Computation **2014**, *2*, 199-220; doi:10.3390/computation2040199

OPEN ACCESS

computation

ISSN 2079-3197

www.mdpi.com/journal/computation

Review

Computation of the Likelihood in Biallelic Diffusion Models Using Orthogonal Polynomials

Claus Vogl

Institute of Animal Breeding and Genetics, University of Veterinary Medicine, Vienna,
Veterinarplatz 1, 1210 Vienna, Austria; E-Mail: claus.vogl@vetmeduni.ac.at;
Tel.: +43-12-5077-5631

External Editor: Marnix Medema

Received: 14 July 2014; in revised form: 12 October 2014 / Accepted: 16 October 2014 / Published: 14 November 2014

Abstract: In population genetics, parameters describing forces such as mutation, migration and drift are generally inferred from molecular data. Lately, approximate methods based on simulations and summary statistics have been widely applied for such inference, even though these methods waste information. In contrast, probabilistic methods of inference can be shown to be optimal, if their assumptions are met. In genomic regions where recombination rates are high relative to mutation rates, polymorphic nucleotide sites can be assumed to evolve independently from each other. The distribution of allele frequencies at a large number of such sites has been called "allele-frequency spectrum" or "site-frequency spectrum" (SFS). Conditional on the allelic proportions, the likelihoods of such data can be modeled as binomial. A simple model representing the evolution of allelic proportions is the biallelic mutation-drift or mutation-directional selection-drift diffusion model. With series of orthogonal polynomials, specifically Jacobi and Gegenbauer polynomials, or the related spheroidal wave function, the diffusion equations can be solved efficiently. In the neutral case, the product of the binomial likelihoods with the sum of such polynomials leads to finite series of polynomials, *i.e.*, relatively simple equations, from which the exact likelihoods can be calculated. In this article, the use of orthogonal polynomials for inferring population genetic parameters is investigated.

Keywords: site frequency spectrum; mutation; drift; biallelic diffusion; directional selection; orthogonal polynomials; inference

1. Introduction

Population genetics is concerned with the evolution of frequencies of heritable variants (alleles) at specific positions in the genome (loci) in natural and domestic populations. The main forces influencing this evolution of allele frequencies are: mutation, migration, drift, linkage, and selection.

In genomic regions, where recombination rates are high relative to mutation rates, polymorphic nucleotides or sites can be assumed to evolve independently, *i.e.*, linkage can be ignored. The distribution of allele frequencies at a large number of such sites in a sample has been called “allele-frequency spectrum” or “site-frequency spectrum” (SFS). Some classes of sites are assumed not to be influenced directly by selection, e.g., some regions in short introns (non-coding insertions in protein coding genes) and fourfold degenerate sites (sites at third positions in codons that do not influence the amino acid in the polypeptides) [1]. With these classes of sites, only mutation and demographic forces, e.g., drift and migration, are assumed to be relevant.

In most organisms studied so far, e.g., fruit flies (*Drosophila*) or mammals including humans, most sites in moderate samples (up to about 100 individuals) are monomorphic; only at some sites a single allele segregates in the population, while sites with more than two segregating alleles are extremely rare. With such a low proportion of polymorphism, a simple biallelic model is adequate, even though each site may assume four states corresponding to the four bases: adenine, cytosine, guanine, and thymine.

Most naturally, the evolution of allele frequencies is modeled forward in time as a Markovian random walk from one generation to the next. The best known such model, the Wright–Fisher model [2,3], uses diploid individuals and binomial sampling of individuals for the transition to the next generation. For large population sizes and if parameters are scaled appropriately, the Wright–Fisher model and other similar models, e.g., the Moran model [4], converge to the same diffusion equation. In population genetics, the use of diffusion equations is associated with the work of Motoo Kimura (1924–1994) [5,6]. In the diffusion limit, allele frequency counts are usually replaced by the allelic proportion x, a continuous quantity ranging between zero and one. Often, solutions that are difficult or impossible to derive with the discrete models can be obtained relatively easily with the diffusion approach. The diffusion model can either be taken as an approximation to the discrete models or as a model in its own right.

While occasionally new results are presented, this article is mainly a review. Population genetic parameters are inferred from site frequency spectra with the diffusion approach. Note that sudden changes in parameters, such as the (effective) population size, may lead to discontinuous jumps in the process, which cannot naturally be modeled by diffusion. These are not subject of this article. Neither are alternative approaches, such as branching processes. In particular, orthogonal polynomials (modified Jacobi and Gegenbauer polynomials) are used to solve the diffusion equation. Furthermore, the use of the oblate spheroidal wave function for solving a model with directional selection and drift is presented. Other methods to analyze diffusion models, such as the calculation of moments [7,8], are not covered. Due to the importance of Ewen’s book [9] in the field of theoretical population genetics, subjects not covered in the book receive special attention. In particular, data analysis with the diffusion approach is reviewed also for equilibrium data, which generally do not require the use of orthogonal polynomials, but the equilibrium distributions may serve as prior distributions in a Bayesian context.

The currently available tools implementing these approaches are limited: functions for Jacobi and Gegenbauer polynomials as well as the oblate spheroidal wave function are available in the formula manipulation programs “Mathematica” [10] and “Maple” [11] for computation and visualization. Song and Steinrücken [12] provide methods and an implementation to solve the Kolmogorov backward equation using modified Jacobi and Gegenbauer polynomials.

2. Mutation and Drift Diffusion

2.1. Moran and Diffusion Models

Assume a population of N haploid individuals; each may assume the state of zero or one, corresponding to the two arbitrarily labeled alleles. With the decoupled Moran model [13–15], either (i) (**mutation**) at a rate of $\mu = \mu_0 + \mu_1$, a random individual i is picked to mutate to type one with probability $\alpha = \mu_1/\mu$ or to type zero with probability $\beta = \mu_0/\mu$; or (ii) (**genetic drift**) at a rate of one, a random individual i is replaced by another random individual j. Setting $\theta = \mu N$, the rate of change of the allelic proportion x of the mean per unit time is caused by mutation:

$$\mathrm{M}_{\delta x} = \frac{1}{N^2}\theta(\alpha - x)N \tag{1}$$

and that of the variance by genetic drift:

$$\mathrm{V}_{\delta x} = \frac{2}{N^2}x(1-x)N^2 \tag{2}$$

Scaling space with $1/N$ and time with $1/N^2$ and taking the appropriate limits, the Kolmogorov forward (or Fokker–Planck) generator of the process becomes:

$$\mathcal{L}_f = \left(\frac{\partial^2}{\partial x^2}x(1-x)\right) - \left(\frac{\partial}{\partial x}\theta(\alpha - x)\right) \tag{3}$$

The forward diffusion equation:

$$\frac{\partial}{\partial t}\phi(x,t) = \mathcal{L}_f\phi(x,t) \tag{4}$$

then describes the evolution of the probability of the allelic proportion x forward in time t, *i.e.*, in the same temporal direction as the transitions in the discrete Wright–Fisher and Moran models. By contrast, coalescence theory looks backward in time (e.g., [16]). (In the following, the use of orthogonal polynomials to solve this diffusion equation will be explained. This is necessarily rather technical; however, for most results only rather elementary mathematical manipulations are needed.)

2.2. Solution of the Mutation-Drift Diffusion Using Modified Jacobi Polynomials

2.2.1. Relationship of the Forward and Backward Diffusion Equation; Sturm–Liouville Form

While this article focuses on the Kolmogorov forward equation, some results can more easily be derived using the Kolmogorov backward generator:

$$\mathcal{L}_b = x(1-x)\frac{\partial^2}{\partial x^2} + \theta(\alpha - x)\frac{\partial}{\partial x} \tag{5}$$

On the interval $[0,1]$ we are looking for solutions of the Kolmogorov backward equation:

$$\frac{\partial}{\partial t}\phi(x,t) = \mathcal{L}_b\phi(x,t) \tag{6}$$

of the form $\phi(x,t) = e^{-\lambda_i t}f_i(x)$:

$$-\lambda_i f_i(x) = \left(x(1-x)\frac{d^2}{dx^2}f_i(x)\right) + \left(\theta(\alpha - x)\frac{d}{dx}f_i(x)\right) \tag{7}$$

where i indexes the eigenvectors.

