Stockholms universitet

A phase-type distribution approach to coalescent theory

Tuure Antonangeli

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2015:3
Matematisk statistik
Maj 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2015:3**
http://www.math.su.se

A phase-type distribution approach to coalescent theory

Tuure Antonangeli*

May 2015

Abstract

A phase-type distribution is defined as the distribution of the time until a continuous time Markov chain with a finite state space reaches an absorbing state. These distributions are fully determined by the infinitesimal generator matrix of the Markov chain, together with a probability vector declaring the probability for the Markov chain to start in a specific state. In this paper, we propose a strategy for analyzing population genetics using phase-type distributions. In particular, we consider a continuous time coalescent process tracking the ancestry of a population sampled at present time, and apply the phase-type approach to determine the properties for such coalescent processes. These properties include the tree height and the total tree length, as well as mutation probabilities. We begin by applying this strategy to the Kingman's coalescent, and we thereafter extend the method to a population complication known as the symmetric island model. The advantage of this approach is that it leads to clear and intuitive derivations for the properties, and we show that the proposed strategy results in a very compact matrix analytic description of the coalescent.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: tuure.antonangeli@hotmail.com. Supervisor: Ola Hössjer.

Acknowledgements

This paper constitutes a master's thesis of 30 ECTS credits in mathematical statistics at the Department of Mathematics at Stockholm University. I would like to thank my supervisor Prof. Ola Hössjer for his help and patience with my never-ending questions, Dr. Pieter Trapman for some very useful comments, as well as all my friends at the department. But above all,

Thank you, Sandra & Elli.

Contents

1 Introduction

Phase-type distributions have, since their introduction in 1975, been used extensively in many fields of applied probability, such as queuing and renewal theory, as well as in insurance risk theory [Bobbio et al., 2003, Asmussen et al., 1996]. However, they have received little or no attention in the field of population genetics. An incitement for extending the applications of phase-type distributions to population genetics is given in [Hössjer et al., 2014], which briefly mentions the existence of a link between phase-type distributions and coalescent theory with general structured populations. Another incitement is given in [Wooding and Rogers, 2002], which motivates the use of matrix-analytic methods when investigating population genetics with variable population size. As will become evident, phase-type distributions and matrix-analytic methods are closely interconnected.

The aim of this thesis is to apply a matrix-analytic framework to coalescent theory using phase-type distributions. In particular, we will derive known results on the Kingman's coalescent tree using a phase-type distribution approach, and we will particularly focus on the distribution and moments of the tree height and the total tree length. Moreover, we will broaden the framework by introducing mutations, as well as by considering a population complication known as the symmetric island model. Although these topics have already been widely discussed, it is still of interest to formalize a connection between coalescent theory and phase-type distributions, since a phase-type distribution approach often results in elegant and compact matrix expressions.

The outline of this study is as follows: We begin by presenting phase-type distributions in Section 2, with a formal definition, and a detailed derivation of their properties, supplemented by a few examples. Thereafter, we provide an introduction to coalescent theory in Section 3, with emphasis on the Kingman's coalescent, mutations and the symmetric island model. Finally, the two theories of phase-type distributions and coalescent theory are joined together in Section 4. The material in sections 2 and 3 is self contained with a thorough introduction to the subjects, and no prior knowledge in either phase-type distributions or coalescent theory is necessary. However, for the reader interested in a better insight in the two fields, we refer to e.g. [Latouche and Ramaswami, 1999] or [Neuts, 1981] on phase-type distributions, and [Durrett, 2008] or [Hein et al., 2005] on coalescent theory.

2 Phase-Type Distributions

Phase-type distributions are a class of continuous distributions[1] with support on $[0, \infty)$, which describe the time until a continuous time Markov chain with a finite state space reaches an absorbing state. Equivalently, if a distribution is of phase-type, then there

[1] Phase-type distributions have a discrete-time equivalent, which represents the number of transitions until a discrete-time Markov chain reaches an absorbing state. Discrete-time phase-type distributions will not be discussed in this study, and by a phase-type distribution, we will explicitly refer to the continuous case.

Table 1: Notation for selected matrix quantities.

Symbol	Definition
$\mathbf{X}$, $\mathbf{x}$	A matrix or a (column) vector
$\mathbf{X}'$	Transpose of $\mathbf{X}$
$\mathbf{X}_{ij}$	Entry i, j of $\mathbf{X}$
$\mathbf{X}_{:j}$	j^{th} column of $\mathbf{X}$
$\mathbf{X}_{:i..j}$	Columns i through j of $\mathbf{X}$
$\mathbf{I}_{n\times n}$	Identity matrix of order n
$\mathbf{1}_n$ ($\mathbf{0}_n$)	Column vector of length n, where each entry equals 1 (0)
$\mathbf{0}_{n\times n}$	$n \times n$ matrix, where each entry equals 0

exists an intrinsic continuous time Markov chain representation of the distribution.

We will demonstrate using matrix-analytic methods (see e.g. [Latouche and Ramaswami, 1999, Bellman, 1960]) that the properties of phase-type distributions, such as the density and distribution function, as well as the moments, are fully characterized by 1) the generator matrix of the underlying Markov chain and 2) a probability vector stating the probabilities for the initial state. In fact, without the use of matrix notation it is generally difficult to find simple, closed form expressions for the properties of phase-type distributed random variables. Using the fact that a distribution is of phase-type, awkward integral computations become convenient matrix expressions.

2.1 Definition

Consider a continuous time Markov chain (CTMC) $\{X(t);\ t \geq 0\}$ with a finite state space $S = \{0, 1, 2, ..., n\}$, in which $\{0\}$ is an absorbing state, whereas states $\{1, 2, ..., n\}$ are transient. In other words, $\{X(t);\ t \geq 0\}$ has the properties that the total time it spends in state $i \in \{1, 2, ..., n\}$ is almost surely bounded, and once $\{X(t);\ t \geq 0\}$ reaches state $\{0\}$, it never leaves this state [Norris, 1997]. With such a state space, the generator matrix $\mathbf{Q}$ for the CTMC can be block partitioned as

$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{0}_n' \\ \mathbf{t} & \mathbf{T} \end{pmatrix} \tag{2.1}$$

where $\mathbf{0}_n$ is a column vector of order n with all entries equal to zero (for a complete description of the matrix notation used in this study, see Table 1), and $\mathbf{t}$ is a column vector of order n in which the i^{th} entry equals the conditional intensity for the chain to enter the absorbing state $\{0\}$, when the chain is in the transient state $\{i\}$. Moreover, $\mathbf{T}$ is an $n \times n$ matrix in which the entries $\mathbf{T}_{i,j\neq i}$, $i, j \in \{1, ..., n\}$, are the transition rates from state i to state j and the negative diagonal elements, $-\mathbf{T}_{ii}$, are the rates for leaving state i. Let $\tilde{\boldsymbol{\alpha}} = (\alpha_0, ..., \alpha_n)'$ be the initial probability vector, i.e. the vector where α_i is the probability that the CTMC starts in state $i \in S$, and define $\boldsymbol{\alpha} := (\alpha_1, ..., \alpha_n)'$ as the vector containing the initial probabilities for the transient states. Observe that, since

the row sums of $\mathbf{Q}$ are zero, and since $\tilde{\boldsymbol{\alpha}}$ is a probability vector, the conditions

$$\mathbf{T}\mathbf{1}_n + \mathbf{t} = \mathbf{0}_n \tag{2.2a}$$
$$\tilde{\boldsymbol{\alpha}}'\mathbf{1}_{n+1} = 1 \tag{2.2b}$$

clearly hold.

Now let T be a random variable denoting the time it takes for the CTMC $\{X(t);\ t \geq 0\}$ specified above to reach the absorbing state $\{0\}$. The distribution of T is called a *Phase-type distribution* [Neuts, 1981, Latouche and Ramaswami, 1999]. This class of continuous distributions is formalized in Definition 2.1 below:

Definition 2.1. *The distribution of the time T until a continuous time Markov chain $\{X(t),\ t \geq 0\}$ with a generator matrix given in Equation (2.1) reaches the absorbing state $\{0\}$ is called a* phase-type distribution *with representation $(\boldsymbol{\alpha}, \mathbf{T})$. We write $T \sim PH(\boldsymbol{\alpha}, \mathbf{T})$.*

2.2 Density and distribution

The cumulative distribution function (CDF) for the random variable T given in Definition 2.1 can, by conditioning on the initial state, be computed as

$$\begin{aligned} F_T(t) &= \mathrm{P}(T \leq t) \\ &= \mathrm{P}(X(t) = 0) \\ &= \sum_{i=0}^{n} \mathrm{P}(X(0) = i)\ \mathrm{P}(X(t) = 0 | X(0) = i) \\ &= \sum_{i=0}^{n} \alpha_i\ P_{i0}(t) = \tilde{\boldsymbol{\alpha}}'\ \mathbf{P}_{:1}(t) \end{aligned}$$

where $\mathbf{P}_{:1}(t)$ is the first column of the matrix $\mathbf{P}(t)$ with entries

$$P_{ij}(t) = \mathrm{P}(X(t) = j | X(0) = i), \tag{2.3}$$

$i, j \in S$. The matrix $\mathbf{P}(t)$ satisfies the *backward differential equation*

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{Q}\mathbf{P}(t), \tag{2.4}$$

solved by

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{Q}^k t^k, \tag{2.5}$$

which is well defined for all square matrices $\mathbf{Q}$ [Norris, 1997, Ch 2.10], [Resnick, 2002, Ch. 5.4]. In order to derive $\mathbf{P}(t)$ and $\mathbf{P}_{:1}(t)$, and thus also $F_T(t)$, we make the following proposition:

Proposition 2.1. *For the matrix* $\mathbf{Q}$ *defined in Equation (2.1), we have*

$$\mathbf{Q}^m = \begin{pmatrix} 0 & \mathbf{0}_n' \\ -\mathbf{T}^m\mathbf{1}_n & \mathbf{T}^m \end{pmatrix}$$

for $m = 1, 2, ...$

Proof. The proof is made by induction:

- For $m = 1$ we have

$$\mathbf{Q}^1 = \begin{pmatrix} 0 & \mathbf{0}_n' \\ -\mathbf{T}^1\mathbf{1}_n & \mathbf{T}^1 \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{0}_n' \\ \mathbf{t} & \mathbf{T} \end{pmatrix}$$

  by Condition (2.2a).

- Assume the proposition is true for $m = k$, so that

$$\mathbf{Q}^k = \begin{pmatrix} 0 & \mathbf{0}_n' \\ -\mathbf{T}^k\mathbf{1}_n & \mathbf{T}^k \end{pmatrix}.$$

- For $m = k + 1$ we then have

$$\begin{aligned} \mathbf{Q}^{k+1} &= \mathbf{Q}^k \cdot \mathbf{Q} \\ &= \begin{pmatrix} 0 & \mathbf{0}_n' \\ -\mathbf{T}^k\mathbf{1}_n & \mathbf{T}^k \end{pmatrix} \cdot \begin{pmatrix} 0 & \mathbf{0}_n' \\ \mathbf{t} & \mathbf{T} \end{pmatrix} \\ &= \begin{pmatrix} 0 & \mathbf{0}_n' \\ \mathbf{T}^k\mathbf{t} & \mathbf{T}^{k+1} \end{pmatrix} \\ &= \begin{pmatrix} 0 & \mathbf{0}_n' \\ -\mathbf{T}^k\mathbf{T}\mathbf{1}_n & \mathbf{T}^{k+1} \end{pmatrix} \\ &= \begin{pmatrix} 0 & \mathbf{0}_n' \\ -\mathbf{T}^{k+1}\mathbf{1}_n & \mathbf{T}^{k+1} \end{pmatrix}, \end{aligned}$$

which completes the proof. □

If we let $\mathbf{I}_{(n+1)\times(n+1)}$ denote the $(n+1) \times (n+1)$ identity matrix, the solution to Equation (2.4) becomes

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{Q}^k t^k = \mathbf{I}_{(n+1)\times(n+1)} + \sum_{k=1}^{\infty} \frac{1}{k!} \begin{pmatrix} 0 & \mathbf{0}_n' \\ -\mathbf{T}^k\mathbf{1}_n & \mathbf{T}^k \end{pmatrix} t^k$$

by Proposition 2.1. Inserting the sum and all the summands into the matrix yields

$$\begin{aligned}\mathbf{P}(t) &= \mathbf{I}_{(n+1)\times(n+1)} + \begin{pmatrix} 0 & \mathbf{0}_n' \\ -\left\{\sum_{k=1}^{\infty} \frac{1}{k!}(\mathbf{T}t)^k\right\}\mathbf{1}_n & \sum_{k=1}^{\infty} \frac{1}{k!}(\mathbf{T}t)^k \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0}_n' \\ -\left\{\sum_{k=1}^{\infty} \frac{1}{k!}(\mathbf{T}t)^k\right\}\mathbf{1}_n & \sum_{k=0}^{\infty} \frac{1}{k!}(\mathbf{T}t)^k \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0}_n' \\ \mathbf{1}_n - e^{\mathbf{T}t}\mathbf{1}_n & e^{\mathbf{T}t} \end{pmatrix}, \end{aligned} \tag{2.6}$$

indicating that the first column of the matrix $\mathbf{P}(t) = e^{\mathbf{Q}t}$ equals

$$\mathbf{P}_{:1}(t) = \begin{pmatrix} 1 \\ \mathbf{1}_n - e^{\mathbf{T}t}\mathbf{1}_n \end{pmatrix}.$$

Consequently, the CDF for the time until absorption equals

$$\begin{aligned} F_T(t) &= \tilde{\boldsymbol{\alpha}}' \ \mathbf{P}_{:1}(t) \\ &= \alpha_0 + (\alpha_1, \alpha_2, ..., \alpha_n)(\mathbf{1}_n - e^{\mathbf{T}t}\mathbf{1}_n) \\ &= \alpha_0 + (\alpha_1, \alpha_2, ..., \alpha_n)\mathbf{1}_n - (\alpha_1, \alpha_2, ..., \alpha_n)e^{\mathbf{T}t}\mathbf{1}_n \end{aligned}$$

which, by applying Condition (2.2b) simplifies to

$$F_T(t) = 1 - \boldsymbol{\alpha}' e^{\mathbf{T}t}\mathbf{1}_n \tag{2.7}$$

for $t \geq 0$ and $\boldsymbol{\alpha}' = (\alpha_1, ..., \alpha_n)$. Note that for the derivation of $F_T(t)$ it is of no concern whether or not we assume that $\alpha_0 = 0$, although this generally is the case. For an alternative derivation of $F_T(t)$ see e.g. [Neuts, 1981].

Taking the derivative of Equation (2.7) with respect to t and using Condition (2.2a) gives the probability density function (PDF) for $T \sim \text{PH}(\boldsymbol{\alpha}, \mathbf{T})$ as

$$f_T(t) = \boldsymbol{\alpha}' e^{\mathbf{T}t}\mathbf{t}, \tag{2.8}$$

for $t \geq 0$.

2.3 Laplace transform and moments

In order to derive the Laplace transform of a phase type distributed random variable, we begin by making the following proposition:

Proposition 2.2. *If the matrix* $\mathbf{T}$ *given in Equation (2.1) is invertible, then* $\lim_{t\to\infty} e^{\mathbf{T}t} = \mathbf{0}_{n\times n}$*, where* $\mathbf{0}_{n\times n}$ *is the* $n \times n$ *matrix with all elements equal to zero.*

Proof. [Latouche and Ramaswami, 1999].

