the preceding chapters I followed Tavaré *et al.* [118] and set the sampling fractions so that the ratio between any two values was fixed, that is, I set $\alpha_i = \alpha p_i$ where $\boldsymbol{p} = (p_1, \ldots, p_k)$ is a fixed vector. The values used are given in Table 3.2.

Tavaré *et al.* took the subjective Bayesian approach that these values are the opinions of the authors and other people are free to insert their own values as they please. They did not try to justify their choice, although they did partially test the robustness of their conclusions to this choice by trying one other value for $\boldsymbol{p}$. Fixing $\boldsymbol{p}$ in this way is one of the main assumptions of their work that is open to criticism by others [113].

The Bayesian solution to this problem is to give the sampling fraction for each interval a prior distribution, and then find the posterior distributions. One way to do this is to give the same prior distribution to each parameter α_i, where the prior distribution depends on hyperparameter $\boldsymbol{\psi}$. This leads to a hierarchical model, with a structure as shown in plate notation in Figure 5.1.

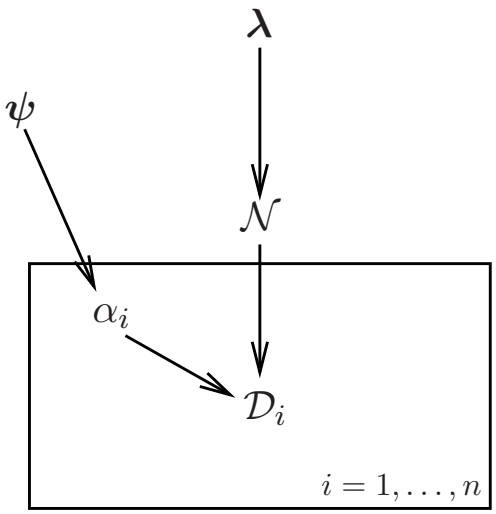

Figure 5.1: *A plate diagram for our model. Here, the plate or box indicates replication of the enclosed parameters. The number in the bottom right corner gives the number of replicates.*

In theory, we are now able to proceed with a rejection-based ABC algorithm similar to algorithms E and F to do inference for all of the parameters. In practice, however, there are problems. The advantage of fixing the ratios of the α_i is that we can take into account the structure of the data. Although none were given in the paper, it is possible to imagine reasons why Tavaré *et al.* may have decided to set the p_i as they did. For example, if the number of rock formations accessible from the Late-Oligocene and Pre-Eocene epochs (intervals 9 and 14) was fewer than those available from other intervals, this would justify the decision to set p_9 and p_{14} to be small (intuitively, less accessible rock means less chance of discovering fossils [92, 93]). What is clearly not acceptable is to examine the data, find that there are only a few fossil finds from these two epochs and then decide to fix the p_i to be small.

Allowing the α_i to vary freely has added fourteen more dimensions to the parameter space. Because rejection-based ABC chooses parameters from the prior distribution, the increased number of dimensions causes the acceptance rate to plummet to low values as the majority of simulations will be with parameter values from regions of low likelihood. An intuitive analogy is of looking for a needle in a multidimensional haystack: the lower the number of dimensions, the more likely we are to find the needle. And so it is with rejection based ABC: the lower the number of parameters, the more likely we are to randomly pick parameter values that have high posterior probability.

5.2.1 The Algorithms

Our aim is to simulate from the posterior distribution of our parameters given the data, i.e., from

$$\pi(\boldsymbol{\lambda}, \tau, \mathcal{N}, \boldsymbol{\alpha}|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau, \mathcal{N})\mathbb{P}(\mathcal{N}|\tau, \boldsymbol{\lambda})\pi(\tau)\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\alpha})$$

where $\boldsymbol{\lambda} = (\lambda, \gamma, \rho)$ are the growth parameters, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{14})$ is a vector of sampling fractions, and $\mathcal{N} = (N_1, \ldots, N_{14})$ represents information about the underlying tree structure, with N_i denoting the number of species that lived at any point during interval i.

An advantage of including the tree structure, $\mathcal{N}$, in the equations above is that parts of the likelihood equation become computable. We cannot use a Gibbs sampler (or any other MCMC algorithm) as the distribution of $\mathcal{N}$ is unknown and thus the conditional distributions of τ and $\boldsymbol{\lambda}$ are unknown. However, we can use the approximate MCMC algorithms, J and K, proposed earlier.

We split the parameter $\boldsymbol{\theta} = (\tau, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \mathcal{N})$ into two parts, $\boldsymbol{\alpha}$ and $(\boldsymbol{\lambda}, \tau, \mathcal{N})$. We are able to write down the conditional distributions required for the Gibbs sampler for the fourteen $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{14})$ parameters:

$$\begin{aligned}
\pi(\boldsymbol{\alpha}|\mathcal{D}, \boldsymbol{\lambda}, \tau, \mathcal{N}) &\propto \pi(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau, \mathcal{N}|\mathcal{D}) \\
&\propto \mathbb{P}(\mathcal{D}|\tau, \boldsymbol{\lambda}, \mathcal{N}, \boldsymbol{\alpha})\mathbb{P}(\mathcal{N}|\tau, \boldsymbol{\lambda})\pi(\tau)\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\alpha}) \\
&\propto \pi(\boldsymbol{\alpha})\mathbb{P}(\mathcal{D}|\mathcal{N}, \boldsymbol{\alpha}) \\
&\propto \pi(\boldsymbol{\alpha}) \prod_{i=1}^{14} \alpha_i^{D_i}(1-\alpha_i)^{N_i - D_i}.
\end{aligned}$$

We can see that if we choose to make the prior distributions for the $\{\alpha_i\}$ independent, the posterior distributions will also be independent. Moreover, note that we can use conjugate beta distributions for the α_i. If α_i has a Beta(a, b) distribution then

$$\begin{aligned}
\pi(\boldsymbol{\alpha}|\mathcal{D}, \boldsymbol{\lambda}, \tau, \mathcal{N}) &\propto \prod_{i=1}^{14} \alpha_i^{D_i}(1-\alpha_i)^{N_i - D_i}\pi_\beta(\alpha_i; a, b) \\
&\propto \prod_{i=1}^{14} \alpha_i^{D_i}(1-\alpha_i)^{N_i - D_i}\alpha_i^{a-1}(1-\alpha_i)^{b-1} \\
&\propto \prod_{i=1}^{14} \alpha_i^{D_i + a - 1}(1-\alpha_i)^{N_i - D_i + b - 1} \\
&\propto \prod \pi_\beta(\alpha_i\,; D_i + a, N_i - D_i + b).
\end{aligned}$$

Here, $\pi_\beta(x\,; a, b)$ denotes the probability density function of a Beta(a, b) random variable. Note that using beta prior distributions has led to the posterior distributions also being

beta distributions. The posterior mean of α_i is then

$$\mathbb{E}(\alpha_i|\mathcal{D},\mathcal{N}) = \frac{D_i + a}{N_i + a + b} \tag{5.2}$$

If we choose a and b to be small compared with N_i and D_i, then Equation (5.2) is approximately equal to the ratio of the number of fossil species to the true number of species, which is the value we intuitively expect the sampling fractions to take:

$$\mathbb{E}(\alpha_i|\mathcal{D},\mathcal{N}) \approx \frac{D_i}{N_i}.$$

The other conditional distribution $\pi(\tau, \boldsymbol{\lambda}, \mathcal{N}|\mathcal{D}, \boldsymbol{\alpha})$ is not known, but we can simulate from it using a rejection-ABC algorithm as done in previous chapters.

Algorithm L: Inferring a Single Split Time

L1 Suppose we are at $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}^{(t)}, \tau^{(t)}, \mathcal{N}^{(t)})$ after t iterations.

L2 Propose $\boldsymbol{\alpha}'$ from

$$\pi(\boldsymbol{\alpha}|\mathcal{D}, \boldsymbol{\lambda}^{(t)}, \tau^{(t)}, \mathcal{N}^{(t)}) = \prod \text{Beta}(a + D_i, N_i^{(t)} - D_i + b) \tag{5.3}$$

L3 Propose $(\boldsymbol{\lambda}', \tau')$ from prior $\pi(\tau)\pi(\boldsymbol{\lambda})$ and simulate tree $\mathcal{N}'$ from $\mathbb{P}(\mathcal{N}'|\tau', \boldsymbol{\lambda}')$. If $N_i < D_i$ for any i repeat step L3.

L4 Simulate fossil counts $\mathcal{D}'$.

L5 Accept $(\boldsymbol{\lambda}', \tau', \mathcal{N}')$ if $\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$ and set $\boldsymbol{\theta}^{(t+1)} = (\boldsymbol{\lambda}', \boldsymbol{\alpha}', \tau', \mathcal{N}')$. If $\rho(\mathcal{D}, \mathcal{D}') > \epsilon$ stay at $(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}^{(t)}, \tau^{(t)}, \mathcal{N}^{(t)})$ and return to step L3.

To initialise the algorithm I choose $\boldsymbol{\lambda}, \tau$ and $\boldsymbol{\alpha}$ from their respective prior distributions. The parameters a and b used in the beta prior distributions are known as hyperparameters. Their value should not matter greatly as long as both are small, as they will be dominated by the data, as Equation (5.2) shows.

In order to infer the joint distribution of two split points we must again use the optimal subtree selection idea from Chapter 4 to find the subtree that most closely matches the anthropoid subtree counts.

Algorithm M: Inferring Two Divergence Times

M1 Suppose we are at $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}^{(t)}, \tau^{(t)}, \tau^{*(t)}, \mathcal{N}^{(t)})$ after t iterations.

M2 Propose $\boldsymbol{\alpha}'$ from

$$\pi(\boldsymbol{\alpha}|\mathcal{D}, \boldsymbol{\lambda}^{(t)}, \tau^{(t)}, \mathcal{N}^{(t)}) = \prod \text{Beta}(a + D_i, N_i^{(t)} - D_i + b).$$

M3 Propose $(\boldsymbol{\lambda}', \tau')$ from prior $\pi(\tau)\pi(\boldsymbol{\lambda})$ and simulate tree $\mathcal{N}'$ from $\mathbb{P}(\mathcal{N}'|\tau', \boldsymbol{\lambda}')$. If $N_i < D_i$ for any i repeat step M3.

M4 Simulate fossil counts $\mathcal{D}'$.

M5 Find the subtree with fossil finds $\mathcal{A}'$ that minimize the metric $\rho(\mathcal{A}, \mathcal{A}')$, and measure its temporal gap τ^*.

M6 Accept $(\boldsymbol{\lambda}', \tau', \tau^*, \mathcal{N}')$ if $\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon_1$ and $\rho(\mathcal{A}, \mathcal{A}') \leq \epsilon_2$, update the parameter so that $\boldsymbol{\theta}^{(t+1)} = (\boldsymbol{\lambda}', \boldsymbol{\alpha}', \tau', \tau^*, \mathcal{N}')$ and return to step M1. Otherwise stay at $(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}^{(t)}, \tau^{(t)}, \tau^{*(t)}, \mathcal{N}^{(t)})$ and return to step M3.

A note about the simulations

In order to make correct inferences we must ensure not only that we have enough output, but also that the Markov Chain has converged. We also have the problem that output from these algorithms is correlated. More details and some diagnostic tests are performed in Section 5.4.1. Each run starts from a different position and so different runs are independent of one another. In the following few chapters, I give the results from many different scenarios but so as not to become too repetitive I will not report on convergence details for every run, unless there is a problem.

A problem observed during testing is that we often need to simulate an unusually large number of trees until the first is accepted. Once the first successful observation has been made, the acceptance rate increases and we need only simulate a much smaller number of trees per successful tree. The reason for this is that each run starts with parameter values drawn from the prior distributions, and only once the first tree has been accepted does the sampler reach regions of high posterior probability. By using a warm-up period, we can greatly speed up the simulations. By this, I mean that we start with large ϵ values and decrease the value after each acceptance until we are at the level we desire. This greatly speeds up the simulations.

There are two main approaches to MCMC. The first is to run one long chain and thin the output (recommended by Raftery *et al.* [100]). The second is to run many shorter chains and take the final value from each run (recommended by Gelfand *et al.* [48]). I have taken a mixture of the two approaches and run several independent medium length chains and then thinned and combined the output from each one.

5.3 Dating the Primate Divergence Time

In this section, I repeat the analysis of the crown-primate data first done in Section 3.2. Instead of giving the sampling fractions a fixed ratio $\alpha_i = \alpha p_i$, I give each α_i an independent Uniform$[0, 1]$ prior distribution. Note that this is equivalent to letting $a =$

$b = 1$ in the beta prior distributions, as these then reduce to uniform priors. Hence, we can still take advantage of the conjugate nature of the problem. The other parameters will be given the same prior distributions as in Equation (3.2).

Using Algorithm L with the standard metric, the usual problem that observed N_0 values are too small is more severe than before. N_0 now has a posterior median value of 35 species with an upper quartile of 41. The reason the problem has become worse is due to the additional freedom we have introduced by adding 13 extra parameters. The first few N_0 values accepted on each run are of a similar size to those seen in previous sections ($N_0 \approx 200$). But after a short period N_0 tends to smaller values, while at the same time the $\alpha-$Markov chains move to larger values consistent with such small N_0.

The key to solving this problem lies, as before, in augmenting the data $\mathcal{D}$ with $\{N_0 = 376\}$. However, using the population-adjusted metric alone does not solve the problem as it did in the previous chapter. Initial runs using ρ_p ran extremely slowly with a very low acceptance rate. This is caused by the warm-up period: as the tolerance ϵ decreases the $\alpha-$Markov chains become stuck in regions of low posterior density with N_0 values that are too small. I found that the simplest solution is to build in a hard constraint on N_0, such as requiring that N_0 is larger than some fixed value. The results presented in Table 5.1 and Figure 5.2 are obtained using the constraint $N_0 \geq 200$, and with the population-adjusted metric ρ_p with $\epsilon = 0.15$. The acceptance rate was 13828 simulated trees per accepted tree, which after a burn-in period of 50 results and thinning by taking every second result, leaves 2452 results[(1)].

(i)

	Min.	LQ	Median	Mean	UQ	Max.
N_0	301	355	370	368	382	438
τ	0.1	7.4	14.0	19.6	24.4	99.7
ρ	0.013	0.074	0.140	0.182	0.271	0.497
γ	0.050	0.0072	0.0095	0.0096	0.0119	0.0150
$1/\lambda$	2.00	2.43	2.67	2.62	2.85	3.00

(ii)

k	α_k
1	0.059
2	0.061
3	0.063
4	0.073
5	0.022
6	0.037
7	0.054
8	0.037
9	0.0070
10	0.047
11	0.10
12	0.25
13	0.58
14	0.026

Table 5.1: *Summary statistics when dating the primate divergence time using the approximate-Gibbs sampler (algorithm L). Obtained using ρ_p, $N_0^{main} \geq 200$, and $\epsilon = 0.15$ (n=2542); (i) posterior summary statistics, (ii) posterior means of the sampling fractions.*

[(1)]Issues of convergence and burn-in are dealt with in Section 5.4.1

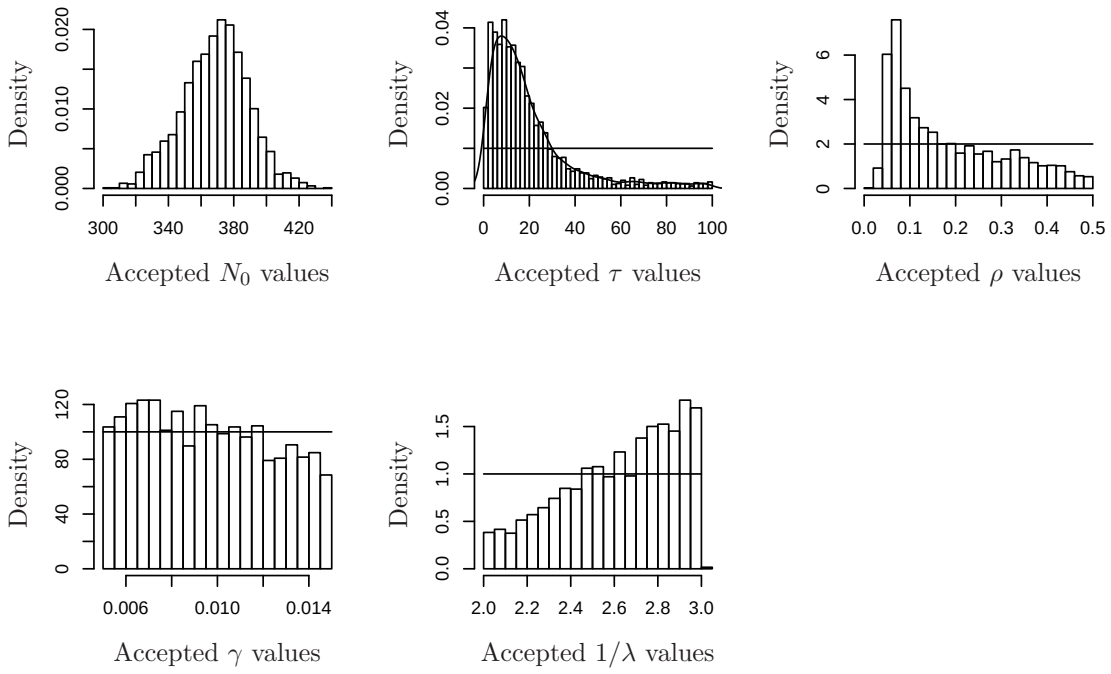

Figure 5.2: *Marginal posterior plots for the primate divergence time when using the approximate-Gibbs sampler. Obtained using* ρ_p, $N_0^{main} \geq 200$, *and* $\epsilon = 0.15$.

The hard constraint $N_0 \geq 200$ has no effect on the results once the chain has converged. In fact, once the warm-up period has finished we are able to remove this constraint without any effect on the accepted observations. If we run the algorithm using the standard metric and the constraint $N_0 > 376$ we observe almost identical marginal distributions for $\tau, \boldsymbol{\alpha}$ and ρ. However, the posterior distribution for N_0 then supports values that are unrealistically large, with an upper quartile of 954. Most experts would agree that there are not 900 primate species in the biota.

Comparing these results to those obtained when $\alpha_i = \alpha p_i$ (Figure 3.4) we can see that allowing α_i to vary freely leads to smaller estimates for the temporal gap τ. This is probably due to the fact that α_{14} is no longer constrained to be much smaller than the other α_i values (in Chapter 3 we set $p_{14} = 0.1$ whereas the average p value was 0.66). Setting $p_{14} = 0.1$ has the effect of allowing only a few fossils to be found before the Eocene, thus allowing τ to take larger values.

If we compare the sampling fractions we found here with the ratios given by the $\boldsymbol{p}$ values in Table 3.2 we can see there is a reasonable agreement. For intervals one to ten the ratios roughly match those assumed before. However, intervals 11 to 14 do not. Intervals 11, 12 and 13 have average sampling fractions that are much larger that the values for other intervals. $\bar{\alpha}_{13}$ is about 10 times as large as $\bar{\alpha}_1$, whereas previously we have modelled them

as taking the same value ($p_1 = p_{13} = 1$). Also α_{14}, although smaller than all the other sampling fractions, is larger than assumed previously.

Finally, although I concluded in the previous chapter that performing local-linear regression improves the accuracy, I do not use this technique in the remainder of the thesis as it is not obvious how to apply the method to the techniques of this chapter.

5.4 Dating the Primate and Anthropoid Divergence Times

We now move our attention to estimating the joint distribution of the primate and anthropoid crown divergence times using Algorithm M (cf. Sections 4.1.1 and 4.2). When using the standard metric, ρ_s, we see the usual problem of N_0 values that are too small (mean of 128) along with an extremely low acceptance rate. The low acceptance rate can be explained by looking at the posterior distribution of the growth parameter ρ, shown in Figure 5.3, which is clearly bimodal. If we look at those runs which have $\rho < 0.2$, then the posterior mean value of N_0 increases from 128 to 273 and the peak at $\rho \approx 0.1$ agrees with the results presented later, obtained using the population-adjusted metric. The reason

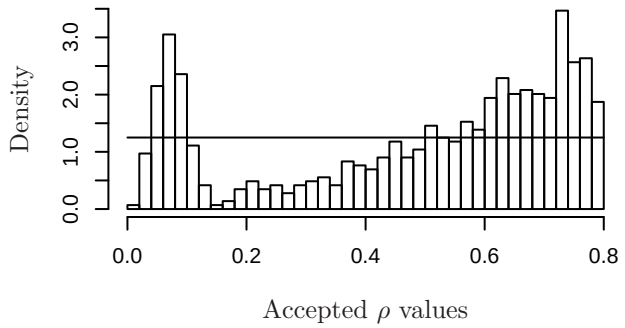

Figure 5.3: *The bimodal marginal posterior distribution of ρ when dating the primate divergence time using the standard metric and the approximate-Gibbs sampler.*

the acceptance rate is slow is that there are two stable solutions: one with small N_0 and large α values that corresponds to large ρ values, and the other with large N_0 and small α values corresponding to small ρ values. The MCMC sampler on the $\alpha-$chains becomes stuck in one of the two stable solutions, and can not easily jump to the other. This is the classic MCMC problem of poor mixing, which inevitably leads to slow convergence rates.

Table 5.2 shows the effect of using different combinations of metric and constraint in Algorithm M. Notice the variation in acceptance rates when using different combinations of metric. Any requirement on the anthropoid diversity, N_0^{sub}, leads to a low acceptance rate. This means that using ρ_p for the anthropoid subtree is not practicable as the simulations run too slowly. The reason for the low acceptance rate is that our model does not fit the data as well as it might. The subtree population does not grow as quickly in our model

as the data suggests it should. I present a solution to this in Chapters 7 and 8. For now though, a trade-off is made between the desire to achieve accurate results, and the need to produce those results in a finite amount of time. I use the population-adjusted metric, ρ_p, for the main tree and the standard metric, ρ_s, on the subtree. The constraint $N_0^{main} \geq 200$ is needed to aid convergence.

Main Tree	Sub-tree	Restrictions on N_0		$\boldsymbol{\epsilon}$	1/Acceptance rate	Median value of		
		Main Tree	Subtree			τ	τ^*	N_0
ρ_s	ρ_s	-	-	(0.3, 0.3)	130400	27.0	13.3	77
ρ_s	ρ_s	-	-	(0.4, 0.4)	51900	17.0	19.2	131
ρ_s	ρ_s	$N_0 \geq 376$	-	(0.4, 0.4)	35300	7.4	16.7	904
ρ_s	ρ_s	$N_0 \geq 376$	-	(0.5, 0.5)	1200	9.7	19.7	665
ρ_s	ρ_s	$N_0 \geq 376$	$N_0 \geq 281$	(0.7, 0.7)	1342	11.7	21.5	708
ρ_p	ρ_s	$N_0 \geq 200$	-	(0.4, 0.4)	17781	14.1	17.1	385
ρ_p	ρ_p	$N_0 \geq 200$	$N_0 \geq 100$	(0.5, 0.5)	32930	12.3	16.9	443
ρ_p	ρ_p	$N_0 \geq 200$	$N_0 \geq 100$	(0.5, 0.6)	9220	14.2	18.3	412
ρ_p	ρ_p	$N_0 \geq 200$	$N_0 \geq 100$	(0.6, 0.6)	5790	13.6	18.6	436

Table 5.2: *Summary of the effects of using different combinations of metric and restrictions on N_0, based on short runs for each scenario.*

We can increase the acceptance rate by reducing the range of the prior distributions used for ρ and τ. This does not have an effect on the posterior distributions as the range used in prior distributions (3.2) had zero posterior probability in the tails. The prior distributions used in this section are

$$\begin{aligned} \tau &\sim U[0, 50] \\ \rho &\sim U[0, 0.4] \\ \gamma &\sim U[0.005, 0.015] \\ 1/\lambda &\sim U[2, 3] \\ (a, b) &= (1, 1). \end{aligned}$$

A summary of the results obtained when using $\boldsymbol{\epsilon} = (0.4, 0.4)$ is shown below in Table 5.3, and Figure 5.4, and two plots of the joint distribution of τ and τ^* are shown in Figure 5.5. The acceptance rate was 6392 successfully simulated trees per accepted tree, which left 4930 results.

The 90% credibility intervals for τ and τ^* are $(0, 25.1)$ My and $(0, 25.2)$ My respectively. The posterior estimate of the primate divergence time is 69.8 My ago with a 90% credibility interval of $(54.8, 79.9)$ My. The posterior estimate of the anthropoid divergence time is 54.7 My ago with a 90% credibility interval of $(37.0, 62.2)$ My.

Comparing these results with those of Chapter 4 (Table 4.7 and Figure 4.7), we can

(i)	Min.	LQ	Median	Mean	UQ	Max.
N_0^{main}	200	334	384	385	437	619
$N_0^{subtree}$	2	104	150	151	196	364
τ	0.6	9.5	13.5	15.0	18.9	49.5
τ^*	0.1	13.8	17.1	17.7	21.0	49.9
ρ	0.017	0.064	0.080	0.083	0.098	0.238
γ	0.0050	0.0070	0.0092	0.0095	0.0119	0.0150
$1/\lambda$	2.00	2.24	2.48	2.49	2.74	3.00

(ii)

k	α_k
1	0.059
2	0.060
3	0.063
4	0.072
5	0.023
6	0.038
7	0.062
8	0.048
9	0.011
10	0.081
11	0.20
12	0.49
13	0.81
14	0.020

Table 5.3: *Summary statistics when dating the primate and anthropoid divergence times using the approximate-Gibbs sampler. Obtained using ρ_p for the main tree and ρ_s for the subtree, $N_0^{main} \geq 200$ and $\boldsymbol{\epsilon} = (0.4, 0.4)$ (n=4930); (i) posterior summary statistics (ii) posterior means of the sampling fractions.*

see that the posterior distributions for τ and τ^* are very similar to those obtained when using fixed ratios for the α_i and using local-linear regression on the output. The posterior mean value of τ has decreased by 3.5 My and the tail is a little shorter, but overall the agreement is good.

Figure 5.6 contains plots of the posterior distributions for the sampling fractions. Notice that the sampling rates vary a large amount from interval to interval. The maximum value of 0.81 occurs in the Early-Eocene epoch, and the minimum value is 0.011 which occurs in the Late-Oligocene. This variation cannot be taken at face value, however, as the length of each epoch is different. In the next chapter I improve this situation by using a more realistic Poisson sampling scheme.

