universität
wien

# MAGISTERARBEIT

Titel der Magisterarbeit

## „Calculation of the transition density of allele frequencies in probabilistic models in population genetics“

Verfasserin
Julia Theresa Csar, Bakk.rer.soc.oec.

angestrebter akademischer Grad

Magistra der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, 2013

| | |
|---|---|
| Studienkennzahl lt. Studienblatt: | A 066 951 |
| Studienrichtung lt. Studienblatt: | Magisterstudium Statistik |
| Betreuer / Betreuerin: | Ao.Univ.Prof. Mag. Dr. Andreas Futschik |

## Zusammenfassung

Diese Diplomarbeit widmet sich der Berechnung der Wahrscheinlichkeitsdichte der Allelfrequenzen in einer Population über die Zeit innerhalb eines bi-allelischen Modells mit Selektion und Drift. Dafür wird von einem Diffusionsmodell ausgegangen, welches das Wright-Fisher- und das Moran-Modell approximiert. Diese Modelle werden kurz vorgestellt. Die Dichte der Allelfrequenzen ist eine Lösung der Kolmogorov Vorwärts- und Rückwärtsgleichungen, die sich aus dem Diffusionsmodell ergeben. Eine Lösung zu diesen Gleichungen wurde bereits von Kimura (1955) gefunden. Der Lösungsweg von Kimura (1955) und ein ähnlicher von Song and Steinrücken (2012) wird beschrieben und ein weiterer Lösungansatz vorgestellt. Der neue Lösungsansatz basiert auf der Transformation der Kolmogorov Vorwärtsgleichung in eine Differentialgleichung, die durch Sphäroidfunktionen gelöst wird. Im Anwendungskapitel werden die unterschiedlichen Dichtefunktionen anhand von Simulationen und Graphiken verglichen.

**Abstract**

This thesis is about the derivation of a formula for the density of allele frequencies of a biallelic model with selection and drift. The starting point is the diffusion model approximating the Wright-Fisher and the Moran model. A short summary of the theory of these models is given in this thesis. The allele frequency transition density is a solution of the Kolmogorov forward and backward equation resulting from the diffusion approximation. A solution by Kimura (1955) already exists. The method of Kimura (1955) and a similar method of Song and Steinrücken (2012) are described and a further approach is proposed. In the chapter on applications, the resulting functions are compared by simulations and graphs.

# Contents

# Chapter 1

# Introduction

The evolution of allele-frequencies of a biallelic model with selection and drift over time in a population is described by the transition density function of the Wright-Fisher or Moran models. Both models can be approximated by a diffusion process. For this bi-allelic model with selection and drift, the Kolmogorov forward and backward equations have been derived. The only parameter is the scaled selection coefficient $\gamma \propto sN$. The allelic proportion given the starting distribution of $\phi(x|p,t,\gamma)$ can be found for all times by solving the Kolmogorov forward equation. Kimura (1955) gave a spectral representation of such a solution using the Gegenbauer polynomials. Selection was accounted for by expanding into a Taylor series. After Kimura (1955), the problem has not been considered for many years. Only lately the topic got more attention; e.g. Mano (2009) used the solution of Kimura (1955). Recently Song and Steinrücken (2012) proposed a similar, but slightly different way to find the solution and an algorithm for the computation that was not based on a Taylor series expansion, but on the solution of an infinite-dimensional system of linear equations. In this thesis it is shown that the differential equation can be solved using the spheroidal wave functions by transforming the forward equation into Sturm-Liouville form. Connections to the approaches of Kimura (1955) and Song and Steinrücken (2012) are pointed out. The spheroidal wave functions and their eigenvalues are implemented in Mathematica (Weisstein, 2013c) and in a Mathematica package implemented by Falloon (2003) and are thus easily available to population geneticists. This thesis is organised as follows: in Chapter 2 the Wright-Fisher and Moran model are described and in Chapter 3 the theory for diffusion processes is summarised. The main part of this thesis is contained in Chapter 4, where the calculation of the allele frequency density is described. In the Chapter **??**, a summary of the used mathematical background is given. The different solutions are compared using graphs and simulations and the Mathematica code is provided.

# Chapter 2

# Probabilistic Models in Population Genetics

In population genetics the *Wright-Fisher* and the *Moran* model are used to describe the change in allele frequencies in a population over generations. We are only considering the bi-allelic selection drift models without mutation.

The main difference between these two models is that in the Wright-Fisher model, the generations are strictly non-overlapping. The Moran model has the advantage that some properties can be calculated exactly where in the Wright-Fisher model they can only be approximated (Baake, 2008).

## 2.1 Assumptions

In both models it is assumed that the population is *finite*; in the Wright-Fisher model individuals are assumed to be *diploid*, in the Moran model *haploid.* Only bi-allelic models are considered, which means that at the relevant locus, there are only two different possible alleles, $A_1$ and $A_2$. The population is assumed to be *random mating* with only one mating type, so the population is *monoecious* (Wakeley, 2009). In a model with selection, we assume that the allele $A_2$ is favored, without loss of generality. The number of individuals having the allele $A_1$ is $Y$ and the proportion is $x = \frac{Y}{2N}$, where $N$ is the population size. The interest lies in the change in the frequency of the alleles. The number of alleles of type $A_1$ at generation $t$ is denoted as $Y_t$ and the relative frequency as $x_t$ $(0 < x_t < 1)$. The relative frequency of the allele $A_2$ is then $1 - x_t$ (Charlesworth and Charlesworth, 2010; Ewens, 2000).

### 2.1.1 Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium gives the proportion of the different diploid genotypes in the population, if the population is random mating. When the relative frequency of allele $A_1$ is $x_t$ the frequencies of the genotypes are given in Table2.1 (Ewens, 2000). Hardy Weinberg equilibrium is reached in a single generation, when the generations are non-overlapping. When the generations are overlapping Hardy Weinberg equilibrium is reached only asymptotically (Charlesworth and Charlesworth, 2010).

| genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| frequencies | $x_t^2$ | $2x_t(1-x_t)$ | $(1-x_t)^2$ |

Table 2.1: Hardy Weinberg equilibrium

### 2.1.2 Fitness associated with Genotypes

With each genotype a different fitness score is associated. The fitness of genotype $A_1A_1$ is $w_{11}$ and of the other genotypes $w_{12}$ and $w_{22}$, respectively. The *mean fitness* in the population is $\bar{w}$. The *relative fitness* is calculated as the total fitness divided by the mean fitness $\bar{w}$. The fitness is dependent on the *selection parameter $s$* and the *heterozygous dominance $h$*. Common ways to model the fitness parameters are given in table 2.1.2 (Ewens, 2000). *Fitness 4* is an adaptation of *fitness 1* which is used by Song and Steinrücken (2012) and will be used later.

$$\bar{w} = x_t^2 w_{11} + 2x_t(1-x_t)w_{12} + (1-x_t)^2 w_{22}$$

Table 2.2: Fitness parameters

| genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| fitness 1 | 1 | $1+sh$ | $1+s$ |
| fitness 2 | $1-s_1$ | 1 | $1-s_2$ |
| fitness 3 | 1 | $1+s_1$ | $1+s_2$ |
| fitness 4 | 1 | $1+2sh$ | $1+2s$ |

## 2.2 Wright-Fisher Model

The change in allele frequency $x$ is modelled in a random mating diploid and finite population. The genes in generation $t+1$ are derived by random sampling with replacement from generation $t$ (Ewens, 2000). The offspring of generation $t$ replaces all individuals, such that the generations are completely non-overlapping.

This is a Markov process where the number of alleles $A_1$ is the Markovian variable ($Y(t) = 2Nx_t$). The transition probability from generation $t$ with $i$ alleles of type 1 to generation $t+1$ with $j$ alleles of type 1 is $p_{ij}$ as given in equation 2.1 ($x_t = \frac{i}{2N}$) and is binomial. In a haploid population $2N$ is replaced by $N$

$$p_{ij} = \binom{2N}{j} x_t^j (1 - x_t^{2N-j}) \,. \tag{2.1}$$

The expected value and variance then are $E[Y_{t+1}] = 2Nx_t = Y_t$ and $Var[Y_{t+1}] = 2Nx_t(1-x_t) = Y_t(1-x_t)$ (Ewens, 2000). Assuming that the different genotypes have different fitness values ($w_{11}$, $w_{12}$ and $w_{22}$) the transition probabilities can be written as

$$p_{ij} = \binom{2N}{j} \eta_i^j (1 - \eta_i)^{2N-j} \tag{2.2}$$

where $\eta_i$ is given as (Ewens, 2000):

$$\eta_i = \frac{w_{11}x_t^2 + w_{12}x_t(1-x_t)}{\bar{w}} \,. \tag{2.3}$$

The change in the allele frequencies from one generation to the next is $\delta x$, such that $x_{t+1} = x_t + \delta x$ and can be calculated as follows (Ewens, 2000):

$$x_{t+1} - x_t = \frac{x_t(1-x_t)}{\bar{w}}(w_{11}x_t + w_{12}(1-2x_t) - w_{22}(1-x_t)) \,.$$

The variance of $\delta x$ is then $V_{\delta x} = \frac{x(1-x)}{2N}$ (Charlesworth and Charlesworth, 2010).

## 2.3 Moran Model

In contrast to the Wright-Fisher model where the population of one generation completely replaces the preceding generation, the generations in the Moran model are overlapping. To model drift, two haploid individuals are chosen at random with replacement from the population at each step. The first chosen individual reproduces and it's duplicate replaces the second individual (Wakeley, 2009). So every birth event is coupled with a death event (Baake, 2008).

If the number of alleles of type 1 ($A_1$) in a *haploid* population is $i$ at time step $t$ ($Y_t = i$), then there are three possible values at time step $t+1$: the number of alleles can stay the same or change by one. This results in a Markov process with a tridiagonal transition matrix with transition probabilities $p_{ij}$ in the case without selection or mutation.

$$p_{ij} = \begin{cases} x(1-x) & \text{if } j = i+1 \text{ or } j = i-1 \\ x^2 + (1-x)^2 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

The probability of the number of alleles changing by one is $2x(1-x)$, half up and half down, and the probability of staying at the same value is $x^2 + (1-x)^2$.

With this knowledge the expectation and the variance of the frequency of allele $A_1$ can be computed. Without selection the mean frequency and the variance are $E[Y_{t+1}] = Y_t$ and $Var[Y_{t+1}] = 2x_t(1-x_t)$.

Selection is modelled with parameter $s$ for example as in Vogl and Clemente (2012), where selection increases the transition probability $p_{i,i+1}$ by $sx(1-x)$. The probability of staying at $i$ is then reduced by the same value.

# Chapter 3

# Diffusion Approximation of Probabilistic Models in Population Genetics

The diffusion model in population genetics describes the stochastic process of allele frequency change within one or more populations over time $(X_t)_{t\geq 0}$ (Kimura, 1955). Starting point is for example the *Wright-Fisher* model or the *Moran* model (see Chapter 2). The probability density function $\phi(x|p,t)$ of the allele frequency implied by the diffusion process is of interest. $x$ is the frequency of allele $A_1$ at time $t$. At time $t = 0$ the allele frequency is $p$. This probability density function is found as a solution to the Kolmogorov forward and backward equation of the diffusion process. The two equations are presented in Subsection 3.7.

In this chapter the diffusion approximations for the two most used probabilistic models in population genetics (the Wright-Fisher and the Moran model) are derived and studied in more detail. The Wright-Fisher and the Moran model share the same diffusion limit up to a factor 2. To make results comparable in literature often the factor 2 is included in the selection or time parameters of the Moran Model (Baake, 2008). The theory of diffusion processes is summarized mainly from Karlin and Taylor (1981).

The population genetic models explained in the preceding chapter are in discrete time, however exponential waiting times can be added to the Moran model. For further calculation a continuous time process would be easier. To transform to continuous time the diffusion theory is used and the resulting model is then called a *diffusion process* or *diffusion model.* To derive the diffusion approximation of the Wright-Fisher model, the Markov chain describing the model has to be approximated in continuous time. The properties at the boundaries play an important role and have to be taken into account.

A diffusion process is a strong Markov process with continuous sample paths. To show that a standard Markov process is a diffusion it is sufficient to show that the Dynkin condition is satisfied (Karlin and Taylor, 1981).

## 3.1 Definition

A *diffusion process* is a continuous time stochastic process, which possesses the *strong Markov property* and for which the sample paths $X_t$ are almost everywhere *continuous functions* of $t$ (Karlin and Taylor, 1981). The process is *regular*, which means that when starting from any point $p$ within the domain of definition of the process, reaching any other point within has positive probability. $X_t$ is the frequency of one allele in a bi-allelic model at time $t$. The behaviour at the boundaries is dependent on the mutation rate. In a model without mutation the boundaries are *absorbing,* with mutation the boundaries are *regular.*

A process $x$ is called a stochastic process, if the probability that the change of $x$ within a time interval $(t, t + \delta t)$ is bigger than $\epsilon$, is of order smaller than $\delta t$ $(o(\delta t))$, with $\epsilon > 0$ (Crow and Kimura, 1970).

$$P[|X_t - X_{t+\delta t}| > \epsilon] = o(\delta t) \tag{3.1}$$

In population genetics the interesting process is the change in relative allele frequencies; therefore the diffusion process used is of course defined on the interval $[0, 1]$.

A *standard Markov process* satisfying the *Dynkin condition* is a diffusion process. The Dynkin condition is given in equation 3.2:

$$\lim_{h \downarrow 0} \frac{1}{h} P[|X_{t+h} - p| > \epsilon \,|\, X_t = p] = 0 \qquad \forall p \in [0, 1] \text{ and } \epsilon > 0\,. \tag{3.2}$$

The diffusion process can be fully defined by the boundary conditions and the infinitesimal parameters: the *infinitesimal mean* $\mu(x)$ and the *infinitesimal variance* $\sigma^2(x)$. In our context, $\mu(x)$ is the mean change in allele frequencies and $\sigma^2(x)$ is the variance of the change. In the theory of diffusion processes, $\mu(x)$ is also called the *drift part* and $\sigma^2(x)$ the *diffusion part* of the process. Since the term *drift* may be misleading in a population genetics context, we will refer to $\mu(x)$ as the *infinitesimal mean.*

If $\triangle_{\delta t} X_t = X_{t+\delta t} - X_{\delta t}$ is the change of the process over the interval $[t, t+h]$, then the *infinitesimal mean* and the *infinitesimal variance* are defined as (Karlin and Taylor, 1981)

$$\lim_{\delta t \downarrow 0} \frac{1}{\delta t} E[\triangle_{\delta t} X_t \,|\, X_t = x] = \mu(x,t) = \mu(x) \tag{3.3}$$

$$\lim_{\delta t \downarrow 0} \frac{1}{\delta t} E[(\triangle_{\delta t} X_t)^2 \,|\, X_t = x] = \sigma^2(x,t) = \sigma^2(x)\,. \tag{3.4}$$

We are only considering *time homogeneous* processes, such that $\mu(x)$ and $\sigma^2(x)$ are independent of $t$.

Usually the following equation is satisfied for higher moments ($r = 3, 4, 5, \dots$) (Karlin and Taylor, 1981):

$$\lim_{\delta t \downarrow 0} \frac{1}{\delta t} E[|\triangle_{\delta t} X_t|^r | X_t = x] = 0\,. \tag{3.5}$$

In table 3.1 the infinitesimal mean and infinitesimal variance in different population genetics models are shown. Selection is modelled by the selection coefficient $s$ and the dominance parameter $h$. In models without dominance the coefficient $h$ is $\frac{1}{2}$. Mutation is assumed to be absent. The scaling with $2N$ of the infinitesimal parameters is unfortunately not consistent in literature.

The infinitesimal parameters can be derived from different models (see 2.1.2). From this the scale and speed function can be calculated (equations 3.9 and 3.11), which give further information about the behaviour of the process.

Table 3.1: some examples for $\mu(x)$ and $\sigma^2(x)$ in different models

| | $\mu(x)$ | $\sigma^2(x)$ |
|---|---:|---:|
| only drift | $0$ | $x(1-x)\frac{1}{2N}$ |
| with selection and dominance | $sx(1-x)(x+h(1-2x))$ | $x(1-x)\frac{1}{2N}$ |
| with selection and no dominance ($h = \frac{1}{2}$) | $sx(1-x)\frac{1}{2}$ | $x(1-x)\frac{1}{2N}$ |

## 3.2 Transition Probability

To construct the diffusion process, first the discrete time Markov chain describing the change in allele frequencies over generations has to be defined.