The forward equation can be transformed to Sturm–Liouville or self-adjoint form by substituting $x^{\alpha\theta-1}(1-x)^{\beta\theta-1}f_i(x) = g_i(x)$:

$$\begin{aligned}
-\lambda_i g_i(x) &= \frac{d^2}{dx^2}\left(x(1-x)g_i(x)\right) - \frac{d}{dx}\left(\theta(\alpha-x)g(x)\right)\\
-\lambda_i x^{\alpha\theta-1}(1-x)^{\beta\theta-1}f_i(x) &= \frac{d^2}{dx^2}\left(x(1-x)x^{\alpha\theta-1}(1-x)^{\beta\theta-1}f_i(x)\right)\\
&\quad - \frac{d}{dx}\left(\theta(\alpha-x)x^{\alpha\theta-1}(1-x)^{\beta\theta-1}f_i(x)\right)\\
&= \frac{d}{dx}\left(x^{\alpha\theta}(1-x)^{\beta\theta}\frac{d}{dx}f_i(x)\right)\\
&\quad + \frac{d}{dx}\left(\theta\left(-\alpha+x\right)x^{\alpha\theta-1}(1-x)^{\beta\theta-1}f_i(x)\right)\\
&\quad - \frac{d}{dx}\left(\theta\left(-\alpha+x\right)x^{\alpha\theta-1}(1-x)^{\beta\theta-1}f_i(x)\right)\\
&= \frac{d}{dx}\left(x(1-x)x^{\alpha\theta-1}(1-x)^{\beta\theta-1}\frac{d}{dx}f_i(x)\right)\\
&= \frac{d}{dx}\left(x^{\alpha\theta}(1-x)^{\beta\theta}\frac{d}{dx}f_i(x)\right)
\end{aligned} \tag{8}$$

For Sturm–Liouville equations [17] of the form:

$$-\lambda w(x)f(x) = \frac{d}{dx}\left(p(x)\frac{d}{dx}f(x)\right) - q(x)f(x) \tag{9}$$

it can be shown that all eigenvalues λ_i are real and can be ordered such that $\lambda_0 < \lambda_1 < \lambda_2 < \cdots < \lambda_i < \cdots \to \infty$. Corresponding to each eigenvalue λ_i is a unique (up to a normalization constant) eigenfunction $f_i(x)$, which has exactly i zeros in the interval. The normed eigenfunctions form an orthonormal basis:

$$\int_a^b f_i(x)f_j(x)\,w(x)\,dx = \delta_{i,j} \tag{10}$$

where $\delta_{i,j}$ denotes the Kronecker delta, *i.e.*, $\delta_{i,j}$ is zero for $i \neq j$ and one for $i = j$, of the Hilbert space $L^2([a,b], w(x)dx)$. The function $w(x)$ is called the weight function.

The Kolmogorov backward Equation (7) can be obtained from the above Sturm–Liouville equation (the last line of Equation (8)):

$$\begin{aligned}-\lambda_i x^{\theta_1 - 1}(1-x)^{\theta_0 - 1} f_i(x) &= \frac{d}{dx}\left(x^{\alpha\theta}(1-x)^{\beta\theta}\frac{d}{dx} f_i(x)\right) \\ &= x^{\alpha\theta}(1-x)^{\beta\theta}\frac{d^2}{dx^2} f_i(x) \\ &\qquad + x^{\alpha\theta - 1}(1-x)^{\beta\theta - 1}\theta(\alpha - x)\frac{d}{dx} f_i(x) \\ -\lambda_i f(x) &= x(1-x)\frac{d^2}{dx^2} f_i(x) + \theta(\alpha - x)\frac{d}{dx} f_i(x)\end{aligned} \tag{11}$$

Thus, multiplication with the weight function:

$$w^{(\theta,\alpha)}(x) = x^{\alpha\theta - 1}(1-x)^{\beta\theta - 1} \tag{12}$$

transforms a solution of the backward equation into that of the forward equation (see also Formula (1.7) in [18]).

2.2.2. Modified Jacobi Polynomials

The backward Equation (7) is closely related to the differential function fulfilled by the classical Jacobi polynomials [19]. One can either modify the backward Equation (7) to fit the Jacobi polynomials (e.g., [18]) or the Jacobi polynomials to fit the backward equation (e.g., [12]). We will follow the latter strategy.

Define the modified Jacobi polynomials:

$$R_i^{(\theta,\alpha)}(x) = P_i^{(\alpha\theta - 1, \beta\theta - 1)}(2x - 1) \tag{13}$$

where the $P_i^{(\alpha,\beta)}(z)$ are the classical Jacobi polynomials [19]. It can be shown that these modified Jacobi polynomials fulfill the backward Equation (7) with the corresponding eigenvalues:

$$\lambda_i = i(i + \theta - 1) \tag{14}$$

With the weight function $w^{(\theta,\alpha)}(x)$, the modified Jacobi polynomials are orthogonal:

$$\int_0^1 R_i^{(\theta,\alpha)}(x) R_j^{(\theta,\alpha)}(x)\, w^{(\theta,\alpha)}(x)\, dx = \Delta_i^{(\alpha,\theta)}\delta_{i,j} \tag{15}$$

where $\delta_{i,j}$ denotes the Kronecker delta, *i.e.*, $\delta_{i,j}$ is zero for $i \neq j$ and one for $i = j$. The proportionality constant depends on i, θ, and α:

$$\Delta_i^{(\alpha,\theta)} = \frac{\Gamma(i + \alpha\theta)\Gamma(i + \beta\theta)}{(2i + \theta - 1)\Gamma(i + \theta - 1)\Gamma(i + 1)} \tag{16}$$

The set of $R_i^{(\theta,\alpha)}(x)$ forms a basis of the Hilbert space $L^2([0,1], w^{(\theta,\alpha)}(x)dx)$ [12].

For $i \geq 1$, the $R_i^{(\theta,\alpha)}(x)$ satisfy the recurrence relation:

$$\begin{aligned} R_{i+1}^{(\theta,\alpha)}(x)\frac{(i+1)(i-1+\theta)}{(2i+\theta)(2i-1+\theta)} &= \\ R_i^{(\theta,\alpha)}(x)\left(x - \tfrac{1}{2} + \frac{\theta^2(\beta^2 - \alpha^2) - 2\theta(\beta - \alpha)}{2(2i+\theta)(2i-2+\theta)}\right) & \\ - R_{i-1}^{(\theta,\alpha)}(x)\frac{(i-1+\alpha\theta)(i-1+\beta\theta)}{(2i-1+\theta)(2i-2+\theta)} & \end{aligned} \tag{17}$$

while $R_0^{(\theta,\alpha)}(x) = 1$ and $R_1^{(\theta,\alpha)}(x) = \theta(x - \alpha)$ [12].

Recall that multiplication with the weight function $w^{(\theta,\alpha)}(x)$ transforms an eigenvector of the backward equation into that of the forward equation. If $\theta > 0$, the forward equation has a stationary beta density $\mathrm{beta}(x \mid \alpha\theta, \beta\theta)$ proportional to the weight function:

$$\Pr(x \mid \theta, \alpha, \beta, t \to \infty) = \frac{1}{\Delta_0^{(\alpha,\theta)}} w^{(\theta,\alpha)}(x)\, R_0^{(\theta,\alpha)}(x) = \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\, x^{\alpha\theta-1}(1-x)^{\beta\theta-1} \tag{18}$$

2.2.3. Series Expansion; Approximation of Functions by Orthogonal Polynomials

We have now established that the backward density is given by an expansion of the form:

$$f(x \mid \theta, \alpha, \beta, t) = c_0 + \sum_{i=1}^{\infty} e^{-i(i+\theta-1)\,t}\, c_i\, R_i^{(\theta,\alpha)}(x) \tag{19}$$

The constants c_i need to be determined such that the initial conditions are met, *i.e.*, a probability density $f(x)$, defined within the interval, is represented by the series expansion:

$$f(x) = c_0 + \sum_{i=1}^{\infty} c_i\, R_i^{(\theta,\alpha)}(x) \tag{20}$$

The coefficients c_i in an expansion up to order n are determined by minimizing a weighted least squares error function.