Suppose that the matrix $\mathbf{T}$ is invertible and define $\boldsymbol{\xi}(t) = (\xi_1(t), ..., \xi_n(t))' = e^{\mathbf{T}t}\mathbf{1}_n$, i.e. $\boldsymbol{\xi}(t)$ is a column vector of order n containing the row sums of $e^{\mathbf{T}t}$. Since the row sums of (2.6) equal 1, it follows that the row sums in $e^{\mathbf{T}t}$ are in $[0, 1]$ for all $t \geq 0$. Therefore,

$$0 \leq \xi_i(t) \leq 1 \tag{2.9}$$

for all $i = 1, ..., n$, and for all $t \geq 0$. Furthermore,

$$e^{\mathbf{T}t} = \mathbf{I}_{n\times n} + \int_{x=0}^{t} \mathbf{T}e^{\mathbf{T}x}\, dx, \tag{2.10}$$

since

$$\frac{d}{dt}e^{\mathbf{T}t} = \frac{d}{dt}\int_{x=0}^{t} \mathbf{T}e^{\mathbf{T}x}\, dx,$$

and since $e^{\mathbf{T}t}\big|_{t=0} = \mathbf{I}_{n\times n}$. Multiplying both sides of Equation (2.10) by $\mathbf{T}^{-1}$ from the left and by $\mathbf{1}_n$ from the right, we obtain

$$\begin{aligned}
&\mathbf{T}^{-1}e^{\mathbf{T}t}\mathbf{1}_n = \mathbf{T}^{-1}\left(\mathbf{I}_{n\times n} + \int_{x=0}^{t} \mathbf{T}e^{\mathbf{T}x}\, dx\right)\mathbf{1}_n \\
\implies &\mathbf{T}^{-1}\boldsymbol{\xi}(t) = \mathbf{T}^{-1}\mathbf{I}_{n\times n}\mathbf{1}_n + \mathbf{T}^{-1}\mathbf{T}\left(\int_{x=0}^{t} e^{\mathbf{T}x}\, dx\right)\mathbf{1}_n \\
\implies &\mathbf{T}^{-1}\boldsymbol{\xi}(t) = \mathbf{T}^{-1}\mathbf{1}_n + \left(\int_{x=0}^{t} e^{\mathbf{T}x}\, dx\right)\mathbf{1}_n.
\end{aligned} \tag{2.11}$$

Since by assumption of invertibility $|\mathbf{T}^{-1}_{ij}| < \infty$ for all $i, j = 1, ..., n$, and since (2.9) holds, the entries in the column vector on the left hand side of Equation (2.11) are bounded for all $t \geq 0$. However, due to the fact that the entries in $e^{\mathbf{T}x}$ are non-negative, the entries in

$$\int_{x=0}^{t} e^{\mathbf{T}x}\, dx$$

are increasing in $t \geq 0$, but bounded, since the left hand side is bounded and $\mathbf{T}^{-1}\mathbf{1}_n$ is constant in t. Consequently, since $e^{\mathbf{T}x}$ is a linear combination of exponential functions, we have

$$\lim_{t\to\infty} e^{\mathbf{T}t} = \mathbf{0}_{n\times n}.$$

□

The Laplace transform for $T \sim PH(\boldsymbol{\alpha}, \mathbf{T})$, where $\mathbf{T}$ is assumed to be invertible, is defined as

$$\mathscr{L}_T(s) = \mathrm{E}\left[e^{-sT}\right] = \int_{t=0}^{\infty} f_T(t)\ e^{-st}\ dt$$

for $s \geq 0$, which equals

$$\begin{aligned}
\mathscr{L}_T(s) &= \int_{t=0}^{\infty} \boldsymbol{\alpha}' e^{\mathbf{T}t}\mathbf{t}\ e^{-st}\ dt \\
&= \boldsymbol{\alpha}' \int_{t=0}^{\infty} e^{(\mathbf{T}-s\mathbf{I}_{n\times n})t}\ dt\ \mathbf{t} \\
&= \boldsymbol{\alpha}'(\mathbf{T}-s\mathbf{I}_{n\times n})^{-1}\left[e^{(\mathbf{T}-s\mathbf{I}_{n\times n})t}\right]_{t=0}^{\infty}\mathbf{t} \\
&= \boldsymbol{\alpha}'(\mathbf{T}-s\mathbf{I}_{n\times n})^{-1}\left[e^{\mathbf{T}t}e^{-s\mathbf{I}_{n\times n}t}\right]_{t=0}^{\infty}\mathbf{t}.
\end{aligned}$$

Note that since the real part of each eigenvalue of $\mathbf{T}$ is negative [Latouche and Ramaswami, 1999, p.44], the inverse $(\mathbf{T}-s\mathbf{I}_{n\times n})^{-1}$ exists for all $s \geq 0$. By Proposition 2.2, we have

$$\lim_{t\to\infty} e^{\mathbf{T}t}e^{-s\mathbf{I}_{n\times n}t} = \mathbf{0}_{n\times n}$$

for $s \geq 0$, whereas $e^{\mathbf{T}t}e^{-s\mathbf{I}_{n\times n}t}\big|_{t=0} = \mathbf{I}_{n\times n}$. Hence, the Laplace transform for a phase-type distributed random variable T equals

$$\mathscr{L}_T(s) = \boldsymbol{\alpha}'(s\mathbf{I}_{n\times n}-\mathbf{T})^{-1}\mathbf{t}, \quad s \geq 0.$$

The derivative of $\mathscr{L}_T(s)$ with respect to s is given by

$$\begin{aligned}
\frac{d\mathscr{L}_T(s)}{ds} &= \boldsymbol{\alpha}'\frac{d}{ds}\Big\{(s\mathbf{I}_{n\times n}-\mathbf{T})^{-1}\Big\}\mathbf{t} \\
&= -\boldsymbol{\alpha}'(s\mathbf{I}_{n\times n}-\mathbf{T})^{-1}\frac{d}{ds}\Big\{s\mathbf{I}_{n\times n}-\mathbf{T}\Big\}(s\mathbf{I}_{n\times n}-\mathbf{T})^{-1}\mathbf{t} \\
&= -\boldsymbol{\alpha}'(s\mathbf{I}_{n\times n}-\mathbf{T})^{-1}\ \mathbf{I}_{n\times n}\ (s\mathbf{I}_{n\times n}-\mathbf{T})^{-1}\mathbf{t} \\
&= -\boldsymbol{\alpha}'(s\mathbf{I}_{n\times n}-\mathbf{T})^{-2}\ \mathbf{t},
\end{aligned}$$

and it is readily checked that

$$\frac{d^k\mathscr{L}_T(s)}{ds^k} = (-1)^k k!\ \boldsymbol{\alpha}'(s\mathbf{I}_{n\times n}-\mathbf{T})^{-k-1}\ \mathbf{t} \tag{2.12}$$

for $k \geq 1$. Thus, by Equation (2.12) we obtain The k^{th} moment of T as

$$\mathrm{E}[T^k] = (-1)^k \frac{d^k \mathscr{L}_T(s)}{ds^k}\bigg|_{s=0} = (-1)^k k!\ \boldsymbol{\alpha}'\ \mathbf{T}^{-k}\mathbf{1}_n \tag{2.13}$$

for $k \geq 1$.

2.4 Examples

Example 2.1. *Exponential distribution*

Consider a CTMC $\{X(t),\ t \geq 0\}$ with state space $\{0, 1\}$ and with a generator matrix

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 \\ \lambda & -\lambda \end{pmatrix}, \tag{2.14}$$

where $\lambda > 0$, and assume that the initial state is $\{1\}$ with probability 1, i.e. that $\tilde{\boldsymbol{\alpha}}' = (0, 1)$, and hence $\boldsymbol{\alpha}' = (1)$. The matrix $\mathbf{Q}$ can be partitioned according to Equation (2.1) with $\mathbf{T} = (-\lambda)$ and $\mathbf{t} = (\lambda)$. Furthermore, let T be the time it takes for the Markov chain to reach the absorbing state $\{0\}$, so that $T \sim PH(1, -\lambda)$. The distribution, density and expected value of T are given by equations (2.7), (2.8) and (2.13) respectively:

$$\begin{aligned} F_T(t) &= 1 - \boldsymbol{\alpha}' e^{\mathbf{T}t}\mathbf{1}_1 = 1 - e^{-\lambda t}, \quad t \geq 0, \\ f_T(t) &= \boldsymbol{\alpha}' e^{\mathbf{T}t}\mathbf{t} = \lambda e^{-\lambda t}, \quad t \geq 0, \\ \mathrm{E}[T] &= (-1)^1\ 1!\ \boldsymbol{\alpha}'\mathbf{T}^{-1}\mathbf{1}_1 = \tfrac{1}{\lambda}. \end{aligned}$$

We recognize these as defining properties of the exponential distribution with an intensity parameter λ, that is, $T \sim \mathrm{Exp}(\lambda)$. We conclude that the exponential distribution is a phase-type distribution with an inherent CTMC representation.

Example 2.2. *Erlang distribution*

Consider a CTMC $\{X(t),\ t \geq 0\}$ with state space $\{0, 1, ..., n\}$ and with a generator matrix

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & -\lambda & \lambda & 0 & \dots & 0 \\ 0 & 0 & -\lambda & \lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda & \lambda \\ \lambda & 0 & 0 & \dots & 0 & -\lambda \end{pmatrix}, \tag{2.15}$$

and assume that the initial state is $\{1\}$ with probability 1, i.e. that $\tilde{\boldsymbol{\alpha}}' = (0, 1, \mathbf{0}'_{n-1})$ and hence $\boldsymbol{\alpha}' = (1, \mathbf{0}'_{n-1})$. The matrix $\mathbf{T}$ in this case equals

$$\mathbf{T} = \begin{pmatrix} -\lambda & \lambda & 0 & \dots & 0 \\ 0 & -\lambda & \lambda & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -\lambda & \lambda \\ 0 & 0 & \dots & 0 & -\lambda \end{pmatrix},$$

whereas $\mathbf{t}' = (\mathbf{0}'_{n-1}, \lambda)$. It is easy to verify that

$$\boldsymbol{\alpha}'\mathbf{T}^k\mathbf{t} = (-1)^{k+n-1}\binom{k}{n-1}\lambda^{k+1}$$

for $k = n-1, n, n+1, ...$ and 0 otherwise. By Equation (2.8), the density function for the time it takes for the CTMC $\{X(t),\ t \geq 0\}$ to reach the absorbing state $\{0\}$ thus equals

$$\begin{aligned} f_T(t) &= \boldsymbol{\alpha}' \sum_{k=0}^{\infty} \frac{1}{k!}\ (\mathbf{T}t)^k\ \mathbf{t} \\ &= \sum_{k=n-1}^{\infty} \frac{1}{k!}(-1)^{k+n-1}\binom{k}{n-1}\lambda^{k+1}t^k \\ &= \sum_{k=n-1}^{\infty} \frac{1}{k!}(-1)^{k+n-1}\frac{k!}{(n-1)!(k-n+1)!}\lambda^{k+1}t^k \\ &= \frac{\lambda}{(n-1)!}\sum_{k=n-1}^{\infty}(-1)^{k+n-1}\frac{1}{(k-n+1)!}(\lambda t)^k. \end{aligned}$$

Shifting the summation index by setting $l = k - n + 1$ gives

$$\begin{aligned} f_T(t) &= \frac{\lambda}{(n-1)!}\sum_{l=0}^{\infty}(-1)^{l+2(n-1)}\ \frac{1}{l!}\ (\lambda t)^{l+n-1} \\ &= \frac{\lambda^n}{(n-1)!}t^{n-1}\sum_{l=0}^{\infty}(-1)^l\frac{1}{l!}(\lambda t)^l \\ &= \frac{\lambda^n}{(n-1)!}\ t^{n-1}\ e^{-\lambda t} \end{aligned}$$

for $t \geq 0$, which we recognize as the density function of an Erlang distribution with parameters n and λ. Thus we conclude that, just as the exponential distribution, the Erlang distribution is of phase-type, with an underlying CTMC representation.

2.5 Multiple absorbing states

In this section we will demonstrate that the theory of phase-type distributions discussed in sections 2.2 and 2.3, can be extended to the case where the underlying Markov chain possesses $m \geq 2$ absorbing states. Consider thus a CTMC $\{X(t);\ t \geq 0\}$ with a finite state space $S = \{1, 2, ..., m+n\}$, such that the first m states are absorbing, whereas the remaining n states are transient. The state space S can therefore be partitioned into the set of absorbing states $\underline{S} = \{1, ..., m\}$ with $|\underline{S}| = m$, and the set of transient states $\overline{S} = \{m+1, ..., m+n\}$ with $|\overline{S}| = n$. Consequently, with multiple absorbing states, the corresponding intensity matrix can be block partitioned as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{0}_{m\times m} & \mathbf{0}_{m\times n} \\ \mathbf{t} & \mathbf{T} \end{pmatrix},$$

where $\mathbf{t}$ is an $n \times m$ matrix, in which element $\mathbf{t}_{ij}$, $i = 1, ..., n$, $j = 1, ..., m$ is the conditional intensity for the Markov chain to enter the absorbing state $\{j\}$, when the Markov chain is in the transient state $\{i\}$. $\mathbf{T}$ in turn, is an $n \times n$ matrix where the entries $\mathbf{T}_{k,l\neq k}$, $k, l \in \overline{S}$, are the conditional intensities for transitions between the transient states, and where the negative diagonal entries $-\mathbf{T}_{kk}$ are the intensities for leaving state $k \in \overline{S}$. We assume for our purposes that $\mathbf{T}$ is invertible.

As in the case with only one absorbing state, we let $\tilde{\boldsymbol{\alpha}} = (\alpha_1, ..., \alpha_{m+n})'$ be the initial probability vector, i.e. the vector where element α_i is the probability that the initial state for the Markov chain is $\{i\}$. Furthermore, we define $\boldsymbol{\alpha} := (\alpha_{m+1}, ..., \alpha_{m+n})'$ as the initial probability vector for the transient states. We note that Conditions (2.2a) (2.2b) modify to

$$\mathbf{T}\mathbf{1}_n + \mathbf{t}\mathbf{1}_m = \mathbf{0}_n \tag{2.16a}$$

$$\tilde{\boldsymbol{\alpha}}'\mathbf{1}_{m+n} = 1 \tag{2.16b}$$

since the rows of $\mathbf{Q}$ sum to 0, and since $\tilde{\boldsymbol{\alpha}}$ is a probability vector.

Suppose, as before, that the random variable T denotes the time it takes for $\{X(t);\ t \geq 0\}$ to reach $\underline{S}$, assuming that the initial state is given by the probability vector $\tilde{\boldsymbol{\alpha}}$. The CDF for T equals

$$F_T(t) = \mathrm{P}(T \leq t) = \mathrm{P}(X(t) \in \underline{S}).$$

By conditioning on the initial state $X(0)$, we obtain

$$
\begin{aligned}
F_T(t) &= \sum_{j \in S} \mathrm{P}(X(0) = j)\ \mathrm{P}(X(t) \in \underline{S} | X(0) = j) \\
&= \sum_{j \in S} \sum_{k \in \underline{S}} \mathrm{P}(X(0) = j)\ \mathrm{P}(X(t) = k | X(0) = j) \\
&= \sum_{j \in S} \sum_{k \in \underline{S}} \alpha_j\ P_{jk}(t) \\
&= \tilde{\boldsymbol{\alpha}}'\ \mathbf{P}_{:1..m}(t)\ \mathbf{1}_m,
\end{aligned}
$$

where $\mathbf{P}_{:1..m}(t)$ denotes the matrix consisting of the first m columns of the matrix $\mathbf{P}(t)$ with elements $P_{ij}(t)$ defined in Equation (2.3). In order to derive $\mathbf{P}_{:1..m}(t)$, we note that

$$
\mathbf{Q}^k = \begin{pmatrix} \mathbf{0}_{m \times m} & \mathbf{0}_{m \times n} \\ \mathbf{T}^{k-1}\mathbf{t} & \mathbf{T}^k \end{pmatrix}. \tag{2.17}
$$

The proof of Equation (2.17) is done analogously to the proof of Proposition 2.1 and will therefore be omitted.

