Reconstruction of Probability Distributions

in Population Genetics

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

in Statistics

The University of Auckland, 2012

Jing Liu

Department of Statistics

The University of Auckland

New Zealand

Abstract

For a range of models in population genetics, we demonstrate that moments of the stationary distribution can be obtained without knowing the stationary distribution itself, using the diffusion approximation. We introduce the maximum entropy principle to use these acquired moments to reconstruct the density of the stationary distribution. This procedure is illustrated by reconstructing the stationary distribution for a two-locus model with linkage and recurrent mutation. Using the reconstructed stationary distribution, the mean and the variance of a linkage disequilibrium measure r^2 are evaluated for the model.

We then propose a novel method for reconstructing unknown distributions analytically based on the maximum entropy principle. Given a sequence of moments expressed in terms of the underlying population parameter, this new method offers a likelihood function for parameter estimation by expressing the density of observable quantities as an explicit function of the data and the parameter.

Acknowledgements

I would like to gratefully and sincerely thank Rachel Fewster and Yong Wang for their supervision and guidance, Stephen Cope for his help in the area of parallel computing, Lisa Chen and Benny Zhu for their useful discussions, Ben Stevenson for proofreading, and my department for providing its support. Funding for this project was provided by The University of Auckland and The New Zealand Institute of Mathematics and its Applications.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivation

This PhD project was motivated by a recent development in the use of the diffusion approximation. **Song and Song (2007)** proposed an elegant procedure to compute the expectation of a common linkage disequilibrium (LD) measure r^2 at steady state. The LD measure r^2 is defined as

$$r^2 = \frac{D^2}{p(1-p)q(1-q)}\,, \tag{1.1}$$

where p and q are frequencies of alleles A_1 and B_1 at diallelic loci labelled A and B respectively, and $D = p_1 - p_1^2 - p_1p_2 - p_1p_3 - p_2p_3$ is the usual coefficient of linkage disequilibrium, with p_1, p_2 and p_3 being the frequencies of gamete-types A_1B_1, A_1B_2, and A_2B_1.

In **Song and Song (2007)**'s paper, they describe a method of finding relevant moments of the stationary distribution of $(p,\ q,\ D)$ using the diffusion approximation, without first finding the stationary distribution itself. Their work generated a lot of interest, both in the results for $\mathbb{E}\left(r^2\right)$ that they obtained and in the elegance of the method that they used.

This PhD project was initially intended to investigate estimation of effective population size in extant populations, but it assumed a life of its own. We were initially interested in obtaining reliable estimates of effective population size N_e to aid conservation of endangered species. It was thought that the poor performance of a common estimator for N_e was a direct consequence of an ill-approximated functional link between $\mathbb{E}\left(r^2\right)$ and N_e. Therefore, this PhD project initially planned to study the method of diffusion approximation in population genetics, in particular the method proposed by **Song and Song (2007)**. We aimed to extend their ideas in order to derive a better functional link between $\mathbb{E}\left(r^2\right)$ and N_e, using genotype data instead of the gametic data required by **Song and Song (2007)**, which is not commonly available in the conservation context. Ultimately, we aimed to propose a better estimator for N_e.

During the early stages of our research we identified that the poor performance of the current estimator for N_e is largely due to the sample size being too small relative to the population size (**Russell and Fewster, 2009**), and not to the link between $\mathbb{E}(r^2)$ and N_e as originally thought. However, studying **Song and Song (2007)**'s general approach became a goal in itself, and we became interested in the possibilities for extending or generalising their method.

1.2 Purpose

We were largely interested in the demonstration by **Song and Song (2007)** that information contained in a finite sequence of moments from an unknown distribution can be used to compute other properties of the distribution, $\mathbb{E}(r^2)$ in their case. Their method is very specific for calculating $\mathbb{E}(r^2)$ from a particular model, and it cannot easily be extended to evaluate other expectations which do not resemble $\mathbb{E}(r^2)$ in a linear way. For example, the variance $\mathbb{V}(r^2)$ cannot readily be evaluated using their method. We decided to investigate whether we could develop a method that uses the information contained in a finite sequence of moments of an unknown distribution to evaluate any expectation or other property of that distribution.

The original Wright-Fisher model in population genetics is ultimately guaranteed to reach fixation, hence the primary interests are in the probability of extinction (or fixation) of a certain allele, the expected time to this extinction (or fixation), and the rate of loss of genetic variability. However, for other Wright-Fisher-type models which exhibit a steady state other than fixation, the interest centres on the stationary distribution of allele frequencies.

It is many years since the stationary distribution for *single locus* models was first found using the diffusion approximation. However, to the best of our knowledge, there is no general approach to the present day for obtaining the stationary distribution for *multiple locus* models with recombination. Our aim is to determine a general method for reconstructing an unknown distribution from a finite sequence of its moments. Using such a method, the stationary distribution can be reconstructed for the model in **Song and Song (2007)**, hence the variance of the LD measure $\mathbb{V}\left(r^2\right)$ as well as its mean $\mathbb{E}\left(r^2\right)$ can be evaluated and studied in terms of recombination and mutation. We will only consider the original Wright-Fisher model and its generalisations under ideal breeding conditions, so we will not make any distinction between effective and census population size in this thesis.

1.3 Aims

The moments found by **Song and Song (2007)** using the diffusion approximation are given in terms of p, q and D. The parametrisation p, q and D is obtained by transforming the original diffusion approximation which is specified in terms of gametic frequencies (p_1, p_2 and p_3). The parametrisation p, q and D is convenient to work with when we are dealing with D. However, it is not convenient when we are reconstructing the distribution, because p, q and D have an irregular support set. Reconstructing a distribution on a regular region, such as that underlying p_1, p_2 and p_3, is much simpler.

Therefore we first need to study the diffusion approximation, which makes it possible to find moments without first finding the stationary distribution. Then we need to find the stationary moments in the original parametrisation (p_1, p_2 and p_3). Finally, we need to find or derive a method of reconstructing the stationary distribution using the information contained in the stationary moments.

1.4 Outline

This thesis is organised as follows. In **Chapter 2**, we discuss some stochastic models in population genetics and show some preliminary results in terms of finding their stationary distributions. We provide a review of the diffusion approximation and an example derivation of a diffusion operator in **Chapter 3**. In **Chapter 4**, we first derive the moments needed for a two-locus model with linkage, using the diffusion approximation. We then introduce the maximum entropy principle and the traditional procedure of applying it. We show that the maximum entropy principle enables us to replicate **Song and Song (2007)**'s results for $\mathbb{E}\left(r^2\right)$, and supply new results for $\mathbb{V}\left(r^2\right)$, up to a limit of computational tractability. At the end of **Chapter 4**, we propose a novel method of analytically reconstructing an unknown distribution if a finite sequence of its analytic moments is available. We complete the thesis with some discussion of the new method and some suggestions for future work in **Chapters 5** and **6**.

2

Preliminaries: Wright-Fisher Models

In this chapter, we first discuss one of the most important stochastic models for genetic drift in population genetics, namely the Wright-Fisher model. Later, we describe some standard generalisations of it to incorporate mutation, selection, and recombination as well as genetic drift. Migration will not be explicitly mentioned, however it can be easily incorporated by modelling it in a similar fashion to mutation. The primary reason for the discussion of these models is to give notation and state some relevant results for later usage. These models are standard and the relevant results are well known; we do not introduce anything novel here. In the second half of this chapter, we show preliminary results on some direct methods of finding the stationary distribution for generalised Wright-Fisher models.

2.1 Models and Notation

2.1.1 Wright-Fisher model

We consider three stochastic models of reproduction, all of which are standard generalisations of the original Wright-Fisher model. The original Wright-Fisher model describes the process of genetic drift in a finite population and it is the most popular stochastic model for reproduction in population genetics, despite it being highly idealised. See standard textbooks such as **Crow and Kimura (1970)**, **Nei (1987)** and **Ewens (2004)** for a detailed discussion of stochastic models in population genetics. In this section, we provide only essential details pertaining to both the original Wright-Fisher model and its generalisations, in which some assumptions are relaxed.

The original Wright-Fisher model was used implicitly by **Fisher (1930)** and explicitly by **Wright (1931)** to model genetic drift. Biologically, the original Wright-Fisher model considers a monoecious diploid population with N individuals in an isolated colony. Every individual effectively has an infinite capacity to produce gametes, and each has an equal chance of contributing successful gametes to the next generation. Conceptually, an infinite number of juveniles are considered to be produced for every generation, but only N juveniles are kept in the colony to populate the next generation. Therefore, at any given time there are N individuals in the colony. The original Wright-Fisher model assumes neutrality (no selective differences between alleles) and no mutation. Its key assumptions are summarised in **Table 2.1**, some of which are far from realistic. Despite this, even in its original form, the Wright-Fisher model succeeds in capturing the essence of genetic inheritance.

Table 2.1: Assumptions of the Wright-Fisher Model.

1. Diploid
2. Monoecious reproduction with an infinite number of gametes
3. Non-overlapping generations
4. Random mating
5. Finite and constant population size
6. No selection
7. No mutation

Diploid means that the organism has two matching sets of chromosomes, and so they possess two alleles at every locus. *Monoecious* means that the organism can produce zygotes with any individual of the population; there is no clear female/male distinction during reproduction, and the two gametes which form the zygote may even be from the same individual.

Mathematically, although Fisher and Wright did not use this terminology, the stochastic process they defined through this model is a Markov chain; the transition probability of genetic composition changing from the current generation t to the next generation $t+1$ does *not* depend on the changes made in previous generations ($t-1$, $t-2$, ..., 2, 1, 0). In the simplest case, given there exist only two alleles (A_1 and A_2, say) at locus A, the transition probability of going from x copies of A_1 at generation t to y^* copies of A_1 at generation $t+1$ is given by the binomial probability distribution $Y^* \mid X \sim \text{Bin}\,(2N, \varphi^*(x))$, where $\varphi^*(x)$ is the expected proportion of A_1 in the generation $t+1$. Let $\boldsymbol{P^*}$ denote the transition matrix of probabilities of going from state x to state y^* in one generation.

The superscript $*$ is used for the quantities that do not involve mutation. This is to distinguish them from similar quantities for models that do involve mutation. This notation will be used consistently throughout the thesis unless specified otherwise.

Given that x is the number of copies of A_1 in the current generation, the expected proportion of A_1 in the next generation is equal to the number of copies of A_1 in the current generation divided by the total number of gametes. Hence $\varphi^*(x)$ is given by,

$$\varphi^*(x) = \frac{x}{2N}\,. \tag{2.1}$$

Equation 2.1 is a direct result of assumptions 2 and 4 in **Table 2.1**; it can be considered that gametes are sampled randomly and independently with replacement to form the zygotes at the time of conception under these two assumptions. These are the essential assumptions of the Wright-Fisher model, which we assume without relaxing throughout the entire thesis. These two assumptions guarantee that the transition matrix is given by binomial probability mass functions with an appropriate expected proportion, or multinomial probability mass functions with an appropriate vector of expected proportions. Therefore, working out the expected proportion, or the vector of expected proportions, gives us the transition matrix for the corresponding model that we are to consider.

For the original Wright-Fisher model, the transition matrix $\boldsymbol{P}^*$ is given by,

$$\begin{aligned} p^*_{xy} &= \mathbb{P}\left(Y^* = y^* \mid X = x\right) \\ &= \binom{2N}{y^*}\Big(\varphi^*(x)\Big)^{y^*}\Big(1-\varphi^*(x)\Big)^{2N-y^*}\,, \end{aligned} \tag{2.2}$$

where $\varphi^*(x)$ is given by **Equation 2.1** , Y^* is the number of copies of A_1 in generation $t+1$, and X is the number of copies of A_1 in generation t. The Markov chain is time-homogeneous so we omit subscripts t and $t+1$ to avoid notational clutter. The letters

x and y will be used to distinguish the current and the next generation unless specified otherwise.

It can be deduced from $\boldsymbol{P}^*$ that there are two absorbing states, 0 and $2N$. Therefore fixation is guaranteed ultimately under the original Wright-Fisher model.

We focus on three generalised Wright-Fisher models, where mutation and selection are considered. All three models are standard generalisations of the Wright-Fisher model; we are not going to extend them but merely use the three generalisations to demonstrate our approach. We will refer to them as the *s*ingle *l*ocus model with *m*utation (SLM), the *s*ingle *l*ocus model with mutation and *s*election (SLS) and the *t*wo-*l*ocus *d*iallelic model with mutation and recombination (TLD) respectively.

Table 2.2: Three generalised Wright-Fisher models.

	Name	Abbreviation
1.	Single locus model with mutation	SLM
2.	Single locus model with mutation and selection	SLS
3.	Two-locus diallelic model with mutation and recombination	TLD

In the next three subsections, each of the three models will be made more explicit. In particular, the type of mutation and selection will be made clear for each. Some key features of the three generalised Wright-Fisher models are summarised in **Table 2.3**.

Table 2.3: Basics of SLM, SLS and TLD.

Abbreviation	# Loci	# Alleles	Mutation	Selection	Recombination
SLM	Single	Multiallelic	Equal	No	No
SLS	Single	Diallelic	Non-Equal	Yes	No
TLD	Two	Diallelic	Equal	No	Yes

2.1.2 Single locus model with mutation, SLM

Firstly, we consider k rather than just two distinct allelic types present at locus A, where $2 \leq k < \infty$. Let $A_1, A_2, \ldots, A_k$ denote the corresponding allelic types at locus A, and suppose the population consists of N individuals. The current genetic composition at locus A can be described by a k-dimensional vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$, where x_i is the number of copies of allelic type A_i, and by definition,

$$x_1 + x_2 + x_3 + \cdots + x_k = 2N \,.$$

Let $\boldsymbol{y}^*$ denote the genetic composition at locus A in the next generation, where $\boldsymbol{y}^*$ is a k-dimensional vector with its element y_i^* being the number of copies of allelic type A_i in the next generation. Let $\varphi_i^*(\boldsymbol{x})$ denote the expected proportion of allelic type A_i in the next generation, and $\boldsymbol{\varphi}^*(\boldsymbol{x})$ denote the vector $(\varphi_1^*(\boldsymbol{x}), \varphi_2^*(\boldsymbol{x}), \cdots, \varphi_k^*(\boldsymbol{x}))$.

So far this is just a multi-dimensional generalisation of the original Wright-Fisher model. With the same argument that gives us **Equation 2.1** , $\varphi_i^*(\boldsymbol{x})$ is given by,

$$\varphi_i^*(\boldsymbol{x}) = \frac{x_i}{2N} \,. \tag{2.3}$$

Let $\boldsymbol{P}^*$ denote the transition matrix of going from $\boldsymbol{x}$ to $\boldsymbol{y}^*$ in one generation. $\boldsymbol{P}^*$ is given by multinomial, rather than the aforementioned binomial, mass functions. Hence,

$$
\begin{aligned}
p^*_{\boldsymbol{x}\boldsymbol{y}^*} &= \mathbb{P}\left(\boldsymbol{Y}^* = \boldsymbol{y}^* \mid \boldsymbol{X} = \boldsymbol{x}\right) \\
&= \frac{(2N)!}{y_1^*! \cdots y_k^*!} \left(\varphi_1^*(\boldsymbol{x})\right)^{y_1^*} \left(\varphi_2^*(\boldsymbol{x})\right)^{y_2^*} \cdots \left(\varphi_k^*(\boldsymbol{x})\right)^{y_k^*},
\end{aligned} \tag{2.4}
$$

where $\varphi_i^*(\boldsymbol{x})$ is given by **Equation 2.3** .

Suppose now there is a reversible recurrent mutation force acting on the population after the formation of the zygotes, such that any allelic type A_i might mutate to any of the remaining $k-1$ allelic types with known mutation probabilities. For k distinct allelic types at locus A, we need a k-by-k matrix $\boldsymbol{U}$ in general to define the mutation structure, where the element u_{ij} is the probability of allelic type A_i mutating to A_j.

We consider an equal (or symmetric) mutation model in which $u_{ij} = u \ \forall \, i, j$ where $i \neq j$, so that the mutation matrix defined in **Equation 2.5** describes the mutation structure for the current model.

$$
\boldsymbol{U} = \begin{pmatrix}
1-(k-1)\,u & u & \dots & u \\
u & 1-(k-1)\,u & \dots & u \\
\vdots & \vdots & \ddots & \vdots \\
u & u & \dots & 1-(k-1)\,u
\end{pmatrix}. \tag{2.5}
$$

By definition a probability is between 0 and 1 inclusive, and each row of $\boldsymbol{U}$ must sum to 1. Hence the probability of a certain allelic type A_i remaining unmutated is $1-(k-1)\,u$, and therefore $0 < u \leq \frac{1}{k-1}$.

Given the mutation force defined by $\boldsymbol{U}$, we will have a new transition matrix $\boldsymbol{P}$ instead of $\boldsymbol{P}^*$.

Let $\boldsymbol{y}$ denote the genetic composition for the population in the next generation in the presence of this mutation force, where $\boldsymbol{y}$ is a k-dimensional vector with its element y_i being the number of copies of allelic type A_i in the next generation. Let $\varphi_i(\boldsymbol{x})$ be the expected proportion of allelic type A_i in the next generation in the presence of this mutation force, and $\boldsymbol{\varphi}(\boldsymbol{x})$ denote the vector $(\varphi_1(\boldsymbol{x}), \varphi_2(\boldsymbol{x}), \cdots, \varphi_k(\boldsymbol{x}))$.

Then $\varphi_i(\boldsymbol{x})$ has the following forms :

$$\begin{aligned}
\varphi_i(\boldsymbol{x}) &= \overbrace{\{1-(k-1)\,u\}\,\frac{x_i}{2N}}^{A_i \quad \text{stays as} \quad A_i} + \overbrace{\sum_{i\neq j} u\frac{x_j}{2N}}^{A_j \quad \text{mutates to} \quad A_i} \\
&= \frac{x_i}{2N} - (k-1)\,u\frac{x_i}{2N} + u\left(1-\frac{x_i}{2N}\right) \\
&= \frac{x_i}{2N}\,(1-ku) + u\,. \qquad (2.6)
\end{aligned}$$

Hence the transition probabilities of going from $\boldsymbol{x}$ to $\boldsymbol{y}$ in the next generation are given by the following,

$$\begin{aligned}
p_{\boldsymbol{x}\boldsymbol{y}} &= \mathbb{P}\,(\boldsymbol{Y}=\boldsymbol{y} \mid \boldsymbol{X}=\boldsymbol{x}) \\
&= \frac{(2N)!}{y_1!\cdots y_k!}\Big(\varphi_1(\boldsymbol{x})\Big)^{y_1}\Big(\varphi_2(\boldsymbol{x})\Big)^{y_2}\cdots\Big(\varphi_k(\boldsymbol{x})\Big)^{y_k}, \qquad (2.7)
\end{aligned}$$

where $\varphi_i(\boldsymbol{x})$ is given by **Equation 2.6** .

The transition matrix $\boldsymbol{P}$ defines this generalised Wright-Fisher model, and we will refer to this model as the single locus model with mutation (SLM) for the rest of the thesis. The SLM model corresponds to an irreducible and aperiodic Markov chain, hence there exists a unique stationary distribution. We include this model mainly to demonstrate the performance of our method when the underlying stationary distribution is multivariate.

2.1.3 Single locus model with mutation and selection, SLS

For this model, we consider a diallelic locus where selection occurs in addition to mutation. Suppose there are N individuals and there are allele types A_1 and A_2 at locus A. Let x denote the number of copies of allelic type A_1 just before conception. Under assumptions 2 and 4 in **Table 2.1**, gametes are not only sampled randomly and independently with replacement, but chosen gametes are also combined randomly and independently to form a zygote at the time of conception for the next generation. Hence the probability of having a certain zygote A_i/A_j is equal to

$$\mathbb{P}_{(A_i/A_j)} = \begin{cases} \Big(\mathbb{P}\left(A_i\right)\Big)^2 & \text{for } i = j, \\ 2\,\mathbb{P}\left(A_i\right)\mathbb{P}\left(A_j\right) & \text{otherwise.} \end{cases}$$

Without a selection force, it is taken that only N zygotes are formed. The expected zygotic proportions of A_1/A_1, A_1/A_2 and A_2/A_2 at the time of conception are given by the following:

$$\mathbb{P}\left(A_1/A_1\right) = \left(\frac{x}{2N}\right)^2 \tag{2.8}$$

$$\mathbb{P}\left(A_1/A_2\right) = 2\left(\frac{x}{2N}\right)\left(1 - \frac{x}{2N}\right) \tag{2.9}$$

$$\mathbb{P}\left(A_2/A_2\right) = \left(1 - \frac{x}{2N}\right)^2 . \tag{2.10}$$

It can be considered that all N zygotes survive to reach maturity and have an equal chance to populate the next generation.

In the presence of a selection force, it can be understood conceptually that an infinite number of zygotes are formed according to the expected zygotic proportions in **Equations 2.8–2.10**, but *not* all of them survive until the next breeding season, at which time only N of them are selected to populate the next generation.

Let a selection force act on the population such that the ratio of A_1/A_1, A_1/A_2 and A_2/A_2 surviving zygotes is given by,

$$\text{Surviving Ratio} \quad = \quad (1+s_1) \,:\, (1+s_2) \,:\, 1\,, \tag{2.11}$$

where s_1 and s_2 are known as selection coefficients.

Mathematically, s_1 and s_2 are bounded between -1 and ∞, where 0 indicates selectively neutral and -1 indicates complete lethality. Approaching infinity indicates an extremely high selective advantage. However, the biologically plausible range is usually in the order of 1×10^{-4} to 1×10^{-3}.

Let $\varphi^*(x)$ denote the expected proportion of allelic type A_1 at the next conception time. With the above selection scheme, $\varphi^*(x)$ is given by,

$$\begin{aligned}\varphi^*(x) &= \frac{\overbrace{(1+s_1)\left(\frac{x}{2N}\right)^2}^{\text{Homozygote}} + \overbrace{(1+s_2)\left(\frac{x}{2N}\right)\left(\frac{2N-x}{2N}\right)}^{\text{Heterozygote}}}{(1+s_1)\left(\frac{x}{2N}\right)^2 + 2(1+s_2)\left(\frac{x}{2N}\right)\left(\frac{2N-x}{2N}\right) + \left(\frac{2N-x}{2N}\right)^2} \\ &= \frac{(1+s_1)\,x^2 + (1+s_2)\,x\,(2N-x)}{(1+s_1)\,x^2 + 2\,(1+s_2)\,x\,(2N-x) + (2N-x)^2}\,.\end{aligned} \tag{2.12}$$

Now, let us also consider a reversible recurrent mutation force that acts in addition to the selection force on the population between two consecutive breeding seasons. Let A_1 mutate to A_2 with probability u_2, and an opposite mutation force act simultaneously which turns A_2 to A_1 with probability u_1:

$$A_1 \xrightarrow{u_2} A_2\,, \qquad A_2 \xrightarrow{u_1} A_1\,.$$

The corresponding mutation matrix $\boldsymbol{U}$ is given by:

$$\boldsymbol{U} = \begin{pmatrix} 1-u_2 & u_2 \\ u_1 & 1-u_1 \end{pmatrix}\,. \tag{2.13}$$

Let $\varphi\left(x\right)$ denote the expected allele proportion of A_1 at locus A at the next conception time in the presence of this mutation force. Then $\varphi\left(x\right)$ is given by the following,

$$\begin{aligned} \varphi\left(x\right) &= \left(1-u_2\right)\varphi^*\left(x\right) + u_1\left(1-\varphi^*\left(x\right)\right) \\ &= \left(1-u_1-u_2\right)\varphi^*\left(x\right) + u_1\,, \end{aligned} \tag{2.14}$$

where $\varphi^*\left(x\right)$ is defined in **Equation 2.12** .

Let y denote the number of copies of A_1 in the population in the presence of selection and mutation after conception in the next generation. The transition probabilities of going from x to y in a generation are given by,

$$
\begin{aligned}
p_{xy} &= \mathbb{P}\left(Y = y \mid X = x\right) \\
&= \binom{2N}{y} \left(\varphi\left(x\right)\right)^{y} \left(1 - \varphi\left(x\right)\right)^{2N-y} ,
\end{aligned} \tag{2.15}
$$

where $\varphi\left(x\right)$ is defined by **Equation 2.14** .

The above transition matrix $\boldsymbol{P}$ defines this model, which we refer to as the single locus model with mutation and selection (SLS) for the rest of the thesis. The SLS model corresponds to an irreducible aperiodic Markov chain, hence there exists a unique stationary distribution. We include this model mainly to demonstrate the performance of our method for a stationary distribution which depends on selection as well as mutation.