$T_{ij}$ is the *transition probability* from state $i$ to state $j$ of the underlying Markov chain. $\phi(x|p,t) \simeq T_{px}^{(t)}$ is the probability of moving from the starting point $p$ to $x$ within $t$ generations. This probability is defined recursively in equation 3.6.

$$\phi(j|p,t+1) = \sum_i \phi(i|p,t) T_{i,j} \tag{3.6}$$

The initial distribution at time $t = 0$ is a function with point mass at $x = p$. This can be expressed by the Dirac delta distribution.

$$p(x|p, t = 0) = \delta(x - p) = \begin{cases} 1 & \text{if } x = p \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

Since we are interested in small changes in $x$ during small changes in $t$ the equation 3.6 is rewritten to

$$\phi(x + \delta x|p, t + \delta t) = \int_0^1 \phi(x|p, t)\phi(x + \delta x|x, \delta t)dx\,, \tag{3.8}$$

where $t \pm \delta t$ is a very small time interval.

## 3.3 Speed and Scale function

The scale function $S(x)$ of the diffusion process is defined in equation 3.9 and the speed function $m(x)$ in equation 3.11. For more details on these functions see Ewens (2000, Section 4.7) and Karlin and Taylor (1981, p. 194-195). In Karlin and Taylor (1981) the integrals are defined indefinite (starting from $-\infty$) and in Ewens (2000) the integrals start at some arbitrary value $x_0 \neq -\infty$. Which definition of the integral boundaries is more appropriate, depends on the context.

The speed and scale function will recur in the following sections in the solutions of different problems.

$$S(x) = \int_{x_0}^{x} e^{-2\int_{x_0}^{y} \frac{\mu(z)}{\sigma^2(z)}dz} dy = \int_{x_0}^{x} s(y)dy \tag{3.9}$$

$$s(x) = e^{-\int_{x_0}^{x} \frac{2\mu(y)}{\sigma^2(y)}dy} \tag{3.10}$$

$$m(x) = 2\int_{x_0}^{x} \sigma^2(y)^{-1} e^{\int_{x_0}^{y} \frac{2\mu(z)}{\sigma^2(z)}dz} dy = \frac{1}{\sigma^2(x)s(x)} \tag{3.11}$$

### 3.3.1 Scale function

The scale function can be used to rescale the process $(X_t)_{t\geq 0}$, which is defined on the interval $(l, u)$, to a process $Y_t = S(X_t)$ defined on the interval $(S(l), S(r))$. The infinitesimal parameters of the new process $(Y_t)_{t\geq 0}$ are then given as (Karlin and Taylor, 1981):

$$\mu_y(y) = \frac{1}{2}\sigma^2(y)S''(y) + \mu(y)S'(y) = 0 \tag{3.12}$$

and

$$\sigma_y^2(y) = \sigma^2(y)S'(y)^2 \,. \tag{3.13}$$

In the case with selection and no mutation the equation can be simplified for special values of $h$. For no dominance ($h = \frac{1}{2}$) the scale function becomes $S_{h=\frac{1}{2}}(x)$ in 3.14 ($c$ is some arbitrary constant) (Ewens, 2000).

$$S_{h=\frac{1}{2}}(x) = \int_{x_0}^{x} e^{sy} dy = \frac{1}{s}[e^{sx} - e^{sx_0}] \tag{3.14}$$

### 3.3.2 Speed function

As the name suggests the speed function can be used to estimate the time spent in an interval. For a diffusion process in its natural scale the derivative of the speed function $m(x)$ is

$$\frac{dm(x)}{dx} = \frac{2}{\sigma^2(x)} \,. \tag{3.15}$$

Larger values of $\frac{dm(x)}{dx}$ point to a longer mean time for leaving an interval (Ewens, 2000). The time, that the process will spend in the interval $[x-\epsilon, x+\epsilon]$ when starting in $x$ is proportional to the speed function, with $\epsilon > 0$

$$E[min(T_{x-\epsilon}, T_{x+\epsilon})] \propto m(x) \tag{3.16}$$

## 3.4 Hitting Times

The process $(X_t)_{t\geq 0}$ reaches $a$ the first time at time $T_a$ and $T_{a,b} = min(T_a, T_b)$, is the time at which either $a$ or $b$ have been reached. The probability that the process reaches $b$ before $a$ is $u_{a,b}(p)$, when starting in $p$. (Karlin and Taylor, 1981).

$$u_{a,b}(p) = P[T_b < T_a | X_0 = p] \qquad a < p < b \tag{3.17}$$

The mean time to reach either $a$ or $b$ when starting in $p$ is $v_{a,b}(p)$ (Karlin and Taylor, 1981).

$$v_{a,b}(p) = E[T_{a,b} | X_0 = p] \tag{3.18}$$

The functions $u_{a,b}(p)$ and $v_{a,b}(p)$ can be found by solving differential equations. The solutions are explained in the next subsections, which are a summary of the methods explained in Karlin and Taylor (1981).

### 3.4.1 Probability to reach $b$ before $a$

The probability of reaching $b$ before $a$ is $u_{a,b}(p)$ given in equation 3.17 (Karlin and Taylor, 1981). The probability of reaching $b$ before $a$ when starting in $a$ is of course 0. Therefore the boundary conditions are: $u_{a,b}(a) = 0$ and $u_{a,b}(b) = 1$.

The probability $u_{a,b}(p)$ can be expressed conditionally on the value of the process at time $t'$ $(X_{t'})$, where the error term $o(t')$ is of smaller order than $t'$ (Karlin and Taylor, 1981).

$$u_{a,b}(p) = E[u_{a,b}(X_{t'})|X_0 = p] + o(t') \tag{3.19}$$

$\triangle X$ is the change in $x$ during the time interval $[0, t']$ $(\triangle X = X_{t'} - p)$. Expanding in a Taylor series leads to:

$$u_{a,b}(X_{t'}) = u_{a,b}(p + \triangle X) = u_{a,b}(p) + \triangle X u'_{a,b}(p) + \frac{1}{2}(\triangle X)^2 u''_{a,b}(p) + \dots$$

Since $(X_t)_{t \geq 0}$ is a diffusion process and the change in time $t'$ can be expressed by the infinitesimal mean and its higher moments are 0 for $t' \to 0$.

$E[\triangle X|X_0 = p] = \mu(p)t' + o(t')$ and $E[(\triangle X)^2|X_0 = p] = \sigma^2(p)t' + o(t')$.

Hence expanding in a Taylor series and expressing the change by the infinitesimal mean and variance leads to the following calculation:

$$\begin{aligned}
u_{a,b}(p) &= E[u_{a,b}(X_{t'})|X_0 = p] \\
&= E[u_{a,b}(p) + \triangle X u'_{a,b}(p) + \frac{1}{2}(\triangle X)^2 u''_{a,b}(p)|X_0 = p] + o(t') \\
&= u_{a,b}(p) + \mu(p)t' u'_{a,b}(p) + \frac{1}{2}\sigma^2(p)t' u''_{a,b}(p) + o(t') \\
0 &= \mu(p)t' u'_{a,b}(p) + \frac{1}{2}\sigma^2(p)t' u''_{a,b}(p) + o(t')\,.
\end{aligned}$$

For $t' \to 0$ we can conclude that $u_{a,b}(p)$ satisfies the differential equation (Karlin and Taylor, 1981):

$$0 = \mu(x)\frac{\partial u(x)}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2 u(x)}{\partial x^2} \qquad a < x < b \tag{3.20}$$

The solution of this differential equation is explained in Karlin and Taylor (1981, 3.10) and can be seen in equation 3.21. Note that the result does not depend on the lower limit of the integration of the scale function $S(.)$ (equation 3.9).

$$u_{a,b}(x) = \frac{S(x) - S(a)}{S(b) - S(a)} \tag{3.21}$$

### 3.4.2 Mean Time to reach $a$ or $b$

The mean time to reach $a$ or $b$ when starting at some value $p$ with $(a < p < b)$ is given in equation 3.18 by $v_{a,b}(p)$. The function $v_{a,b}(p)$ is a special case of $w_{a,b}(x)$, where $g(x) = 1 \, \forall x$ (Karlin and Taylor, 1981).

$$w_{a,b}(p) = E[\int_0^{T_{a,b}} g(X_r)dr \,|\, X_0 = p] \tag{3.22}$$

From the fact that the diffusion process $(X_t)_{t \geq 0}$ satisfies the Markov property, it follows that

$$E[\int_t^{\prime T_{a,b}} g(X_r)dr \,|\, X_t' = z] = E[\int_0^{T_{a,b}} g(X_r)dr \,|\, X_0 = z] = w_{a,b}(p) \tag{3.23}$$

and, since the sample paths and the function $g$ are continuous, the expected value can be approximated by (Karlin and Taylor, 1981)

$$E[\int_0^{t'} g(X_r)dr \,|\, X_0 = p] = g(x)t' + o(t') \,. \tag{3.24}$$

By using the Markov property as above the function $w(p)$ can be expressed as (Karlin and Taylor, 1981)

$$w_{a,b}(p) = E[\int_0^{t'} g(X_r)dr \,|\, X_0 = p] + E[w_{a,b}(X_{t'}) \,|\, X_0 = p] \,. \tag{3.25}$$

Expanding $w_{a,b}(p + \triangle X)$ in Taylor series leads to:

$$E[w_{a,b}(X_{t'}) \,|\, X_0 = p] = w_{a,b}(p) + \mu(x)w'_{a,b}(x)t' + \frac{1}{2}\sigma^2(p)w''_{a,b}(p)t' + o(t') \tag{3.26}$$

Plugging this into 3.25 gets:

$$0 = g(p)t' + \mu(p)w'_{a,b}(p)t' + \frac{1}{2}\sigma^2(p)w''_{a,b}(p)t' + o(t') \tag{3.27}$$

when letting $t' \to 0$ it can be seen that $w_{a,b}$ satisfies the differential equation 3.28.

$$-g(x) = \mu(x)\frac{dw}{dx} + \frac{1}{2}\sigma^2(x)\frac{d^2w}{dx^2} \qquad a < x < b \tag{3.28}$$

As stated before $v(.)$ is a special case of $w(.)$ where the function $g(x) = 1 \, \forall x$. Therefore the differential equation satisfied by $v_{a,b}(p)$ is equation 3.29.

$$-1 = \mu(x)\frac{dv}{dx} + \frac{1}{2}\sigma^2(x)\frac{d^2v}{dx^2} \qquad a < x < b \tag{3.29}$$

The solutions to the differential equations are explained in Karlin and Taylor (1981, 3.11 and 3.10) and are

$$w_{a,b}(x) = 2[u_{a,b}(x) \int_x^b (S(b) - S(\xi))m(\xi)g(\xi)d\xi \tag{3.30}$$

$$+ (1 + u(x)) \int_a^x (S(\xi) - S(a))m(\xi)g(\xi)d\xi] \tag{3.31}$$

$S(.)$ is the scale function (equation 3.9) and $m(.)$ is the speed function (equation 3.11).

## 3.5 Probability of Absorption

$P(X_t = 0, X_0 = p)$ is the probability of reaching 0 after $t$ generations, when the starting frequency was $p$. $P(X_t = 1|X_0 = p)$ is the probability of reaching 1 after $t$ generations. For these probabilities the following differential equation is satisfied (Karlin and Taylor, 1981).

$$\frac{\partial}{\partial t}P(X_t = 0|X_0 = p) = \mu(p)\frac{\partial}{\partial p}P(X_t = 0, X_0 = p) + \frac{1}{2}\sigma^2(p)\frac{\partial^2}{\partial p^2}P(X_t = 0|X_0 = x) \tag{3.32}$$

For $t \to \infty$ $P(X_t = 0|X_0 = p) \to P_0(p)$ and the equation becomes (Karlin and Taylor, 1981)

$$0 = \mu(p)\frac{\partial}{\partial p}P_0(p) + \frac{1}{2}\sigma^2(p)\frac{\partial^2}{\partial p^2}P_0(x) \tag{3.33}$$

$$\frac{\partial^2}{\partial p^2}P_0(p) = -2\frac{\mu(p)}{\sigma^2(p)}\frac{\partial P_0(p)}{\partial p}\,. \tag{3.34}$$

It is obvious that $P_0(0) = 1$ and $P_0(1) = 0$ since 0 and 1 are absorbing states, with $\xi \in [0, 1]$.

$$\frac{\delta}{\delta p}P_0(x) = C.\exp(-1\int_\xi^p)\frac{a(y)}{b(y)}dy \tag{3.35}$$

$$P_0(p) = \frac{\int_p^1 \exp(-2\int_\xi z\frac{\mu(y)}{\sigma^2(y)}dy)dz}{\int_0^1 \exp(-2\int_\xi z\frac{\mu(y)}{\sigma^2(y)}dy)dz} \tag{3.36}$$

The same calculation can be done for $P_1(p)$. So the probability of absorption can be expressed as a function of the scale function $S(.)$ (given in equation 3.9).

$$P_0(p) = \frac{S(1) - S(p)}{S(1) - S(0)} \tag{3.37}$$

$$P_1(p) = \frac{S(p) - S(0)}{S(1) - S(0)} \tag{3.38}$$

Similarly the probability of hitting $b$ before $a$ can be calculated $(0 < a < p < b)$. $T_a$ is a hitting time of $a$ (so $X_{T_a} = a$).

$$P[T_b < T_a | X_o = p] = \frac{S(p) - S(a)}{S(b) - S(a)} \tag{3.39}$$

If the scale function $S(x)$ is a linear function, the diffusion is said to be in its *natural* or *canonical scale.* In this case the probability $P[T_a < T_b | X_0 = p] = \frac{b-p}{b-a}$ is proportional to the distances (Karlin and Taylor, 1981).

## 3.6 Boundaries

The functions $u_{a,b}(z)$ and $v_{a,b}(z)$ are used to gain more information about the behaviour of the process at the boundary point $z$. $u_{a,b}(z)$ is the probability of reaching $b$ before $a$ for $0 < a < x < b < 1$ and $v_{a,b}(x)$ is the estimated time to reach either $a$ or $b$ (see section 3.4 for more information). In our population genetics context the boundaries are 0 and 1. For further information on these function see Ewens (2000, Section 4.7) and Karlin and Taylor (1981, p. 192-202).

To learn more about the boundaries let $a$ decrease to 0 and $b$ increase to 1 separately in those functions.

The probability of reaching $b$ before reaching the boundary 0 can be calculated asymptotically by $\lim_{a \to 0} u_{a,b}(x)$, where $0 < a < x < b < 1$. Remember that $u_{a,b}(x) = \frac{S(x)-S(a)}{S(b)-S(a)}$. Therefore the boundary is called *attracting* if $\lim_{a \to 0} S(x) - S(a) < \infty$ with $0 < a < x < 1$, since this would mean that the boundary will be reached with positive probability before reaching any other point $b$ ($0 < a < x < b < 1$) (Karlin and Taylor, 1981).

The boundary is called *attainable* when it can be reached in finite time. An attainable boundary is also attracting, but an attracting boundary is not necessarily attainable (Karlin and Taylor, 1981).

$$\begin{aligned}
\lim_{a \to 0} v_{a,b}(x) &= E[T_{a,b} | X_0 = x], \qquad 0 < a < x < b < 1, \text{ b fixed} && (3.40)\\
&= \lim_{a \to 0} 2(u_{a,b}(x) \int_x^b (S(b) - S(\xi))m(\xi)d\xi && (3.41)\\
&\quad + (1 + u_{a,b}(x)) \int_a^x (S(\xi) - S(a))m(\xi)d\xi) && (3.42)
\end{aligned}$$

Since 0 is an attracting boundary $\lim_{a \to 0} u_{a,b}(x) < \infty$, which leaves the condition that $\Sigma(0) = \lim_{a \to 0} \int_a^x (S(\xi) - S(a))m(\xi)d\xi$ has to be finite. If and only if $\Sigma(0) < \infty$ the boundary 0 is attainable (Karlin and Taylor, 1981).