Since the following considerations hold generally for all orthogonal polynomials, we now use arbitrary intervals between a and b, the symbol $f_i(x)$ for the ith orthogonal polynomial, and $w(x)$ for the weight function associated with the $f_i(x)$:

$$E(c_0, \ldots, c_n) = \int_a^b w(x) \left(f(x) - \sum_{i=0}^{n} c_i f_i(x) \right)^2 dx \tag{21}$$

Differentiating with respect to c_i:

$$\frac{dE(c_0, \ldots, c_n)}{dc_i} = -2 \int_a^b w(x) f_i(x) \left(f(x) - \sum_{j=0}^{n} c_j f_j(x) \right) dx \tag{22}$$

and setting equal to zero, we get:

$$\int_a^b w(x) f_i(x) f(x)\, dx = \sum_{i=0}^{n} \int_a^b f_i(x) c_j w(x) f_j(x)\, dx \tag{23}$$

From the orthogonality relation, we have $\int_a^b f_i^2(x) w(x)\, dx = \Delta_i \delta_{i,j}$. Thus, we set the coefficients for the backward equation to:

$$c_i = \frac{1}{\Delta_i} \int_a^b w(x) f_i(x) f(x)\, dx \tag{24}$$

The forward expansion can be obtained from the backward expansion by multiplication with the weight function (see Equation (11)):

$$f(x \mid \theta, \alpha, \beta, t) = w^{(\theta,\alpha)}(x) \left(c_0 + \sum_{i=1}^{\infty} e^{-i(i+\theta-1)\,t}\, c_i\, R_i^{(\theta,\alpha)}(x) \right) \tag{25}$$

As in the case of the backward expansion, the constants c_i are determined such that the initial conditions are met, *i.e.*, an initial probability density $f(x)$, defined within the interval between a and b, is represented by the series expansion of orthogonal polynomials $f_i(x)$ with the weight function $w(x)$:

$$f(x) = w(x)\left(c_0 + \sum_{i=1}^{\infty} c_i\, f_i(x)\right) \tag{26}$$

The coefficients are now determined by minimizing the weighted least squares error function:

$$E(c_0, \dots, c_n) = \int_0^1 w(x)^{-1}\left(f(x) - \sum_{i=0}^{n} c_i w(x) f_i(x)\right)^2 dx \tag{27}$$

With similar considerations as for the backward case, we find:

$$c_i = \frac{1}{\Delta_i}\int_a^b f_i(x) f(x)\, dx \tag{28}$$

Returning to the mutation drift diffusion, we note that often an initial density corresponding to a Dirac delta function at a point p in $[0,1]$, $f(x) = \delta(x-p)$, is considered (e.g., [20]); then:

$$c_i(p) = \frac{R_i(p)}{\Delta_i} \tag{29}$$

Substituting these coefficients into Equation (25), we get:

$$f(x\,|\,\theta,\alpha,p,t) = w^{(\theta,\alpha)}(x)\left(c_0 + \sum_{i=1}^{\infty} e^{-i(i+\theta-1)\,t}\; R_i(x)\frac{R_i(p)}{\Delta_i}\right) \tag{30}$$

This corresponds to Formula (4.68) in [9], where the eigenfunctions are assumed to be normed, such that division by the proportionality constant Δ_i is unnecessary. (Note that exchanging x and p transforms the right side of this equation into that of the corresponding backward equation; compare Formula (5) in [12], where the backward equation is used.)

Returning to the modified Jacobi polynomials, we note that, from the orthogonality relation Equation (15) and $R_0^{(\theta,\alpha)}(x) = 1$, it can be deduced for all $i \geq 1$ and thus also for all times:

$$0 = \int_0^1 R_i^{(\theta,\alpha)}(x)\, R_0^{(\theta,\alpha)}(x)\, w(x)\, dx = \int_0^1 R_i^{(\theta,\alpha)}(x)\, w^{(\theta,\alpha)}(x)\, dx \tag{31}$$

Therefore the probability mass over the whole interval $[0,1]$ comes only from the equilibrium term, *i.e.*, the beta density Equation (18); all other terms $R_i^{(\theta,\alpha)}(x)\, w^{(\theta,\alpha)}(x)$ with $i \geq 1$ shift this mass within the interval. Note that a polynomial times a beta density results in a weighted sum of beta densities.

2.2.4. Example: A Change in the Scaled Mutation Rate with Modified Jacobi Polynomials

As an example, assume that the population had been in equilibrium with parameters α and θ_a, to switch to a new mutation bias θ_c at time t_c. Then the expansion until time t_c contains only the equilibrium beta density. The change of the mutation bias necessitates a change in the eigenvectors from $w^{(\theta_a,\alpha)}(x)\, R_i^{(\theta_a,\alpha)}(x)$ to $w^{(\theta_c,\alpha)}(x)\, R_i^{(\theta_c,\alpha)}(x)$. The coefficients for the new eigensystem are (compare Formula (28)):

$$c_i = \frac{1}{\Delta_i^{\theta_c,\alpha}} \int_0^1 R_i^{(\theta_c,\alpha)}(x) \frac{1}{\Delta_0^{\theta_a,\alpha}} w^{(\theta_a,\alpha)}(x)\, R_0^{(\theta_a,\alpha)}(x)\, dx \tag{32}$$

The evolution of the proportion $f(x)$ between t_c and the present time is given by the series expansion Equation (25) with the c_i from Equation (32).

While one such change may not be too cumbersome to implement in a computer program, approximating rapidly changing population sizes by many piecewise linear changes can be, since then equilibrium has not been reached and for each change a sum over all terms in the expansion is needed, such that Equation (32) needs to be modified to:

$$c_i = \frac{1}{\Delta_i} \int_0^1 R_i^{(\theta_c,\alpha)}(x) w^{(\theta_a,\alpha)} \sum_i \tau_i(t) R_i^{(\theta_a,\alpha)}(x)\, dx \tag{33}$$

where the $\tau_i(t)$ are the time-dependent coefficients.

2.3. Statistics of Site Frequency Spectra

2.3.1. Equilibrium

For $\theta > 0$, the beta density $\mathrm{beta}(x \,|\, \alpha\theta, \beta\theta)$ is the equilibrium or stationary solution of the forward diffusion process [3].

Given a single sample of size $M \ll N$ with a frequency y of the first allelic type, the likelihood conditional on the population allelic proportion x is naturally a binomial:

$$\Pr(y \,|\, x, M) = \binom{M}{y} x^y (1-x)^{M-y} \tag{34}$$

The joint distribution of y and x after multiplication with the equilibrium beta density Equation (18) is:

$$\Pr(y, x \,|\, \alpha, \theta, M) = \binom{M}{y} \frac{\Gamma(\alpha\theta)\Gamma(\beta\theta)}{\Gamma(\theta)} x^{y+\alpha\theta-1}(1-x)^{M+\beta\theta-y-1} \tag{35}$$

Integrating out x results in the likelihood, a beta-binomial compound distribution:

$$\begin{aligned} \Pr(y \,|\, \alpha, \theta, M) &= \binom{M}{y} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \int_0^1 x^{y+\alpha\theta-1}(1-x)^{M-y+\beta-1}\, dx \\ &= \binom{M}{y} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \frac{\Gamma(y+\alpha\theta)\Gamma(M-y+\beta\theta)}{\Gamma(M+\theta)} \end{aligned} \tag{36}$$

Site frequency spectra (SFS) can be considered samples of identical sample size M from L biallelic loci, indexed by l ($1 \le l \le L$), with the allelic proportions x_l drawn independently from a beta density with common α and θ. Let $L_0, \ldots, L_M$ represent the counts of alleles of the first type in the samples. The likelihood then is a product of beta-binomials:

$$\Pr(L_0, \ldots, L_M \,|\, \alpha, \theta, M) = \frac{L!}{\prod_{i=0}^{M} L_y!} \prod_{y=0}^{M} \left(\binom{M}{y} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \frac{\Gamma(y+\alpha\theta)\Gamma(M-y+\beta\theta)}{\Gamma(M+\theta)} \right)^{L_y} \tag{37}$$

Interest is centered on obtaining (maximum-likelihood) estimates of θ and α given the vector of allelic counts $(L_0, \ldots, L_M)$ or, in a Bayesian context, their posterior distribution given a suitable prior. As a function of α, the distribution is a polynomial; as a function of θ, the distribution is a rational function. A rational function can be integrated by partial fraction decomposition. If auxiliary variables that count the number of mutations in each allelic class conditional on θ, α, y and M are introduced, an expectation maximization algorithm may be used for finding the maximum likelihood estimates [21].

2.3.2. Outside Equilibrium

If the population size or the mutation bias has changed recently, a population will be outside equilibrium. Then instead of the equilibrium beta density Equation (18) the expansion in Equation (32) or Equation (33) needs to be used. Since in both cases, the series conforms to a weighted sum of beta distributions, integration to obtain the likelihood can be performed relatively easily; however, the author is not aware of an implementation of this algorithm. With formula manipulation programs, e.g., “Mathematica” [10] or “Maple” [11], the Jacobi polynomials are readily available, such that it is possible to program these algorithms relatively easily.

3. Selection and Drift Diffusion with Mutations from the Boundaries

Another tradition in theoretical population genetics follows the fate of a single mutant allele to calculate, e.g., the probability of fixation of the mutant allele or the time until its fixation or loss [9]. While the allele is polymorphic, directional selection with a scaled strength of γ and drift are the forces usually considered. Importantly, mutation is assumed to be negligible within the polymorphic region, *i.e.*, in $1/N \leq x \leq (N-1)/N$. While only drift (or selection and drift) governs the dynamics within the polymorphic region, mutations may be considered as boundary terms. The fact that no mutations are considered within the polymorphic region means that expansions using the Gegenbauer polynomials instead of the Jacobi polynomials can be used. This change simplifies calculations.