2.1.4 Two-locus diallelic model with mutation and recombination, TLD

Finally, and most importantly, we consider two diallelic loci A and B, at which there are alleles A_1, A_2 and B_1, B_2 respectively. There are thus four possible types of gamete, A_1B_1, A_1B_2, A_2B_1 and A_2B_2, and we will refer to these as types 1 to 4 respectively. Suppose the population size is N, and let vector $\boldsymbol{x}$ denote the genetic composition of the population in the current generation, where its elements x_1, x_2, x_3 and x_4 are the number of gametes of types 1 to 4 in the current generation as shown in **Table 2.4**.

By definition:

$$x_1 + x_2 + x_3 + x_4 = 2N.$$

Table 2.4: Possible types of gamete for two diallelic loci.

	Type of gamete	Current count
1.	A_1B_1	x_1
2.	A_1B_2	x_2
3.	A_2B_1	x_3
4.	A_2B_2	x_4

We again consider the case that gametes are chosen randomly and independently. In the absence of recombination and mutation, the expected proportions of gametes of types 1 to 4 in the next generation are given by the following:

$$\frac{x_1}{2N}, \quad \frac{x_2}{2N}, \quad \frac{x_3}{2N} \quad \text{and} \quad \frac{x_4}{2N}. \tag{2.16}$$

There are ten possible genotypes and each genotype has been given an index in **Table 2.5**. Let vector $\boldsymbol{g} = (g_1, g_2, \ldots, g_{10})$, where g_i denotes the number of copies of genotype i in the current generation.

By definition:

$$g_1 + g_2 + \cdots + g_{10} = N\,.$$

The gametic and the genotype counts are genetic descriptions of the same population, so they must match for every generation. Given our definitions of $\boldsymbol{x}$ and $\boldsymbol{g}$, the identities between gametic and genotype counts can be identified; see **Table 2.6**.

Gametes are sampled randomly and independently with replacement to form zygotes

Table 2.5: Possible genotypes for two diallelic loci.

Index	Genotype	Count, $\boldsymbol{g}_t$	Type of zygote	$\mathbb{E}\left(\boldsymbol{g}_{t+1}\right)$
1	A_1B_1/A_1B_1	g_1	Homozygous	$\frac{x_1}{2N}\frac{x_1}{2N}$
2	A_1B_1/A_1B_2	g_2	Heterozygous	$2\frac{x_1}{2N}\frac{x_2}{2N}$
3	A_1B_1/A_2B_1	g_3	Heterozygous	$2\frac{x_1}{2N}\frac{x_3}{2N}$
4	A_1B_1/A_2B_2	g_4	Heterozygous	$2\frac{x_1}{2N}\frac{x_4}{2N}$
5	A_1B_2/A_1B_2	g_5	Homozygous	$\frac{x_2}{2N}\frac{x_2}{2N}$
6	A_1B_2/A_2B_1	g_6	Heterozygous	$2\frac{x_2}{2N}\frac{x_3}{2N}$
7	A_1B_2/A_2B_2	g_7	Heterozygous	$2\frac{x_2}{2N}\frac{x_4}{2N}$
8	A_2B_1/A_2B_1	g_8	Homozygous	$\frac{x_3}{2N}\frac{x_3}{2N}$
9	A_2B_1/A_2B_2	g_9	Heterozygous	$2\frac{x_3}{2N}\frac{x_4}{2N}$
10	A_2B_2/A_2B_2	g_{10}	Homozygous	$\frac{x_4}{2N}\frac{x_4}{2N}$

where $\mathbb{E}\left(\boldsymbol{g}_{t+1}\right)$ denotes expected genotype proportion in the next generation without recombination and mutation.

Table 2.6: Identities between gamete type frequencies and genotype frequencies.

$$
\begin{aligned}
x_1 &= 2g_1 + g_2 + g_3 + g_4 \\
x_2 &= 2g_5 + g_6 + g_7 + g_2 \\
x_3 &= 2g_8 + g_9 + g_6 + g_3 \\
x_4 &= 2g_{10} + g_9 + g_7 + g_4
\end{aligned}
$$

at the time of conception for the next generation. Without recombination and mutation, the expected *genotype* proportions in the next generation are therefore gained from the product of the corresponding expected *gametic* proportions in the next generation. This leads to **Equation 2.17** .

$$\mathbb{P}\left(A_iB_j/A_mB_n\right) = \begin{cases} \left\{\mathbb{P}\left(A_iB_j\right)\right\}^2 & \text{for } i = m \text{ and } j = n, \\ 2\,\mathbb{P}\left(A_iB_j\right)\mathbb{P}\left(A_mB_n\right) & \text{otherwise.} \end{cases} \tag{2.17}$$

Therefore it is sufficient to consider gametic counts x_1, x_2, x_3 and x_4. In the absence of recombination and mutation, the expected genotype proportions can be expressed in terms of the current gametic counts x_1, x_2, x_3 and x_4, see **Table 2.5**.

Now let us consider recombination. Recombination, also known as crossover, is a process of exchanging alleles between two uniting gametes during meiosis. In the presence of recombination, the two uniting gametes A_iB_j and A_mB_n might lead to a zygote of the type A_iB_n/A_mB_j as well as a zygote of the type A_iB_j/A_mB_n, where the possibility of having A_iB_n/A_mB_j is due to crossover. Recombination is a relative phenomenon between two or more loci, hence it is only meaningful when we are considering more than one locus. Recombination may happen more than once at the same locus for the same pair of gametes. Only an odd number of crossovers at the same locus on the same pair of gametes will change the genetic composition.

The recombination fraction between two loci, that is to say the probability of there being an odd number of crossovers between them, is defined as C. Biologically, C is bounded between 0 and 0.5. Two loci effectively become one locus when the lower bound $C = 0$ is achieved (complete linkage). The upper bound $C = 0.5$ is achieved if the loci are unlinked, for example they are on a different pair of chromosomes.

For the next generation, let $\varphi_1^*(\boldsymbol{x})$, $\varphi_2^*(\boldsymbol{x})$, $\varphi_3^*(\boldsymbol{x})$ and $\varphi_4^*(\boldsymbol{x})$ denote the expected proportions of gametes of types 1 to 4 in the presence of recombination but not yet mutation. We need to consider the expected genotype proportions in order to work out

$\varphi_i^*(\boldsymbol{x})$. The expected gamete type frequency $\varphi_1^*(\boldsymbol{x})$ consists of four parts. Part 1 in **Equation 2.18** represents individuals that are homozygous in *both loci* and *both of their gametes* are A_1B_1; this part is not affected by crossover. Part 2 represents individuals that are homozygous in *exactly one locus* and have *exactly one gamete* being A_1B_1; this part is also not affected by crossover. Part 3 represents individuals that would have one gamete of A_1B_1 given that an *even* number of crossovers occurred. Part 4 represents individuals that would have one gamete being A_1B_1 given that an *odd* number of crossovers occurred. Collecting all these contributions gives us $\varphi_1^*(\boldsymbol{x})$:

$$\begin{aligned}
\varphi_1^*(\boldsymbol{x}) &= \left(\frac{\overbrace{2\left(\frac{x_1}{2N}\right)^2 N}^{\text{part 1}} + \overbrace{\left(2\frac{x_1}{2N}\frac{x_2}{2N}\right) N + \left(2\frac{x_1}{2N}\frac{x_3}{2N}\right) N}^{\text{part 2}}}{2N} \right) \\
&\quad + \left(\frac{\overbrace{(1-C)\left(2\frac{x_1}{2N}\frac{x_4}{2N}\right) N}^{\text{part 3}} + \overbrace{C\left(2\frac{x_2}{2N}\frac{x_3}{2N}\right) N}^{\text{part 4}}}{2N} \right) &\quad (2.18) \\
&= \left(\frac{x_1}{2N}\right)^2 + \frac{x_1}{2N}\frac{x_2}{2N} + \frac{x_1}{2N}\frac{x_3}{2N} + (1-C)\frac{x_1}{2N}\frac{x_4}{2N} + C\frac{x_2}{2N}\frac{x_3}{2N}. &\quad (2.19)
\end{aligned}$$

Given that $x_1 + x_2 + x_3 + x_4 = 2N$, **Equation 2.19** can be simplified to :

$$\varphi_1^*(\boldsymbol{x}) = \frac{x_1}{2N} - C\left(\overbrace{\frac{x_1}{2N}\frac{x_4}{2N}}^{\text{part 1}} - \overbrace{\frac{x_2}{2N}\frac{x_3}{2N}}^{\text{part 2}} \right). \qquad (2.20)$$

Part 1 and part 2 in **Equation 2.20** represent respectively the loss and gain of A_1B_1

due to crossover.

The quantity $\left(\frac{x_1}{2N}\frac{x_4}{2N} - \frac{x_2}{2N}\frac{x_3}{2N}\right)$ is usually known as the coefficient of linkage disequilibrium. D is a function of the gametic proportions, so we will write $\left(\frac{x_1}{2N}\frac{x_4}{2N} - \frac{x_2}{2N}\frac{x_3}{2N}\right)$ as $D(\boldsymbol{x})$ to emphasise the dependency.

Note that

$$\frac{x_4}{2N} = 1 - \frac{x_1}{2N} - \frac{x_2}{2N} - \frac{x_3}{2N},$$

so

$$D\left(\boldsymbol{x}\right) = \frac{x_1}{2N} - \left(\frac{x_1}{2N}\right)^2 - \frac{x_1}{2N}\frac{x_2}{2N} - \frac{x_1}{2N}\frac{x_3}{2N} - \frac{x_2}{2N}\frac{x_3}{2N}. \tag{2.21}$$

Using similar working, $\varphi_1^*\left(\boldsymbol{x}\right)$, $\varphi_2^*\left(\boldsymbol{x}\right)$, $\varphi_3^*\left(\boldsymbol{x}\right)$ and $\varphi_4^*\left(\boldsymbol{x}\right)$ are given by,

$$\varphi_1^*\left(\boldsymbol{x}\right) = \left(\frac{x_1}{2N} - CD(\boldsymbol{x})\right) \tag{2.22}$$

$$\varphi_2^*\left(\boldsymbol{x}\right) = \left(\frac{x_2}{2N} + CD(\boldsymbol{x})\right) \tag{2.23}$$

$$\varphi_3^*\left(\boldsymbol{x}\right) = \left(\frac{x_3}{2N} + CD(\boldsymbol{x})\right) \tag{2.24}$$

$$\varphi_4^*\left(\boldsymbol{x}\right) = \left(\frac{x_4}{2N} - CD(\boldsymbol{x})\right). \tag{2.25}$$

Now suppose there are reversible recurrent mutation forces present at both loci in addition to recombination. Let us consider the same equal mutation structure we considered for the SLM model in **subsection 2.1.2** for both loci, and let us assume the mutation probabilities are u for both loci.

$$A_1 \underset{u}{\overset{u}{\rightleftharpoons}} A_2 \,, \qquad\qquad B_1 \underset{u}{\overset{u}{\rightleftharpoons}} B_2 \,.$$

Therefore both loci have the same mutation matrix $\boldsymbol{U}$ defined by **Equation 2.26** .

$$\boldsymbol{U} = \begin{pmatrix} (1-u) & u \\ u & (1-u) \end{pmatrix} , \tag{2.26}$$

where u is the mutation probability.

Let $\varphi_1(\boldsymbol{x})$, $\varphi_2(\boldsymbol{x})$, $\varphi_3(\boldsymbol{x})$ and $\varphi_4(\boldsymbol{x})$ denote the expected proportions of gametes of types 1 to 4 respectively in the presence of recombination and the equal mutation structure defined by $\boldsymbol{U}$. Then we have,

$$\begin{aligned} \varphi_1(\boldsymbol{x}) &= \varphi_1^*(\boldsymbol{x})(1-u)^2 + \varphi_2^*(\boldsymbol{x})\,u\,(1-u) + \varphi_3^*(\boldsymbol{x})\,u\,(1-u) + \varphi_4^*(\boldsymbol{x})\,u^2 \\ &= \frac{x_1}{2N}(1-u)^2 + \left(\frac{x_2}{2N} + \frac{x_3}{2N}\right) u\,(1-u) + \frac{x_4}{2N}u^2 - CD(\boldsymbol{x})\,(1-2u)^2 \,, \end{aligned} \tag{2.27}$$

where $\varphi_1^*(\boldsymbol{x})$, $\varphi_2^*(\boldsymbol{x})$, $\varphi_3^*(\boldsymbol{x})$ and $\varphi_4^*(\boldsymbol{x})$ are defined by **Equations 2.22–2.25**.

Similarly, $\varphi_2(\boldsymbol{x})$, $\varphi_3(\boldsymbol{x})$ and $\varphi_4(\boldsymbol{x})$ can be given by,

$$\varphi_2(\boldsymbol{x}) = \frac{x_2}{2N}(1-u)^2 + \left(\frac{x_1}{2N} + \frac{x_4}{2N}\right)u(1-u) + \frac{x_3}{2N}u^2 + CD(\boldsymbol{x})(1-2u)^2 \tag{2.28}$$

$$\varphi_3(\boldsymbol{x}) = \frac{x_3}{2N}(1-u)^2 + \left(\frac{x_1}{2N} + \frac{x_4}{2N}\right)u(1-u) + \frac{x_2}{2N}u^2 + CD(\boldsymbol{x})(1-2u)^2 \tag{2.29}$$

$$\varphi_4(\boldsymbol{x}) = \frac{x_4}{2N}(1-u)^2 + \left(\frac{x_2}{2N} + \frac{x_3}{2N}\right)u(1-u) + \frac{x_1}{2N}u^2 - CD(\boldsymbol{x})(1-2u)^2 . \tag{2.30}$$

Let the vector $\boldsymbol{y}$ denote the genetic composition of the population in the next generation in the presence of recombination and mutation, where the elements y_1, y_2, y_3 and y_4 are the number of copies of gametes of types 1 to 4 in the next generation. Hence the transition probabilities of going from $\boldsymbol{x}$ to $\boldsymbol{y}$ are given by the multinomial probability distribution defined by,

$$\begin{aligned} p_{\boldsymbol{xy}} &= \mathbb{P}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}\right) \\ &= \frac{(2N)!}{y_1!y_2!y_3!y_4!}\Big(\varphi_1(\boldsymbol{x})\Big)^{y_1}\Big(\varphi_2(\boldsymbol{x})\Big)^{y_2}\Big(\varphi_3(\boldsymbol{x})\Big)^{y_3}\Big(\varphi_4(\boldsymbol{x})\Big)^{y_4}, \end{aligned} \tag{2.31}$$

where $\varphi_1(\boldsymbol{x})$, $\varphi_2(\boldsymbol{x})$, $\varphi_3(\boldsymbol{x})$ and $\varphi_4(\boldsymbol{x})$ are defined by **Equations 2.27–2.30**.

The above transition matrix $\boldsymbol{P}$ defines this model, which we will refer to as the two-locus diallelic model with mutation and recombination (TLD) for the rest of the thesis. The TLD is an irreducible aperiodic Markov chain, so there exists a unique stationary distribution. We include this model mainly to demonstrate that our method can be used to compute an analytic approximation for stationary distributions in models where no analytic approaches are available.

2.2 Stationary Distributions for Discrete Processes

In this section, we study some methods of directly obtaining the stationary distribution of the discrete generalised Wright-Fisher models without using the diffusion approximation. We will show preliminary results for each method, outline when these approaches are adequate, and discuss their limitations.

2.2.1 Analytic computation

Analytically, for extremely small N, the stationary distributions for the generalised Wright-Fisher models can be found element-wise as rational functions of population parameters by solving the stationary conditions directly using the corresponding transition matrix.

For example, let us consider the SLS model for an extremely small population of just two individuals ($N = 2$) in the case where the mutation forces are equal ($u_1 = u_2 = u$) and there is no selection force ($s_1 = s_2 = 0$). See **Tables 2.2** and **2.3** for a short summary of the SLS model, and see **Equation 2.15** for the general form of the transition matrix. There are only five possible states for this model; there could be either 0, 1, 2, 3 or 4 copies of A_1 at locus A. The corresponding transition matrix for the model is therefore,

$$\boldsymbol{P} = \begin{pmatrix} (1-u)^4 & 4u(1-u)^3 & 6u^2(1-u)^2 & 4u^3(1-u) & u^4 \\ \left(\frac{3}{4}-\frac{1}{2}u\right)^4 & 4\left(\frac{1}{4}+\frac{1}{2}u\right)\left(\frac{3}{4}-\frac{1}{2}u\right)^3 & 6\left(\frac{1}{4}+\frac{1}{2}u\right)^2\left(\frac{3}{4}-\frac{1}{2}u\right)^2 & 4\left(\frac{1}{4}+\frac{1}{2}u\right)^3\left(\frac{3}{4}-\frac{1}{2}u\right) & \left(\frac{1}{4}+\frac{1}{2}u\right)^4 \\ \frac{1}{16} & \frac{1}{4} & \frac{3}{8} & \frac{1}{4} & \frac{1}{16} \\ \left(\frac{1}{4}+\frac{1}{2}u\right)^4 & 4\left(\frac{1}{4}+\frac{1}{2}u\right)^3\left(\frac{3}{4}-\frac{1}{2}u\right) & 6\left(\frac{1}{4}+\frac{1}{2}u\right)^2\left(\frac{3}{4}-\frac{1}{2}u\right)^2 & 4\left(\frac{1}{4}+\frac{1}{2}u\right)\left(\frac{3}{4}-\frac{1}{2}u\right)^3 & \left(\frac{3}{4}-\frac{1}{2}u\right)^4 \\ u^4 & 4u^3(1-u) & 6u^2(1-u)^2 & 4u(1-u)^3 & (1-u)^4 \end{pmatrix}, \tag{2.32}$$

where u is the probability of mutation between A_1 and A_2.

Let $\boldsymbol{\pi}$ denote the stationary distribution, which can be found by solving the stationary condition,

$$\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{P} - \boldsymbol{I}) = \mathbf{0}\,,$$

where $\boldsymbol{I}$ denotes the identity matrix and $\mathbf{0}$ denotes a vector of zeros.

We obtain the following stationary distribution using Maple:

$$\boldsymbol{\pi} = \left(\frac{1}{\Lambda}\right)\begin{pmatrix} \frac{1}{2}\left(16u^4-32u^3+72u^2-56u+29\right) \\ -32u(u-1)\left(4u^2-4u+5\right) \\ 12u(u-1)\left(48u^4-96u^3+88u^2-40u-9\right) \\ -32u(u-1)\left(4u^2-4u+5\right) \\ \frac{1}{2}\left(16u^4-32u^3+72u^2-56u+29\right) \end{pmatrix}, \tag{2.33}$$

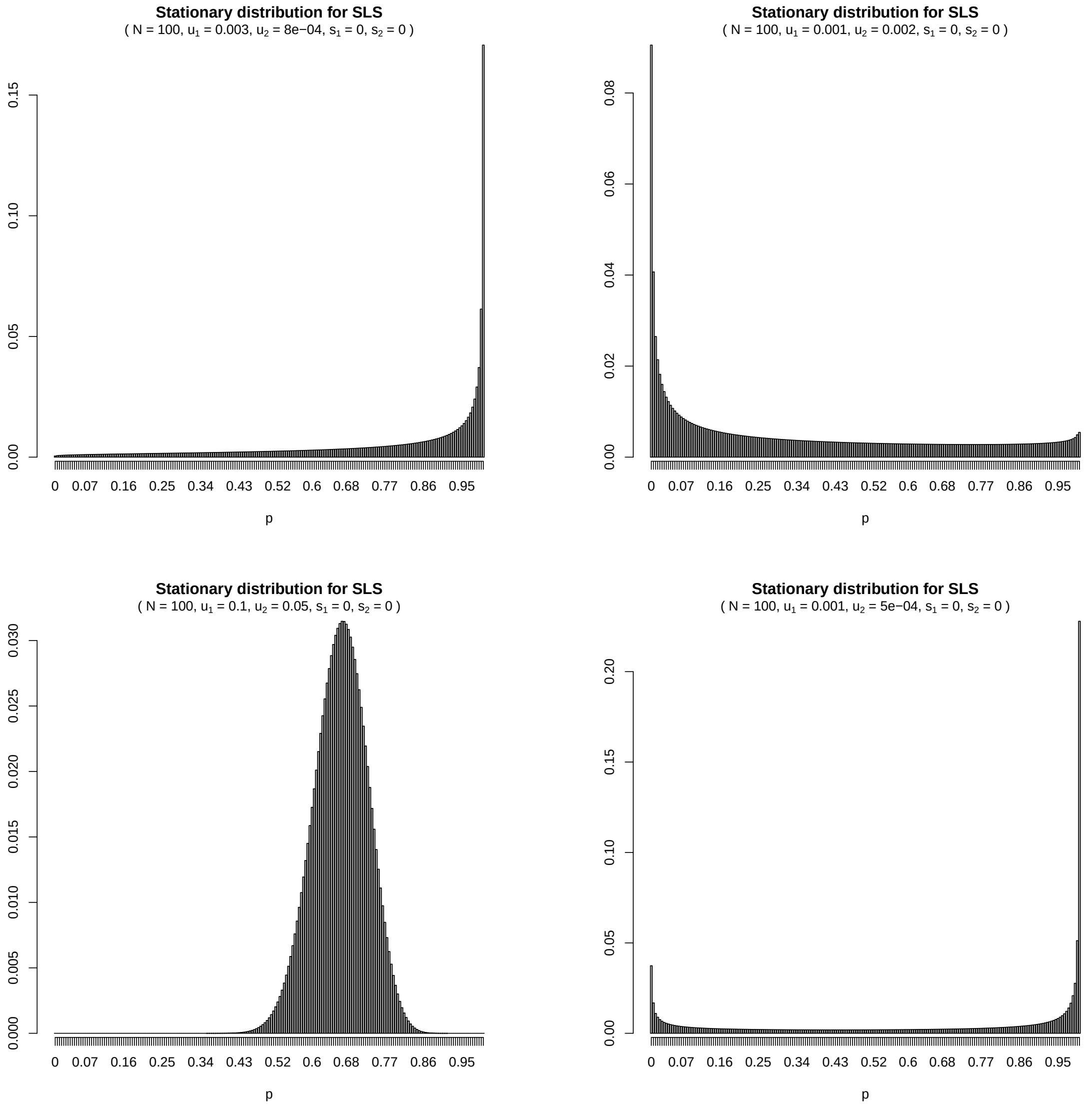
where $\Lambda = 576u^6 - 1728u^5 + 1968u^4 - 1056u^3 - 132u^2 + 372u + 29$.

It appears to be impossible to obtain the stationary distribution in a single closed expression for all elements. Analytic formulae, such as **Equation 2.33** , can be obtained for most of the generalised Wright-Fisher models, but as N increases such formulae rapidly become too computationally expensive to find, even with a computer algebra system such as Maple. The analytic form also becomes so vastly complicated that it has little additional benefit over an accurate numerical solution.

2.2.2 Numerical computation

We have the option to solve the stationary condition numerically instead of symbolically. Numerical solutions allow us to deal with more complex generalised Wright-Fisher models, including those with much larger N.

Figure 2.1: Numerical stationary distributions of the SLS model.



When N is in the hundreds, we can numerically solve the stationary condition almost instantaneously. **Figure 2.1** shows four stationary distributions that are numerically determined. All four are solutions to the SLS model, but with different sets of population parameters. It can be seen that the stationary distributions for such models are highly variable even with only recurrent reversible mutation. When plausible, numerical solutions have the advantage of being simple and quick, and neither approximations nor simulations are necessary. These solutions are useful when we want to study the effect of changing population parameters on the corresponding stationary distributions.

However, there still exists a limit, in terms of the size of N, beyond which solving the transition matrices numerically becomes computationally infeasible due to the discrete nature of the problem, especially when dealing with high dimensions. For the TLD model, the support of the stationary distribution is a three-dimensional simplex grid. There are $\binom{2N+4-1}{3}$ possible states for N individuals, so for the specific case of 10 individuals there are 1771 possible states. It is clearly inefficient and problematic to solve the stationary condition with thousands of rows for a population of just a few individuals. Therefore we need further tools to study the case where N is arbitrarily large.