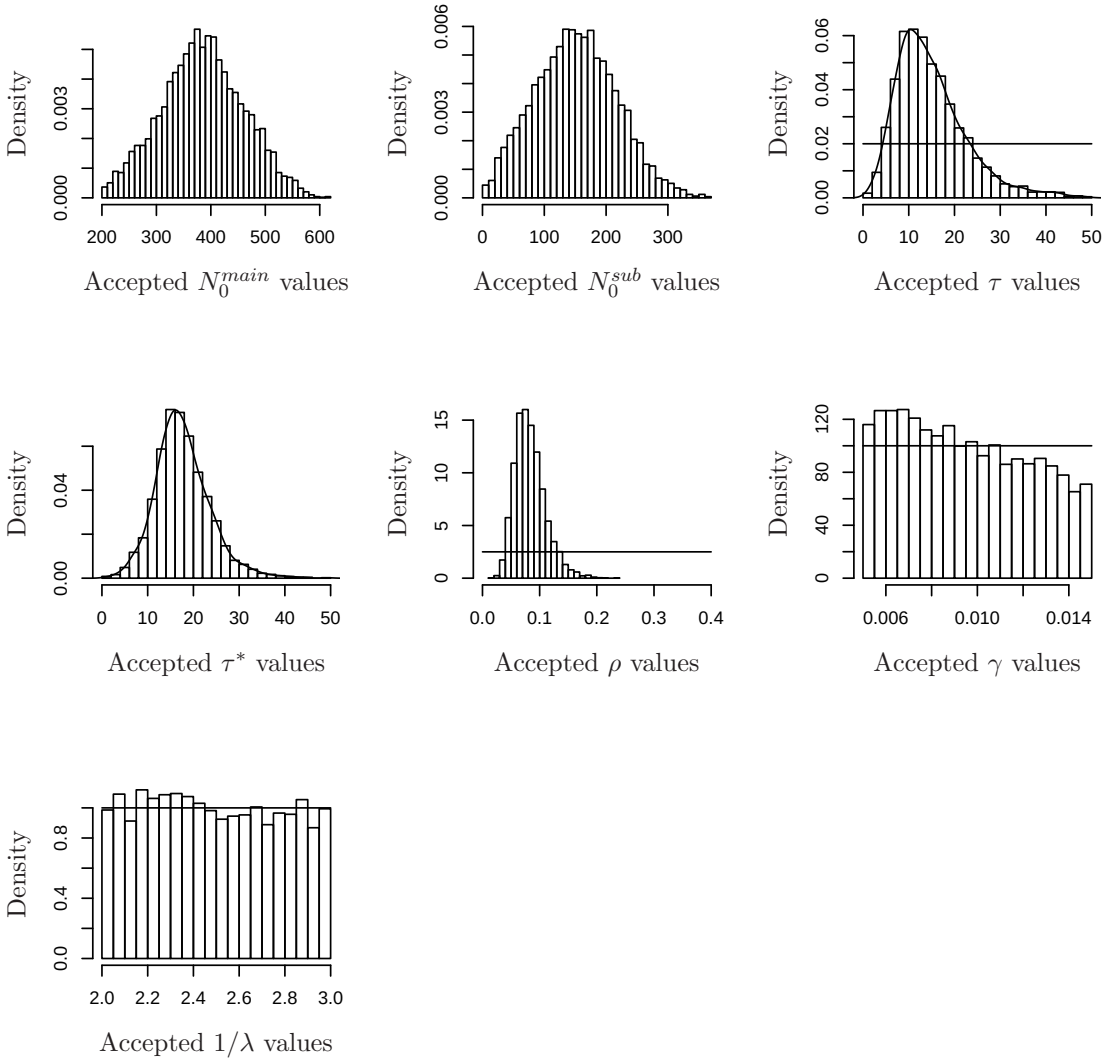

Figure 5.4: *Plots of the posterior distributions when dating the primate and anthropoid divergence times using the approximate-Gibbs sampler. Obtained using ρ_p for the main tree and ρ_s for the subtree.*

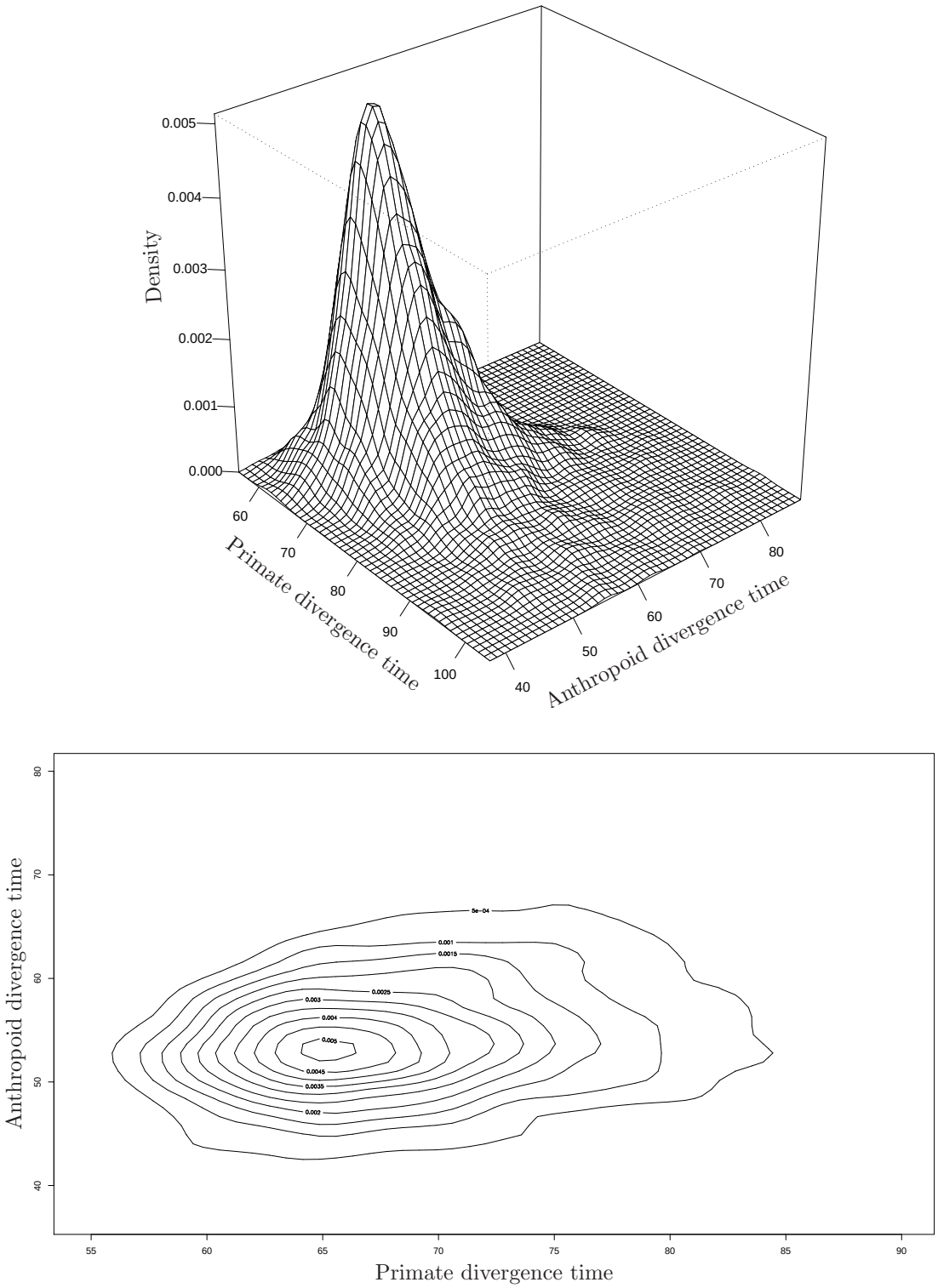

Figure 5.5: *Plots of the joint posterior distribution of the primate and anthropoid divergence times. These distributions were obtained using Algorithm M with ρ_p for the main tree and ρ_s for the subtree.*

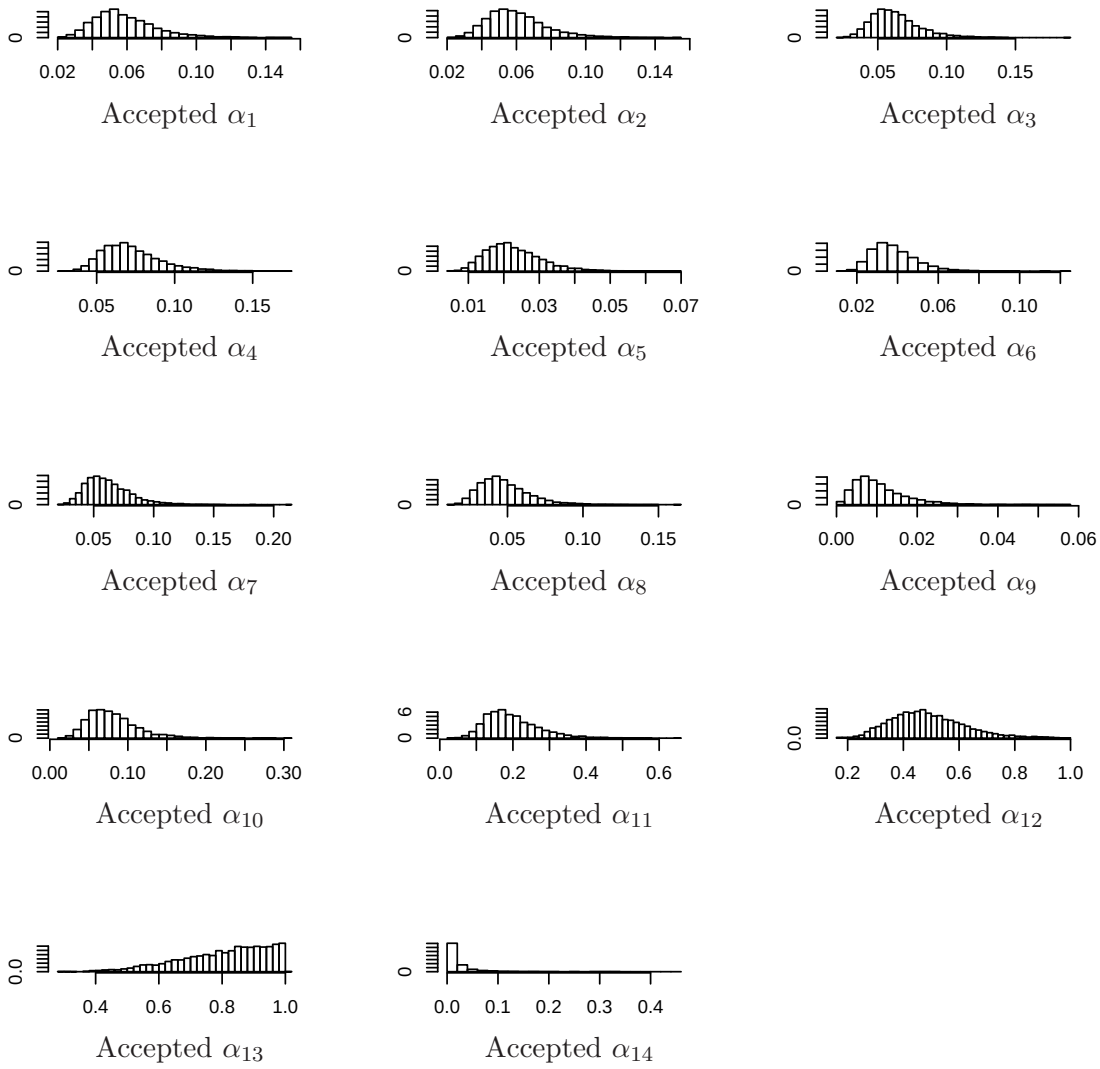

Figure 5.6: *Plots of the posterior distributions of the sampling fractions, α_i, when dating the primate and anthropoid divergence times using the approximate-Gibbs sampler. The prior distributions are $\alpha_i \sim U[0, 1]$.*

5.4.1 MCMC Difficulties and Concerns

The advantage of using Markov Chain Monte Carlo rather than the rejection algorithm is that by correlating the output more time is spent simulating with parameter values with higher posterior probabilities. MCMC, however, does suffer from potentially serious drawbacks: the output from an MCMC algorithm is not stationary, and it is difficult (perhaps impossible) to know whether any given chain has converged. Ergodic averages are used to estimate the equilibrium distribution and are known to converge in the limit (see Meyn and Tweedie [84]). Difficulties in assessing convergence arise due to the correlated nature of the output slowing the speed at which the chain explores the state space, and the fact that the output is only a sample from a distribution, rather than the distribution itself. Cowles *et al.* [34] and Brooks *et al.* [24] give a review of different convergence diagnostics.

There are also several implementation decisions that need to be made, such as how many results we need, where to start each chain, whether to use a burn-in, and whether to thin the output. In this section I address some of these issues and introduce some diagnostic tests to try to find any problems.

One of the most common difficulties faced when using MCMC is that the chain fails to mix well and consequently it fails to converge to equilibrium. When we use MCMC, a Markov chain which converges to the target distribution of interest is simulated on the parameter space. In our case, this is the posterior distribution $\pi(\theta|\mathcal{D})$. If the chain starts from a region in the tail of the posterior distribution, it can take a long time to converge to equilibrium. For this reason, some authors propose that the results from some initial transient phase, often called a *burn-in* period, are discarded in the hope that we are then left with output from the equilibrium distribution. Others, such as Geyer, point out that burn-in is only a method for finding a good starting point for the Markov chain, and recommend starting each new run where the previous run finished. He goes on to suggest that a good point is any you would not mind having in your sample, and that the problem can be avoided entirely if we condition on the starting point in a Bayesian fashion.

Another complication is that the output is correlated. A common approach is to *thin* the output by taking every n^{th} value, where n is chosen so that samples are approximately independent. Figure 5.7 below shows plots of the autocorrelation function, $R(l) = \mathbb{E}[(X_i - \mu)(X_{i+l} - \mu)]/\sigma^2$, for four of the parameters. Note that the correlation quickly drops to a negligible level for all four parameters, as it does for the parameters not shown. The quick decay of the correlation suggest that we do not need to thin the results too much, if at all, as our interest lies solely in τ, τ^* and $\boldsymbol{\alpha}$. From now on, I thin the output by taking every second value.

A maximally informative sample would be a set of independent draws from the distribution of interest. Autocorrelation decreases the amount of information in the sample;

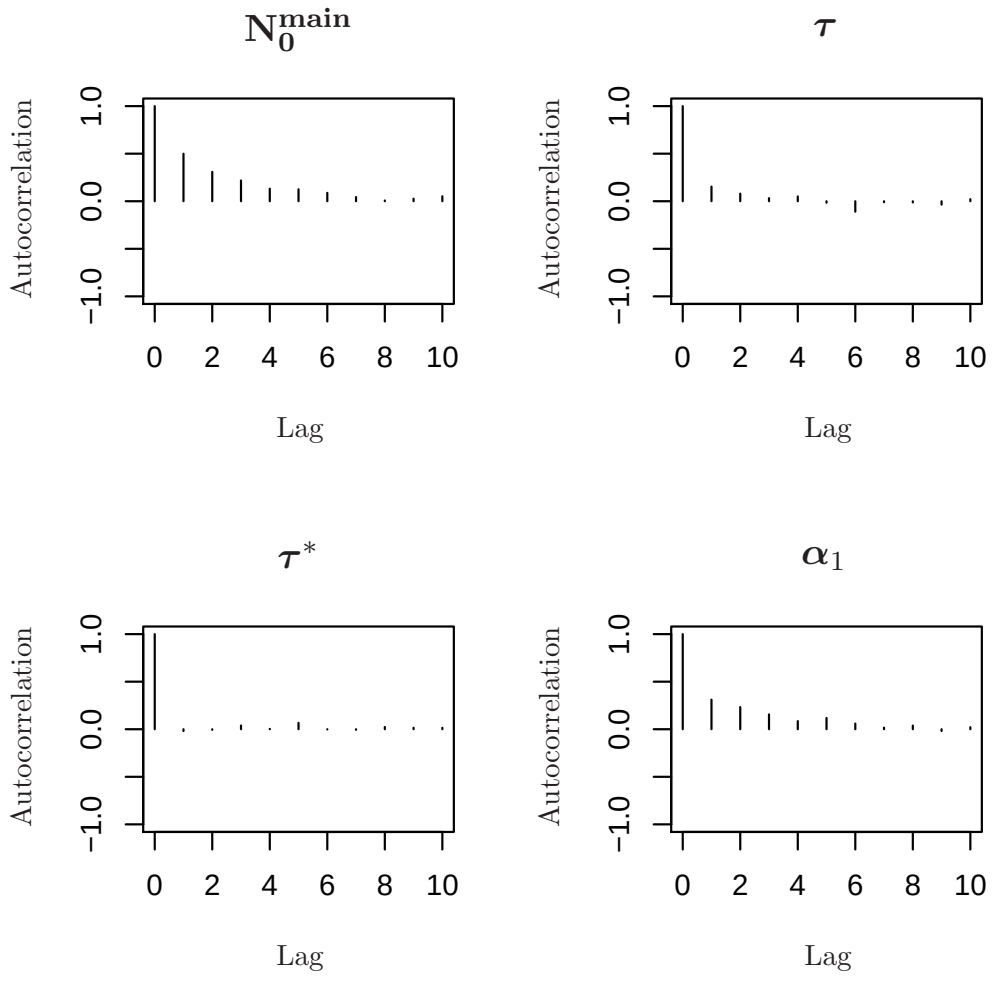

Figure 5.7: *Plots of the autocorrelation function for four different parameters when using Algorithm M. The decay of the correlation is given across a range of different lags.*

the size of the correlation determines the effective sample size (see Liu [73] page 126). For the chain analysed here, which is 2000 samples long, we have marginal effective sample sizes as follows:

N_0^{main}	τ	N_0^{sub}	τ^*	ρ	α_1
568	1560	1625	2000	1142	692

This means that the 2000 data points only contain as much information about τ as 1560 independent draws from the posterior distribution would. Notice that for parameters with a longer tailed correlation structure, such as α_1, the marginal effective sample size is much smaller than 2000. Conversely, for parameters with no correlation between values, such as τ^*, the effective sample size is equal to the true sample size.

The effective sample size determines the standard error of our estimates. Because of the correlated nature of the output, we use the time-series standard error which takes account of the correlation between consecutive values. This is also known as the asymptotic standard error and is given by the square root of the spectral density estimate divided by the sample size. The small correlation between successive values means that for our data, the two measures of standard error are nearly identical. The standard error in the estimate of the posterior mean for each parameter is as follows:

N_0^{main}	τ	N_0^{sub}	τ^*	ρ	γ	$1/\lambda$	α_1	α_2
0.60	0.061	0.53	0.048	0.00022	0.000023	0.0021	0.00016	0.00015

These are all small and suggest that we have large enough sample sizes to have confidence in the estimates[(2)].

Raftery *et al.* [100] give a method for estimating the run length needed to estimate quantiles to within a certain accuracy when using the Gibbs sampler. For example, suppose we wish to estimate the quantile $q = \mathbb{P}(\theta \leq u|\mathcal{D})$ to within an accuracy of $\pm r$ with probability s. They construct a two-state Markov chain $Z_t = \mathbb{I}_{\theta_t \leq u}$, where u is the quantile of interest, and consider its transition function and convergence rate. An R-function `raftery.diag()` is available in the package *coda* [95] to implement this technique. For our output, if we wish to estimate the median to within ± 0.01, with probability 0.95, the Raftery-Lewis method predicts that we need 10783 independent results (so we will need more than this if correlated output is used). If we wish to predict the upper 95% quantile with the same accuracy, we need 2065 results. Brooks *et al.* [24] points out that this method is only valid when estimating the run length needed when inferring quantiles, and should not be relied upon for other quantities of interest such as the mean. The method also suggests how long a burn in period is needed. Taking a burn in of 1000 iterations appears to be sufficient.

Raftery *et al.* [100] also suggests that posterior correlation between parameters can cause MCMC algorithms to converge slowly, and that one should reparameterise models to decrease correlations as much as possible. Table 5.4 shows the correlations between some of the parameters used in our model. None of these are greater than 0.5.

How well the chain mixes refers to the speed with which the sampler explores the posterior distribution. It can be difficult to diagnose poorly mixing chains; if a chain stays in one region and converges slowly, it may look as though it has converged, even when it has not. A simple diagnostic test is to run independent chains with a variety of starting points. After a possibly transient initial period, output from different chains should be indistinguishable. I checked this after each run and to the naked eye each chain seems to be from the same distribution. Another quick ad-hoc test is to look at the trace of the

[(2)] Confidence subject to the proviso of having used an approximate inference technique, our choice of model, etc.

	N_0^{main}	τ	N_0^{sub}	τ^*	ρ	γ	$1/\lambda$
N_0^{main}	1	-0.19	0.35	0.060	0.046	-0.13	0.099
τ		1	-0.025	0.30	-0.50	-0.13	0.19
N_0^{sub}			1	0.019	-0.049	-0.062	0.0037
τ^*				1	-0.10	-0.027	0.30
ρ					1	-0.023	-0.047
γ						1	-0.020
$1/\lambda$							1

Table 5.4: *Correlations between different parameters when using the approximate-Gibbs sampler (Algorithm M) to date the primate and anthropoid divergence times.*

output for each parameter of interest and to check that the sample space is being explored efficiently. Figure 5.8 shows the trace of the first 5000 results for three of the parameters of interest. Note that the chain moves quickly around the whole space, which suggests that the sampler is mixing well and converges quickly.

Many authors have given more complex diagnostic tests to check that chains have reached convergence. The lack of any general technique for a priori prediction of convergence times leads Brooks *et al.* [24] to assert that 'it is necessary to carry out some form of statistical analysis in order to assess convergence', and they recommend the tests of Gelman and Rubin [49] (generalised by Brooks and Gelman [23]) and of Raftery and Lewis. However, these tests do not detect whether a chain has converged, but try to detect when it has not.

Gelman *et al.* [49] propose using multiple independent runs which have starting points that are over-dispersed with respect to the target density to diagnose a lack of convergence. The method gives an estimate of how close we are to convergence and how much we could improve if we run the chains for longer. The coda function `gelman.diag`() reports the value R_c, known as the *potential scale reduction factor* (PSRF). If R_c is close to one, Gelman *et al.* argue that we can conclude that the set of observations is close to the target distribution. For all of the parameters reported in this section, we had a maximum PRSF of 1.01, and so conclude that the chains have converged.

Brooks *et al.* [23] point out that the approach of Gelman *et al.* ignores some of the available information when computing R_c. They propose a iterated graphical approach to address some of these problems and this has been implemented in the R-function `gelman.plot`(). Figure 5.9 shows the output when this is applied to a selection of parameters from our output. The plots show whether R_c, the potential scale reduction factor, has really converged to one, or whether it is still fluctuating. The plot shows that we can be satisfied that our chains really have converged.

Finally, it is worth noting that there are a variety of opinions about the value of running convergence diagnostics. Geyer refers to the total "bogosity of MCMC convergence

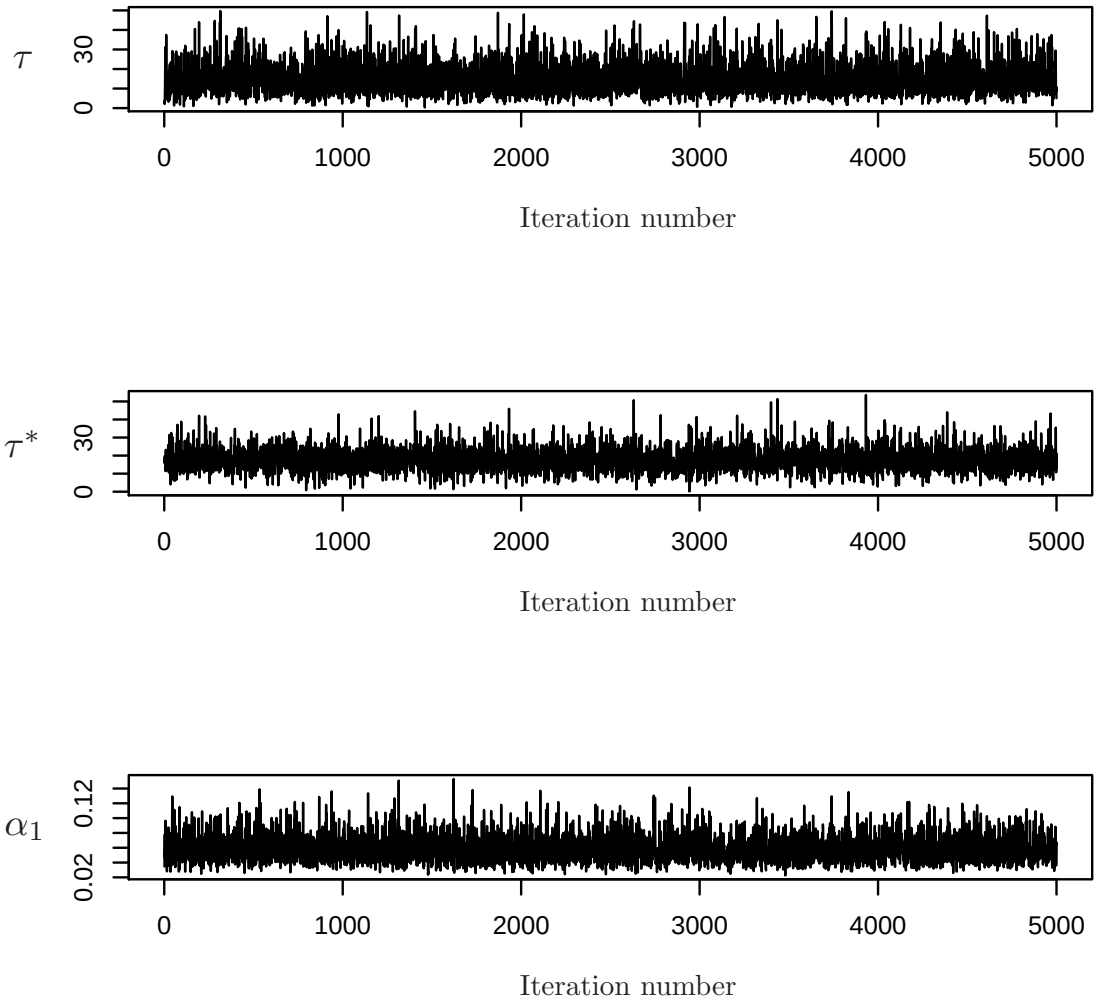

Figure 5.8: *Trace of the MCMC output from Algorithm M for τ, τ^* and α_1.*

diagnostics" and points out that they are only capable of finding gross and embarrassing problems. No test is guaranteed to work and so a variety of tests should be used. There is good reason to believe that the algorithms we use in this chapter should converge quickly, and that the output should not require too much thinning. The rejection-based ABC step samples parameters from their prior distributions and ensures that the autocorrelation function of the chain tends quickly to zero over a very short lag.

5.5 Returning to the Two Trees Situation

In Chapter 4 I gave a simulation approach which allowed us to give different growth parameters to the haplorhine and strepsirrhine sides of the speciation tree. In this chapter I have shown how to remove the fixed ratio assumption that was previously used on the sampling fractions. This section contains the results from re-examining the two trees situ-

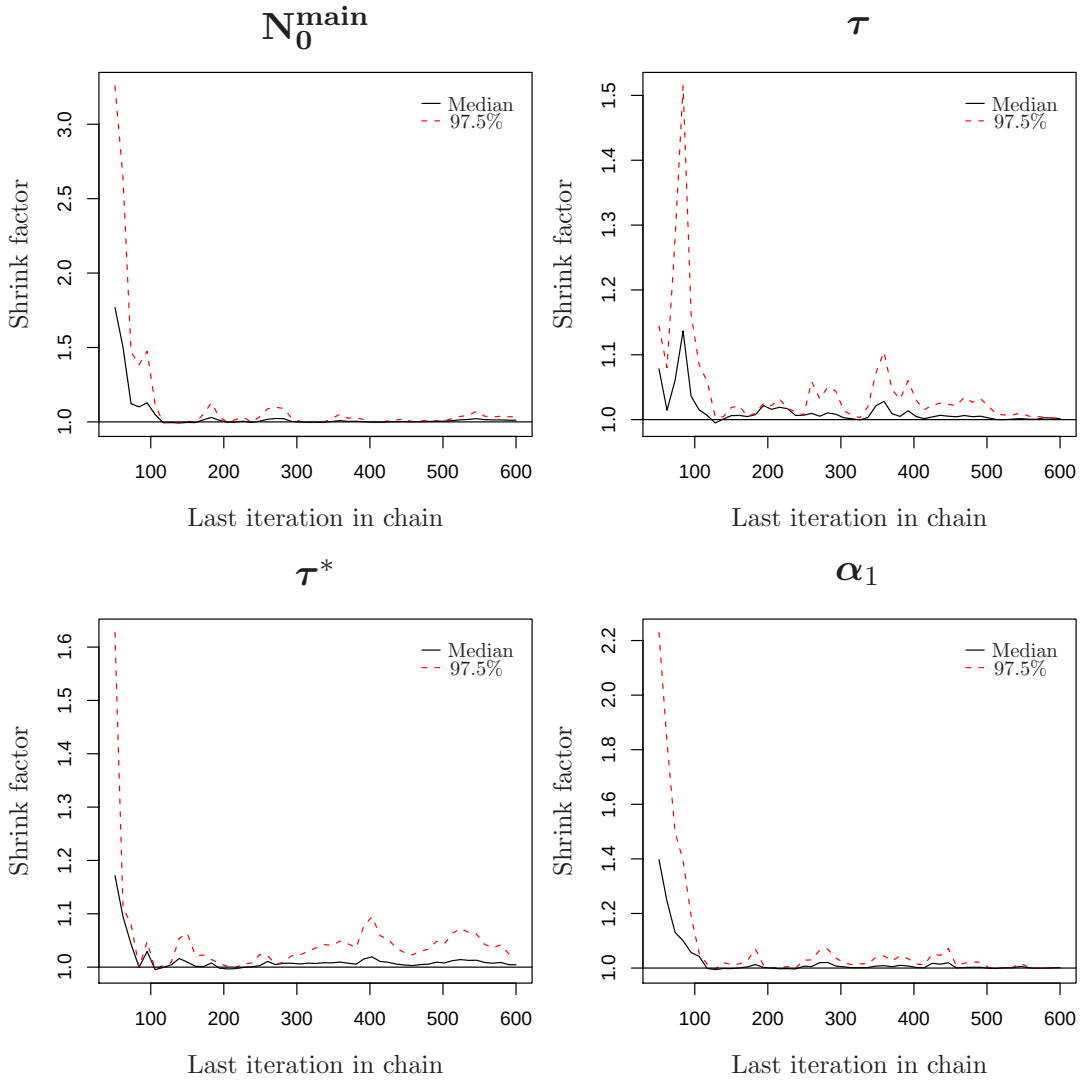

Figure 5.9: *Plots showing the convergence of the potential scale reduction factor for the output from Algorithm M. This suggests our chains have reached convergence.*

ation but with sampling fractions that are allowed to vary freely on both the strepsirrhine and haplorhine subtrees. A similar calculation to that given in Section 5.2.1 shows that the natural Gibbs update scheme for the strepsirrhine and haplorhine sampling fractions, $\boldsymbol{\alpha}^S$ and $\boldsymbol{\alpha}^H$, is to let

$$\begin{aligned} \alpha_i^S &\sim \text{Beta}(a + S_i, b - S_i + N_i^S) \\ \alpha_i^H &\sim \text{Beta}(a + H_i, b - H_i + N_i^H) \end{aligned}$$

where S_i and H_i represent the number of strepsirrhine and haplorhine fossils in interval i (see Table 4.5). Parameters $\mathcal{N}^S = (N_1^S, \ldots, N_{14}^S)$ and $\mathcal{N}^H = (N_1^H, \ldots, N_{14}^H)$ are the total number of strepsirrhine and haplorhine species in each interval in the previously accepted

tree.