Two different ways of approaching the problem with selection are presented: in the Appendix, the forward equation is transformed to the spheroidal wave equation, for which excellent computation tools are available; in the main text, the strategy of Song and Steinrücken [12] is followed, which will likely be more familiar to population geneticists.

From the corresponding Moran model, e.g., [15], the change in the mean is now inferred to be:

$$\mathrm{M}_{\delta x} = \frac{1}{N^2}\gamma x(1-x)N \tag{38}$$

After scaling, the corresponding forward diffusion equation is for $1/N < x < (N-1)/N$:

$$\frac{\partial}{\partial t}\phi(x,t) = \left(\frac{\partial^2}{\partial x^2}x(1-x) - \frac{\partial}{\partial x}\gamma x(1-x)\right)\phi(x,t) \tag{39}$$

Again from the corresponding Moran model, it can be deduced that the flow from $x = 1/N$ to $x = 0$ consists of drift and that between $x = (N-1)/N$ and $x = 1$ of selection and drift; after appropriate scaling:

$$\begin{cases} \frac{F(1/N)}{dt} = \frac{N-1}{N}\phi(1/N,t) \\ \frac{F((N-1)/N)}{dt} = (1+\gamma/N)\frac{N-1}{N}\phi((N-1)/N,t) \end{cases} \tag{40}$$

By multiplication with the weight function $w(x) = e^{\gamma x}x^{-1}(1-x)^{-1}$ and substituting the series expansion, the forward Equation (39) can be transformed to Sturm–Liouville form:

$$\begin{aligned} -\lambda_i e^{\gamma x}x^{-1}(1-x)^{-1}f_i(x) &= \left(\frac{d^2}{dx^2}x(1-x) - \frac{d}{dx}\gamma x(1-x)\right) e^{\gamma x}x^{-1}(1-x)^{-1}f_i(x) \\ &= \frac{d^2}{dx^2}e^{\gamma x}\phi(x,t) - \gamma\frac{d}{dx}e^{\gamma x}f_i(x) \\ &= \frac{d}{dx}\left(e^{\gamma x}\frac{d}{dx}f_i(x)\right) \end{aligned} \tag{41}$$

From this the backward equation can be obtained:

$$\begin{aligned} -\lambda_i e^{\gamma x}x^{-1}(1-x)^{-1}f_i(x,t) &= \frac{d}{dx}\left(e^{\gamma x}\frac{d}{dx}f_i(x)\right) \\ &= \gamma e^{\gamma x}\frac{d}{dx}\phi(x,t) + e^{\gamma x}\frac{d^2}{dx^2}f_i(x) \\ -\lambda_i f_i(x) &= \gamma x(1-x)\frac{d}{dx}f_i(x) + x(1-x)\frac{d^2}{dx^2}f_i(x) \end{aligned} \tag{42}$$

Again, we are looking for an eigensystem of this Sturm–Liouville problem. We proceed indirectly, by first obtaining a solution for the neutral system, *i.e.*, without selection, and then deriving eigenvectors as linear combinations of this eigensystem.

3.1. Pure Drift within the Polymorphic Region

Consider the pure drift forward generator:

$$\mathcal{L}_f = \frac{\partial^2}{\partial x^2}x(1-x) \tag{43}$$

and the corresponding backward generator:

$$\mathcal{L}_b = x(1-x)\frac{\partial^2}{\partial x^2} \tag{44}$$

As before, either the generator can be modified to that of the classical Gegenbauer polynomials [19] as in [20], or the Gegenbauer polynomials to fit the generator. Choosing the latter strategy again, the orthogonal polynomials solving the backward equation are the modified Gegenbauer polynomials [12] with the weight function:

$$w(x) = x^{-1}(1-x)^{-1} \tag{45}$$

and the proportionality constant:

$$\Delta_i = \frac{i+1}{(i+2)(2i+3)} \tag{46}$$

The first two polynomials are $G_0(x) = -x(1-x)$ and $G_1(x) = x(1-x)(2-4x)$ and the recurrence relation to calculate all other polynomials is:

$$G_{i+1}(x)\frac{(i+3)(i+1)}{2(i+2)(2i+3)} = G_i(x)\left(x-\tfrac{1}{2}\right) - G_{i-1}(x)\frac{(i+1)}{2(2i+3)} \tag{47}$$

Furthermore, the eigenvalues are:

$$\lambda_i = (i+2)(i+1) \tag{48}$$

As before, the eigenvectors of the forward series expansion $U_i(x)$ can be obtained from those of the backward expansion by multiplication with the weight function $w(x)$:

$$U_i(x) = (x(1-x))^{-1} G_i(x) \tag{49}$$

Since all eigenvalues are greater than zero, there is no equilibrium term in this expansion. Without replenishing mutations, probability weight is lost continually towards the boundaries zero and one. It is convenient to already include this behavior into the eigenfunctions by including boundary terms at zero and one [22,23], where the deduction made corresponds to what is expected to fix eventually. One can show that:

$$\begin{cases} \int_0^1 U_i(x)\,dx = (U_i(0) + U_i(1))/\lambda_i \\ \int_0^1 (1-x) U_i(x)\,dx = U_i(0)/\lambda_i = \frac{(-1)^i}{i+2} \\ \int_0^1 x U_i(x)\,dx = U_i(1)/\lambda_i = \frac{1}{i+2} \end{cases} \tag{50}$$

Therefore the forward eigenvectors $H_i(x)$ are defined as:

$$H_i(x) = \frac{(-1)^i}{i+2}\,\delta(x) + U_i(x) + \frac{1}{i+2}\,\delta(1-x) \tag{51}$$

where $\delta(x)$ is the Dirac delta function.

A probability density $f(x)$ defined between zero and one can be represented by an expansion of the $H_i(x)$:

$$f(x) = b_1\delta(x-1) + b_0\delta(x) + \sum_{i=2}^{\infty} (c_i H_i(x)) \tag{52}$$

where:

$$\begin{cases} b_0 = \int_0^1 x f(x)\,dx \\ b_1 = 1 - b_0 = \int_0^1 (1-x) f(x)\,dx \end{cases} \tag{53}$$

Should $f(x)$ have point masses at the boundaries, these are included in this integration. The coefficients c_i can be calculated using:

$$c_i = \frac{1}{\Delta_i} \lim_{N\to\infty} \int_{1/N}^{1-1/N} x(1-x) U_i(x) f(x)\,dx \tag{54}$$

where the limit indicates that the integration includes only the polymorphic region, *i.e.*, no point masses at the boundaries.

In contrast to the classical solution of Kimura [20], this solution also accounts for the dynamics at the boundaries. For the case of an initial Dirac delta distribution at a proportion p inside the polymorphic region, Tran *et al.* [23] derive the analogous eigenexpansion using the classical Gegenbauer polynomials (Equation (20) in [23]).

For real data, we do not know the true proportion p and rather have an estimate of p given a sample. Then polynomials are more useful as initial distributions, which will be illustrated with examples below.

3.1.1. Equilibrium of Mutations from the Boundaries and Drift; Outgroup Information

With information from a closely related species, *i.e.*, outgroup information, and small scaled mutation rates, polymorphic alleles can be polarized into ancestral (already present in the outgroup) or derived (not observed in the outgroup and thus arisen by a recent mutation). The requirements on the outgroup are strict. If the outgroup is too closely related to the focal species, polymorphism may still segregate in the outgroup (in a phylogenetic context this is known as "incomplete lineage sorting"). If the outgroup is too far from the focal species, fixed differences may have established or multiple mutations may have occurred. Both too close and too far relationships thus obscure polarization. Information from different, closely related species allows for better inference of ancestral states [24,25]. In practice, however, the monomorphic classes in the ingroup are usually combined.