2.2.3 Normal approximation

There are special cases for which it is possible to find the exact mean and variance of a stationary distribution without knowing the exact form of the stationary distribution itself. For these special cases, a normal approximation might be adequate: in other words, to approximate the stationary distribution using a normal distribution with the correct mean and variance. Let us again consider the SLS model: see **Tables 2.2** and **2.3** for a short summary of SLS and see **Equation 2.15** for the general form of the transition matrix. We will use $\boldsymbol{\pi}^{\mathsf{T}} = (\pi_0, \pi_1, \pi_2, \ldots, \pi_{2N})$ to denote the stationary distribution. At stationarity:

$$\boldsymbol{\pi}^{\mathsf{T}} = \boldsymbol{\pi}^{\mathsf{T}} \boldsymbol{P}\,, \tag{2.34}$$

where $\boldsymbol{P}$ is the transition matrix defined by **Equation 2.15** .

Let μ be the mean of the stationary distribution and $\boldsymbol{\xi}$ be a vector containing values for all the possible states, that is, $\boldsymbol{\xi}^{\mathsf{T}} = (0, 1, 2, 3, \ldots, 2N)$. Therefore, by definition μ is given by:

$$\begin{aligned} \mu &= \boldsymbol{\pi}^{\mathsf{T}} \boldsymbol{\xi} \\ &= \boldsymbol{\pi}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{\xi}\,. \end{aligned} \tag{2.35}$$

Each row of $\boldsymbol{P}$ is given by the binomial distribution in **Equation 2.15** for a specific value of x, where x is the number of copies of allele type A_1 in generation t. Hence each component of the vector $\boldsymbol{P}\boldsymbol{\xi}$ is the mean of the binomial distribution for the corresponding x. The vector $\boldsymbol{P}\boldsymbol{\xi}$ is therefore given by :

$$\boldsymbol{P\xi} = \begin{pmatrix} \sum_{y=0}^{2N} y\binom{2N}{y}\{\varphi(0)\}^y\{1-\varphi(0)\}^{2N-y} \\ \sum_{y=0}^{2N} y\binom{2N}{y}\{\varphi(1)\}^y\{1-\varphi(1)\}^{2N-y} \\ \vdots \\ \sum_{y=0}^{2N} y\binom{2N}{y}\{\varphi(2N)\}^y\{1-\varphi(2N)\}^{2N-y} \end{pmatrix}$$

$$= \begin{pmatrix} 2N\varphi(0) \\ 2N\varphi(1) \\ \vdots \\ 2N\varphi(2N) \end{pmatrix}, \tag{2.36}$$

where $\varphi(x)$ is defined by **Equation 2.14**.

When there is no selection pressure ($s_1 = s_2 = 0$), then :

$$2N\varphi(x) = x(1-u_2) + (2N-x)u_1 . \tag{2.37}$$

Hence combining **Equations 2.35–2.37**, we obtain,

$$
\begin{aligned}
\mu &= \boldsymbol{\pi}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{\xi} \\
&= \sum_{x=0}^{2N} \pi_x \left\{ x \left(1 - u_2\right) + \left(2N - x\right) u_1 \right\} \\
&= \left(1 - u_2\right) \sum_{x=0}^{2N} x \pi_x - u_1 \sum_{x=0}^{2N} x \pi_x + 2N u_1 \sum_{x=0}^{2N} \pi_x \\
&= \left(1 - u_2\right) \mu - u_1 \mu + 2N u_1 \, . \qquad (2.38)
\end{aligned}
$$

By collecting μ we obtain,

$$
\mu = \frac{2N u_1}{u_1 + u_2} \, . \qquad (2.39)
$$

Let σ^2 and ν denote respectively the variance and the second moment of the stationary distribution, and let $\boldsymbol{\zeta}^{\mathsf{T}} = \left(0^2, 1^2, 2^2, 3^2, \ldots, (2N)^2\right)$. Using a similar argument we can derive,

$$
\begin{aligned}
\nu &= \boldsymbol{\pi}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{\zeta} \\
&= \sum_{x=0}^{2N} \pi_x \left[2N \varphi\left(x\right) \left\{1 - \varphi\left(x\right)\right\} + \left\{2N \varphi\left(x\right)\right\}^2 \right] \\
&= \frac{\left(2N - 1\right)\left(1 - u_1 - u_2\right)^2}{2N} \nu \\
&\qquad + \left(1 - 2u_1 + 4N u_1\right)\left(1 - u_1 - u_2\right) \mu \\
&\qquad\qquad + 2N u_1 \left(1 - u_1 + 2N u_1\right) \, . \qquad (2.40)
\end{aligned}
$$

We are able to calculate the variance σ^2 by combining **Equations 2.39–2.40**:

$$\begin{aligned}\sigma^2 &= \nu - \mu^2 \\ &= \frac{4N^2 u_1 u_2}{(u_1+u_2)^2 \left\{2N - (2N-1)(1-u_1-u_2)^2\right\}} .\end{aligned} \tag{2.41}$$

The mean μ and the variance σ^2 in **Equations 2.39** and **2.40** essentially agree with the derivation by **Ewens (1969)**, in which the same mean and a formula based on a few initial terms of a Taylor series expansion for the variance were given. We choose to use the exact variance instead of its approximation by a Taylor series.

Using these expressions for the mean and the variance of the stationary distribution, we investigated the possibility of approximating the underlying stationary distribution using a normal distribution. This clearly is not a good idea for non-symmetric distributions such as three of those shown in **Figure 2.1** . However, we found that for a given set of population parameters, the stationary distribution becomes more and more symmetric and bell-shaped as N increases. **Figure 2.2** shows such a pattern.

The stationary distributions in **Figure 2.2** are numerically determined. All four cases have the same mutation probabilities ($u_1 = 0.01$ and $u_2 = 0.005$), but N increases from 50 to 500 from the top left to the bottom right. For this set of mutation probabilities, the stationary distribution becomes approximately symmetric and bell-shaped for N as small as 500.

Figure 2.2: Shape of the stationary distribution of the SLS model as N increases, for fixed values of u_1, u_2, s_1, and s_2.

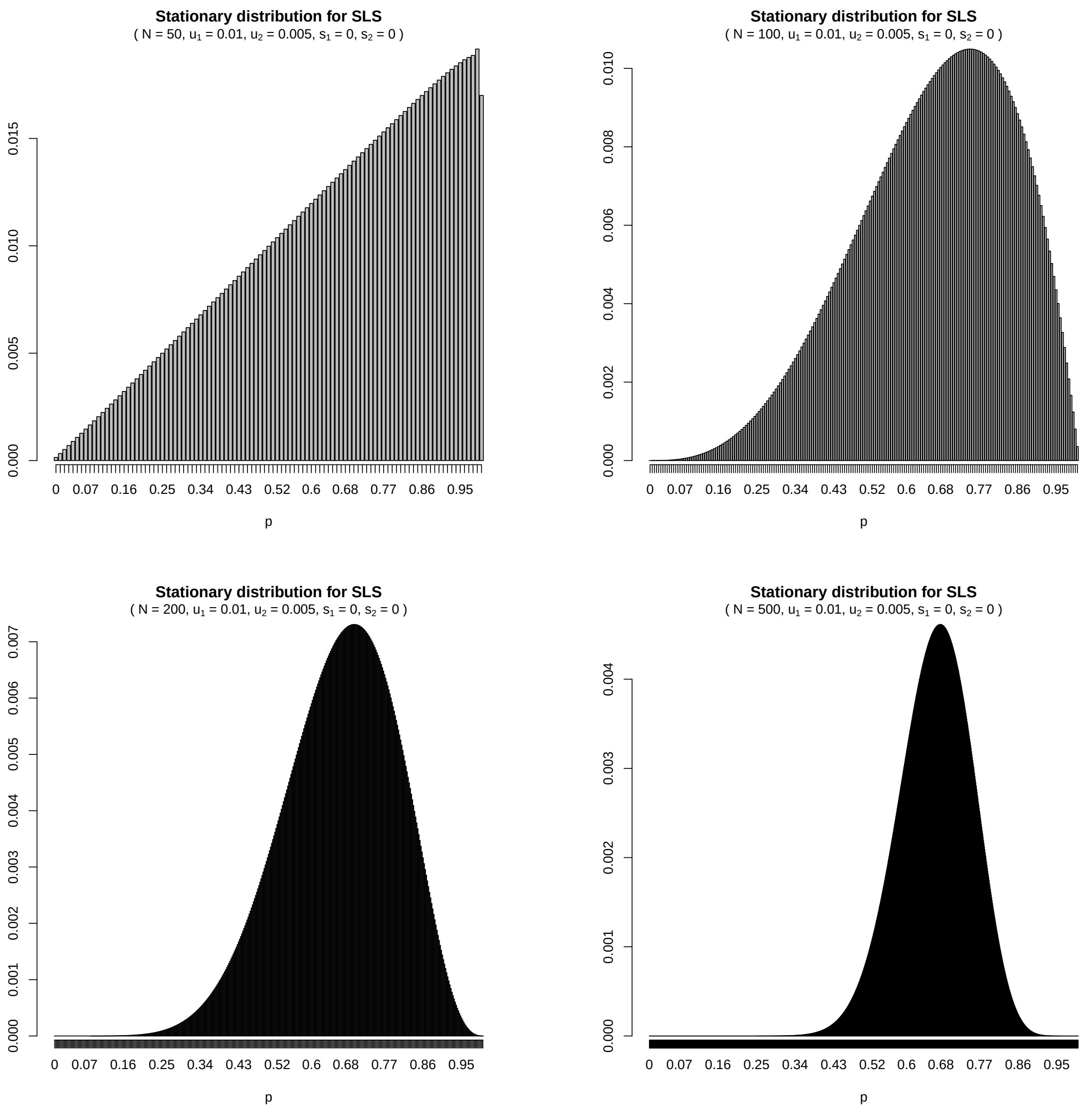
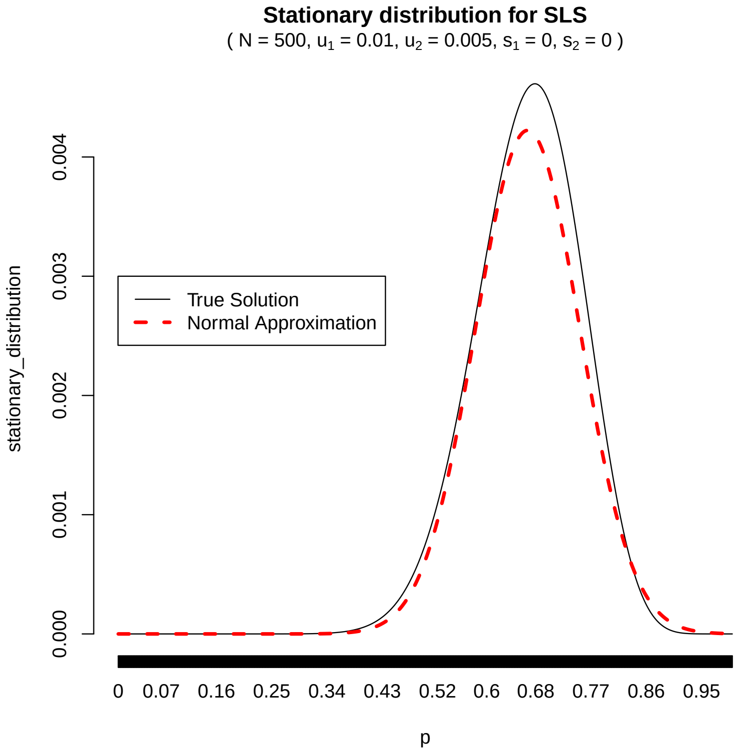


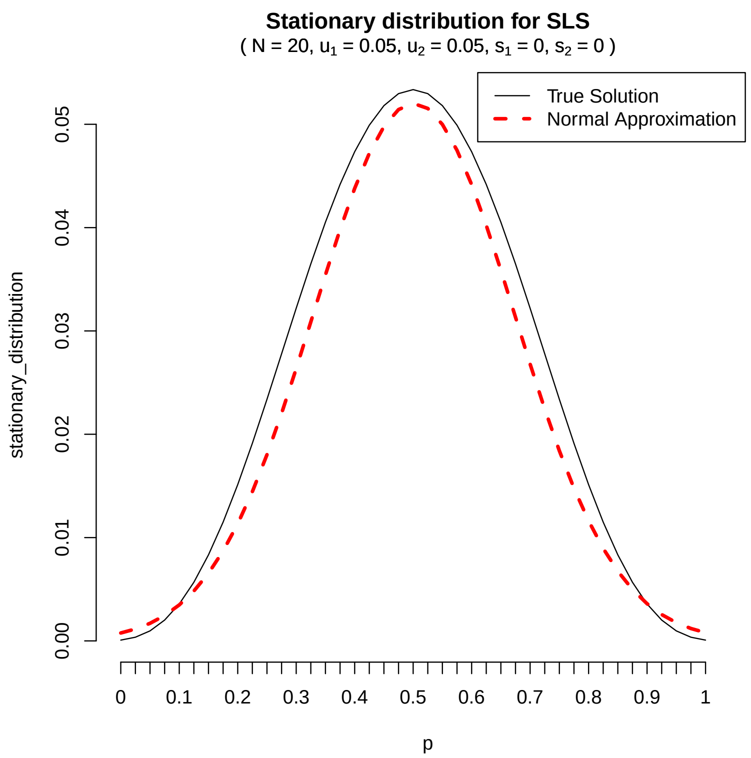
Figure 2.3: Normal approximation for the SLS model in the absence of selection when $N = 500$.



The normal approximation is not valid for moderate N, but it might be adequate for large N where we have difficulties with obtaining a numerical solution. **Figure 2.3** shows both the numerically calculated and normal approximation of the stationary distribution for $N = 500$. If we take the numerical solution to be the true solution, the normal approximation manages to capture the shape, but slightly underestimates the true solution at the peak, for N as small as 500.

The performance of the normal approximation depends on the size of the population parameters relative to the population size N. If the mutation rates u_1 and u_2 are equal or close, creating a reasonably symmetric stationary distribution, the normal approximation gives a reasonable approximation for N as small as 50. **Figure 2.4** shows a normal approximation against a symmetric stationary distribution for $N = 20$. A reasonable approximation is achieved even for this very small value of N.

Figure 2.4: Normal approximation for the SLS model in the absence of selection when $N = 20$.



Without knowing an explicit expression for the underlying discrete stationary distribution, it is only possible to determine the mean and variance for cases where the expected frequency of y is a linear function of x. In the presence of selection, the expected frequency of y becomes a rational function of x, therefore it is difficult to derive the mean and the variance in the same way.

The normal approximation is only a special tool for simple models with relatively large N. However, it shows that knowledge of just two moments of the stationary distribution could lead to a viable approximation to the whole distribution under certain conditions. We will pursue this idea further in the later chapters, where more than two moments are used to reconstruct the stationary distribution.

2.3 Summary

For various generalisations of Wright-Fisher models, solving the stationary conditions analytically can only be done when N is very small. The complexity of the analytic expression for the stationary distribution, as well as the computational cost of obtaining it, grow quickly as N increases. Therefore, this method is neither practical nor useful in general.

For reasonably large N, numerical routines can be used to solve the stationary conditions directly. This offers simple and useful output for investigating the effects of population parameters on the stationary distribution. However, as N increases to become arbitrarily large, numerical solutions gradually lose both accuracy and efficiency. They are particularly problematic for the stationary distribution of a k-dimensional random variable, where the number of possible states grows rapidly in an exponential fashion as N and k both increase.

For arbitrarily large N, the normal approximation offers a simple solution for some models where the two previous approaches cease to be feasible. However, the normal approximation is not feasible for all models, and its performance varies from case to case.

Fortunately there exists a general approach for an arbitrarily large N, which proves to be adequate for a wide range of models in population genetics. This involves approximating the discrete process of genetic drift by a continuous-time continuous-space diffusion process. It has proved to be a robust approach for many models and has been widely used. We will provide a literature review on this subject in the next chapter.

3

Diffusion Approximation

In this chapter, we first provide a history of the development of diffusion theory in population genetics. We then outline its fundamental idea and state the stationary condition under the diffusion approximation. We show an intuitive derivation of some relevant diffusion operators in the subsequent section. In the later sections of this chapter, we provide a literature review of stationary distributions of diffusion processes. Finally, we describe a relevant recent development in diffusion theory.

3.1 Basics

3.1.1 A brief history

The mathematics of the discrete Wright-Fisher processes can be greatly simplified by considering their continuous counterparts. The discrete random process of genetic drift in finite populations due to sexual reproduction was first approximated mathematically using a continuous random process by **Fisher et al. (1922, 1930)**. A differential equation, later known as the Kolmogorov equation or the diffusion equation, was introduced to describe the continuous random process. This continuous process, later named the diffusion process, was further studied by **Wright (1931, 1945)** to obtain stationary distributions, and he laid the foundations for further usage of the diffusion process in population genetics.

Diffusion theory in population genetics was substantially extended and many results were verified by Motoo Kimura. For a summary paper see **Kimura (1964)**, and see **Watterson (1996)** for a detailed review of Kimura's contribution. After his pioneering work in the 1950s, there was a series of papers on single locus models by himself and others, see **Kimura (1962)**, **Kimura and Crow (1964)**, **Ewens (1964a)**, **Kimura and Ohta (1969)** and **Ewens (1972)**. Later Kimura and his student Ohta focused on multiple loci and the interaction between them, see **Ohta and Kimura (1969a, 1969b, 1970, 1971)**.

Besides aiding study of important quantities in single locus and multiple locus models, such as the fixation probabilities, the expected time to fixation, the degree of polymorphism supported by mutation, linkage disequilibrium, and so on, the diffusion approximation also allows stationary and transient distributions to be recovered for various models in population genetics. For transient distributions under diffusion, see **Kimura (1955a, 1956)**, **Ewens (1963a, 1963b, 1964b)** and **Griffiths (1979, 1980, 1981)**. To avoid duplication we provide a literature review on stationary distributions in the later **Section 3.2**.

3.1.2 Intuition

In population genetics, the rationale behind the diffusion approximation is based on three observations. Firstly, when considering the process of genetic sampling for a species from one mating season to another on an evolutionary time scale, the time interval between two mating seasons seems infinitesimal compared with time frame of its existence. Therefore time can be treated as a continuous variable, and there is no concept of generation needed mathematically.

Secondly, we can choose to measure time in units that relate to the population size N, because the time can be treated as a continuous variable. The purpose of this will be clear shortly.

Lastly, when the population size N is small, there is only a small number of possible allele frequencies $\frac{x}{2N}$ can take between 0 and 1. However as N increases to be arbitrarily large, the gaps between possible values of $\frac{x}{2N}$ become arbitrarily small. Therefore allele frequency can be approximated as a continuous quantity for large N.

The diffusion approximation can be understood and constructed by studying the diffusion process through the Kolmogorov equations. There exists a substantial and rigorous theory of diffusion processes, however we are interested only in one particular area of its application, namely the expectations of functions of diffusion variables. Hence we only include a brief explanation of the relevant results and assume without derivation the existence and the uniqueness of the relevant diffusion process itself.

The Wright-Fisher model and its generalisations are discrete-time discrete-space Markov chains. The diffusion approximation for these models uses the three observations above to approximate a discrete-time discrete-space Markov chain by a continuous-time continuous-space process. In terms of the expectations of functions of diffusion variables, the key results can be understood and derived by rescaling the time and the space axes of the original discrete Markov chain, and making a link between the time units and the population size.

For our purpose, the diffusion approximation provides a so-called master equation, which states that the expectation of a diffusion operator $\mathcal{L}$ acting on any well-behaved function Φ of the diffusion variables is always zero at stationarity (also known as being at equilibrium or in a steady state). That is to say, the expectation of the special quantity $\mathcal{L}\left\{\Phi\left(p\right)\right\}$ is always zero with respect to the stationary distribution of the diffusion variable p. We will derive this stationary condition in the next section. A master equation has the following form:

$$\mathbb{E}\left[\mathcal{L}\left\{\Phi\left(p\right)\right\}\right] = 0\,, \tag{3.1}$$

where the diffusion variable p is defined as:

$$p = \lim_{N\to\infty} \frac{x}{2N}\,.$$

To derive a diffusion operator $\mathcal{L}$, and hence the master equation for any particular model, we need to consider the original discrete Markov chain as N approaches infinity.

3.1.3 Diffusion operator

There are many references regarding the derivation of the diffusion approximation. **Kimura et al. (1955b)** gives an elementary derivation based on the geometric interpretation of the process involved, and a more mathematical version in terms of the Kolmogorov equation is given by **Kimura (1964)**. A series of papers, **Ethier and Nagylaki (1980, 1988, 1989)**, give a more general and rigorous treatment. **Ewens (2004)** also provides an intuitive derivation, but many details are omitted.

Here, we follow the approach in **Ewens (2004)**, however we provide some additional details to clarify and explain the derivation. The SLS model is used to demonstrate the derivation of the corresponding diffusion operator given the discrete transition matrix. Using this derivation, we hope to illustrate that deriving and applying the diffusion approximation can be quite simple and requires only limited mathematics, and the reason for the stationary condition in **Equation 3.1** is intuitive.

The SLS model defines a Markov chain with discrete state space $\{0, 1, 2, \ldots, 2N\}$ over time space $\{0, 1, 2, \ldots\}$. Let x_0, x and y denote the initial state at time 0, intermediate state at time t, and final state at time $t + 1$ respectively. Let us rescale the state space by a factor of $(2N)^{-1}$, which gives the new variables p_0, p and δp defined as,

$$p_0 = \frac{x_0}{2N} \tag{3.2}$$

$$p = \frac{x}{2N} \tag{3.3}$$

$$p + \delta p = \frac{y}{2N} \,. \tag{3.4}$$

Now let us rescale the time space by the same factor of $(2N)^{-1}$, and consider a new Markov chain evolving over time points $\{0\delta t, 1\delta t, 2\delta t, \ldots, t - \delta t, t, t + \delta t, \ldots\}$, where

$$\delta t = \frac{1}{2N} \,. \tag{3.5}$$

We will later consider p_0, p, $p+\delta p$ and δt as $N \to \infty$, but they remain discrete for now.

Let $h(t;p_0)$ denote the expectation of $\Phi(p)$ at time t given the initial state p_0, where $\Phi(p)$ is any well behaved function which depends only on p. Thus we have the following:

$$h(t;p_0) = \sum_p \mathbb{P}(p \mid p_0, t)\, \Phi(p) \tag{3.6}$$

$$= \mathbb{E}_p \{\Phi(p)\} \ , \tag{3.7}$$

where $\mathbb{P}(p \mid p_0, t)$ denotes the probability of being at state p at time t given the initial state being p_0 at time 0, and $\mathbb{E}_p$ denotes the corresponding expectation with respect to p.

Notice that we omit p_0 and t from the notation in **Equation 3.7** , to avoid notational clutter. The function $\Phi(p)$ depends on neither the time t nor the initial state p_0, but the probability $\mathbb{P}(p \mid p_0, t)$ does, and therefore $h(t;p_0)$ depends on the time and the initial state. Therefore this notation is not to indicate that time is not relevant, but merely for notational convenience.

Now let us consider the function h at the time $t+\delta t$. An increment in terms of time δt induces an increment in state of δp and hence we have the following,

$$h(t+\delta t;p_0) = \mathbb{E}_{p,\delta p} \{\Phi(p+\delta p)\} \ . \tag{3.8}$$

If we separate the jump into an intermediate step p and a final step δp, then we can use the law of total expectation to give:

$$h\left(t+\delta t;p_0\right)=\mathbb{E}_p\left[\mathbb{E}_{\delta p|p}\left\{\Phi\left(p+\delta p\right)\right\}\right]\ . \tag{3.9}$$

Suppose N is very large, and thus δp would almost be continuous around zero. Therefore, we can approximate $\Phi\left(p+\delta p\right)$ using a Taylor series expansion centered at the point of origin $\delta p=0$,

$$\begin{aligned} h\left(t+\delta t;p_0\right) &= \mathbb{E}_p\left[\mathbb{E}_{\delta p|p}\left\{\Phi\left(p\right)+\delta p\Phi'\left(p\right)+\frac{1}{2}\left(\delta p\right)^2\Phi''\left(p\right)+R_2\left(\delta p;p\right)\right\}\right] \\ &= \mathbb{E}_p\left[\Phi\left(p\right)+\mathbb{E}_{\delta p|p}\left(\delta p\right)\Phi'\left(p\right)+\frac{1}{2}\,\mathbb{E}_{\delta p|p}\left\{\left(\delta p\right)^2\right\}\Phi''\left(p\right)\right]+\mathbb{E}_{p,\delta p}\left\{R_2\left(\delta p;p\right)\right\}\ , \end{aligned} \tag{3.10}$$

where $R_2\left(\delta p;p\right)$ is the remainder term for the second-order Taylor series expansion.