Running the simulations with prior distributions (3.2), hyper parameters both equal to one ($a = b = 1$), and using the population-adjusted metric, leads to an extremely low acceptance rate. The acceptance rate starts off at a comparable value to previous simulations, but in a short amount of time it decreases to values which make simulation infeasible. A possible reason for this is that the α-Markov chains eventually move to a region where large α values are accepted, and once this has happened the chain becomes stuck and unable to return to the small values seen in previous sections.

The simulations can be dramatically speeded up by using different values for the hyper-parameters a and b. With $a = 0.1, b = 1$ the acceptance rate when using the population adjusted metric on both data sets (i.e., the haplorhine and strepsirrhine data) with a tolerance of $\boldsymbol{\epsilon} = (0.3, 0.3)$, gives an acceptance rate of 4743 successfully simulated trees per accepted tree. A summary of the results is given in Table 5.5 and Figure 5.10.

(i)	Min.	LQ	Median	Mean	UQ	Max.
N_0^H	166	238	266	267	293	395
N_0^S	50	75	85	85	93	123
τ	0.8	23.7	46.7	48.6	73.6	99.7
ρ_H	0.010	0.171	0.269	0.269	0.266	0.500
ρ_S	0.010	0.235	0.336	0.321	0.322	0.500
γ_H	0.0050	0.0070	0.0094	0.0096	0.0119	0.0150
γ_S	0.0050	0.0071	0.0094	0.0096	0.0119	0.0150
$1/\lambda$	2.00	2.38	2.62	2.58	2.81	3.00

k	α_k^H	α_k^H
1	0.080	0.0010
2	0.082	0.00090
3	0.085	0.00096
4	0.091	0.022
5	0.026	0.0078
6	0.041	0.015
7	0.059	0.0088
8	0.030	0.021
9	0.0048	0.00048
10	0.028	0.019
11	0.035	0.089
12	0.072	0.13
13	0.12	0.17
14	0.00090	0.00076

Table 5.5: *A summary of the posterior distributions when using different growth parameters and different $\boldsymbol{\alpha}$ for the haplorhine and strepsirrhine sides of the tree. Obtained using ρ_p with $\boldsymbol{\epsilon} = (0.3, 0.3)$, (n=1893); (i) summary statistics, (ii) posterior means.*

Comparing the sampling fractions above with those seen in Table 5.3 shows that the haplorhine sampling fractions show much less variation, which is more realistic than the large variation seen previously. The strepsirrhine sampling fractions, however, show much more variation with at least two orders of magnitude difference between some values. This is a consequence of the large unsampled taxonomic family[(3)] in the strepsirrhine suborder.

If we use the same sampling fractions on both sides of the tree then the acceptance rate is much lower. Using the same simulation conditions but with $\boldsymbol{\epsilon} = (0.4, 0.4)$ and $\boldsymbol{\alpha}^S = \boldsymbol{\alpha}^H$, the acceptance rate is 12037 simulated trees per accepted tree. This is much

[(3)]Recall that there are no lemurs fossils, but that there are at least 85 modern lemur species.

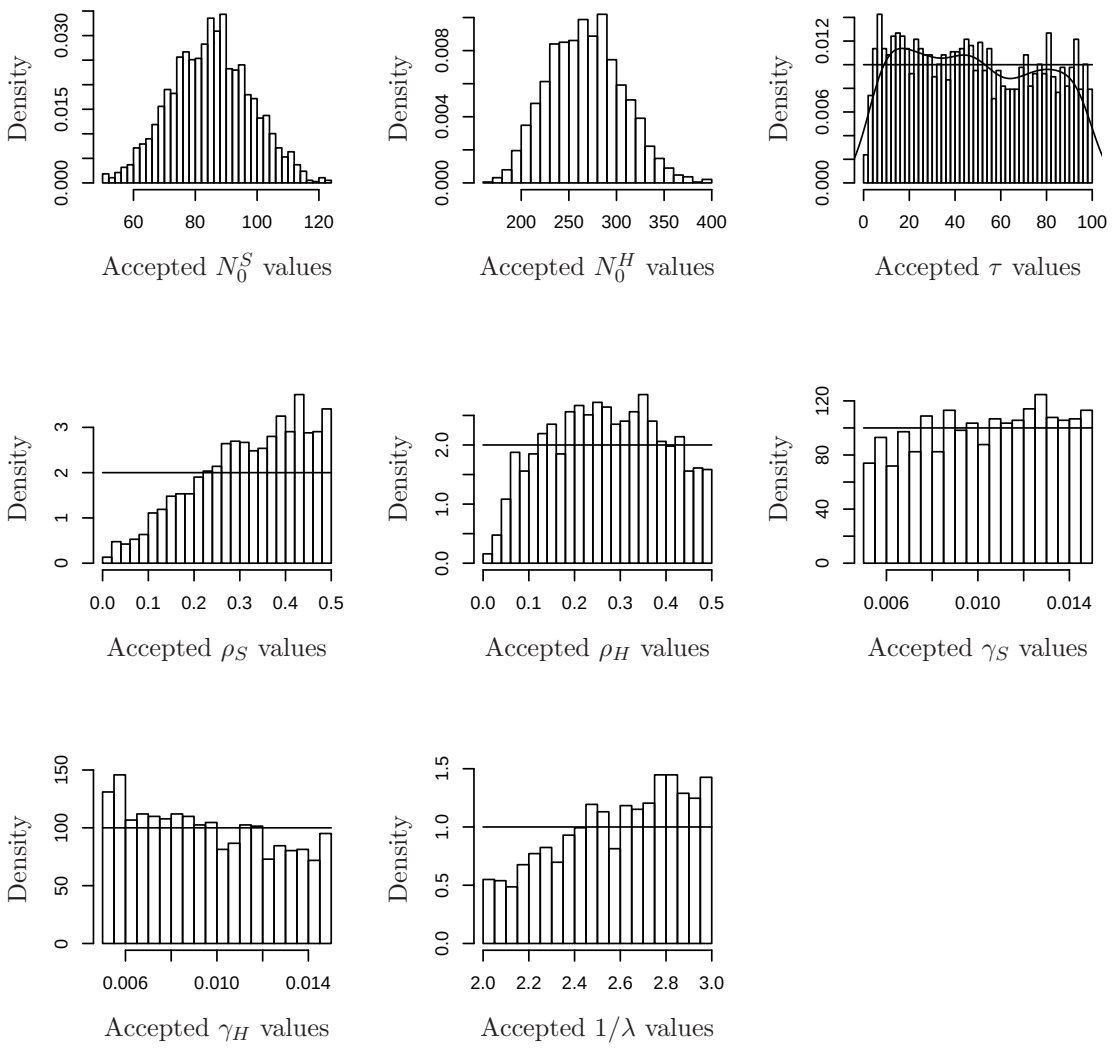

Figure 5.10: *Plots of the posterior distributions when using the approximate-Gibbs sampler and modelling the haplorhini and strepsirrhini suborders separately. The results were obtained using different* $\boldsymbol{\alpha}$ *and different growth parameters for the haplorhini and the strepsirrhini. Metric* ρ_p *was used for both trees with* $\boldsymbol{\epsilon} = (0.3, 0.3)$.

slower than when we allow the $\boldsymbol{\alpha}$ to take different values on each side of the tree. This is to be expected, as allowing $\boldsymbol{\alpha}$ to vary on each side has introduced 14 more parameters and so we should expect that the model will fit the data better.

Unfortunately, however, using so many parameters (34 in total) cannot be supported by the data. The data set we have is limited. Breaking the fossil counts down into haplorhine and strepsirrhine species gives us 30 different data points, and so to expect a model with 34 parameters to learn anything meaningful is unrealistic. We can see what effect such a large number of parameters has had by examining the posterior distribution of τ in Figure 5.10. There is very little difference between the uniform prior distribution and the marginal posterior distribution. This is because all the variation in the data

can be explained by the 28 different sampling fractions, and so consequently we lose any biotic signal about the temporal gap τ. Too many parameters also leads to confounding between the parameters, where we are no longer able to distinguish between the effect of each variable.

These results show that, given our limited data set, we are at the limit of what is possible in terms of the complexity of our model. It is not possible to increase the number of parameters in the model any further without making the output meaningless. Consequently, the work that follows this chapter focusses mainly on improving the model, rather than adding complexity. Improvements are made by making more realistic assumptions and correcting some of the simplistic approaches that have been necessary thus far.

Chapter 6

Modelling Extensions

In this chapter I extend and improve the model that lies at the heart of our approach, by introducing a more realistic model for the discovery of fossils. Rather than each species having an equal probability of being preserved, I let the preservation probability depend upon the length of time for which that species lived. In Section 6.2 I introduce a model for the mass extinction event that wiped out the dinosaurs at the Cretaceous-Tertiary boundary 65 My ago. In Section 6.3 I outline the possibility for a model selection approach that emerges naturally from the ABC algorithm, before applying this in Section 6.4 to determine which form of growth curve best fits the data.

6.1 A Poisson Sampling Scheme

The binomial fossil find model introduced in Section 2.2 has been used exclusively up until this point. Under this model each species alive during a particular epoch has an equal probability of being discovered as a fossil. This leads to a simple binomial structure, so that if N_i species lived during interval i, then $D_i \sim \text{Bin}(N_i, \alpha_i)$.

It is, however, more realistic to use a model in which the probability of a species being preserved depends upon the length of time that species lived for. So for example, a species

that only lived for 1 My should be less likely to be discovered as a fossil than one that lived for 10 My. One way to do this is to use a *Poisson sampling scheme*. We assume that fossil finds occur as the events of a Poisson point process on the branches of the tree. As we are only interested in whether a species is discovered or not, we do not care how many times a given species is found. Label each of the N_k species in interval k from 1 to N_k, then let

$$I_j^{(k)} = \begin{cases} 1 & \text{if we find species } j \text{ in interval } k, \\ 0 & \text{otherwise.} \end{cases}$$

Then, if species j lives for time $t_j^{(k)}$ in interval k, let $\mathbb{P}(I_j^{(k)} = 1) = 1 - e^{-\beta_k t_j^{(k)}}$ where $\{\beta_k\}_{k=1,\ldots,14}$ are the *sampling rates*[1] for each interval. The $\{I_j^{(k)}\}_{j=1,\ldots N_k}$ are independent Bernoulli random variables with parameters $1-e^{-\beta_k t_j^{(k)}}$, where $t_j^{(k)}$ is a random parameter. The number of fossil finds is the sum of these random variables:

$$D_k = \sum_{j=1}^{N_k} I_j,$$

i.e., the number of fossils found in interval k is the sum of N_k independent Bernoulli random variables where each variable has a different parameter.

Our aim is to find a Gibbs update step for the β_k similar to the beta conjugate priors used in Section 5.2.1. The posterior distribution of β_k is

$$\begin{aligned} \pi(\beta_k|\mathcal{D}, \boldsymbol{\lambda}, \tau, \mathcal{N}) &\propto \pi(\beta_k)\mathbb{P}(\mathcal{D}|\mathcal{N}, \beta_k) \\ &\propto \pi(\beta_k) \prod_{j=1}^{N_k} (1 - e^{-\beta_k t_j^{(k)}})^{I_j} (e^{-\beta_k t_j^{(k)}})^{1-I_j} \end{aligned}$$

where the final term on the right is the probability of observing any particular set of fossil finds, $\{I_j^{(k)}\}_{j=1,\ldots,N_k}$. Because the $t_i^{(k)}$ terms are all different, no conjugate prior distribution exists, which means we are not able to use standard Gibbs updates. We are, however, able to write down a Metropolis-Hastings update step, but due to the acceptance step (H2 in Algorithm H) the algorithm slows down considerably. Instead, it is possible to approximate $\mathcal{D}'$ by a Poisson distribution

$$D'_k = \sum_{j=1}^{N_k} I_j^{(k)} \approx \mathcal{P}o\left(\sum_{j=1}^{N_k} (1 - e^{-\beta_k t_j^{(k)}})\right)$$

[1] It no longer makes sense to call the β_k sampling fractions, as unlike the α_k, they do not represent the fraction of species preserved.

so that

$$\pi(\beta_k|\mathcal{D},\boldsymbol{\lambda},\tau,\mathcal{N}) \propto \pi(\beta_k)\left(\sum_{j=1}^{N_k}(1-e^{-\beta_k t_j^{(k)}})\right)^{D_i}\exp\left(-\sum_{j=1}^{N_k}(1-e^{-\beta_k t_j^{(k)}})\right).$$

The Kullback-Liebler divergence between the distribution of the sum of Bernoulli variables and the Poisson approximation, has been bounded above by Kontoyiannis *et al.* [69], so that

$$D(P_{S_n}||Po(\lambda)) \leq \frac{1}{\lambda}\sum_{i=1}^{n}\frac{p_i^3}{1-p_i},$$

where S_n is the sum of n independent Bernoulli(p_i) random variables and $\lambda = \sum p_i$.

Unfortunately, even with this approximation, there is still no conjugate distribution we can use. However, by approximating the sum of parameters in the Poisson distribution to first order, we find that

$$\sum_{j=1}^{N_k}(1-e^{-\beta_k t_j^{(k)}}) \approx \beta_k\sum_{j=1}^{N_k}t_j^{(k)} = \beta_k L_k$$

where L_k is the total length of all the lineages in interval k. This gives

$$\begin{aligned}\pi(\beta_k|\mathcal{D},\boldsymbol{\lambda},\tau,\mathcal{N}) &\propto \pi(\beta_k)\mathcal{P}o(\beta_k L_k)\\ &\propto \pi(\beta_k)e^{-\beta_k L_k}(\beta_k L_k)^{D_k}\end{aligned}$$

and so we can use conjugate gamma prior distributions, $\beta_k \sim \Gamma(a,b)$. The posterior distributions are then also gamma distributions, but with different parameter values.

$$\beta_k|\mathcal{D},\mathcal{N} \sim \Gamma(D_k+a, L_k+b).$$

When a and b are small the posterior means of the sampling rates are again close to the value initially hoped for:

$$\mathbb{E}(\beta_k|\mathcal{D},\mathcal{N}) = \frac{D_k+a}{L_k+b} \approx \frac{D_k}{L_k}. \tag{6.1}$$

6.1.1 Primate Crown Divergence Time

In this section I focus on dating the primate divergence time. Throughout this section the following set of prior distributions are used:

$$\begin{aligned}\tau &\sim U[0,50]\\ \gamma &\sim U[0.005,0.015]\\ \rho &\sim U[0,0.2]\\ 1/\lambda &\sim U[2,3].\end{aligned} \tag{6.2}$$

I have reduced the prior range for τ and ρ, as experience has shown that all of the posterior mass is concentrated in these regions. Extending the prior ranges for τ and ρ has no effect other than to slow down the simulations. We also need to choose the hyperparameters a and b for the gamma prior distributions given to the sampling rates. I tried numerous different combinations, none of which changed the posterior quantiles for N_0, τ, ρ or γ by more than a couple of percent. Different values for a and b did lead to different posterior values for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, however. It also had a large effect on the acceptance rate, with smaller values leading to lower acceptance rates.

Initially the simulations suffered from the same problem as found when using a binomial sampling scheme, namely that the observed values of the modern diversity were too small (the posterior median of N_0 was 50). Using the population-adjusted metric and imposing the constraint $N_0 > 200$, as used in Chapter 5, did not solve the problem. The reason for this is that N_0 converges on the lower bound of 200, so that most of the accepted runs have $N_0 \approx 200$. A solution is to require N_0 to be greater than some larger value, such as $N_0 \geq 360$. The value of 360 is chosen to be large enough so as to stop the chain converging to the wrong place, and slightly less that the 'known' diversity of 376 to account for the uncertainty surrounding this value.

Table 6.1 and Figure 6.1 give a summary of the posterior distributions obtained using $(a, b) = (5, 50)$ with the population-adjusted metric and $\epsilon = 0.4$. These hyperparameter values are small enough so that the posterior is largely determined by the data (cf. Equation (6.1)), but large enough so that the simulations run sufficiently quickly. The acceptance rate was 5362 simulated trees per accepted tree, which after burn-in and thinning leaves 3568 accepted results. The acceptance rate is much lower for the Poisson sampling scheme than for the binomial sampling scheme, suggesting that it is a poorer fit. The smallest tolerance value that can feasibly be used is $\epsilon = 0.4$, whereas for the binomial scheme $\epsilon = 0.15$ was used.

Comparing these results to the analysis done using the binomial sampling scheme (Figure 5.2), we can see that the posterior distribution for τ has a shorter tail when using Poisson sampling, although both medians are roughly the same. The posterior distribution of ρ shows the most clearly defined change from its prior distribution seen so far, which may be due to the constraint on N_0.

The largest change to previous analyses, as expected, is to the sampling rates. If we rank the epochs according to the size of the average sampling rate β_i, we find a markedly different order to that seen in Chapter 5 and that given by Tavaré *et al.* (Table 3.2). The values given in Table 6.1 agree much more closely with what we might intuitively expect.

Raup [101] argued that the fossil record is biased towards more recent species because there is more available rock from the more recent past (this is the so called *pull of the recent* argument). The inferred sampling rates largely show that here. Aside from small

(i)	Min.	LQ	Median	Mean	UQ	Max.
N_0	360	366	374	377	384	476
τ	0.1	9.6	13.6	14.7	18.5	43.9
ρ	0.017	0.056	0.066	0.067	0.077	0.177
γ	0.0050	0.0064	0.0084	0.0089	0.0110	0.0150
$1/\lambda$	2.00	2.24	2.49	2.49	2.74	3.00

(ii)

k	β_k
1	0.25
2	0.26
3	0.23
4	0.13
5	0.053
6	0.030
7	0.053
8	0.038
9	0.018
10	0.071
11	0.19
12	0.25
13	0.41
14	0.033

Table 6.1: *Summary statistics when using the Poisson sampling scheme to date the primate divergence time. Obtained with ρ_p, $N_0 \geq 360$, and $\epsilon = 0.4$ (n=3568); (i) posterior summary statistics (ii) posterior means of the sampling fractions.*

fluctuations, the β_i decrease from the most recent epoch, all the way back to the Early-Oligocene. They then rise again through the Eocene. However, the relative increase through these epochs is not as large and therefore not as unrealistic as that found with the binomial sampling scheme (cf. Section 5.3).

6.1.2 Joint Distribution of the Primate and Anthropoid Divergence Times

I now find the joint distribution of τ and τ^* using the Poisson sampling scheme. Initially, I tried using the same strategy as in Section 5.4 and used ρ_p on the main tree and ρ_s on the subtree. Although this runs with a reasonably high acceptance rate, the observed values of N_0^{sub} are too far from the known value of 281. Instead, I used the population-adjusted metric on both the main tree and the subtree. Using the hard constraint $N_0^{main} \geq 360$, and $\boldsymbol{\epsilon} = (0.5, 0.5)$ gives an acceptance rate of 4966 simulated trees per accepted tree, which after burn-in and thinning leaves 4815 results. A summary of the posterior distribution is shown below in Table 6.2 and Figure 6.2.

Unlike in the previous section, we now find that using the Poisson sampling scheme is a better fit than the binomial scheme. Using the same metrics and tolerances with binomial sampling led to an acceptance rate of 32930 simulated trees per accepted tree (see Table 5.2) which is more than six times slower than the simulations done here.

Comparing these results with those obtained when using the binomial sampling scheme in Section 5.4, we can see that the marginal posterior distributions for the two divergence times, τ and τ^*, are almost identical for both sampling models. The sampling rates, β_i, are very similar to those found in Table 6.1, which differs from Chapter 5 where we found

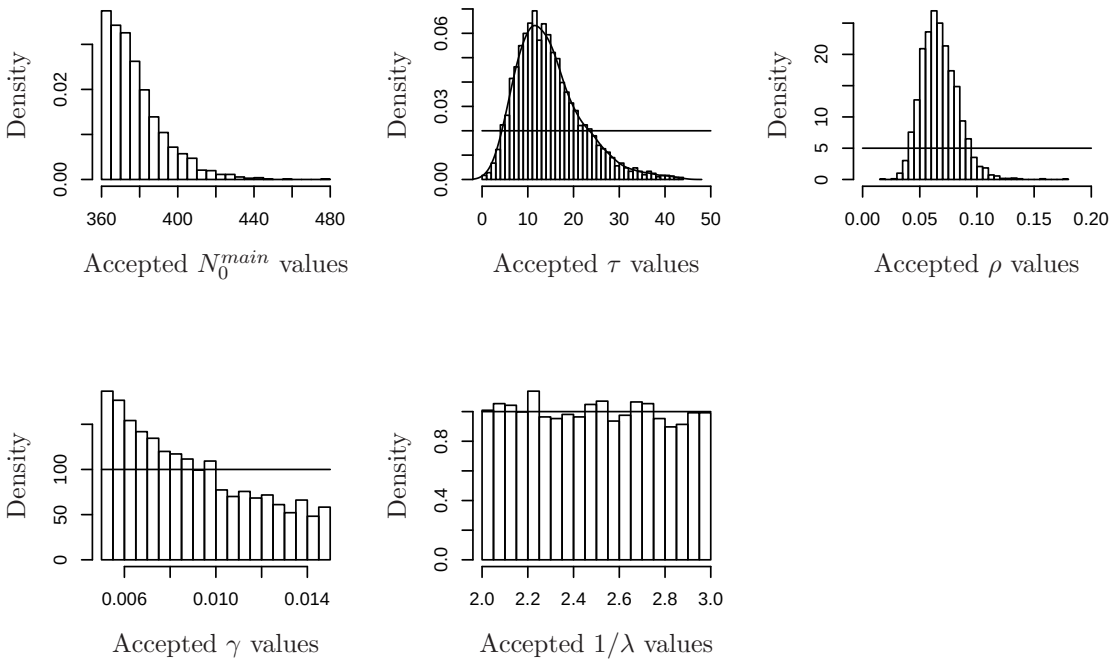

Figure 6.1: *Plots of the posterior distributions when dating the primate divergence time using a Poisson sampling scheme with ρ_p, $N_0 > 360$, $\epsilon = 0.4$, and $\Gamma(5, 50)$ priors for the β_i.*

that dating two split points caused a large change in the sampling fractions. Similarly, under the Poisson sampling scheme the posterior distribution of τ is the same whether dating both the primate and anthropoid divergence times, or just the primate divergence time. This is not what we observed in Chapter 5, but is clearly a desirable outcome.

As before, using the population-adjusted metric to measure the accuracy of the simulated anthropoid fossil counts slows down the simulations (by a factor of 10 in this case). Even using ρ_p, the observed values of the anthropoid diversity, N_0^{sub}, do not closely match the known diversity of 281 species. This is a problem with our model and is addressed in Chapter 7. Using ρ_p on the subtree slows the acceptance rate by a factor of 10 compared to using ρ_s. This is because our model is not fitting as well as it could.

6.2 Modelling the K-T Mass Extinction

The Cretaceous-Tertiary (K-T) boundary 65 My ago marks the end of the Mesozoic era, often called the *age of the dinosaurs*, and the start of the Cenozoic era and the dominance of bird and mammal species. It also marks one of the five big mass extinction events [103]. Although some authors wonder whether most fluctuations in diversity are really artefacts

(i)	Min.	LQ	Median	Mean	UQ	Max.
N_0^{main}	360	373	388	393	408	549
N_0^{sub}	103	169	187	189	208	303
τ	0.2	9.8	13.9	15.1	19.0	49.8
τ^*	0.6	11.2	15.0	15.4	19.0	49.8
ρ	0.022	0.059	0.070	0.073	0.084	0.191
γ	0.0050	0.0067	0.0087	0.0091	0.0113	0.0150
$1/\lambda$	2.00	2.21	2.44	2.46	2.70	3.00

(ii)

k	β_k
1	0.25
2	0.26
3	0.22
4	0.12
5	0.050
6	0.027
7	0.047
8	0.033
9	0.016
10	0.065
11	0.18
12	0.25
13	0.41
14	0.033

Table 6.2: *Summary statistics when dating the primate and anthropoid divergence times using Poisson sampling. Obtained with ρ_p for both trees, $N_0^{main} \geq 360$, and $\boldsymbol{\epsilon} = (0.5, 0.5)$ (n=4815); (i) posterior summary statistics (ii) posterior means of the sampling fractions.*

of the fossil record [92, 109], most now believe that a large mass extinction, known as the *K-T crash*, did take place at the K-T boundary. This led to the extinction of the dinosaurs, probably due to either a extraterrestrial impact or a flood basalt event [2].

Many authors believe the K-T crash to be of key importance in the evolutionary history of primates and other mammalian species. It is suggested that mammals evolved to fill the evolutionary niche presented by the dinosaurs demise [3]. Others, however, believe that mammals, possibly including some primate species, coexisted with the dinosaurs [19, 33, 90].

If the dinosaurs were killed by an extraterrestrial impact, then any primate species would almost certainly have been affected as well. A large asteroid collision would have thrown huge clouds of debris into the atmosphere inhibiting photosynthesis and starving many species. For this reason, in this section I give a model for the K-T crash and examine the effects on our predictions.

I model the effect of the K-T crash on the primates by specifying the probability, $\mathbb{P}_C$, that each primate species becomes extinct at the K-T boundary. The parameter $\mathbb{P}_C$, which I will refer to as the *cull-probability*, can then be treated in the same way as the other parameters by giving it a prior distribution. Our palaeontologist collaborators suggest that at least 50% of primate species became extinct at this time and so we will use the prior distribution $\mathbb{P}_C \sim U[0.5, 1]$. Note that there is no fossil evidence of Cretaceous primates and so any a priori belief about the proportion that became extinct at the K-T boundary is based solely on examining other taxonomic families.