## 3.7 Kolmogorov forward and backward equation

The probability density function of the allele frequencies $x$ over time is dependent on the initial allele frequency $p$ at time $t = 0$. The allele frequency at time $t$ has the density $\phi(x|p,t)$. This function can be found as a solution of the *Kolmogorov forward equation* of the diffusion process. In physics the Kolmogorov forward equation is more commonly known as the Fokker-Planck equation. The forward equation is

$$\frac{\partial \phi(x|p,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2(x)\phi(x|p,t)) - \frac{\partial}{\partial x}(\mu(x)\phi(x|p,t)), \qquad 0 < x < 1 \tag{3.43}$$

and the backward equation is

$$\frac{\partial \phi(x|p,t)}{\partial t} = \frac{1}{2}\sigma^2(p)\frac{\partial^2}{\partial p^2}\phi(x|p,t) + \mu(p)\frac{\partial}{\partial p}\phi(x|p,t), \qquad 0 < x < 1 \tag{3.44}$$

(Crow and Kimura, 1970, p. 373). The time parameter is often rescaled to $\tau \propto Nt$. The solutions of the forward and backward equation are connected to each other, such that if $E_n(x)$ is an eigenfunction of the generator of the backward equation then $\frac{E_n(x)}{\sigma^2(x)}$ is an eigenfunction of the generator of the forward equation. $\mathcal{L}_B$ is the backward generator and the forward generator is $\mathcal{L}_F$.

$$\mathcal{L}_F f(x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2(x)f(x)) - \frac{\partial}{\partial x}(\mu(x)f(x))$$

$$\mathcal{L}_B f(x) = \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}f(x) + \mu(x)\frac{\partial}{\partial x}f(x)$$

By defining $\sigma^2(x)$ and $\mu(x)$ as $\sigma^2(x) = \frac{x(1-x)}{2N}$ and $\mu(x) = sx(1-x)$ or as any other possible function in table 3.1 the relationship of the eigenfunctions can be obtained:

$$\begin{aligned}\sigma^2(x)\mathcal{L}_F \frac{f(x)}{\sigma^2(x)} &= \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}f(x) + \mu(x)\frac{\partial}{\partial x}f(x) \\ &= \mathcal{L}_B f(x)\,.\end{aligned}$$

If $E_n(x)$ is an eigenfunction of $\mathcal{L}_B$ with associated eigenvalues $\lambda_n$, then $\mathcal{L}_B E_n(x) = -\lambda_n E_n(x)$. From the above equation it follows that $\mathcal{L}_F \frac{E_n(x)}{\sigma^2(x)} = -\lambda_n \frac{E_n(x)}{\sigma^2(x)}$ and thus it is shown that $\frac{E_n(x)}{\sigma^2(x)}$ is an eigenfunction of $\mathcal{L}_F$ with associated eigenvalues $\lambda_n$.

### 3.7.1 Derivation of the Kolmogorov forward equation

The probability density of the frequency of $A_1$ at time $t$, given that the frequency is $p$ at time $t = 0$ is $\phi(x|p,t)$. That the function satisfies the Kolmogorov forward equation is shown in Crow and Kimura (1970, p. 373-382) and is summarised here. The allele frequency at time $t$ is in the interval $x \pm \frac{1}{2}h$. After some small time change $\delta t$ the frequency will change to $x + h$ with probability $m(x,t)\delta t$ by systematic pressure and $h$ is a real number greater than 0. $v(x,t)\delta t$ is the probability that it moves outside the interval $[x \pm \frac{1}{2}h]$ by random fluctuation.

From this the probability that the gene frequency does not leave the interval $[x \pm \frac{1}{2}h]$ can be calculated in equation 3.45. The probability includes the probability of being in this interval $(\phi(x,t)h)$, the probability of moving from an upper or lower frequency by random fluctuation to $[x \pm \frac{1}{2}h]$ and the probability of moving there by systematic pressure $(m(x,t)\delta t)$. The probability of moving away from this interval by random fluctuation or systematic pressure is subtracted $(v(x,t) + m(x,t)\delta t\phi(x,t)h)$(Karlin and Taylor, 1981).

$$\begin{aligned}\phi(x,t+\delta t)h = \phi(x,t)h - v(x,t) + m(x,t)\delta t\phi(x,t)h \\ + \frac{1}{2}v(x-h,t)\delta t\phi(x-h,t)h + \frac{1}{2}v(x+h,t)\delta t\phi(x+h,t)h \\ + m(x-h,t)\delta t\phi(x-h,t)h \qquad (3.45)\end{aligned}$$

The change in $\phi(x,t+\delta t)h$ due to random change in $\delta t$ is therefore (Karlin and Taylor, 1981)

$$\frac{1}{2}v(x-h,t)\delta t\phi(x-h,t)h + \frac{1}{2}v(x+h,t)\delta t\phi(x+h,t)h\,. \qquad (3.46)$$

So the variance in the change in $x$ per $\delta t$ due to random change is (Karlin and Taylor, 1981)

$$\begin{aligned}\sigma^2(x) = V(x,t)\delta t &= h^2\frac{1}{2}v(x,t)\delta t + (-h)^2\frac{1}{2}v(x,t)\delta t \qquad (3.47)\\ &= h^2v(x,t) \qquad (3.48)\end{aligned}$$

and the mean change is(Karlin and Taylor, 1981)

$$\mu(x) = M(x,t)\delta t = hm(x,t)\delta t\,. \qquad (3.49)$$

Substituting $\mu(x)$ and $\sigma^2(x)$ into equation 3.45 and letting $h \to 0$ and $\delta t \to 0$ leads to the Kolmogorov forward equation (Karlin and Taylor, 1981).

$$\frac{\delta\phi(x|p,t)}{\delta t} = \frac{1}{2}\frac{\delta^2}{\delta x^2}(\sigma^2(x)\phi(x|p,t)) - \frac{\delta}{\delta x}(\mu(x)\phi(x|p,t)) \qquad (3.50)$$

Plugging the mean and the variance of a process with selection and without mutation into the forward equation in equation 4.4 leads to equation 3.51 (Karlin and Taylor, 1981; Crow and Kimura, 1970).

$$\frac{\delta\phi}{\delta t} = \frac{1}{4N}\frac{\delta^2}{\delta x^2}\{x(1-x)\phi\} - s\frac{\delta}{\delta x}\{x(1-x)\phi\} \tag{3.51}$$

In a model without selection and without mutation the forward equation is (Crow and Kimura, 1970, p.383).

$$\frac{\delta\phi}{\delta t} = \frac{1}{4N}\frac{\delta^2}{\delta x^2}\{x(1-x)\phi\}\,. \tag{3.52}$$

### 3.7.2 Kolmogorov Backward Equation

For the backward equation the process is considered into the opposite direction. Opinions differ on whether in this setting the starting allele frequency $p$ is a random variable or not (Crow and Kimura, 1970; Ewens, 2000).

The Kolmogorov backward equation for the process as before is shown in equation3.53 (Crow and Kimura, 1970, p. 373).

$$\frac{\delta\phi(p,x;t)}{\delta t} = \frac{1}{2}\sigma^2(p)\frac{\delta^2}{\delta p^2}\phi(p,x;t) + \mu(p)\frac{\delta}{\delta p}\phi(p,x;t) \tag{3.53}$$

The backward equation is harder to interpret but solutions to the backward equation can be transformed to solutions of the forward equation and vice versa. Song and Steinrücken (2012) proposed a solution to finding the density $\phi(x|p,t)$ with the backward equation as the starting point. This solution is summarised in Section 4.2.

### 3.7.3 Behaviour at the boundaries

The boundaries are $x = 0$ and $x = 1$, which act as absorbing barriers (Crow and Kimura, 1970). Denote $f(x,t) = \phi(x|p,t)\delta x$. When substituting $\delta x = \frac{1}{2N}$, $f(x,t)$ gives an approximation of the probability that the gene frequency is $x$ at time $t$ (this approach is presented in Crow and Kimura (1970)).

It is useful to look at $P(x,t) = -\frac{1}{2}\frac{\delta}{\delta x}(\sigma^2(x)\phi(x|p,t)) - \frac{\delta}{\delta x}\mu(x)\phi(x|p,t)$, where $-\frac{\delta}{\delta x}P(x,t)$ can be interpreted as the rate of probability mass flow across the point $x$ per generation.

So in the case of absorbing barriers we have that

$$\frac{df(0,t)}{dt} = -P(0,t)\frac{df(1,t)}{dt} = P(1,t)\,. \tag{3.54}$$

## 3.8 Spectral Representation

To find a representation of the allele frequency density by solving the Kolmogorov forward or backward equation the spectral representation is useful. The backward generator $\mathcal{L}_B$ is self-adjoint and there exist solutions to $\mathcal{L}_B E(x) = -\lambda E(x)$, with a unique sequence of $\lambda_n$ with $\lambda_n \to \infty$ as $n \to \infty$ (Song and Steinrücken, 2012). These eigenvalues $-\lambda_n$ with $0 \leq n < \infty$ are called the spectrum of $\mathcal{L}_B$ (Karlin and Taylor, 1981) and their associated eigenfunctions $E_n(x)$ form a basis in $L^2([0,1], m(x))$ (Song and Steinrücken, 2012). The speed function and the scale function can give insight into the characteristics of a diffusion process (see section 3.3) $m(x) = \frac{1}{\sigma^2(x)s(x)}$ is the speed function of the process with $s(x) = \mathrm{e}^{-\int_{x_0}^{x} \frac{2\mu(y)}{\sigma^2(y)} dy}$ (Karlin and Taylor, 1981). The Kolmogorov backward equation can be expressed by the backward generator $\mathcal{L}_B$ as

$$\frac{\partial \phi(x|p,t)}{\partial t} = \mathcal{L}_B \phi(x|p,t)\,. \tag{3.55}$$

Then it is obvious that the function $f_n(p,t) = e^{-\lambda_n t} E_n(p)$ satisfies the backward equation. Since $\mathcal{L}_B$ is a linear operator, also a linear combination of $e^{-\lambda_n t} E_n(p)$ is a solution. The spectral representation of the transition density is thus given by

$$\phi(x|p,t) = \sum_{n=0}^{\infty} c_n(x) e^{-\lambda_n t} E_n(p)\,; \tag{3.56}$$

and the coefficients $c_n(x)$ are set to satisfy the initial condition. For the initial condition $\phi(x|p,0) = \delta(p-x)$ the spectral representation is

$$\phi(x|p,t) = \sum_{n=0}^{\infty} e^{-\lambda_n t} m(x) \frac{E_n(p) E_n(x)}{\langle E_n, E_n \rangle_m} \tag{3.57}$$

(Song and Steinrücken, 2012; Karlin and Taylor, 1981).

# Chapter 4

# Calculation of the probability density function of allele frequencies

In this chapter different solutions to the Kolmogorov forward and backward equation are presented. The theory of models in population genetics is given in Chapter 2 and the derivation and theory of diffusion processes can be found in Chapter 3. The first solution for the allele frequency density presented is the solution found by Kimura (1955) in form of a spectral representation using the Gegenbauer polynomials. A second solution by Song and Steinrücken (2012) is presented and in the last section of this chapter a third solution is proposed. The latter solution is based on transforming the Kolmogorov forward equation into the differential equation solved by the angular oblate spheroidal wave functions. For all three solutions, computation is discussed.

Table 4.1: Notation Kimura

| | |
|---|---|
| selection | $s$ |
| time | $t$ |
| infinitesimal variance | $\sigma^2(x) = \frac{1}{2N}x(1-x)$ |
| infinitesimal mean | $\mu(x) = sx(1-x)$ |

## 4.1 Solution by Kimura

Kimura (1955) proposed a solution to the Kolmogorov forward equation of the diffusion process in order to find a representation of the transition density of the allele frequencies. The solution is also presented in Kimura (1964) and Crow and Kimura (1970); the main steps are summarized here. Notation differs among literature. For this section the notation given in table 4.1 is used.

### 4.1.1 Only Drift

First we consider the case without selection and without mutation; i.e. the model where the only force is drift. In this model the infinitesimal mean is $\mu(x) = 0$ and the infinitesimal variance is $\sigma^2(x) = \frac{x(1-x)}{2N}$, where $\sigma^2(x)$ is the binomial variance corresponding to $2N$ alleles (Kimura, 1964).

The Kolmogorov forward equation is then given as

$$\frac{\partial \phi(x \,|\, p, t)}{\partial} = \frac{1}{4N} \frac{\partial^2}{\partial x^2} (x(1-x)\phi(x \,|\, p, t)) \,, \tag{4.1}$$

with $p$ corresponding to the starting frequency ($\phi(x \,|\, p, 0) = \delta(x-p)$), i.e., the initial condition.

Kimura (1964) assumes that the solution is of the form $\phi = TX$ where $T$ is a function depending only on $t$ and $X$ is a function depending only on $x$. Inserting this into the Kolmogorov forward equation leads to the equation

$$\frac{1}{T}\frac{\partial T}{\partial t} = \frac{1}{4NX}\frac{\partial^2}{\partial x^2}\left(x(1-x)X\right) \,.$$

Since the left side depends on $t$ only and the right side depends on $x$ only, both sides have to be equal to a constant $(-\lambda)$ and the equation can be separated into two ordinary differential equations (Kimura, 1964). The first differential equation is

$$\frac{dT}{dt} = -\lambda T \,,$$

from which it follows that $T \propto \mathrm{e}^{-\lambda t}$. The second equation is

$$x(1-x)\frac{d^2X}{dx^2} + 2(1-2x)\frac{dX}{dx} - (2-4N\lambda)X = 0 \,,$$

which is the hypergeometric equation (Abramowitz, 1972, 15.5.1)

$$x(1-x)X'' + [\gamma - (\alpha+\beta+1)x]X' - \alpha\beta X = 0$$

with $\alpha = \frac{3+\sqrt{1+16N\lambda}}{2}$ and $\beta = \frac{3-\sqrt{1+16N\lambda}}{2}$. It is necessary to find a solution that is finite at the end points $x = 0$ and $x = 1$ (Kimura, 1964). Therefore the possible values for $\lambda$ are $\lambda_i = \frac{i(i+1)}{4N}$ and then $X_i \propto F(2+i, 1-i, 2, x)$. Thus $X_i$ can be written in terms of the Gegenbauer polynomials as

$$X_i = \frac{i(i+1)}{2} F(i+2, 1-i, 2, \frac{1-z}{2}) = C_i^{1.5}(z), \qquad z = 1-2x\,.$$

The solution can then be expressed as a linear combination of all possible values for $\lambda_i$ and $X_i$:

$$\phi(x|p,t) = \sum_{i=1}^{\infty} c_i C_i^{1.5}(z)\, \mathrm{e}^{-\frac{i(i+1)t}{4N}}\ . \tag{4.2}$$

The coefficients $c_i$ are found such that the initial condition is satisfied ($\phi(x|p,t) = \delta(x-p)$). The coefficients then are

$$c_i = 4p(1-p)\frac{2i+1}{i(i+1)} C_i^{1.5}(1-2p)$$

and the full solution is

$$\phi(x|p,t) = \sum_{i=1}^{\infty} 4p(1-p)\frac{2i+1}{i(i+1)} C_i^{1.5}(1-2p) C_i^{1.5}(1-2x)\, \mathrm{e}^{-\frac{i(i+1)t}{4N}} \tag{4.3}$$

(Kimura, 1964).

### 4.1.2 With Selection

In a bi-allelic model with selection the infinitesimal mean is $\mu(x) = sx(1-x)$ and the variance is $\sigma^2(x) = \frac{x(1-x)}{2N}$. Then the Kolmogorov forward equation is

$$\frac{\partial\phi(x|p,t)}{\partial t} = \frac{1}{4N}\frac{\partial^2}{\partial x^2}(x(1-x)\phi(x|p,t)) - s\frac{\partial}{\partial x}(x(1-x)\phi(x|p,t))\,. \tag{4.4}$$

The solution of the allele frequency density is assumed to be of the form

$$\phi \propto e^{2\gamma x} W(x) e^{-\lambda t}\ , \tag{4.5}$$

where $W(x)$ is a function of $x$ only and $\gamma = Ns$ (Kimura, 1955).

$$\begin{aligned}
\frac{\partial \phi}{\partial t} &= -\lambda e^{2\gamma x} W e^{-\lambda t} \\
\frac{\partial \phi}{\partial x} &= e^{2\gamma x} e^{-\lambda t}(2\gamma W + W') \\
\frac{\partial}{\partial x}\{x(1-x)\phi\} &= e^{2\gamma x} e^{-\lambda t}((1-2x)W + 2\gamma x(1-x)W + x(1-x)W') \\
\frac{\partial^2}{\partial x^2}\{x(1-x)\phi\} &= e^{2\gamma x} e^{-\lambda t}(W(4\gamma^2 x(1-x) - 2 + 4\gamma - 8\gamma x) + W'(4\gamma x(1-x) + 2(1-2x)) \\
&\quad + W'' x(1-x))
\end{aligned}$$

Insertion of these results into the forward equation (4.4) leads to

$$0 = [-4\gamma^2 x(1-x) - 2 + 4\lambda N]W + 2(1-2x)W' + x(1-x)W''\,. \tag{4.6}$$

Substituting $x = \frac{1-z}{2}$ in the above equation results in equation 4.7. From $0 < x < 1$ it follows that $-1 < z < 1$, such that

$$(1-z^2)W'' - 4zW' + ((4N\lambda - 2 - \gamma^2) + \gamma^2 z^2)W = 0\,. \tag{4.7}$$

Solutions to equations of this type have been studied by J. A. Stratton (1954) with the parameters $a = 1$ and $b = 4N\lambda - 2 - \gamma^2$. For the case where $\gamma = 0$ the equation reduces to the differential equation known for the Gegenbauer polynomials. For equations such as equation 4.7 with $\gamma^2 > 0$, J. A. Stratton (1954) proposes a solution of the form

$$W_i(z) = \sideset{}{'}\sum_{n=0,1} f_n^i C_{n+1}^{1.5}(z)\,, \tag{4.8}$$

where $i = 0, 1, 2, \ldots$ and $f_n^i$ are constants. $\sum_n'$ is a primed sum, which means that the summation is over even/odd values of $n$ if $k$ is even/odd (Crow and Kimura, 1970). The solution is obtained by expanding into a power series around the selection coefficient $\gamma$. For higher values of $\gamma$, results are not very accurate (Song and Steinrücken, 2012).