We need to evaluate the conditional expectations $\mathbb{E}_{\delta p|p}\left(\delta p\right)$ and $\mathbb{E}_{\delta p|p}\left\{\left(\delta p\right)^2\right\}$. Applying the standard formula for the mean of a binomial distribution, and using **Equations 3.2–3.4**, we have the following,

$$\begin{aligned} \mathbb{E}_{\delta p|p}\left(\delta p\right) &= \frac{1}{2N}\,\mathbb{E}_{y|x}\left(y\right)-p \\ &= \varphi\left(x\right)-p\,, \end{aligned} \tag{3.11}$$

where $\varphi\left(x\right)$ is defined by **Equations 2.12** and **2.14**.

Rewriting **Equations 2.12** and **2.14** to give $\varphi\left(x\right)$ in terms of p, $\mathbb{E}_{\delta p|p}\left(\delta p\right)$ is,

$$\mathbb{E}_{\delta p|p}\left(\delta p\right) = (1 - u_1 - u_2)\,\frac{(1+s_1)p^2 + (1+s_2)p(1-p)}{1 + s_1 p^2 + 2 s_2 p(1-p)} + u_1 - p. \tag{3.12}$$

We can use the Taylor series expansion again, this time on $\mathbb{E}_{\delta p|p}\left(\delta p\right)$ centered at the point of zero for all the population parameters ($u_1 = 0$, $u_2 = 0$, $s_1 = 0$ and $s_2 = 0$). This gives the following,

$$\begin{aligned}\mathbb{E}_{\delta p|p}\left(\delta p\right) = {} & (1-p)\,u_1 - p u_2 + p^2\,(1-p)\,s_1 \\ & + p\,(1-p)\,(1-2p)\,s_2 + O\left(\frac{1}{N^2}\right)\end{aligned} \tag{3.13}$$

where $O(z)$ is the usual big O notation (Landou notation). We note that Maple is used to do Taylor series expansions and most of the algebraic manipulations in this thesis.

The population parameters u_1, u_2, s_1 and s_2 are generally assumed to be $O\left(\frac{1}{N}\right)$. Thus higher order terms in the Taylor series in **Equation 3.13** are denoted by $O\left(\frac{1}{N^2}\right)$. This relationship between the population parameters and N is artificial: it is a result of linking time δt and the population size N together, that is,

$$\frac{1}{2N} = \delta t\,.$$

Biologically, a link between the time per generation and the parameters, u_1, u_2, s_1 and s_2, seems more sensible than this link with N; more time per generation means more chance for mutation and selection forces to act on the population during each generation.

Let θ_1 and θ_2 denote the scaled mutation rates and α_1 and α_2 give the scaled selection rates, where $\theta_1 = 2Nu_1$, $\theta_2 = 2Nu_2$, $\alpha_1 = 2Ns_1$ and $\alpha_2 = 2Ns_2$. The expectation $\mathbb{E}_{\delta p|p}\left(\delta p\right)$ is given by,

$$\begin{aligned}\mathbb{E}_{\delta p|p}\left(\delta p\right) = & \frac{\left\{(1-p)\,\theta_1 - p\theta_2\right\}}{2N} \\ & + \frac{p\,(1-p)\left\{p\alpha_1 + (1-2p)\,\alpha_2\right\}}{2N} + O\left(\frac{1}{N^2}\right). \end{aligned} \tag{3.14}$$

Similarly, we can approximate $\mathbb{E}_{\delta p|p}\left\{(\delta p)^2\right\}$ using a Taylor series expansion. If we take the standard formula for the variance of the binomial distribution, and rearrange for p, this expansion around the origin is given by,

$$\begin{aligned}\mathbb{E}_{\delta p|p}\left\{(\delta p)^2\right\} &= \left(\frac{1}{2N}\right)^2\left[\mathbb{V}\mathrm{ar}_{y|x}\left(y\right) + \left\{\mathbb{E}_{y|x}\left(y\right)\right\}^2\right] - \frac{p}{N}\,\mathbb{E}_{y|x}\left(y\right) + p^2 \\ &= \frac{p\,(1-p)}{2N} + \frac{s_1}{2N}\,(1-2p)\,p^2\,(1-p) + \frac{s_2}{2N}\,(1-2p)^2\,(1-p) \\ &\quad + \frac{u_1}{2N}\,(1-2p)\,(1-p) - \frac{u_2}{2N}\,(1-2p)\,p^2 + R_2\left(u_1, u_2, s_1, s_2; p\right) \end{aligned} \tag{3.15}$$

Again, population parameters u_1, u_2, s_1 and s_2 are assumed to be $O\left(\frac{1}{N}\right)$, thus terms involving $\frac{u_1}{2N}$, $\frac{u_2}{2N}$, $\frac{s_1}{2N}$ and $\frac{s_2}{2N}$ or higher order terms are $O\left(\frac{1}{N^2}\right)$. Therefore, we have the following,

$$\mathbb{E}_{\delta p|p}\left\{(\delta p)^2\right\} = \frac{p\,(1-p)}{2N} + O\left(\frac{1}{N^2}\right) \tag{3.16}$$

Now, substituting $\mathbb{E}_{\delta p|p}\left(\delta p\right)$ in **Equation 3.14** and $\mathbb{E}_{\delta p|p}\left\{\left(\delta p\right)^2\right\}$ in **Equation 3.16** into **Equation 3.10**, we get,

$$h\left(t+\delta t;p_0\right)=\mathbb{E}_p\left\{\Phi\left(p\right)\right\}+\frac{1}{2N}\,\mathbb{E}_p\left[\left\{M(p)+O\left(\frac{1}{N}\right)\right\}\Phi'\left(p\right)\right]$$
$$+\frac{1}{2N}\,\mathbb{E}_p\left[\left\{\frac{1}{2}V(p)+O\left(\frac{1}{N}\right)\right\}\Phi''\left(p\right)\right]+\mathbb{E}_{p,\delta p}\left\{R_2\left(\delta p;p\right)\right\}\ , \tag{3.17}$$

where $M(p)$ and $V(p)$ are given by **Equation 3.18–3.19**.

$$M(p)=\left\{\left(1-p\right)\theta_1-p\theta_2\right\}+p\left(1-p\right)\left\{p\alpha_1+\left(1-2p\right)\alpha_2\right\}\ , \tag{3.18}$$

$$V(p)=p\left(1-p\right)\ . \tag{3.19}$$

The term $M(p)$ is known as the drift coefficient or the mean of the diffusion variable, and the term $V(p)$ is known as the diffusion coefficient or the variance of the diffusion variable.

In general, it is assumed that the higher order moments of δp approach zero as $N\to\infty$. More precisely, the following is assumed,

$$\mathbb{E}_{\delta p|p}\left(\left|\delta p\right|^i\right)=o\left(\frac{1}{N}\right),\qquad \text{where}\quad i\geq 3. \tag{3.20}$$

Because the remainder term $\mathbb{E}_{p,\delta p}\left\{R_2\left(\delta p;p\right)\right\}$ is a polynomial of the higher order moments with the derivatives of $\Phi\left(p\right)$, which are functions only of p, the following must be true:

$$\mathbb{E}_{p,\delta p}\left\{R_2\left(\delta p;p\right)\right\} = O\left(\frac{1}{N}\right). \tag{3.21}$$

Simplifying and rearranging **Equation 3.17** gives,

$$\frac{h\left(t+\delta t;p_0\right)-\mathbb{E}_p\left\{\Phi\left(p\right)\right\}}{\frac{1}{2N}} = \mathbb{E}_p\left\{M(p)\Phi'\left(p\right)+\frac{1}{2}V(p)\Phi''\left(p\right)\right\}+O\left(\frac{1}{N}\right), \tag{3.22}$$

where $M(p)$ and $V(p)$ are defined by **Equations 3.18–3.19**.

By definition, $\mathbb{E}_p\left\{\Phi\left(p\right)\right\} = h\left(t;p_0\right)$ and $\frac{1}{2N} = \delta t$, and so,

$$\frac{h\left(t+\delta t;p_0\right)-h\left(t;p_0\right)}{\delta t} = \mathbb{E}_p\left\{M(p)\Phi'\left(p\right)+\frac{1}{2}V(p)\Phi''\left(p\right)\right\}+O\left(\frac{1}{N}\right), \tag{3.23}$$

where $M(p)$ and $V(p)$ are defined by **Equations 3.18–3.19**.

Now let consider $N \to \infty$, so $\delta t \to 0$. Then the limit of the left hand side of **Equation 3.23** is,

$$\lim_{\delta t\to 0}\frac{h\left(t+\delta t;p_0\right)-h\left(t;p_0\right)}{\delta t} = \frac{dh\left(t;p_0\right)}{dt}.$$

As $N \to \infty$, the higher order terms on the right hand side disappear. Thus we have the following,

$$\frac{dh\,(t;p_0)}{dt} = \mathbb{E}_p \left\{ M(p)\Phi'\,(p) + \frac{1}{2}V(p)\Phi''\,(p) \right\}\,, \tag{3.24}$$

where $M(p)$ and $V(p)$ are defined by **Equations 3.18–3.19**.

Equation 3.24 actually means the following in a more explicit notation:

$$\frac{dh\,(t;p_0)}{dt} = \sum_p \mathbb{P}\,(p \mid p_0, t) \left\{ M(p)\Phi'\,(p) + \frac{1}{2}V(p)\Phi''\,(p) \right\}\,. \tag{3.25}$$

We have omitted t and p_0 from the notation for convenience, because they do not affect any of the expansions or orders of approximation in the previous derivation. However, it is important to note that t and p_0 do affect the distribution with respect to which $\mathbb{E}_p$ is calculated. This dependency on t is actually the reason that we can reach **Equation 3.24**, which is certainly not true if t and p are discrete. However, as $N \to \infty$ and thus $\delta t \to 0$, the distribution gradually changes and approaches a continuous distribution which makes **Equation 3.24** exact. This is the reason why other authors use the term "the diffusion limit of a certain model" for the diffusion approximation of the model.

We only consider processes that admit a stationary distribution, in which case an important observation to make is that the rate of change of the expectation of any well-behaved function of the random variable with respect to time must be zero at stationarity. That is,

$$\frac{dh\,(t;p_0)}{dt} = 0\,. \tag{3.26}$$

This completes the derivation of the diffusion operator $\mathcal{L}$ for the SLS model:

$$\mathcal{L} = M(p)\frac{\partial}{\partial p} + \frac{1}{2}V(p)\frac{\partial^2}{\partial p^2}\,, \tag{3.27}$$

where $M(p)$ and $V(p)$ are defined by **Equations 3.18–3.19**.

Therefore, the master equation for the SLS model is,

$$\mathbb{E}_p\left\{M(p)\frac{\partial\Phi\,(p)}{\partial p} + \frac{1}{2}V(p)\frac{\partial^2\Phi\,(p)}{\partial p^2}\right\} = 0\,, \tag{3.28}$$

where $M(p)$ and $V(p)$ are defined by **Equations 3.18–3.19**, and $\Phi\,(p)$ is any well-behaved function of p only.

Other models possessing a single diffusion variable p will have a master equation of the same form, but with different $M(p)$ and $V(p)$.

A generalisation of the master equation for a multi-dimensional diffusion variable is straightforward. For linearly independent random variables x_1, x_2, …, x_{k-1}, it has the following form:

$$\mathbb{E}_{\boldsymbol{p}}\left\{\sum_{i=1}^{k-1} M_i(\boldsymbol{p})\frac{\partial\Phi\,(\boldsymbol{p})}{\partial p_i} + \frac{1}{2}\sum_{i=1}^{k-1} V_i(\boldsymbol{p})\frac{\partial^2\Phi\,(\boldsymbol{p})}{\partial p_i^2} + \sum_{i=1}^{k-2}\sum_{j>i}^{k-1} W_{ij}(\boldsymbol{p})\frac{\partial^2\Phi\,(\boldsymbol{p})}{\partial p_i\partial p_j}\right\} = 0\,, \tag{3.29}$$

where $\boldsymbol{p}$ is a vector of the diffusion variables p_i for the corresponding discrete random variable $\boldsymbol{x}$, and $M_i(\boldsymbol{p})$, $V_i(\boldsymbol{p})$ and $W_{ij}(\boldsymbol{p})$ are the mean, the variance and the covariance of the corresponding diffusion variables respectively.

The master equation for the SLM model and the TLD model both have the above form, but each incorporates a different set of $M_i(\boldsymbol{p})$, $V_i(\boldsymbol{p})$ and $W_{ij}(\boldsymbol{p})$, and $\boldsymbol{p}$ has a different biological meaning in each case. Terms $M_i(\boldsymbol{p})$, $V_i(\boldsymbol{p})$ and $W_{ij}(\boldsymbol{p})$ can be determined from the corresponding conditional expectations $\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left(\delta p_i\right)$, $\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left\{\left(\delta p_i\right)^2\right\}$, and $\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left(\delta p_i \delta p_j\right)$. See **Appendix A.1** for the corresponding conditional expectations of SLM, and **Appendix A.2** for TLD.

For the SLM model, we have the following,

$$M_i(\boldsymbol{p}) = \theta\left(1 - kp_i\right) \tag{3.30}$$

$$V_i(\boldsymbol{p}) = p_i(1 - p_i) \tag{3.31}$$

$$W_{ij}(\boldsymbol{p}) = -p_i p_j\,, \tag{3.32}$$

where k is the number of distinct allelic types, and θ is the scaled mutation rate ($\theta = 2Nu$). The terms p_1, p_2, …, p_{k-1} are the corresponding diffusion variables for the discrete random variables x_1, x_2, …, x_{k-1}. In this case, the discrete random variables x_1, x_2, …, x_{k-1} are defined as allele type counts for alleles A_1, A_2, …, A_{k-1} respectively in **Subsection 2.1.2**.

For the TLD model, we have the following,

$$M_1(\boldsymbol{p}) = -2\theta p_1 + \theta p_2 + \theta p_3 - \rho D\left(\boldsymbol{p}\right) \tag{3.33}$$

$$M_2(\boldsymbol{p}) = \theta - 3\theta p_2 - \theta p_3 + \rho D\left(\boldsymbol{p}\right) \tag{3.34}$$

$$M_3(\boldsymbol{p}) = \theta - \theta p_2 - 3\theta p_3 + \rho D\left(\boldsymbol{p}\right) \tag{3.35}$$

$$V_i(\boldsymbol{p}) = p_i(1 - p_i) \tag{3.36}$$

$$W_{ij}(\boldsymbol{p}) = -p_i p_j\,, \tag{3.37}$$

where θ is the scaled mutation rate ($\theta = 2Nu$) and ρ is the scaled recombination rate ($\rho = 2NC$). The terms p_1, p_2 and p_3 are the corresponding diffusion variables for the discrete random variables x_1, x_2 and x_3. In this case, the discrete random variables x_1, x_2 and x_3 are defined as gamete type counts for A_1B_1, A_1B_2 and A_2B_1 respectively in **Subsection 2.1.4**. Here, $D(\boldsymbol{p})$ is the usual coefficient of linkage disequilibrium, $D(\boldsymbol{p}) = p_1 - p_1^2 - p_1p_2 - p_1p_3 - p_2p_3$, corresponding to **Equation 2.21** in the discrete case. The random variable x_4 for gamete type A_2B_2, and its counterpart p_4, is not included because $x_4 = 1 - x_1 - x_2 - x_3$.

Traditionally, the diffusion approximation for the TLD model is transformed from p_1, p_2 and p_3 into p, q and D, where

$$
\begin{aligned}
p &= p_1 + p_2 \\
q &= p_1 + p_3 \\
D &= p_1 - p_1^2 - p_1p_2 - p_1p_3 - p_2p_3 \,.
\end{aligned}
$$

Here, p and q are the continuous allelic frequencies for A_1 and B_1 at loci A and B respectively, and D is the usual coefficient of linkage disequilibrium.

Initially, **Ohta and Kimura (1969a)** used this transformation to obtain the variance of D. Most recently **Song and Song (2007)** used the same transformation to evaluate the expectation of r^2. The parametrisation p, q and D is convenient to work with when we are dealing with D, but it is not convenient when we are reconstructing the distribution. This is because p, q and D together have an irregular support. Reconstructing a distribution on a regular region, such as the support of p_1, p_2 and p_3, is usually much simpler. Therefore we will use the original parametrisation p_1, p_2 and p_3 instead of the usual parametrisation

p, q and D for the TLD model under the diffusion approximation.

3.2 Stationary Distribution

3.2.1 Classical approach

For a single locus model, a series of papers by **Wright (1931, 1937, 1945)** and **Kimura (1964)** provide a general approach for finding the stationary distribution of allele proportions under the diffusion approximation. They show that the stationary distribution $\pi(p)$ under the diffusion approximation has the form,

$$\pi(p) = \frac{1}{\Lambda V(p)} \exp\left(2 \int \frac{M(p)}{V(p)}\, dp\right), \tag{3.38}$$

where Λ denotes a normalising constant, which is generally unknown in closed form, and $M(p)$ and $V(p)$ are the mean and the variance of the diffusion variable under consideration.

The continuous distribution $\pi(p)$ is an approximation to the discrete $\boldsymbol{\pi}$ in **Chapter 2**, but it is the exact solution under the diffusion approximation. That is, $\pi(p)$ is the exact solution to the Kolmogorov equation or the diffusion equation at steady state.

Using $M(p)$ and $V(p)$ in **Equations 3.18–3.19**, the stationary distribution for SLS under the diffusion approximation is given by,

$$\pi(p) = \frac{1}{\Lambda} p^{2\theta_1 - 1}(1-p)^{2\theta_2 - 1} \exp\left(2\alpha_2 p + (\alpha_1 - 2\alpha_2)\, p^2\right), \tag{3.39}$$

where Λ is determined such that $\displaystyle\int_0^1 \pi(p)\, dp = 1$.

The stationary distribution for the SLM model under the diffusion approximation has the following form; see **Wright (1968)** for more details:

$$\pi\left(p\right) = \frac{\Gamma\left(2k\theta\right)}{\left\{\Gamma\left(2\theta\right)\right\}^k} \prod_{j=1}^{k} p_j^{2\theta - 1} \, . \tag{3.40}$$

Hence, $\pi\left(p\right)$ is a symmetric Dirichlet distribution $\mathrm{Dir}\left(2\theta\right)$ for the SLM model under the diffusion approximation.

For multiple-locus models, in which all loci are *unlinked*, the stationary distribution under the diffusion approximation for the diallelic case is given by **Wright (1937)**, and for the multiallelic case by **Wright (1949)**. Diffusion theory for multiple loci was also studied by **Kimura et al. (1955b)** and **Ethier (1979)**, but for many years no stationary distribution under the diffusion approximation was found for models involving *linked* loci such as the TLD model.

Although the stationary distribution was not found, some other important quantities were evaluated for multiple loci in the meantime, such as the variance of linkage disequilibrium D for the TLD model at steady state by **Ohta and Kimura (1969a)**, and fixation times and probabilities for an independent-locus model by **Littler and Good (1978)**. The first result for the stationary distribution of linked-locus models under the diffusion approximation occurred ten years later, when **Ethier and Nagylaki (1989)** obtained the stationary distribution for two models of linked loci under certain conditions. However, there still exists no general approach for models involving multiple linked loci.

3.2.2 Recent developments

The diffusion approximation has seen a recent resurgence in popularity: see **Cherry and Wakeley (2003)**, **Song and Song (2007)**, **Jenkins and Song (2009)**, and most recently **Etheridge and Lemaire (2011)**. This PhD project was initiated by one of these developments. **Song and Song (2007)** proposed an elegant procedure to compute the expectation of the linkage disequilibrium coefficient r^2 for the TLD model at steady state. We were largely inspired by their idea of breaking r^2 into an infinite series of monomials, and evaluating the expectation of each using the diffusion approximation:

$$\begin{aligned} r^2 &= \frac{D^2}{p\,(1-p)\,q\,(1-q)} \\ &= 4\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} D^2 p^m q^n\,, \end{aligned} \tag{3.41}$$

where p and q are the allelic frequencies for allele type A_1 and B_1 at locus A and B respectively,

$$p = p_1 + p_2 \quad , \quad q = p_1 + p_3\,,$$

and D is the usual coefficient of linkage disequilibrium,

$$D = p_1 - p_1^2 - p_1 p_2 - p_1 p_3 - p_2 p_3\,.$$

Partial fraction decomposition, convergent series expansion, and the symmetry of the

TLD model are used to reach the expansion in **Equation 3.41**; see **Song and Song (2007)** for details. Therefore their method is very specific for evaluating $\mathbb{E}\left(r^2\right)$, and it cannot be easily extended to evaluate other expectations. However, the idea behind their method can be generalised.

In **Song and Song (2007)**, the expectation of r^2 at steady state is evaluated without first finding the stationary distribution. Intuitively, each moment $\mathbb{E}\left(D^2 p^m q^n\right)$ provides a small piece of information regarding the stationary distribution, and the infinite series provides the information regarding r^2 (a complicated function of genetic composition). Putting these together gives the expectation of r^2 at steady state. The method treats the diffusion approximation as a tool for evaluating expectations, rather than a description of the continuous reproduction process.

The motivation behind our approach is that a series of expectations may contain all the information we require to compute the entire stationary distribution. Whereas **Song and Song (2007)** need a series of expectations to evaluate the expectation of a complicated function of genetic composition, we shall attempt to reconstruct the entire stationary distribution using a similar series of expectations.

4

Method for Reconstructing Distributions

In this chapter, we first describe a method for finding expectations in analytic form using the diffusion approximation without first finding the stationary distribution, for models in population genetics. We then study methods of reconstructing a density function using a sequence of its moments. Next we focus on one of the methods of reconstructing a density function, namely the maximum entropy principle (Maxent). In the final section, we propose a new method of reconstructing a density function to incorporate not only numerical values of its expectations but also analytic formulae of its expectations, therefore obtaining the analytic form of the density function in terms of the random variable and its parameter.

4.1 Stationary Moments

4.1.1 Single locus model

The method of finding relevant expectations at stationarity using the diffusion approximation without first finding the stationary distribution was pioneered by Ohta and Kimura. The idea is to choose a few relevant functions $\Phi(\boldsymbol{p})$ intelligently, so that applying the master equation with these functions generates a solvable system of linear equations in terms of the desired expectations. See **Ohta and Kimura (1969a, 1969b, 1970, 1971)** for details.

Taking the SLS model as an example, substituting $\Phi(p) = p^n$ into the master equation **Equation 3.28** gives the following recursive relationship regarding expectations at steady state:

$$
\begin{aligned}
&n\left(2\alpha_2 - \alpha_1\right)\mathbb{E}\left(p^{n+2}\right) + n\left(\alpha_1 - 3\alpha_2\right)\mathbb{E}\left(p^{n+1}\right) \\
&\qquad = n\left(\theta_1 + \theta_2 - \alpha_2 + \frac{n}{2} - \frac{1}{2}\right)\mathbb{E}\left(p^{n}\right) - n\left(\theta_1 + \frac{n}{2} - \frac{1}{2}\right)\mathbb{E}\left(p^{n-1}\right),
\end{aligned}
\tag{4.1}
$$

where n takes an integer value.

If selection forces are absent ($\alpha_1 = \alpha_2 = 0$), **Equation 4.1** immediately reduces to,

$$\mathbb{E}\left(p^n\right) = \frac{2\theta_1 + n - 1}{2\theta_1 + 2\theta_2 + n - 1}\,\mathbb{E}\left(p^{n-1}\right) . \tag{4.2}$$

As $\mathbb{E}\left(p^0\right) = 1$,

$$\mathbb{E}\left(p\right) = \frac{\theta_1}{\theta_1 + \theta_2} ,$$

where θ_1 and θ_2 are the scaled mutation rates.

Solving the recursion relation in **Equation 4.2** gives the following equation for the SLS model without selection at steady state:

$$\mathbb{E}\left(p^n\right) = \frac{\Gamma\left(2\theta_1 + n\right)\Gamma\left(2\theta_1 + 2\theta_2\right)}{\Gamma\left(2\theta_1 + 2\theta_2 + n\right)\Gamma\left(2\theta_1\right)} , \tag{4.3}$$

where $\Gamma\left(z\right)$ is the usual Gamma function.