Using the prior distributions (6.2), $(a, b) = (5, 50)$, Poisson sampling, the hard constraint

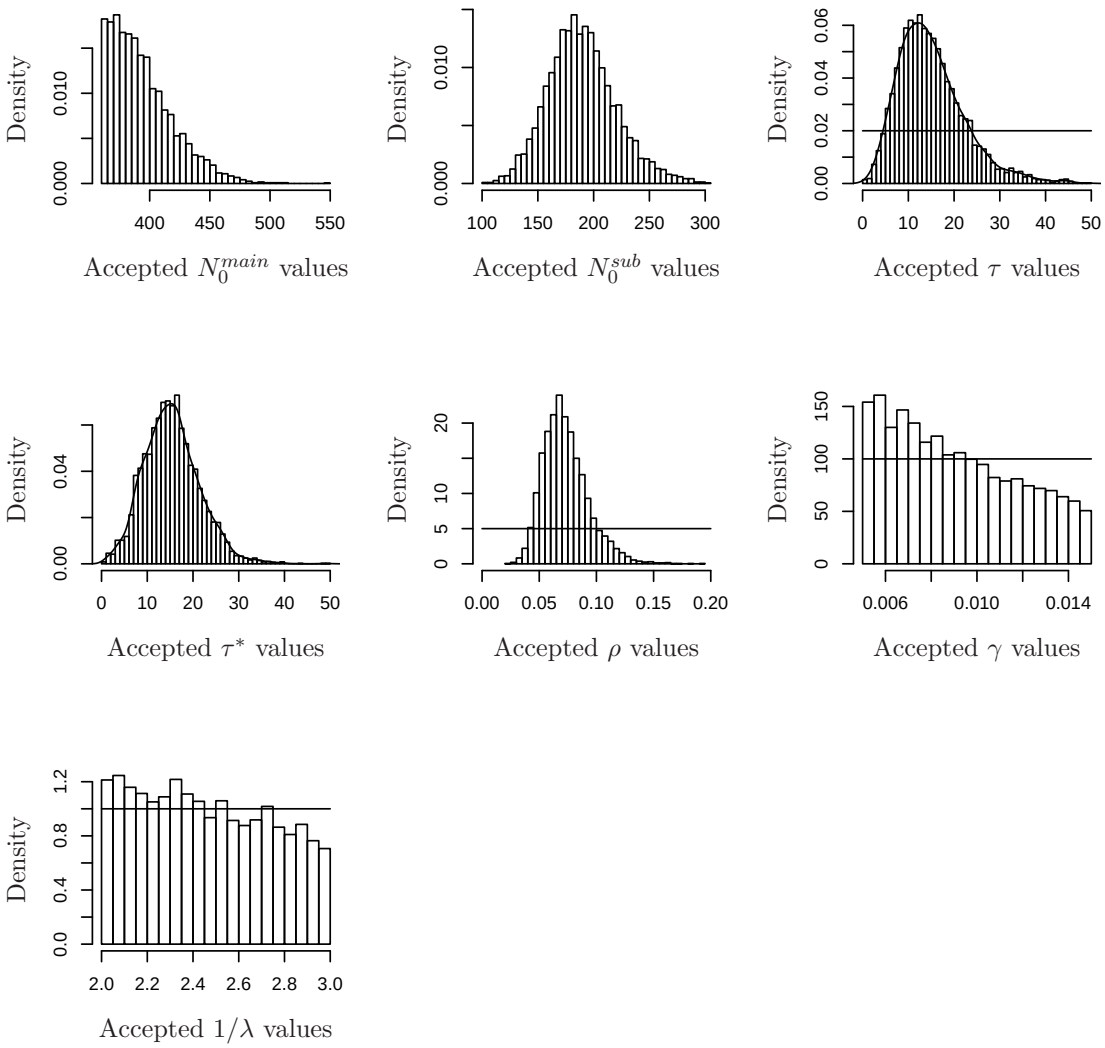

Figure 6.2: *Plots of the posterior distributions when dating the primate and anthropoid divergence times using Poisson sampling, obtained with ρ_p for the main tree and the subtree, $N_0^{main} \geq 360$, $(a, b) = (5, 50)$ and $\boldsymbol{\epsilon} = (0.5, 0.5)$.*

$N_0 \geq 360$, and the population-adjusted metric for both trees with $\boldsymbol{\epsilon} = (0.5, 0.5)$, we find the posterior distributions shown in Table 6.3 and Figure 6.3. The acceptance rate is 3098 successfully simulated trees per accepted tree, which after burn-in and thinning leaves 5819 results. Notice that this model fits the data better than the model in which the K-T mass extinction event is not considered[(2)].

Comparing the posterior distribution for τ with the results found in Section 6.1 we can see that $\tau < 10.2$ My is much more likely when modelling the K-T crash. The value

[(2)] In Section 6.1, the acceptance rate when using the same priors, metrics and constraints, was 4966 simulated trees per accepted tree. This is lower than observed here, which indicates a poorer model fit (cf. Section 6.3.

(i)	Min.	LQ	Median	Mean	UQ	Max.
N_0^{main}	360	374	391	397	413	559
N_0^{sub}	95	169	189	191	211	315
τ	0.1	6.7	8.8	11.5	13.9	49.7
τ^*	0.1	10.4	14.1	14.2	17.7	51.2
ρ	0.019	0.064	0.076	0.080	0.92	0.197
γ	0.0050	0.0068	0.0090	0.0093	0.0117	0.0150
λ	2.00	2.17	2.37	2.41	2.62	3.00
$\mathbb{P}_C$	0.500	0.592	0.702	0.718	0.836	1.000

(ii)

k	β_k
1	0.25
2	0.25
3	0.22
4	0.12
5	0.049
6	0.026
7	0.046
8	0.033
9	0.016
10	0.066
11	0.18
12	0.26
13	0.44
14	0.044

Table 6.3: *Summary statistics when modelling the K-T crash with cull probability* $\mathbb{P}_C \sim U[0.5, 1]$*, using* $N_0^{main} \geq 360$ *and* ρ_p *with* $\boldsymbol{\epsilon} = (0.5, 0.5)$ *(n=5819); (i) posterior summary statistics (ii) posterior means of the sampling fractions.*

$\tau = 10.2$ My is significant as this value corresponds to the K-T boundary. Examining the distribution more closely we can see that there is a clear change point at this time.

A posterior estimate of the probability that primates and dinosaurs coexisted can be obtained from the output by looking at the proportion of accepted τ values that are larger than 10.2 My, i.e., all those values that correspond to the primates diverging before the K-T boundary. For the simulations above we find that out of 5819 results, 1760 have $\tau > 10.2$ My, which gives the posterior probability $1760/5819 = 0.30$ that primates existed in the Cretaceous. If we return to the simulations reported in Table 6.2, we find that $\mathbb{P}(\tau > 10.2|\mathcal{D}) \approx 0.73$. Modelling the mass extinction at the K-T boundary has made the posterior probability that Cretaceous primates existed less than half what it was previously.

Examining the posterior distribution of τ^* we can see that there is also a change point in this distribution at $\tau^* = 28$ My, which again corresponds to the K-T crash. This can be seen more clearly by forcing the simulations to have Cretaceous primates by giving τ a $U[10.2, 100]$ prior distribution. Plots of the marginal distributions for τ and τ^* are shown in Figure 6.4. This allows us to more clearly see the change point for τ^*, and we can see that modelling the K-T crash makes the posterior estimate of the existence of Cretaceous anthropoids much less likely.

Finally, as expected, there is limited information about $\mathbb{P}_C$ contained in its posterior distribution. If we use a fixed value for $\mathbb{P}_C$ in the range $[0.5, 0.9]$ the posterior distributions do not change much. If we use a larger prior range for $\mathbb{P}_C$ such as $\mathbb{P}_C \sim U[0, 1]$, then the same effects are observed, but with less extreme change points.

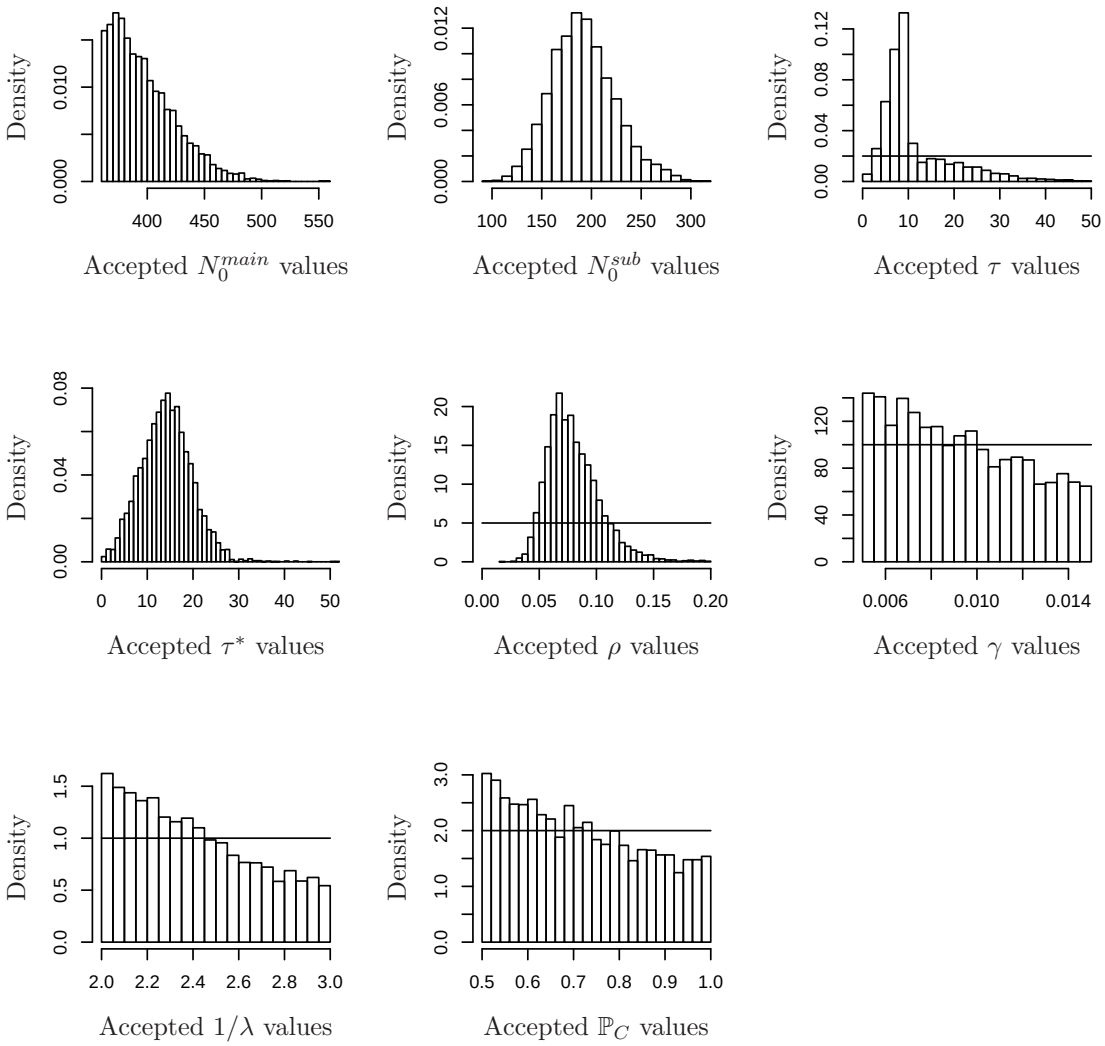

Figure 6.3: *Plots of the posterior distributions when modelling the K-T crash with* $\mathbb{P}_C \sim U[0.5, 1]$. *Results are obtained with the Poisson fossil find model,* $N_0^{main} \geq 200$, *and* ρ_p *with* $\boldsymbol{\epsilon} = (0.5, 0.5)(n = 5819)$.

6.3 Model Selection via ABC

During the modelling process, there are many decisions to be taken. We must choose the model, the prior distributions, and the inference technique, etc. One of the most important decisions, in terms of the effect on the results, concerns which model to use. There are often several different models available which we label $\mathcal{M}_1, \ldots, \mathcal{M}_k$ say. While it is often acceptable to declare that we are using model $\mathcal{M}_1$ and that all results are dependent upon this fact (cf. Goldstein's [53] subjective Bayesian approach), it is valuable to know which model best explains the data. Another common approach is to find the relative a posteriori probabilities for each model and then to take a weighted average of the predictions.

Many different model selection approaches have been used over the previous decades.

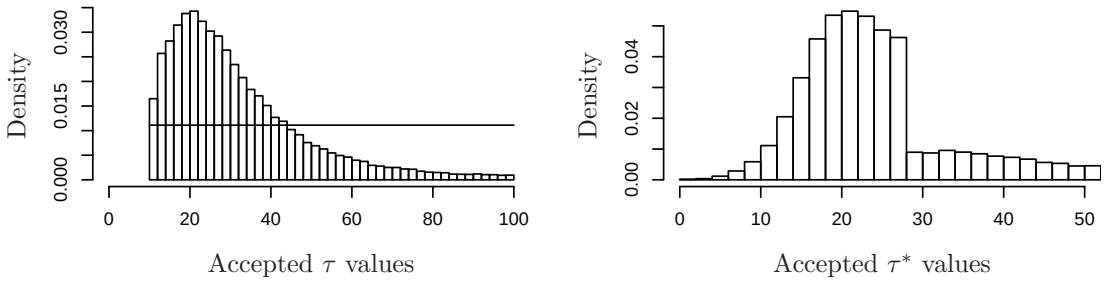

Figure 6.4: *Posterior plots of the primate and anthropoid temporal gaps when we condition on having Cretaceous primates by giving τ a $U[10.2, 100.0]$ prior distribution. Obtained using $\mathbb{P}_C = 0.75$ and ρ_p with $\boldsymbol{\epsilon} = (0.5, 0.5)$.*

Wasserman [122] gives an excellent review of some of the simpler techniques and a review of more recent developments is given by Clyde and George [30]. The problem can be described as follows: suppose we have models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ each with prior probability $\pi(\mathcal{M}_i)$. Each model gives a specification of the probabilities $\mathbb{P}(\mathcal{D}|\theta_i, \mathcal{M}_i)$ and prior distributions $\pi(\theta_i|\mathcal{M}_i)$, where θ_i is the multidimensional parameter used in model i. The aim is to find the posterior model probabilities, $\pi(\mathcal{M}_i|\mathcal{D})$.

For the problems considered in this thesis, where the likelihood function $\mathbb{P}(\mathcal{D}|\theta_i, \mathcal{M})$ is unknown, most model selection approaches I know of are impossible to apply. For example, the simplest and most popular methods are a set of asymptotic approximations, such as the Akaike Information Criterion [1] (AIC), the Bayesian (or Schwartz) Information Criterion [107] (BIC), or more recently the Deviance Information Criterion [114] (DIC), all of which depend on being able to explicitly calculate the likelihood function. Model selection is often a trade-off between complexity and goodness of fit, the idea being that models with more parameters should fit better, but are more complex. The guiding principle is that of *Occam's razor*, which states that simpler models are preferable to more complex models.

Bayesian model selection is based on calculating the Bayes factors, which Jeffreys [61] proposed as an alternative to hypothesis testing. If we wish to select between models $\mathcal{M}_1$ and $\mathcal{M}_2$, we use the Bayes factor for $\mathcal{M}_1$ to $\mathcal{M}_2$:

$$B_{12} = \frac{\mathbb{P}(\mathcal{D}|\mathcal{M}_1)}{\mathbb{P}(\mathcal{D}|\mathcal{M}_2)} = \frac{\int \pi(\theta_1|\mathcal{M}_1)\mathbb{P}(\mathcal{D}|\theta_1, \mathcal{M}_1)d\theta_1}{\int \pi(\theta_2|\mathcal{M}_2)\mathbb{P}(\mathcal{D}|\theta_2, \mathcal{M}_2)d\theta_2}.$$

The reason this quantity is used can be seen by writing

$$\frac{\mathbb{P}(\mathcal{M}_1|\mathcal{D})}{\mathbb{P}(\mathcal{M}_2|\mathcal{D})} = \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \frac{\mathbb{P}(\mathcal{D}|\mathcal{M}_1)}{\mathbb{P}(\mathcal{D}|\mathcal{M}_2)}$$

so that the *posterior odds = prior odds* $\times$ *Bayes Factor*.

The size of B_{12} records how far the data change our prior beliefs in $\mathcal{M}_1$ in preference over $\mathcal{M}_2$ (often called the strength of evidence for model one over model two). Jeffreys gives a scale for the interpretation of Bayes factors. Hypothesis testing via Bayes factors has several advantages over traditional Pearson χ^2-significance tests: they allow easy comparison of non-nested models; they give strength of evidence (or a measure of fit) for a model (whereas Neyman-Pearson tests only reject or accept models); and they can be used to give posterior model probabilities which can then be used to average over different models. See Kass and Raftery [64] for more details.

Calculation of Bayes factors is often either hard or impossible, as they depend upon the normalising constants $\mathbb{P}(\mathcal{D}|\mathcal{M}_i)$ which are usually difficult to compute with any accuracy. Various methods have been suggested: Schwartz [107] gives an asymptotic approximation for the logarithm of the Bayes factors, Chib [29] gives an approach based on rearranging Bayes formula, and Newton and Raftery [89] give an importance sampling approach to estimating the required integrals. These methods all rely on knowing the likelihood function, and in the case of the latter two, are often reported to be unstable.

The ABC approach to estimating Bayes factors is based on the acceptance rate of the algorithms. The exact rejection algorithms A and B in Chapter 3 both have acceptance rates equal to $\mathbb{P}(\mathcal{D}|\mathcal{M})$. By using two different models we can, in theory at least, get an estimate of the Bayes factor by taking the ratio of two acceptance rates.

When using the approximate algorithms C and D, the acceptance rate is equal to $\mathbb{P}(\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon|\mathcal{M})$. By looking at the ratio of the acceptance rates we approximate the value

$$\frac{\mathbb{P}(\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon|\mathcal{M}_1)}{\mathbb{P}(\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon|\mathcal{M}_2)}$$

which should tend to the Bayes factor as $\epsilon \downarrow 0$. At present, there is no theory to show that either estimate will be robust. In the next section I try to apply this idea to choosing a growth curve for primate evolution.

6.4 Using Different Growth Curves

Up until this point of the thesis I have exclusively used a logistic growth curve to model the diversity as this is the model most commonly used in the palaeontology literature (cf. Section 2.1.1). There are, however, numerous other growth curves that could have been used to determine the birth probabilities. In this section I try a couple of other growth curves before using the methodology of the previous section to indicate which best fits the data.

The method outlined in Section 6.3 only works for the rejection-based ABC algorithms. It is not clear how to extend the methods to the hybrid ABC-MCMC algorithms of Chapter 5, and so for this reason I shall use Algorithm E from Chapter 4 throughout this section.

I also use the simpler binomial fossil-find model with no K-T crash in order to keep everything simple and to help aid comparison with previous results. Throughout this section I use the prior distributions $\tau \sim U[0, 100]$, $\alpha \sim U[0, 0.3]$, and the population-adjusted metric.

6.4.1 Linear Growth

In this section I use a linear growth curve to model the primate diversity through time. Equating the linear growth curve

$$\mathbb{E}(Z_t) = at + b$$

with the expected diversity for the general binary branching process (Equation (2.9)) gives, after differentiating, the branching probabilities

$$p_2(t) = \frac{1}{2} + \frac{a}{2\lambda(at+b)}, \quad p_0(t) = \frac{1}{2} - \frac{a}{2\lambda(at+b)}.$$

The parameter b is determined by the initial population size, and so we fix $b = 2$ for all of the simulations. Experimentation with the prior range for the gradient parameter a leads us to try giving a a $U[0, 50]$ prior distribution. Changing the prior range has little effect, as can be seen from the plots of its marginal posterior distribution. Using the population-adjusted metric with $\epsilon = 0.15$ gives the results summarised in Figure 6.5.

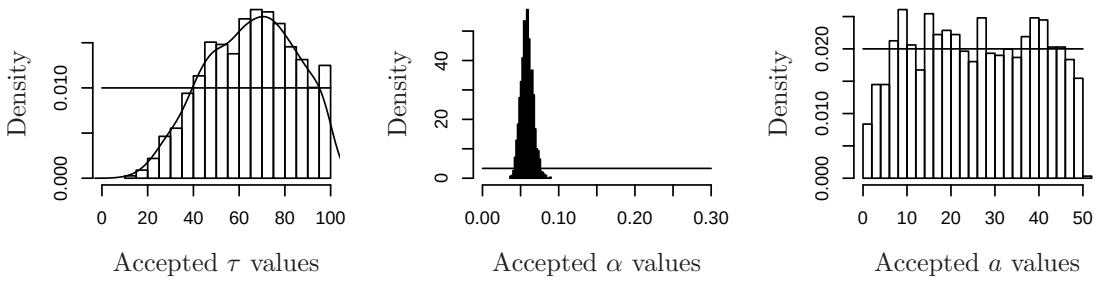

Figure 6.5: *Marginal posterior distributions when using a linear growth curve for the primate population growth. Obtained using ρ_p with $\epsilon = 0.15$ ($n = 1553$).*

I am not proposing linear growth as a realistic model, and so I will not give much attention to these results other than to say that the divergence times suggested by the marginal distribution for τ are unrealistic. The large range extension is due to the slow linear population growth that does not allow the population to reach a large enough value to account for the number of Eocene fossils in a short period of time. Acceptance rates for different values of ϵ are shown in Table 6.4.

6.4.2 Exponential Growth

I also tried fitting an exponential growth curve. Equating

$$\mathbb{E}(Z_t) = 2e^{kt}$$

with Equation (2.9) and differentiating leads to the branching probabilities

$$p_2(t) = \frac{1}{2} + \frac{k}{2\lambda}, \;\; p_0(t) = \frac{1}{2} - \frac{k}{2\lambda}.$$

Using exponential growth is equivalent to using a homogeneous Galton-Watson process, which can be seen by noting that the branching probabilities do not depend upon time.

The main difficulty when using an exponential growth curve is that the population size grows extremely quickly and so care needs to be taken to discard trees once they become too large. Giving k a uniform prior, $k \sim U[0, 0.05]$, and using ρ_p with $\epsilon = 0.2$ leads to the results summarised in Figure 6.6. Again note that observed τ values are unrealistic, as is to be expected. Exponential population growth is not feasible over an extended period due to the limited amount of resources available.

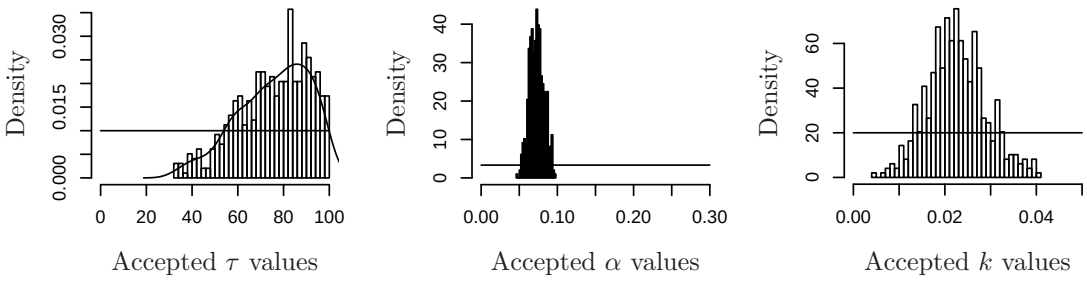

Figure 6.6: *Plots of the marginal posterior distributions when using exponential growth curves for the primate population growth. Obtained using ρ_p with $\epsilon = 0.2$ ($n = 490$).*

6.4.3 Logistic Growth

Throughout this thesis I have used logistic growth (Equation (2.10)) to describe the primate diversity. However, I have not yet given the results from dating just the primate divergence time using the population-adjusted metric. Using prior distributions $\rho \sim U[0, 0.5]$ and $\gamma \sim U[0.005, 0.015]$ with $\epsilon = 0.1$ gives the results summarised in Figure 6.7.

Notice that the marginal posterior distribution for the sampling fractions α is almost identical for all three growth curves. This is because the constraint on the diversity imposed by the population-adjusted metric ensures that only a small range of α values are able to explain both the fossil counts and the fact that $N_0 = 376$ species.

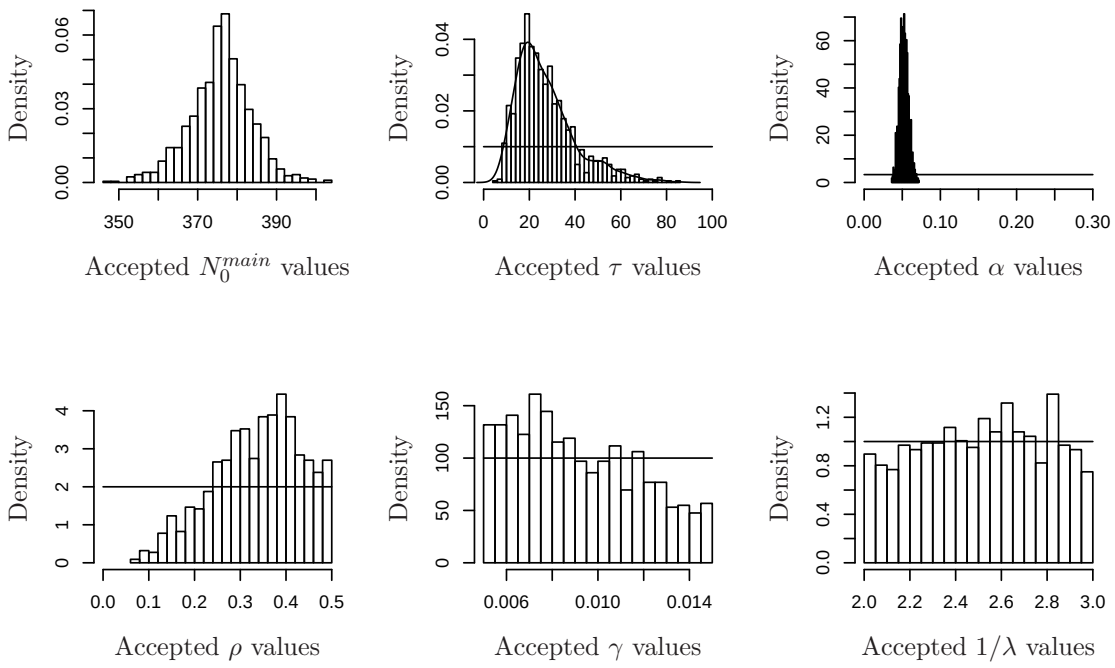

Figure 6.7: *Plots of the marginal posterior distributions when using a logistic curve for the primate population growth. Obtained using ρ_p with $\epsilon = 0.1$.*

6.4.4 Acceptance Rates

I now give the acceptance rates for each model. Before doing this however, we should consider what quantity we should use to calculate the acceptance rates. We have a choice; either the reciprocal of the number of trees simulated per accepted tree, or the reciprocal of the number of successfully simulated trees per accepted tree. By a successfully simulated tree I mean a tree that survives to the present with both sides extant (i.e., with extant haplorhine and strepsirrhine species; cf. Section 2.3.1.). The value we use makes a difference to the size of the Bayes factor between the models. For example, using a linear growth curve leads to more trees going extinct during the simulations than using the logistic growth curve does. And so using the first quantity to calculate the Bayes factor B_{LoLi} (evidence for logistic over linear growth) gives a larger value than using the second quantity.

Here I use the second quantity, the number of successfully simulated trees per accepted tree, as I believe that this is what we are interested in. Our model is $\widetilde{Z}$ from Chapter 2, i.e., trees that survive on both sides. Because we are unable to simulate from this model we use the trick of simulating Z values and throwing away results where $Z = 0$. The acceptance rate should therefore be based on what percentage of successful trees are

Acceptance Rates			
ϵ	Logistic Growth	Linear Growth	Exponential Growth
0.3	72	127	387
0.2	451	1109	11925
0.15	2964	11759	23419
0.1	135238	∞	∞

Bayes Factors			
ϵ	B_{LoExp}	B_{LoLi}	B_{LiExp}
0.3	5.4	1.8	3.1
0.2	26.4	2.5	10.8
0.15	7.9	4.0	2.0

Table 6.4: *Acceptance rates and Bayes factors for the three different models; linear, logistic and exponential population growth. Acceptance rates given as number of successfully simulated trees per accepted tree.*

accepted.

Table 6.4 shows how the acceptance rate varies with ϵ for the three different growth curves, and the approximate Bayes factors between the three models. We can see that for $\epsilon = 0.15$, B_{LoExp} and B_{LoLi} are both greater than three, which Jeffries interprets as substantial evidence in favour of logistic growth over either of the other two forms.

For this toy example we would not need to perform model selection as there are physical reasons why exponential and linear growth are unrealistic. The analysis is included here, however, as an example of how it could be used in other situations.

There are many problems with this approach. Firstly, Table 6.4 shows how unstable the estimate of the Bayes factor can be, with large fluctuations as ϵ varies. Secondly, the acceptance rate varies with the prior range of values used. If we used the prior distribution $\tau \sim U[0, 50]$ then the evidence in favour of logistic growth becomes much stronger. I am not sure how to solve either of these two problems, or indeed whether they can be solved. For this reason, I think the ideas expressed in Section 6.3 can only be used as a tool for the modeller to compare the relative fit of two similar models, rather than as a formal tool to use in model selection.