The solution of the forward equation with selection is $\phi(x|p,t) = \sum_{i=0}^{\infty} c_i e^{-\lambda_i t + 2cx} W_i(z)$ (Crow and Kimura, 1970). This solution is in the spectral representation, where $\lambda_i$ are eigenvalues and $W_i(z)$ are the corresponding eigenfunctions. The coefficients $c_i$ are calculated such that the initial condition ($\phi(x|p,0) = \delta(x-p)$) is satisfied (Kimura, 1964). The orthogonality relation of the eigenfunctions is

$$\int_{-1}^{1} (1-z^2) W_i(z) W_j(z) dz = \delta_{ij} \sideset{}{'}\sum_{n=0,1} (f_n^i)^2 \frac{(n+2)!}{n!(2n+3)}\,.$$

Then the coefficients are

$$c_i = \frac{(1-r^2)e^{-\gamma(1-r)}W_i(r)}{\sum'_{n=0,1}\frac{(n+1)(n+2)}{2n+3}(f_n^i)^2}\,, \tag{4.9}$$

where $r = 1 - 2p$ (Crow and Kimura, 1970). The primed summation is defined as above. The values for the coefficients $f_n^i$ and the eigenvalues $\lambda_i$ can be found in J. A. Stratton (1954).

Since the work of Kimura (1955) and J. A. Stratton (1954) many developments have been made in the field of spheroidal wave functions. Meixner and Schäfke (1954) and Flammer (1957) were most influential and used a different notation. There the definition of spheroidal wave functions is based on a linear combination of Legendre polynomials and not Gegenbauer polynomials.

## 4.2 Solution by Song and Steinrücken (2012)

Unlike Kimura's approach, the solution of Song and Steinrücken (2012) is not based on perturbation around the selection coefficient. Therefore the method is applicable also in cases with strong selection. Song and Steinrücken (2012) first refer to the Wright-Fisher model with the dominance parameter $h = \frac{1}{2}$ and later also to versions with a different dominance parameter $h \neq \frac{1}{2}$. In this chapter only the case with dominance parameter $h = \frac{1}{2}$ is considered.

Table 4.2: Notation Yun Song

| | |
|---|---|
| selection | $s$ |
| scaled selection | $\gamma = 2Ns$ |
| time | $t$ |
| time in generations | $\tau = 2Nt$ |
| infinitesimal variance | $\sigma^2(p) = p(1-p)$ |

### 4.2.1 Methods

The backward generator $\mathcal{L}_B$ of a one-dimensional diffusion process on $[0,1]$ is:

$$\mathcal{L}_B f(p) = \frac{1}{2}\sigma^2(p)\frac{\partial^2}{\partial p^2}f(p) + \mu(p)\frac{\partial}{\partial p}f(p) \tag{4.10}$$

where $f$ is a twice continuously differentiable bounded function over $[0,1]$ (Song and Steinrücken, 2012). Using the backward generator $\mathcal{L}_B$ on the allele frequency density function $\phi(x \,|\, p, \tau)$ leads to the Kolmogorov backward equation. With time rescaled to $\tau = 2Nt$ the infinitesimal variance is $\sigma^2(p) = p(1-p)$ and the mean is $\mu(p) = 2\gamma p(1-p)[p + h(1-2p)]$. In the case with no dominance, i.e. $h = \frac{1}{2}$, this reduces to $\mu(p) = \gamma p(1-p)$. Ewens (2000) uses the same notation as Song and Steinrücken (2012). This definition is also equivalent to the definition used by Crow and Kimura (1970) only with the different scaled time parameter $\tau = 2Nt$, as is shown by the following calculation.

$$\begin{aligned}
\frac{\partial\phi(x\,|\,p,t)}{\partial t} &= \frac{1}{4N}p(1-p)\frac{\partial^2}{\partial p^2}\phi(x\,|\,p,t) + sp(1-p)\frac{\partial}{\partial p}\phi(x\,|\,p,t) \\
2N\frac{\partial\phi(x\,|\,p,t)}{\partial t} &= \frac{1}{2}p(1-p)\frac{\partial^2}{\partial p^2}\phi(x\,|\,p,t) + 2Nsp(1-p)\frac{\partial}{\partial p}\phi(x\,|\,p,t) \\
&\Downarrow \tau = 2Nt \\
\frac{\partial\phi(x\,|\,p,\tau)}{\partial \tau} &= \frac{1}{2}p(1-p)\frac{\partial^2}{\partial p^2}\phi(x\,|\,p,\tau) + 2Nsp(1-p)\frac{\partial}{\partial p}\phi(x\,|\,p,\tau)
\end{aligned}$$

The generator of the backward equation is self-adjoint with respect to the speed function

$$m(p) = \frac{1}{\sigma^2(p)S(p)} \tag{4.11}$$

with the scale function $S(p)$

$$S(p) = e^{-\int_{p_0}^{p} \frac{2\mu(z)}{\sigma^2(z)}dz} \,. \tag{4.12}$$

Self adjointness means that $\langle \mathcal{L}f, g\rangle_m = \langle f, \mathcal{L}g\rangle_m$ for $f, g \in L^2([0,1],\rho)$, where $L^2([0,1],\rho)$ is the space of real-valued functions on $[0,1]$ (Song and Steinrücken, 2012). $X_\tau$ is a stochastic process on $[0,1]$ with the density function $\phi(x \,|\, p,\tau)$. Where $\tau \geq 0$ is the continuous time variable, $p$ the starting allele frequency at time $\tau = 0$ and $x$ the allele frequency at time $\tau$ and the random variable of the process.

The Kolmogorov backward equation, which is generated by $\mathcal{L}_B$, is satisfied by the density function $\phi(x \,|\, p,\tau)$.

The spectral representation of $\phi(x \,|\, p,\tau)$ is a linear combination of the eigenfunctions and eigenvectors of $\mathcal{L}_B$. Assume $B_n(p)$ is an eigenfunction of $\mathcal{L}_B$ and $-\lambda_n$ the according eigenvalue, then the function $e^{-\lambda_n\tau}B_n(p)$ satisfies the Kolmogorov backward equation generated by $\mathcal{L}_B$. Since $\mathcal{L}_B$ is a linear operator, also a linear combination of $e^{-\lambda_n\tau}B_n(p)$ is a solution. The coefficients for the linear combination $c_n(x)$ in the spectral representation in equation 4.13 are chosen such that the initial condition is satisfied ($\phi(x \,|\, p, 0) = \delta(x-p)$).

$$\phi(x \,|\, p,\tau) = \sum_{n=0}^{\infty} c_n(x)e^{-\lambda_n\tau}B_n(p) \tag{4.13}$$

Since $\lambda_n \to \infty$ as $n \to \infty$, it is enough to sum until some reasonable high $n$, this is discussed in chapter 5 in more detail. The *initial condition* states that $x = p$ at time $t = 0$, which means that $\phi(0,p,x) = \delta(x-p)$. The coefficients satisfying this starting condition are:

$$c_n(x) = \frac{m(x)B_n(x)}{\langle B_n, B_n\rangle_m} \,. \tag{4.14}$$

To find the solution to the eigenvalue function, orthogonal polynomials are used. For this purpose Song and Steinrücken (2012) define modified Gegenbauer polynomials, which are denoted by $G_n(x)$ and are defined over $0 < x < 1$. The classical Gegenbauer polynomials $C_n^\alpha(z)$ are defined over $-1 < z < 1$.

$$G_n(x) = -x(1-x)P_n^{1,1}(2x-1) = -x(1-x)R_n^{(2,2)}(x) \,. \tag{4.15}$$

Here $P_n^{(a,b)}(x)$ are the classical Jacobi polynomials and $R_n^{(a,b)}(x)$ are modified Jacobi polynomials (Song and Steinrücken, 2012). In Abramowitz (1972, 22.5.20) a relationship between the Jacobi polynomials and the Gegenbauer polynomials ($C_n^\alpha(x)$) is given. For $\alpha = 1.5$ the equation becomes

$$P_n^{(1,1)}(x) = \frac{2}{n+2}C_n^{1.5}(x) \tag{4.16}$$

and the relationship between the modified and classical Gegenbauer polynomials can easily be obtained (see equation 4.17).

$$G_n(x) = -x(1-x)\frac{2}{n+2}C_n^{1.5}(2x-1)\,. \tag{4.17}$$

From the differential equation satisfied by the Gegenbauer polynomials a differential equation satisfied by the modified Gegenbauer polynomials can be obtained (see equation 4.18).

$$x(1-x)\frac{d^2G_n(x)}{dx^2} + (n+2)(n+1)G_n(x) = 0\,. \tag{4.18}$$

### 4.2.2 Diffusions with Genic Selection and No Mutation

Only the solution in the case without mutation, with selection, but without dominance is studied in more detail here. Without dominance $h = \frac{1}{2}$ and without selection the backward equation has the generator $\mathcal{L}_0$. The modified Gegenbauer polynomials $G_n(p)$ are eigenfunctions of this generator

$$\mathcal{L}_0 f(p) = \frac{1}{2}p(1-p)\frac{\partial^2 f(p)}{\partial p^2}\,. \tag{4.19}$$

and they satisfy the differential equation in equation 4.18. Using the backward generator on $G_n(p)$ leads to the following equation

$$\mathcal{L}_0 G_n(p) = \frac{1}{2}p(1-p)\frac{\partial^2}{\partial p^2}G_n(p) = -\frac{1}{2}(n+2)(n+1)G_n(p)\,. \tag{4.20}$$

and the corresponding eigenvalues are $\lambda_n = \frac{1}{2}(n+2)(n+1)$.
With genic selection ($\gamma = 2Ns$) the backward generator with $\sigma^2(p) = p(1-p)$ and $\mu(p) = \gamma p(1-p)$ is $\mathcal{L}_B$.

$$\mathcal{L}_B f(p) = \frac{1}{2}p(1-p)\frac{\partial^2}{\partial p^2}f(p) + \gamma p(1-p)\frac{\partial}{\partial p}f(p)\,. \tag{4.21}$$

Substituting the infinitesimal mean and variance into the equations for the scale function $S(p)$ and the speed function $m(p)$ leads to

$$m(p) = \frac{e^{2\gamma p}}{p(1-p)}\,. \tag{4.22}$$

This leads to a solution for the eigenfunctions and associated eigenvalues of the diffusion part of the backward generator $\mathcal{L}_0$. As shown before the backward generator $\mathcal{L}_B$ is self-adjoint with respect to $m(.)$. The eigenfunctions of $\mathcal{L}_B$ are $B_n(p)$ and the associated eigenvalues are $\lambda_n$. They satisfy the following equations:

$$\mathcal{L}_B B_n(p) = -\lambda_n B_n(p) \tag{4.23}$$

and

$$\int_0^1 B_n(p)B_m(p)m(p)dp \propto \delta_{n,m}\,. \tag{4.24}$$

Song and Steinrücken (2012) next consider the functions $H_n(p)$, which are orthogonal with respect to the same weight function $m(.)$:

$$H_n(p) = e^{-\gamma p} G_n(p)\,. \tag{4.25}$$

$H_n(p)$ are not eigenfunctions of $\mathcal{L}_B$, but $B_n(p)$ can be expressed as a linear combination of $H_n(p)$:

$$B_n(p) = \sum_{m=0}^{\infty} u_{n,m} H_m(p)\,. \tag{4.26}$$

The backward generator used on $H_n(p)$ leads to the equation:

$$\mathcal{L}_B H_n(p) = -e^{-\gamma p}[\lambda_n G_n(p) + \frac{1}{2}\gamma^2 p(1-p) G_n(p)] \tag{4.27}$$

and

$$\begin{aligned}
\mathcal{L}_B B_n(p) &= \sum_{m=0}^{\infty} u_{n,m} \mathcal{L} H_m(p) && (4.28)\\
&= -\sum_{m=0}^{\infty} u_{n,m} e^{-\gamma p}[\lambda_n + \frac{1}{2}\gamma^2 p(1-p)] G_n(p) && (4.29)\\
&= \sum_{m=0}^{\infty} u_{n,m}[\lambda_n + \frac{1}{2}\gamma^2 p(1-p)] G_n(p)\,. && (4.30)
\end{aligned}$$

The eigenfunctions $B_n(p)$ are also called the *oblate spheroidal wave functions*, in the case of selection and no recurrent mutation.

Because the $B_n(p)$ are eigenfunctions of $\mathcal{L}_B$ and $\lambda_n$ are the associated eigenvalues, we have:

$$\sum_{m=0}^{\infty} v_{n,m}[\lambda_n + \frac{1}{2}\gamma^2 p(1-p)] G_n(p) = \lambda_n \sum_{m=0}^{\infty} u_{n,m} G_m(p)\,. \tag{4.31}$$

Then it can be shown that

$$\frac{1}{2}\gamma^2 p(1-p) G_m(p) = a_m^{(-2)} G_{m-2}(p) + a_m^{(0)} G_m(p) + a_m^{(+2)} G_{m+2}(p) \tag{4.32}$$

with

$$\begin{aligned}
a_m^{(-2)} &= \gamma^2 \frac{1}{8} \frac{m(m+1)}{(2m+1)(2m+3)} \text{ where } m \geq 2 && (4.33)\\
a_m^{(0)} &= \gamma^2 \frac{1}{4} \frac{(m+1)(m+2)}{(2m+1)(2m+5)} && (4.34)\\
a_m^{(+2)} &= -\gamma^2 \frac{1}{8} \frac{(m+1)(m+4)}{2m+3)(2m+5)} && (4.35)\\
& && (4.36)
\end{aligned}$$

and

$$\lambda_k u_{n,k} + a_{k+2}^{(-2)} u_{n,k+2} + a_{k-2}^{(+2)} u_{n,k-2} = \Lambda_n u_{n,k}\,. \tag{4.37}$$

**Algorithm**

Song and Steinrücken (2012) propose an algorithm for the calculation of the eigenvalues and eigenfunction of $\mathcal{L}_B$. First the equation 4.37 can be rewritten in matrix form:

$$\begin{pmatrix} \lambda_0 + a_0^{(0)} & 0 & a_2^{(-2)} & 0 & 0 & \cdots & \\ 0 & \lambda_1 + a_1^{(0)} & 0 & a_3^{(-2)} & 0 & \cdots & \\ a_0^{(+2)} & 0 & \lambda_2 + a_2^{(0)} & 0 & a_4^{(-2)} & \cdots & \\ 0 & a_1^{(+2)} & 0 & \lambda_3 + a_3^{(0)} & 0 & \cdots & \\ 0 & 0 & a_2^{(+2)} & 0 & \lambda_4 + a_4^{(0)} & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots \end{pmatrix} \begin{pmatrix} u_{n,0} \\ u_{n,1} \\ u_{n,2} \\ u_{n,3} \\ u_{n,4} \\ \vdots \end{pmatrix} = \Lambda_n \begin{pmatrix} u_{n,0} \\ u_{n,1} \\ u_{n,2} \\ u_{n,3} \\ u_{n,4} \\ \vdots \end{pmatrix} \tag{4.38}$$

such that:

$$M u_n = \Lambda_n u_n \, . \tag{4.39}$$

Let $M^{[D]}$ be the upper left part of the matrix $M$ with $D$ rows and $D$ columns. Song and Steinrücken (2012) show empirically that the eigenvalues and eigenfunctions of $M^{[D]}$ converge quickly. The equation $M^{[D]} u_n^{[D]} = \Lambda_n^{[D]} u_n^{[D]}$ can be separated into two systems, one only consisting of the odd rows and columns and one system containing only the even rows and columns. This follows directly from the structure of the matrix $M$.

The form of the matrix above corresponds to the primed sums in the solution by Crow and Kimura (1970) and J. A. Stratton (1954).

## 4.3 New Solution Method with Spheroidal Wave Functions

In this section a solution to the Kolmogorov forward equation is presented using the spheroidal wave functions. This is achieved by transforming the Kolmogorov forward equation into a Sturm-Liouville form.