The matrix $\mathbf{P}(t)$ is, according to Equation (2.5), obtained as

$$
\mathbf{P}(t) = \mathbf{I}_{(m+n)\times(m+n)} + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{Q}^k t^k.
$$

Using Equation (2.17) we obtain

$$
\begin{aligned}
\mathbf{P}(t) &= \mathbf{I}_{(m+n)\times(m+n)} + \sum_{k=1}^{\infty} \frac{1}{k!} \begin{pmatrix} \mathbf{0}_{m \times m} & \mathbf{0}_{m \times n} \\ \mathbf{T}^{k-1}\mathbf{t} & \mathbf{T}^k \end{pmatrix} t^k \\
&= \mathbf{I}_{(m+n)\times(m+n)} + \begin{pmatrix} \mathbf{0}_{m \times m} & \mathbf{0}_{m \times n} \\ \sum_{k=1}^{\infty} \left\{ \frac{1}{k!} \mathbf{T}^k t^k \right\} \mathbf{T}^{-1}\mathbf{t} & \sum_{k=1}^{\infty} \left\{ \frac{1}{k!} \mathbf{T}^k t^k \right\} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}_{m \times m} & \mathbf{0}_{m \times n} \\ \left(e^{\mathbf{T}t} - \mathbf{I}_{n \times n}\right)\mathbf{T}^{-1}\mathbf{t} & e^{\mathbf{T}t} \end{pmatrix},
\end{aligned} \tag{2.18}
$$

and we recognize that the first m columns of $\mathbf{P}(t)$ thus equal

$$
\mathbf{P}_{:1..m}(t) = \begin{pmatrix} \mathbf{I}_{m \times m} \\ \left(e^{\mathbf{T}t} - \mathbf{I}_{n \times n}\right)\mathbf{T}^{-1}\mathbf{t} \end{pmatrix}.
$$

Therefore we obtain

$$
\begin{aligned}
F_T(t) &= \tilde{\boldsymbol{\alpha}}' \, \mathbf{P}_{:1..m}(t) \, \mathbf{1}_m \\
&= \tilde{\boldsymbol{\alpha}}' \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \big(e^{\mathbf{T}t} - \mathbf{I}_{n\times n}\big)\mathbf{T}^{-1}\mathbf{t}\mathbf{1}_m \end{pmatrix} \\
&= \tilde{\boldsymbol{\alpha}}' \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \big(\mathbf{I}_{n\times n} - e^{\mathbf{T}t}\big)\mathbf{1}_n \end{pmatrix} \\
&= \alpha_1 + \cdots + \alpha_m + (\alpha_{m+1}, ..., \alpha_{m+n})\big(\mathbf{I}_{n\times n} - e^{\mathbf{T}t}\big)\mathbf{1}_n \\
&= 1 - \boldsymbol{\alpha}' e^{\mathbf{T}t}\mathbf{1}_n,
\end{aligned}
$$

which agrees with Equation (2.7). Hence, $T \sim \mathrm{PH}(\boldsymbol{\alpha}, \mathbf{T})$ as before. However, since Condition 2.2a is no longer valid, the density function for T is not identical to Equation (2.8) when the number of absorbing states is $m \geq 2$, but rather

$$f_T(t) = \boldsymbol{\alpha}' e^{\mathbf{T}t}\mathbf{t}\mathbf{1}_m, \tag{2.19}$$

by Equation (2.16a). Computing the Laplace transform similarly as in Section 2.3, by using Equation (2.19) instead of Equation (2.8), it is easy to verify that the moments of T are still given by Equation (2.13).

2.6 Absorption probabilities

For the purpose of this study, when the number of absorbing states is $m \geq 2$, it is also of interest to derive the absorption probability vector $\boldsymbol{\phi}$, with entries

$$\phi_j = \mathrm{P}(X(T) = j)$$

for $j \in \underline{S}$, that is, the vector in which the j^{th} entry is the probability that the CTMC $\{X(t);\ t \geq 0\}$ is absorbed into state j. To do this, we note that for a transient state $i \in \overline{S}$ and an absorbing state $j \in \underline{S}$, we have by Equation (2.18)

$$P_{ij}(t) = \Big[\big(e^{\mathbf{T}t} - \mathbf{I}_{n\times n}\big)\,\mathbf{T}^{-1}\mathbf{t}\Big]_{ij},$$

and since

$$\frac{d}{dt}\big(e^{\mathbf{T}t} - \mathbf{I}_{n\times n}\big)\,\mathbf{T}^{-1}\mathbf{t} = e^{\mathbf{T}t}\ \mathbf{T}\ \mathbf{T}^{-1}\mathbf{t} = e^{\mathbf{T}t}\mathbf{t},$$

we have

$$P'_{ij}(t) = \left[e^{\mathbf{T}t}\mathbf{t}\right]_{ij}.$$

Thus we can write

$$\begin{aligned}\phi_j &= \sum_{i\in\underline{S}} \alpha_i \delta_{ij} + \sum_{i\in\overline{S}} \alpha_i \int_{t=0}^{\infty} P'_{ij}(t)\ dt \\ &= \alpha_i \mathbb{I}(j\in\underline{S}) + \sum_{i\in\overline{S}} \alpha_i \int_{t=0}^{\infty} \left[e^{\mathbf{T}t}\mathbf{t}\right]_{ij}\ dt \\ &= \alpha_i \mathbb{I}(j\in\underline{S}) + \sum_{i\in\overline{S}} \alpha_i \left[\int_{t=0}^{\infty} e^{\mathbf{T}t} dt\ \mathbf{t}\right]_{ij}, \end{aligned} \tag{2.20}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. By Proposition 2.2, the integral in (2.20) reduces to

$$\int_{t=0}^{\infty} e^{\mathbf{T}t} dt = -\mathbf{T}^{-1},$$

implying that

$$\phi_j = \alpha_i \mathbb{I}(j\in\underline{S}) - \sum_{i\in\overline{S}} \alpha_i \left[\mathbf{T}^{-1}\mathbf{t}\right]_{ij}. \tag{2.21}$$

In matrix notation, Equation (2.21) can for all $j \in \underline{S}$ be written as

$$\boldsymbol{\phi} = (\alpha_1, ..., \alpha_{m+n}) \begin{pmatrix} \mathbf{I}_{m\times m} \\ -\mathbf{T}^{-1}\mathbf{t} \end{pmatrix},$$

and in particular, if we set $\alpha_i = 0$ for all $i \in \underline{S}$, i.e. discarding the possibility for the Markov chain to start in an absorbing state, we remain with the simple form

$$\boldsymbol{\phi} = -\boldsymbol{\alpha}'\mathbf{T}^{-1}\mathbf{t}. \tag{2.22}$$

3 Coalescent theory

3.1 Preliminaries

This section (3.1) provides a brief introduction to the concepts and terminology in population genetics that are needed in this study. For a more detailed account on basic genetics, see e.g. [Khoury et al., 1993, Section 1] and [Durrett, 2008, Chapter 1.1].

The DNA (deoxyribonucleic acid) contains the hereditary information of an organism. It can, for our purposes, be viewed as a non-random sequence of four different nucleotides; adenine (A), guanine (G), cytosine (C) and thymine (T). This sequence is

contained mainly[2] in the chromosomes of the organism, which in turn lie within the nucleus of a cell. The DNA contains distinct genetic *loci*, that is, subsequences of nucleotides which have the specific task of producing enzymes or structural proteins. These subsequences will for simplicity be referred to as *genes*. The length of the DNA sequence, and the number of genes within the DNA, varies for different organisms. For instance, the human DNA is made up of approximately $3 \cdot 10^9$ nucleotides, and the subsequences within genes consist of a few thousands or ten thousands of nucleotides.

Haploid organisms contain one single copy of its genetic material, and reproduce in such way that the DNA is copied to the offspring from one and only one parent. These are typically more primitive, single-cell organisms such as bacteria. More complex organisms, such as humans, are *diploid*. Diplod organisms require two parents for reproduction and thus contain two sets of genetic information, one from each parent. Some plants are *polyploid*, meaning that they contain multiple sets of DNA. In this study we will for simplicity restrict ourselves to haploid organisms, although it is not difficult to extend the modeling to the diploid case [Durrett, 2008, Ch. 1.2] [Hein et al., 2005, Sect. 1.4].

The objective of coalescent theory is to trace the ancestry of a gene throughout generations backwards in time, by means of stochastic modeling [Hein et al., 2005]. More precisely, one considers a particular genetic locus within the DNA of distinct individuals sampled from a large population at present time. Thereafter, given some probabilistic population dynamics such as variations in offspring size, the theory examines the behaviour of the corresponding genealogies, until the most recent common ancestor (MRCA) of the sampled genes is found.

In the modeling, we will initially consider a population of haploid individuals with non-overlapping generations of equal size N. This population is assumed to lack any kind of social or geographical structure, and in each generation the individuals are all assumed equally likely to survive and to produce offspring. These assumptions provide a model for population genetics known as the *Wright-Fisher model*, introduced by S. Wright and R. Fisher in the early 1930's [Hein et al., 2005]. Although highly idealized, the Wright-Fisher model is very appealing and widely used in mathematical modeling due to its simplicity. For instance, under the Wright-Fisher model, k individuals randomly selected from generation g share a common parent in generation $g-1$ with probability $1/N^{k-1}$.

In section 3.2, we will closely examine the properties of the genealogical tree that results from sampling n genes in generation 0 (present time), and thereafter tracing their parental genes throughout preceding generations, assuming a Wright-Fisher model. In particular, we will demonstrate that when the tree is observerd in continuous time, it manifests exponentially distributed times until the lineages of two genes meet. The continuous time tree that arises from the Wright-Fisher model is known as the *Kingman's coalescent*.

After the Kingman's coalescent has been presented, Section 3.3 introduces *mutations* as a component in the coalescent process. A mutation will be assumed to be a random interchange of one nucleotide into another within a gene, and this interchange occurs as

[2]Excluding mitochondrial DNA (mtDNA), residing in the mitochondria.

an error when the haploid organism copies its DNA in order to produce offspring. If the alternation in nucleotides changes the type of enzyme or structural protein that the gene is coded for, the mutation is said to be *nonsynonymous*, thus altering a characteristic of the organism. We will demonstrate how the exponential distribution plays a key role also in the occurrence of mutations.

Finally, in Section 3.4, we will introduce a population complication known as the *symmetric island model*, where the organism population is subdivided into distinct islands. The organisms are then assumed to migrate between the islands according to some random process. Although the Wright-Fisher model is in itself no longer valid, since the Wright-Fisher model assumes the absence of geographical structure, it is assumed that the conditions for the Wright-Fisher model *within* each island are met.

3.2 Kingman's coalescent

The genealogy of n genes can, under the Wright-Fisher model, be portrayed as follows: Each sampled gene in generation $g = 0, -1, -2, ...$ "chooses" its parent uniformly at random from the genes in generation $g - 1$. When $j \geq 2$ genes choose the same parent, they *coalesce*, meaning that the genes have found their MRCA, and at this point the sample size reduces to $n - j + 1$. Starting from generation 0, and repeating the procedure throughout generations backwards in time, until all n genes have coalesced, results in a discrete-step *coalescence tree*, an example of which is illustrated in Figure 3.1.

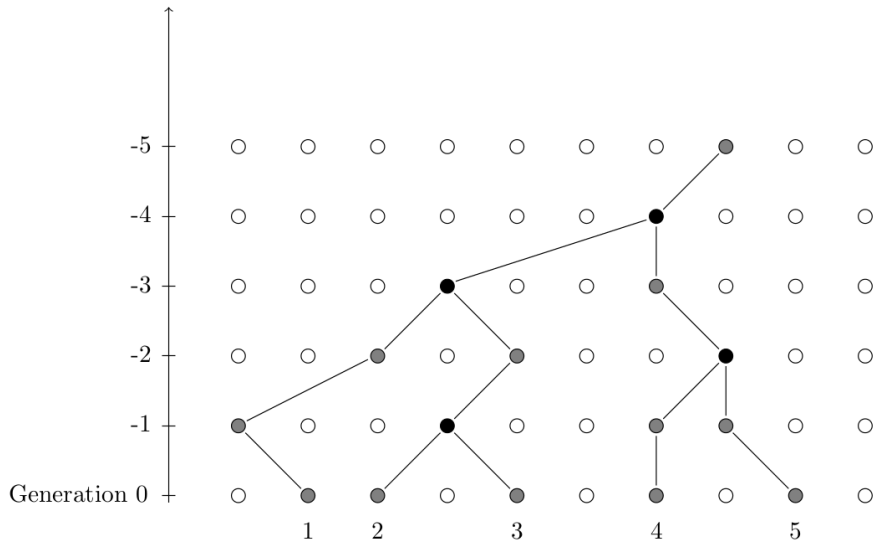


Figure 3.1: A coalescence tree with $n = 5$ sampled genes. The genes are sampled at generation 0 (present), and we observe the lineages backwards in time. A gray circle indicates a sampled gene, and a black circle indicates a coalescence event between two genes. For instance, genes 2 and 3 coalesce in generation -1, and all 5 genes have coalesced in generation -4.

In order to investigate the behaviour of such a coalescence tree under the Wright-Fisher model, we begin by closely examining the transition from generation 0 to gener-

ation -1, assuming a sample of n genes from generation 0, and a total population of N genes per generation. Throughout this study, n is assumed to satisfy $n \ll N$.

Define a *coalescence event* as the event that two or more genes choose the same parent in one generation shift backwards in time. Thereafter define π_{nk} as the probability that one single coalescence event takes place, such that exactly k genes choose the same parent. Furthermore, let M be the event that more than one coalescence event occurs. The probability π_n that any coalescence event will occur in the transition from generation 0 to generation -1 hence equals

$$\pi_n = \pi_{n2} + \pi_{n3} + \cdots + \pi_{nn} + \mathrm{P}(M).$$

Proposition 3.1. *π_n satisfies*

$$\pi_n = \pi_{n2} + o(N^{-1}) = \binom{n}{2} \cdot \frac{1}{N} + o(N^{-1}),$$

where $o(N^{-1})$ satisfies $\lim_{N\to\infty} o(N^{-1})N = 0$.

Proof. Let M_{ij} be the event that genes i and j have the same parent, whereas all other genes have different parents. Then all M_{ij}, $i, j = 1, ..., n$ are disjoint events with equal probability, so that

$$\begin{aligned}
\pi_{n2} &= \sum_{1\le i<j\le n} \mathrm{P}(M_{ij}) \\
&= \binom{n}{2} \mathrm{P}(M_{12}) \\
&= \binom{n}{2} \cdot \frac{1}{N^2} \cdot N \cdot \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{n-2}{N}\right) \\
&= \binom{n}{2} \cdot \frac{1}{N} + o(N^{-1}),
\end{aligned}$$

whereas, under the assumption that $n \ll N$, we have

$$\begin{aligned}
0 < \pi_{nn} < \pi_{n,n-1} < \cdots < \pi_{n3} &= \binom{n}{3} \cdot \frac{1}{N^3} \cdot N \cdot \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{n-3}{N}\right) \\
&\le \binom{n}{3} \cdot \frac{1}{N^2} = o(N^{-1}).
\end{aligned}$$

Now, define $\tilde{M}$ as the event where two distinct coalescence events take place, where two genes choose the same parent. It is clear that $M \subset \tilde{M}$, and thus

$$\begin{aligned}
0 < \mathrm{P}(M) \le \mathrm{P}(\tilde{M}) &= \binom{n}{2} \cdot \frac{1}{N^2} \cdot N \cdot \binom{n-2}{2} \frac{1}{(N-1)^2} \cdot (N-1) \\
&= \binom{n}{2} \cdot \binom{n-2}{2} \cdot \frac{1}{N(N-1)} = o(N^{-1}).
\end{aligned}$$

Therefore,

$$\pi_n = \binom{n}{2}\frac{1}{N} + o(N^{-1}).$$

□

A natural question that arises is, given that we have a sample of n genes in the current generation of size N, what is the distribution of the number of generations backwards in time before we observe a coalescence event? To answer this, we note that the associated repeated experiment has a *memoryless* characteristic; the probability for a coalescence event remains unchanged in the following generation shift, if no coalescence occurred in the current one. This leads us to the conclusion that J_n – the number of generation shifts that need to occur until the first coalescence event is observed, given n sampled genes – has a "First Success" distribution with success probability π_n, i.e.

$$\mathrm{P}(J_n > k) = (1 - \pi_n)^k \tag{3.1}$$

for $k = 1, 2, ...$ If we now instead let $t_n = J_n/N$ indicate one unit of time, we obtain

$$\begin{aligned}\mathrm{P}(t_n > t) &= \mathrm{P}(J_n/N > t)\\ &= \mathrm{P}(J_n > Nt)\\ &= (1 - \pi_n)^{\lfloor Nt \rfloor}\end{aligned}$$

by Equation (3.1). Under the assumption that $N \gg 1$ we have that $0 < \pi_n = \binom{n}{2}\frac{1}{N} + o(N^{-1}) \ll 1$, and hence we can approximate $1 - \pi_n$ by $e^{-\pi_n}$. This yields

$$\begin{aligned}\mathrm{P}(t_n > t) &\approx \left(e^{-\pi_n}\right)^{Nt}\\ &= \left(e^{-\binom{n}{2}\frac{1}{N} - o(N^{-1})}\right)^{Nt}\\ &= e^{-\binom{n}{2}t - o(N^{-1})Nt}\\ &\longrightarrow e^{-\binom{n}{2}t}, \quad \text{as} \quad N \longrightarrow \infty.\end{aligned}$$

The conclusion is that the rescaled time $t_n = J_n/N$, until a coalescence event occurs in the lineages of n genes, converges in distribution to an exponential distribution with intensity parameter $\binom{n}{2}$, as N grows large. Furthermore, by Proposition 3.1, the probability that the coalescence event is of the type where only 2 of the n lineages coalesce, converges to 1. The resulting continuous time coalescence tree can be illustrated with the *Kingman's coalescent*, in which the tree is depicted in the rescaled continuous time, and the horizontal movement of a lineage between generations seen in Figure 3.1 is disregarded. An example of a Kingman's coalescent for an initial sample of 4 genes is illustrated in Figure 3.2.