6.5 Future Modelling Extensions

There is disagreement amongst primatologists about the early evolutionary history of primates. One of the areas of contention concerns the Eocene primates. Nearly all primates from this epoch belonged to either the adapiform infraorder, often known as lemuroids, or the omomyiforms which are often known as tarsioids [81]. One of the key unanswered questions in primatology concerns how these infraorders are related to the extant primates

[105]. Previously, there was widespread belief that adapiforms were relatives of modern strepsirrhine primates and that omomyiforms were relatives of modern haplorhine primates. However, new fossil discoveries have led to this picture being questioned [78]. An alternative theory is that the adapiforms and the omomyiforms formed a separate outgroup. Although the idea is not pursued here, it is possible to apply the model selection approach above to see which evolutionary scenario best fits the model. For example, if we examine the primate data without the omomyiform and adapiform fossil counts (Table 6.5; derived from Table 1.1 by summing over the columns for each row, but excluding the stem strepsirrhine and stem haplorhine columns) we find that the acceptance rate is much higher than that observed previously. For example, using the prior distributions (3.2), the population-adjusted metric on the main tree and the standard metric on the subtree, with $\boldsymbol{\epsilon} = (0.4, 0.4)$ gives an acceptance rate of 210 successfully simulated trees per accepted tree. This is much faster than in Section 5.4 where the acceptance rate was 6392 successfully simulated trees per accepted tree. This shows that the reduced data set fits the logistic growth model much better than the full data set. Plots of the posterior distributions of the temporal gaps are shown in Figure 6.8. Note how the range extension for the primates (τ) is smaller than seen previously (cf. Figure 5.4), but that the range extension for the anthropoids (τ^*) is comparable to before. Both results are as expected. In order to construct a full model selection argument between the four evolutionary scenarios given in Martin [78], we would need to better develop the two trees approach and model the adapiforms and omomyiforms as a subtree that dies out towards the end of the Eocene. This is beyond the scope of this thesis.

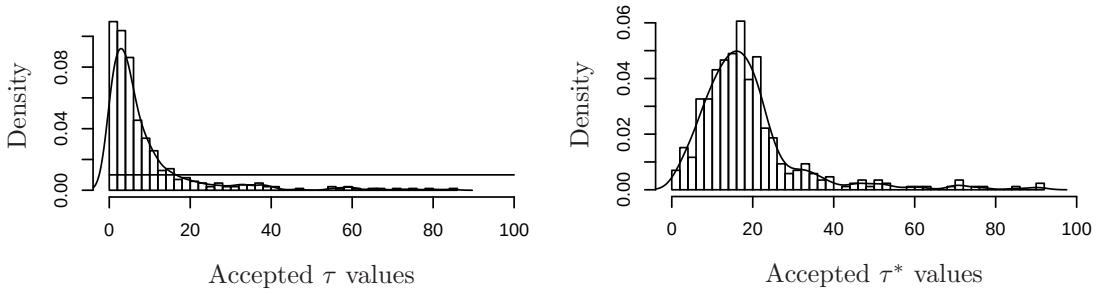

Figure 6.8: *Plots of the posterior temporal gaps for the primates and the anthropoids when using the data set without the adapiform and omomyiform data. Obtained using ρ_p for the main tree, ρ_s for the subtree with $\epsilon = (0.4, 0.4)$.*

A further modelling extension, would be to use a more general growth curve. Logistic growth, although widely accepted, is very simple and may be ignoring some of the complexities hidden in the data. It may be possible to take into account some of the

Table 6.5: *Fossil counts for the strepsirrhine and haplorhine crown species. These data do not include any adapiform or omomyiform fossil counts.*

	Number of strepsirrhine and haplorhine crown fossils	Number of anthropoid fossils
Extant	376	281
Late-Pleistocene	22	22
Middle-Pleistocene	28	28
Early-Pleistocene	30	30
Late-Pliocene	43	40
Early-Pliocene	12	11
Late-Miocene	35	34
Middle-Miocene	45	43
Early-Miocene	34	28
Late-Oligocene	2	2
Early-Oligocene	18	6
Late-Eocene	9	2
Middle-Eocene	13	0
Early-Eocene	2	0
Pre-Eocene	0	0

fluctuations that are believed to have taken place throughout the Cenozoic. For example, we could take into consideration the two possible mass extinctions that took place during the Cenozoic. The first occurred at the Eocene-Oligocene boundary and is known as the *Grand Coupure*. All European primates became extinct at this time, possibly due to climate change. The second extinction took place during the Miocene. We could model this growth curve either by using a more general curve shape or by allowing the birth and death probabilities p_2 and p_0 to vary from epoch to epoch.

It is of course, simple to date other divergence times using the data set we have available. We could for example, without much difficulty, find the joint distribution of the primate and catarrhine divergence times. There would likely be some convergence problems to overcome, but in theory this would pose no problem. I do not believe it is possible to find the joint distribution of three or more divergence times at present as the computer time required would be too great. However, with further methodological developments and increases in computer power it may become possible in the future.

Chapter 7

Conditioned Galton-Watson Trees

In previous chapters I have stressed the importance of conditioning in our calculations. For example, in Chapter 2 I noted that because we know that there are extant primates, we should model speciation using branching processes conditioned to have a non-zero modern diversity. Similarly, in Chapter 4 we adjusted the metric used in the ABC algorithms so that we were estimating the posterior distribution of the parameters conditioned on the data and on the fact that there are 376 modern primate species.

Another implicit act of conditioning has also been performed since Section 4.1. In that section we took into account the fact that we knew that the anthropoids formed a subtree in the main primate tree. To take this into account I proposed the optimal subtree selection algorithm to find which subtree looked most like the known anthropoid subtree. A better and more theoretically sound approach would have been to simulate a primate tree conditional on the fact that the anthropoid subtree originated at a given time. That is the aim of this chapter.

Although many authors have considered conditioned branching processes before, no one has looked at processes conditioned to have a subtree originate at a particular point in time. Kennedy [68] considered discrete time homogeneous Galton-Watson processes conditioned on the total number of progeny; Geiger [45] considered continuous time binary Galton-Watson trees conditioned on extinction and non-extinction; and Chauvin *et al.* [28]

considered Markovian branching particle systems in $\mathbb{R}^d$ conditioned to have a particle in a particular place at a given time. Waugh [124] considers general conditioned Markov processes. Unfortunately, none of these approaches is directly applicable to our situation, although Chauvin *et al.* provides the inspiration for the approach set out in Section 7.3 of this chapter.

This chapter contains a description of the structure of continuous time inhomogeneous branching processes conditioned to have a birth at a particular point in time. Section 7.1 contains a calculation that motivates the formal theory which is contained later in Section 7.3. Section 7.2 contains an approximate method for simulating conditioned trees, which I worked on before discovering a more efficient and exact method. The final part of the chapter, Section 7.3, contains a description of the fish-bone like structure of the conditioned trees. In order to prove that this is the case, a measure space on trees is described, and the Galton-Watson measure is explained. The proof is rather technical and relies on representing the birth process of the tree as a random point measure on the line and using Palm theory to perform the conditioning. The approach works for general offspring distributions, not just the binary trees considered so far. Details of how to implement this approach, and the results obtained, are given in Chapter 8.

7.1 The Conditioned Tree

In this section I give a heuristic derivation to illustrate what effect conditioning has on the Galton-Watson process. The arguments used in this section are not intended as a formal proof, but instead as calculations to motivate what follows in Section 7.3.

The aim is to discover the structure of Galton-Watson processes conditioned to have a birth (i.e., a split point) at some fixed time y. Initially, I only consider Galton-Watson processes which begin with a single lineage at time 0. We can deal with processes begun with more than one lineage by an application of the size-biased lemma given later in this section. As in Chapter 2, let $X(t)$ denote the size at time t of a tree begun from a single species at time zero, and let $Z(t)$ denote the size of the tree begun from two species. Let $B(y)$ denote the event that a species dies and is replaced by two new species at time y, and let $B(y, y+h)$ denote the event that a birth occurs between times y and $y+h$. As we are interested in the continuous time process, a priori, $\mathbb{P}(B(y)) = 0$ for any given y. Therefore, in order to determine the conditional probability $\mathbb{P}(X(t) = k|B(y))$, we must consider the limit $\lim_{h\downarrow 0} \mathbb{P}(X(t) = k|B(y, y+h))$[(1)].

It is necessary to split the calculation into two parts: namely, before and after the birth at time y. Firstly, consider the growth of the tree before the conditional split point, i.e.,

[(1)] Feller [42] contains a useful discussion about conditioning on events which have probability 0.

$t < y$:

$$\mathbb{P}(X(t) = j \mid B(y)) = \mathbb{P}(X(t) = j). \lim_{h \downarrow 0} \frac{\mathbb{P}(B(y, y+h) \mid X(t) = j)}{\mathbb{P}(B(y, y+h))}.$$

By conditioning on the value $X(y)$, and recalling that for a population of size n the time to the next death has an $\text{Exp}(n\lambda)$ distribution, we can see that

$$\begin{aligned}
\mathbb{P}(X(t) = j | B(y)) &= \mathbb{P}(X(t) = j) \lim_{h \downarrow 0} \frac{\sum_{k=0}^{\infty} \mathbb{P}(B(y, y+h) | X(y) = k) \mathbb{P}(X(y) = k | X(t) = j)}{\sum_{k=1}^{\infty} \mathbb{P}(B(y, y+h) \mid X(y) = k) \mathbb{P}(X(y) = k)} \\
&= \mathbb{P}(X(t) = j) \lim_{h \downarrow 0} \frac{\sum (k\lambda h + o(h))\, p_2(y)\, \mathbb{P}(X(y) = k \mid X(t) = j)}{\sum (k\lambda h + o(h))\, p_2(y)\, \mathbb{P}(X(y) = k)}
\end{aligned}$$

where $p_2(t)$ is the probability that a death at time t leads to the birth of two new species. By dividing both the numerator and the denominator by h, recalling that $\lim(a_n/b_n) = \lim a_n / \lim b_n$, and exchanging the order of the summation and the limit, we find that

$$\begin{aligned}
\mathbb{P}(X(t) = j | B(y)) &= \mathbb{P}(X(t) = j) \frac{\sum \lim(k\lambda + o(h)/h)\, \mathbb{P}(X(y) = k \mid X(t) = j)}{\sum \lim(k\lambda + o(h)/h)\, \mathbb{P}(X(y) = k)} \\
&= \mathbb{P}(X(t) = j) \frac{\sum k \mathbb{P}(X(y) = k \mid X(t) = j)}{\sum k \mathbb{P}(X(y) = k)} \\
&= \mathbb{P}(X(t) = j) \frac{\mathbb{E}(X(y) \mid X(t) = j)}{\mathbb{E}X(y)}.
\end{aligned}$$

The branching property implies that $\mathbb{E}(X(y) \mid X(t) = j) = j\mathbb{E}(X(y) \mid X(t) = 1)$ and we can use the tower property to note that

$$\begin{aligned}
\mathbb{E}X(y) &= \mathbb{E}[\mathbb{E}(X(y) \mid X(t))] \\
&= \mathbb{E}[X(t)\mathbb{E}(X(y) \mid X(t) = 1)] \\
&= \mathbb{E}X(t)\mathbb{E}(X(y) \mid X(t) = 1)
\end{aligned}$$

and that therefore

$$\mathbb{P}(X(t) = j \mid B(y)) = \frac{j\mathbb{P}(X(t) = j)}{\mathbb{E}X(t)}. \tag{7.1}$$

Now consider the structure of the process after the conditioned split point, i.e., for $t > y$:

$$\begin{aligned}
\mathbb{P}(X(t) = j \mid B(y)) &= \sum_{k=1}^{\infty} \mathbb{P}(X(t) = j, X(y_+) = k+1 \mid B(y)) \\
&= \sum \mathbb{P}(X(t) = j \mid X(y_+) = k+1) \mathbb{P}(X(y_-) = k \mid B(y)) \\
&= \sum \mathbb{P}(X(t) = j \mid X(y) = k+1) \frac{k\mathbb{P}(X(y) = k)}{\mathbb{E}X(y)}.
\end{aligned} \tag{7.2}$$

where y_+ denotes the time just after the birth $B(y)$ and y_- represents the time just before. Taken together, Equations (7.1) and (7.2) give the distribution of the size process

of conditioned Galton-Watson trees.

Equation (7.1) shows that the conditioning leads to what Lyons *et al.* [74] and others call size-biased Galton-Watson trees. A size-biased random variable is defined as follows [106]:

Definition. *Let X be a positive random variable with density f such that $\mathbb{E}X < \infty$, then the size-biased version, $\widehat{X}$ say, has density*

$$\widehat{f}(x) = \frac{xf(x)}{\mathbb{E}X}$$

If X is a discrete random variable with $\mathbb{P}(X = j) = p_j$, then

$$\mathbb{P}(\widehat{X} = j) = \frac{jp_j}{\mathbb{E}X}$$

Size-biased distributions occur naturally in a variety of situations. For example, if the random variables $X_1, X_2, \ldots$ represent the life times of components in a renewal process (and are independent and identically distributed), then the probability distribution of the life time of the component in use at some fixed time t is the size-biased distribution of X. This is the source of the well known *waiting-time* paradox.

The calculation of Equations (7.1) and (7.2) are for Galton-Watson trees begun with a single species, whereas we are interested in trees begun with two ancestors at time zero. The following result will be used to justify our simulation approach.

Lemma. *Let $X_1, X_2, \ldots, X_k$ be discrete independent positive random variables with $\mathbb{E}X_j = \lambda_j$, 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:*

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

where the random variable J has distribution

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

and is independent of the X_j.

This result is well known. However, I have been unable to find a proof of it and so for completeness I include one below.

Proof. Let Y be a positive discrete random variable with distribution $\mathbb{P}(Y = y) = p_y$ and probability generating function (pgf) $\phi_Y(u)$. Then the size-biased variable $\widehat{Y}$ has pgf

$$\phi_{\widehat{Y}}(u) = \sum_y \frac{yu^y p_y}{\mathbb{E}Y} = \frac{u}{\mathbb{E}Y}\frac{d}{du}\sum_y u^y p_y = \frac{u\phi'_Y(u)}{\phi'_Y(1)}. \tag{7.3}$$

The X_i are independent, and so we can write

$$\phi_S(u) = \mathbb{E}(u^{\sum X_i}) = \prod_{i=1}^{k} \phi_i(u)$$

where $\phi_i(u)$ denotes the pgf of X_i. Using Equation (7.3) we find that the pgf of the size-biased sum, $\widehat{S}$, is

$$\phi_{\widehat{S}}(u) = \frac{u\frac{d}{du}\prod \phi_i(u)}{\frac{d}{du}\prod \phi_i(u)|_{u=1}} = \frac{u\sum_j \phi_j'(u)\prod_{i\neq j}\phi_i(u)}{\sum \lambda_i}. \tag{7.4}$$

Now we find the pgf of the random variable $X_1 + \ldots + X_{J-1} + \widehat{X}_J + X_{J+1} + \ldots + X_k$ given in the lemma:

$$\begin{aligned}
\phi_{\widehat{X}_d}(u) = \mathbb{E}(u^{X_1+\ldots+\widehat{X}_J+\ldots+X_k}) &= \mathbb{E}(\mathbb{E}(u^{X_1+\ldots+\widehat{X}_j+\ldots+X_k}|J=j)) \\
&= \mathbb{E}\Big(\frac{u\phi_J'(u)}{\phi_J'(1)}\prod_{i\neq J}\phi_i(u)\Big) \\
&= \frac{u\sum_j \phi_j'(u)\prod_{i\neq j}\phi_i(u)}{\sum \lambda_i}
\end{aligned} \tag{7.5}$$

Observing that Equations (7.4) and (7.5) agree shows that $\widehat{S}$ and $X_1 + \ldots + X_{J-1} + \widehat{X}_J + X_{J+1} + \ldots + X_k$ do indeed have the same distribution. □

Remarks:

1. I have only stated the result for discrete random variables as that is all we shall require, however the result also holds for continuous random variables. The proof is similar, except Laplace transforms must be used rather than probability generating functions.

2. This lemma says that to simulate a size-biased Galton-Watson tree begun with two lineages, we can simulate two independent trees, one of which is a standard Galton-Watson tree and the other a size-biased Galton-Watson tree, both begun with one lineage:
$$\widehat{Z}(t) =_d X_1(t) + \widehat{X}_2(t).$$

We can use Equations (2.5) and (2.17) to calculate the distribution of $\widehat{X}$ and $\widehat{Z}$. We find that

$$\begin{aligned}
\mathbb{P}(\widehat{X}(t) = j) &= j(1-\eta)^2\eta^{j-1} \\
\mathbb{P}(\widehat{Z}(t) = j) &= j(1-\eta)^2\eta^{j-2}(\frac{1}{2}(1-\xi)(1-\eta)^{j-1} + \xi\eta).
\end{aligned}$$

As discussed in Chapter 2, our interest is in $\widetilde{Z}(t)$, the tree conditioned on non-extinction, rather than in $Z(t)$. Calculating the size-biased distribution of $\widehat{\widetilde{Z}}$ using Equation (2.18) gives

$$\mathbb{P}(\widehat{\widetilde{Z}}(t) = j) = \frac{1}{2}j(j-1)(1-\eta)^3\eta^{j-2}.$$

This allows us to note that the independence of different sides of the tree carries over to trees conditioned on non-extinction, i.e.,

$$\widehat{\widetilde{Z}}(t) =_d \widehat{X}(t) + \widetilde{X}(t).$$

A Problem

The calculations above give values of $\mathbb{P}(Z(t) = j|B(y))$. It is not clear whether knowing the distribution of the size process is sufficient to uniquely determine the structure of the tree. There could conceivably be various ways to simulate trees with this size law, but which lead to trees with different structures. Ideally, we would like a result similar to the one given in Geiger [46] which says something along the lines of

$$\mathbb{P}(\mathcal{T} \in \mathrm{d}t|B) = \frac{\mathbb{P}(\mathcal{T} \in \mathrm{d}t)Z_t(\mathcal{T})}{\mathbb{E}Z_t(\mathcal{T})},$$

where $\mathcal{T}$ represents the tree, $\mathrm{d}t$ a subset of the tree space, and $Z_t(\mathcal{T})$ the size of tree $\mathcal{T}$ at time t. This requires a measure space on the type of trees we are interested in (i.e., those extant at time t), and a measure for Galton-Watson trees. This is given in Section 7.3. In the next section, I give an approximate method for simulating size-biased random variables and trees.

7.2 Size-Biased Simulation Methods

In this section I give a couple of algorithms for simulating size-biased versions of a random variable, all of which depend upon being able to simulate observations from the original distribution, $F(\cdot)$. One option for sampling from $\widehat{F}(\cdot)$ is to use a simple Metropolis independence sampler:

1. Suppose we are currently at x. Propose a move to $y \sim F(\cdot)$
2. Accept the move with probability

$$r = \min\left(1, \frac{q(y \to x)\pi(y)}{q(x \to y)\pi(x)}\right) = \min\left(1, \frac{y}{x}\right).$$

As discussed in Chapter 5, MCMC algorithms often have a high start-up cost as we must wait for convergence. Thus, in situations where only a single observation is required from

each distribution, this approach will be too slow.

An alternative approach is to use a rejection-based algorithm. Suppose $X \leq C$ almost surely (a.s.). Then the following rejection algorithm gives observations from the size-biased distribution:

1. Draw $X \sim F(\cdot)$.
2. Accept X with probability $r = \frac{X}{C}$.

The proof that this algorithm gives draws from the size-biased distribution $\widehat{F}(\cdot)$ is essentially identical to those given in Section 3.1, and so is not reproduced here. The condition that $X \leq C$ a.s. will not be true in many cases of interest, including the conditional Galton-Watson trees considered earlier. In these cases we can either resort to using MCMC (which requires no such conditions) or we can use an approximate form of the rejection method. The following algorithm is just one such approximate method:

1. Simulate $X \sim F(\cdot)$
2. Accept X with probability $r = \min(1, \frac{X}{C})$.

The acceptance rate of this algorithm, assuming that F admits a density f, is

$$\begin{aligned} A = \mathbb{P}(I=1) &= \int_{x:x\leq C} \frac{x}{C} f(x)\mathrm{d}x + \int_{x:x>C} f(x)\mathrm{d}x \\ &= \frac{1}{C}\left(\int_0^\infty x f(x)\mathrm{d}x - \int_{x:x>C} (x-C) f(x)\mathrm{d}x\right) \\ &= \frac{1}{C}(\mathbb{E}X - \epsilon_C) \end{aligned}$$

where $\epsilon_C = \mathbb{E}(X-C)^+ = \int_{x:x>C}(x-C)\, f(x)\mathrm{d}x$. Let $\hat{f}_1$ denote the probability density function of observations from this algorithm. Then

$$\hat{f}_1(x) = \begin{cases} \frac{xf(x)}{CA} & \text{if } x \leq C \\ \frac{f(x)}{A} & \text{if } x > C \end{cases}$$

We can calculate the accuracy of this approximation by calculating the total variation distance between $\widehat{f}$ and $\hat{f}_1$:

$$\begin{aligned} d_{TV}(\widehat{f}, \hat{f}_1) &= \frac{1}{2}\int_0^\infty |\widehat{f}(x) - \hat{f}_1(x)|\mathrm{d}x \\ &= \frac{1}{2}\left(\int_{x:x\leq C}\left|\frac{xf(x)}{CA} - \frac{xf(x)}{\mathbb{E}X}\right|\mathrm{d}x + \int_{x:x>C}\left|\frac{f(x)}{A} - \frac{xf(x)}{\mathbb{E}X}\right|\mathrm{d}x\right) \\ &= \frac{1}{2}\left(\frac{\epsilon_C}{\mathbb{E}X - \epsilon_C}\mathbb{P}(\widehat{X}\leq C) + \int_{x:x>C} f(x)\left|\frac{x\epsilon_C - \mathbb{E}X(x-C)}{(\mathbb{E}X-\epsilon_C)\mathbb{E}X}\right|\mathrm{d}x\right) \end{aligned}$$

It does not seem possible to evaluate this integral, but we can use the triangle inequality to bound the total variation distance.

$$\begin{aligned} d_{TV}(\widehat{f}, \hat{f}_1) &\leq \frac{1}{2}\left(\frac{\epsilon_C}{\mathbb{E}X - \epsilon_C} \cdot \mathbb{P}(\widehat{X} \leq C) + \int_{x:x>C} f(x)\left(\frac{x\epsilon_C + \mathbb{E}X(x-C)}{(\mathbb{E}X - \epsilon_C)\mathbb{E}X}\right) \mathrm{d}x\right) \\ &= \frac{1}{2}\left(\frac{\epsilon_C}{\mathbb{E}X - \epsilon_C} \cdot \mathbb{P}(\widehat{X} \leq C) + \frac{\epsilon_C}{\mathbb{E}X - \epsilon_C} + \frac{\epsilon_C}{\mathbb{E}X - \epsilon_C} \cdot \mathbb{P}(\widehat{X} > C)\right) \\ &= \frac{\epsilon_C}{\mathbb{E}X - \epsilon_C} \end{aligned}$$

By controlling the size of C we can control the size of the error for this approximation. But, as usual, the more accurate we wish to be, the longer we will need to run the computations. So a potential method for simulating size-biased Galton-Watson trees is as follows:

1. Simulate a standard Galton-Watson tree until time y.
2. With probability $\min(1, Z(y)/n)$ accept the tree. If rejected, return to step 1.
3. Pick a random species from those alive at time y and replace it with two new species.
4. Simulate the tree into the future as required.

By calculating that

$$\begin{aligned} \epsilon_C &= \sum_{n=C}^{\infty} (n - C)\mathbb{P}(Z(t) = n) | Z(y) = 2) \\ &= \left(\frac{1-\xi}{1-\eta}\right) \eta^{C-1}(C(1-\eta)(1-\xi) + 2\eta) \end{aligned}$$

we can choose n large enough to guarantee that the error is less than some predetermined value.

There are several problems with this approach. Firstly, it is too slow. If we require a total variation distance of less than 0.01, this algorithm is about two orders of magnitude slower than the method of previous chapters. Secondly, for the reason stated at the end of the previous section, it is not clear that we are generating the correct distribution with this algorithm. In the next section I give a description of the fish-bone like structure of conditioned trees, which will then allow us to use a more efficient simulation approach.

7.3 The Fish-Bone Process

Size-biased Galton-Watson trees have been considered in the past by several other authors: Lyons, Pemantle and Peres [74] used size-biased trees to give conceptual proofs of various

$L \log L$ criteria[2] for discrete time Galton-Watson trees; Geiger [45, 46] used size-biased trees to give a Poisson process representation of critical continuous time Galton-Watson trees and to prove various limit laws; Chauvin, Rouault, and Wakolbinger [28] considered Markovian branching particle systems in $\mathbb{R}^d$ conditioned to be in a specific location at a given time, and found that this conditioning leads to size-biased tree structures.

All of these authors found that size-biased trees have what I will call a *fish-bone* structure. Geiger [46] gave the structure for critical size-biased Galton-Watson trees to be as follows: the initial ancestor (the root of the tree) lives for an Exp(λ) period of time before giving birth to a size-biased number of offspring, $\widehat{L}$. One of these offspring is chosen at random and after another Exp(λ) period it then gives birth to a size-biased number of offspring, and so on. This defines a central *spine*, or *distinguished path* of individuals from the root to time t. Offspring not part of this spine give rise to a standard number of offspring, L, and evolve as independent Galton-Watson trees. See Figure 7.1 for an illustration of this structure. Notice that because the offspring distribution on the distinguished path is size-biased, the tree cannot become extinct. Geiger proved that given a size-biased tree with at least two lineages at time t, the distinguished path cannot be recovered, and that the final member of the distinguished path is uniformly distributed across all lineages extant at time t.

It is clear that if we are able to find a similar construction for trees conditioned on having a birth at time y, our simulations will be much more efficient than those based on the rejection algorithm. Unfortunately however, the arguments given by other authors are not readily adaptable to our scenario. The method of Chauvin *et al.* [28] most closely matches our requirements, and the notation and measure space I use is based on their approach. However, they consider a different type of conditioning (conditioning particles in a branching particle system to occupy a given place at a given time) and only consider the homogeneous case, whereas we require trees conditioned to have a birth at a given time and are interested in the inhomogeneous case where the offspring probabilities $\{p_k(t)\}_{k=0,1,2,\ldots}$ are functions of time.

7.3.1 A Measure Space and the Galton-Watson Distribution

I now give a description of a tree sample space, a σ-algebra on this space and a description of the Galton-Watson measure on this σ-algebra. The description is fairly technical due to the amount of information that we are required to keep track of. The space must describe the tree topology and the lengths of each branch. Because the birth probabilities $p_k(t)$ are inhomogeneous, the space must also record when in time each tree originated. I shall also describe a pruning operator which will allow subtrees to be pruned from the main tree and

[2] Throughout this section, L denotes the number of offspring of a given lineage, and has distribution $\{p_k(t)\}_{k=0,1,\ldots}$. A *critical tree* has $\mathbb{E}L = 1$, a *subcritical tree* $\mathbb{E}L < 1$, and a *supercritical tree* $\mathbb{E}L > 1$.