Table 4.3: Notation

| | |
|---|---|
| selection | $s$ |
| scaled selection | $\gamma = Ns$ |
| time | $t$ |
| time | $\tau = 4Nt$ |

The Kolmogorov forward equation is

$$\frac{\delta\phi(x|p,t)}{\delta t} = \frac{1}{4N}\frac{\delta^2}{\delta x^2}\{x(1-x)\phi(x|p,t)\} - s\frac{\delta}{\delta x}\{x(1-x)\phi(x|p,t)\}\,. \tag{4.40}$$

With rescaling of time to $\tau$ such that $\tau = 4Nt$ the rescaled equation in equation 4.41 is obtained.

$$\frac{\delta\phi(x|p,\tau)}{\delta\tau} = \frac{\delta^2}{\delta x^2}\left(x(1-x)\phi(x|p,\tau)\right) - 4\gamma\frac{\delta}{\delta x}\left(x(1-x)\phi(x|p,\tau)\right)\,. \tag{4.41}$$

For this purpose we used $\tau = 4Nt$ and not $2Nt$ like Song and Steinrücken (2012) did, also the scaled selection $\gamma$ differs by a factor of two. The density is assumed to be of the form $\phi(x|p,\tau) \propto e^{2\gamma x}v(x)e^{-\lambda\tau}$. Inserting this into the scaled forward equation (4.41) leads to:

$$x(1-x)\frac{d^2v(x)}{dx^2} + 2(1-2x)\frac{dv(x)}{dx} - \left(2 + 4\gamma^2 x(1-x) - \lambda\right)v(x) = 0\,. \tag{4.42}$$

By substituting $x = \frac{1-z}{2}$ we obtain:

$$(1-z^2)\frac{d^2v(\frac{1-z}{2})}{dz^2} - 4z\frac{dv(\frac{1-z}{2}))}{dz} + \left(\lambda - 2 - \gamma^2(1-z^2)\right)v(\frac{1-z}{2}) = 0\,. \tag{4.43}$$

This equation can be transformed to a Sturm-Liouville form. This can be achieved by setting $g(z)(1-z^2)^{-\frac{1}{2}} = v(\frac{1-z}{2})$ and substituting this in equation 4.43 leads to the following calculation;

$$\begin{aligned}
0 &= (1-z^2)\frac{d^2(1-z^2)^{-1/2}g(z)}{dz^2} - 4z\frac{d(1-z^2)^{-1/2}g(z)}{dz} \\
&\quad + \left(\lambda - 2 - \gamma^2(1-z^2)\right)(1-z^2)^{-1/2}g(z) \\
0 &= (1-z^2)\frac{d^2g(z)}{dz^2} + 2z\frac{dg(z)}{d}z + (1+3z^2(1-z^2)^{-1})g(z) \\
&\quad - 4z\frac{dg(z)}{dz} - 4z^2(1-z^2)^{-1}g(z) + \left(\lambda - 2 - \gamma^2(1-z^2)\right)g(z) \\
0 &= \frac{d}{dz}\left((1-z^2)\frac{dg(z)}{dz}\right) + \left(\lambda - \gamma^2(1-z^2) - \frac{1}{1-z^2}\right)g(z)\,. \qquad (4.44)
\end{aligned}$$

### 4.3.1 In the Model with Selection

With selection, i.e. $\gamma^2 > 0$, the differential equation 4.43 is in a form that can be transformed to the differential equation solved by *angular oblate speroidal wave functions* (Abramowitz, 1972, 21.6.4). Note that the index $m$ of the spheroidal wave equations can be omitted since $m = 1$ always.

$$\frac{d}{dz}[(1-z^2)\frac{d}{dz}S_n(c,z)] + (\lambda_n^S + c^2z^2 - \frac{1}{1-z^2})S_n(c,z) = 0 \qquad (4.45)$$

$\mathcal{L}_Z$ is the generator of the differential equation in equation 4.44 and $\mathcal{L}_S$ the generator of the differential equation solved by the angular oblate spheroidal wave functions in equation 4.45. The eigenfunctions of $\mathcal{L}_S$ are the angular oblate spheroidal wave functions $S_n(z)$ with the associated eigenvalues $\lambda_n^S$.

$$\mathcal{L}_Z f(z) = \frac{d}{dz}\left((1-z^2)\frac{df(z)}{dz}\right) + \left(-\gamma^2 + \gamma^2z^2 - \frac{1}{1-z^2}\right)f(z) \qquad (4.46)$$

$$\mathcal{L}_S f(z) = \frac{d}{dz}\left((1-z^2)\frac{df(z)}{dz}\right) + \left(\gamma^2z^2 - \frac{1}{1-z^2}\right)f(z)\,, \qquad (4.47)$$

The eigenvalues and eigenvectors for $\mathcal{L}_Z$ and $\mathcal{L}_S$ then are

$$\begin{aligned}
\mathcal{L}_S S_n &= \lambda_n^S S_n \\
\mathcal{L}_Z S_n &= (\lambda_n^S + \gamma^2)S_n\,.
\end{aligned} \qquad (4.48)$$

From this, it follows that the eigenfunctions are identical and the eigenvalues differ by $\gamma^2$. The solution can therefore be expressed in the spectral respresentation by the eigenfunctions $S_n(z)$ and the associated eigenvalues $\lambda_n = \lambda_n^S + \gamma^2$.

The angular oblate Spheroidal wave functions are normalised by the Meixner-Schäfke scheme as (Abramowitz, 1972; Meixner and Schäfke, 1954)

$$\int_{-1}^{1}[S_n(c,z)]^2dz = \frac{2n(n+1)}{2n+1}\,. \qquad (4.49)$$

Now only the transformation from $g(z)$ to $\phi(x|p,t)$ is needed. This is obtained by using the relations $\phi(x) = e^{2\gamma x} g(z)(1-z^2)^{-\frac{1}{2}} e^{-\lambda\tau}$, $x = \frac{1-z}{2}$ and $g(z)(1-z^2)^{-\frac{1}{2}} = v(\frac{1-z}{2})$

$$\phi(x|p,\tau) \propto g(z)\frac{1}{2}(x(1-x))^{-\frac{1}{2}} e^{2\gamma x} e^{-\lambda\tau}\,. \tag{4.50}$$

Then the solution to the allele frequency density is

$$\begin{aligned}\phi(x|p,\tau) = \sum_{n=0}^{\infty} & S_n(1-2p) S_n(1-2x)\frac{1}{2}(x(1-x))^{-\frac{1}{2}} \\ & e^{2\gamma x} e^{-\lambda_n \tau}\frac{2n+1}{2n(n+1)}\end{aligned} \tag{4.51}$$

### 4.3.2 Model without Selection

When there is no selection $\gamma^2 = 0$ and equation (4.44) reduces to the differential equation solved by the Legendre polynomials.

$$0 = \frac{d}{dz}\left((1-z^2)\frac{dg(z)}{dz}\right) + \left(l(l+1) - \frac{m^2}{1-z^2}\right) g(z)\,. \tag{4.52}$$

The Legendre polynomials are orthogonal polynomials that satisfy $\int_{-1}^{1} P_n(z)^2 dz = \frac{2}{2n+1}$. The transformation from $g(z)$ back to $\phi(x|p,t)$ has to be taken into account as before

$$\phi(x|p,\tau) \propto g(z)\frac{1}{2}(x(1-x))^{-\frac{1}{2}} e^{-\lambda\tau}\,, \tag{4.53}$$

and the solution can be written as

$$\phi(x|p,\tau) = \sum_{i=0}^{\infty} \frac{2i+1}{2} P_i(1-2p) P_i(1-2x)\frac{(x(1-x))^{\frac{1}{2}}}{2}\,\mathrm{e}^{-i(i+1)\tau}\,. \tag{4.54}$$

## 4.4 Calculation

The spheroidal wave functions defined by Meixner and Schäfke (1954) and Flammer (1957) and their eigenvalues are implemented in many software packages, e.g. in Mathematica (Weisstein, 2013c) and in the Mathematica package *spheroidal* by Falloon (2003) and are thus easily available. In the appendix the functions implemented in Mathematica are used for simulation. The oblate spheroidal wave functions are implemented in Mathematica as the function *SpheroidalPS* with parameter $-i\gamma$, with $i = \sqrt{-1}$. Calculation of the functions and their eigenvalues is still an issue (see Kirby (2006)), although the difficulties arise mainly in the calculation of radial functions and for high values of $\gamma$. Here only angular functions are used, but for higher values of $\gamma$ difficulties may still arise. Eigenvalues can be calculated to high precision using a series expansion around $\gamma = 0$; the formulas derived in this way are very accurate for small values of $\gamma$; better than parts per million (Sullivan and Thompson, 1999). The eigenvalues can be calculated exactly by using a continued fraction, which is computationally expensive (Sullivan and Thompson, 1999).

# Chapter 5

# Applications

In this chapter the solution of Kimura (1955) is compared to the solution in section 4.3 using graphs. The time parameter is always scaled to $\tau = 4Nt$ in all results in this chapter. The analysis is made using the software *Wolfram Mathematica 9*. The code used for creating the figures is presented in Appendix B.

## 5.1 Allele Frequency Density without Selection

### 5.1.1 Gegenbauer Polynomials

The solution of the allele frequency density is an infinite sum. Truncating the sum causes the density to be very inaccurate close to $\tau = 0$. For $\tau > 0$, the coefficient $e^{-i(i+1)\tau}$ gets very small very fast, so for the computation of the function a minimum value for $e^{-i(i+1)\tau}$ is set, such that the computation stops as soon as this value is exceeded. Therefore the polynomials of a higher degree are only used for very small values of $\tau$. In figure 5.1 the number of polynomials is shown for different minimum values for $d = e^{-i(i+1)\tau}$.

The Gegenbauer Polynomials for some different values of $n$ can be seen in figures 5.2 and 5.3.

The Gegenbauer polynomials are multiplied by the term $C_n^{1.5}(1-2p)$, which is constant since $p$ is the starting allele frequency. In figure 5.4 it can be seen that the sign of this term changes depending on $n$. The value also increases, but this does not have too much influence since $e^{-i(i+1)/2}$ decreases much faster.

The resulting allele frequency density without selection by Kimura (1955) is shown in figure 5.5. Since there is no mutation the boundaries 0 and 1 are absorbing. The area under the allele frequency density curve in the interval $]0,1[$ can be interpreted as the probability of both alleles being present in the population at time $t$. The probability

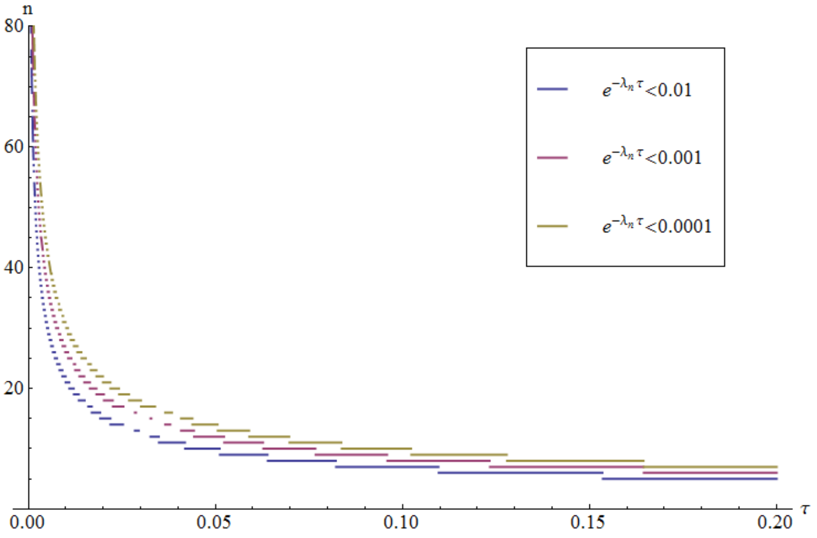

Figure 5.1: This graph shows the number of summands needed to achieve a given accuracy. As measure for accuracy the expression $e^{-\lambda_n \tau}$ is used. As soon as $e^{-\lambda_n \tau}$ is smaller than some value (i.e. 0.01, 0.001 and 0.0001) the sum is truncated. For example when $\tau = 0.05$ summation until $n = 10$ is needed to achieve that $e^{-\lambda_n 0.05} < 0.01$, $n = 12$ for $e^{-\lambda_n 0.05} < 0.001$ and $n = 14$ for $e^{-\lambda_n 0.05} < 0.0001$. At $\tau = 0.001$ the difference is much bigger; $n = 68$ is needed for $e^{-\lambda_n 0.001} < 0.01$, $n = 83$ for $e^{-\lambda_n 0.001} < 0.001$ and $n = 96$ for $e^{-\lambda_n 0.001} < 0.0001$.

mass at the boundaries is increasing by $-\frac{\partial}{\partial t}\int_0^1 \phi(x|p,t)dx$.

In Kimura (1955) it is shown that the increase in the probability mass absorbed at the boundaries is proportional to the amount present there:

$$-\frac{\partial}{\partial t}\int_0^1 \phi(x|p,t)dx = \frac{\phi(1|p,t) + \phi(0|p,t)}{4N}\,. \tag{5.1}$$

The coefficients in the solution are chosen such that there is at time 0 a point mass at $p$. To reach this point mass exactly it would be necessary to sum up all values of the infinite sum. Here only the first linear combinations of Gegenbauer polynomials are used until $e^{-i(i+1)\tau}$ is smaller then 0.0001 . Figure 5.6 is calculated using the Mathematica function *Integrate* on the allele frequency density. Since at time 0 a point mass is approximated the result of the integration is far from 1.

The change over time in this area under the curve is shown in figure 5.6 and the differences are shown in figure 5.7. As calculated by Kimura it can be seen, that the area is decaying faster with time.

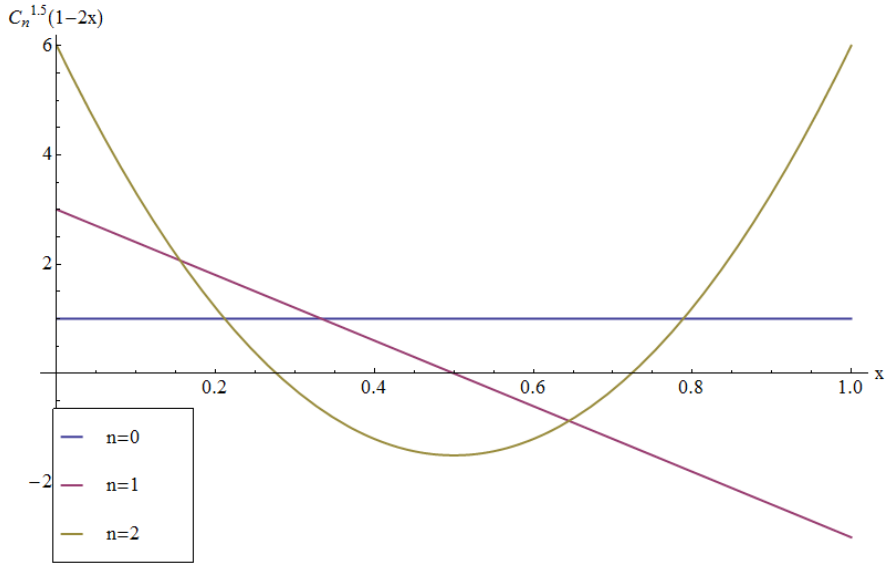

Figure 5.2: Gegenbauer Polynomials $C_n^{1.5}(1-2x)$ for different values of $n$

### 5.1.2 Legendre Polynomials

In the solution in section 4.3, the allele frequency density is expressed by the Legendre polynomials. The resulting allele frequency density can be seen in figure 5.8. The Legendre Polynomials $P_n(1-2x)$ are shown in figures 5.9 and 5.10 for different values of $n$. For choosing when to truncate the infinite sum, the same strategy as with the Gegenbauer Polynomials is used.

## 5.2 Allele Frequency Density with Selection

The density of allele frequencies in the model with selection is expressed as a linear combination of spheroidal wave functions. In Mathematica the angular oblate spheroidal wave functions are implemented in *SpheroidalPS* and *SpheroidalQS* with the parameter $-i\gamma$ ($i = \sqrt{-1}$). The spheroidal wave functions in *SpheroidalPS* are linear combinations of the Legendre polynomials of the first kind; in *SpheroidalQS* the Legendre polynomials of the second kind are used (Weisstein, 2013c). Since the Legendre polynomials of the second kind are singular at the boundaries $-1$ and 1, the spheroidal wave functions of the first kind are needed. The spheroidal wave functions can be seen in figures 5.11 and 5.12 for different values of $n$.

The allele frequency density for a selection coefficient $\gamma = 0.1$ can be see in figure 5.13. The density when the selection coefficient is stronger with $\gamma = 3$ is shown in figure 5.14 and for weak selection $\gamma = 1$ in figure 5.15.