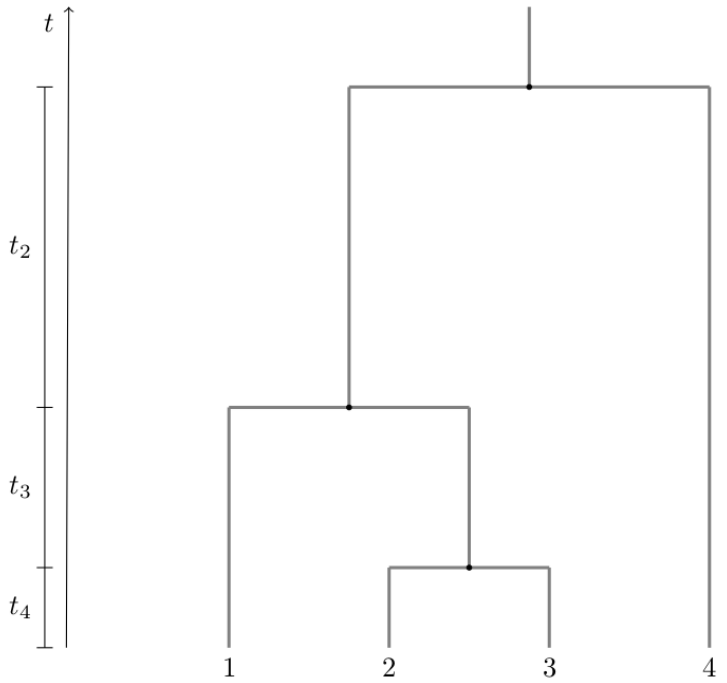


Figure 3.2: A Kingman's coalescent tracking the ancestral lineages of $n = 4$ genes. The time t_i when the number of ancestral lineages equals i ($i = 2, 3, 4$) is exponentially distributed with rate $\binom{i}{2}$. A black dot ($\cdot$) signifies the coalescence of two genes.

3.3 Mutations

The Wright-Fisher model can be extended by allowing mutations to occur when the gene makes a transition between generations. Thus, let $0 < u \ll 1$ be the probability that a mutation occurs in one gene during the transition from one generation to the previous, and define $\theta = 2Nu$. Just as the number of generation transitions until a gene for the first time coalesces with another, the number μ^* of generation transitions until a mutation occurs within a gene, has a First succes distribution with success probability u, i.e. $\mu^* \sim \mathrm{Fs}(u)$ with

$$\mathrm{P}(\mu^* > k) = (1-u)^k$$

for $k = 1, 2, \ldots$. By letting $\mu = \mu^*/N$, we obtain

$$\mathrm{P}(\mu > t) = (1-u)^{\lfloor Nt \rfloor},$$

and, since under the assumption that $0 < u \ll 1$ we have $1 - u \approx e^{-u}$, we immediately obtain

$$\mathrm{P}(\mu > t) \approx e^{-\frac{\theta}{2}t},$$

which means that the time until the first mutation within a lineage occurs is approximately exponentially distributed with rate $\theta/2$.

When a mutation has taken place, the number of transitions until the following mutation occurs has again a First Success distribution with success probability u, and hence the process restarts from the beginning. By this argument, it is easy to see that within one lineage, the time between any two subsequent mutations is exponentially distributed with rate $\theta/2$. This implies that on each vertical line on the Kingman's coalescent, as exemplified in Figure 3.2, the mutations form a homogeneous Poisson process where the expected number of occurrences equals $\theta/2$ multiplied by the length of the vertical line.

On the Kingman's coalescent, one can rearrange the vertical lines, or *branches*, by stacking one after the other, to form a single line of length $\tau = \sum_{j=2}^{n} jt_j$. We call τ the *total tree length*. Assuming that the coalescent consists of b distinct branches, the interval $[0, \tau]$ can be partitioned into b subintervals, along which mutations occur according to a homogeneous Poisson process with intensity $\theta/2$. It is then easy to show that the mutations within the subintervals form a homogeneous Poisson process with intensity $\theta/2$ on the whole interval $[0, \tau]$. In particular, if we let S_n denote the total number of mutations on the tree[3], we have

$$S_n|\tau = x \sim \text{Po}\left(\frac{\theta x}{2}\right). \tag{3.2}$$

A hypothetical realization of a Kingman's coalescent with mutations is illustrated in Figure 3.3.

3.4 Symmetric island model

Consider a set of k islands; $I_1, I_2, ..., I_k$, each containing equally large gene populations of size N, constant throughout nonoverlapping generations, and suppose that we sample n_i genes from I_i, $i \in \{1, ..., k\}$. Within each island I_i, $i \in \{1, ..., k\}$, the reproduction of the individuals occurs according to the rules of the Wright-Fisher model, that is, each of the organisms of generation l chooses its parent uniformly at random with replacement from the N parents in generation $l-1$ within the same island, and all organisms choose their parent independently of each other. When shifting to continuous time, the resulting genealogy within each island can be represented with Kingman's coalescent, so that the rate at which coalescence occurs within an island, when the number of ancestral lineages within the island is n_i, equals $\binom{n_i}{2}$.

What we refer to as "island" represents a geographically isolated region. In particular, this geographical isolation implies that the genealogies of two genes can *not* coalesce when the genes are on separate islands. However, we do allow for an occasional migration to occur from one island to another. We quantify this migration with the set of

[3]In population genetics, S_n is normally referred to as *the number of segregating sites* when working with the *infinite sites model*, see e.g. [Durrett, 2008, Hein et al., 2005] for details. However, for our purposes it is sufficient to consider S_n simply as the the total number of mutations on a Kingman's coalescent.

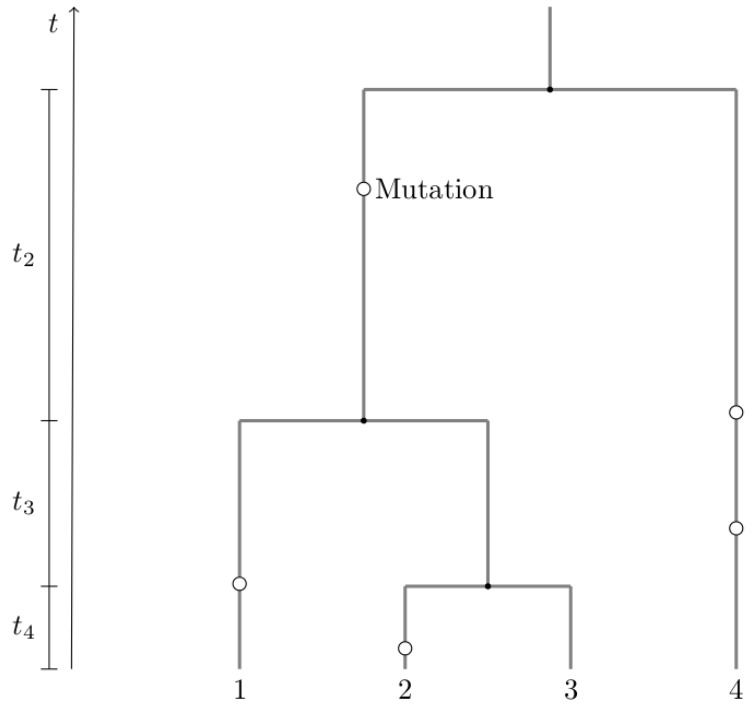


Figure 3.3: A Kingman's coalescent with $n = 4$ and with mutations (marked with $\circ$) allowed. The time between mutations is exponentially distributed with rate $\theta/2$ along each branch, and thus the mutations form a Poisson process on the tree. The number of branches in this illustration is $b = 6$.

migration probabilities $\{m_{ij}\}$, such that m_{ij} is the probability that an individual in I_i is a descendant of an individual in I_j; $i, j \in \{1, ..., k\}$. For simplicity, let

$$m_{ii} = 1 - m, \text{ and} \tag{3.3a}$$
$$m_{ij} = m/(k-1) \tag{3.3b}$$

for $i \neq j$, and for some m satisfying $0 < m \ll 1$. This means that in each I_i, $i \in \{1, ..., k\}$, and in each generation transition, a gene migrates to some $I_{j \neq i}$ with some small probability m, and the gene chooses the island to which it migrates uniformly at random. With probability $1-m$ the parent gene is from the same island as the offspring. This design leads to a model widely used in population genetics known as the *symmetric island model* [Durrett, 2008, Ch. 4.6].

The genealogy of $n = \sum_{i=1}^{k} n_i$ sampled genes can, in discrete time, be depicted similarly as the genealogical tree for the general Wright-Fisher model illustrated in Figure 3.1. The main difference is that the total population of kN genes needs to be partitioned into separate regions representing the islands. A schematic example of such a discrete time tree is shown in Figure 3.4 for the special case $k = 2$, with $N = 4$, $n_1 = 3$ and $n_2 = 2$.

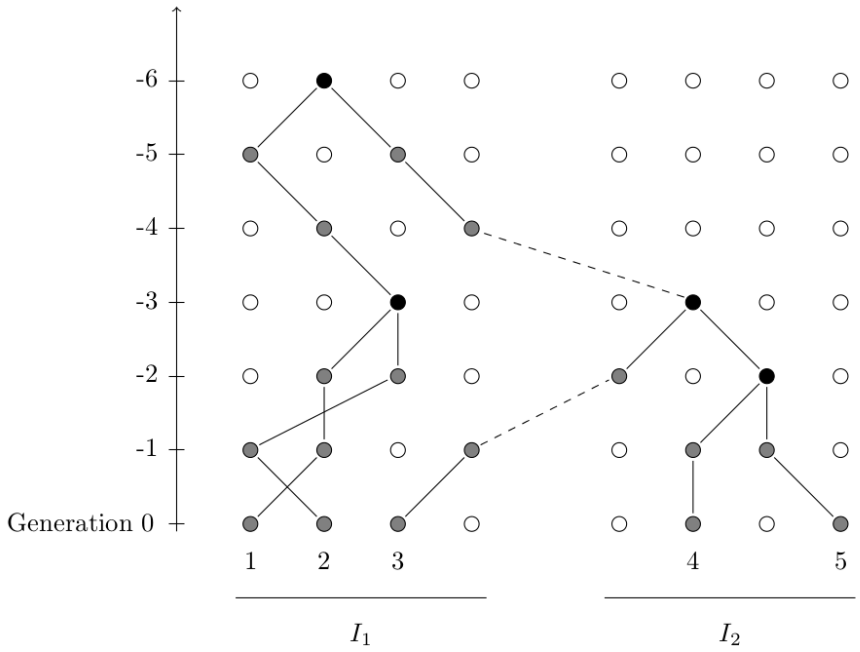


Figure 3.4: A hypothetical discrete time coalescent tree for a symmetric island model with $k = 2$ islands; I_1 and I_2. Genes 1,2 and 3 are sampled from I_1, whereas genes 4 and 5 are sampled from I_2. A dashed line indicates a gene migration. For instance, Gene 3 is a descendant of a parent gene from I_2, and hence Gene 3 migrates from I_2 to I_1 between generations -2 and -1. Finally, the complete sample of 5 genes coalesces in generation -6.

Equations (3.3a) and (3.3b) imply that, for one gene in I_i, the number ν^* of generations backwards in time until it migrates to any other island is $\mathrm{Fs}(m)$, so that

$$\mathrm{P}(\nu^* > x) = (1-m)^x$$

for $x = 1, 2, ...$ As before, let $\nu = \nu^*/N$, that is, count continuous time in units of N generations. This implies

$$\mathrm{P}(\nu > t) = (1-m)^{\lfloor Nt \rfloor}.$$

Now define $M = 2Nm$. This, together with the assumption that $0 < m \ll 1$ yields

$$\mathrm{P}(\nu > t) \approx e^{\frac{M}{2}t}, \tag{3.4}$$

i.e. the time until a gene from I_i emigrates is approximately exponentially distributed with rate $M/2$. Since the genes choose the island to which they migrate uniformly at random with repetition, Equation (3.4) also implies that a gene migrates from I_i to $I_{j\neq i}$ with rate $\frac{M}{2(k-1)}$. Further, when the sample from I_i consists of n_i genes, a migration from I_i to $I_{j\neq i}$ occurs with rate $\frac{Mn_i}{2(k-1)}$.

The symmetric island model can obviously be extended by allowing mutations to occur along the genealogies. Recall from Section 3.3 that mutations, which are assumed to take place independently of migration, occur according to a Poisson process with rate $\theta/2$ along the coalescent tree. By the memoryless property of the exponential distribution we thus conclude that, under the symmetric island model, the lineage of a gene is constantly susceptible to 3 possible events:

1. *coalescence*, which occurs after an exponential time with rate 1,
2. *mutation*, which occurs after an exponential time with rate $\theta/2$, and
3. *migration*, which occurs after an exponential time with rate $M/2$.

A continuous time coalescent tree for the symmetric island model with $k = 2$, $n_1 = 4$ and $n_2 = 2$, with mutations allowed, is exemplified in Figure 3.5.

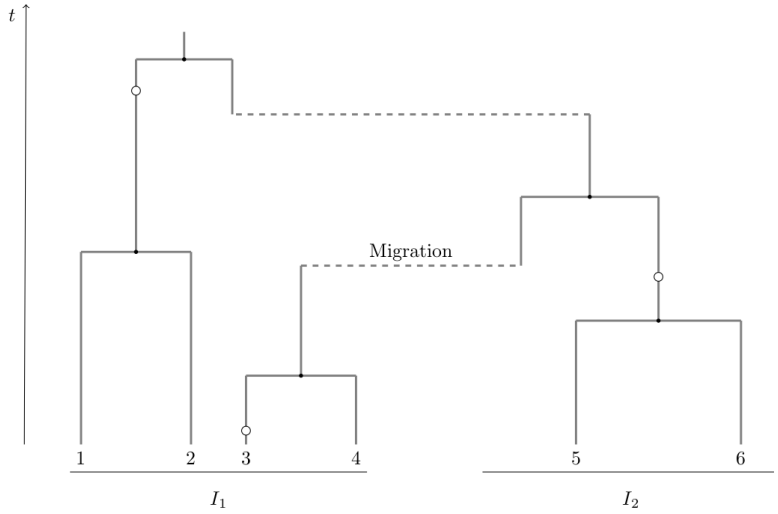


Figure 3.5: A hypothetical continuous time coalescent tree for a symmetric island model with $k = 2$ islands, with mutations. The tree traces the ancestry of 6 genes sampled at time $t = 0$. Genes 1,2,3 and 4 are sampled from I_1, whereas genes 5 and 6 are sampled from I_2. A circle ($\circ$) indicates a mutation, whereas a dot ($\cdot$) indicates a coalescence of two genes. A dashed line indicates a gene migration between the two islands.

4 The phase-type coalescent

The relationship between coalescent theory and phase-type distributions is evident by the persistency of the exponential distribution throughout Section 3. Indeed, in this section we will establish that many problems in coalescent theory can be tackled using a phase-type distribution approach. We begin by demonstrating in Section 4.1, that the basic properties of the Kingman's coalescent are readily derivable using phase-type

distributions; the tree height will be analyzed in Section 4.1.1, and thereafter we extend the applications of phase-type distributions to mutations in Section 4.1.2.

Finally, in Section 4.2, we will investigate two special cases of the symmetric island model presented in Section 3.4, using phase-type distributions as a starting point. Firstly, we will analyze the case where the number of islands is held fixed at $k = 2$, whereas the initial sample size is allowed to be an arbitrary number n. Secondly, we will examine the case where the number of islands is allowed to be an arbitrary k, whereas the initial sample is held fixed at $n = 2$ genes.