Suppose again that time is scaled in units of N and that mutations arise from the ancestral state at a constant rate ϑ. Then the equilibrium distribution of proportions is ϑ/x, *i.e.*, inversely proportional to the distance from the origin (compare Formula (9.18) in [9]). Note that the integral from $1/N$ to $(N-1)/N$, *i.e.*, over the range for polymorphic alleles, is approximately:

$$\int_{1/N}^{1-1/N} \vartheta/x \, dx = \vartheta \log(x) \,|_{1/N}^{(N-1)/N} = \vartheta \log(N-1) \tag{55}$$

Representing the probability mass at the boundary zero with a delta function, we thus arrive at the following density:

$$\Pr(x \,|\, \theta, N) = \delta(x)(1 - \vartheta \log(N-1)) + \vartheta/x \tag{56}$$

For a sample of polymorphic alleles of size M, with y the number of derived alleles observed in the sample, assume again a binomial likelihood. Combining this likelihood with the density (56) and integrating out the allelic proportion x in the limit $N \to \infty$ results in:

$$\begin{aligned} \Pr(y \,|\, \vartheta, M) = \int_0^1 \binom{M}{y} x^y (1-x)^{M-y} \vartheta/x \, dx \\ = \vartheta \frac{M}{M-y} \end{aligned} \tag{57}$$

Let $L_0, \ldots, L_M$ represent the counts of derived alleles in the samples. With site frequency data and in equilibrium, all information is contained in the number of polymorphic alleles $L_p = \sum_{i=1}^{M-1} L_i$ as opposed to the monomorphic alleles. The likelihood of a polymorphic sample then becomes:

$$\Pr(L_p \,|\, \alpha, \theta, M, L) = \vartheta \frac{L!}{L_p!(L-L_p)!} (\vartheta \sum_{i=1}^{M-1} 1/i)^{L_p} (1 - \vartheta \sum_{i=1}^{M-1} 1/i)^{L-L_p} \tag{58}$$

The maximum likelihood estimator of ϑ is:

$$\hat{\vartheta} = \frac{L_p}{L \sum_{i=1}^{M-1} 1/i} \tag{59}$$

This estimator coincides with the Ewens–Watterson estimator [26,27]. It can also be derived using the Poisson Random Fields (PRF) approach [28].

While the similarities between Equations (36) and (57) are obvious, the underlying models are different; in particular, θ and ϑ are defined differently.

3.1.2. Equilibrium of Mutations from the Boundaries and Drift; No Outgroup Information

Assuming no outgroup information and equilibrium, *i.e.*, the same model as used for deriving Equation (36), but small scaled mutation rates and the other Poisson Random Field (PRF) assumptions, RoyChoudhury and Wakely [29] derive the distribution of polymorphic sites in a sample of L loci and M haploid individuals to be Poisson with mean:

$$2L\alpha\beta\theta\sum_{i=1}^{M-1}1/i \tag{60}$$

If we set $\vartheta = 2\alpha\beta\theta$, this leads to a maximum likelihood estimator of variability that is identical to that of Ewens and Watterson:

$$\hat{\vartheta} = \frac{L_p}{L\sum_{i=1}^{M-1}1/i} \tag{61}$$

The same estimator can also be derived without assuming a PRF, but instead expanding the likelihood (Equation (37)) into a power series in ϑ at zero, keeping only terms up to first order in ϑ [21]. The likelihood (Equation (37)) then becomes proportional to:

$$\Pr(L_0, L_p, L_M \,|\, \alpha, \vartheta, L, M) \propto (\beta - \frac{1}{2}\vartheta\sum_{i=1}^{M-1}\frac{1}{i})^{L_0}(\vartheta\sum_{i=1}^{M-1}\frac{1}{i})^{L_p}(\alpha - \frac{1}{2}\vartheta\sum_{i=1}^{M-1}\frac{1}{i})^{L_M} \tag{62}$$

With a model with mutations from both boundaries, the equilibrium density analogous to that in Equation (56) is [30]:

$$\Pr(x \,|\, \alpha\theta, N) = \delta(x)(\beta - \alpha\beta\theta\log(N)) + \frac{\vartheta}{x(1-x)} + \delta(x)(\alpha - \alpha\beta\theta\log(N)) \tag{63}$$

The joint distribution of this density and the binomial likelihood is for polymorphic alleles, *i.e.*, $1 \le y \le (M-1)$:

$$\Pr(x, y \,|\, \alpha\theta, M) = \vartheta\binom{M}{y}x^{y-1}(1-x)^{M-y-1} \tag{64}$$

Integrating this joint distribution over x in the limit of $N \to \infty$ also results in the marginal distribution Equation (62).

The two monomorphic classes L_0 and L_M may be combined to obtain a marginal likelihood, from which the same maximum likelihood estimator as in Equation (61) can be derived. As long as $\vartheta\sum_{i=1}^{M-1}\frac{1}{i} \ll 1$ (which is usually fulfilled), the Poisson approximation can be derived from the marginal likelihood of L_p given L, M, and ϑ [21]:

$$\begin{aligned}\Pr(L_p \,|\, \vartheta, L, M) &= \binom{L}{L_p}(\vartheta\sum_{i=1}^{M-1}\frac{1}{i})^{L_p}(1-\vartheta\sum_{i=1}^{M-1}\frac{1}{i})^{L-L_p} \\ &\approx \frac{(L\vartheta\sum_{i=1}^{M-1}\frac{1}{i})^{L_p}e^{-\vartheta\sum_{i=1}^{M-1}\frac{1}{i}}}{L_p!}\end{aligned} \tag{65}$$

Then $\frac{L_p}{H_M}$ is a maximum likelihood estimator for $L\vartheta$, which corresponds to the parameter θ of RoyChoudhury and Wakeley [29].

Only with small scaled mutation rates, maximum likelihood estimators of ϑ can be obtained relatively easily. With most real data, small scaled mutation rates, *i.e.*, $\vartheta < 0.0125$, are usually observed. This is also the parameter range, where use of outgroup data would enhance analyses, but the outgroups are never ideal. In fact, data will usually conform to a "joint frequency spectrum", where sample sizes in different populations or species may differ. If data from a second population come from a single diploid individual in Hardy–Weinberg equilibrium, the haploid sample size there will be two. Use of such data requires non-equilibrium approaches as in [31]. For small scaled mutation rates, a probabilistic model using orthogonal polynomials is formulated in [30].

3.1.3. Example for the Use of Gegenbauer Polynomials: Evolve and Resequence

As is obvious from the preceding subsections, samples from a single time point from a population assumed to be in an equilibrium of mutations from the boundaries and drift do not require orthogonal polynomials. To demonstrate the use of orthogonal polynomials, data that might occur in an "Evolve and Resequence" experiment, e.g., [32], will be analyzed in this subsection. Assume that a base population of, e.g., $N = 200$ fruit flies (*Drosophila melanogaster*) is taken from a wild population that is assumed to be in equilibrium. In a cage, the population evolves without selection for t/N generations. Within the short times customary in such studies, mutations are unlikely and can be ignored. At a certain locus, the initial sample size from the base population is $M = 5$; $y = 3$ alleles are of the first allelic type. Conditional on the allelic proportion x, the likelihood is binomial and the joint distribution is given in Equation (35):

$$\begin{aligned}\Pr(y = 3, x \,|\, M = 5, \alpha, \theta) &= \alpha\beta\theta \binom{3}{2} x^3(1-x)^2\, x^{-1}(1-x)^{-1} \\ &= 3\,\alpha\beta\theta\, x^2(1-x)\end{aligned} \tag{66}$$

This is actually a polynomial of degree three and proportional to a $\mathrm{beta}(x \,|\, 3, 2)$ density, which can be represented exactly by a series of the modified Gegenbauer polynomials H_i up to the appropriate degree. The loss of variation from a $\mathrm{beta}(x \,|\, 3, 2)$ distribution within the polymorphic region over t/N generations is shown in Figure 1.

Consider two time points and an even smaller sample size that allows calculation in the text. In particular, assume that the sample size of the initial sample is $M_0 = 3$ with two alleles of the first type $y_0 = 2$. Thus the joint distribution of the sample y_0 and the allelic proportions x is:

$$\begin{aligned}\Pr(y_0 = 2, x \,|\, M_0 = 3, \alpha, \theta, t = 0) &= \alpha\beta\theta \binom{3}{2} x^2(1-x)\, x^{-1}(1-x)^{-1} \\ &= 3\,\alpha\beta\theta\, x\end{aligned} \tag{67}$$

This polynomial can be represented by the modified Gegenbauer polynomials of degree up to one: $c_1 = -\frac{3}{4}\alpha\beta\theta$ and $c_0 = -\frac{3}{2}\alpha\beta\theta$. At time t_1, before considering the second sample, the probability mass of the joint density has diminished in the polymorphic region:

$$\Pr(y_0 = 2, 0 < x < 1 \,|\, M_0 = 3, \alpha, \theta, t = t_1) = \alpha\beta\theta \left(-\frac{3}{2}\, e^{-2t_1}(-1) - \frac{3}{4}\, e^{-6t_1}(2-4x) \right) \tag{68}$$