If the mutation forces are equal ($\theta_1 = \theta_2 = \theta$), in addition to selection forces being absent ($\alpha_1 = \alpha_2 = 0$), the SLS model reduces to the diallelic SLM model ($k = 2$). Using **Equation 4.3** , we obtain the following equation for the diallelic SLM model at steady state:

$$\mathbb{E}\left(p^n\right) = \frac{\Gamma\left(2\theta + n\right)\Gamma\left(4\theta\right)}{\Gamma\left(4\theta + n\right)\Gamma\left(2\theta\right)} . \tag{4.4}$$

Let m_i denote the ith stationary moment,

$$m_i = \mathbb{E}\left(p^i\right) = \int_0^1 p^i \pi(p) dp\,, \tag{4.5}$$

where $\pi(p)$ is the corresponding stationary distribution. Let M_n denote the sequence of stationary moments up to and including the nth stationary moment,

$$M_n = \{m_i : i = 0,\ 1,\ 2,\ \ldots,\ n\}\ , \tag{4.6}$$

where the first element of the sequence, m_0, is always 1 by definition. The sequence M_n for the SLS model in the absence of selection can be generated using **Equation 4.3**, and the sequence M_n for the diallelic SLM model can be generated using **Equation 4.4**.

4.1.2 Two-locus model

It may also be possible to find the moments of a multivariate diffusion variable at steady state using a similar method. **Song and Song (2007)** recover the stationary moments in terms of p, q and D for the TLD model, where p and q are continuous allelic frequencies, and D is the usual coefficient of linkage disequilibrium. We now show how to recover the stationary moments in terms of p_1, p_2 and p_3 for the TLD model, where p_1, p_2 and p_3 are continuous gametic frequencies. See **Equation 3.29** and **3.33** for the TLD master equation in terms of p_1, p_2 and p_3.

Let $\boldsymbol{p^i}$ denote the following monomial,

$$\boldsymbol{p}^{\boldsymbol{i}} = p_1^{i_1} p_2^{i_2} p_3^{i_3} \,, \tag{4.7}$$

where $\boldsymbol{i} = (i_1, i_2, i_3)$, and i_1, i_2 and i_3 can only take non-negative integer values.

Let $\boldsymbol{P}_d$ denote the sequence of monomials of degree up to and including d, that is,

$$\boldsymbol{P}_d = \left\{ \boldsymbol{p}^{\boldsymbol{i}} : i_1 + i_2 + i_3 \leq d \right\} \,, \tag{4.8}$$

where d can only take non-negative integer values, and the elements of the sequence $\boldsymbol{P}_d$ follow a graded lexicographic order. First, the monomials are ordered based on the degree (the sum of all exponents). Any ties are broken by comparing the first exponent of p_1. If these are also equal, exponents of p_2 are compared, and so on. We refer to $\boldsymbol{P}_d$ as a sequence of monomials of order d.

The expectation of the monomial $\boldsymbol{p}^{\boldsymbol{i}}$ at steady state gives the stationary moments for the TLD model. Let $\boldsymbol{m}^{\boldsymbol{i}}$ denote the following stationary moments,

$$\boldsymbol{m}^{\boldsymbol{i}} = \int_{\Delta^3} \boldsymbol{p}^{\boldsymbol{i}} \pi\left(\boldsymbol{p}\right) d\boldsymbol{p} \,, \tag{4.9}$$

where $\displaystyle\int_{\Delta^3} d\boldsymbol{p}$ denotes the multiple integral over the standard 3-simplex region. The standard simplex region is defined as the following:

$$\Delta^3 = \left\{ (p_1, p_2, p_3) \in \mathbb{R}^3 : \sum_{i=1}^{3} p_i \leq 1 \quad \text{and} \quad p_i \geq 0 \quad \text{for} \quad i = 1, 2, 3 \right\} .$$

The sequence given by the expectation of each element in the sequence $\boldsymbol{P}_d$ at steady state represents a sequence of stationary moments for the TLD model following a graded lexicographic order. Let $\boldsymbol{M}_d$ denote the sequence of stationary moments of order up to and including d, that is,

$$\boldsymbol{M}_d = \left\{\boldsymbol{m}^{\boldsymbol{i}} : i_1 + i_2 + i_3 \leq d\right\}. \tag{4.10}$$

The number of monomials of the same degree i in the trivariate case is given by the binomial coefficient,

$$\binom{2+i}{i}.$$

Therefore the total number of elements in the sequence $\boldsymbol{P}_d$ is,

$$n_d = \sum_{i=0}^{d} \binom{2+i}{i}.$$

The sequence $\boldsymbol{M}_d$ corresponds to $\boldsymbol{P}_d$, thus it has the same number of elements.

Using Maple, we identified that the master equation defined by **Equations 3.29** and **3.33** can be used to find analytic formulae for $\boldsymbol{m}^{\boldsymbol{i}}$ for any $\boldsymbol{i}$. By analytic, we mean that $\boldsymbol{m}^{\boldsymbol{i}}$ is expressed in terms of the population parameters θ and ρ. When a monomial of degree d is substituted as $\Phi(\boldsymbol{p})$ into the master equation, it leads to a linear equation in terms of stationary moments of the orders $d+1$, d and $d-1$ in general. If we use a number of different monomials, we obtain a system of linear equations.

By examining the linear system using Maple, we identified that the system of linear equations constructed by putting every element of the sequence $\boldsymbol{P}_{2(d+1)}$ as $\Phi(\boldsymbol{p})$ in the master equation, is partially solvable up to and including $\boldsymbol{m}^{\boldsymbol{i}}$ of the order of d. That is, all the stationary moments of the sequence $\boldsymbol{M}_d$ can be determined by considering all the monomials of the sequence $\boldsymbol{P}_{2(d+1)}$ in the master equation. The TLD model is symmetric in the sense that the two loci are under the same mutation force, so some moments are equal. This symmetry is used while solving the system of equations. For example, the first order stationary moments for the TLD model are equal, because there is no force favouring any one gamete type (e.g. A_1B_1) over any of the others (A_1B_2, A_2B_1, or A_2B_2). Thus:

$$\mathbb{E}(p_1) = \mathbb{E}(p_2) = \mathbb{E}(p_3) = \frac{1}{4}\,. \tag{4.11}$$

Similarly, using Maple, we have the following for the second order stationary moments. The three expectations $\mathbb{E}(p_1^2)$, $\mathbb{E}(p_2^2)$ and $\mathbb{E}(p_3^2)$ are equal and have the following forms:

$$\frac{4096\theta^4 + 1536\theta^3\rho + 128\theta^2\rho^2 + 4352\theta^3 + 1216\rho\theta^2}{4\Lambda} + \frac{64\rho^2\theta + 1568\theta^2 + 304\rho\theta + 8\rho^2 + 216\theta + 26\rho + 9}{4\Lambda}\,.$$

The two expectations $\mathbb{E}(p_1p_2)$ and $\mathbb{E}(p_1p_3)$ are equal and have the following form:

$$\frac{\theta\left(1024\theta^3 + 384\rho\theta^2 + 32\rho^2\theta + 704\theta^2 + 192\rho\theta + 8\rho^2 + 144\theta + 26\rho + 9\right)}{\Lambda}\,.$$

The expectation $\mathbb{E}(p_2 p_3)$ has the following form:

$$\frac{8\theta^2 \left(128\theta^2 + 48\rho\theta + 4\rho^2 + 72\theta + 14\rho + 9\right)}{\Lambda}.$$

The term Λ for the three expressions of the second order moments is the same and has the following form:

$$\Lambda = (8\theta + 1)\left(2048\theta^3 + 768\rho\theta^2 + 64\rho^2\theta + 1280\theta^2 + 304\rho\theta + 8\rho^2 + 216\theta + 26\rho + 9\right).$$

Recall that θ is the scaled mutation rate, and ρ is the scaled recombination rate.

Notice that $\mathbb{E}(p_1 p_2) = \mathbb{E}(p_1 p_3) \neq \mathbb{E}(p_2 p_3)$. The terms $p_1 p_2$ and $p_1 p_3$ represent the proportions of individuals that are heterozygous at exactly one locus, respectively with genotypes A_1B_1/A_1B_2 and A_1B_1/A_2B_1. By contrast, $p_2 p_3$ represents individuals that are heterozygous at both loci with genotype A_1B_2/A_2B_1, so this quantity has a different expectation. These both differ from the expectations of p_1^2, p_2^2 and p_3^2, which represent individuals that are homozygous at both loci.

See **Table B.1** for additional moments of the TLD model. In general, we have observed that moments of TLD are rational functions of the scaled mutation rate θ and the scaled recombination rate ρ. However, a general closed formula for $\mathbb{E}\left(\boldsymbol{p}^{\boldsymbol{i}}\right)$ in terms of θ, ρ, i_1, i_2 and i_3 has not been found.

A summary of the procedure for finding the stationary moments of the TLD model in terms of p_1, p_2 and p_3 is given in **Table 4.1**.

Table 4.1: Procedures for determining the stationary moments for the TLD model.

1	Consider all monomials $\Phi(\boldsymbol{p}) = \boldsymbol{p}^{\boldsymbol{i}}$ for $i_1 + i_2 + i_3 \leq 2(d+1)$, and insert each in the master equation $\mathbb{E}\left[\mathcal{L}\left\{\Phi\left(p\right)\right\}\right] =0$. Obtain a $n_{2(d+1)} \times n_{2(d+1)}$ linear system $\boldsymbol{A}$ involving $\boldsymbol{m}^{\boldsymbol{i}}$ of order up to and including $2d+2$.
2	Use the symmetry in the TLD model to reduce the number of unknowns in $\boldsymbol{A}$.
3	Solve the reduced $\boldsymbol{A}$ to obtain all of the moments $\boldsymbol{m}^{\boldsymbol{i}}$ of order up to and including d.

4.2 Moment Problem

The problem of reconstructing a density function using knowledge of its moments is a special case of the problem of inverting an integral transform. Questioning the existence and the uniqueness, as well as finding such a density, is known as the moment problem, see **Shohat and Tamarkin (1943)**. In terms of our setting, it is known as the Hausdorff moment problem, because the underlying distribution is defined on a bounded simplex region.

Consider first a univariate probability density function $\pi(p)$, and a finite sequence of its moments

$$M_n = \{m_i : i = 0, 1, 2, \ldots, n\} \ ,$$

where m_i is defined in **Equation 4.5**.

We do not aim to recover the true underlying $\pi\left(p\right)$, but seek a general method of finding an approximation $\tilde{\pi}_n(p)$ which matches the first n moments,

$$m_i = \int_0^1 p^i \tilde{\pi}_n(p)\, dp \quad \text{for} \quad i = 0, 1 \ldots, n\,. \tag{4.12}$$

We also require $\tilde{\pi}_n$ to converge weakly to π as $n \to \infty$, that is,

$$\lim_{n\to\infty} \int_0^p \tilde{\pi}_n(z)\, dz = \int_0^p \pi(z)\, dz\,, \tag{4.13}$$

for all points $p \in (0, 1)$ at which $\int_0^p \pi(z)\, dz$ is continuous.

There are several possibilities for reconstructing such a $\tilde{\pi}_n(p)$. The first possibility is to use a certain type of orthogonal polynomial expanding $\pi(p)$,

$$\pi(p) = \lambda_0 b_0(p) + \lambda_1 b_1(p) + \lambda_2 b_2(p) + \cdots\,,$$

where $b_0(p), b_1(p), \ldots$ are members of some orthogonal polynomial basis and $\lambda_0, \lambda_1, \ldots$ are expansion coefficients.

The series is then truncated at $n+1$ terms, and the expansion coefficients are determined by solving the system of linear equations specified by the first n moment constraints; see **Kendall, Stuart and Ord (1991)**. The rate of convergence of $\tilde{\pi}_n(p)$ to $\pi(p)$ is largely affected by the choice of weight function with respect to which the orthogonal polynomial basis is defined,

$$0 = \int_0^1 b_i(p) b_j(p) w(p) dp \,,$$

where $w(p)$ is the weight function.

In practice, the selection of a weight function that allows for efficient and stable convergence is rather difficult in the absence of any additional knowledge regarding the unknown distribution $\pi(p)$. An inadequate choice of weight function leads to a poor choice of orthogonal polynomial, and in turn may lead to a highly oscillating approximation, non-positive measure or even a singular linear system.

Another alternative is the Padé approximation, which is widely used in the fields of physics and engineering to solve moment problems. Our understanding is that it is more powerful than the previous method, and has a stable convergence. See **Amindavar and Ritcey (1994)** for an application of the Padé approximation. The Padé approximation has been used very recently in the field of population genetics by **Jenkins and Song (2009)**, **Jenkins and Song (2011)**, and **Bhaskar and Song (2011)**.

4.2.1 Maximum entropy principle

A different and competitive approach, the maximum entropy principle, is considered in this thesis. It has gained attention in recent years in various scientific fields (**Tagliani (1999)**; **Wu (2003)**; **Abramov (2009)**), but it has not previously been used in population genetics, as far as we know. It is regarded as the least-biased solution to the moment problem in terms of entropy. The maximum entropy principle and its generalisation (Kullback's minimum cross entropy principle) has been developed in areas of statistical mechanics, computer science, economics, and finance. See **Mead and Papanicolaou**

(1984) and **Borwein and Lewis (1991)** for some results on the convergence of the maximum entropy (Maxent) approach.

The entropy of a distribution is understood to be a measure of uncertainty or ignorance. The fundamental idea of the maximum entropy principle is that the distribution with the maximum amount of uncertainty is the most honest choice after all constraints have been taken into account. See Appendix **C** for an intuitive discussion of entropy, its rationale, and the theory behind the maximum entropy principle.

Mathematically, the maximum entropy (Maxent) distribution $\tilde{\pi}_n(p)$ is the distribution that maximises the expectation of the negative logarithm of its own density function while satisfying all of the moment constraints. This gives rise to the following variational problem:

$$\begin{aligned} &\text{Maximise} \qquad I[\tilde{\pi}_n] = -\int_0^1 \tilde{\pi}_n(p) \ln \tilde{\pi}_n(p)\, dp\,, \\ &\text{Subject to} \qquad m_i = \int_0^1 p^i\, \tilde{\pi}_n(p)\, dp \quad \text{for} \quad i = 0, 1 \ldots, n\,. \end{aligned} \tag{4.14}$$

The corresponding Lagrange function is

$$L = -\int_0^1 \tilde{\pi}_n(p) \ln \tilde{\pi}_n(p)\, dp - \sum_{i=0}^{n} \lambda_i \left(\int_0^1 p^i \tilde{\pi}_n(p)\, dp - m_i \right). \tag{4.15}$$

Substituting this into the Euler-Lagrange equation allows the Maxent distribution $\tilde{\pi}_n(p)$ to be solved. In terms of our formulation, it leads to the univariate distribution

$$\tilde{\pi}_n(p) = \exp\left(\lambda_0 + \lambda_1 p + \lambda_2 p^2 + \lambda_3 p^3 + \cdots + \lambda_n p^n\right) , \tag{4.16}$$

where the λ_is are Lagrange multipliers and are obtained by solving the following unconstrained minimisation problem,

$$\underset{\boldsymbol{\lambda}}{\arg\min} \left\{ \int_0^1 \exp\left(\sum_{i=0}^{n} \lambda_i p^i\right) dp - \sum_{i=0}^{n} \lambda_i m_i \right\} , \tag{4.17}$$

where m_i is the ith order moment of $\pi(p)$ and $\boldsymbol{\lambda}$ is the vector containing values of λ_i.

Equation 4.16 is the general solution of Maxent given a sequence of moments of a univariate distribution. The Maxent solution $\tilde{\pi}_n(p)$ can be generalised to cover multivariate cases. For multivariate $\tilde{\pi}_d(\boldsymbol{p})$, the sequence of univariate monomials in **Equations 4.16** and **4.17** is replaced by a sequence of multivariate monomials, and the univariate moments are replaced by the necessary multivariate moments.

Given a sequence of trivariate moments of order up to and including d defined previously in **Section 4.1** for the TLD model,

$$\boldsymbol{M}_d = \left\{ \boldsymbol{m}^{\boldsymbol{i}} : i_1 + i_2 + i_3 \leq d \right\} ,$$

the Maxent solution $\tilde{\pi}_d(\boldsymbol{p})$ has the form

$$\tilde{\pi}_d(\boldsymbol{p}) = \exp\left(\sum_{j=0}^{d} \sum_{i_1=0}^{j} \sum_{i_2=0}^{j-i_1} \sum_{i_3=0}^{j-i_1-i_2} \lambda_{(i_1,i_2,i_3)} \boldsymbol{p}^{\boldsymbol{i}} \right) , \tag{4.18}$$

where $\boldsymbol{p^i}$ is defined by **Equation 4.7** and the $\lambda_{(i_1,i_2,i_3)}$s are obtained by solving the unconstrained minimisation problem

$$\arg\min_{\boldsymbol{\lambda}} \left\{ \int_{\Delta^3} \tilde{\pi}_d\left(\boldsymbol{p}\right) d\boldsymbol{p} - \sum_{j=0}^{d} \sum_{i_1=0}^{j} \sum_{i_2=0}^{j-i_1} \sum_{i_3=0}^{j-i_1-i_2} \lambda_{(i_1,i_2,i_3)} \boldsymbol{m^i} \right\}, \tag{4.19}$$

where $\boldsymbol{m^i}$ is defined by **Equation 4.9** and $\boldsymbol{\lambda}$ denotes the vector containing values of $\lambda_{(i_1,i_2,i_3)}$.

Let $\tilde{\pi}_n\left(p\right)$ and $\tilde{\pi}_d\left(\boldsymbol{p}\right)$ denote respectively the univariate and the multivariate Maxent distribution for the rest of the thesis.

4.3 Numerical Maximum Entropy Solution

We have sequences of stationary moments in analytic form for the SLS model in the absence of selection, for the diallelic SLM model, and for the TLD model, see **Section 4.1**. We refer to these sequences as sequences of analytic moments, and sequences of moments evaluated at a specific population parameter are referred to as sequences of numerical moments.

Often in physics and engineering, where the Maxent principle is traditionally used, only sequences of numerical moments are available from experimental data. Hence, $\boldsymbol{\lambda}$ in **Equation 4.16** is typically determined by numerically solving the optimisation problem in **Equation 4.17**. See **Poland (2000)** for an example where the maximum entropy principle is used successfully to reconstruct a molecular energy distribution from experimental data.

In this section, we first show the traditional method of solving the optimisation defined by **Equation 4.17** . We then show some results on the application of the maximum entropy principle for the SLS model and the TLD model under the diffusion approximation, using their sequences of numerical moments. We refer to these solutions as numerical maximum entropy solutions. We will propose a novel method of incorporating a sequence of analytic moments in the next section. When a sequence of analytic moments is considered using the maximum entropy principle, we shall refer to the solution as an analytic maximum entropy solution.

4.3.1 Gaussian quadrature

The integral inside the objective function in **Equations 4.17** and **4.19** is not available in closed form. Being part of an objective function, this integral needs to be evaluated many times with high accuracy for numerical optimisation. Gaussian quadrature has been proposed for evaluating the integral by various authors since the earliest application of the Maxent method. We found that Gaussian-Legendre quadrature provides a satisfactory solution in the univariate case.

The usual Gaussian-Legendre quadrature is over the interval $[-1, 1]$. However, all of the problems we consider involve proportions, and hence a quadrature over the interval $[0, 1]$ is needed. Using the standard technique of interval shifting and scaling, we can obtain the appropriate nodes z_j and weights w_j:

$$\begin{aligned}\int_0^1 f(z)\, dz &= \frac{1}{2}\int_{-1}^1 f\left(\frac{1}{2}z + \frac{1}{2}\right) dz \\ &\approx \frac{1}{2}\sum_j w_j^* f\left(\frac{1}{2}z_j^* + \frac{1}{2}\right) \,, \end{aligned} \tag{4.20}$$

where z_j^* and w_j^* are the usual Gaussian-Legendre nodes and weights respectively; see **Hildebrand (1987)** for a derivation of the Gaussian-Legendre nodes and weights. We thus have the following Gaussian-Legendre nodes and weights for the integral over the interval $[0, 1]$:

$$z_j = \frac{1}{2}\left(z_j^* + 1\right)$$

$$w_j = \frac{1}{2} w_j^* .$$

Let p_j^i denote p to the power of i evaluated at the jth node. Applying the Gaussian-Legendre quadrature to **Equation 4.17** , the optimisation problem reduces to

$$\arg\min_{\boldsymbol{\lambda}} \left\{ \sum_j w_j \exp\left(\sum_{i=0}^{n} \lambda_i p_j^i \right) - \sum_{i=0}^{n} \lambda_i m_i \right\} . \tag{4.21}$$

Similarly, the multivariate case in **Equation 4.19** can be reduced to

$$\arg\min_{\boldsymbol{\lambda}} \left\{ \sum_l w_l \tilde{\pi}_d\left(\boldsymbol{p}_l\right) - \sum_{j=0}^{d} \sum_{i_1=0}^{j} \sum_{i_2=0}^{j-i_1} \sum_{i_3=0}^{j-i_1-i_2} \lambda_{(i_1,i_2,i_3)} \boldsymbol{m}^{\boldsymbol{i}} \right\} , \tag{4.22}$$

where w_l is the weight of a multidimensional Gaussian cubature for the corresponding lth node on the simplex region Δ^3. Determining the nodes and the weights of an efficient Gaussian cubature on the simplex is rather difficult. We converted a procedure originally written by **Greg von Winckel** in Matlab to generate the necessary nodes and weights. It appears to be highly accurate and efficient with all the examples we examined. The corresponding algorithm for the procedure is briefly described in his recent paper **Clason and von Winckel (2011)**.

When the objective function involves a high dimensional integral, however, it is unrealistic to expect an evaluation method to be both efficient and accurate if the dimension keeps increasing. This is the weakness of the maximum entropy approach: it severely suffers from the curse of dimensionality. In terms of our problem, this means that there is little chance of accurately reconstructing any stationary distribution of more than three variables with current computing power. Special treatment of the integral is needed for high dimensional cases. We will not consider this issue in this thesis, and restrict ourselves to three variables at most. Indeed, even with three variables, the integral is barely evaluated to a satisfactory level of accuracy with current computing power.

4.3.2 Chebyshev form

In the above discussion of the maximum entropy solution for the stationary distribution, the univariate density $\tilde{\pi}_n(p)$ is in the form of a power function, with terms involving

$$p^i \quad \text{for} \quad i = 0, 1, \ldots, n\,.$$

The trivariate density $\tilde{\pi}_d(\boldsymbol{p})$ has monomial terms:

$$p_1^{i_1} p_2^{i_2} p_3^{i_3} \quad \text{for} \quad \{i_1, i_2, i_3 \in \mathbb{Z}_{\geq 0} : i_1 + i_2 + i_3 \leq d\}\ .$$

These forms are due to the fact that we started with the corresponding power moments,

$$m_i = \mathbb{E}\left(p^i\right) \quad \text{for} \quad i = 0, 1, \ldots, n\,,$$

and the corresponding monomial moments

$$\boldsymbol{m}^{\boldsymbol{i}} = \mathbb{E}\left(p_1^{i_1} p_2^{i_2} p_3^{i_3}\right) \quad \text{for} \quad \{i_1, i_2, i_3 \in \mathbb{Z}_{\geq 0} : i_1 + i_2 + i_3 \leq d\}\ .$$

These are natural and simple forms for the Maxent distribution: we will refer to them as the original forms of the Maxent distribution. However, this is not the best form to work with numerically. The optimisation in **Equation 4.21** and **Equation 4.22** using power moments leads to an ill-posed numerical problem. This is because we are interested in proportions, $0 \leq p \leq 1$, so the power function p^i becomes extremely small as i becomes large. Therefore the contributions from high order power moments need many significant digits to capture. The level of machine precision puts a limit on how many high-order power moments can be considered, and any higher order power moments past this level are redundant.

By contrast, Chebyshev polynomials are more stable to compute numerically. The maximum entropy approach using Chebyshev moments (the expectation of Chebyshev polynomials) is better conditioned as there is no redundancy. See **Wheeler, Prais and Blumstein (1974)** for a study of power moments and modified moments, where modified moments refer to the expectation of polynomials in general. In particular, **Silver and Röder (1997)** found that the Hessian for the optimisation has a much flatter eigenvalue spectrum when using Chebyshev moments than when power moments are used. See **Bandyopadhyay, Bhattacharya, Biswas and Drabold (2005)** and **Biswas and

Bhattacharya (2010) for studies of Maxent using Chebyshev polynomials to reconstruct various densities.