4.1 Properties of the Kingman's coalescent

4.1.1 Tree height

Section 3.2 demonstrated that on the Kingman's coalescent (as exemplified in Figure 3.1), the time it takes for a coalescence event to occur when the number of ancestral lineages equals i, is asymptotically exponentially distributed with rate $\binom{i}{2}$, when the total population size N becomes large. It therefore becomes intuitive to regard the continuous time coalescent as a CTMC on $[0, \infty)$ with a state space consisting of the integers $1, ..., n$; that is, the state of the Markov chain denotes the number of ancestors of the sample at any given time $t \geq 0$.

Thus, with the notation used in Section 2, let $\{X(t);\ t \geq 0\}$ be a CTMC with state space $S = \{1, 2, ..., n\}$, where $X(t)$ represents the ancestral size of the sample at time t on the Kingman's coalescent. In this case states $\{2, ..., n\}$ are transient, since by Proposition 3.1, the Markov chain makes a transition from state i to state $i - 1$ with probability 1 for all $2 \leq i \leq n$. When the Markov chain is in state $\{1\}$, it means that no more coalescence events can take place, making $\{1\}$ an absorbing state. The generator matrix $\mathbf{Q}$ for $\{X(t);\ t \geq 0\}$ thereby equals

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ \binom{2}{2} & -\binom{2}{2} & 0 & \dots & 0 \\ 0 & \binom{3}{2} & -\binom{3}{2} & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \binom{n}{2} & -\binom{n}{2} \end{pmatrix}, \tag{4.1}$$

which can be block partitioned into

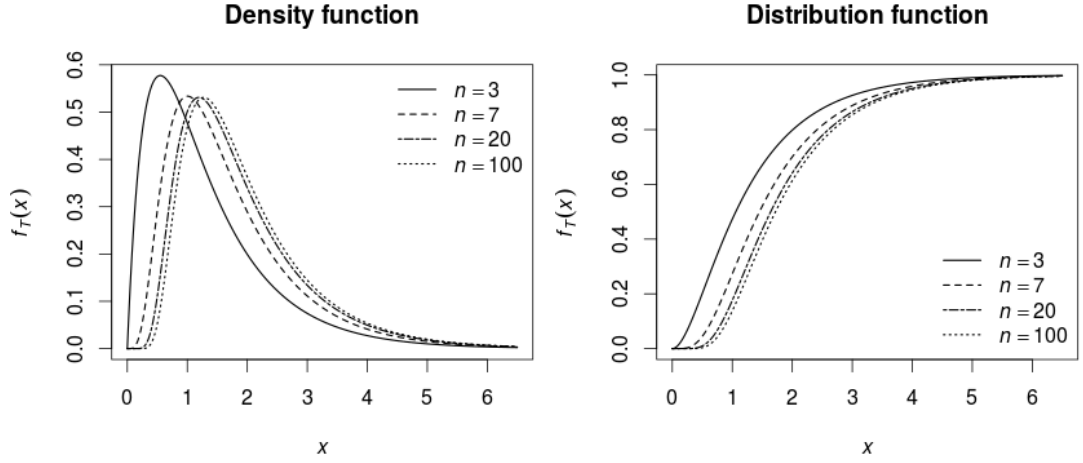
$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{0}'_{n-1} \\ \mathbf{t} & \mathbf{T} \end{pmatrix},$$

where $\mathbf{t}' = (1, \mathbf{0}'_{n-2})$ and

$$\mathbf{T} = \begin{pmatrix} -\binom{2}{2} & 0 & 0 & \dots & 0 \\ \binom{3}{2} & -\binom{3}{2} & 0 & \dots & 0 \\ 0 & \binom{4}{2} & -\binom{4}{2} & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \binom{n}{2} & -\binom{n}{2} \end{pmatrix}. \tag{4.2}$$

Furthermore, since with probability 1 the initial sample consists of n genes, we have the initial probability vector for the transient states $\boldsymbol{\alpha}' = (\mathbf{0}'_{n-2}, 1)$.

Let T_{nh} be the time it takes for $\{X(t);\ t \geq 0\}$ to reach the absorbing state $\{1\}$. In terms of the coalescent, T_{nh} thus represents the time at which the lineages of an initial sample of n genes have coalesced, which corresponds to the height of the coalescent. According to Definition 2.1, $T_{nh} \sim \text{PH}(\boldsymbol{\alpha}, \mathbf{T})$, and the CDF and the PDF for T_{nh} are explicitly given by equations (2.7) and (2.8) respectively. Their graphs are illustrated in figures 4.1a and 4.1b by a computer implementation of (2.8) and (2.7), for initial sample sizes $n = 3, 7, 20, 100$. By equations (4.1) and (2.14), the case with $n = 2$ results in the exponential distribution with rate 1.



(a) PDF of T_{nh}. **(b)** CDF of T_{nh}.

Figure 4.1: The PDF and the CDF of the height T_{nh} of a Kingman's coalescent when the initial sample consists of $n = 3, 7, 20, 100$ individuals.

It is easy to verify that the negative inverse of the matrix $\mathbf{T}$ given in Equation (4.2) equals the lower triangular matrix

$$-\mathbf{T}^{-1} = \begin{pmatrix} \binom{2}{2}^{-1} & & & & \\ \binom{2}{2}^{-1} & \binom{3}{2}^{-1} & & & \\ \binom{2}{2}^{-1} & \binom{3}{2}^{-1} & \binom{4}{2}^{-1} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \binom{2}{2}^{-1} & \binom{3}{2}^{-1} & \binom{4}{2}^{-1} & \dots & \binom{n}{2}^{-1} \end{pmatrix},$$

which can be used to determine the moments of T_{nh} by Equation (2.13). In particular, with $\mathbf{t}' = (1, \mathbf{0}'_{n-2})$ and $\boldsymbol{\alpha}' = (\mathbf{0}'_{n-2}, 1)$, we immediately obtain

$$\mathrm{E}[T_{nh}] = \sum_{2 \le i \le n} \binom{i}{2}^{-1} = 2 - \frac{2}{n},$$

which agrees with existing literature (see e.g. [Durrett, 2008, Sect. 1.2.1]).

4.1.2 Mutations

The model can be extended by introducing mutations, which, according to the results obtained in Section 3.3, occur along each lineage after an exponentially distributed time with rate $\theta/2$, with $\theta = 2Nu$, and where $0 < u \ll 1$ is the probability that a mutation occurs in one gene during the transition from one generation to the previous. This means that on the Kingman's coalescent, when the number of ancestral lineages equals i, a mutation occurs with rate $i\theta/2$. We therefore consider the CTMC $\{X(t);\ t \ge 0\}$ on the state space $S = \{0, 1, ..., n\}$, that is, the same state space as before, but with an additional absorbing state $\{0\}$, such that $X(t) = 0$ if a mutation has occurred at time t. The corresponding intensity matrix for this extended model equals

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \frac{2\theta}{2} & \binom{2}{2} & -\sigma_2 & 0 & \dots & 0 \\ \frac{3\theta}{2} & 0 & \binom{3}{2} & -\sigma_3 & & \vdots \\ \vdots & \vdots & & \ddots & \ddots & 0 \\ \frac{n\theta}{2} & 0 & \dots & 0 & \binom{n}{2} & -\sigma_n \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times(n-1)} \\ \mathbf{t} & \mathbf{T} \end{pmatrix},$$

where $\sigma_k = -\left[\frac{k\theta}{2} + \binom{k}{2}\right]$ for $k = 2, ..., n$. Consequently, we have

$$\mathbf{t} = \begin{pmatrix} \frac{2\theta}{2} & \binom{2}{2} \\ \frac{3\theta}{2} & 0 \\ \vdots & \vdots \\ \frac{n\theta}{2} & 0 \end{pmatrix} \tag{4.3}$$

and

$$\mathbf{T} = \begin{pmatrix} -\sigma_2 & 0 & 0 & \dots & 0 \\ \binom{3}{2} & -\sigma_3 & 0 & \dots & 0 \\ 0 & \binom{4}{2} & -\sigma_4 & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \binom{n}{2} & -\sigma_n \end{pmatrix}. \tag{4.4}$$

Let T be the time it takes for the CTMC to reach one of the absorbing states $\{0\}$ or $\{1\}$. Observe that since mutations have been introduced, the random variable T no longer equals the height of the coalescent tree. Nonetheless, according to the results from Section 2.5 where phase-type distributions with multiple absorbing states were studied, we have $T \sim \text{PH}(\boldsymbol{\alpha}, \mathbf{T})$, with $\boldsymbol{\alpha} = (\mathbf{0}'_{n-2}, 1)$ and $\mathbf{T}$ given by Equation (4.4). The CDF and the PDF of T are given by equations (2.7) and (2.19), respectively. These functions are illustrated in Figure 4.2 for some selected values of n and θ that correspond to different forms of the PDF of T.

The probability that a mutation occurs before the complete coalescence of a sample of n genes is explicitly given by Equation (2.22): Recall that the i^{th} entry in $\boldsymbol{\phi} = (\phi_0, \phi_1)$, $i \in \{0, 1\}$, is the probability that the CTMC $\{X(t);\ t \geq 0\}$ is absorbed into state i, and note that absorbtion into state $\{0\}$ implies that a mutation occurs before complete coalescence. These quantities can be considered particularly important, since $\phi_0 = 1 - \phi_1$ is the probability that the n initially sampled genes are *identical*, i.e. the gene has been passed from the MRCA to the sampled individuals throughout generations, without the occurrence of a mutation in the genealogies. In particular, ϕ_0 is given by

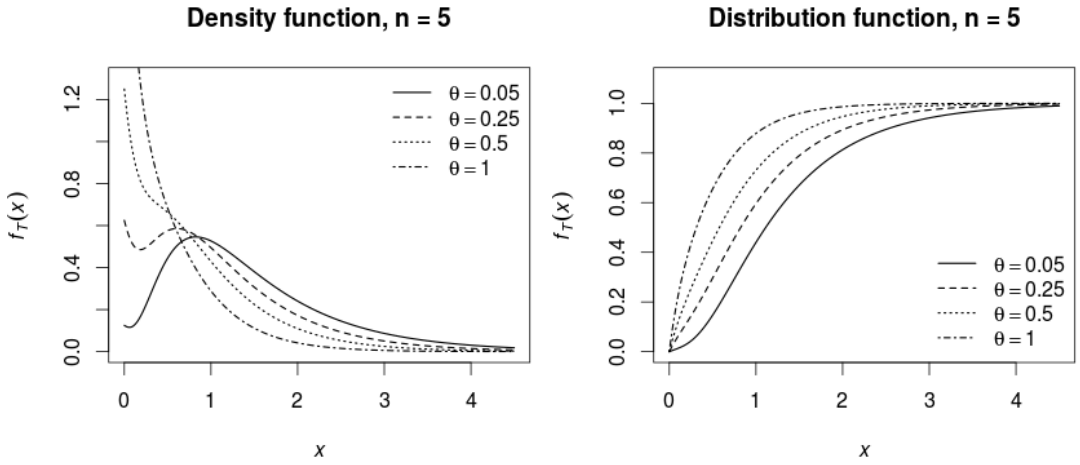
$$\phi_0 = -\boldsymbol{\alpha}'\mathbf{T}^{-1}\mathbf{t}_{:1}, \tag{4.5}$$

where $\boldsymbol{\alpha}' = (\mathbf{0}'_{n-2}, 1)$, $\mathbf{T}$ is given by Equation (4.4), and

$$\mathbf{t}'_{:1} = \left(\frac{2\theta}{2}, \frac{3\theta}{2}, ..., \frac{n\theta}{2} \right)$$

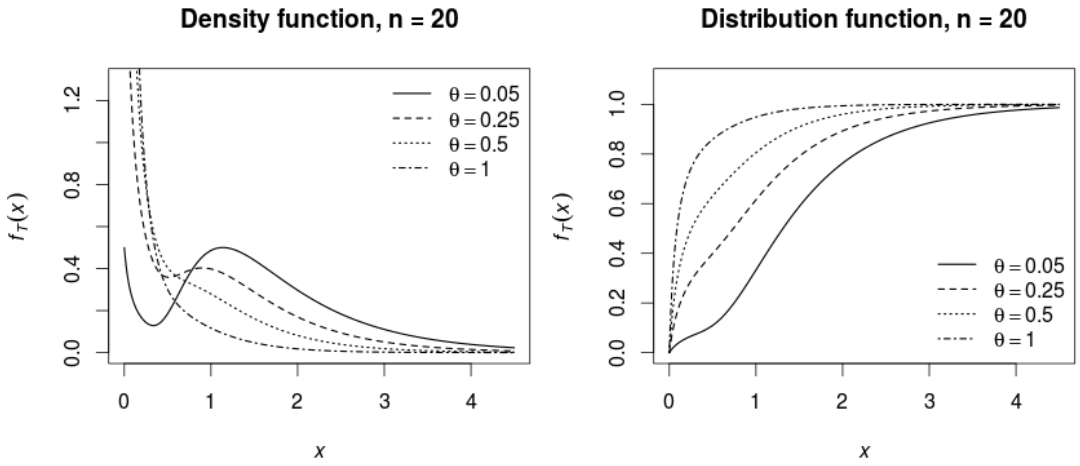
is the first column in the matrix $\mathbf{t}$ given in Equation (4.3). To put Equation (4.5) into practice, we generate ϕ_0 as a function of θ for some selected values of n, by a computer implementation of Equation (4.5) (Figure 4.3). It is clearly seen that the probability that a mutation occurs before the complete coalescence of the sample converges to 1 with increasing mutation rate, and that the convergence is more rapid with a larger initial sample size.

With the introduction of mutations, computing the expected total tree length, $\text{E}[\tau]$ (where τ was described in Section 3.3), with a phase-type distribution approach becomes straightforward: By recalling Equation (3.2) and by conditioning on τ, we have



(a) PDF of T, $n = 5$, $\theta = 0.05, 0.25, 0.5, 1$.

(b) CDF of T, $n = 5$, $\theta = 0.05, 0.25, 0.5, 1$.



(c) PDF of T, $n = 20$, $\theta = 0.05, 0.25, 0.5, 1$.

(d) CDF of T, $n = 20$, $\theta = 0.05, 0.25, 0.5, 1$.

Figure 4.2: The PDF (left), and the CDF (right) for T, computed according to equations (2.19) and (2.7) respectively.

$$\begin{aligned}\phi_0(\theta) &= \mathrm{P}(S_n = 0) \\ &= \int_{x=0}^{\infty} \mathrm{P}(S_n = 0 | \tau = x) f_\tau(x)\ dx \\ &= \int_{x=0}^{\infty} e^{-\frac{\theta x}{2}} f_\tau(x)\ dx \end{aligned} \tag{4.6}$$

and thus

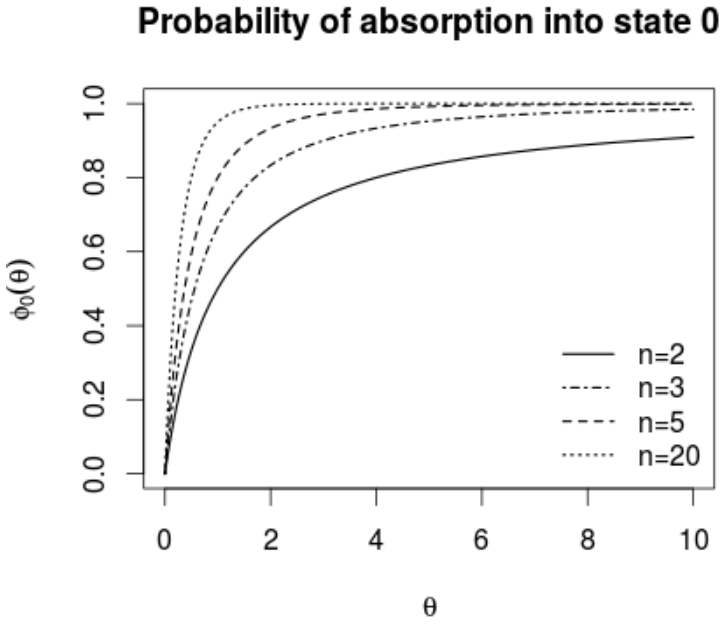


Figure 4.3: The probability ϕ_0 that a mutation occurs before the complete coalescence of n sampled genes, as a function of θ.

$$\phi_0(\theta) = \mathscr{L}_\tau(\theta/2),$$

where $\mathscr{L}_\tau(\cdot)$ denotes the Laplace transform of τ. Hence, by evaluating the k^{th} derivative of $\phi_0(\theta)$ at $\theta = 0$ we obtain the k^{th} moment of τ as

$$\mathrm{E}[\tau^k] = (-2)^k \left.\frac{d^k}{d\theta^k}\phi_0(\theta)\right|_{\theta=0}, \tag{4.7}$$

where $\phi_0(\theta)$ is defined in Equation (4.5). In particular, by Equation (4.7), the expected value of the total tree length equals

$$\mathrm{E}[\tau] = 2\boldsymbol{\alpha}'\mathbf{T}^{-1}\left(\frac{d\mathbf{t}_{:1}}{d\theta} - \frac{d\mathbf{T}}{d\theta}\mathbf{T}^{-1}\mathbf{t}_{:1}\right)\Bigg|_{\theta=0},$$

with $\frac{d\mathbf{t}_{:1}}{d\theta} = \left(\frac{2}{2}, \frac{3}{2}, ..., \frac{n}{2}\right)'$ and

$$\frac{d\mathbf{T}}{d\theta} = \begin{pmatrix} -\frac{2}{2} & 0 & 0 & \dots & 0 \\ 0 & -\frac{3}{2} & 0 & \dots & 0 \\ 0 & 0 & -\frac{4}{2} & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \dots & 0 & -\frac{n}{2} \end{pmatrix}.$$

4.2 Symmetric island model

4.2.1 Two islands, arbitrary sample size

Consider the special case of $k = 2$ in the symmetric island model discussed in Section 3.4. That is, suppose that a hypothetical target population is divided into 2 distinct islands, I_1 and I_2, each containing a total population of N haploid organisms, and that in continuous time, a gene migrates from one island to the other with rate $M/2$. Also, allow for mutations to occur according to the rules stated in Section 3.3. The objective in this section is to use phase-type distributions to analyze the genealocical tree of a sample of $n = n_1 + n_2$ genes, where n_i, $i \in \{1, 2\}$ is the number of genes sampled from island I_i.