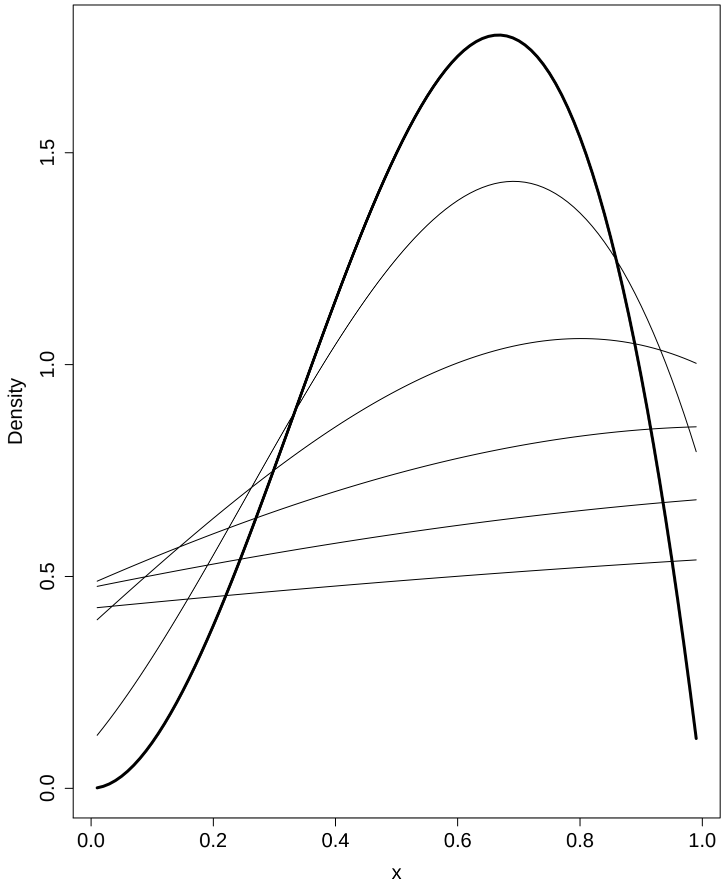
while it has grown at the boundaries:

$$\Pr(y_0 = 2, x = 0 \,|\, M_0 = 3, \alpha, \theta, t = t_1) = \alpha\beta\theta \left(\frac{3}{2} \cdot \frac{1}{2} (1 - e^{-2t_1}) - \frac{3}{4} \cdot \frac{1}{3} (1 - e^{-6t_1}) \right) \tag{69}$$

and

$$\Pr(y_0 = 2, x = 1 \,|\, M_0 = 3, \alpha, \theta, t = t_1) = \alpha\beta\theta \left(\frac{3}{4} (1 - e^{-2t_1}) + \frac{1}{4} (1 - e^{-6t_1}) \right) \tag{70}$$

Figure 1. Distribution of the allelic proportion x starting from a $dbeta(x \,|\, 3, 2)$ distribution (thick line). The thin lines represent the loss of variation through genetic drift at generations $t/N = (0.05, 0.15, 0.25, 0.35, 0.45)$.



The likelihood of a second sample at time t_1 of size $M_1 = 3$ with $y_1 = 3$ alleles of the first type, *i.e.*, a monomorphic sample, is binomial: x^3. The joint probability consists of an interior and a boundary part. From the interior part of the joint distribution, x can be integrated out:

$$\begin{aligned} &\Pr(y_0 = 2, y_1 = 3 \,|\, M_0 = 3, M_1 = 3, \alpha, \theta, t = t_1, 0 < x < 1) = \\ &\alpha\beta\theta \int_0^1 x^3 \left(\frac{3}{2} e^{-2t_1} + \frac{3}{4} e^{-6t_1} (-2 + 4x) \right) dx = \\ &\alpha\beta\theta \left(\frac{3}{8} e^{-2t_1} - \frac{3}{8} e^{-6t_1} + \frac{3}{5} e^{-6t_1} \right) \end{aligned} \tag{71}$$

Summing the interior and the boundary parts, the likelihood of the two samples is obtained:

$$\begin{aligned} &\Pr(y_0 = 2, y_1 = 3 \,|\, M_0 = 3, M_1 = 3, \alpha, \theta, t = t_1) = \\ &\alpha\beta\theta \left(\frac{3}{8}\, e^{-2t_1} - \frac{3}{8}\, e^{-6t_1} + \frac{3}{5}\, e^{-6t_1} + \frac{3}{4}\,(1 - e^{-2t_1}) + \frac{1}{4}\,(1 - e^{-6t_1}) \right) \end{aligned} \tag{72}$$

Note that the parameters $\alpha\beta\theta$ pertain to the base population. With respect to drift during the experiment, the single parameter in this likelihood is t_1, *i.e.*, the time in generations normed by the (effective) population size. Usually, the number of generations is known in "evolve and resequence" studies, such that the (effective) population size N can be estimated. This may or may not coincide with the census population size, which is also usually known.

Note that the above computation is much simpler than the use of the transition probabilities of the Wright–Fisher model with the effective population size N as a parameter [33,34]. As the data at the time that those articles were published (about 2000) were usually microsatellites, rather than single-nucleotide polymorphisms, these methods provide for multiple alleles.

Using the statistical language "R" [35] with the high-precision algebra package "Rmpfr", likelihoods for sample sizes of about 50 can be calculated within minutes with this method.

3.2. Selection and Drift

In the following subsection, the approach of Song and Steinrücken [12] is followed (their section: "Diffusion with Genetic Selection and No Mutation"). While the numerical methods of these authors are less advanced than those implemented in the commercial packages (see the Appendix), their general method based on the modified Jacobi polynomials is also applicable in cases with mutation and dominance.

Our goal is to find eigenfunctions $V_i(x)$ and the associated eigenvectors Λ_i of the backward generator:

$$\mathcal{L}_b(x)V_i(x) = \left(\gamma x(1-x)\frac{d}{dx} + x(1-x)\frac{d^2}{dx^2} \right) V_i(x) = \Lambda_i V_i(x) \tag{73}$$

The $V_i(x)$ are orthogonal with respect to the weight function $w(x) = e^{\gamma x} x^{-1}(1-x)^{-1}$, such that:

$$\int_0^1 V_i(x)V_j(x)w(x)\, dx \propto \delta_{i,j} \tag{74}$$

Substituting $K_i(x) = e^{-\frac{\gamma}{2}x} G_i(x)$ into this equation, it can be verified that the $K_i(x)$ are also orthogonal with respect to the same weight function:

$$\int_0^1 V_i(x)V_j(x)w(x)\, dx = K_i(x)K_j(x)x^{-1}(1-x)^{-1}\, dx = \delta_{i,j}\frac{i+1}{(i+2)(i+3)} \tag{75}$$

Therefore, even though the $K_i(x)$ are not eigenfunctions of the backward generator $\mathcal{L}_b(x)$, linear combinations of the $K_i(x)$ can be used to represent $V_j(x)$:

$$V_j(x) = \sum_{i=0}^{\infty} u_{j,i} K_i(x) \tag{76}$$

where the $u_{j,i}$ are constants to be determined. Substituting K_i into the backward operator results in:

$$\begin{aligned}\mathcal{L}_b K_i(x) &= e^{-\frac{\gamma}{2}x}\left(x(1-x)\frac{d^2}{dx^2}G_i(x) - \tfrac{\gamma^2}{4}x(1-x)G_i(x)\right)\\ &= -e^{-\frac{\gamma}{2}x}\left(\lambda_i G_i(x) + \tfrac{\gamma^2}{4}x(1-x)G_i(x)\right)\end{aligned} \tag{77}$$

Using this result together with Equations (73) and (76) leads to:

$$\sum_{i=0}^{\infty} u_{j,i}\left(\lambda_i G_i(x) + \tfrac{\gamma^2}{4}x(1-x)G_i(x)\right) = \Lambda_i \sum_{i=0}^{\infty} u_{j,i}G_i(x) \tag{78}$$

For $i \geq 0$, with the recurrence relation Equation (47), one can show that:

$$\tfrac{\gamma^2}{4}x(1-x)G_i(x) = a_i^{(-2)}G_{i-2}(x) + a_i^{(0)}G_i(x) + a_i^{(+2)}G_{i+2}(x) \tag{79}$$

where:

$$\begin{cases} a_i^{(-2)} = -\frac{\gamma^2}{16}\frac{i(i+1)}{(2i+1)(2i+3)}\text{, for } i \geq 2\text{, otherwise } 0\\ a_i^{(0)} = \frac{\gamma^2}{8}\frac{(i+1)(i+2)}{(2i+1)(2i+5)}\\ a_i^{(+2)} = -\frac{\gamma^2}{16}\frac{(i+1)(i+4)}{(2i+3)(2i+5)}\,. \end{cases} \tag{80}$$