In terms of our problem, for a univariate model we need to consider the shifted Chebyshev polynomials of the first kind, $T_i^*(p)$, of degree i up to and including n:

$$T_i^*(p) = T_i(2p-1) \quad \text{for} \quad i = 0, 1, \ldots, n\,, \tag{4.23}$$

where T_i is the usual Chebyshev polynomial of the first kind of degree i; see **Appendix D** for details of Chebyshev polynomials.

The reason for considering T_i^* instead of T_i is to adjust the interpolation domain from the interval $(-1, 1)$ to $(0, 1)$. The support of p is $(0, 1)$, so the adjustment is needed to avoid using only half of the interpolation domain and to avoid having to accommodate a discontinuity at $p = 0$. This adjustment allows us to access the full interpolation power of Chebyshev polynomials. Hence, information recovered is at the greatest possible level for a fixed number of moments.

Let m_i^c denote the expectation of $T_i^*(p)$:

$$m_i^c = \mathbb{E}\left\{T_i^*(p)\right\}\ .$$

We will refer to m_i^c as the shifted Chebyshev moment of order i. Let M_n^c denote the sequence of shifted Chebyshev moments of order up to and including n:

$$M_n^c = \left\{m_i^c : i = 0, 1, 2, \ldots, n\right\}\ .$$

Given a sequence of power moments M_n, the corresponding sequence of shifted Chebyshev moments M_n^c can be identified easily by the simple linear transformation,

$$
\begin{pmatrix} m_0^c \\ m_1^c \\ \vdots \\ \vdots \\ m_n^c \end{pmatrix} = \begin{pmatrix} a_{00} & a_{01} & \cdots & \cdots & a_{0n} \\ a_{10} & \cdots & \cdots & \cdots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ a_{n0} & \cdots & \cdots & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ \vdots \\ m_n \end{pmatrix}, \tag{4.24}
$$

where a_{ij} is the coefficient for the term p^j in $T_i^*(p)$. For example, for $i = 0$ we have,

$$
\begin{aligned}
T_0^*(p) &= T_0(2p-1) \\
&= 1 \quad \text{for} \quad 0 \leq p \leq 1\,,
\end{aligned}
$$

thus $a_{00} = 1$ and $a_{0j} = 0$ for $j = 1, 2, \ldots, n$. For $i = 1$,

$$
\begin{aligned}
T_1^*(p) &= T_1(2p-1) \\
&= 2p - 1 \quad \text{for} \quad 0 \leq p \leq 1\,,
\end{aligned}
$$

so $a_{10} = -1$, $a_{11} = 2$ and $a_{1j} = 0$ for $j = 2, 3, \ldots, n$. For $i = 2$,

$$\begin{aligned} T_2^*(p) &= T_2(2p-1) \\ &= 2(2p-1)^2 - 1 \\ &= 8p^2 - 8p + 1 \quad \text{for} \quad 0 \le p \le 1\,, \end{aligned}$$

so $a_{20} = 1$, $a_{21} = -8$, $a_{22} = 8$ and $a_{2j} = 0$ for $j = 3, 4, \ldots, n$.

The remaining elements in the matrix in **Equation 4.24** can be calculated in a similar fashion. Let $\boldsymbol{A}$ be the matrix containing these values. Notice that $\boldsymbol{A}$ is a lower triangular matrix, so the linear transformation can be done quickly by direct forward substitution.

With this new sequence of moments, $\{m_i^c : i = 0, 1, 2, \ldots, n\}$, we have the following modified optimisation problem instead of the one defined by **Equation 4.21** :

$$\underset{\boldsymbol{\lambda}^c}{\arg\min} \left\{ \sum_j w_j \exp\left(\sum_{i=0}^{n} \lambda_i^c T_{ij}^* \right) - \sum_{i=0}^{n} \lambda_i^c m_i^c \right\}\,, \tag{4.25}$$

where $\boldsymbol{\lambda}^c$ is the vector containing values λ_i^c and $T_{ij}^* = T_i^*(p_j)$ denotes the shifted Chebyshev polynomial of the first kind of degree i evaluated at the jth quadrature node. The vector $\boldsymbol{\lambda}^c$ forms a new set of Lagrange multipliers corresponding to the following Maxent distribution $\tilde{\pi}_n^c(p)$:

$$\tilde{\pi}_n^c(p) = \exp\left(\lambda_0^c + \lambda_1^c T_1^*(p) + \lambda_2^c T_2^*(p) + \lambda_3^c T_3^*(p) + \cdots + \lambda_n^c T_n^*(p)\right)\,. \tag{4.26}$$

The densities $\tilde{\pi}_n^c(p)$ and $\tilde{\pi}_n(p)$ in **Equation 4.16** specify the same distribution in two different forms. $\boldsymbol{\lambda}$ is a linear transformation of $\boldsymbol{\lambda}^c$. Given $\boldsymbol{\lambda}^c$, $\boldsymbol{\lambda}$ is given by

$$\boldsymbol{\lambda} = \boldsymbol{A}^{\mathsf{T}} \boldsymbol{\lambda}^{c} \,, \tag{4.27}$$

where $\boldsymbol{A}^{\mathsf{T}}$ is the transpose of the matrix $\boldsymbol{A}$ in **Equation 4.24**.

Given a sequence of power moments up to order n, we first determine the corresponding sequence of Chebyshev moments using **Equation 4.24**. Then we can find $\boldsymbol{\lambda}^{c}$ by solving the optimisation problem in **Equation 4.25**. Lastly, we back-transform $\boldsymbol{\lambda}^{c}$ into $\boldsymbol{\lambda}$ using **Equation 4.27** and obtain the Maxent density function in the original form.

For multivariate cases, a similar approach can be implemented by using an appropriate multi-dimensional Chebyshev polynomial. For the 3-variable TLD model, this is $T^{*}_{(i_1,i_2,i_3)}$. Given the specific form of $T^{*}_{(i_1,i_2,i_3)}$, the corresponding matrix $\boldsymbol{A}$ and thus the corresponding sequence of Chebyshev moments in three dimensions is readily obtained, and hence the optimisation in **Equation 4.22** can be replaced by a more stable optimisation with respect to $\lambda^{c}_{(i_1,i_2,i_3)}$. Solving this leads to $\lambda_{(i_1,i_2,i_3)}$ as well as $\lambda^{c}_{(i_1,i_2,i_3)}$ by a similar back transformation using the corresponding $\boldsymbol{A}^{\mathsf{T}}$. We simply use the product of multiple univariate Chebyshev polynomials of the first kind as the $T^{*}_{(i_1,i_2,i_3)}$:

$$T^{*}_{(i_1,i_2,i_3)}\left(\boldsymbol{p}\right) = T_{i_1}\left(2p_1 - 1\right) T_{i_2}\left(2p_2 - 1\right) T_{i_3}\left(2p_3 - 1\right) \,. \tag{4.28}$$

This $T^{*}_{(i_1,i_2,i_3)}$ is known as the tensor product of univariate Chebyshev polynomials of the first kind; see **Barthelmann, Novak and Ritter (2000)** for more details.

4.3.3 Trust region optimisation

Solving the optimisation problem in **Equation 4.25** is the key step to gaining a Maxent solution $\tilde{\pi}_n(p)$, and deserves a brief discussion. Although the chance of finding the optimal $\boldsymbol{\lambda}$ is greatly improved by putting the problem in the Chebyshev form and solving $\boldsymbol{\lambda}^c$ instead, different optimisation algorithms perform quite differently. We will use the univariate case to make a few points, but they apply to multivariate cases as well.

Let f denote the objective function for the optimisation problem,

$$f = \sum_j w_j \exp\left(\sum_{i=0}^{n} \lambda_i^c T_{ij}^*\right) - \sum_{i=0}^{n} \lambda_i^c \, m_i^c \,. \tag{4.29}$$

The gradient $\boldsymbol{g}$ is a vector with $n+1$ components of the following form:

$$\boldsymbol{g} = \begin{pmatrix} \sum_j w_j T_{0j}^* \tilde{\pi}_n^c(p_j) - m_0^c \\ \sum_j w_j T_{1j}^* \tilde{\pi}_n^c(p_j) - m_1^c \\ \vdots \\ \vdots \\ \sum_j w_j T_{nj}^* \tilde{\pi}_n^c(p_j) - m_n^c \end{pmatrix}, \tag{4.30}$$

where $\tilde{\pi}_n^c(p_j) = \exp\left(\sum_{i=0}^{n} \lambda_i^c T_{ij}^*\right)$.

Notice that setting the gradient to zero implies

$$
\begin{aligned}
m_i^c &= \sum_j w_j T_{ij}^* \tilde{\pi}_n^c\,(p_j) \quad \text{for} \quad i = 0, 1, \ldots, n \\
&\approx \int_0^1 T_i^*(p) \tilde{\pi}_n^c(p)\, dp\,.
\end{aligned} \tag{4.31}
$$

The quality of the approximation in **Equation 4.31** depends on the accuracy of the quadrature. It also determines how well the moment constraints are satisfied. Therefore, the accuracy of the quadrature affects the quality of any optimisation procedure.

The Hessian matrix $\boldsymbol{H}$ is a symmetric $(n+1) \times (n+1)$ matrix of the following form:

$$
\boldsymbol{H} = \begin{pmatrix}
\sum_j w_j T_{0j}^* T_{0j}^* \tilde{\pi}_n^c\,(p_j) & \cdots & \cdots & \cdots & \sum_j w_j T_{0j}^* T_{nj}^* \tilde{\pi}_n^c\,(p_j) \\
\sum_j w_j T_{0j}^* T_{1j}^* \tilde{\pi}_n^c\,(p_j) & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\sum_j w_j T_{0j}^* T_{nj}^* \tilde{\pi}_n^c\,(p_j) & \cdots & \cdots & \cdots & \sum_j w_j T_{nj}^* T_{nj}^* \tilde{\pi}_n^c\,(p_j)
\end{pmatrix}. \tag{4.32}
$$

Again, the quality of the Hessian depends on the accuracy of the quadrature.

In general, gradient based optimisation algorithms that incorporate both the gradient and the Hessian tend to perform better than other algorithms. However, in our case, the Hessian matrix is very expensive to compute. The Hessian matrix has many elements, each of which is a sum of many terms. Alternative algorithms that allow for simple computation but take a large number of iterations to converge are not suitable here.

We used a trust region algorithm, as discussed by **Fletcher (1987)** and **Nocedal and Wright (1999)**, and implemented in R by **Geyer (2008)**. Trust region optimisation is a gradient based algorithm considered to be slower than line search algorithms in general as the trust region subproblem within each iteration is harder to solve than the line search subproblem. Usually a line search algorithm takes many more iterations to solve a given problem, but each iteration takes much less time. However, using the trust region algorithm for our problem, we found that the time saved from computing extra iterations outweighs the time lost for each iteration. The trust region algorithm outperformed all other optimisation algorithms we tested, and is the only algorithm that provides both the stability and the speed needed for all the Maxent problems we considered.

The main idea of trust region optimisation is that the gradient and the Hessian are used to build up a quadratic approximation for a region of the objective function around the initial value. If the initial value is in a region where the objective function is adequately approximated, then a direction is chosen according to the approximation to move the initial value to an improved value within the region and the region is expanded so that a big step might be taken later. Otherwise the region is contracted and the optimisation stays at the initial value. Because of the extra computation needed when deciding whether to move or to stay, trust region optimisation is considered to be conservative in its decision to stay. See **Fletcher (1987)** and **Nocedal and Wright (1999)** for a more mathematical discussion of the trust region algorithm.

4.3.4 Results for the SLS model

We consider two models, the SLS model and the TLD model, for which to apply the numerical Maxent approach. The example of the SLS model is mainly to demonstrate the performance of Maxent, because the underlying stationary distribution for the diffusion process is available; see **Equation 3.39** . The TLD model is considered because we are interested in $\mathbb{E}\left(r^2\right)$ and $\mathbb{V}\left(r^2\right)$ at steady state for various values of θ and ρ.

We first show the results for the SLS model. In the presence of selection, the analytic moments for the SLS model are not available by solving the recursion in **Equation 4.1** . However, the underlying stationary distribution $\pi\,(p)$ for the SLS model under the diffusion approximation is known up to a normalisation constant, see **Equation 3.39** . Given the values of its scaled selection rates (α_1 and α_2) and scaled mutation rates (θ_1 and θ_2), $\pi\,(p)$ for SLS can be normalised and its power moments can be obtained by numerical integration of the corresponding power function p^i and the density function in **Equation 3.39** .

For the set of population parameters

$$\alpha_1 = 1\,, \qquad \alpha_2 = 2\,, \qquad \theta_1 = 1\,, \qquad \theta_2 = 1\,,$$

the normalisation constant Λ is 0.5175441583467, and hence $\pi\,(p)$ is

$$\pi\,(p) = \frac{1}{0.5175441583467}\, p(1-p)\exp\left(4p - 3p^2\right)\,, \tag{4.33}$$

and the corresponding sequence of shifted Chebyshev moments of order up to and including 8 is

$$M_8^c = \{1, 0.084090111, -0.652777967, -0.120800686, 0.159699910,$$
$$0.038372820, -0.010044775, -0.002712854, 0.001985952\}\,.$$

The distribution in **Equation 4.33** is taken to be the true $\pi\,(p)$, and we use its sequence of numerical moments to reconstruct it using Maxent. The Maxent distribution,

$\tilde{\pi}_n(p)$ for $n = 2, 4, 6, 8$, is plotted against the true distribution $\pi(p)$ in **Figure 4.1** . In this case, $\tilde{\pi}_n(p)$ converges quickly to $\pi(p)$; with the first eight moments, $\tilde{\pi}_8(p)$ differs only slightly from $\pi(p)$.

In **Figure 4.2** , $\log\{\tilde{\pi}_n(p)\}$ is plotted against $\log\{\pi(p)\}$ to magnify the differences, from which it can be seen that the Maxent method performs worst near the boundaries of the interval. Modern computing power allows us to include more moments beyond the first eight in the optimisation. Using the first 50 moments, $\tilde{\pi}_{50}(p)$ does not differ markedly from $\pi(p)$ even on the logarithmic scale; see **Figure 4.3** .

The convergence rate in terms of the number of moments needed is largely dependent on the shape of the underlying $\pi(p)$. A fairly symmetric and smooth $\pi(p)$, such as that in the example above, requires fewer moments to reconstruct to a given level of accuracy than a highly asymmetric or spiked $\pi(p)$.

For the SLS model with the same scaled selection rates ($\alpha_1 = 1, \alpha_2 = 2$), but different mutation rates ($\theta_1 = 0.4$ and $\theta_2 = 0.6$), $\pi(p)$ has an irregular shape (**Figure 4.4**), and requires a greater number of moments to reconstruct. In practice, an underlying distribution that is sharp at the boundaries is generally the hardest shape to reconstruct, besides a highly oscillatory $\pi(p)$. When $\pi(p)$ is difficult to reconstruct, a higher order quadrature (more nodes) as well as a higher order sequence of moments is needed to provide the extra accuracy to capture the sharpness at the boundaries or the fluctuations in $\pi(p)$. This is because Gaussian quadrature of order n gives an exact result for polynomials of degree up to $2n - 1$. We are of course not considering an integral of polynomials, but the general idea still applies: a quadrature of a higher order is needed as the integrand becomes more complex, for the accuracy of the quadrature to remain at the same level.

Figure 4.1: Maxent densities $\tilde{\pi}_2(p)$, $\tilde{\pi}_4(p)$, $\tilde{\pi}_6(p)$ and $\tilde{\pi}_8(p)$ for the SLS model with equal mutation shown with the known true density $\pi(p)$.

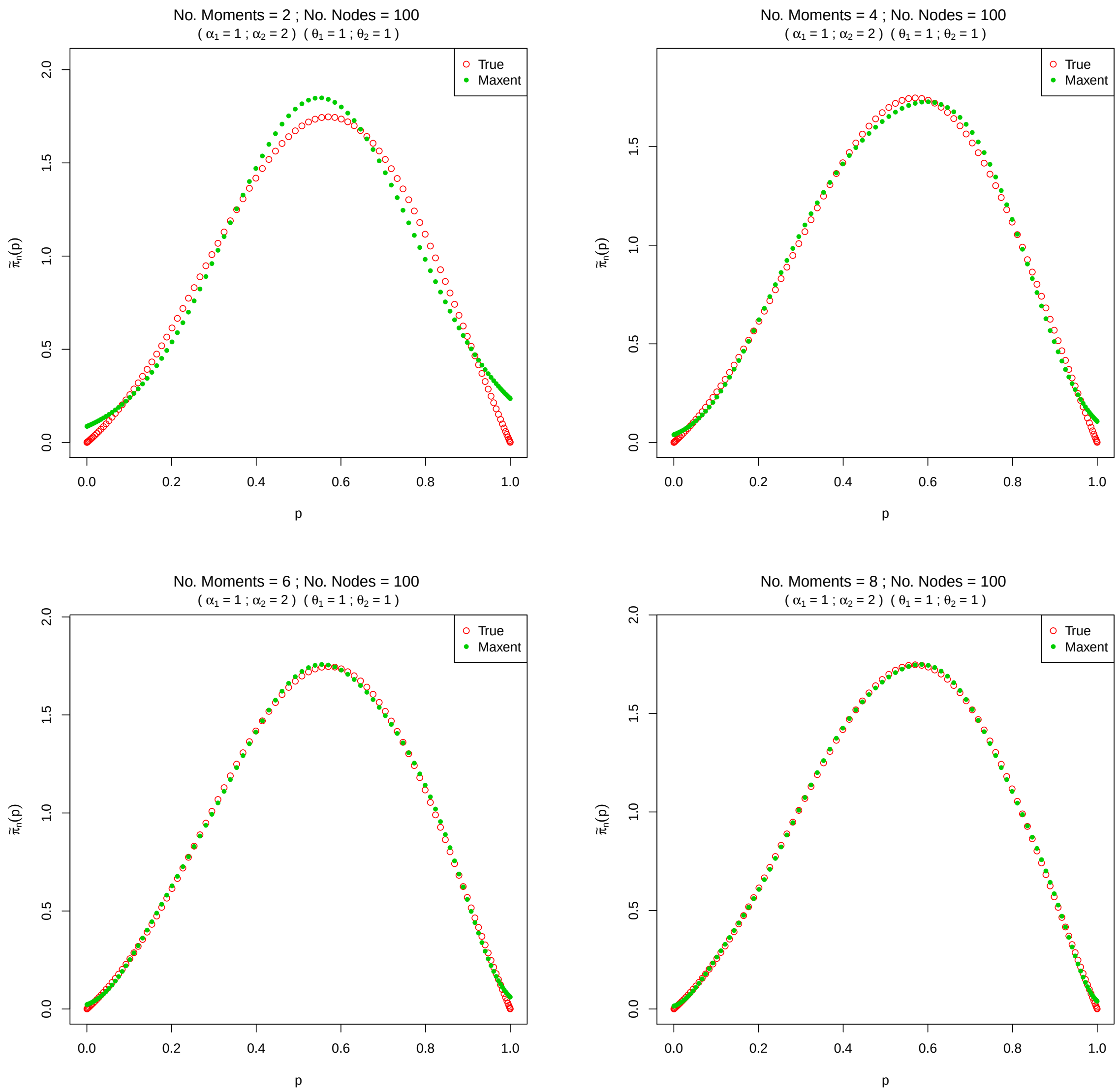


Figure 4.2: The Maxent density using moments up to order 8, plotted with the known true density. $\tilde{\pi}_8(p)$ is shown on the left panel, and $\log\{\tilde{\pi}_8(p)\}$ is shown on the right panel.

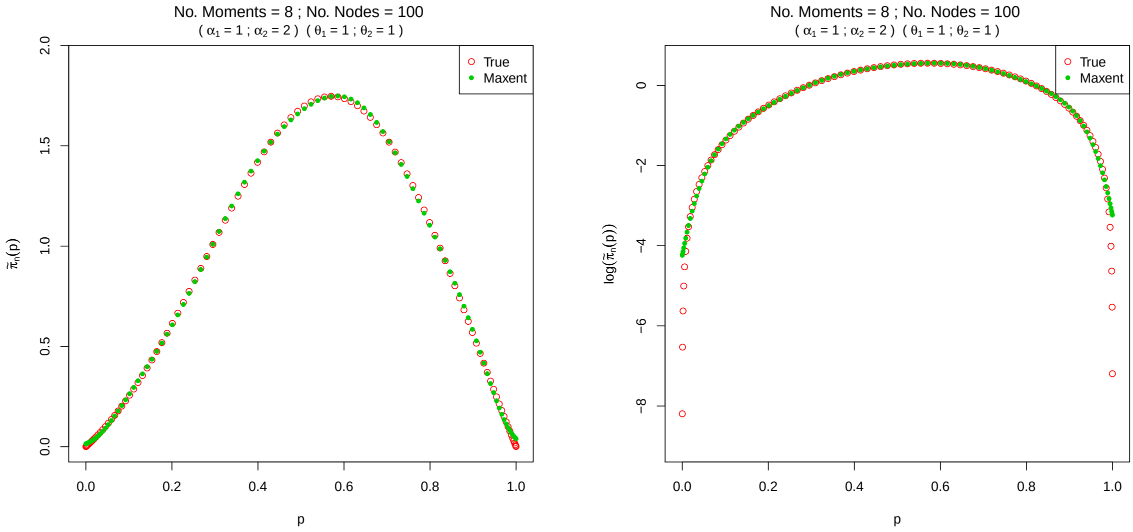


Figure 4.3: The Maxent density using moments up to order 50, plotted with the known true density. $\tilde{\pi}_{50}(p)$ is shown on the left panel, and $\log\{\tilde{\pi}_{50}(p)\}$ is shown on the right panel.