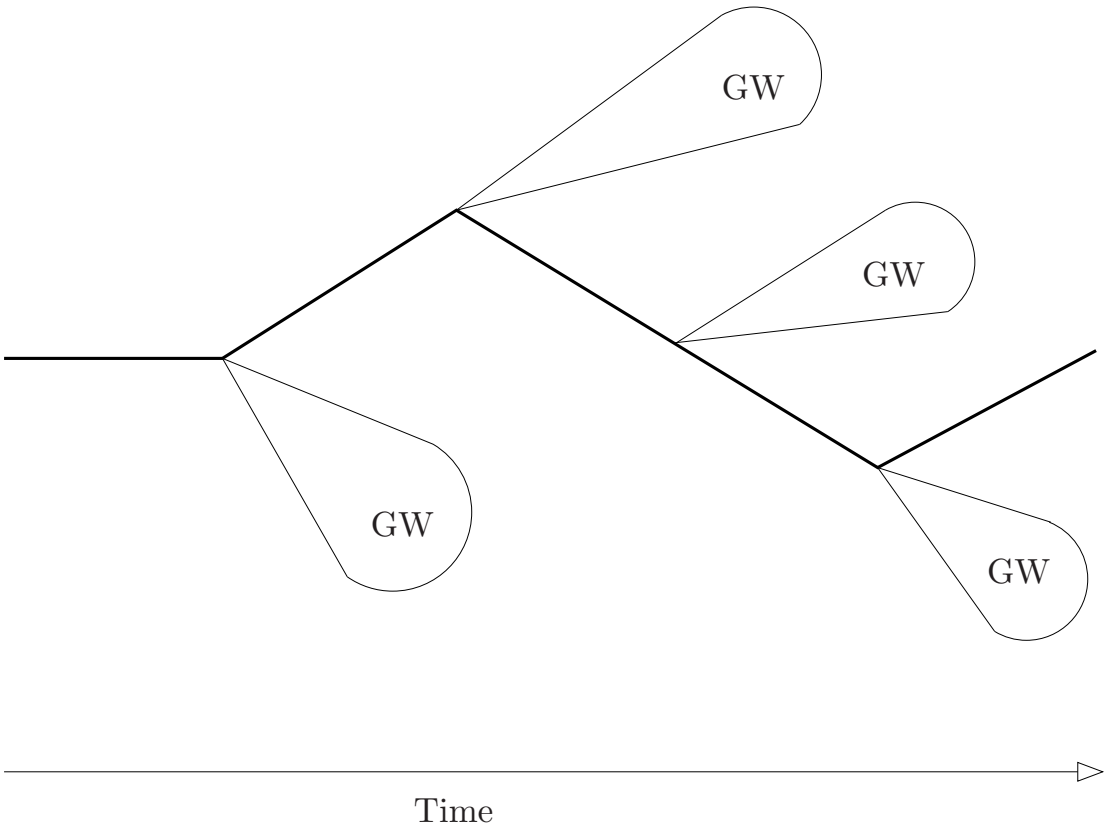

Figure 7.1: *The fish-bone structure of size-biased trees. The spine is shown by the thick lineage down the centre. Offspring distributions on this spine are size-biased. GW denotes the standard Galton-Watson trees that branch off the central spine.*

considered as separate members of the tree space which will be useful when conditioning on a birth at a particular point in time.

Firstly, we need a notation that allows us to describe the tree topology. Let Ø denote the root of the tree and let $\nu_\emptyset$ be the number of offspring, which we label as $1, 2, \ldots, \nu_\emptyset$. We then denote the offspring of individual k as $k1, k2, \ldots, k\nu_k$ and so on. More formally, let $U = \{\emptyset\} \cup \bigcup_{n=1}^{\infty} \mathbb{N}^n$ be the space of individuals, where $\mathbb{N}^n$ represents the Cartesian product of $\mathbb{N}$ with itself n times.

Following Neveu [88], we call a subset $\tilde{\omega} \subseteq U$ a tree if

1. $\emptyset \in \tilde{\omega}$ (the root belongs to the tree).

2. For all $u, v \in \tilde{\omega}, (uv \in \tilde{\omega}) \Rightarrow (u \in \tilde{\omega})$ (the ancestor u of the individual uv belongs to the tree).

3. For all $u \in \tilde{\omega}$ there exists $\nu_u \in \mathbb{N} \cup \{0\}$ such that $uj \in \tilde{\omega}$ for $j = 1, 2, \ldots \nu_u$ (ν_u is the number of offspring of individual u and $\{uj\}_{j \leq \nu_u}$ are its offspring).

Let $\tilde{\Omega}$ denote the set of trees described above. Denote by $\tilde{\Omega}_u$ the subset of $\tilde{\Omega}$ which contains only those trees containing u:

$$\tilde{\Omega}_u = \{\tilde{\omega} \in \tilde{\Omega} : u \in \tilde{\omega}\}$$

As we are using the continuous time Galton-Watson process, we need to include information about the life duration of each species which is represented in the tree by the length of each branch. We define the *branch length* information for $\tilde{\omega}$ to be a collection $(\sigma_u : u \in \tilde{\omega})$ of values in $\mathbb{R}^+$. We label the birth time of the root as $s_\emptyset$ where $s_\emptyset$ takes values in $\mathbb{R}$. We then define a continuous-time tree to be the triple

$$\omega = (\tilde{\omega}, (\sigma_u : u \in \tilde{\omega}), s_\emptyset)$$

and let Ω denote the set of all such trees. We can define the subset of Ω containing only those trees which contain u by letting p represent the canonical projection of Ω into $\tilde{\Omega}$, such that $p(\omega) = \tilde{\omega}$ and then letting $\Omega_u = p^{-1}(\tilde{\Omega}_u)$. The values σ_u and ν_u can be considered mappings with σ_u the branch length mapping $\sigma_u : \Omega_u \to \mathbb{R}^+$, and ν_u the offspring mapping $\nu_u : \Omega_u \to \mathbb{N} \cup \{0\}$. A couple of other mappings that will be required are the *birth* and *death* time mappings, defined on Ω_u as

$$\begin{aligned}
\text{Birth Mapping: } & S_u : \Omega_u \to \mathbb{R} \\
& \text{s.t. } \; S_u = S_v + \sigma_v \text{ and } \quad S_\emptyset = s_\emptyset \\
\text{Death Mappling: } & D_u : \Omega_u \to \mathbb{R} \\
& \text{s.t. } \; D_u = D_v + \sigma_u \text{ and } D_\emptyset = s_\emptyset + \sigma_\emptyset \\
& \qquad \text{where } v \text{ is } u\text{'s parent}
\end{aligned}$$

This completes the description of the space Ω of continuous time branching trees. For notational convenience, from this point on I let $s = s_\emptyset$, $\sigma = \sigma_\emptyset$ and $\nu = \nu_\emptyset$ represent the birth time, branch length, and number of offspring of the root.

I now define a σ-algebra on Ω. The choice needs to be fine enough to ensure that all the subsets of interest are included. For example, it must include the subset of trees which have births at time y. This is achieved by letting $\mathscr{F}$ be the σ-algebra on Ω generated by $S_\emptyset$ and $\{\Omega_u, \sigma_u\}$ for all $u \in U$. This ensures that the mappings D_u, S_u and σ_u are all measurable functions from $(\Omega, \mathscr{F})$ to $(\mathbb{R}, \mathscr{B})$ or $(\mathbb{R}^+, \mathscr{B}^+)$ where $\mathscr{B}$ (or $\mathscr{B}^+$) is the Borel σ-algebra on $\mathbb{R}$ (or $\mathbb{R}^+$). Together, the double $(\Omega, \mathscr{F})$ form a measurable space.

Because we are considering trees that evolve through time, it will be useful to have a filtration $\{\mathscr{F}_t\}_{t\in\mathbb{R}}$ on $\mathscr{F}$. A filtration is an increasing sequence of σ-algebras (i.e., $\mathscr{F}_t$ is a σ-algebra for all t and if $s \leq t$, then $\mathscr{F}_s \subseteq \mathscr{F}_t \subseteq \mathscr{F}$). In order to define a filtration we shall need the *killing operator* M_t. For $t \in \mathbb{R}^+$, let $M_t : \Omega \to \Omega$ be the killing operator

defined by

$$\begin{aligned} p(M_t(\omega)) &= \{u \in p(\omega) : S_u(\omega) - S_\emptyset(\omega) \leq t\} \\ \sigma_u(M_t(\omega)) &= \sigma_u(\omega) \\ S_\emptyset(M_t(\omega)) &= S_\emptyset(\omega) \end{aligned}$$

so that $M_t(\omega)$ only includes those individuals in the tree ω which were born before time $s_\emptyset + t$. This allows us to define the filtration $\mathscr{F}_t = M_t(\mathscr{F})$ on $\mathscr{F}$. The σ-algebra $\mathscr{F}_0$ then represents the knowledge contained by the root only, namely ν, σ and $s_\emptyset$.

Before we can define the Galton-Watson probability distribution (i.e., measure) on $(\Omega, \mathscr{F})$, we need one final mapping, the *pruning operator* T_u. For a tree $\omega \in \Omega$, $T_u(\omega)$ is the subtree which has root u and birth-time equal to the birth-time of u. Formally, for all $u \in U$, let $T_u : \Omega_u \to \Omega$, with behaviour defined by

$$\begin{aligned} p(T_u(\omega)) &= \{v \in U : uv \in p(\omega)\} \\ \sigma_v(T_u(\omega)) &= \sigma_{uv}(\omega) \\ S_\emptyset(T_u(\omega)) &= S_u(\omega). \end{aligned}$$

We are now finally able to consider the probability distribution of Galton-Watson trees. Neveu [88] proved that there exists a unique family of probability distributions $\{P_s\}_{s\in\mathbb{R}}$ on $(\Omega, \mathscr{F})$ such that

(1) $P_s(\{\omega \in \Omega : S_\emptyset \neq s\}) = 0$, i.e., P_s only gives positive probability to trees which originate at time s.

(2) Under P_s, the root has length $\sigma \sim \text{Exp}(\lambda)$, and conditional on a death taking place at time t, the root has ν offspring with distribution $(p_k(t) : k = 0, 1, \ldots)$.

(3) The branching property holds: conditional on knowing σ and ν, the subtrees formed by the offspring, $T_1, T_2, \ldots T_\nu$, are independent and identically distributed Galton-Watson trees, with distribution $P_{s+\sigma}$.

Conditions (2) and (3) describe the structure we have been using throughout this thesis.

The branching property can be expressed mathematically by looking at expectations of measurable functions: if $\{f_i\}_{i\in\mathbb{N}}$ is a sequence of non-negative measurable functions $f_i : \Omega \to \mathbb{R}$, then

$$\mathbb{E}_s(\prod_{i=1}^{\nu} f_i \circ T_i | \mathscr{F}_\sigma) = \prod_{i=1}^{\nu} \mathbb{E}_{s+\sigma}(f_i) \tag{7.6}$$

where $\mathbb{E}_s$ denotes expectation with respect to probability measure P_s, i.e.,

$$\mathbb{E}_s f = \int_\Omega f(\omega) P_s(\mathrm{d}\omega).$$

This illustrates a result I shall rely upon later, namely that two measures μ_1 and μ_2 agree on $(\Omega, \mathscr{F})$ if and only if for all non-negative measurable functions f the equality $\int f d\mu_1 = \int f d\mu_2$ holds.

7.3.2 The Fish-Bone Theorem

We are now in a position to state the main result of this chapter:

Theorem 1. *The Galton-Watson process, begun at time s, and conditioned on having a death at time $y > s$, has the following properties:*

(i) σ has an inhomogeneous exponential $(\lambda m(\cdot))$ distribution truncated at y. That is, it has probability density function

$$g_s(\sigma) = \lambda m(s+\sigma) e^{-\lambda \int_s^{s+\sigma} m(t)\mathrm{d}t} . \mathbb{I}_{s+\sigma<y} + e^{-\lambda \int_s^{y} m(t)\mathrm{d}t} . \delta_{y-s}(\sigma) \tag{7.7}$$

where $m(t) = \sum_{k=0}^{\infty} k p_k(t)$, is the mean of the offspring distribution, L, at time t.

(ii) If $s + \sigma < y$ (first death is before time y)

(a) ν, the number of offspring of the distinguished species, has a $\left\{ \frac{k p_k(\sigma+s)}{m(\sigma+s)} \right\}_{k=1,2,\ldots}$ distribution,

(b) one of the subtrees, T_j say, is choosen uniformly from $\{T_1, \ldots, T_\nu\}$. T_j is then a Galton-Watson tree born at time $s+\sigma$ conditioned to have a split point at time y. The other subtrees, T_i for $i \neq j$, are independent identically distributed Galton-Watson trees born at time $s+\sigma$, i.e., they have distribution $P_{s+\sigma}$.

(iii) If $s + \sigma = y$ (first death is at time y), then ν has the usual $(p_k(y) : k = 0, 1, \ldots)$ offspring distribution. The offspring subtrees, $T_1, \ldots, T_\nu$, all evolve as independent identically distributed Galton-Watson trees begun at time y.

Remarks:

1. Part (ii)(a) of the theorem shows that lineages on the distinguished path have a size-biased number of offspring, $\widehat{L}$.

2. The distribution of lifetimes on the distinguished path is what I have called a truncated inhomogeneous exponential distribution. Suppose that $m(t)$ is a positive integrable function with $\int_s^\infty m(t)\mathrm{d}t = \infty$ for all s. We say that X has an inhomogeneous exponential distribution with rate $m(\cdot)$ started at time s if

$$\mathbb{P}_s(X > x) = \exp\Big(- \int_s^{s+x} m(t)\mathrm{d}t \Big),$$

and we write $X \sim \text{Exp}_s(m(\cdot))$. Differentiating $1 - \mathbb{P}_s(X > x)$ with respect to x, truncating at time y and adding an atom at this point so that the density integrates to one, leads to the density $g_s(\sigma)$ given by Equation (7.7) in part (i) of the theorem. Here, $\delta_\cdot(\cdot)$ denotes the Dirac delta function, which is defined by its behaviour when integrated:

$$\int_A \delta_x(y)\mathrm{d}y = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

3. This theorem describes Galton-Watson trees conditioned on having a death at time y. To condition on a having a birth at time y, simply change the offspring distribution in step (*iii*) of the theorem so that ν has offspring distribution $\{p_k(y)/(1 - p_0(y))\}$ for $k = 1, 2, \ldots$.

The proof of the theorem works by constructing a random point measure on $\mathbb{R}$ with atoms at the death times of the species in the tree. In this way, each tree ω defines a point measure on $\mathbb{R}$ (see Figure 7.2). We then look at the distribution of the point measure conditioned on it having an atom (i.e., a death) at a fixed time y. The conditioning is done by using the Palm approach to random point measures (see Kallenberg [63] or Daley *et al.* [36] for example).

Let M denote a random point measure on $\mathbb{R}$ with

$$\begin{aligned} M &: \Omega \times \mathbb{R} \to \mathbb{N} \cup \{0\} \\ \text{s.t. } M(\omega, \mathrm{d}y) &= \sum_{u:\ u \in \omega} \delta_{D_u(\omega)}(\mathrm{d}y). \end{aligned}$$

An intuitive way to think of M is as follows: Tyche, the Goddess of Chance, picks a random tree ω from Ω according to the law P_s. This then defines the point measure $M(\omega, \cdot)$ on $\mathbb{R}$ with atoms at the death times of individuals in the tree. A point measure is simply a measure which counts the number of points in a set. This description treats M as a map from $\Omega \to \mathcal{N}$, where $\mathcal{N}$ is the set of integer valued measures on $\mathbb{R}$. Alternatively, we can view $\{M(A)\}_{A \in \mathscr{B}}$ as a collection of integer valued random variables (i.e., maps from Ω to $\mathbb{N} \cup \{0\}$).

The Campbell measure, C_s, is a product measure on $(\Omega \times \mathbb{R}, \mathscr{F} \otimes \mathscr{B})$ where $\otimes$ denotes the product σ-algebra. It is defined as

$$C_s(\mathrm{d}\omega \times \mathrm{d}y) = P_s(\mathrm{d}\omega)M(\omega, \mathrm{d}y).$$

If f is any $\mathscr{B}$ measurable function $f : \mathbb{R} \to \mathbb{R}$, and g is a $\mathscr{F}$ measurable function $g : \Omega \to \mathbb{R}$,

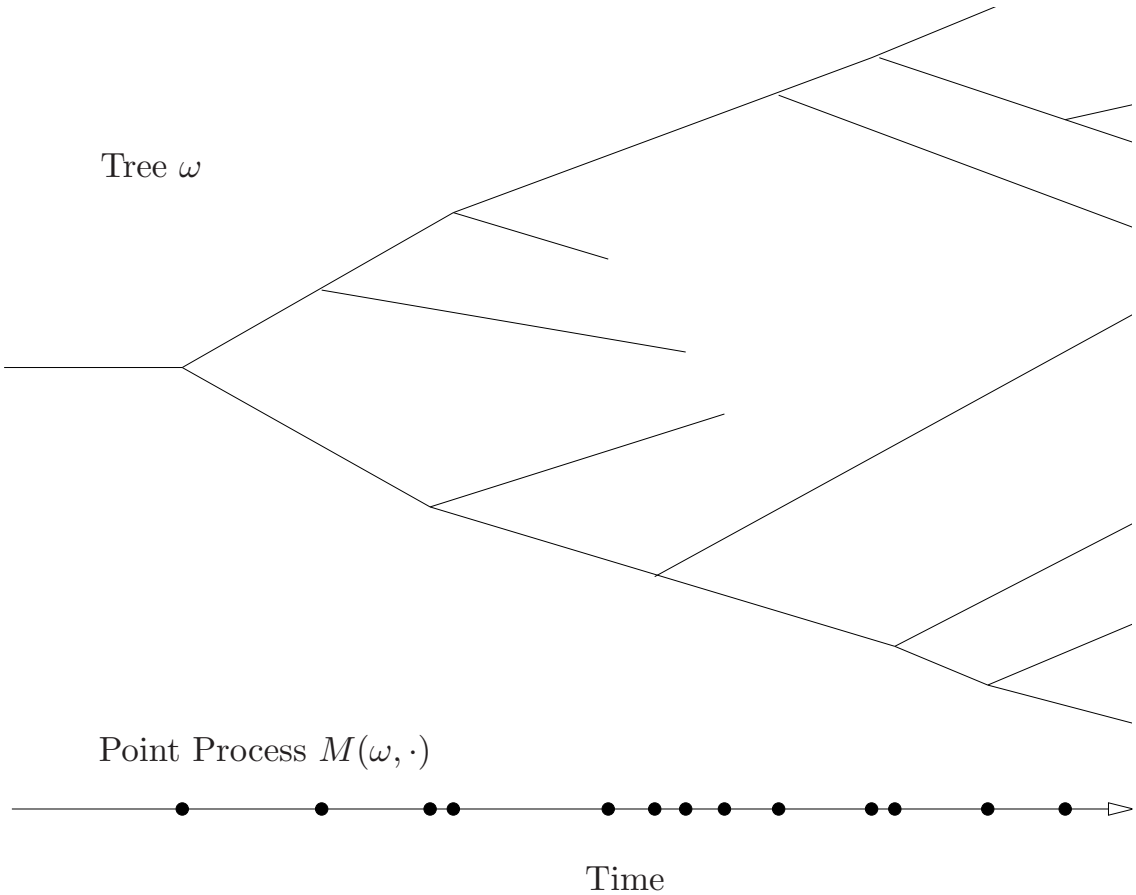

Figure 7.2: *Point process representation of a tree. Every death in the tree, ω, is represented as an atom in the point process $M(\omega)$.*

then

$$\begin{aligned}\int_{\Omega\times\mathbb{R}} f(y)g(w)C_s(\mathrm{d}\omega\times\mathrm{d}y) &= \int_{\Omega\times\mathbb{R}} f(y)g(\omega)P_s(\mathrm{d}\omega)M(\omega,\mathrm{d}y)\\ &= \int_{\Omega\times[s,\infty)} f(y)g(\omega)P_s(\mathrm{d}\omega)M(\omega,\mathrm{d}y),\end{aligned}$$

where the restriction from $\mathbb{R}$ to $[s,\infty)$ can be assumed because of condition (1) in the description of the Galton-Watson measure. The Palm distributions on Ω, $\mathcal{Q}^s_y$ for $y > s$, are found by a disintegration of C_s with respect to its first marginal. Theorem 15.3.3 in Kallenberg [63] guarantees that we can decompose the σ-finite measure C_s on $\Omega\times\mathbb{R}$ into a σ-finite measure $v_s(\mathrm{d}y) = \mathbb{E}_s M(\mathrm{d}y)$ on $\mathbb{R}$ and a kernel $\mathcal{Q}^s_\cdot(\cdot)$ from $\mathbb{R}$ to Ω such that

$$\int_{\Omega\times[s,\infty)} f(y)g(\omega)P_s(\mathrm{d}\omega)M(\omega,\mathrm{d}y) = \int_{\Omega\times[s,\infty)} f(y)g(\omega)\mathcal{Q}^s_y(\mathrm{d}\omega)\mathbb{E}_s M(\mathrm{d}y) \tag{7.8}$$

Remarks:

1. The Palm distributions $\mathcal{Q}^s_y$ can be thought of as the conditional distributions of the

point measure M given that there is an atom at time y. In terms of the branching process, this is the conditional distribution of a Galton-Watson process given that there is a death at time y.

2. To say $\mathcal{Q}^s_\cdot(\cdot)$ is a kernel means that $\mathcal{Q}^s_y(\mathrm{d}\omega)$ is a probability measure on Ω for all $y > s$ and $\mathcal{Q}^s_\cdot(A)$ is a $\mathscr{B}$ measurable function on $\mathbb{R}$ for all $A \in \mathscr{F}$. The existence and almost sure uniqueness of $(\mathcal{Q}^s_y)_{y>s}$ are consequences of the Radon-Nikodým Theorem.

3. We can calculate $\mathbb{E}_s M(\mathrm{d}y)$ as follows:

$$\begin{aligned}
\mathbb{E}_s M(\mathrm{d}y) &= \sum_{n=1}^{\infty} \mathbb{E}(\text{number of deaths in } (y, y+\mathrm{d}y) | M_y = n)\mathbb{P}(M_y = n) \\
&= \sum_{n=1}^{\infty} (\lambda n.\mathrm{d}y + o(\mathrm{d}y))(1-\xi(s,y))(1-\eta(s,y))\eta(s,y)^{n-1} \text{ by Equation (2.5)} \\
&= \frac{1-\xi(s,y)}{1-\eta(s,y)}\lambda.\mathrm{d}y \\
&= \lambda e^{\lambda \int_s^y (m(t)-1)\mathrm{d}t}.\mathrm{d}y \qquad \text{by Equation (2.9)}
\end{aligned}$$

4. An alternative way to write Equation (7.8) is to split the integration over the product space and to write the increment with the integral

$$\begin{aligned}
\int_{\Omega\times[s,\infty)} f(y)g(\omega)\mathcal{Q}^s_y(\mathrm{d}\omega)\mathbb{E}_s M(\mathrm{d}y) &= \int_s^\infty f(y)\left[\int_\Omega g(\omega)\mathcal{Q}^s_y(\mathrm{d}\omega)\right]\mathbb{E}_s M(\mathrm{d}y) \\
&= \int_s^\infty \mathbb{E}_s M(\mathrm{d}y) f(y) \int_\Omega \mathcal{Q}^s_y(\mathrm{d}\omega)g(\omega)
\end{aligned}$$

The following two results shall be required in the proof of the theorem:

$$e^{-\lambda\theta}\mathbb{E}_{s+\theta}M(\mathrm{d}y) = e^{-\lambda\int_s^{s+\theta} m(t)\mathrm{d}t}.\mathbb{E}_s M(\mathrm{d}y) \tag{7.9}$$

$$e^{-\lambda(y-s)} = e^{-\lambda\int_s^y m(t)\mathrm{d}t}.\mathbb{E}_s M(\mathrm{d}y) \tag{7.10}$$

The crux of the problem is to identify the Palm probabilities $\mathcal{Q}^s_y$ for all $y > s$. Measures are uniquely determined by their behaviour when integrating measurable functions. The σ-algebra $\mathscr{F}$ is generated by functions of the form

$$f_0(\omega)\mathbb{I}_{\nu(\omega)=k}\prod_{i=1}^{\nu(\omega)} f_i \circ T_i(\omega) \tag{7.11}$$

where f_0 is $\mathscr{F}_\sigma$ measurable, and the f_i are $\mathscr{F}$ measurable. To prove the theorem it is sufficient to show that $\mathcal{Q}^s_y$ behaves on functions of this form in the way described by the

theorem.

Proof of Theorem 1. Under the Galton-Watson distribution P_s, we can decompose M into its subtrees, and an atom at the time of the first death:

$$M(\omega, \mathrm{d}y) = \sum_{j=1}^{\nu(\omega)} M(T_j(\omega), \mathrm{d}y)\mathbb{I}_{y>s+\sigma(\omega)} + \delta_{s+\sigma(\omega)}(\mathrm{d}y) \tag{7.12}$$

Then on functions of the form (7.11) multiplied by Borel measurable $f(\cdot)$

$$\begin{aligned}
I &= \int_{\Omega\times[s,\infty)} f(y) f_0(\omega)\mathbb{I}_{\nu(\omega)=k} \prod_{i=1}^{\nu(\omega)} (f_i \circ T_i)(\omega) P_s(\mathrm{d}\omega) M(\omega, \mathrm{d}y) \\
&= \int_{\Omega\times[s,\infty)} f(y) f_0 \mathbb{I}_{\nu=k} \prod_{i=1}^{\nu} (f_i \circ T_i) \left(\sum_{j=1}^{\nu} M(T_j(\omega), \mathrm{d}y)\mathbb{I}_{y>s+\sigma(\omega)} + \delta_{s+\sigma(\omega)}(\mathrm{d}y) \right) P_s(\mathrm{d}\omega) \\
&= \mathbb{E}_s \Bigg[f_0 \mathbb{I}_{\nu=k} \Bigg(\sum_{j=1}^{k} \prod_{\substack{i=1 \\ i\neq j}}^{k} (f_i \circ T_i) \times (f_j \circ T_j) \int_s^\infty f(y)\mathbb{I}_{y>s+\sigma(\omega)} M(T_j, \mathrm{d}y) \\
&\qquad\qquad + \prod_{i=1}^{k} (f_i \circ T_i) \int_s^\infty f(y)\delta_{s+\sigma}(\mathrm{d}y) \Bigg) \Bigg].
\end{aligned}$$

Then, using the tower property of expectation $\mathbb{E}_s(f) = \mathbb{E}_s(\mathbb{E}_s(f|\mathscr{F}_\sigma))$, we can see that the integral above is equal to

$$\begin{aligned}
&= \mathbb{E}_s \Bigg[p_k(s+\sigma) f_0 \mathbb{E}_s \Bigg(\sum_{j=1}^{k} \prod_{\substack{i=1 \\ i\neq j}}^{k} (f_i \circ T_i) \times (f_j \circ T_j) \int_s^\infty f(y)\mathbb{I}_{y>s+\sigma(\omega)} M(T_j, \mathrm{d}y) \\
&\qquad\qquad + \prod_{i=1}^{k} (f_i \circ T_i) \int_s^\infty f(y)\delta_{s+\sigma}(\mathrm{d}y)|\mathscr{F}_\sigma \Bigg) \Bigg] \\
&\qquad\qquad \text{as } \nu \text{ and } f_0 \text{ are } \mathscr{F}_\sigma \text{ measurable} \\
&= \mathbb{E}_s \Bigg[p_k(s+\sigma) f_0 \Bigg(\sum_{j=1}^{k} \prod_{\substack{i=1 \\ i\neq j}}^{k} \mathbb{E}_{s+\sigma}(f_i) \times \mathbb{E}_{s+\sigma}(f_j \int_s^\infty f(y)\mathbb{I}_{y>s+\sigma} M(\mathrm{d}y)) \\
&\qquad\qquad + \prod_{i=1}^{k} \mathbb{E}_{s+\sigma}(f_i) \times \mathbb{E}_{s+\sigma}(\int_s^\infty f(y)\delta_{s+\sigma}(\mathrm{d}y)) \Bigg) \Bigg]
\end{aligned} \tag{7.13}$$

by the branching property (7.6) and treating $M(\mathrm{d}y)$ as a random point measure on $\mathbb{R}$. Looking at the right hand term on both lines, and using the decomposition (7.8) and the

properties of the Dirac-delta function, we can see that

$$\begin{aligned}\mathbb{E}_{s+\sigma}(f_j \int_s^\infty f(y)\mathbb{I}_{y>s+\sigma}M(\mathrm{d}y)) &= \int_{\Omega\times[s,\infty)} f_j(\omega)f(y)P_{s+\sigma}(\mathrm{d}\omega)\mathbb{I}_{y>s+\sigma}M(\omega,\mathrm{d}y)\\ &= \int_{\Omega\times[s,\infty)} f_j(\omega)f(y)\mathcal{Q}_y^{s+\sigma}(\mathrm{d}\omega)\mathbb{I}_{y>s+\sigma}\mathbb{E}_{s+\sigma}M(\mathrm{d}y)\\ \mathbb{E}_{s+\sigma}(\int_s^\infty f(y)\delta_{s+\sigma}(\mathrm{d}y)) &= f(s+\sigma)\end{aligned}$$