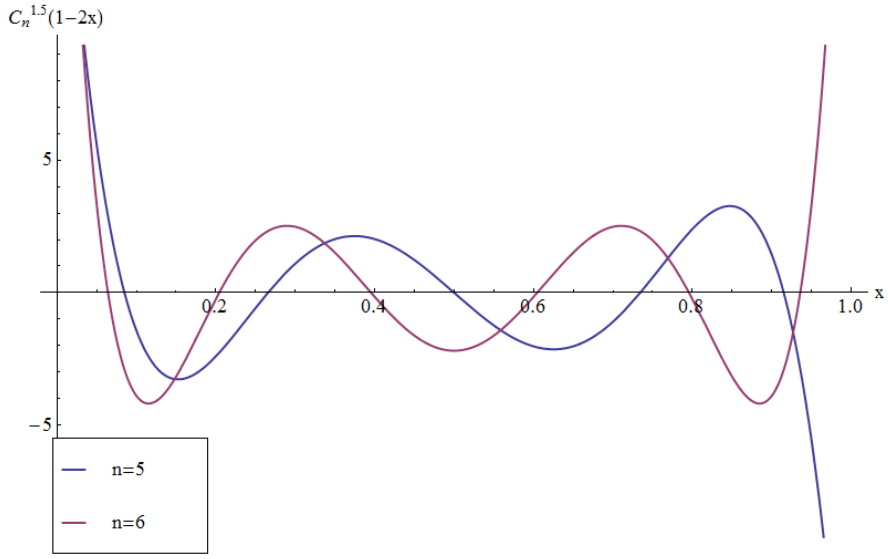

Figure 5.3: Gegenbauer Polynomials $C_n^{1.5}(1-2x)$ for different values of $n$

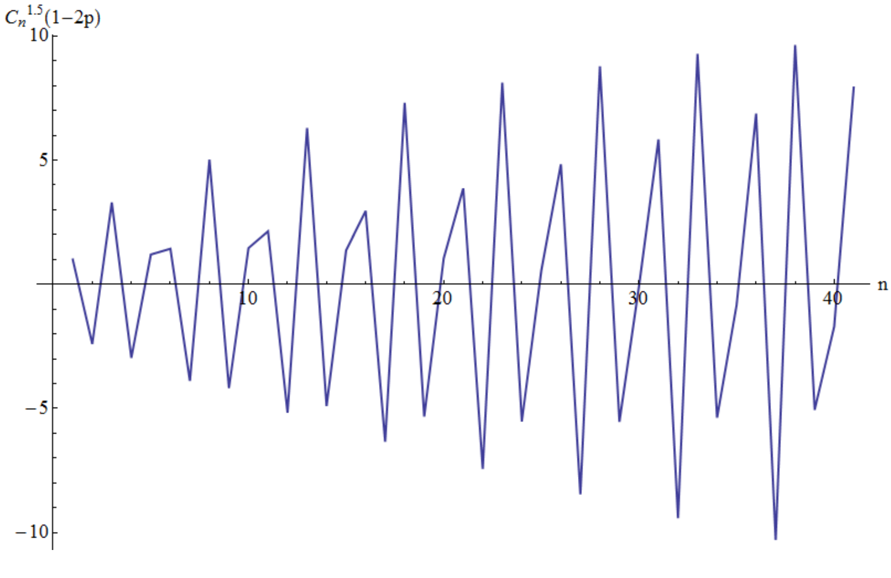

Figure 5.4: Gegenbauer Polynomials $C_n^{1.5}(1-2p)$ with $p = 0.4$ with $n$ on the x-axis. The characteristic jumping of the function from positive to negative values can be observed.

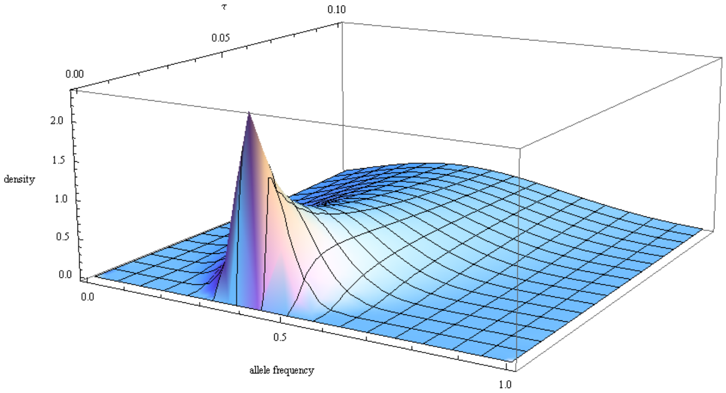

Figure 5.5: Visualisation of the allele frequency density solution by Kimura in three dimensions; the time $\tau$, the density and the allele frequency $x$. The starting allele frequency at time $\tau = 0$ is $p = 0.4$. Where $\tau = 0$ a delta distribution at 0.4 is approximated and decays quickly with time. This is the density without selection $\gamma = 0$.

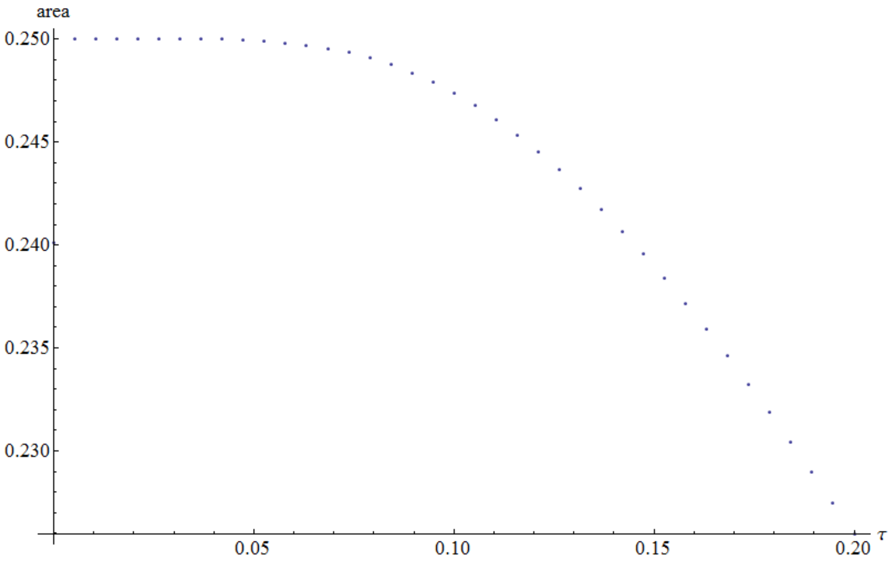

Figure 5.6: The allele frequency density decays quickly with time, as it can be seen in figure 5.5. The area under the curve at each time can be interpreted as the probability of the two alleles existing at the same time. The integral is shown in this figure.

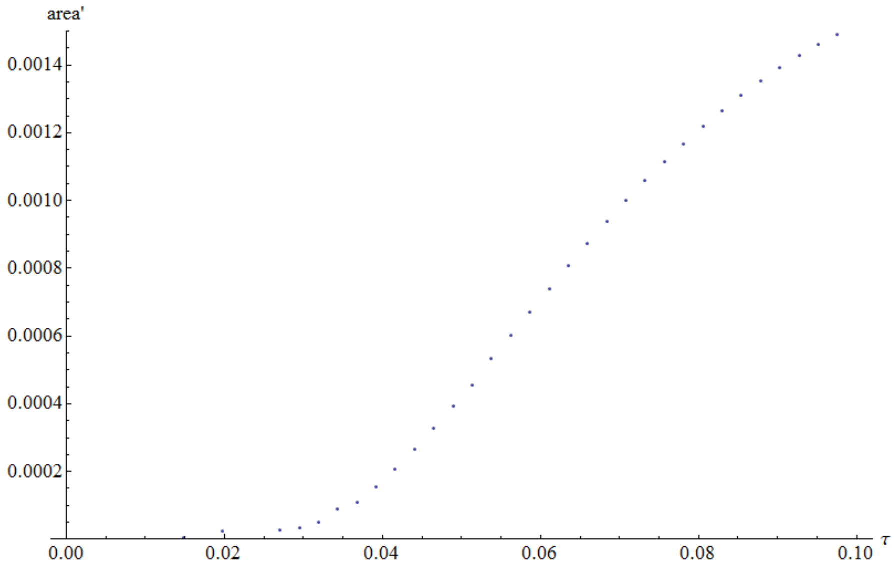

Figure 5.7: The area under the allele frequency density function at each time can be interpreted as the probability of the two alleles existing at the same time. The integral is shown in figure 5.6. In this figure the change in the area is shown. As already shown by Kimura (1955) the loss of probability mass increases with time.

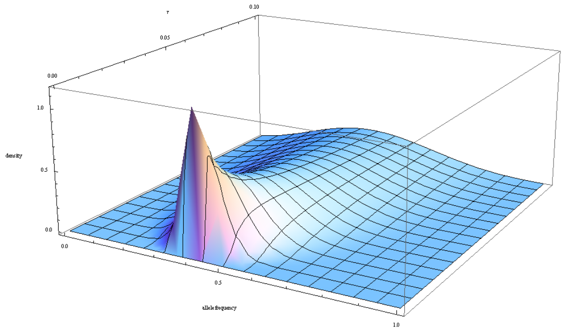

Figure 5.8: The allele frequency density by the solution in section 4.3 using the Legendre polynomials, without selection ($\gamma = 0$).

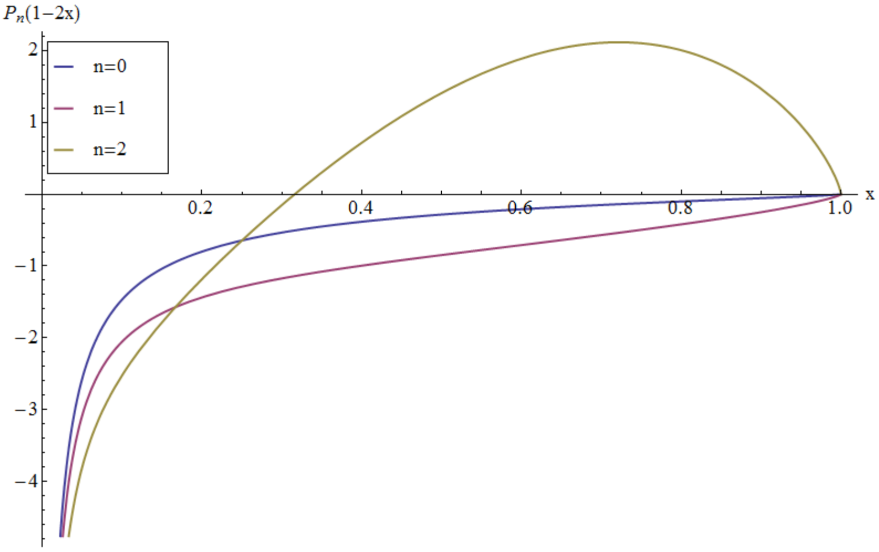

Figure 5.9: Legendre Polynomials $P_n(1 - 2x)$ for different values of $n$

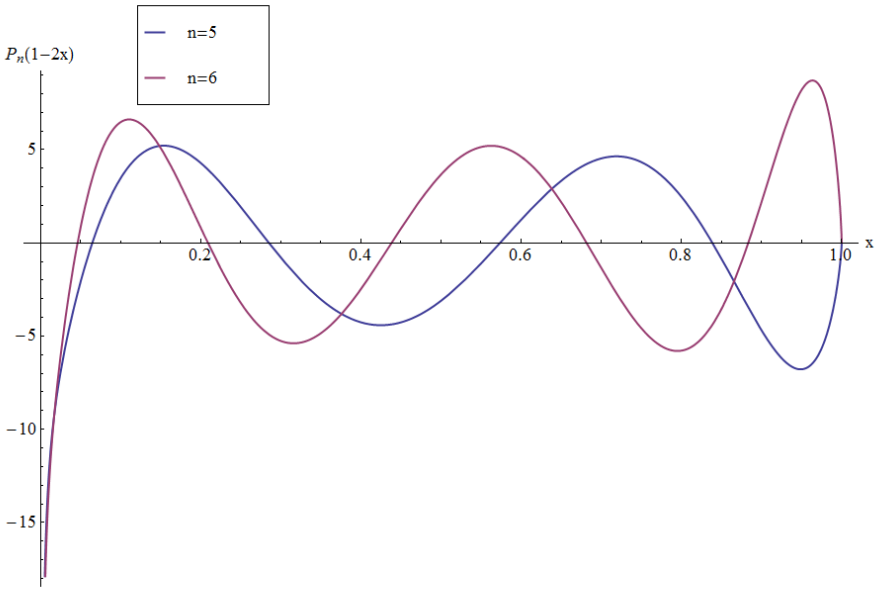

Figure 5.10: Legendre Polynomials $P_n(1 - 2x)$ for different values of $n$

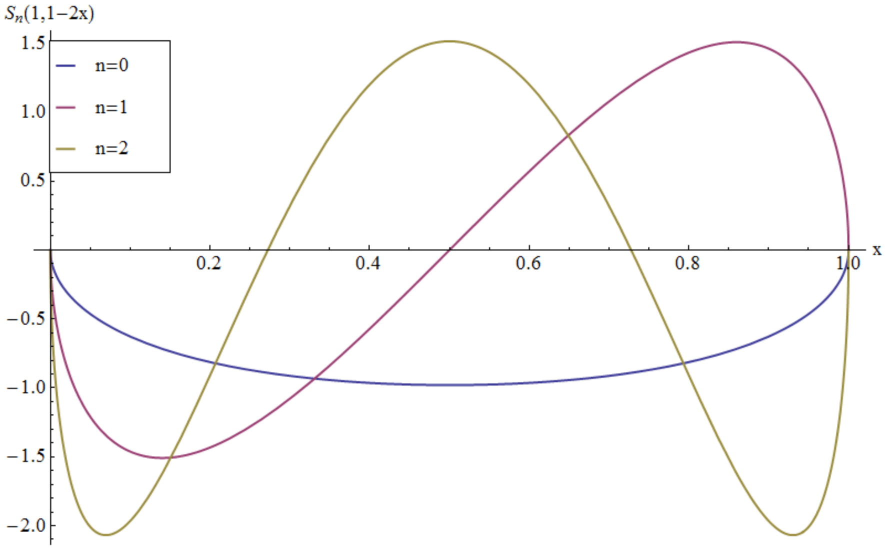

Figure 5.11: angular oblate spheroidal wave functions with different values of $n$

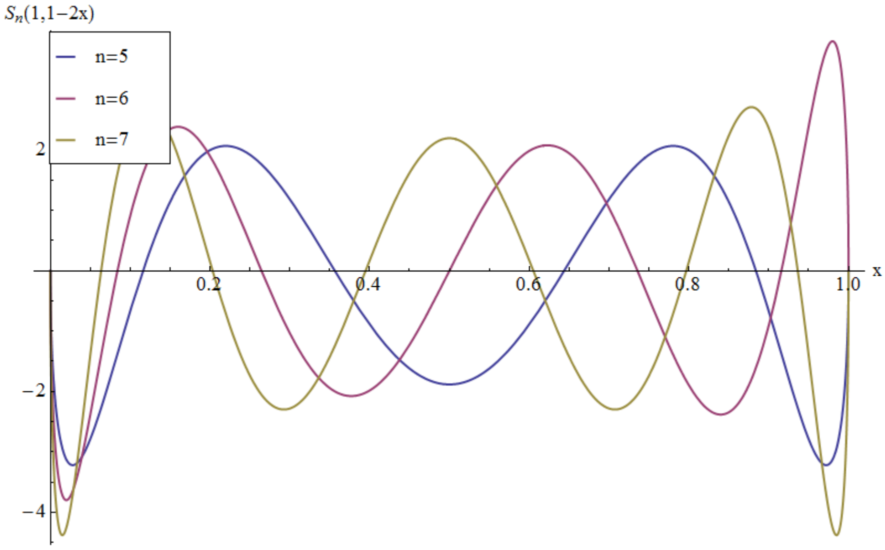

Figure 5.12: angular oblate spheroidal wave functions with different values of $n$

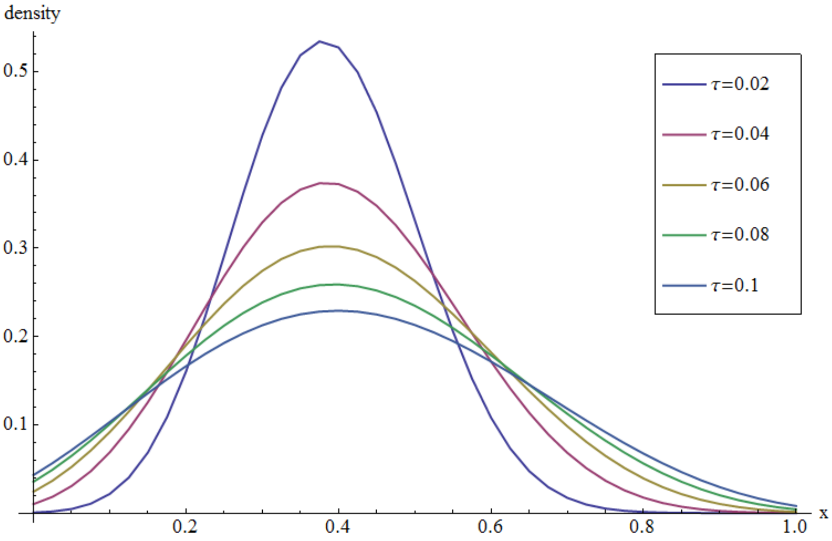

Figure 5.13: Allele frequency density with selection coefficient $\gamma = 0.1$ at different time points $\tau$.