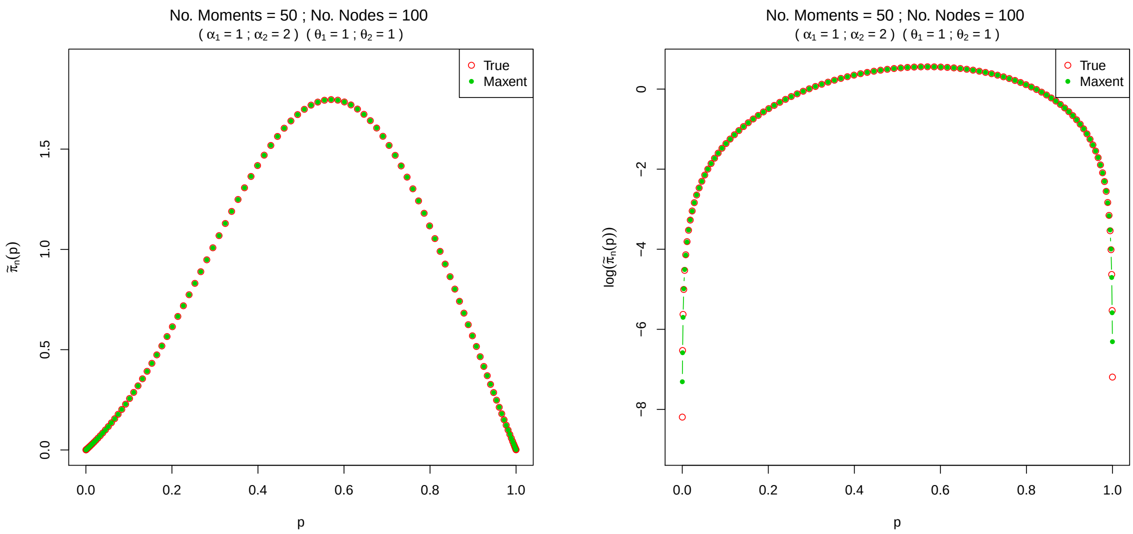
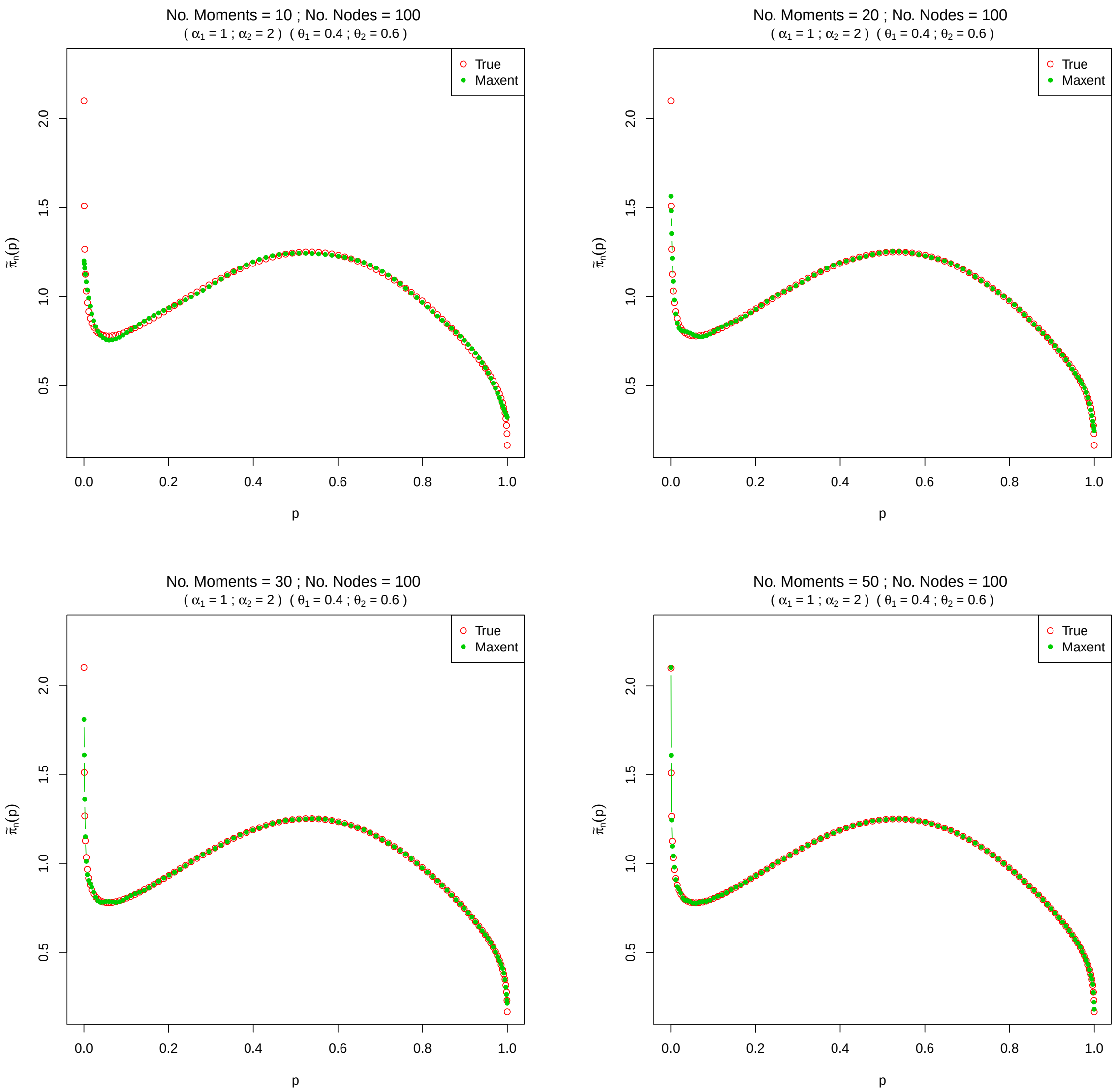


Figure 4.4: Maxent densities $\tilde{\pi}_{10}(p)$, $\tilde{\pi}_{20}(p)$, $\tilde{\pi}_{30}(p)$ and $\tilde{\pi}_{50}(p)$ for the SLS model with unequal mutation rates, shown with the known true density $\pi(p)$.



4.3.5 Expectation and variance of r^2 in the TLD model

The numerical stationary moments for the TLD model are obtained by evaluating the analytic stationary moments at various θ and ρ. See **Table 4.1** for the procedure of recovering the analytic stationary moments for the TLD model under the diffusion approximation.

After transforming to the shifted Chebyshev moments, we apply the numerical Maxent approach for the TLD model, and we determine its Maxent $\tilde{\pi}_d\left(\boldsymbol{p}\right)$ at steady state for each pair of θ and ρ using a sequence of moments of order up to and including 8. With the Maxent densities, $\tilde{\pi}_d\left(\boldsymbol{p}\right)$, we compute the following integral numerically and obtain a table of $\mathbb{E}\left(r^2\right)$, the expected linkage disequilibrium measure r^2,

$$\begin{aligned}\mathbb{E}\left(r^2\right) &= \int_{\Delta^3} r^2(\boldsymbol{p})\pi(\boldsymbol{p})\,d\boldsymbol{p} \\ &\approx \int_{\Delta^3} r^2(\boldsymbol{p})\tilde{\pi}_d(\boldsymbol{p})\,d\boldsymbol{p}\,, \end{aligned} \tag{4.34}$$

where r^2 is a function of p, q and D, which are, in turn, functions of p_1, p_2 and p_3:

$$r^2 = \frac{D^2}{p\left(1-p\right)q\left(1-q\right)}\,, \tag{4.35}$$

where p, q and D are

$$\begin{aligned} p &= p_1 + p_2 \\ q &= p_1 + p_3 \\ D &= p_1 - p_1^2 - p_1p_2 - p_1p_3 - p_2p_3\,. \end{aligned}$$

Similarly, we compute the following integral numerically and obtain a table of $\mathbb{V}\left(r^2\right)$, the variance of the linkage disequilibrium measure r^2,

$$\begin{aligned}\mathbb{V}\left(r^2\right) &= \int_{\Delta^3} \left\{r^2(\boldsymbol{p}) - \mathbb{E}\left(r^2\right)\right\}^2 \pi(\boldsymbol{p})\, d\boldsymbol{p} \\ &\approx \int_{\Delta^3} \left\{r^2(\boldsymbol{p}) - \mathbb{E}\left(r^2\right)\right\}^2 \tilde{\pi}_d(\boldsymbol{p})\, d\boldsymbol{p}\,, \end{aligned} \tag{4.36}$$

where $\mathbb{E}\left(r^2\right)$ is obtained from the computation in **Equation 4.34** .

We compare our evaluation of the expectation of r^2 against the evaluation by **Song and Song (2007)**. There, the authors produced a table of the expectation of the linkage disequilibrium measure r^2 for the TLD model at steady state across a range of values for scaled mutation θ and scaled recombination ρ.

It can be seen in **Table 4.2** that the Maxent computation of $\mathbb{E}\left(r^2\right)$ and that in **Song and Song (2007)** are generally identical when θ is large, but differences become obvious as θ decreases. These differences are due to the fact that the shape of $\pi\left(\boldsymbol{p}\right)$ becomes extremely sharp close to the boundary of the simplex region as θ decreases. In terms of population genetics, fixation becomes more imminent as θ decreases, the underlying stationary distribution becomes closer to degeneracy, and our computation becomes less accurate as this happens. This is due to the fact that the method of Maxent is not equipped to reconstruct degenerate distributions.

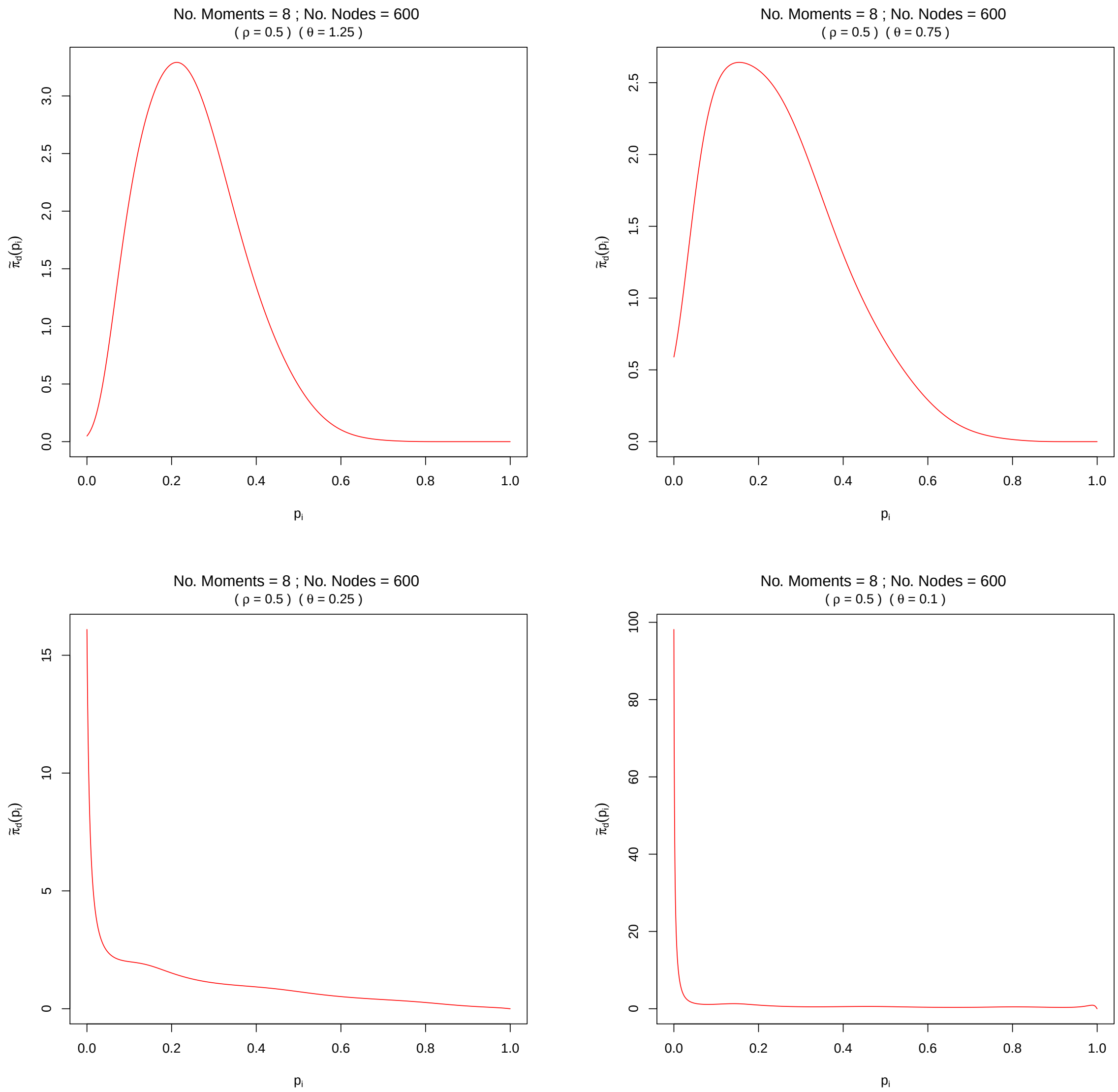
We show the Maxent marginal density of a single gamete type p_i as θ decreases in **Figure 4.5** . The Maxent solution provides the joint density $\tilde{\pi}_d\left(\boldsymbol{p}\right)$, and the marginal density of p_1 is determined by numerically integrating out p_2 and p_3. The TLD model is symmetric in the sense that p_1, p_2, and p_3 have the same marginal distribution, so only one marginal distribution is plotted in **Figure 4.5** . The plots show the spiked nature of

Table 4.2: Comparison of r^2 between the Maxent method, Song & Song's method and coalescent simulations performed by **Song and Song (2007)**.

ρ	θ										
	0.0125	0.0250	0.0500	0.0750	0.1000	0.1250	0.2500	0.5000	0.7500	1.0000	1.2500
(a) *Maxent method using moments up to 8th order for $\tilde{\pi}_d(\boldsymbol{p})$ and $\mathbb{E}[r^2]$ derived numerically with 216 million nodes*											
0.00	0.016	0.036	0.065	0.083	0.094	0.101	0.105	0.081	0.063	0.051	0.042
0.25	0.011	0.025	0.048	0.063	0.076	0.083	0.092	0.075	0.059	0.048	0.041
0.50	0.008	0.019	0.040	0.053	0.063	0.071	0.082	0.069	0.056	0.046	0.039
1.25	0.005	0.012	0.026	0.036	0.044	0.050	0.062	0.056	0.047	0.040	0.035
2.50	0.004	0.008	0.018	0.025	0.030	0.035	0.045	0.043	0.038	0.033	0.030
5.00	0.002	0.006	0.012	0.019	0.020	0.023	0.029	0.030	0.027	0.025	0.023
(b) **Song and Song (2007)** *computation of $\mathbb{E}[r^2]$*											
0.00	0.008	0.024	0.056	0.079	0.094	0.103	0.106	0.081	0.063	0.051	0.042
0.25	0.006	0.018	0.043	0.062	0.076	0.085	0.093	0.075	0.059	0.048	0.041
0.50	0.005	0.014	0.035	0.052	0.064	0.072	0.083	0.069	0.056	0.046	0.039
1.25	0.003	0.009	0.024	0.036	0.045	0.052	0.063	0.056	0.047	0.040	0.035
2.50	0.002	0.006	0.016	0.025	0.031	0.036	0.045	0.043	0.038	0.033	0.030
5.00	0.001	0.004	0.011	0.016	0.020	0.023	0.030	0.030	0.027	0.025	0.023
(c) *Average r^2 from* **Song and Song (2007)** *coalescent simulations, with no restriction on segregation*											
0.00	0.013	0.033	0.069	0.095	0.102	0.111	0.105	0.077	0.057	0.050	0.040
0.25	0.009	0.024	0.056	0.075	0.088	0.095	0.091	0.072	0.056	0.045	0.039
0.50	0.007	0.019	0.046	0.063	0.075	0.080	0.085	0.067	0.053	0.044	0.038
1.25	0.005	0.014	0.032	0.044	0.057	0.059	0.067	0.056	0.047	0.040	0.035
2.50	0.003	0.009	0.023	0.032	0.039	0.043	0.050	0.045	0.039	0.034	0.031
5.00	0.002	0.006	0.015	0.022	0.026	0.029	0.034	0.033	0.030	0.027	0.026
(d) *Average r^2 from* **Song and Song (2007)** *coalescent simulations, conditioned on segregation at both sites*											
0.00	0.131	0.128	0.126	0.125	0.121	0.119	0.106	0.077	0.057	0.050	0.040
0.25	0.093	0.093	0.097	0.099	0.102	0.103	0.095	0.072	0.056	0.045	0.039
0.50	0.076	0.078	0.081	0.082	0.088	0.091	0.084	0.067	0.053	0.044	0.038
1.25	0.051	0.052	0.057	0.059	0.062	0.066	0.067	0.056	0.047	0.040	0.035
2.50	0.037	0.038	0.041	0.042	0.046	0.048	0.051	0.045	0.039	0.034	0.031
5.00	0.024	0.026	0.028	0.029	0.031	0.032	0.035	0.033	0.030	0.027	0.026

$\pi(\boldsymbol{p})$ close to $p_i = 0$ as θ becomes small. A subsidiary peak near $p_i = 1$ can also be seen in the last panel.

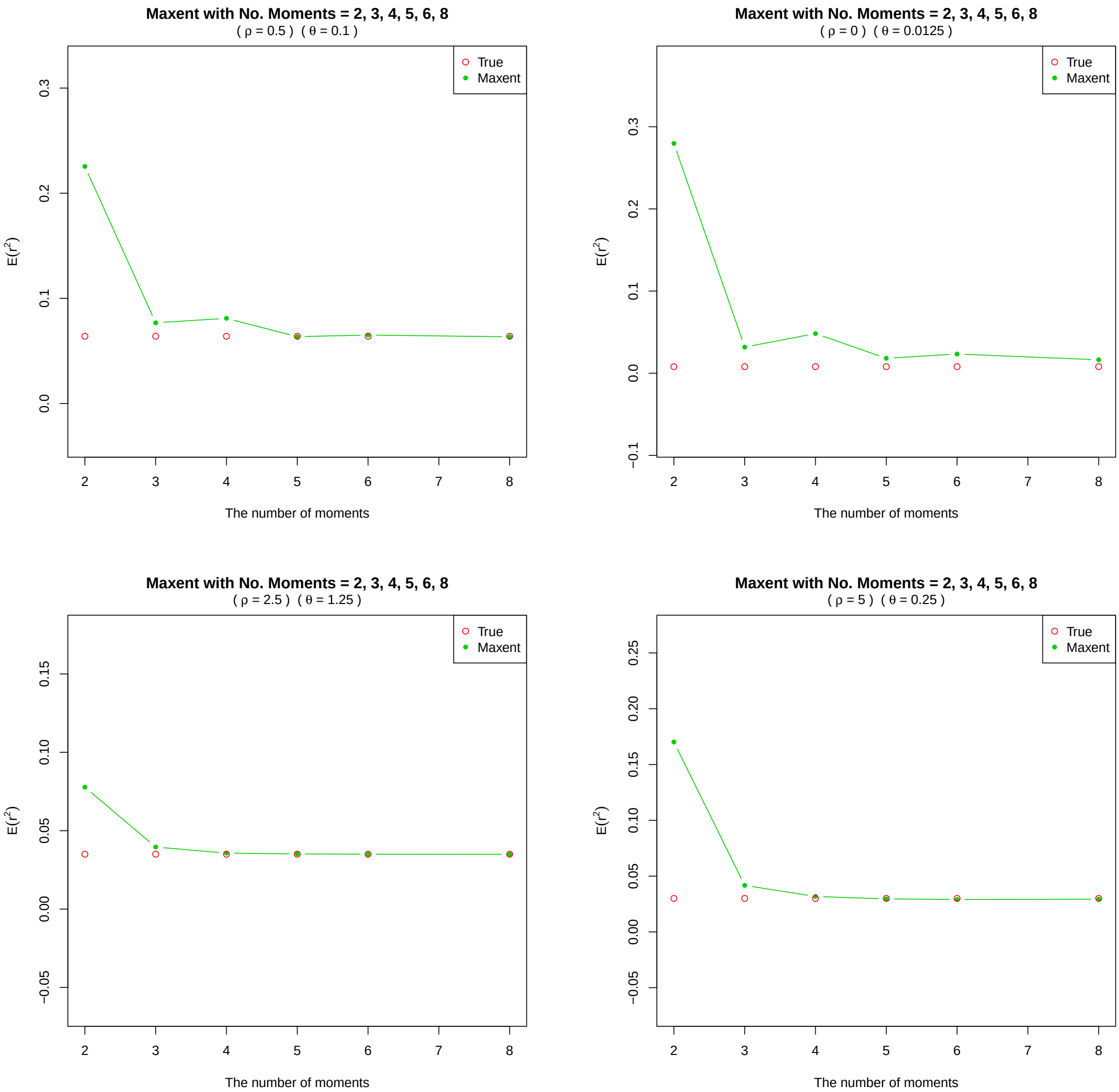
Figure 4.5: Maxent marginal density of a single gametic frequency p_i for the TLD model as θ decreases. 600 quadrature nodes in one dimension corresponds to a total of 216 million nodes in three dimensions.



Taking $\mathbb{E}(r^2)$ computed by **Song and Song (2007)** to be the true values, we show plots of our computation of $\mathbb{E}(r^2)$ as we use an increasing number of moments in **Figure 4.6**.

It can be seen that our computation becomes closer to the true value as the number of moments increases. This suggests that accuracy in our method is primarily limited by computer power, which restricts the number of moments it is feasible to use. With unlimited computer power the trends suggest that the Maxent method could be very accurate.

Figure 4.6: Maxent computation of $\mathbb{E}\left(r^2\right)$ for various θ and ρ as the number of moments increases.



We can compute $\mathbb{V}(r^2)$ using **Equation 4.36**. **Table 4.3** gives $\mathbb{V}(r^2)$ for a range of θ and ρ values computed using $\tilde{\pi}_d(\boldsymbol{p})$ of order $d = 8$.

Our computation of $\mathbb{V}(r^2)$ shows that it is a strictly decreasing function of ρ, but it is more complicated as a function of θ. It initially increases as θ increases from zero, but starts to decrease after reaching a certain maximum point. This maximum point of θ after which $\mathbb{V}(r^2)$ starts to decrease depends on the ρ value of the underlying system.

In the same way that large numbers of moments are needed to reconstruct densities with sharp edges in the univariate case, a sequence of moments of high order is required to achieve the accuracy needed to match the computation in **Song and Song (2007)** for small θ. Using a sequence of moments up to order 8 is apparently not enough for very small θ. However, demands on computer power increase rapidly for multivariate cases as the order increases.

Table 4.3: Variance of r^2 computed by finding $\tilde{\pi}_8(\boldsymbol{p})$ and using numerical integration.

ρ	θ										
	0.0125	0.0250	0.0500	0.0750	0.1000	0.1250	0.2500	0.5000	0.7500	1.0000	1.2500
0.00	0.005934	0.015332	0.02908	0.03516	0.03711	0.03569	0.02436	0.01175	0.00687	0.00454	0.003212
0.25	0.003226	0.008784	0.01805	0.02248	0.02500	0.02526	0.01902	0.00999	0.00612	0.00413	0.002963
0.50	0.002119	0.005896	0.01350	0.01653	0.01837	0.01897	0.01531	0.00862	0.00544	0.00377	0.002743
1.25	0.000972	0.002723	0.00609	0.00812	0.00943	0.00993	0.00905	0.00587	0.00404	0.00293	0.002213
2.50	0.000487	0.001292	0.00295	0.00388	0.00459	0.00476	0.00482	0.00356	0.00266	0.00206	0.001619
5.00	0.000241	0.000626	0.00136	0.00272	0.00198	0.00209	0.00211	0.00174	0.00142	0.00117	0.000976

4.4 Analytic Maximum Entropy Solution

4.4.1 Derivation

For models such as the SLS model in the absence of selection, the SLM model and the TLD model, where the sequence of stationary moments in analytic form is available in terms of population parameters, we propose a method of obtaining the analytic forms of $\tilde{\pi}_n(p)$ and $\tilde{\pi}_d(\boldsymbol{p})$ in terms of the population parameters and the random variable by extending the numerical Maxent approach. We shall use the diallelic SLM to outline this method.

We showed in **Section 4.2** that the univariate $\tilde{\pi}_n(p)$ has the following general form,

$$\tilde{\pi}_n(p) = \exp\left(\lambda_0 + \lambda_1 p + \lambda_2 p^2 + \lambda_3 p^3 + \cdots + \lambda_n p^n\right) , \tag{4.37}$$

where the λ_is are Lagrange multipliers and are obtained by solving the following unconstrained minimisation problem:

$$\underset{\boldsymbol{\lambda}}{\arg\min} \left\{ \int_0^1 \exp\left(\sum_{i=0}^{n} \lambda_i p^i\right) dp - \sum_{i=0}^{n} \lambda_i m_i \right\} . \tag{4.38}$$

Equations 4.37 and **4.38** determine the Maxent stationary distribution of the diallelic SLM model given its numerical sequence of moments at stationarity. In **Section 4.1**, the moments of the diallelic SLM model were shown to have the following form:

$$\mathbb{E}\left(p^{n}\right)=\frac{\Gamma\left(2\theta+n\right)\Gamma\left(4\theta\right)}{\Gamma\left(4\theta+n\right)\Gamma\left(2\theta\right)}\,. \tag{4.39}$$

Each of the individual moments depends on θ, so the sequence of analytic moments, M_n, depends on θ. This sequence solely determines the values of the λ_is through the optimisation. Thus the λ_is can be considered as functions of θ for the SLM model. We shall write $\lambda_i\left(\theta\right)$ rather than just numerical λ_i when the sequence of analytic moments is available, where $\lambda_i\left(\theta\right)$ denotes the unknown function in terms of θ for $i=0,1,2,\ldots,n$. The form of the function $\lambda_i\left(\theta\right)$ is likely to be different for different i, so $n+1$ unknown functions need to be determined instead of $n+1$ unknown coefficients. We have the following for $\tilde{\pi}_n\left(p;\theta\right)$:

$$\tilde{\pi}_n\left(p;\theta\right)=\exp\left(\sum_{i=0}^{n}\lambda_i\left(\theta\right)p^i\right)\,. \tag{4.40}$$

For the diallelic model, this leads to moment constraints of the following forms,

$$m_j\left(\theta\right)=\int_0^1 p^j\exp\left(\sum_{i=0}^{n}\lambda_i\left(\theta\right)p^i\right)dp \quad \text{for} \quad j=0,\,1\ldots,\,n\,, \tag{4.41}$$

where $m_j\left(\theta\right)=\dfrac{\Gamma\left(2\theta+j\right)\Gamma\left(4\theta\right)}{\Gamma\left(4\theta+j\right)\Gamma\left(2\theta\right)}$.

Differentiating both sides of **Equation 4.41** with respect to θ, we have

$$m_j'(\theta) = \frac{\partial}{\partial \theta}\left[\int_0^1 p^j \exp\left\{\sum_{i=0}^{n} \lambda_i(\theta)\, p^i\right\} dp\right], \tag{4.42}$$

where $m_j'(\theta)$ denotes the derivative of $m_j(\theta)$ with respect to θ, and $j = 0, 1, 2 \ldots, n$.