Note that σ is considered a constant in the three lines above as they appear within an expectation in Equation (7.13). Hence

$$I = \mathbb{E}_s\Bigg[p_k(s+\sigma)f_0\Bigg(\sum_{j=1}^k \prod_{\substack{i=1\\ i\neq j}}^k \mathbb{E}_{s+\sigma}(f_i) \times \int_\Omega\int_s^\infty f_j(\omega)f(y)\mathcal{Q}_y^{s+\sigma}(\mathrm{d}\omega)\mathbb{I}_{y>s+\sigma}\mathbb{E}_{s+\sigma}M(\mathrm{d}y))$$
$$+\prod_{i=1}^k \mathbb{E}_{s+\sigma}(f_i)\times f(s+\sigma)\Bigg)\Bigg]$$

By making explicit the first expectation, and writing the increment with the integrand (i.e., write $\int_{B\times A} f(x,y)\mathrm{d}x\times\mathrm{d}y = \int_A \mathrm{d}y\int_B \mathrm{d}x f(x,y)$ so that it is clear what we are integrating over), we can see that I is equal to

$$\int_0^\infty \mathrm{d}\theta\lambda e^{-\lambda\theta}p_k(s+\theta)f_0(\theta,k)\Bigg(\sum_{j=1}^k \prod_{\substack{i=1\\ i\neq j}}^k \mathbb{E}_{s+\theta}(f_i)\int_\Omega \mathcal{Q}_y^{s+\theta}(\mathrm{d}\omega)f_j\int_s^\infty \mathbb{E}_{s+\theta}M(\mathrm{d}y)f(y)\mathbb{I}_{y>s+\theta}$$
$$+\prod_{i=1}^k \mathbb{E}_{s+\theta}(f_i)\times f(s+\theta)\Bigg)$$

Then, using Fubini's Theorem to justify changing the order of integration, and using Equation (7.9), we can see that I equals

$$\int_s^\infty \mathbb{E}_s M(\mathrm{d}y)f\int_0^{y-s}\mathrm{d}\theta\lambda e^{-\lambda\int_s^{s+\theta}m(t)\mathrm{d}t}p_k(s+\theta)f_0(\theta,k)\sum_{j=1}^k\prod_{\substack{i=1\\ i\neq j}}^k \mathbb{E}_{s+\theta}(f_i)\int_\Omega \mathcal{Q}_y^{s+\theta}(\mathrm{d}\omega)f_j$$
$$+\int_s^\infty \mathrm{d}y\lambda e^{-\lambda(y-s)}p_k(y)f_0(y-s,k)\prod_{i=1}^k \mathbb{E}_y(f_i)\times f(y)$$

where the final term comes from using the substitution $y = s+\theta$. Using Equation (7.10)

and rearranging, we can see that

$$I = \int_s^\infty \mathbb{E}_s M(\mathrm{d}y) f(y) \left[\int_0^{y-s} \mathrm{d}\theta \lambda m(s+\theta) e^{-\lambda \int_s^{s+\theta} m(t)\mathrm{d}t} \frac{k p_k(s+\theta)}{m(s+\theta)} f_0(\theta, k) \right.$$
$$\times \frac{1}{k} \sum_{j=1}^{k} \prod_{\substack{i=1 \\ i \neq j}}^{k} \mathbb{E}_{s+\theta}(f_i) \int_\Omega \mathcal{Q}_y^{s+\theta}(\mathrm{d}\omega) f_j$$
$$\left. + \lambda e^{-\lambda \int_s^y m(t)\mathrm{d}t} p_k(y) f_0(y-s, k) \prod_{i=1}^{k} \mathbb{E}_y(f_i) \right].$$

Comparing this expression with Equation (7.8) we can see that $\mathcal{Q}_y^s$ acts as described in the theorem, thus completing the proof. □

Remarks:

1. For clarity, note that we can write the final expression in the proof as

$$I = \int_s^\infty \mathbb{E}_s M(\mathrm{d}y) f(y) \left(\int_0^{y-s} \mathrm{d}\theta g(\theta) \left[\mathbb{I}_{\theta<y-s} \frac{k p_k(s+\theta)}{m(s+\theta)} f_0(\theta, k) \right.\right.$$
$$\times \frac{1}{k} \sum_{j=1}^{k} \prod_{\substack{i=1 \\ i \neq j}}^{k} \mathbb{E}_{s+\theta}(f_i) \int_\Omega \mathcal{Q}_y^{s+\theta}(\mathrm{d}\omega) f_j$$
$$\left.\left. + \mathbb{I}_{\theta=y-s} p_k(y) f_0(y-s, k) \prod_{i=1}^{k} \mathbb{E}_y(f_i) \right] \right)$$

   where $g(\theta)$ is the density given in Equation (7.7).

2. It is not obvious how to extend this result to trees conditioned to have two or more specified split points. The difficulty arises when considering whether or not later split points lie on the same distinguished path back to the root as the first split point does. For an application such as the evolutionary trees considered in this thesis, conditioning on two or more split points is possible as we know in advance whether later split points lie in the same subtree as the first split. In this case the theorem extends in the obvious manner.

3. Note that after the death at time y, all branches evolve as independent identically distributed Galton-Watson trees.

In this chapter I have described the structure of Galton-Watson trees when they are conditioned to have a branching point at a specified time. The resulting fish-bone structure clearly suggests a method of simulation. In the next chapter I give details of how to

simulate such trees and I give the results from a full Bayesian analysis of the distribution of the primate and anthropoid divergence times.

Chapter 8

A Full Bayesian Analysis

In Chapter 4 I gave the optimal subtree selection algorithm, which has since been used to infer the joint distribution of the primate and anthropoid divergence times. The idea was that we could infer the anthropoid divergence time by simulating primate trees and then finding the subtree that looked most like the anthropoid data. This ad hoc approach is, in some sense, unsatisfactory. For the parameters $\rho, \gamma, \lambda, \alpha$ and τ, the approach has been to choose values from prior distributions, simulate a tree forwards in time, and then decide whether to reject or accept the parameters into the posterior distribution according to how close the simulations are to reality.

The approach used for the anthropoid temporal gap τ^* is different; I have been simulating a tree and then determining τ^* according to some optimality criterion. Consequently, it is not clear what the resulting distribution for τ^* represents: it is not a posterior distribution.

The ideas in the previous chapter have shown that it is possible to simulate speciation trees conditional on a subtree originating at a given point in time. In this chapter I implement this approach. This allows us to choose τ^* from a prior distribution and then use the ABC approach to infer its posterior distribution as we have been doing for the other parameters. This allows us to find the joint posterior distribution of two divergence times using a full Bayesian analysis for the first time.

Section 8.1 contains the details of how to simulate conditioned trees, and includes an approximate method for simulating inhomogeneous exponential distributions. Section 8.2 gives the results from the simulations.

8.1 Fish-Bone Simulations

In this section I show how to simulate the conditioned Galton-Watson process described in Chapter 7. One of the requirements is to simulate inhomogeneous $\text{Exp}_s(\lambda m(\cdot))$ random variables, as this is the lifetime distribution of species on the distinguished path (Theorem 1 step (i)). The distribution function of this variable is

$$F(x) = \mathbb{P}_s(X \leq x) = 1 - \exp\Big(- \int_s^{s+x} m(t)\mathrm{d}t\Big) \tag{8.1}$$

where s, in our case, is the world time at which the species originates.

The function $m(t) = \sum kp_k(t)$ is the mean of the offspring distribution at time t. We require that $\int_s^\infty m(t)\mathrm{d}t = \infty$ in order that $F(x)$ represents a distribution function. For logistic population growth, we found in Chapter 2 that the mean offspring distribution

$$m(t) = 1 + \frac{\frac{\rho}{2\lambda}(1-\gamma)}{1-\gamma+\gamma e^{\rho t}}$$

gave the required diversity.

One method for simulating $\text{Exp}_s(\lambda m(\cdot))$ random variables is to use the inverse of the distribution function, as described in Liu [73]. If U is a uniform random variable on $[0,1]$ then $F^{-1}(U)$ has distribution $F(\cdot)$. Unfortunately, the distribution $F(x) = \mathbb{P}_s(X < x) = 1 - \exp\big(- \lambda \int_s^{s+x} m(t)dt\big)$ has no closed form expression for its inverse for the value of $m(\cdot)$ above. Instead, we could numerically find $F^{-1}(u) = \inf\{x : F(x) \geq u\}$ if desired. However, because the function $m(t)$ changes slowly in comparison to the average lifetime distribution, we can use a simple approximation. For a species born at time t, I approximate its branch length by a standard $\text{Exp}(\lambda m(t))$ distribution, where $m(t)$ is treated as a constant.

We are now in a position to simulate conditioned Galton-Watson trees. Starting from time zero with two species, one of which is labelled as *distinguished* and the other as *undistinguished*, we can use the following algorithm:

Algorithm N: Simulating Conditioned Galton-Watson Trees

N1 For each undistinguished species, replace it with L offspring after an $\text{Exp}(\lambda)$ period, where L has distribution $\{p_k(d)\}_{k=0,1,2,\ldots}$ and d is the death time of that species. Label all the offspring as *undistinguished*.

N2 If there is a distinguished species, let s denote its birth time, and generate $\sigma \sim \text{Exp}_s(\lambda m(\cdot))$. Let $d = s + \sigma$ denote the proposed death time.

(a) If $d < y$, replace the distinguished species with $\widehat{L}$ species, where $\widehat{L}$ has size-biased distribution $\{kp_k(d)/m(d)\}_{k=1,2,\ldots}$. Uniformly choose one of these species and label it as *distinguished*. Label all other offspring as *undistinguished*.

(b) If $d > y$, kill the distinguished species at time y, and replace it by L new species. Label all offspring as *undistinguished*.

N3 Return to step N1 until we have simulated far enough into the future.

This algorithm describes how to simulate conditioned Galton-Watson trees with general offspring distributions. We, however, are only interested in binary trees, where each species has either zero or two offspring, which is a special case of the algorithm above. We are now able to write down the complete inference algorithm:

Algorithm O: A Full Bayesian Analysis

O1 Draw parameters $\tau, \tau^*, \boldsymbol{\alpha}, \gamma, \lambda, \rho$ from their prior distributions.

O2 Simulate a tree starting with two species at time $-54.8 - \tau$ My ago conditional on a birth at time $y = -37 - \tau^*$ My ago.

O3 Check that both sides of the tree survive to the present, and also that both sides of the subtree originating at time y survive to the present.

O4 Simulate fossil finds and count the number of fossils on the complete tree ($\mathcal{D}'$) and on the subtree ($\mathcal{S}'$).

O5 If $\rho(\mathcal{D}, \mathcal{D}') < \epsilon_1$ and $\rho(\mathcal{S}, \mathcal{S}') < \epsilon_2$, accept the parameters. Otherwise reject. Return to step O1.

This algorithm is approximately two orders of magnitude quicker than the approximate size-biasing algorithm I gave in Section 7.2. Comparing the output from the two algorithms acts as a quick check that the calculations of the previous chapter are correct.

8.1.1 Prior Distributions

Throughout this chapter, I use the prior distributions shown below, which are similar to those used in previous chapters, with the exception that we are now able to specify a prior distribution for τ^*.

$$\tau \sim U[0, 100]$$
$$\tau^* \sim U[0, 30]$$

$$\begin{aligned}\beta &\sim U[0, 0.6]\\ \gamma &\sim U[0.005, 0.02]\\ \rho &\sim U[0, 0.8]\\ 1/\lambda &\sim U[2, 3]\end{aligned}$$

The prior for τ^* was chosen after discussions with our primatologist collaborators. The oldest anthropoid fossil is from the Late-Eocene, and so 37 My is an upper bound on the divergence time. For a lower bound, it was felt that the anthropoid divergence time is somewhat better constrained by the fossil record than the primates and that an upper bound of 67 My ago was sufficient. Assuming a flat prior distribution for the temporal gap gives a Uniform$[0, 30]$ prior distribution for τ^*. As shall be seen from the results, little difference is made by using a larger range extension. Note that for small τ values, large τ^* values would mean that the anthropoid subtree originated before the primate tree originated, which is clearly impossible. Thus in practice, the prior used for τ^* is really a uniform distribution on $[0, \min(30, 17.8 + \tau)]$, i.e., the event $\tau^* > \tau$ has zero prior probability.

8.2 Results

In this section, I give the results generated by Algorithm O. As is perhaps to be expected by now, the initial simulations were unsuccessful. The simulations ran extremely slowly, even when large values of $\boldsymbol{\epsilon}$ were used, and the accepted results had unacceptably small values of N_0^{main} and N_0^{sub}, the extant numbers of primates and anthropoids. Upon examination, it was discovered that this is due to the assumption made in Chapter 2 that the population diversity grows logistically. This assumption requires that there is an initial phase of rapid population growth, followed by a long period of constant population size[(1)]. Once the initial period of population growth is finished, the offspring probabilities, $p_0(t)$ and $p_2(t)$, are both approximately equal to $1/2$. From Equation (2.11) we can see that $p_0(t) \uparrow 1/2$ and $p_2(t) \downarrow 1/2$ as t tends to infinity. At the point the subtree originates I found that both values were usually very close to $1/2$. This means that the subtree is essentially a critical Galton-Watson process ($\mathbb{E}L = 1$). The expected population size for critical processes equals the initial population size, which in this case is two species. Even when conditioning on non-extinction, the population size only grows linearly with time.

A solution to observing unrealistically small N_0^{sub} values is to assume that both the anthropoid and the non-anthropoid population diversity grow logistically. Scientifically, this does not present any problem as there is no consensus about which taxonomic level we

[(1)] It is only the expected population size that remains constant, and so in practice, the population size will fluctuate for each simulation

should model as having logistic growth. So modelling the primate order, or the anthropoid parvorder, as having logistic growth are both valid approaches. Of course, a consequence of allowing both anthropoid and non-anthropoid populations to grow logistically is that the primate population will no longer show logistic growth.

In the results that follow, I assumed that the two different growth curves take different parameter values, as I could see no reason for why they should be the same. Note that previously this was not possible, as we did not know in advance which subtree represented the anthropoids. All of the results given in this chapter are obtained using the Poisson sampling scheme introduced in Chapter 6.

8.2.1 Fixed Sampling Rates

In this section I present results obtained using the fixed sampling rate approach from Chapter 3, where $\beta_i = \beta p_i$. The notation used is as follows: N_0^{main} is the extant primate diversity and N_0^{sub} the extant anthropoid diversity. N_1^{main} and N_2^{main} represent the extant diversity on either side of the main primate tree, with side one representing the haplorhini (includes the anthropoid subtree), and side two the strepsirrhini. Parameters ρ^{main} and γ^{main} are the growth parameters for the non-anthropoids, and ρ^{sub} and γ^{sub} the growth parameters for the anthropoid subtree. Using the population-adjusted metric on both trees with $\boldsymbol{\epsilon} = (0.4, 0.4)$ gives the results shown in Table 8.1 and Figure 8.1.

	Min.	LQ	Median	Mean	UQ	Max.
N_0^{main}	228	360	391	393	425	553
N_1^{main}	165	296	328	329	361	496
N_2^{main}	5	41	58	64	81	231
N_0^{sub}	140	233	260	259	284	430
τ	0.0	17.9	36.0	40.3	60.3	99.9
τ^*	0.0	2.6	5.9	7.6	10.8	30.0
β	0.057	0.186	0.279	0.305	0.416	0.600
ρ^{main}	0.003	0.397	0.533	0.516	0.657	0.800
ρ^{sub}	0.041	0.148	0.235	0.270	0.365	0.791
γ^{main}	0.0050	0.0096	0.0133	0.0131	0.0167	0.0200
γ^{sub}	0.0050	0.0078	0.0111	0.0115	0.0150	0.0200
$1/\lambda$	2.00	2.31	2.56	2.54	2.80	3.00

Table 8.1: *A summary of the posterior distributions when using the conditioned Galton-Watson process from Chapter 7. The results were obtained using the Poisson sampling model with fixed ratios for the sampling rates and ρ_p for both trees with $\boldsymbol{\epsilon} = (0.4, 0.4)$ ($n = 4392$).*

The acceptance rate was 6682 successfully simulated trees per accepted tree, which is much higher than that found in Chapter 4 when using the binomial sampling scheme and the optimal subtree selection algorithm. The fact that the conditioned Galton-Watson trees (with anthropoid logistic growth) are a better fit has allowed us to use the population-

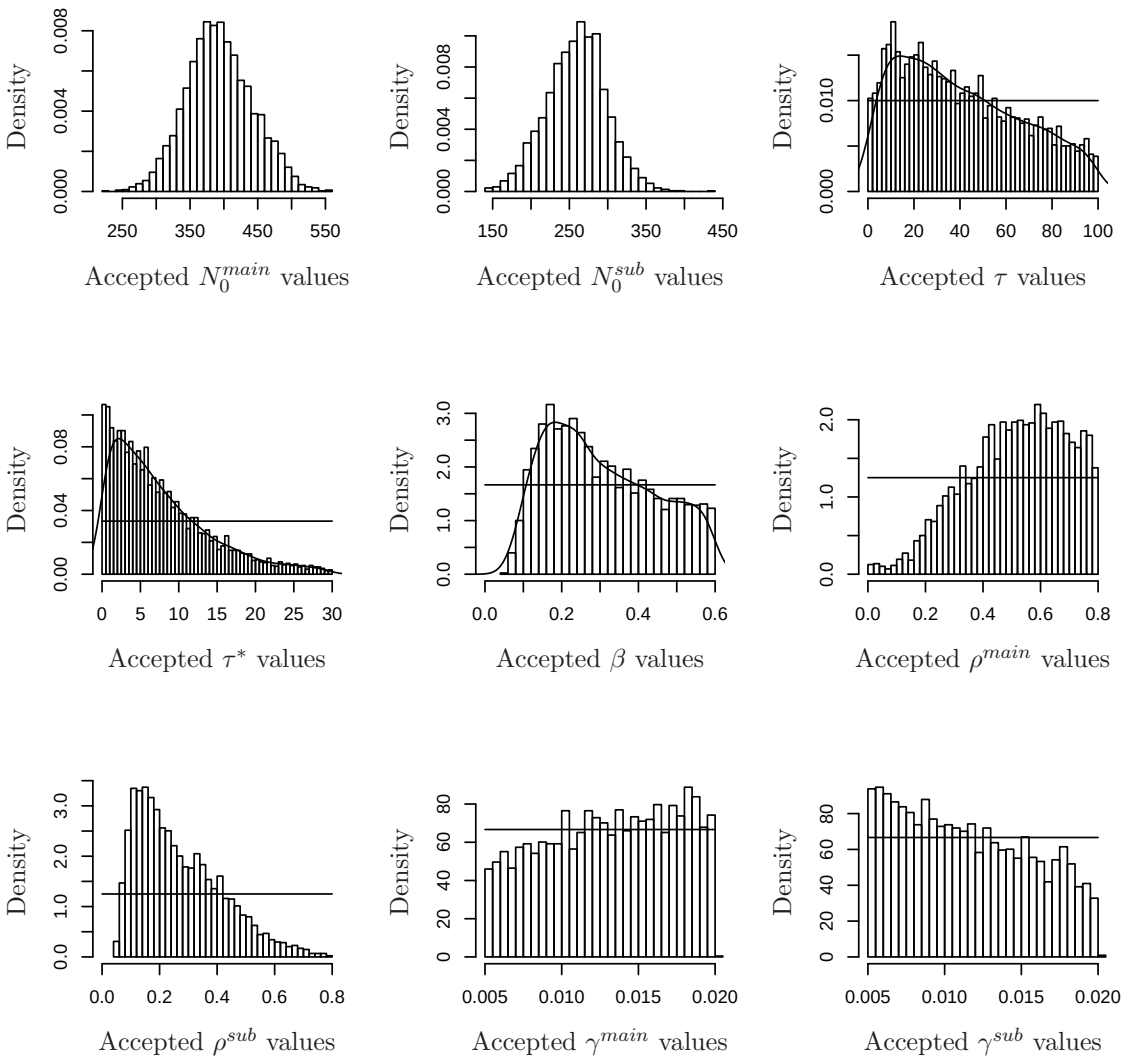

Figure 8.1: *Plots of the posterior distributions obtained when using the conditioned Galton-Watson process from Chapter 7. Obtained using the Poisson sampling scheme with $\beta_i = \beta p_i$ and using ρ_p for both trees with $\boldsymbol{\epsilon} = (0.4, 0.4)$.*

adjusted metric on the subtree as well as the main tree so that the anthropoid diversity is taken into account.

The main difference between these results and those obtained in previous chapters is the shorter range extension for τ^*. The 90% credibility interval for the anthropoid divergence time is now $(53.7, 37)$ My ago. The posterior distribution for τ is now more disperse than previously observed, however.

For these simulations the same fossil sampling rates were used throughout the tree so that all species in the same interval had an equal chance (per million years) of being

preserved as a fossil. The data, however, suggest otherwise. Table 4.1 shows that 75% of modern species are anthropoids yet 97% of Pliocene and Pleistocene (i.e., from 5.3 My ago to the present) fossil species are anthropoids. This suggests that perhaps the two groups were subject to different sampling rates.

Now that our simulation approach specifies in advance which subtree will represent the anthropoids, we are able to vary the parameters used on the subtree. This was not possible previously as it was not known at the time of simulation which tree would represent the subtree. In the results above, I allowed the growth parameters to differ on the main tree and the subtree. We can also allow the sampling rates to differ on the subtree and the main tree. Using the same fixed ratios for each side of the tree, $\beta_i^{main} = \beta^{main} p_i$ and $\beta_i^{sub} = \beta^{sub} p_i$, and using the population adjusted metric on each tree with $\boldsymbol{\epsilon} = (0.4, 0.4)$ leads to the results shown in Table 8.2 and Figure 8.2.

	Min.	LQ	Median	Mean	UQ	Max.
N_0^{main}	207	328	358	358	386	539
N_1^{main}	153	293	324	325	354	492
N_2^{main}	1	10	24	34	47	284
N_0^{sub}	133	247	274	272	297	419
τ	0.0	12.0	27.5	34.2	52.0	100.0
τ^*	0.0	3.1	6.7	8.5	12.4	29.9
β^{main}	0.014	0.050	0.077	0.113	0.133	0.600
β^{sub}	0.026	0.058	0.075	0.083	0.100	0.270
ρ^{main}	0.001	0.440	0.586	0.553	0.696	0.800
ρ^{sub}	0.062	0.141	0.236	0.281	0.391	0.760
γ^{main}	0.0050	0.0101	0.0139	0.0134	0.0172	0.020
γ^{sub}	0.0050	0.0071	0.0112	0.0115	0.0152	0.0200
$1/\lambda$	2.00	2.35	2.61	2.58	2.81	3.00

Table 8.2: *A summary of the posterior distributions when using fixed sampling fractions with the conditioned Galton-Watson process, but with different β values on each side of the tree. Obtained using ρ_p for both trees with $\boldsymbol{\epsilon} = (0.4, 0.4)$ ($n = 4150$).*

The acceptance rate was 34021 successfully simulated trees per accepted tree. This is lower than when using the same β on each tree as the extra parameter means that the sample space has an extra dimension which we must now search, and so the ABC algorithm thus misses the target (areas of high posterior density) more often.

The main change between the results obtained using the same sampling rates on each tree and using different rates, is on the posterior distribution of β, the sampling rate. In Figure 8.2 the posteriors are tightly constrained around peaks at about 0.06 fossils per million years. Whereas in Figure 8.1, the posterior for β is constrained away from the origin and is much more widely dispersed.

One of the problems with this analysis is that I have assumed that the ratio of the sampling rates is fixed the same for both trees, i.e., $\beta_i^{main} = \beta^{main} p_i$ and $\beta_i^{sub} = \beta^{sub} p_i$.

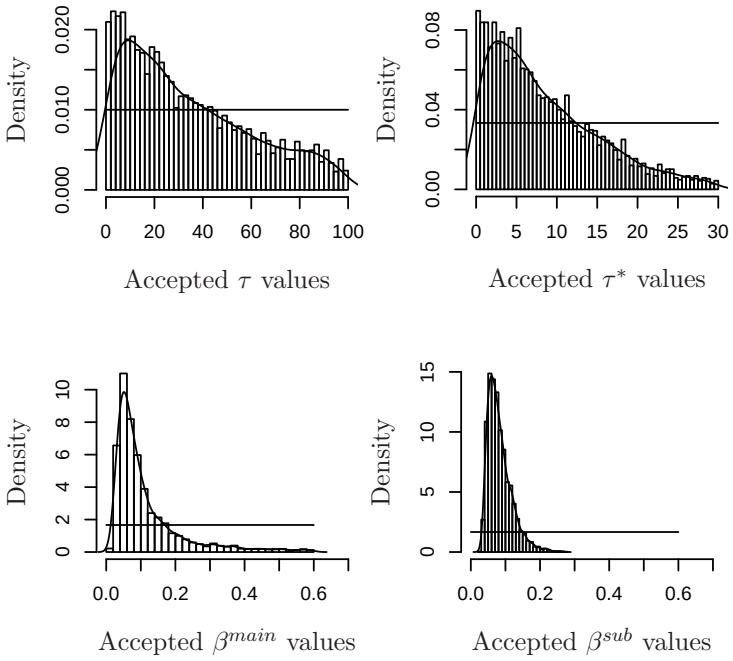

Figure 8.2: *Marginal posterior distributions for the parameters of interest when using different sampling rates for the main tree and the subtree in the conditioned Galton-Watson model of Chapter 7. Obtained using ρ_p for both trees with $\boldsymbol{\epsilon} = (0.4, 0.4)$.*

In the next section I try to remove this assumption and suggest why this is not feasible.

8.2.2 Free Sampling Fractions

The obvious way to proceed from the previous section is to allow the sampling fractions $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{14})$ to take independent values and use the ABC-MCMC hybrid sampler described in Chapters 5 and 6 to infer their posterior distributions. Unfortunately, this approach is unsuccessful. Using the population-adjusted metric ρ_p on both trees with $\epsilon = (0.6, 0.6)$ leads to the marginal posterior distributions shown in Figure 8.3.

We can see that the posterior distributions for the two temporal gaps, τ and τ^*, have barely changed from their original prior distributions. I believe there are two main reasons why this approach has failed to provide any information about the temporal gaps.

Firstly, simulating successful conditioned trees takes much longer than using the model used in previous chapters. Because of this I have been forced to use larger tolerance values than was previously necessary, thus providing less pressure on the simulated data to look like the real data.

Secondly, and probably more importantly, I believe that we are seeing the effects of confounding between the parameters. Since using the conditioned Galton-Watson trees

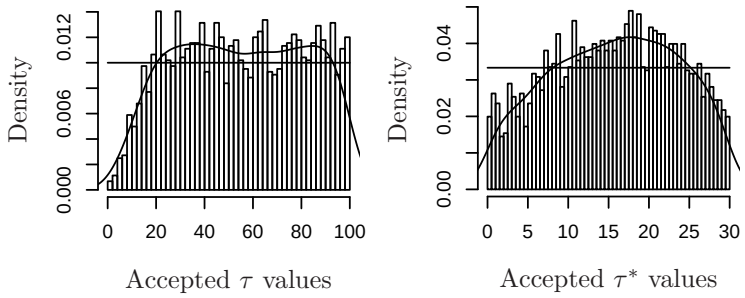

Figure 8.3: *Marginal posterior distributions obtained using the conditioned Galton-Watson process and allowing the sampling fractions to vary freely. Inference was done using the approximate-Gibbs sampler from Chapter 5. The large number of model parameters used have stopped us learning anything about the primate and anthropoid divergence times.*

I have introduced new parameters into the problem. For the simulations in this section we used 21 parameters. Since the data only consists of 28 data points this is too large a number to be supported by the data. Having too many parameters means that the effect of each is confounded with the others. In conclusion, I believe we are at the limit of what the data can support.

8.2.3 Using previous parameter estimates

An alternative approach that brings together some of the previous inferences is to fix some of the parameters to have a specific value and to repeat the simulations. As discussed in Chapter 4, fixing parameter values does not fully account for all of the uncertainty, but the practice is commonplace. Fixing the sampling rates, $\beta_1, \ldots, \beta_{14}$, to be equal to a multiple of those those given in Table 6.2, and taking the growth parameters to be equal to the posterior modal values from Figure 8.1, we find the results given in Table 8.3 and Figure 8.4.