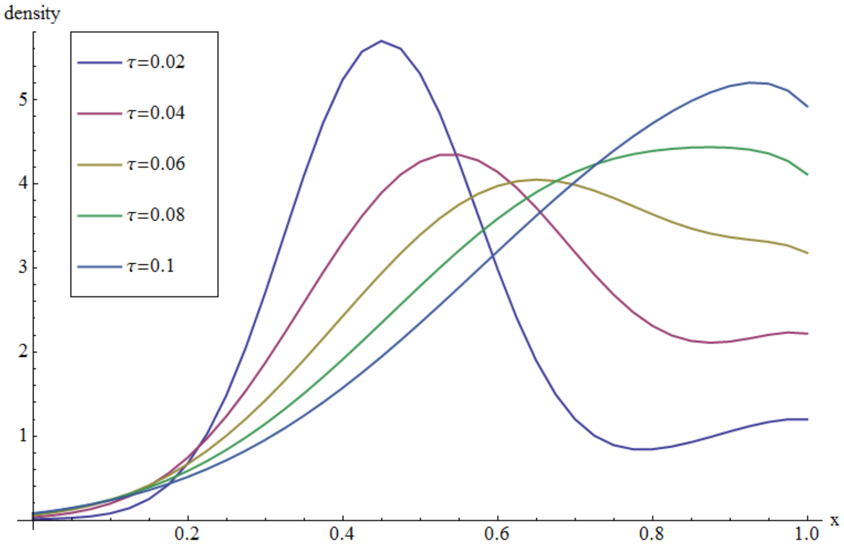

Figure 5.14: Allele frequency density with selection coefficient $\gamma = 3$ at different time points $\tau$.

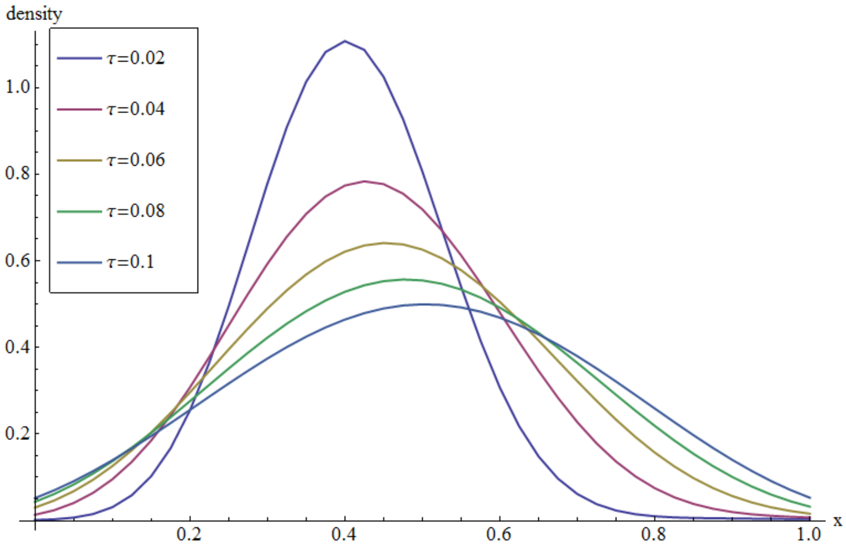

Figure 5.15: Allele frequency density with selection coefficient $\gamma = 1$ at different time points $\tau$.

# Chapter 6

# Summary and Discussion

In this thesis the methods of Kimura (1955), Song and Steinrücken (2012) and a further approach for obtaining the allele frequency density are presented. Kimura (1955) was the first to present a solution to this problem. His solution in the case with selection is based on the oblate spheroidal wave functions defined by J. A. Stratton (1941) and J. A. Stratton (1954). The eigenvalues and eigenfunctions are found by Kimura (1955) by expanding into a Taylor series around the selection coefficient $\gamma$. Hence, the solution by Kimura is inaccurate for high values of $\gamma$ (Song and Steinrücken, 2012). The solution by Song and Steinrücken (2012) also uses the definition for spheroidal wave functions by J. A. Stratton (1954), but the eigenvalues are found by solving a finite dimensional matrix problem. The computational possibilities have of course increased significantly since Kimura (1955) and so Song and Steinrücken (2012) were able to present an algorithm for the eigenvalue problem, implemented in the programming language `C`. At the time as J. A. Stratton (1954) published updated tables for the spheroidal wave functions, the spheroidal wave functions got more attention, because of their applications in physics and quantum mechanics. As a result there are many different notations for spheroidal wave functions. In Abramowitz (1972) an overview of different notations is given. The notations by Meixner and Schäfke (1954) and Flammer (1957) are most common by now. The spheroidal wave functions defined by Meixner and Schäfke (1954) and Flammer (1957) and their eigenvalues are implemented in many software packages (e.g. in Mathematica (Weisstein, 2013c) and in the Mathematica package *spheroidal* by Falloon (2003)) and are used in the chapter on applications. The spheroidal wave functions used by Kimura (1955), Song and Steinrücken (2012) and J. A. Stratton (1954) are given as a linear combination of Gegenbauer polynomials, but the oblate angular spheroidal wave functions are more commonly expressed as linear combinations of Legendre Polynomials.

In section 4.3 it is shown that the foward equation can be transformed such that the spheroidal wave functions defined by Meixner and Schäfke (1954) and Flammer (1957) can be used. Since those functions are implemented in many software packages, these methods are now easily available to population geneticists.

# Appendix A

# Mathematical Background

An overview and short summary of the needed mathematical background for this thesis is given here.

## A.1 Markov Process

A markov process is a stochastic process that satisfies the markov property. The *markov property* states, that the future path is only dependend on the current state of the process and not on the preceding path. In other words, the markov process is *memory-less* (Norris, 1997). $(\Omega, \mathcal{F}, P)$ is the probability space with filtration $(\mathcal{F}_t, t > 0)$. Then the Markov property can be formulated as

$$P(X_t \in A|\mathcal{F}_s) = P(X_t \in A|X_s) \tag{A.1}$$

with $s < t$ and $A$ a subset of the measurable space of $(X_t)_{t\geq 0}$ (Norris, 1997). The *strong markov property* states that the process $(X_s)_{s>\tau}$ is independent of $\mathcal{F}_\tau$ and that $(X_{\tau+t} - X_\tau)_{t\geq 0}$ is equal in distribution to $(X_t)_{t\geq 0}$.

A *standard Markov process* is a strong Markov process with the following properties (Norris, 1997):

- $(X_t)$ is right continuous

$$lim_{t\downarrow s} X_t = X_s, \qquad \forall s \geq 0$$

- the left limits of $(X_t)$ exist

$$lim_{t\uparrow s} X_t \text{ exist} \qquad \forall s > 0$$

- $(X_t)$ is quasi-left continuous

$$lim_{n\to\infty} X_{T_n} = X_T \qquad \text{for } T_1 \leq T_2 \leq \cdots \leq T < \infty$$

$T_n$ are Markov times (Norris, 1997).

## A.2 Orthogonal Polynomials

A system of polynomials $F_n(x)$ is called orthogonal on the interval $a \leq 0 < b$ if

$$< F_n, F_m >_w = \int_a^b w(x) F_n(x) F_m(x) dx = 0 \, n \neq m; n, m = 0, 1, 2, \ldots$$

$w(x)$ is the weight function and $n$ is the degree of the polynomial (Abramowitz, 1972).

$$F_n(x) = k_n x^n + k'_n x^{n-1} + \cdots$$

$$\int_a^b w(x) F_n^2(x) dx = h_n$$

$F_n$ is orthogonal with respect to $w(.)$.

### A.2.1 Differential Equations

Orthogonal polynomials satisfy differential equations of the form:

$$g_2(x) F''_n + g_1(x) F'_n a_n F_n = 0$$

with $g_2(x)$ and $g_1(x)$ independent of $n$ and $a_n$ is a constant depending only on $n$.

### A.2.2 Recurrence relation

The polynomials satisfy a recurrence relation of the form:

$$F_{n+1} = (a_n + x b_n) F_n - c_n F_{n-1}$$

where $b_n = \frac{k_{n+1}}{k_n}$, $a_n = b_n(\frac{k'_{n+1}}{k_{n+1}} - \frac{k'_n}{k_n})$ and $c_n = \frac{k_{n+1} k_{n-1} h_n}{k_n^2 h_{n-1}}$.
For more details on orthogonal polynomials see Abramowitz (1972).

### A.2.3 Gegenbauer Polynomials

The Gegenbauer polynomials $C_n^\alpha$ are also called *ultraspherical* polynomials. They are orthogonal polynomials defined with $z$ in the interval $[-1, 1]$ and with the weight function $w(z) = (1 - z^2)^{\alpha - \frac{1}{2}}$ (Abramowitz, 1972, p.774). Some special values of the Gegenbauer polynomials are (Abramowitz, 1972):

$$C_n^\alpha(1) = \binom{n + 2\alpha - 1}{n} \, \alpha \neq 0$$

$$C_n^0(1) = \frac{2}{n}$$

$$C_0^0(1) = 1 \, .$$

$$h_n = \begin{cases} \frac{\pi 2^{1-2\alpha}\Gamma(n+2\alpha)}{n1(n+\alpha)[\Gamma(\alpha)]^2} & \alpha \neq 0 \\ \frac{2\pi}{n^2} & \alpha = 0 \end{cases} \tag{A.2}$$

The general form of the Gegenbauer Polynomials is given in equation A.3. Where F(a,b,c;x) is the hypergeometric function, often referred to as ${}_2F_1(a,b,c;x)$ given in equation A.4. $(a)_n$ is the Pochhammer function, which is defined as $(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)} = a(a+1)(a+2)\dots(a+n-1)$ with $(-a)_n = (-1)^n(a)_n$.

$$C_n^\alpha(z) = \frac{(2\alpha)_n}{n!} F(-n, 2\alpha+n, \alpha+\frac{1}{2}; \frac{1-z}{2}) \tag{A.3}$$

$$F(a,b,c;x) = \sum_{k=0}^{\infty} \frac{(a)_k(b)_k}{(c)_k} x^k \frac{1}{k!} \tag{A.4}$$

The differential equation solved by the Gegenbauer Polynomials is (Abramowitz, 1972):

$$(1-z^2)\frac{d^2C_n^\alpha(z)}{dz^2} - (2a+1)z\frac{dC_n^\alpha(z)}{dz} + n(n+2*a)C_n^\alpha(z) = 0 \tag{A.5}$$

For the calculation of the derivatives, the differential relation in equation A.6 can be used (Abramowitz, 1972, p.783).

$$(1-x^2)\frac{d}{dx}C_n^\alpha(x) = -nxC_n^\alpha(x) + (n+2a-1)C_{n-1}^\alpha(x) \tag{A.6}$$

The recurrence relation of the Gegenbauer Polynomials is (Abramowitz, 1972, p.782):

$$(n+1)C_{n+1}^\alpha = 2(n+\alpha)zC_n^\alpha(z) - (n+2\alpha-1)C_{n-1}^\alpha(z) \tag{A.7}$$

**Different Definitions of the Gegenbauer Polynomials**

In the literature, the Gegenbauer polynomials are defined in different ways and modified versions are introduced. The different definitions are in fact only special cases of the general Gegenbauer Polynomials defined in Abramowitz (1972). The important difference is that the modified polyonmials are rescaled to $[0,1]$.

### A.2.4 Legendre Polynomial

$P_n(x)$ is the Legendre or spherical polynomial defined for $-1 < x < 1$ as

$$P_n(x) = \binom{2n}{n}(\frac{x-1}{2})^n F(-n, n+1; 1, \frac{1-x}{2}) \tag{A.8}$$

in terms of the hypergeometric function $F(.,.;.,.)$ (Abramowitz, 1972, 22.5.49). $P_n(x)$ is the solution to the differential equation

$$(1-x^2)\frac{d^2P_n(x)}{dx^2} - 2x\frac{dP_n(x)}{dx} + n(n+1)P_n(x)\,. \tag{A.9}$$

For the Gegenbauer polynomials and the Legendre polynomials the following relation holds

$$P_n(x) = C_n^{1/2}(x) \tag{A.10}$$

(Abramowitz, 1972, 22.5.36).

## A.3 Legendre Functions

The Legendre functions satisfy the differential equation in equation A.11 (Abramowitz, 1972, 8.1.1).

$$(1-z^2)\frac{d^2w}{dz^2} - 2z\frac{dw}{dz} + [v(v+1) - \frac{\mu^2}{1-z^2}]w = 0 \tag{A.11}$$

The variable $v$ is the degree and $\mu$ is the order of the Legendre function.The Legendre functions can be expressed in terms of the hypergeometric function (Abramowitz, 1972, 8.1.2).

$$P_v^\mu(z) = \frac{1}{\Gamma(1-\mu)}\left[\frac{z+1}{z-1}\right]^{\frac{1}{2}\mu} F(-v, v+1; 1-\mu; \frac{1-z}{2}),\ |1-z| < 2 \tag{A.12}$$

There are Legendre functions of the first and of the second kind. The Legendre functions of the second kind are singular at the origin. The Legendre functions of the first kind can be simplified to the Legendre polynomials (Weisstein, 2013b,a).

## A.4 Spheroidal Wave Functions

The spheroidal wave functions can be expressed as a linear combination of Legendre functions of first or second kind (Flammer, 1957; Meixner and Schäfke, 1954; Abramowitz, 1972; Weisstein, 2013c). The prolate spheroidal wave functions satisfy the differential equations (Abramowitz, 1972, 21.6)

$$\frac{d}{d\eta}[(1-\eta^2)\frac{d}{d\eta}S_{mn}(c,\eta)] + (\lambda_{mn} - c^2\eta^2 - \frac{m^2}{1-\eta^2})S_{mn}(c,\eta) = 0 \tag{A.13}$$

and

$$\frac{d}{d\xi}[(\xi^2-1)\frac{d}{d\xi}R_{mn}(c,\xi)] - (\lambda_{mn} - c^2\xi^2 + \frac{m^2}{\xi^2-1})R_{mn}(c,\xi) = 0\,. \tag{A.14}$$

$R_{mn}(z)$ is the radial and $S_{mn}(z)$ the angular spheroidal wave function. Their differential equations are equal but are defined over different ranges of $z$. The angular spheroidal wave functions are defined over the interval $z \in [-1,1]$ (Weisstein, 2013c).

The differential equations satisfied by the oblate spheroidal wave functions can be derived from the prolate differential equations by the transformation $\xi \rightarrow \pm i\xi$ and $c \rightarrow \pm ic$ (Abramowitz, 1972).
If $c = 0$ the spheroidal wave functions are equal to the Legendre Polynomials (Weisstein, 2013c).

# Appendix B

# Mathematica Code

In this chapter the mathematica code used for the graphics in chapter 5 is given and explained. For the legends the mathematica package `Plot Legends`[1] is used. This chapter is organized like chapter 5 such that the code corresponding to the respective figures can be found easily.