To do this, define a CTMC $\{X(t);\ t \geq 0\}$ on the two dimensional state space $S = \{(i, j) \in \mathbb{N}^2 :\ 0 \leq i + j \leq n\}$, where $X(t) = (i, j)$, $i + j \neq 0$, signifies that there are i ancestral lineages on I_1 and j ancestral lineages on I_2 at time $t \geq 0$. If $X(t) \in \{(i, j) :\ i+j = 1\}$, then clearly either $i = 1$ and $j = 0$, or $i = 0$ and $j = 1$, meaning that there is only one lineage left at time t, and no more coalescence events can take place. Therefore we define $(1, 0)$ and $(0, 1)$ as absorbing states. Furthermore, let state $(0, 0)$ be an absorbing state representing a mutation, in the sense that if $X(t) = (0, 0)$, then a mutation has taken place along some lineage at some time $t \geq 0$, before all lineages have coalesced. The transition intensities for the CTMC are the following:

i. For any state (i, j), such that $i+j \geq 2$, the transition $(i, j) \longrightarrow (0, 0)$, corresponding to a mutation along some lineage, occurs with intensity $\frac{(i+j)\theta}{2}$.

ii. When $i \geq 1$, $j \geq 0$ and $i + j \geq 2$, the transition $(i, j) \longrightarrow (i - 1, j + 1)$ occurs with intensity $\frac{iM}{2}$. This transition corresponds to a migration for one gene from I_1 to I_2. Similarly,

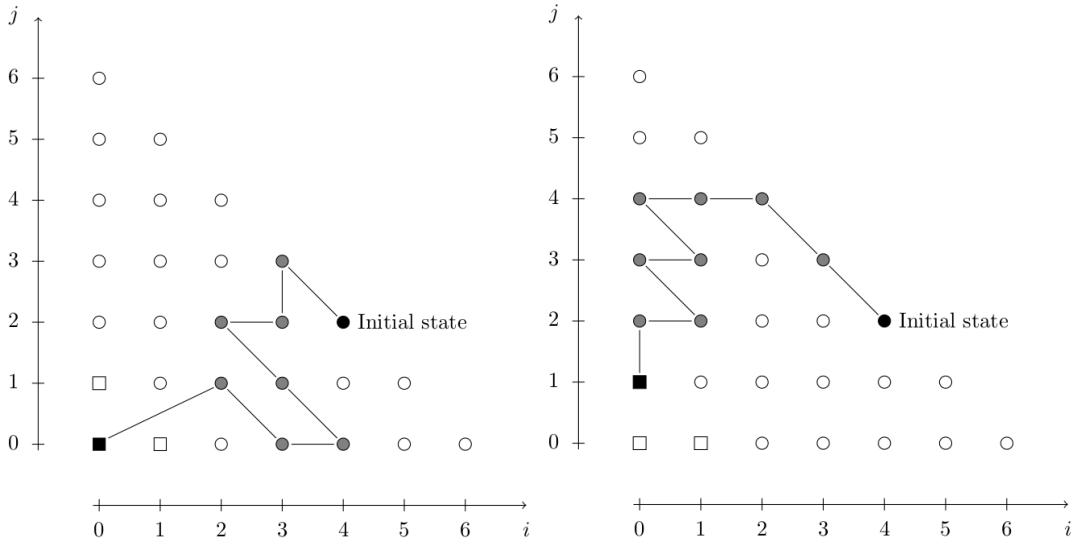
iii. when $i \geq 0$, $j \geq 1$ and $i + j \geq 2$, the transition $(i, j) \longrightarrow (i + 1, j - 1)$ occurs with intensity $\frac{jM}{2}$. This transition corresponds to a migration for one gene from I_2 to I_1.

iv. The transition $(i, j) \longrightarrow (i - 1, j)$ occurs with intensity $\binom{i}{2}$, provided that $i \geq 2$, $j \geq 0$. This transition corresponds to a coalescence event in I_1. Similarly,

v. the transition $(i, j) \longrightarrow (i, j - 1)$ occurs with intensity $\binom{j}{2}$, provided that $j \geq 2$, $i \geq 0$. This transition corresponds to a coalescence event in I_2.

In the large N limit, all other transitions are impossible and thus occur with rate 0.

Figure 4.4 illustrates two hypothetical realizations of the discrete time transitions between states made by the CTMC $\{X(t);\ t \geq 0\}$ on the state space S, until an absorbing state is reached. The illustration in Figure 4.4a shows that the CTMC is absorbed into state $(0,0)$ from state $(2,1)$, indicating that a mutation occurs on one of the lineages on the coalescent at a time when there are 2 lineages in I_1 and 1 lineage in I_2. Figure 4.4b on the other hand, shows a path which is absorbed into state $(0,1)$ from state $(0,2)$, meaning that all lineages coalesce before a mutation occurs, and the final coalescence event occurs in I_2. Note however that these examples do not take into account that $\{X(t);\ t \geq 0\}$ progresses in continuous time, and thereby merely depict *which* transitions are made, and not *when* the transitions are made. In other words, Figure 4.4 illustrates two realizations of the discrete time Markov chain embedded in the CTMC.



(a) Absorption into $(0,0)$. **(b)** Absorption into $(0,1)$.

Figure 4.4: Two examples of the path taken by $\{X(t);\ t \geq 0\}$ before reaching an absorbing state. The initial state (marked with a black circle) is in both examples $X(0) = (4,2)$, indicating that the initial sample sizes from I_1 and I_2 are $n_1 = 4$ and $n_2 = 2$ respectively. The absorbing states $(0,0)$, $(0,1)$ and $(0,0)$ are marked as squares, a black square indicating which state the CTMC is absorbed into.

To determine the generator matrix $\mathbf{Q}$ for $\{X(t);\ t \geq 0\}$, one first needs to specify a convenient ordering of the states in S. With a given order, one can then obtain an explicit generator matrix $\mathbf{Q}$ by applying the transition intensities specified in i-v above. We propose the following ordering: For $(i,j),(k,l) \in S$, let $(i,j) < (k,l)$ if $i+j < k+l$, or if $i+j = k+l$ and $i < k$. In other words, the states are ordered as $(0,0)$, $(0,1)$, $(1,0)$, $(0,2)$, $(1,1)$, $(2,0)$, $(0,3)$, $(1,2)$, $(2,1)$, and so on. As an example, with this ordering, and for the special case $n_1 + n_2 = 3$ the generator matrix would equal

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{2\theta}{2} & \binom{2}{2} & 0 & -\sigma_{(0,2)} & \frac{2M}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{2\theta}{2} & 0 & 0 & \frac{M}{2} & -\sigma_{(1,1)} & \frac{M}{2} & 0 & 0 & 0 & 0 \\ \frac{2\theta}{2} & 0 & \binom{2}{2} & 0 & \frac{2M}{2} & -\sigma_{(2,0)} & 0 & 0 & 0 & 0 \\ \frac{3\theta}{2} & 0 & 0 & \binom{3}{2} & 0 & 0 & -\sigma_{(0,3)} & \frac{3M}{2} & 0 & 0 \\ \frac{3\theta}{2} & 0 & 0 & 0 & \binom{2}{2} & 0 & \frac{M}{2} & -\sigma_{(1,2)} & \frac{2M}{2} & 0 \\ \frac{3\theta}{2} & 0 & 0 & 0 & \binom{2}{2} & 0 & 0 & \frac{2M}{2} & -\sigma_{(2,1)} & \frac{M}{2} \\ \frac{3\theta}{2} & 0 & 0 & 0 & 0 & \binom{3}{2} & 0 & 0 & \frac{3M}{2} & -\sigma_{(3,0)} \end{pmatrix} \tag{4.8}$$

where $\sigma_{(i,j)}$, $(i,j) \in S$, is determined so that the rows of $\mathbf{Q}$ sum to 0. For larger values of n, the generator matrix is obtained similarly, by applying i-v. Note however, that since

$$|S| = \sum_{k=1}^{n+1} k = \frac{(n+1)(n+2)}{2}, \tag{4.9}$$

the dimension of the intensity matrix is $O(n^2)$, and thus $\mathbf{Q}$ rapidly becomes very large.

From Equation (4.8), it is evident that for a total sample size n, the matrix $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times\delta} \\ \mathbf{t} & \mathbf{T} \end{pmatrix}$$

where $\delta := \frac{(n+1)(n+2)}{2} - 3$, and hence for our purposes, the far more important matrices $\mathbf{t}$ and $\mathbf{T}$ are easily extracted from $\mathbf{Q}$. The dimensions of $\mathbf{t}$ and $\mathbf{T}$ are $\delta \times 3$ and $\delta \times \delta$ respectively.

The ordering of the states plays an important role when one wishes to determine the initial sample sizes as well. One obtains the initial sample size n_1 and n_2 by setting the k^{th} entry in the initial probability vector equal to 1, and the remaining entries equal to 0, where k is the number corresponding to state (n_1, n_2) in the specific ordering scheme. Also, since $n_1 + n_2 = n$, the initial sample size implicitly determines the size of the intensity matrix $\mathbf{Q}$ by Equation (4.9). As an example, with the ordering proposed earlier, the initial sample size $n_1 = 4$ and $n_2 = 2$ corresponding to the initial state $(4, 2)$, equals the 26^{th} entry. One appoints this initial sample by setting the initial probability vector $\tilde{\boldsymbol{\alpha}}' = (\mathbf{0}'_{25}, 1, \mathbf{0}'_2)$. The initial probability vector for the transient states is then $\boldsymbol{\alpha} = (\mathbf{0}'_{22}, 1, \mathbf{0}'_2)$. Since $n_1 + n_2 = 6$, the matrix $\mathbf{Q}$ in this case has dimension 28×28.

As is well established by now, phase-type distributions provide a complete description of the time $T \sim \text{PH}(\boldsymbol{\alpha}, \mathbf{T})$ it takes for the CTMC to reach one of the absorbing states. In this section the Markov chain describes the genealogies of a sample of $n_1 + n_2$ genes taken from the two islands, and the absorbing states correspond either to the complete coalescence of the genealogies, or a mutation. If one wishes to exclude mutations from

the model, it is of no restriction to set $\theta = 0$, and in this case, the time until absorption would represent the height of the coalescent. The distribution- and density functions, as well as the moments of T are obtained by applying equations (2.7), (2.19) and (2.13) respectively. Moreover, Equation (2.22) allows us to examine the probability that the genealogy of an initial sample of genes contains a mutation before complete coalescence of the sample is reached, as well as the probability that the complete coalescence occurs on a specific island.

As an example, we consider the initial sample sizes $n_1 = 4$, $n_2 = 2$, and $n_1 = 6$, $n_2 = 0$, and thereafter illustrate the density function f_T for these initial values. We select some values of the migration rate M, and fix the value of the mutation rate to $\theta = 0.05$. Moreover, we illustrate the probabilities for the absorbing states as a continuous function of M. These illustrations are produced by a computer implementation of the methods and results provided in this section, and are shown in Figure 4.5. It can be seen that the PDF, as well as the absorption probabilities are strongly influenced by the selection of the initial conditions.

4.2.2 Arbitrary number of islands, sample of size 2

In theory, it is possible to extend the model considered in Section 4.2.1 to the case where the number of islands is k, by letting the state space consist of all $(i_1, i_2, ..., i_k) \in \mathbb{N}^k$, such that $k \geq 2$ and $0 \leq \sum_j i_j \leq n$. The difficulty with such a state space, however, is that the resulting intensity matrix would quickly become inconveniently large. To avoid this complication we thus proceed by restricting ourselves to the case where the initial sample consists of precisely 2 genes, and by recognizing the convenient symmetries that follow. Indeed, by assuming that a migrating gene "chooses" the island to which it migrates uniformly at random, the size of the state space can be reduced by simply considering whether the 2 genealogies are on the *same*, or on *different* islands.

Excluding mutations

For the moment, exclude mutations from the model, and consider a CTMC $\{X(t);\ t \geq 0\}$ on the state space $S = \{\mathcal{C}, \mathcal{S}, \mathcal{D}\}$, where the states signify the following:

$$X(t) = \begin{cases} \mathcal{C} & \text{if the 2 lineages have coalesced at time } t. \\ \mathcal{S} & \text{if the 2 lineages are on the same island at time } t \\ \mathcal{D} & \text{if the 2 lineages are on different islands at time } t \end{cases} \tag{4.10}$$

We define state $\mathcal{C}$ as an absorbing state.

Coalescence of the lineages is possible only when the lineages are on the same island, and in particular the transition $\mathcal{S} \longrightarrow \mathcal{C}$ occurs at rate $\binom{2}{2} = 1$. Furthermore, since both lineages leave their current island at rate $M/2$, the transition $\mathcal{S} \longrightarrow \mathcal{D}$ occurs at rate M, and since it is assumed that a lineage chooses the island to which it migrates uniformly at random, the rate at which the transition $\mathcal{D} \longrightarrow \mathcal{S}$ occurs is $M/(k-1)$. The possible transitions, and the corresponding transition rates are summarized in Figure 4.6.