For $j \geq 0$, multiplying this system of equations with $G_j(x)$ and integrating with respect to the weight function $x^{-1}(1-x)^{-1}$ yields a system of equations. In matrix form, this system can be written as:

$$\begin{pmatrix} \lambda_0 + a_0^{(0)} & 0 & a_2^{(-2)} & 0 & 0 & \cdots\\ 0 & \lambda_1 + a_1^{(0)} & 0 & a_3^{(-2)} & 0 & \cdots\\ a_0^{(+2)} & 0 & \lambda_2 + a_2^{(0)} & 0 & a_4^{(-2)} & \cdots\\ 0 & a_1^{(+2)} & 0 & \lambda_3 + a_3^{(0)} & 0 & \cdots\\ 0 & 0 & a_2^{(+2)} & 0 & \lambda_4 + a_4^{(0)} & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}\begin{pmatrix} u_{j,0}\\ u_{j,1}\\ u_{j,2}\\ u_{j,3}\\ u_{j,4}\\ \vdots \end{pmatrix} = \Lambda_j \begin{pmatrix} u_{j,0}\\ u_{j,1}\\ u_{j,2}\\ u_{j,3}\\ u_{j,4}\\ \vdots \end{pmatrix} \tag{81}$$

The eigenvalues Λ_j correspond to the eigenvalues of the operator $\mathcal{L}_b$ with the associated eigenvectors $u_j = (u_{j,0}, u_{j,1}, u_{j,2}, u_{j,3}, u_{j,4}, \ldots)$. Note that this band-diagonal system can be subdivided into two independent tridiagonal systems for even and odd i and j. While this system has infinite size, it can be truncated at dimension D to obtain an approximation, with little loss [12]. The eigenvalues of tridiagonal matrices with real coefficients can be obtained relatively quickly. Furthermore, approximate solutions to the eigenvalues can be improved using continued fractions [36]. Nevertheless, so far the only implementation seems to be that in [12], where the backward equation is considered. The eigenvectors of the forward equation can be obtained by multiplication with the weight function $e^{\gamma x}x^{-1}(1-x)^{-1}$. A solution of the forward diffusion equation has not been implemented yet, as far as the author is aware of.

4. Conclusions

A biallelic locus subject to the population genetic forces such as mutation, drift and selection can be modeled using diffusion equations. These diffusion equations can be solved using orthogonal polynomials; the case of pure drift using the Gegenbauer polynomials, the case of mutation and drift using Jacobi polynomials, and the case of selection and drift using spheroidal wave functions. The theory for using series of orthogonal polynomials to solve the corresponding diffusion equations has been elaborated in detail. By adjusting the coefficients in the expansion any initial distribution may be approximated. In genomic regions of relatively high recombination rates and low mutation rates, each polymorphic nucleotide can be assumed to evolve independently. Samples from such regions have been called site frequency spectra. Assuming equilibrium, the joint distribution of the allelic proportion x and the data y of each such site can be modeled as a linear combination of eigenvectors of the forward equation up to an order determined by the sample size. With this, it is thus possible to condition on samples from two time points, as with ancient DNA or "evolve and resequence" studies, or use outgroup information.

A major advantage of using diffusion equations and orthogonal polynomials over competing methods, e.g., approximate Bayesian computation [37], which uses summary statistics, or even alternative methods based on the solution of diffusion equations [31], is that the relevant distributions may be calculated exactly and without loss of information, or can at least be approximated very efficiently. Furthermore, the well-developed theory of orthogonal polynomials connects population genetics to other more advanced disciplines, e.g., theoretical physics.

Acknowledgments

The author wishes to thank the other members of the doctorate college "Population Genetics" for discussions and comments and Juraj Bergmann for critically reading the manuscript. The comments of Joshua Schraiber and an anonymous reviewer helped improve the manuscript.

Appendix. The Oblate Spheroidal Wave Function

In this appendix, the oblate spheroidal wave function is used to solve a directional selection-drift model. The Kolmogorov forward equation is parametrized as follows:

$$\frac{\partial \phi(x,\tau)}{\partial \tau} = \frac{\partial^2}{\partial x^2}\left(x(1-x)\phi(x,\tau)\right) - 4\gamma \frac{\partial}{\partial x}\left(x(1-x)\phi(x,\tau)\right) \tag{A1}$$

where γ is the scaled directional selection coefficient. This equation can be transformed to the spheroidal wave equation, to which a lot of research has been dedicated from about the time of Kimura's work until now [36,38–41].

We will transform the scaled forward Equation (A1) to the Sturm–Liouville form (specifically to the oblate spheroidal wave equation with $m = 1$). Initially,

$$\phi(t,x) = e^{-\lambda t} e^{2\gamma x} v(x) \tag{A2}$$

is substituted into the scaled forward Equation (A1) to obtain:

$$x(1-x)\frac{d^2 v(x)}{dx^2} + 2(1-2x)\frac{dv(x)}{dx} - \left(2 + 4\gamma^2 x(1-x) - \lambda\right) v(x) = 0 \tag{A3}$$

Setting $x = (1-z)/2$ (such that $(-2\frac{\partial}{\partial z}) = \frac{\partial}{\partial x}$, $x(1-x) = (1-z^2)/4$; the boundaries are then -1 and 1), the next equation, used by Kimura [20], is obtained:

$$(1-z^2)\frac{d^2 v((1-z)/2)}{dz^2} - 4z\frac{dv((1-z)/2)}{dz} + \left(\lambda - 2 - \gamma^2(1-z^2)\right) v((1-z)/2) = 0 \tag{A4}$$

It can be transformed to the Sturm–Liouville form by setting $g(z)(1-z^2)^{-1/2} = v((1-z)/2)$, since:

$$\begin{aligned}\frac{d}{dx}(1-z^2)^{-1/2} &= z(1-z^2)^{-3/2}\\ \frac{d^2}{dx^2}(1-z^2)^{-1/2} &= (1-z^2)^{-3/2} + 3z^2(1-z^2)^{-5/2}\end{aligned} \tag{A5}$$

Whence,

$$\begin{aligned}
0 &= (1-z^2)\frac{d^2v((1-z)/2)}{dz^2} - 4z\frac{dv((1-z)/2)}{dz} + \left(\lambda - 2 - \gamma^2(1-z^2)\right) v((1-z)/2)\\
0 &= (1-z^2)\frac{d^2(1-z^2)^{-1/2}g(z)}{dz^2} - 4z\frac{d(1-z^2)^{-1/2}g(z)}{dz}\\
&\quad + \left(\lambda - 2 - \gamma^2(1-z^2)\right)(1-z^2)^{-1/2}g(z)\\
0 &= (1-z^2)(1-z^2)^{-1/2}\frac{d^2g(z)}{dz^2} + (1-z^2)2z(1-z^2)^{-3/2}\frac{dg(z)}{dz}\\
&\quad + (1-z^2)\left((1-z^2)^{-3/2} + 3z^2(1-z^2)^{-5/2}\right)g(z)\\
&\quad - 4z(1-z^2)^{-1/2}\frac{dg(z)}{dz} - 4z^2(1-z^2)^{-3/2}g(z)\\
&\quad + \left(\lambda - 2 - \gamma^2(1-z^2)\right)(1-z^2)^{-1/2}g(z)\\
0 &= (1-z^2)\frac{d^2g(z)}{dz^2} + 2z\frac{dg(z)}{d}z + (1 + 3z^2(1-z^2)^{-1})g(z)\\
&\quad - 4z\frac{dg(z)}{dz} - 4z^2(1-z^2)^{-1}g(z) + \left(\lambda - 2 - \gamma^2(1-z^2)\right)g(z)\\
0 &= \frac{d}{dz}\left((1-z^2)\frac{dg(z)}{dz}\right) + \left(\lambda - 1 - \frac{z^2}{1-z^2} - \gamma^2(1-z^2)\right)g(z)\\
0 &= \frac{d}{dz}\left((1-z^2)\frac{dg(z)}{dz}\right) + \left(\lambda - \frac{1-z^2}{1-z^2} - \frac{z^2}{1-z^2} - \gamma^2(1-z^2)\right)g(z)\\
0 &= \frac{d}{dz}\left((1-z^2)\frac{dg(z)}{dz}\right) + \left(\lambda - \gamma^2(1-z^2) - \frac{1}{1-z^2}\right)g(z)
\end{aligned} \tag{A6}$$

The last line is in Sturm–Liouville form (see Equation (9)). It also corresponds to

$$0 = \frac{d}{dz}\left((1-z^2)\frac{dg(z)}{dz}\right) + \left(\lambda_n^m(\gamma) + c^2(1-z^2) - \frac{m^2}{1-z^2}\right)g(z) \tag{A7}$$

which is generally used for spheroidal wave functions ([19], Chapter 21). As can be seen from Equation (9), the weight function is unity, such that the forward and backward equations are identical. The condition $c^2 < 0$ actually defines the oblate spheroidal wave functions. For $c^2 = 0$, corresponding to the case without selection, Equation (A7) reduces to the differential equation of the associated Legendre function ([19], Chapter 8):

$$0 = \frac{d}{dz}\left((1-z^2)\frac{dg(z)}{dz}\right) + \left(l(l+1) - \frac{m^2}{1-z^2}\right)g(z) \tag{A8}$$

While the spheroidal wave functions and the associated Legendre functions solving the above equations are not strictly polynomials, much of the theory of orthogonal polynomials also applies to them, such that any initial function can be approximated by a series of Legendre functions.