## B.1 Allele Frequency Density without Selection

### B.1.1 Gegenbauer Polynomials

**Polynomials**

The Gegenbauer polynomials are plotted for different values of $n$ in Figures 5.2 and 5.3.

```
Plot[    {GegenbauerC[0, 1.5, 1 - 2*x],
                  GegenbauerC[1, 1.5, 1 - 2*x],
                  GegenbauerC[2, 1.5, 1 - 2*x]}, {x, 0, 1},
         PlotLegend -> {Style["n=0", 18], Style["n=1", 18],
                  Style["n=2", 18]},
         LegendShadow -> None,
         PlotStyle -> AbsoluteThickness[2],
         AxesLabel -> {"x", "C_n^1.5(1-2x)"},
         LegendPosition -> {-0.9, -0.65},
         LegendSize -> 0.35,
         BaseStyle -> {FontSize -> 18}]

Plot[    {GegenbauerC[5, 1.5, 1 - 2*x],
```

[1] `http://reference.wolfram.com/mathematica/PlotLegends/tutorial/PlotLegends.html`

```
                GegenbauerC[6, 1.5, 1 - 2*x]},
        {x,0, 1},
        PlotLegend -> {Style[”n=5”, 18],
                Style[”n=6”, 18]},
        LegendShadow -> None,
        PlotStyle -> AbsoluteThickness[2],
        AxesLabel -> {”x”, ”C_n^1.5(1-2x)”},
        LegendPosition -> {-0.9, -0.63},
        LegendSize -> 0.35,
        BaseStyle -> {FontSize -> 18}]
```

Figure 5.4 was created for different values of $n$ with the following code:

```
ListLinePlot[
        Table[GegenbauerC[n, 1.5, 1 - 2*0.9],
                {n, 0, 40, 1}],
        PlotStyle -> AbsoluteThickness[2],
        AxesLabel -> {”n”, ”C_n^1.5(1-2p)”},
        BaseStyle -> {FontSize -> 18}]
```

**Allele frequency density**

The transition density given by the solution of Crow and Kimura (1970) is calculated.

```
phi[x_, p_, t_, lim_] := Module[{m}, {
        m = Ceiling[(-t + Sqrt[t]
                Sqrt[t - 4 Log[lim]])/(2 t)];
        ParallelSum[(2 i + 1)*p*(1 - p)/(i*(i + 1))*
                GegenbauerC[i - 1, 1.5, 1 - 2*p]*
                GegenbauerC[i - 1, 1.5, 1 - 2*x]*
                Exp[-i*(i + 1)*t], {i, 1, m}]}]
```

Since it is given as an infinte sum it is necessary to introduce a stopping condition. For this stopping condition $m$ is used. The sum is continued until $e^{-\lambda_m \tau}$ is smaller than $lim$. The plot in Figure 5.1 is created by the following code.

$$\text{findm[t\_, lim\_]:=Ceiling}[\frac{\sqrt{t}\sqrt{t-4\log(\text{lim})}-t}{2t}]$$

Plot[{findm[$t$, 0.01], findm[$t$, 0.001], findm[$t$, 0.0001]}, {$t$, 0, 0.2},

PlotLegend → {$e^{-\lambda_n \tau}$<0.01, $e^{-\lambda_n \tau}$<0.001, $e^{-\lambda_n \tau}$<0.0001},

LegendShadow → None,

PlotStyle → AbsoluteThickness[2],

PlotRange → {0, 80},

AxesLabel → {$\tau$, n},

LegendPosition → {0.2, 0},

LegendSize → 0.5,

BaseStyle → {FontSize → 18}]

The three dimensional plot of the transition density in Figure 5.5 is created with:

```
Plot3D[phi[x, 0.4, t, 0.0001], {x, 0.001, 0.999},
        {t, 0.0001, 0.1},
        AxesLabel -> {”allele frequency”, ”tau”, ”density”},
        PlotRange -> Full]
```

### Change in the Probability Mass

For Figures 5.6 and 5.7 the integral was calculated at some points and then the difference was taken.

```
intTable =
        Table[  Integrate[phi[x, 0.4, t, 0.001],
                        {x, 0.001, 0.999}][[1]],
                {t,  0.01, 0.2, 0.005}]

ListPlot[
        intTable,
        AxesLabel -> {”tau”, ”area”},
        DataRange -> {0, 0.2},
        PlotStyle -> AbsoluteThickness[2],
        BaseStyle -> {FontSize -> 18}]

ListPlot[-Differences[intTable],
        AxesLabel -> {”tau”, ”area’”},
        DataRange -> {0.01, 0.1},
        PlotRange -> {0, 0.0015},
        PlotStyle -> AbsoluteThickness[2],
        BaseStyle -> {FontSize -> 18}]
```

### B.1.2 Legendre Polynomials

Those are the plots for the solution presented in section 4.3.

#### Allele frequency density

The allele frequency density (Figure 5.8) is calculated using the following code. As before $m$ is used to set a stopping condition for the otherwise infinte sum.

```
phi[x_, p_, t_, lim_] := Module[{m},
          {m = Ceiling[(-t + Sqrt[t]
                    Sqrt[t - 4 Log[lim]])/(2 t)];
          ParallelSum[(2*i + 1)/2*LegendreP[i, 1 - 2*p]*
                    LegendreP[i, 1 - 2*x]*(x*(1 - x))^(1/2)/2
                    *Exp[-i*(i + 1)*t], {i, 0, m}]}]

Plot3D[
          phi[x, 0.4, t, 0.0001],
          {x, 0.001, 0.999},
          {t, 0.0001, 0.1},
          AxesLabel -> {”allele frequency”, ”tau”, ”density”},
          PlotRange -> Full]
```

#### Polynomials

The Gegenbauer polynomials are plotted for different values of $n$ in Figures 5.9 and 5.10.

```
Plot[     {LegendreP[0, 1.5, 1 - 2*x],
                    LegendreP[1, 1.5, 1 - 2*x],
                    LegendreP[2, 1.5, 1 - 2*x]},
          {x, 0, 1},
          PlotLegend -> {Style[”n=0”, 18],
                    Style[”n=1”, 18], Style[”n=2”, 18]},
          LegendShadow -> None,
           PlotStyle -> AbsoluteThickness[2],
          AxesLabel -> {”x”, ”P_n(1-2x)”},
          LegendPosition -> {-0.9, 0.23},
          LegendSize -> 0.3,
          BaseStyle -> {FontSize -> 18}]

Plot[   {LegendreP[5, 1.5, 1 - 2*x],
                    LegendreP[6, 1.5, 1 - 2*x]},
          {x, 0, 1},
          PlotLegend -> {Style[”n=5”, 18],
                    Style[”n=6”, 18]},
          LegendShadow -> None,
          PlotStyle -> AbsoluteThickness[2],
          AxesLabel -> {”x”,  ”P_n(1-2x)”},
          LegendPosition -> {-0.9, 0.48},
          LegendSize -> 0.3,
          BaseStyle -> {FontSize -> 18}]
```

## B.2 Allele Frequency Density with Selection

### B.2.1 Spheroidal Wave Functions

In Figures 5.11 and 5.12 the oblate spheroidal wave functions are shown for different values of $n$.

```
Plot[   {SpheroidalPS[1, 1, -I*1, 1 - 2*x],
                SpheroidalPS[2, 1, -I*1, 1 - 2*x],
                SpheroidalPS[3, 1, -I*1, 1 - 2*x]},
        {x, 0, 1},
        PlotLegend -> {Style[”n=0”, 18],
                Style[”n=1”, 18], Style[”n=2”, 18]},
        LegendShadow -> None,
        PlotStyle -> AbsoluteThickness[2],
        AxesLabel -> {”x”, ”S_n(1,1-2x)”},
        LegendPosition -> {-0.9, 0.23},
        LegendSize -> 0.3,
        BaseStyle -> {FontSize -> 18}]

Plot[   {SpheroidalPS[5, 1, -I*1, 1 - 2*x],
                SpheroidalPS[6, 1, -I*1, 1 - 2*x],
                 SpheroidalPS[7, 1, -I*1, 1 - 2*x]},
        {x, 0, 1},
        PlotLegend -> {Style[”n=5”, 18],
                Style[”n=6”, 18], Style[”n=7”, 18]},
        LegendShadow -> None,
        PlotStyle -> AbsoluteThickness[2],
        AxesLabel -> {”x”, ”S_n(1,1-2x)”},
        LegendPosition -> {-0.9, 0.25},
        LegendSize -> 0.3,
        BaseStyle -> {FontSize -> 18}]
```

### B.2.2 Allele Frequency Density

The allele frequency density is calculated with the following function:

```
density[x_, c_, t_] :=
        1/2*(x*(1 - x))^(1/2)*Exp[2*c*x]*
        ParallelSum[
                (Re[SpheroidalPS[n, 1, -I*c, 1 - 2*0.4]] *
                Re[SpheroidalPS[n, 1, -I*c, 1 - 2*x]])
                *Exp[-Re[SpheroidalEigenvalue[n, 1, c]
                        + c^2]*t]
                (2*n + 1)/(2*n*(n + 1)),
                {n, 1, 40}]
```

### B.2.3 Allele Frequency Density with $\gamma = 0.1$

The result can be seen in Figure 5.13.

```
list = Table[   density[x, 0.1, t],
                {t, 0.02, 0.1, 0.02},
                {x, 0.1, 0.9, 0.02}]

ListLinePlot[list,
        PlotLegend -> {Style[”tau=0.02”, 18],
                                Style[”tau=0.04”, 18],
                                Style[”tau=0.06”, 18],
                                Style[”tau=0.08”, 18],
                                Style[”tau=0.1”, 18]},
        LegendSize -> 0.6,
        PlotRange -> Automatic,
        DataRange -> {0, 1},
        LegendPosition -> {0.5, -0.1},
        LegendShadow -> None,
        PlotStyle -> AbsoluteThickness[2],
        AxesLabel -> {"x", "density"},
        BaseStyle -> {FontSize -> 18}]
```

### B.2.4 Allele Frequency Density with $\gamma = 1$

The result can be seen in Figure 5.15.

```
list2 = Table[  density[x, 1, t],
                {t, 0.02, 0.1, 0.02},
                {x, 0.1, 0.9, 0.02}]

ListLinePlot[
          list2,
          PlotLegend -> {Style["tau=0.02", 18],
                         Style["tau=0.04", 18],
                         Style["tau=0.06", 18],
                         Style["tau=0.08", 18],
                         Style["tau=0.1", 18]},
          PlotRange -> Automatic,
          DataRange -> {0, 1},
          LegendPosition -> {-0.90, -0.04},
          LegendSize -> 0.6,
          LegendShadow -> None,
          PlotStyle -> AbsoluteThickness[2],
          AxesLabel -> {"x", "density"},
          BaseStyle -> {FontSize -> 18},
          LegendSize -> 0.8]
```

### B.2.5 Allele Frequency Density with $\gamma = 3$

The result can be seen in Figure 5.14.

```
list3 = Table[  density[x, 3, t],
                {t, 0.02, 0.1, 0.02},
                {x, 0.1, 0.9, 0.02}]

ListLinePlot[
list3,
PlotLegend -> {Style[”tau=0.02”, 18],
               Style[”tau=0.04”, 18],
               Style[”tau=0.06”, 18],
               Style[”tau=0.08”, 18],
               Style[”tau=0.1”, 18]},
PlotRange -> Automatic,
DataRange -> {0, 1},
LegendPosition -> {-0.85, -0.05},
LegendSize -> 0.6,
LegendShadow -> None,
PlotStyle -> AbsoluteThickness[2],
AxesLabel -> {”x”, ”density”},
BaseStyle -> {FontSize -> 18}]
```

# Appendix C

# Zusammenfassung

In dieser Magisterarbeit wurden die Methoden zur Berechnung der Wahrscheinlichkeitsdichte der Allelfrequenzen von Kimura (1955) und Song and Steinrücken (2012) erklärt, und eine eigene Methode vorgestellt. Der älteste Lösungsweg stammt von Kimura (1955). Dieser basiert, im Fall mit Selektion, auf der Verwendung von Oblate Sphäroidfunktionen, wie sie von J. A. Stratton (1941) und J. A. Stratton (1954) vorgestellt wurden. Die Eigenwerte und Eigenfunktionen wurden von Kimura (1955) mittels Taylorentwicklung um den Selektionskoeffizenten $\gamma$ berechnet, als Folge davon ist die Lösung von Kimura für höhere Werte von $\gamma$ nicht sehr genau (Song and Steinrücken, 2012). Die Lösung von Song and Steinrücken (2012) basiert ebenfalls auf der Definition der Sphäroidfunktionen von J. A. Stratton (1954), aber die Eigenwerte werden durch die Lösung eines endlich-dimensionalen Gleichungssystems gefunden. Die Möglichkeiten der Computer gestützten Berechnung haben sich seit Kimura (1955) weiterentwickelt und so war es Song and Steinrücken (2012) möglich einen Algorithmus für die Berechnung der Eigenwerte in der Programmiersprache `C` zu implementieren.

Zur selben Zeit als Stratton die Tabellen zur Berechnung der Sphäroidfunktionen publizierte, bekamen die Sphäroidfunktionen, wegen ihrer Anwendungsmöglichkeiten in der Physik und Quantenmechanik, sehr viel Aufmerksamkeit. Daraus resultierte, dass innerhalb kurzer Zeit viele Publikationen mit unterschiedlichen Notationen der Sphäroidfunktionen erschienen. In Abramowitz (1972) gibt es eine Übersicht der unterschiedlichen Notationen. Die Notationen von Meixner and Schäfke (1954) und Flammer (1957) sind heute die üblichen Varianten. Die Sphäroidfunktionen von Meixner and Schäfke (1954) und Flammer (1957) und deren Eigenwerte sind in einigen Softwarepaketen implementiert (z.B. in Mathematica (Weisstein, 2013c) und im Mathematica Paket Falloon (2003)). Diese Funktionen wurden im Kapitel zu den Anwendungen in dieser Arbeit verwendet. Die Sphäroidfunktionen, die von Kimura

(1955), Song and Steinrücken (2012) und J. A. Stratton (1954) verwendet wurden, sind definiert als eine Linearkombination von Gegenbauer Polynomen. Die Sphäroidfunktionen werden aber allgemein als Linearkombination von Legendre Polynomen dargestellt. In Abschnitt 4.3 wird gezeigt, dass die Vorwärtsgleichung in die Differentialgleichung der Sphäroidfunktionen, definiert in Meixner and Schäfke (1954) und Flammer (1957), transformiert werden kann. Da diese Funktionen in vielen Softwarepaketen implementiert wurden, sind diese Methoden nun einfach für PopulationsgenetikerInnen zugänglich.

# Appendix D

# Curriculum Vitae

## Julia Theresa Csar, Bakk. rer. soc. oec

### Education

Primary School Volksschule 2, Rablstraße 24, Wels 1994-1998

Secondary School WRG der Franziskanerinnern, Wels 1998-2003

Secondary School BG/BRG Dr. Schauerstraße 9, Wels 2003-2006

## Higher Education

Bakk. rer. soc. oec in Statistics, University of Vienna 2006–2011

Bachelorstudies in Scientific Computing, University of Vienna 2007–today

Masterstudies in Statistics, University of Vienna 2011–today

### Bachelor's Thesis

*Measurements for Hurricane Intensity* 2011

Statistical analysis of meteorological data associated with hurricanes

*Vorhersage des Risikos einer Rezidiverkrankung von Thrombosepatienten* 2011

Estimation of risk for the recurrence of a thrombosis disease

### Working Experience

Scientific project coworker at the University of Vienna, Institute of Scientific Computing 2008–2010

Logi.Diag—Test Driven Development and Condition Monitoring in Automated Systems

Development and implementation of a statistical compression algorithm, which can be used for real time analysis and preventive maintenance. Implementation of several functions for condition monitoring.

WS2011/12 Tutor ”System- und Modelltheorie” at the University of Vienna

SS2012 Tutor ”Evaluation und Assessment im Bildungsbereich” at the University of Vienna

## Talks

2010 *Prototyping Predictive Maintenance Tools with R*
R User Conference, Gaithersburg, Maryland, USA

2010 *Condition Monitoring in der Automatisierungstechnik*
SPS IPC-Drives Messe, Nürnberg, Germany

# List of Tables

# List of Figures

# Index

# Bibliography

Abramowitz, S. (1972). *Handbook of Mathematical Functions.* National Bureau of Standards.

Baake, B. (2008). Ancestral processes with selection: branching and moran models. *Statistical Models in Biological Sciences, Banach Center Publications*, 80.

Charlesworth, B. and Charlesworth, D. (2010). *Elements of Evolutionary Genetics.* Roberts and Company Publishers.

Crow, J. F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory.* The Blackburn Press.

Ewens, W. J. (2000). *Mathematical Population Genetics*, volume 27. Springer, second edition edition.

Falloon, P. E. (2003). Theory and computation of the spheroidal wave functions. *Journal of Physics.*

Flammer, C. (1957). *Spheroidal Wave Functions.* Stanford University Press.

J. A. Stratton, e. a. (1941). *Elliptic Cylinder and Spheroidal wave functions.* The Technology Press of the Massachusetts Institute of Technology.

J. A. Stratton, e. a. (1954). *Spheroidal wave functions.* The Technology Press of the Massachusetts Institute of Technology.

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

Kimura (1955). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Press*, 20:33–53.

Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability*, 1.

Kirby, P. (2006). Calculation of spheroidal wave functions. *Elsevier.*

Mano, S. (2009). Duality, ancestral and diffusion processes in models with selection. *Theoretical Population Biology*, 75.

Meixner, J. and Schäfke, F. W. (1954). *Mathieusche Funktionen und Sphäroidfunktionen.* Springer-Verlag.

Norris, J. (1997). *Markov Processes.* Cambridge Series in Statistical and Probabilistic Mathematics.

Song, Y. and Steinrücken, M. (2012). A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics*, 190.

Sullivan, F. and Thompson, W. J. (1999). Spheroidal wave functions. *Computing in Science and Engineering.*

Vogl, C. and Clemente, F. (2012). The allele-frequency spectrum in a decoupled moran model with mutation, drift, and directional selection, assuming small mutation rates. *Theoretical Population Biology.*

Wakeley, J. (2009). *Coalescent Theory, An Introduction.* Roberts & Company Publishers.

Weisstein, E. (2013a). Legendre differential equation. *MathWorld – A Wolfram Web Resource.*

Weisstein, E. (2013b). Legendre function of the first kind. *MathWorld – A Wolfram Web Resource.*

Weisstein, E. (2013c). Spheroidal wave function. *MathWorld – A Wolfram Web Resource.*