Ordering the states $\mathcal{C}$, $\mathcal{S}$, $\mathcal{D}$ results in the generator matrix

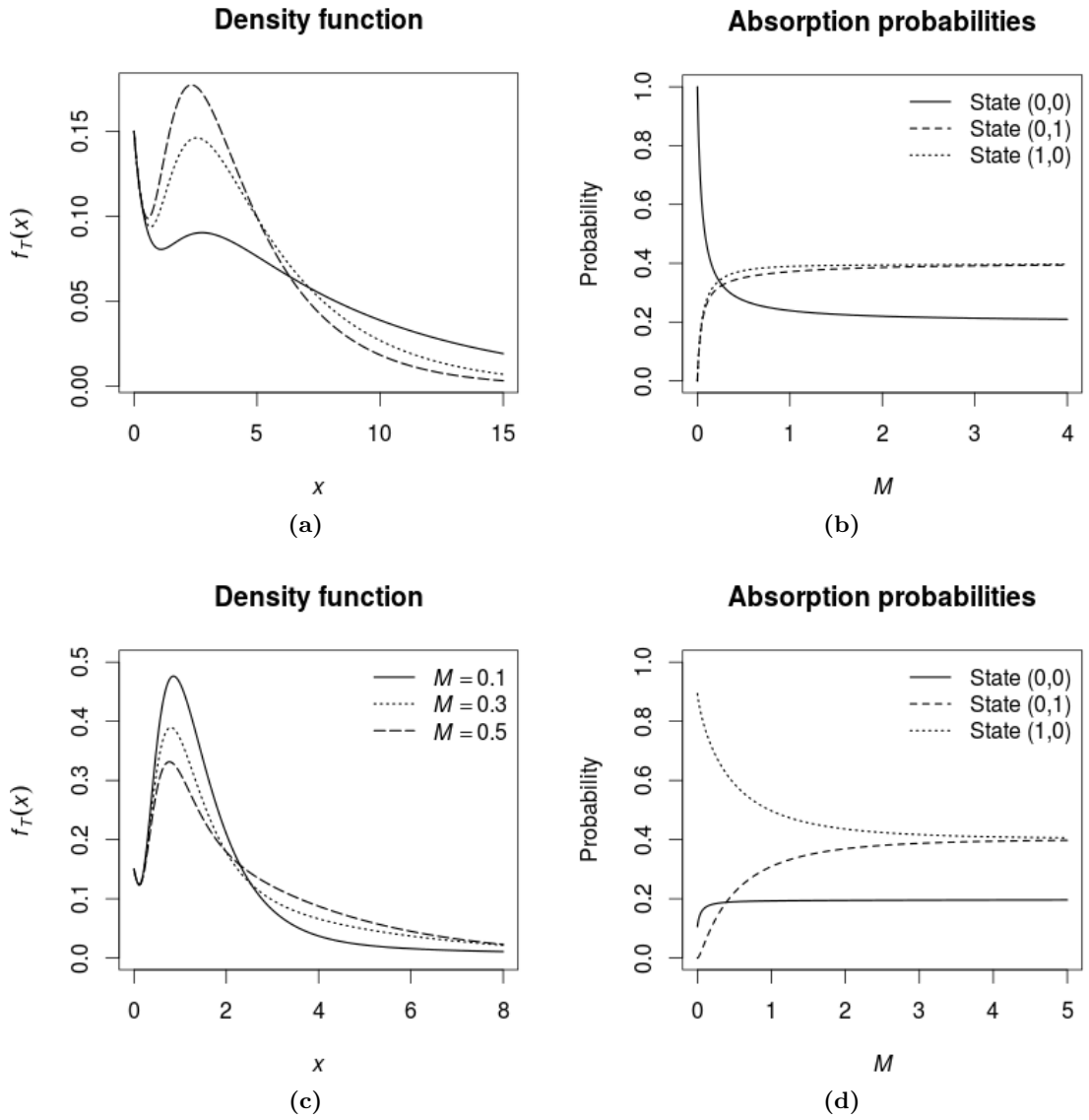


Figure 4.5: The PDF (**(a)** and **(c)**) of the time until $\{X(t);\ t \geq 0\}$ reaches one of the absorbing states (0,0), (0,1) or (1,0), for varying values of the migration rate M. Figures **(b)** and **(d)** illustrate the probabilities of absorption into the states in question, as a function of M. In figures **(a)** and **(b)**, the initial state $X(0)$ equals $(4, 2)$, whereas in figures **(c)** and **(d)**, the initial state $X(0)$ equals $(6, 0)$. In all figures, θ is kept at a constant value of 0.05. Note the change in scale in both axis in figures (**(a)** and **(c)**).

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & -(M+1) & M \\ 0 & \frac{M}{k-1} & -\frac{M}{k-1} \end{pmatrix}, \tag{4.11}$$

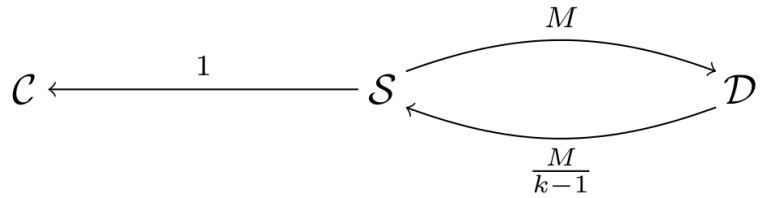


Figure 4.6: An illustration of the state space S. An arrow between two states indicates that a transition between the states in the direction of the arrow is possible, and the rate at which the transition occurs is indicated alongside the arrow.

and thus we have $\mathbf{t}' = (1, 0)$ and

$$\mathbf{T} = \begin{pmatrix} -(M+1) & M \\ \frac{M}{k-1} & -\frac{M}{k-1} \end{pmatrix}.$$

The initial probability vector for the transient states, $\boldsymbol{\alpha}$, determines whether the two initially sampled genes reside on the same island or on different islands; $\boldsymbol{\alpha}' = (1, 0)$ indicates that $X(0) = \mathcal{S}$, whereas $\boldsymbol{\alpha}' = (0, 1)$ indicates that $X(0) = \mathcal{D}$. The matrix $\mathbf{T}$, together with the vectors $\mathbf{t}$ and $\boldsymbol{\alpha}$ determine the PDF and the CDF, as well as the moments for the time until the CTMC reaches the absorbing state $\mathcal{C}$, according to equations (2.8), (2.7) and (2.13) respectively. This time, T_h, thus represents the time it takes for the two sampled genes to reach their MRCA, or in other words the height of the coalescent. The PDF of T_h is illustrated in Figure 4.7, for different values of k and M, as well as for the initial conditions determining whether the two genes are sampled from the same island or from different islands.

The simplicity of the intensity matrix given in Equation (4.11), allows us to explicitly determine the expected value of T_h as a function of the migration rate M, for a model with any given number k of islands. Since the determinant

$$\det(\mathbf{T}) = \frac{M(M+1)}{k-1} - \frac{M^2}{k-1} = \frac{M}{k-1},$$

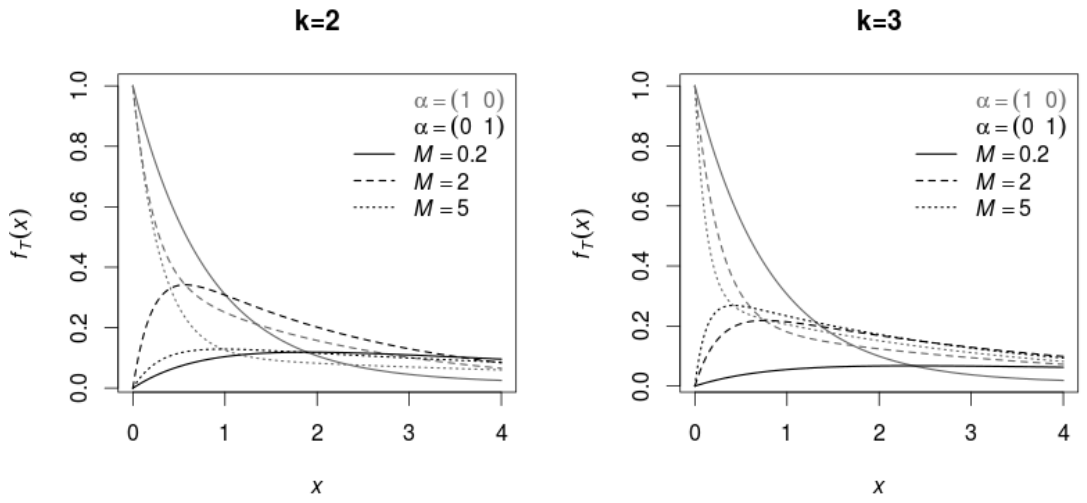
it follows that

$$-\mathbf{T}^{-1} = \frac{k-1}{M} \begin{pmatrix} \frac{M}{k-1} & M \\ \frac{M}{k-1} & M+1 \end{pmatrix} = \begin{pmatrix} 1 & k-1 \\ 1 & \frac{(M+1)(k-1)}{M} \end{pmatrix},$$

and thus that

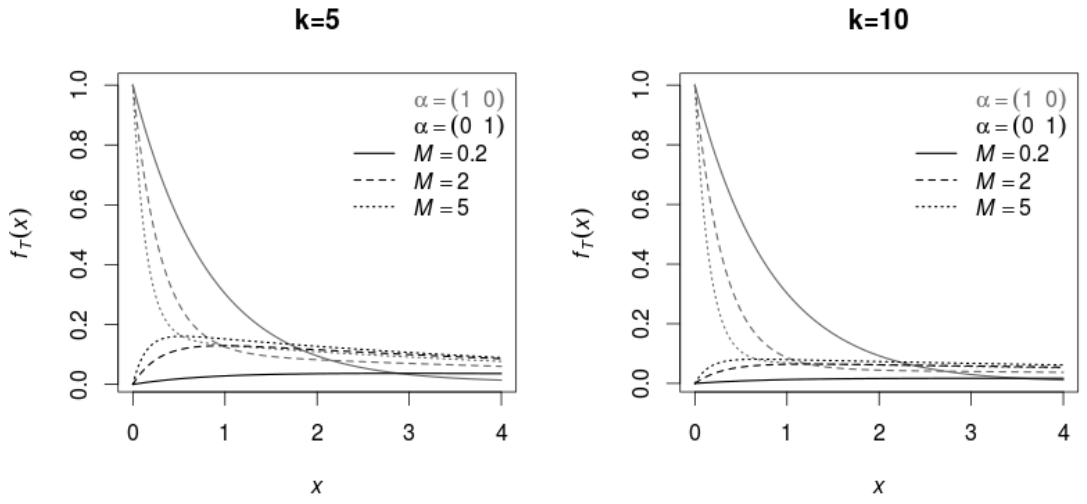
$$-\mathbf{T}^{-1}\mathbf{1}_2 = \begin{pmatrix} 1 & k-1 \\ 1 & \frac{(M+1)(k-1)}{M} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} k \\ \frac{kM+k-1}{M} \end{pmatrix}. \tag{4.12}$$

By Equation (4.12), and by recalling that the moments of a random variable $T \sim \mathrm{PH}(\boldsymbol{\alpha}, \mathbf{T})$ are given by Equation (2.13), the expected value of T_h is given by



(a) PDF with $k = 2$.

(b) PDF with $k = 3$.



(c) PDF with $k = 5$.

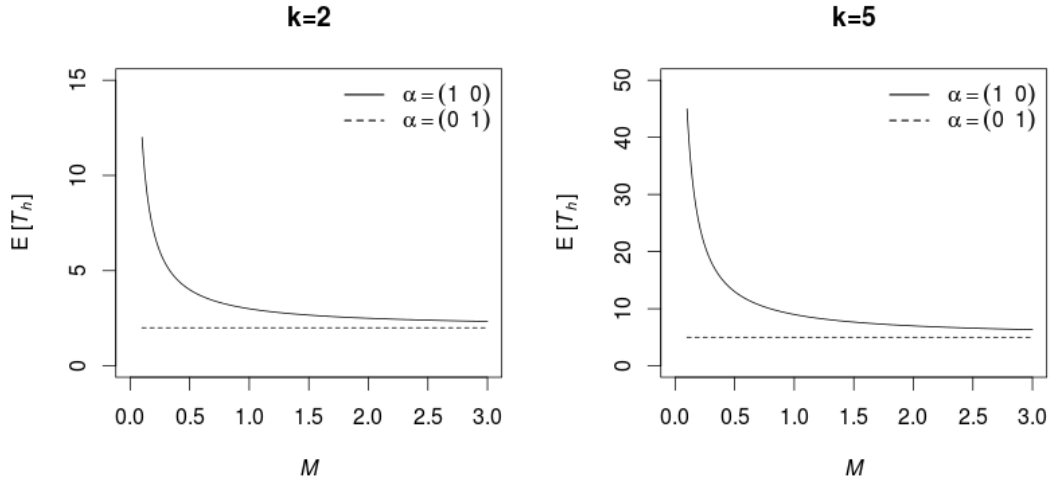
(d) PDF with $k = 10$.

Figure 4.7: The PDF of the time T_h until coalescence of two genes in the symmetric island model with k islands, computed according to Equation (2.8). The functions in gray indicate that the two genes are sampled from different islands, whereas the functions in black indicate that the two genes are sampled from the same island.

$$\mathrm{E}[T_h] = \begin{cases} k & \text{if the initial sample is taken from the same island,} \\ k + \frac{k-1}{M} & \text{if the initial sample is taken from different islands.} \end{cases} \tag{4.13}$$

The first part of Equation (4.13) agrees with Strobeck's theorem, implying that the expected time until the coalescence of two genes sampled from the same island in the sym-

metric island model is independent of the migration rate, see [Durrett, 2008, Sect. 4.5.1] and [Strobeck, 1987]. Notice also that the second part can be obtained from the first by applying Theorem 4.13 in [Durrett, 2008]. Figure 4.8 illustrates the expected value of T_h as a function of the migration rate M for $k = 2$ and $k = 5$. Furthermore, it is easy to check that



(a) $k = 2$. **(b)** $k = 5$.

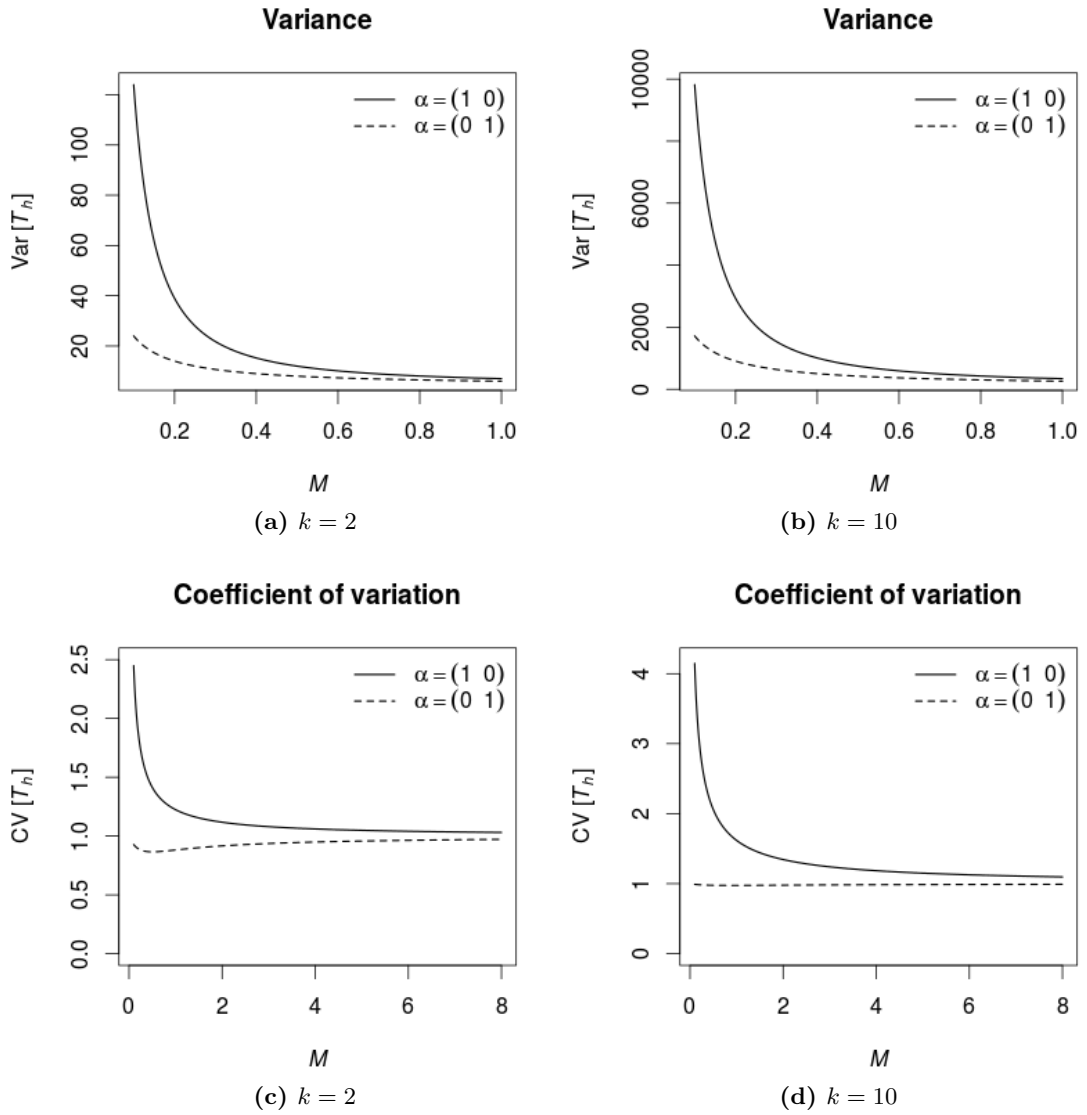
Figure 4.8: The expected value of the tree height T_h as a function of M, for the two cases $k = 2$ (left) and $k = 5$ (right). The dashed line is the expected value given that the two genes are sampled from the same island, which by Equation (4.13) equals k for all $M > 0$. The continuous line is the expected value given that the initial sample is taken from different islands.

$$
\begin{aligned}
(-\mathbf{T}^{-1})^2\mathbf{1}_2 &= \begin{pmatrix} k & k-1+\frac{(M+1)(k-1)^2}{M} \\ 1+\frac{(M+1)(k-1)}{M} & k-1+\frac{(M+1)^2(k-1)^2}{M^2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} 2k-1+\frac{(M+1)(k-1)^2}{M} \\ k+\frac{(M+1)(k-1)}{M}+\frac{(M+1)^2(k-1)^2}{M^2} \end{pmatrix},
\end{aligned}
$$

and thus by Equation (2.13) one eventually obtains

$$
\mathrm{Var}[T_h] = \begin{cases} \frac{Mk^2+2(k-1)^2}{M} & \text{if the initial sample is taken from the same island,} \\ \frac{k^2(M+1)^2-2k(2M+1)+2M+1}{M^2} & \text{if the initial sample is taken from different islands.} \end{cases} \tag{4.14}
$$

We note that Equation (4.14) agrees Theorem 4.14 in [Durrett, 2008]. The variance of T_h, as well as the coefficient of variation $\sqrt{\mathrm{Var}[T_h]}/\mathrm{E}[T_h]$, are illustrated in Figure 4.9, as a function of the migration rate M, when the number of islands equals $k = 2$ and $k = 10$.