Importantly, the computation of spheroidal wave functions has been advanced relatively recently and implemented in commercially available computer packages [36,41]. Computation is also based on a similar band-diagonal system of equations as in Equation (81). The formula manipulation program "Mathematica" [10] defines the spheroidal wave function slightly differently from above [41]:

$$\frac{d}{dz}\left((1-z^2)\frac{dS_{mn}(z)}{dz}\right)+\left(\lambda_{mn}-c^2z^2-\frac{m^2}{1-z^2}\right)S_{mn}(z)=0 \tag{A9}$$

Set

$$L_2S_{mn}=\frac{d}{dz}\left((1-z^2)\frac{dS_{mn}(z)}{dz}\right)+\left(-c^2z^2-\frac{m^2}{1-z^2}\right)S_{mn}(z) \tag{A10}$$

while the original operator $L_1 = L_2 + c^2$. The eigenvalues and eigenvectors are then:

$$\begin{aligned} L_2S_{mn} &= \lambda_{mn}S_{mn} \\ (L_2+c^2)S_{mn} &= (\lambda_{mn}+c^2)S_{mn} \\ L_1S_{mn} &= (\lambda_{mn}+c^2)S_{mn} \end{aligned} \tag{A11}$$

From this, we see that the eigenvectors are identical and the eigenvalues differ by c^2.

The Mathematica package "Spheroidal.m" [42] also defines the spheroidal wave equations, this time with the first operator L_1. Packages are also available for "Maple" [11]. As far as the author is aware of, these tools are the only ones currently available to compute and visualize directional forward and backward selection-drift diffusion models relatively easily.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Parsch, J.; Novozhilov, S.; Saminadin-Peter, S.; Wong, K.; Andolfatto, P. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila. Mol. Biol. Evol.* **2010**, *27*, 1226–1234.
2. Fisher, R. *The Genetical Theory of Natural Selection*; Clarendon Press: Oxford, UK, 1930.
3. Wright, S. Evolution in Mendelian populations. *Genetics* **1931**, *16*, 97–159.
4. Moran, P. Random processes in genetics. *Proc. Camb. Philos. Soc.* **1958**, *54*, 60–71.
5. Kimura, M. *The Neutral Theory of Molecular Evolution*; Cambridge University Press: Cambridge, UK, 1983.
6. Kimura, M. *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers*; University of Chicago Press: Chicago, IL, USA, 1994.
7. Evans, S.; Shvets, Y.; Slatkin, M. Non-equilibrium theory of the allele frequency spectrum. *Theor. Popul. Biol.* **2007**, *71*, 109–119.
8. Zivkovic, D.; Stephan, W. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theor. Popul. Biol.* **2011**, *79*, 184–191.
9. Ewens, W. *Mathematical Population Genetics*, 2nd ed.; Springer: New York, NY, USA, 2004.

10. Wolfram Research, Inc., Mathematica, Version 10.0, Champaign, IL, USA, 2014. Available online: http://wolfram.com/ (accessed on 6 November 2014).
11. Matlab 8.4, The MathWorks Inc., Natick, MA, USA, 2014. Available online: http://www.mathworks.de/ (accessed on 6 November 2014).
12. Song, Y.; Steinrücken, M. A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics* **2012**, *190*, 1117–1129.
13. Baake, E.; Bialowons, R. *Ancestral Processes with Selection: Branching and Moran Models*; Banach Center Publications: Bielefeld, Germany, 2008; Volume 80, pp. 33–52.
14. Etheridge, A.; Griffiths, R. A coalescent dual process in a Moran model with genic selectio. *Theor. Popul. Biol.* **2009**, *75*, 320–330.
15. Vogl, C.; Clemente, F. The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates. *Theor. Popul. Genet.* **2012**, *81*, 197–209.
16. Hein, J.; Schierup, M.; Wiuf, C. *Gene Genealogies, Variation, and Evolution: A Primer in Coalescent Theory*; Oxford University Press: Oxford, UK, 2005.
17. Hazewinkel, M. Sturm-Liouville Theory. In *Encyclopedia of Mathematics*; Springer: New York, NY, USA, 2001.
18. Griffiths, R.; Spanò, D. Diffusion processes and coalescent trees. In *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*; Cambridge University Press: Cambridge, UK, 2010; pp. 358–375.
19. *Handbook of Mathematical Functions*, 9th ed.; Abramowitz, M., Stegun, I., Eds.; Dover: New York, NY, USA, 1970.
20. Kimura, M. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* **1955**, *41*, 144–150.
21. Vogl, C. Estimating the Scaled Mutation Rate and Mutation Bias with Site Frequency Data. *Theor. Popul. Biol.* **2014**, in press.
22. McKane, A.; Waxman, D. Singular solutions of the diffusion equation of population genetics. *J. Theor. Biol.* **2007**, *247*, 849–858.
23. Tran, T.; Hofrichter, J.; Jost, J. An introduction to the mathematical structure of the Wright-Fisher model of population genetics. *Theory Biosci.* **2013**, *132*, 73–82.
24. Clemente, F.; Vogl, C. Unconstrained evolution in short introns?—An analysis of genome-wide polymorphism and divergence data from *Drosophila*. *J. Evol. Biol.* **2012**, *25*, 1975–1990.
25. Clemente, F.; Vogl, C. Evidence for complex selection on four-fold degenerate sites in *Drosophila melanogaster*. *J. Evol. Biol.* **2012**, *25*, 2582–2595.
26. Ewens, W. A note on the sampling theory for infinite alleles and infinite sites models. *Theor. Popul. Biol.* **1974**, *6*, 143–148.
27. Watterson, G. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **1975**, *7*, 256–276.
28. Sawyer, S.; Hartl, D. Population genetics of polymorphism and divergence. *Genetics* **1992**, *132*, 1161–1176.

29. RoyChoudhury, A.; Wakeley, J. Sufficiency of the number of segregating sites in the limit under finite-sites mutation. *Theor. Popul. Biol.* **2010**, *78*, 118–122.
30. Vogl, C. Biallelic Mutation-Drift Diffusion in the Limit of Small Scaled Mutation Rates. *ArXiv E-Prints* **2014**, arXiv:1409.2299.
31. Gutenkunst, R.; Hernandez, R.; Williamson, S.; Bustamante, C. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet.* **2009**, *5*, e1000695.
32. Tobler, R.; Franssen, S.; Kofler, R.; Orozco-Terwengel, P.; Nolte, V.; Hermisson, J.; Schlötterer, C. Massive habitat-specific genomic response in D. melanogaster populations during experimental evolution in hot and cold environments. *Mol. Biol. Evol.* **2014**, *31*, 364–375.
33. Williamson, E.; Slatkin, M. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* **1999**, *152*, 755–761.
34. Anderson, E.; Williamson, E.; Thompson, E. Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics* **2000**, *156*, 2109–2118.
35. R Core Team. *R: A Language and Environment for Statistical Computing*; ISBN 3-900051-07-0; R Foundation for Statistical Computing: Vienna, Austria, 2013.
36. Falloon, P.; Abbott, P.; Wang, J. Theory and computation of spheroidal wave functions. *J. Phys. A Math. Gen.* **2003**, *36*, 5477–5495.
37. Beaumont, M.; Zhang, W.; Balding, D. Approximate Bayesian Computation in Population Genetic. *Genetics* **2002**, *162*, 2025–2035.
38. Stratton, J. *Spheroidal Wave Functions*; The Technology Press of the Massachusetts Institute of Technology: Cambridge, MA, USA, 1954.
39. Meixner, J.; Schäfke, F. *Mathieusche Funktionen und Sphäroidfunktionen*; Springer: Berlin, Germany, 1954. (In German)
40. Flammer, C. *Spheroidal Wave Functions*; Stanford University Press: Palo Alto, CA, USA, 1957.
41. Li, L.W.; Leong, M.S.; Yeo, T.S.; Kooi, P.S.; Tan, K.Y. Computations of spheroidal harmonics with complex arguments: A review with an algorithm. *Phys. Rev. E* **1998**, *58*, 6792–6806.
42. Falloon, P.E. Theory and Computation of Spheroidal Harmonics with General Arguments. Master's Thesis, Department of Physics, The University of Western Australia, Crawley, Australia, 2001.