Removing the uncertainty surrounding these values reduces the number of free parameters in the model, which lessens the effects of the confounding seen previously. The posterior distributions for τ and τ^* do now contain a signal that is different to their prior distributions. We found the 90% credibility interval for the primate divergence time to be $[54.8, 92.1]$ My ago, and the corresponding interval for the anthropoid divergence time to be $[37.0, 52.0]$ My ago.

	Min.	LQ	Median	Mean	UQ	Max.
N_0^{main}	202	315	355	354	392	521
N_1^{main}	134	259	302	301	342	475
N_2^{main}	12	38	51	53	66	128
N_0^{sub}	126	223	263	262	298	425
τ	0.0	5.6	10.7	14.0	20.2	49.9
τ^*	0.0	3.3	6.5	6.9	9.9	25.4
β^{main}	0.78	2.9	4.2	4.7	6.2	10.0

Table 8.3: *A summary of the posterior distributions when using fixed sampling fractions (with values from Table 6.2) and fixed growth parameters with the conditioned Galton-Watson process. Obtained using ρ_p for both trees with $\boldsymbol{\epsilon} = (0.4, 0.4)$, $\rho^{main} = 0.72, \rho^{sub} = 0.265, \gamma^{main} = 0.02, \rho^{sub} = 0.0065, 1/\lambda = 3$, $(n = 2185)$.*

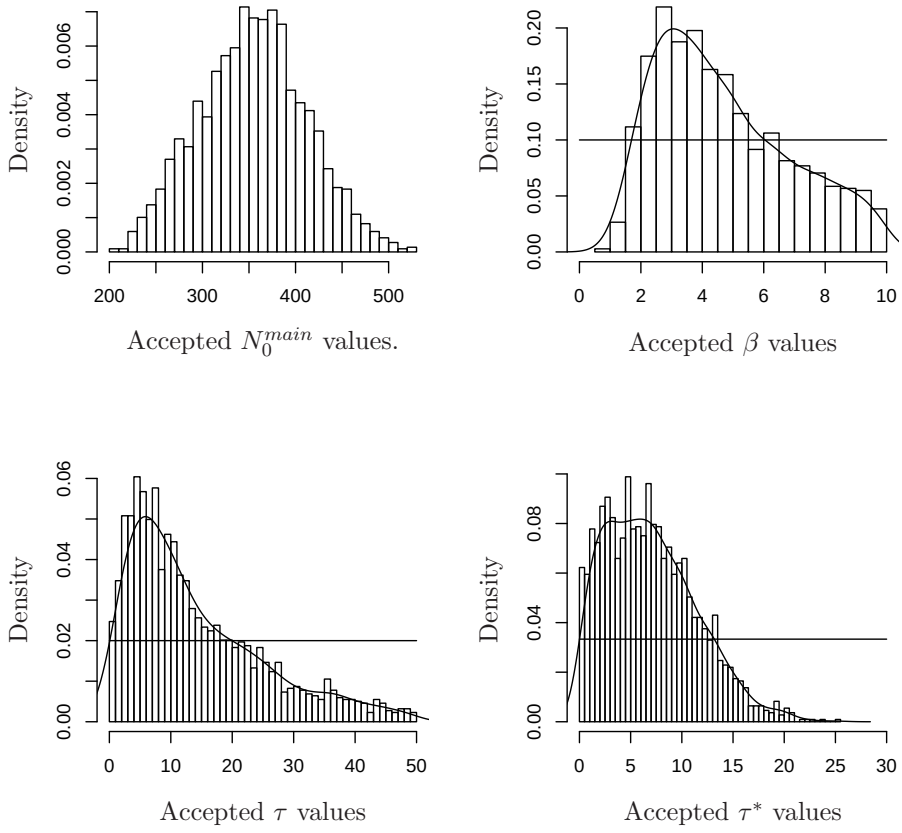

Figure 8.4: *Marginal posterior distributions for the parameters of interest when using fixed growth parameters and fixed sampling rates (from Table 6.2) with the conditioned Galton-Watson process. Obtained using ρ_p for both trees with $\boldsymbol{\epsilon} = (0.4, 0.4)$, $\rho^{main} = 0.72, \rho^{sub} = 0.265, \gamma^{main} = 0.02, \rho^{sub} = 0.0065, 1/\lambda = 3$, $(n = 2185)$.*

Chapter 9

Conclusions and Future Work

The key conclusion from this thesis is that the primate divergence time is not as tightly constrained by the fossil record as previously thought. Previous work [51, 52] has overstated the accuracy of the predictions possible using only fossil evidence; it is not possible using this evidence alone to rule out the possibility that the primates originated well into the Cretaceous and coexisted alongside dinosaurs for many years before their eventual extinction 65 My ago. Throughout this thesis I have used various different models, all of which have suggested that the posterior probability that primates diverged in the Cretaceous is large. For the sake of brevity I only report here what I consider to be the strongest set of inferences, which are those from Chapter 6.1.2.

The 95% credibility interval[(1)] for the primate divergence time was found to be $[83.4, 54.8]$ My ago. The 95% credibility interval for the anthropoid divergence time was $[62.8, 37.0]$ My ago (although if the conditioned Galton-Watson model from Chapter 7 is used, the 95% credibility interval for the anthropoids reduces to $[52.0, 37.0]$ My ago). If the K-T crash is modelled (cf. Section 6.2) these intervals reduce to $[82.0, 54.8]$ and $[60.4, 37.0]$ My ago for

[(1)] The aim of Bayesian inference is to find posterior distributions. To reduce these posteriors to a point estimate or a credibility interval is not philosophically sound. However, as this is the norm, I give the credibility intervals found for the parameters of interest and refer the reader to the plots of the distributions in Chapter 6.

the primate and anthropoid divergence times respectively. While it is to be expected that modelling the K-T mass extinction should make the estimates more recent, the effect is not sufficient to make the likelihood of Cretaceous primates as low as previously claimed.

The work presented here also goes some way towards explaining the difference between the genetic and fossil based estimates. For while the estimates found here do not overlap with recent genetic estimates, they are much closer. For example, in a recent 2007 *Nature* article Bininda-Emonds *et al.* [19] gave a 95% confidence interval of $[90.4, 85.0]$ My ago for the primate divergence time. It should also be remembered that we expect some disparity between genetic estimates and fossil estimates as genetic estimates date the time at which genetic mixing ceased, whereas fossils, at best, date the time at which the diagnostic characters first appeared, which will always be after the genetic date. Benton [14] has suggested that the estimates will in time converge as more fossils are discovered and as genetic estimates better take account of the irregular molecular clock.

In addition, it should be noted that our estimates also cover the dates given by the palaeontology community, which generally place primates firmly in the Cenozoic. It was never going to be possible to disprove that the primates have Cenozoic origins, and nor was it the aim. I have, however, shown that the possible degree of accuracy when making predictions using the fossil record is less than previously thought by many authors. Given the data that are available, the expected temporal gap between the divergence time and the age of the first fossil discovery is larger than is currently accepted.

The guiding principle when developing the methodology was that the inference of divergence times should be based on a sound theoretical platform. The methodology used was based on the work of Tavaré *et al.* [118], but with a new inference mechanism rooted in the Bayesian approach to statistics which finds the posterior distributions of the parameters. I have given special emphasis to using all of the information available in the data, by conditioning on both the modern diversity and the internal structure provided by the number of anthropoid species. Inference of the anthropoid divergence time was done firstly by using an optimality criterion to select the subtree which closest matched the data, and later by conditioning the Galton-Watson model to have a subtree originate at a particular point in time.

I have made several improvements to the model of fossil preservation and discovery used previously. The binomial model in which each species had an equal probability of preservation, regardless of its branch length, was replaced by a Poisson sampling scheme in which longer lived species were more likely to be discovered than shorter lived ones. I also removed the assumption necessary in Tavaré *et al.* which required the specification of a fixed ratio for the sampling rates. This subjective assumption was often then criticised by those who disagreed with their conclusions. For this reason I developed an inference approach which allowed the sampling fractions to vary independently, and for their value

to be determined solely by the data.

Throughout this thesis I have tried to highlight the limit of what is possible in terms of model complexity. For a simple data set such as the primate count data used here, overly complex models can lead to confounding between parameters and the loss of any clear or reliable biotic signal. This was the case at various points in the thesis. I also gave a model of the mass extinction event which took place at the Cretaceous-Cenozoic boundary and found that this made Cretaceous origins for the primates less likely than under a neutral model.

Several statistical contributions have been made in this thesis. A hybrid ABC-MCMC sampler has been developed that allows exact Monte Carlo inference to be used on the aspects of the model for which it is possible, and Approximate Bayesian Computation to be used on parts for which it is not. Using Markov Chain Monte Carlo means that successive outputs from part of the parameter space becomes correlated and consequently more time is spent by the sampler in regions of high posterior probability. This has the effect of increasing the speed and efficiency of the computations as less time is spent searching the tails of the distributions. I also gave an in-depth discussion of the pros and cons of ABC and gave several toy calculations to show how the error in the approximation scales with the choice of tolerance. A simulation study was performed to assess the accuracy of the ABC inference methods for the divergence time problem. The error was acceptable, and it was found that the accuracy could be improved by performing local-linear regression on the output as suggested by Beaumont *et al.* [11].

I have also looked at a possible model selection technique that arises naturally from the ABC algorithm and used this to show that the primate data are explained better by logistic growth than exponential or linear growth.

Finally, I have derived the structure of Galton-Watson processes when conditioned on a birth or death occurring at a specified point in time. These conditioned trees have a fish-bone like structure with a distinguished spine from the root to the species which died at the specified time. Along this root the offspring distribution is the size-biased version of the standard offspring distribution, and the branch lengths of species on the spine have an inhomogeneous exponential distribution with a different parameter to before. Species off the spine evolve as usual. I have given a detailed technical proof of this fact based on the Palm approach to random point measures.

Ultimately, our hope is that this methodology will be built upon by others in the future and used to date the divergence times of other major clades.

Future Work

In this thesis I have developed a methodology for inferring divergence times using data from the fossil record. Applying this to the primates, I found that Cretaceous origins are feasible and perhaps even likely. Ultimately, however, this question can only be definitively answered by further fossil discoveries. New discoveries have the ability to instantly change our understanding of evolutionary history. For example, loris fossil remains found in Egypt in 2003 doubled the documented geological age of lorisform primates from 20 to 40 My [80, 108]. Robert Martin summed up the situation [79] when he wrote the following:

> The very fact that new fossil discoveries can have such a dramatic impact is eloquent testimony to the yawning gaps in our knowledge that still remain.

One of the consequences of the early primate divergence time suggested by our work is that other divergence times, such as the ape-human divergence, may also need to be revised to an earlier date. A recent corroboration of our approach was provided in August 2007 when Suwa *et al.* [116] published findings of a great ape fossil from the late Eocene. The human-ape divergence was previously pegged at about 6 My ago. However, if this discovery is accepted as a genuine ape fossil this finding will push the human-ape divergence time back to at least 10.5 My ago. This is a stark example of the effect a single discovery can have.

Fortunately, we may not need to wait too much longer for stronger evidence. Soligo *et al.* [113] illustrate the almost exponential growth in the number of new fossil primate species discovered per decade over the previous century. If a Cretaceous primate fossil is discovered, then the question will have been resolved. On the other hand, if all new fossil discoveries are from the Cenozoic it adds weight to the argument that primates originated in the last 65 million years.

To improve the methodology, there are various areas that should be pursued. The assumption that the diversity grows logistically is only an approximation. During the Eocene and again during the Miocene there was an increase in global temperatures which enabled primates to spread to the northern continents, before dying out in those locations at the end of these epochs. The population fluctuations, known as the Eocene and Miocene bulges, could be modelled by using a more general growth curve. However, we are again likely find that the data set is too meagre to be able to accurately model these fluctuations. We could also conceivably take into account geographic considerations. The majority of fossils are from Europe and North America. However, many people believe that primates originated in either Africa or Asia and only spread to the Northern continents during warmer climatic conditions. A two state model, with states representing the well sampled northern continents and the less well sampled southern continents, could perhaps better account for the structure of the data.

Another extension would be to take more of the data structure into account by dating three split points simultaneously. In order to do this, however, considerable increases in the speed of the simulations would be required. A massive improvement could be achieved if we could efficiently simulate Galton-Watson trees conditioned on non-extinction. It is not currently known how to do this, other than by simulating standard Galton-Watson trees and rejecting those which become extinct. This approach means that we reject about 95% of the simulations, as we require not only that each side of the main primate tree survives, but also that each side of the subtree survives. Oliver Will used an approach in his PhD thesis which involved simulating trees backwards in time [125]. The model begins at the present time with a number of species equal to the modern diversity, and then runs backwards in time until only one species remains. This has the advantage of naturally taking into account the modern primate diversity and means that we do not need to check for non-extinction. Unfortunately the computational requirements of this approach are large, making inference for this model too costly to perform for the situations of interest studied here.

I have discussed the main areas of future work for the Approximate Bayesian Computation method in Chapter 3. But to recap briefly, before ABC can become a more popular inference tool, a methodology for selecting and assessing good summary statistics needs to be developed, as well as an assessment of the general accuracy of ABC methods.

A feasible and achievable improvement in the accuracy of our estimates of divergence times could be achieved by taking into account the number of times each species had been preserved in the fossil record. This would improve the estimates of the sampling rates for each epoch, and as it is these estimates that are key to estimating divergence times, a potentially large improvement could be made. Wang and Dobson 2006 [121] used an abundance-based coverage estimator to estimate the completeness[(2)] of the dinosaur fossil record. For a given completeness, the expected number of species found in the fossil record once, twice, thrice etc., can be computed and then compared with the known abundances, and adjusted until they are close, giving an estimate of the true completeness. Approximately 40% of primate fossil species are known from only a single fossil sample (personal communication with Robert Martin and Christophe Soligo). A simple change to our model so that each species in the tree could be sampled numerous times would, with the addition of a term to the metric to take this into account, enable us to improve the accuracy of the inferences. At present, however, the data concerning the frequency of fossil finds is unavailable.

I believe that the most critical improvement that could be made to improve the estimates of divergence times is to utilise more of the available information. The grand aim must be to develop a unified approach to estimation that combines the direct information given

[(2)]The percentage of species preserved in the fossil record

by the fossil record and the indirect information given by the DNA of modern primates. There may also be possibilities for exploiting the known phylogenetic relationships between extant species. It is not yet clear how this might be achieved.

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(7):716–723, 1974.

[2] L. W. Alvarez, W. Alvarez, F. Asaro, and H. V. Michel. Extraterrestrial cause for the Cretaceous-Tertiary extinction: experimental results and theoretical interpretation. *Science*, 208:1095–1108, 1980.

[3] J. D. Archibald and D. H. Deutschmann. Quantitative analysis of the timing of the origin and diversification of extant placental orders. *Journal of Mammalian Evolution*, 8:107–124, 2001.

[4] U. Arnason, A. Gullberg, and A. Janke. Molecular timing of primate divergences as estimated by two nonprimate calibration points. *Journal of Molecular Evolution*, 47:718–727, 1998.

[5] U. Arnason, A. Gullberg, A. Janke, and X. F. Xu. Pattern and timing of evolutionary divergences between hominids based on analyses of complete mtDNAs. *Journal of Molecular Evolution*, 43(650-661), 1996.

[6] S. Asmussen and H. Hering. *Branching Processes*. Boston: Birkhauser, 1983.

[7] K. B. Athreya and P. Jagers. *Classical and Modern Branching Processes*. Springer: New York, 1997.

[8] K. B. Athreya and P. E. Ney. *Branching Processes*. Springer-Verlag, Berlin-Heidelberg-New York, 1972.

[9] K. B. Athreya and A. N. Vidyashankar. Branching processes. Technical report, University of Georgia, 1999.

[10] M. A. Beaumont and B. Rannala. The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5:251–261, 2004.

[11] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computatation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[12] R. Bellman and T. E. Harris. On the theory of age-dependent stochastic branching processes. *Proc. Natl. Acad. Sci. USA*, 34:601–604, 1948.

[13] M. J. Benton. Phlogeny of the major tetrapod groups: Morphological data and divergence dates. *Journal of Molecular Evolution*, 30:409–424, 1990.

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

[15] M. J. Benton and F. J. Ayala. Dating the tree of life. *Science*, 300:1698–1700, 2003.

[16] M. J. Benton and P. C. J. Donoghue. Paleontological evidence to date the tree of life. *Molecular Biology and Evolution*, 24(1):26–53, 2007.

[17] J. Berger. The case for Objective-Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[18] J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, Ser. B*, 41(2):113–147, 1979.

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

[20] G. J. Bowen, W. C. Clyde, W.C. Koch, P. L. Ting, J. Alroy, T. Tsubamoto, Y. Q. Wang, and Y. Wang. Mammalian dispersal at the Paleocene/Eocene boundary. *Science*, 295:2062–2065, 2002.

[21] G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[22] S. P. Brooks. Markov chain Monte Carlo method and its applications. *The Statistician*, 47:69–100, 1998.

[23] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

[24] S. P. Brooks and G. O. Roberts. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8(4):319–335, 1998.

[25] P. Brown, T. Sutikna, M. J. Morwood, R. P. Soejono, Jatmiko, E. Wayhu Saptomo, and Rokus Awe Due. A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature*, 431:1055–1061, 2004.

[26] L. Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419–1455, 1964.

[27] G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[28] B. Chauvin, A. Rouault, and A. Wakolbinger. Growing conditioned trees. *Stochastic Processes and their Applications*, 39:117–130, 1991.

[29] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[30] M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

[31] G. C. Conroy. *Primate Evolution*. W. W. Norton and Co.: New York, 1990.

[32] A. Cooper and R. Fortey. Evolutionary explosions and the phylogenetic fuse. *Trends in Ecology and Evolution*, 13(4):151–156, 1998.

[33] A. Cooper and D. Penny. Mass survivals of birds across the Cretaceous-tertiary boundary: Molecular evidence. *Science*, 275:1109–1113, 1997.

[34] M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[35] P. R. Crane and S. Lidgard. Angiosperm diversification and paleolatitudinal gradients in Cretaceous floristic diversity. *Science*, 246:675–678, 1989.

[36] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer series in Statistics. Springer-Verlag,, 1988.

[37] B. de Finetti. *Theory of Probability*, volume 1. Chichester: Wiley, 1974.

[38] B. de Finetti. *Theory of Probability*, volume 2. Chichester: Wiley, 1975.

[39] P. C. J. Donoghue. Saving the stem group - a contradiction in terms? *Paleobiology*, 31(4):553–558, 2005.

[40] M. Dwass. The total progency in a branching process and a related random walk. *Journal of Applied Probability*, 6:682–686, 1969.

[41] S. Easteal. Molecular evidence for the early divergence of placental mammals. *BioEssays*, 21:1052, 1999.

[42] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 2. Wiley and Sons, New York, 1950.

[43] M. Foote. Estimating taxonomic durations and preservation probability. *Paleobiology*, 23(2):278–300, 1997.

[44] M. Foote and D. M. Raup. Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology*, 22(2):121–140, 1996.

[45] J. Geiger. Size-biased and conditioned random splitting trees. *Stochastic Processes and their Applications*, 65:187–207, 1996.

[46] J. Geiger. Poisson point process limits in size-biased Galton-Watson trees. *Electronic Journal of Probability*, 5:1–12, 2000.

[47] J. Geiger and G. Kersting. Depth-first search of random trees, and Poisson point processes. *In [7]*, 1997.

[48] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[49] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511, 1992.

[50] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[51] P. D. Gingerich and M. D. Uhen. Time of origin of primates. *Journal of Human Evolution*, 27:443–445, 1994.

[52] P. D. Gingerich and M. D. Uhen. Likelihood estimation of the time of origin of Cetacea and the time of divergence of Cetacea and Artiodactyla. *Paleontologica Electronica*, 2:1–47, 1998.

[53] M. Goldstein. Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 1(3):403–420, 2006.

[54] F. M. Gradstein, J. G. Ogg, and A. G. Smith, editors. *A Geologic Time Scale*. Cambridge University Press, 2004.

[55] D. Graur and W. Martin. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends in Genetics*, 20(2):80–86, 2004.

[56] C. Groves. *Order Primates*, volume 1, pages 111–184. John Hopkins University Press, Baltimore, 2005.

[57] P. Guttorp. *Statistical Inference for Branching Processes*. John Wiley and Sons, New York, 1991.

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

[59] C. C. Heyde and E. Seneta. *I J Bienaymé: Statistical theory anticipated*. Springer-Verlag, New York, 1977.

[60] P. Jagers. The growth and stabilization of populations. *Statistical Science*, 6(3):269–283, 1991.

[61] H. Jeffreys. *The Theory of Probability*. Oxford University Press, 3rd edition, 1961.

[62] J. B. Johnson and K. S. Omland. Model selection in ecology and evolution. *Trends in Ecology and Evolution*, 19(2):101–108, 2004.

[63] O. Kallenberg. *Random Measures*. Akademie-Verlag, Berlin, 3rd edition, 1983.

[64] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[65] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[66] R. F. Kay, C. Ross, and B. A. Williams. Anthropoid origins. *Science*, 275:797–804, 1997.

[67] D. G. Kendall. On the generalized birth-and-death process. *The Annals of Mathematical Statistics*, 19(1):1–15, 1948.

[68] D. P. Kennedy. The Galton-Watson process conditioned on the total progency. *Journal of Applied Probability*, 12:800–806, 1975.

[69] I. Kontoyiannis, P. Harremoes, and O. Johnson. Entropy and the law of small numbers. *IEEE Transactions of Information Theory Society*, 51(2):466–472, 2005.

[70] S. Kumar and S. B. Hedges. A molecular timescale for vertebrate evolution. *Nature*, 392:917–920, 1998.

[71] S. Lidgard and P. R. Crane. Quantitative analyses of the early angiosperm radiation. *Nature*, 331:344–346, 1988.

[72] D. V. Lindley. The philosophy of statistics. *The Statistician*, 49(3):293–337, 2000.

[73] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, 2001.

[74] R. Lyons, R. Pemantle, and Y. Peres. Conceptual proofs of $l \log l$ criteria for mean behaviour of branching processes. *Annals of Probability*, 23:1125–1138, 1995.

[75] R. H. MacArthur and E. O. Wilson. *The theory of island biogeography*. Princeton University Press, New Jersey, 1967.

[76] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100(26):15324–15328, 2003.

[77] R. D. Martin. *Primate origins and evolution: a phylogenetic reconstruction*. Princeton University Press, New York, 1990.

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

[79] R. D. Martin. New light on primate evolution. In *Ernst Mayr Lecture*. Berlin-Brandenburgische Akademie Der Wissenschaften, 2003.

[80] R. D. Martin. Palaeontology: Combing the primate record. *Nature*, 422:388, 2003.

[81] R. D. Martin. Chinese latern for early primates. *Nature*, 427:22–23, 2004.

[82] R. D. Martin, A. M. MacLarnon, J. L. Phillips, L. Dussubieux, P. R. Williams, and W. B. Dobyns. Comment on "The brain of LB1, Homo floresiensis". *Science*, 312(5776):999, 2006.

[83] R. D. Martin, C. Soligo, and S. Tavaré. Primate origins: Implications of a Cretaceous ancestry. *Folia Primatologica*, 78, 2007.

[84] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. London: Springer-Verlag, 1993.

[85] K. G. Miller, R. G. Fairbanks, and G. S. Mountain. Tertiary oxygen isotope synthesis, sea level history, and continental margin erosion. *Paleoceanography*, 2:1–19, 1987.

[86] R. A. Mittermeier, W. R. Konstant, F. Hawkins, E. E. Louis, and O. Langrand. *Lemurs of Madagascar*, volume 2. Conservation International, 2006.

[87] S. Nee, R. M. May, and P. H. Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society Soc. B*, 344:305–311, 1994.

[88] J. Neveu. Arbres et processus de Galton-Watson. *Annales de l'Institut Henri Poincaré*, 22:199–207, 1986.

[89] M.A. Newton and A. E. Raftery. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Ser. B*, 56(3-48), 1994.

[90] D. Penny and M. J. Phillips. Mass survivals. *Nature*, 446:501–502, 2007.

[91] P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

[92] S. E. Peters and M. Foote. Biodiversity in the Phanerozoic: a reinterpretation. *Paleobiology*, 27:583–601, 2001.

[93] S. E. Peters and M. Foote. Determinants of extinction in the fossil record. *Nature*, 416:420–424, 2002.

[94] V. Plagnol and S. Tavaré. Approximate Bayesian computation and MCMC. In H. Niederreiter, editor, *Proceedings of Monte Carlo and Quasi-Monte Carlo Methods 2002*, pages 99–114. Springer-Verlag, 2004.

[95] M. Plummer, N. G. Best, M. K. Cowles, and S. K. Vines. *CODA: Output analysis and diagnostics for MCMC*. MRC Biostatistics Unit, Cambridge, UK, 2006.

[96] D. J. Poirier. The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, 1(4):969–980, 2006.

[97] A. D. Polyanin, V. F. Zaitset, and A. Moussiaux. *Handbook of First Order Partial Differential Equations*. Taylor and Francis, London, 2002.

[98] J. K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

[99] P. S. Puri. On the homogeneous birth-and-death process and its integral. *Biometrika*, 53(1/2):61–71, 1966.

[100] A. E. Raftery and S. M. Lewis. How many iterations in the Gibbs sampler? In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger, editors, *Bayesian Statistics 4*. Oxford University Press, 1992.

[101] D. M. Raup. Biases in the fossil record of species and genera. *Bulletin of the Carnegie Museum of Natural History*, 13:85–91, 1979.

[102] D. M. Raup, S. J. Gould, T. J. M. Schopf, and D. S. Simberloff. Stochastic models of phlogeny and the evolution of diversity. *Journal of Geology*, 81:525–542, 1973.

[103] D. M. Raup and J. J. Sepkoski. Mass extinctions in the marine forrils record. *Science*, 215(4539):1501–1503, 1982.

[104] L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales*, volume one. John Wiley and Sons, Chichester, 1994.

[105] C. F. Ross. Review of 'The Primate Fossil Record'. *Journal of Human Evolution*, 45:195–201, 2003.

[106] R. L. Scheaffer. Size-biased sampling. *Technometrics*, 14(3):635–644, 1972.

[107] G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[108] E. R. Seiffert, E. L. Simons, and Y. Attia. Fossil evidence for an ancient divergence of lorises and galagos. *Nature*, 422:421–424, 2003.

[109] P. W. Signor and J. H. Lipps. Sampling bias, gradual extinction patterns, and catastrophes in the fossil record. In L. T. Silver and P. H. Schultz, editors, *Geological implications of impacts of large asteroids and comets on the Earth*, volume 190, pages 291–296, 1982.

[110] G. G. Simpson. *The Geography of Evolution*. Chilton Books: Philadelphia, 1965.

[111] S. A. Sisson. Genetics and stochastic simulation *do* mix! *The American Statistician*, 61:112–119, 2007.

[112] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104(6):1760–1765, 2007.

[113] C. Soligo, O. Will, S. Tavaré, C. R. Marshall, and R. D. Martin. *Primate Origins: Adaptations and Evolution*, pages 29–49. Springer: New York, 2007.

[114] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, 64(4):583–639, 2002.

[115] D. J. Spiegelhalter, A. Thomas, and N. G. Best. Winbugs version 1.2 user manual. *MRC Biostatistics Unit*, 1999.

[116] G. Suwa, R. T. Kono, S. Katoh, B. Asfaw, and Y. Beyene. A new species of great ape from the late Miocene epoch in Ethiopia. *Nature*, 448:921–924, 2007.

[117] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times for molecular sequence data. *Genetics*, 145:505–518, 1997.

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

[119] L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1762, 1994.

[120] J. von Neumann. Various techniques used in connection with random digits. *National Bureau of Standards Applied Mathematics Series*, 12:36–38, 1951.

[121] S. C. Wang and P. Dobson. Estimating the diversity of dinosaurs. *Proc. Natl. Acad. Sci. USA*, 103(37):13601–13605, 2006.

[122] L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

[123] H. W. Watson and F. Galton. On the probability of the extinction of families. *Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144, 1875.

[124] W. A. O'N. Waugh. Conditioned Markov processes. *Biometrika*, 45(2):241–249, 1958.

[125] O. Will. *Statistical Inference in the Fossil Record*. PhD thesis, University of Southern California, 2001.