(a) $k = 2$ **(b)** $k = 10$

(c) $k = 2$ **(d)** $k = 10$

Figure 4.9: The variance (above) and the coefficient of variation (below) of T_h as a function of M, computed for the cases $k = 2$ (left) and $k = 10$ (right).

Including mutations

The model can be extended by introducing an additional absorbing state $\mathcal{M}$, that is, by considering a CTMC $\{X(t);\ t \geq 0\}$ on the state space $S = \{\mathcal{C}, \mathcal{M}, \mathcal{S}, \mathcal{D}\}$, where the

states $\mathcal{C}, \mathcal{S}$, and $\mathcal{D}$ are as stated in Equation (4.10), and state $\mathcal{M}$ signifies that $X(t) = \mathcal{M}$ if a mutation has occurred at time $t \geq 0$.

Since mutations occur with equal rate $\theta/2$ on each lineage regardless of which islands the genes reside on, and since there are two lineages at all times t before an absorbing state is reached, we deduce that the transitions $\mathcal{S} \longrightarrow \mathcal{M}$ and $\mathcal{D} \longrightarrow \mathcal{M}$ both occur at rate θ. The transition rates between the states $\mathcal{C}, \mathcal{S}$, and $\mathcal{D}$ remain the same as before. The possible transitions, and the corresponding transition rates are summarized in Figure 4.10.

Ordering the states $\mathcal{C}$, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{D}$, the resulting intensity matrix for the CTMC equals

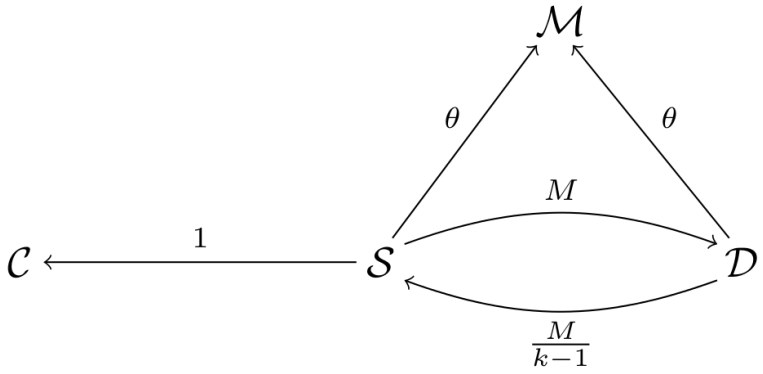


Figure 4.10: An illustration of the state space S when mutations are included in the model.

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & \theta & -(1+\theta+M) & M \\ 0 & \theta & \frac{M}{k-1} & -\left(\theta + \frac{M}{k-1}\right) \end{pmatrix},$$

which in turn provides the matrices

$$\mathbf{t} = \begin{pmatrix} 1 & \theta \\ 0 & \theta \end{pmatrix}, \tag{4.15}$$

and

$$\mathbf{T} = \begin{pmatrix} -(1+\theta+M) & M \\ \frac{M}{k-1} & -\left(\theta + \frac{M}{k-1}\right). \end{pmatrix} \tag{4.16}$$

The PDF, the CDF, as well as the moments for the time T until the CTMC reaches one of the absorbing states $\mathcal{C}$ or $\mathcal{M}$ are as usual obtained with the help of the matrices $\mathbf{T}$ and $\mathbf{t}$, as well as the initial probability vector for the transient states $\boldsymbol{\alpha}$. More importantly though, the probability vector $\boldsymbol{\phi} = (\phi_{\mathcal{C}}, \phi_{\mathcal{M}})$ for the absorption probabilities is given by Equation (2.22). Recall that the entry $\phi_{\mathcal{C}}$ equals the probability that the CTMC is absorbed into state $\mathcal{C}$, whereas the entry $\phi_{\mathcal{M}}$ equals the probability that

the CTMC is absorbed into state $\mathcal{M}$. In other words, with the current model including mutations, $\phi_{\mathcal{C}} = 1-\phi_{\mathcal{M}}$ is the probability that coalescence is reached prior to a mutation, meaning that the two selected genes have been passed on through generations without the occurrence of a mutation, thus making the genes identical. To obtain $\boldsymbol{\phi}$, we note that the inverse of $\mathbf{T}$ given in Equation (4.16) equals

$$-\mathbf{T}^{-1} = \begin{pmatrix} \frac{\theta(k-1)+M}{\theta(1+\theta+M)(k-1)+M(1+\theta)} & \frac{M(k-1)}{\theta(1+\theta+M)(k-1)+M(1+\theta)} \\ \frac{M}{\theta(1+\theta+M)(k-1)+M(1+\theta)} & \frac{(1+\theta+M)(k-1)}{\theta(1+\theta+M)(k-1)+M(1+\theta)} \end{pmatrix}.$$

Multiplying $-\mathbf{T}^{-1}$ with the matrix $\mathbf{t}$ given in Equation (4.15) yields

$$-\mathbf{T}^{-1}\mathbf{t} = \begin{pmatrix} \frac{\theta(k-1)+M}{\theta(1+\theta+M)(k-1)+M(1+\theta)} & 1 - \frac{\theta(k-1)+M}{\theta(1+\theta+M)(k-1)+M(1+\theta)} \\ \frac{M}{\theta(1+\theta+M)(k-1)+M(1+\theta)} & 1 - \frac{M}{\theta(1+\theta+M)(k-1)+M(1+\theta)} \end{pmatrix}.$$

If the initial sample is selected from the same island, we have $\boldsymbol{\alpha}' = (1, 0)$, and thus by Equation (2.22),

$$\phi_{\mathcal{C}} = \frac{\theta(k-1)+M}{\theta(1+\theta+M)(k-1)+M(1+\theta)}, \tag{4.17}$$

whereas if the initial sample is taken from two different islands, we have $\boldsymbol{\alpha}' = (0, 1)$, and hence

$$\phi_{\mathcal{C}} = \frac{M}{\theta(1+\theta+M)(k-1)+M(1+\theta)}. \tag{4.18}$$

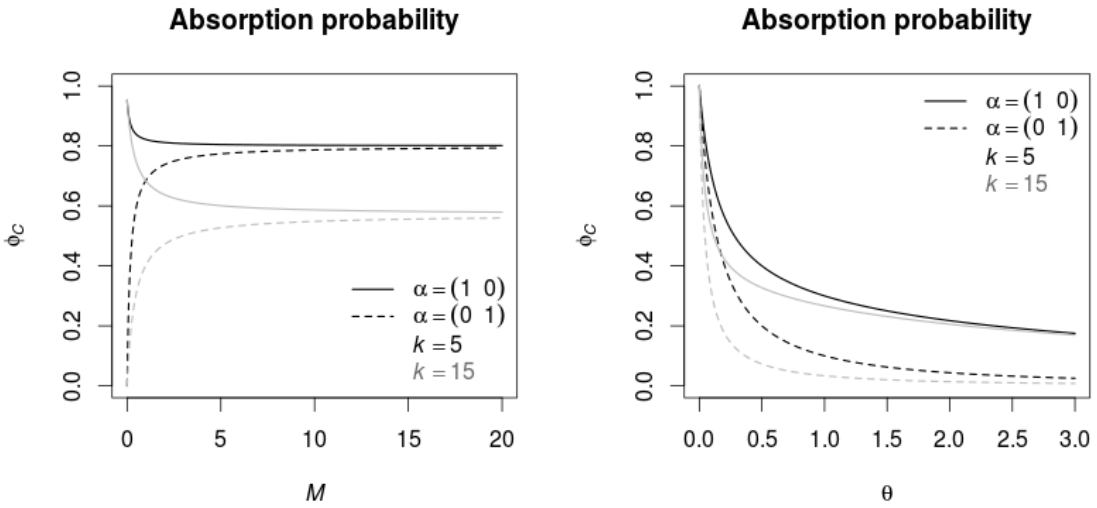
Equations (4.17) and (4.18) are demonstrated in Figure 4.11.

5 Conclusions and discussion

This study has shown that a variety of coalescent models can be analyzed in a matrix-analytic framework, obtained by first identifying an appropriate CTMC, and thereafter applying the theory for phase-type distributions.

Section 4.1 presents an intuitive phase-type distribution approach to the Kingman's coalescent, by considering the coalescent tree as a CTMC $\{X(t),\ t \geq 0\}$, where $X(t)$ represents the number of ancestral lineages at time t. In particular, $X(0)$ equals the number of genes sampled at present time. In Section 4.1.1 we provide matrix expressions for the PDF and the CDF, as well as the moments of the height T_{nh} of the Kingman's coalescent tree with an initial sample of n genes. This is done by stating the appropriate block-partitioned generator matrix and initial probability vector, and thereafter referring to the corresponding equations in Section 2. We then demonstrate that the resulting expected tree height agrees with existing literature.

It is possible to distinguish a similarity between the generator matrix given in Equation (4.1) and the generator matrix given by Equation (2.15) in Example 2.2, where



(a) θ held fixed. **(b)** M held fixed.

Figure 4.11: The probability ϕ_C that the two sampled genes coalesce before a mutation occurs on the lineages, when the genes are originally sampled from the same islands (continuous line) and from different islands (dashed line). The number of islands in the illustrations is $k = 5$, and $k = 15$ (in grayscale). Figure **(a)** illustrates ϕ_C as a function of M, keeping fixed $\theta = 0.1$, whereas Figure **(b)** illustrates ϕ_C as a function of θ, keeping fixed $M = 2$.

the Erlang distribution was discussed. In fact, the distribution of T_{nh} is the *generalized* Erlang distribution, or *hypoexponential* distribution, i.e. the distribution of a sum of exponentially distributed random variables with (possibly) different rates [Bolch et al., 2006]. In our case, $T_{nh} = \sum_i t_i$, where $2 \leq i \leq n$ and each t_i is exponentially distributed with rate $\binom{i}{2}$.

Section 4.1.2 further discusses how the resulting matrix-analytic framework can be applied to the Kingman's coalescent, when mutations are included in the model. In particular, we demonstrate that it is easy to obtain the probability ϕ_0 that a gene is passed from the MRCA to a set of individuals sampled at present time, without the occurrence of a mutation in the genealogies. Also, the relationship between ϕ_0 and the Laplace-transform of the total tree length τ shown in Equation (4.6), results in a convenient matrix expression for the moments of τ, without requiring any detailed information on its distribution.

The application of phase-type distributions to the symmetric island model is discussed in section 4.2. In Section 4.2.1, we examine the special case with two islands and arbitrary sample size, and the initial sample is assumed to consist of some number of individuals from both islands. In particular, we investigate the time it takes either to reach a mutation, or a complete coalescence of the initial sample, as well as the absorption probability distribution. What is not explicitly discussed however, are the tree height, as well as the total tree length. These two quantities can nonetheless be easily

obtained from the model; the time until absorption equals the height of the coalescent, if the mutation rate is set to $\theta = 0$. Furthermore, when $\theta > 0$, the moments of the total tree length can be obtained similarly as in Section 4.1 by implementing Equation (4.7).

It can be argued that the generator matrix for the CTMC in Section 4.2.1 is cumbersome to work with, due to the relatively large matrix dimensions even for comparably small sample sizes. It is in fact possible to reduce the corresponding state space by merging together states that are equivalent by symmetry. This is done by considering state (i, j) to be equivalent with state (j, i), thus diminishing the state space by approximately a half. On the other hand, this reduction of the state space results in some loss of information; for instance, it becomes impossible to deduce which island a complete coalescence or a mutation would occur on.

Finally, in Section 4.2.2, we consider the special case of the symmetric island model where the number of islands equals k, and where the initial sample consists of exactly two genes. The initial sample is taken either from the same island, or from two different islands, depending on the selection of the initial probability vector $\boldsymbol{\alpha}$ for the transient states. While excluding mutations, we determine the matrix forms of the density, distribution and moments of the tree height T, and from these we verify by expanding the matrix expressions, that $\mathrm{E}[T]$ and $\mathrm{Var}[T]$ are consistent with the "non-matrix" forms given in existing literature. Thereafter, when including mutations, we show how the probability that two genes sampled at present time are identical, can very easily be extracted from the model.

Although this study does not in itself provide new results in population genetics – but rather contributes with a new perspective – it is interesting to ascertain that the matrix-analytic framework that results from using phase-type distributions leads to such a compact description of the coalescent. One advantage of our proposed strategy is that it becomes easy to program a computer to perform the matrix calculations to obtain a complete representation of a given coalescent model. Sufficient input variables for such a program would be the initial sample size(s) and an initial probability vector $\boldsymbol{\alpha}$, as well as a generator matrix $\mathbf{Q}$ – which in turn could be a function of potential migration and/or mutation rates, depending on the structure of the coalescent. Another advantage is that the derivation of some specific properties, such as the expectation and variance of the tree height, becomes intuitive and clear – even without extensive experience in population genetics. Moreover, the phase-type perspective that has been introduced, can hopefully be extended to a wider range of population complications that have been left out of this study. For instance, as already hinted by [Wooding and Rogers, 2002], one could apply the framework to time-dependent population sizes. As another example, the phase-type approach could be applied to more general population subdivisions than the symmetric island model. Such population subdivisions, referred to as *stepping stone models* [Hein et al., 2005, Sect. 4.6.3], relax the assumption that a gene migrates to any other island with equal probability, and thus assume a more specific migration dynamic.

References

[Asmussen et al., 1996] Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23:419–441.

[Bellman, 1960] Bellman, R. (1960). *Introduction to matrix analysis*. McGraw-Hill, New York.

[Bobbio et al., 2003] Bobbio, A., Horváth, A., Scarpa, M., and Telek, M. (2003). Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. *Performance evaluation*, 54(1):1–32.

[Bolch et al., 2006] Bolch, G., Greiner, S., de Meer, H., and Trivedi, K. S. (2006). *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons, New York.

[Durrett, 2008] Durrett, R. (2008). *Probability models for DNA sequence evolution*. Springer Science & Business Media, New York.

[Hein et al., 2005] Hein, J., Schierup, M., and Wiuf, C. (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford university press, New York.

[Hössjer et al., 2014] Hössjer, O., Olsson, F., Laikre, L., and Ryman, N. (2014). A new general analytical approach for modeling patterns of genetic differentiation and effective size of subdivided populations over time. *Mathematical Biosciences*, 258:113–133.

[Khoury et al., 1993] Khoury, M. J., Beaty, T. H., and Cohen, B. H. (1993). *Fundamentals of genetic epidemiology*, volume 22. Oxford University Press, New York.

[Latouche and Ramaswami, 1999] Latouche, G. and Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*, volume 5. SIAM, Philadelphia.

[Neuts, 1981] Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Dover Corporation.

[Norris, 1997] Norris, J. (1997). *Markov Chains*, volume 2. Cambridge University Press, Cambridge.

[Resnick, 2002] Resnick, S. I. (2002). *Adventures in stochastic processes*. Birkhäuser, Boston.

[Strobeck, 1987] Strobeck, C. (1987). Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics*, 117(1):149–153.

[Wooding and Rogers, 2002] Wooding, S. and Rogers, A. (2002). The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics*, 161(4):1641–1650.