We assume that $\left\{p^j \exp\left(\sum_{i=0}^{n} \lambda_i(\theta)\, p^i\right)\right\}$ is a well behaved function of p and θ, so we can differentiate with respect to θ under the integral sign:

$$\begin{aligned} m_j'(\theta) &= \int_0^1 \frac{\partial}{\partial \theta}\left\{p^j \exp\left(\sum_{k=0}^{n} \lambda_k(\theta)\, p^k\right)\right\} dp \\ &= \int_0^1 \sum_{i=0}^{n} \lambda_i'(\theta)\, p^{i+j} \exp\left(\sum_{k=0}^{n} \lambda_k(\theta)\, p^k\right) dp, \end{aligned} \tag{4.43}$$

where $\lambda_i'(\theta)$ denotes the derivative of $\lambda_i(\theta)$ with respect to θ. Notice that $\lambda_i(\theta)$ is not a function of p, thus

$$m_j'(\theta) = \sum_{i=0}^{n} \lambda_i'(\theta) \int_0^1 p^{i+j} \exp\left(\sum_{k=0}^{n} \lambda_k(\theta)\, p^k\right) dp\,. \tag{4.44}$$

Notice that $\int_0^1 p^{i+j} \exp\left(\sum_{k=0}^{n} \lambda_k(\theta)\, p^k\right) dp$ is actually the $(i+j)$th moment $m_{i+j}(\theta)$, therefore **Equation 4.44** reduces to

$$m_j'(\theta) = \sum_{i=0}^{n} \lambda_i'(\theta)\, m_{i+j}(\theta) \quad \text{for} \quad j = 0, 1 \ldots, n\,. \tag{4.45}$$

Therefore we have the following system of linear equations in the $\lambda_i'(\theta)$s,

$$\begin{pmatrix} m_0(\theta) & m_1(\theta) & m_2(\theta) & m_3(\theta) & \cdots & m_n(\theta) \\ m_1(\theta) & m_2(\theta) & m_3(\theta) & \cdots & \cdots & m_{n+1}(\theta) \\ m_2(\theta) & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ m_n(\theta) & \cdots & \cdots & \cdots & \cdots & m_{2n}(\theta) \end{pmatrix} \begin{pmatrix} \lambda'_0(\theta) \\ \lambda'_1(\theta) \\ \vdots \\ \vdots \\ \lambda'_n(\theta) \end{pmatrix} = \begin{pmatrix} m'_0(\theta) \\ m'_1(\theta) \\ \vdots \\ \vdots \\ m'_n(\theta) \end{pmatrix}. \tag{4.46}$$

We will refer to this linear system of equations as the first order derivative condition. The matrix in **Equation 4.46** is a Hankel matrix, which is a square matrix with constant skew diagonals:

$$\begin{pmatrix} a & b & c & d & e \\ b & c & d & e & f \\ c & d & e & f & g \\ d & e & f & g & h \\ e & f & g & h & i \end{pmatrix}. \tag{4.47}$$

A linear system given by a $n \times n$ Hankel matrix only has $2n+1$ degrees of freedom, as opposed to n^2 in general, and can be solved in $\mathcal{O}(n^2)$ time, as opposed to $\mathcal{O}(n^3)$ for the general case. See **Freund and Zha (1993)** for a discussion of Hankel matrices and an algorithm for solving the associated linear systems.

The analytic form of $m'_j(\theta)$ can be determined through direct differentiation of **Equation 4.4**,

$$m_j'(\theta) = \frac{2\Gamma(2\theta+j)\Gamma(4\theta)}{\Gamma(4\theta+j)\Gamma(2\theta)}\left\{\Psi(2\theta+j)+2\Psi(4\theta)-\Big(2\Psi(4\theta+j)+\Psi(2\theta)\Big)\right\}, \quad (4.48)$$

where Ψ denotes the digamma function (**Abramowitz and Stegun 1964, Section 6.3.6**).

Therefore the function $\lambda_i'(\theta)$ can be obtained in the form of a rational function of θ by solving the first order derivative condition in **Equation 4.46** . For example, with the first two moments ($n = 2$), using Maple, we obtain the $\lambda_i'(\theta)$s in the following form:

$$\begin{pmatrix} \lambda_0'(\theta) \\ \lambda_1'(\theta) \\ \lambda_2'(\theta) \end{pmatrix} = \begin{pmatrix} -\dfrac{2(4\theta+3)}{4\theta+1} \\ \dfrac{2(4\theta+3)}{\theta} \\ -\dfrac{2(4\theta+3)}{\theta} \end{pmatrix}. \quad (4.49)$$

By integrating each $\lambda_i'(\theta)$ with respect to θ, we have $\lambda_i(\theta)$ up to a constant of integration, which depends neither on p nor θ. Let us define $\tilde{\lambda}_i(\theta)$ in the following way:

$$\begin{aligned} \lambda_i(\theta) &= \int \lambda_i'(\theta)\, d\theta \\ &= \tilde{\lambda}_i(\theta) + c_i\,, \end{aligned} \quad (4.50)$$

where c_i is the constant of integration.

For the $\lambda_i'(\theta)$s in **Equation 4.49** , the corresponding $\tilde{\lambda}_i(\theta)$s have the following form:

$$\begin{pmatrix} \tilde{\lambda}_0(\theta) \\ \tilde{\lambda}_1(\theta) \\ \tilde{\lambda}_2(\theta) \end{pmatrix} = \begin{pmatrix} -2\theta - \ln(4\theta + 1) \\ 8\theta + 6\ln(\theta) \\ -8\theta - 6\ln(\theta) \end{pmatrix} \tag{4.51}$$

In general, we have the following $\tilde{\pi}_n(p;\theta)$ in terms of $\tilde{\lambda}_i(\theta)$,

$$\tilde{\pi}_n(p;\theta) = \exp\left(\sum_{i=0}^{n} \left(\tilde{\lambda}_i(\theta) + c_i\right) p^i\right), \tag{4.52}$$

where the c_is are the only unknown quantities that need to be determined.

We compute the constants c_i by using the numerical Maxent approach in the last section for a sensible value of θ. We shall discuss the choice of θ shortly. For the present denote this value of θ by θ^*.

For a chosen θ^*, the constant of integration is given by

$$c_i = \lambda_i(\theta^*) - \tilde{\lambda}_i(\theta^*), \tag{4.53}$$

where $\lambda_i(\theta^*)$ is obtained by using the numerical Maxent approach described in **Section 4.3**, and $\tilde{\lambda}_i$ is the analytic function obtained by performing the indefinite integration in **Equation 4.50**.

This completes the analytic form for $\tilde{\pi}_n(p;\theta)$ in **Equation 4.52**, which is now an analytic density for the random variable p in terms of the parameter θ. We refer to this analytic form as the Maxent distribution centred at θ^*, and refer to the specific choice of parameter θ^* as the centre of the approximation or simply the centre.

4.4.2 Example using the first two moments

As an example, when the number of moments is 2 ($n = 2$) and the centre is chosen to be 1 ($\theta^* = 1$), the analytic Maxent distribution $\tilde{\pi}_2(p;\theta)$ takes the following form:

$$\tilde{\pi}_2(p;\theta) = \exp\left[c_0 - 2\theta - \ln(4\theta+1) + \{c_1 + 8\theta + 6\ln(\theta)\}\, p + \{c_2 - 8\theta - 6\ln(\theta)\}\, p^2\right] , \tag{4.54}$$

where we obtain by numerical computation:

$$c_0 = 2.2244121 \tag{4.55}$$

$$c_1 = -0.5031736 \tag{4.56}$$

$$c_2 = 0.5031736\,. \tag{4.57}$$

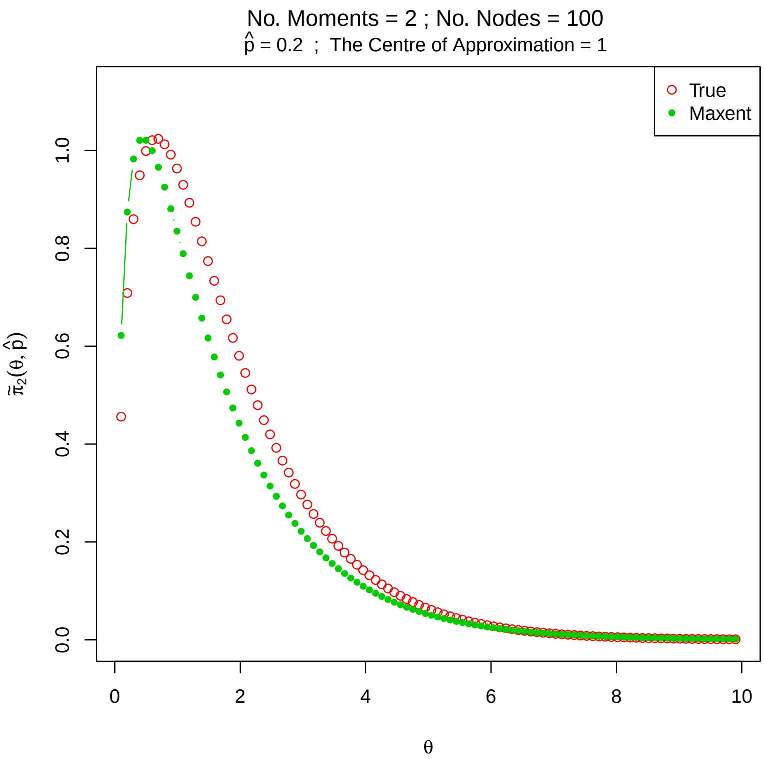
Simplifying **Equation 4.54** , we have:

$$\tilde{\pi}_2(p;\theta) = \frac{\theta^{6p(1-p)}}{4\theta+1} \exp\left\{c_0 - 2\theta + \{c_1 + 8\theta\}\, p + \{c_2 - 8\theta\}\, p^2\right\} , \tag{4.58}$$

where c_0, c_1 and c_2 are given in **Equations 4.55–4.57**.

Figure 4.7 shows the likelihood function $\tilde{\pi}_2(\theta;\hat{p})$ of **Equation 4.58** plotted against the true likelihood function $\pi(\theta;\hat{p})$ of **Equation 3.40** under the diffusion approximation for an observed value $\hat{p} = 0.2$. Only two moments are used here to showcase the analytic form. In practice, a larger number of moments would be used to reconstruct the underlying distribution to a reasonable degree of accuracy.

Figure 4.7: Likelihood function derived from $\tilde{\pi}_2\,(p;\theta)$ centred at $\theta^* = 1$ against the true likelihood function for $\hat{p} = 0.2$.

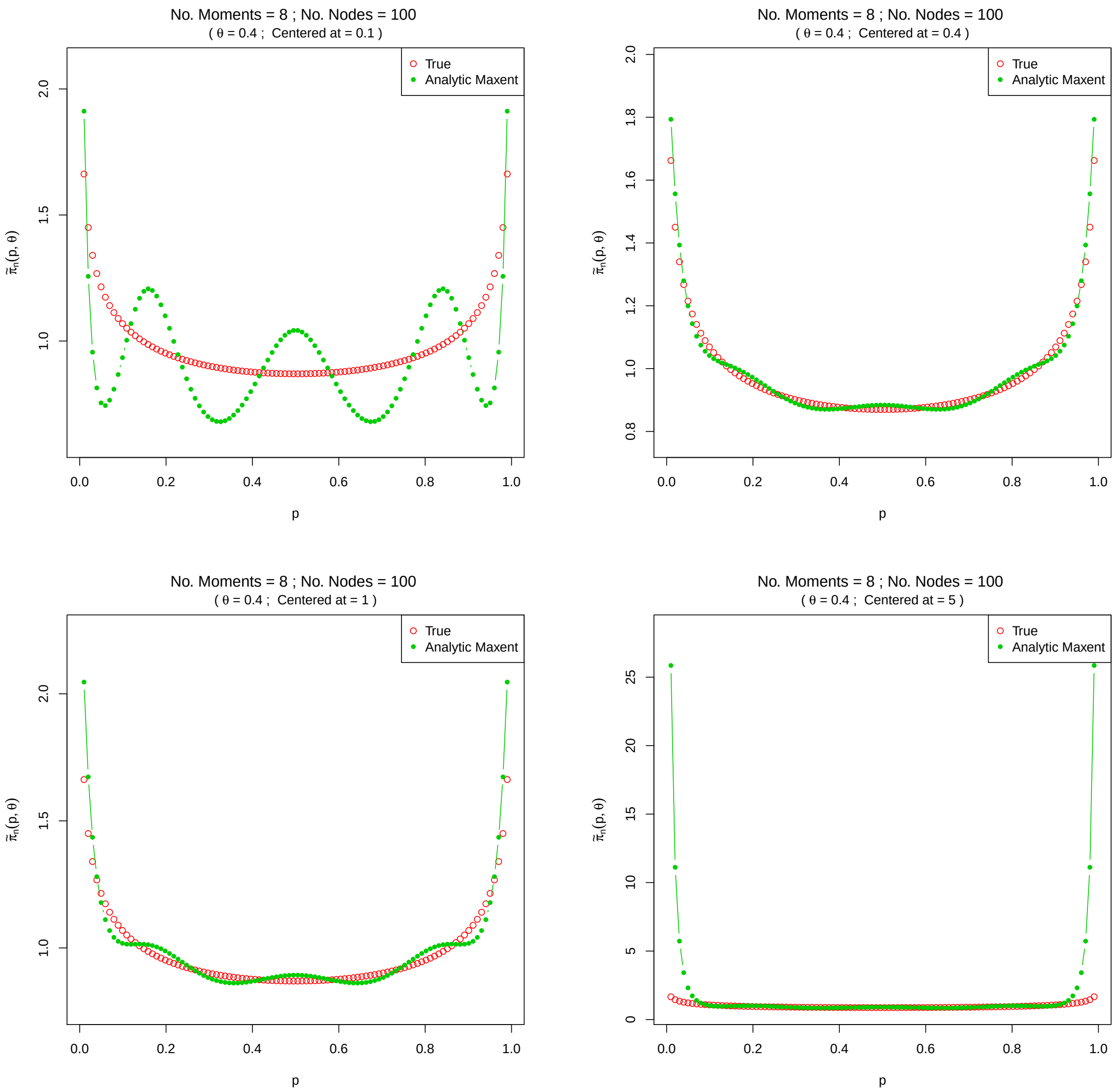


4.4.3 The centre of the approximation

As the name suggests, the analytic Maxent distribution centred at θ^* depends on the choice of θ^*. We first investigate the effect of different centres θ^* on the shape of $\tilde{\pi}_n\,(p)$ with the same θ. We then investigate the performance of analytic Maxent $\tilde{\pi}_n\,(p)$ as we change θ, given a sensible choice of θ^*.

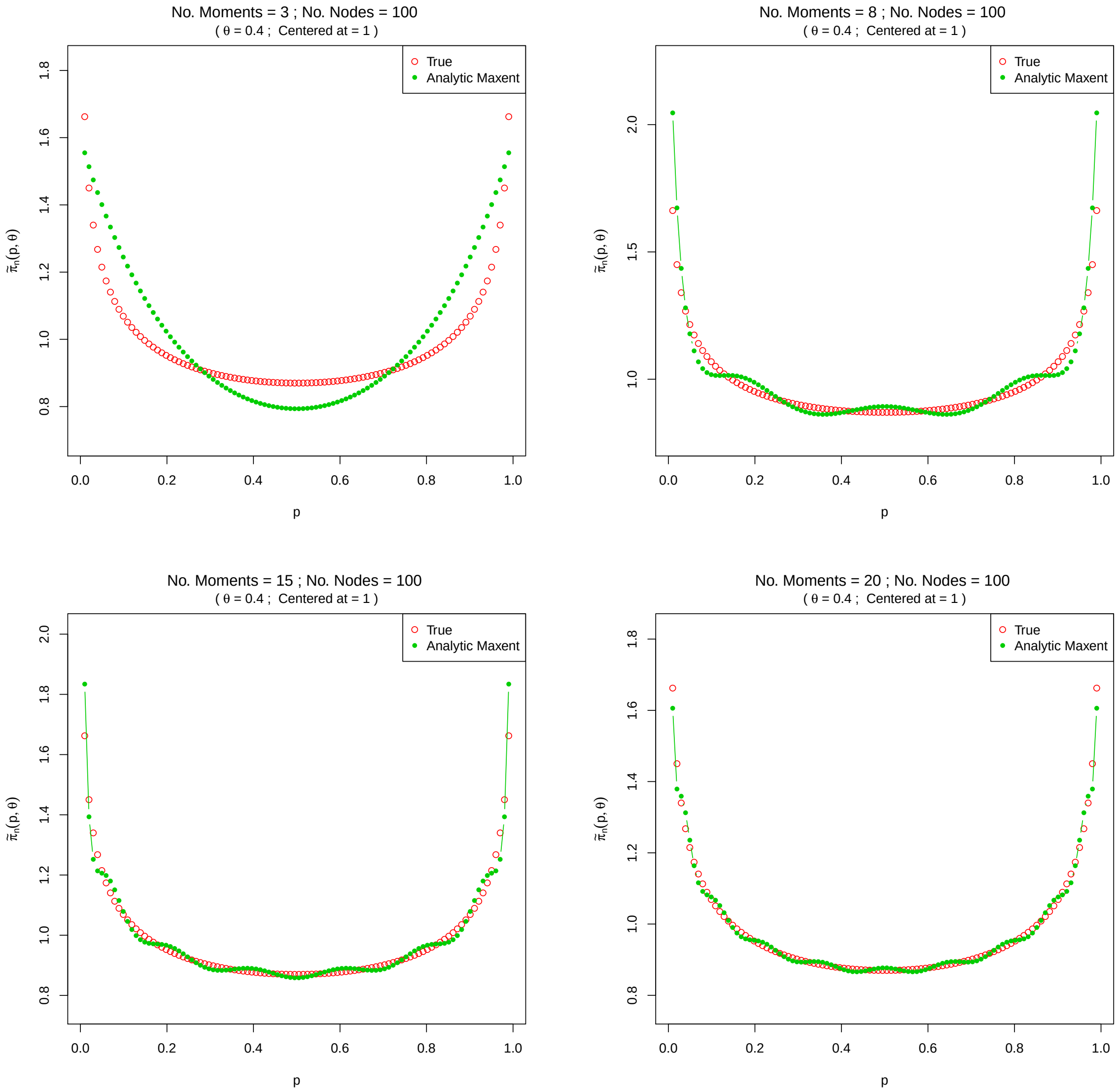
When a small number of moments is used, the choice of the centre affects the shape of $\tilde{\pi}_n\,(p;\theta)$ and thus the quality of the approximation. **Figure 4.8** shows the shape of the analytic $\tilde{\pi}_8\,(p;\theta)$ for a single parameter value $\theta = 0.4$ but at different centres $\theta^* \in \{0.1, 0.4, 1, 5\}$. With a moment sequence up to only order 8, the shape of the distribution differs greatly with different centres.

Figure 4.8: The shape of analytic Maxent distributions, $\tilde{\pi}_8\,(p;\theta)$, centered at different θ^* using a sequence of moments up to and including order 8.



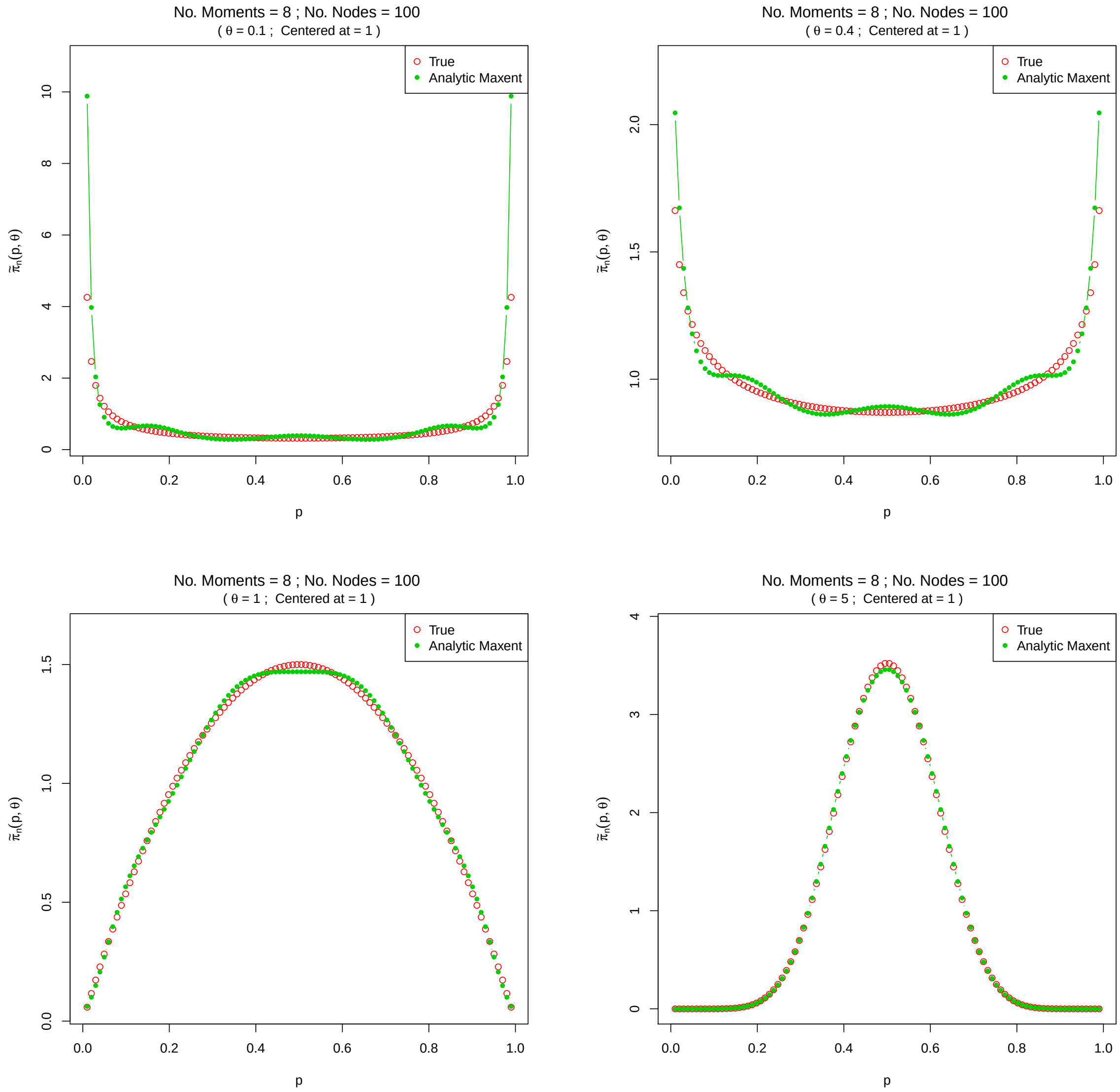
However, the choice of θ^* becomes less important as the number of moments increases, as shown in **Figure 4.9** . When only three moments are used, and the true $\theta = 0.4$, the choice of $\theta^* = 1$ affects the analytic Maxent $\tilde{\pi}_n\,(p;\theta)$ greatly, and we do not have a good approximation. However, by the time we reach 20 moments, the choice of θ^* becomes less influential, and we have a reasonable approximation using $\theta^* = 1$, see the bottom right panel of **Figure 4.9** .

Figure 4.9: Convergence of the analytic Maxent distribution, $\tilde{\pi}_n(p;\theta)$, with a fixed centre $\theta^* = 1$, as the number of moments used increases, $n \in \{3, 8, 15, 20\}$.



With a chosen centre of $\theta^* = 1$ and sufficiently many moments, only 8 moments in this case, **Figure 4.10** shows that the analytic $\tilde{\pi}_n(p;\theta)$ successfully captures the shape of $\pi(p;\theta)$ for various parameter values $\theta \in \{0.1, 0.4, 1, 5\}$. The approximation gradually worsens as the parameter θ moves further away from the center θ^*. This is more apparent when approximating a distribution that is sharp at the boundaries.

Figure 4.10: The shape of analytic Maxent distributions $\tilde{\pi}_8\,(p;\theta)$ centred at $\theta^* = 1$, for different values of θ, using a sequence of moments up to and including order 8.



There is a limit to the number of moments we can include, due to the need for calculating $\lambda_i\,(\theta^*)$ as in **Section 4.3**. The choice of centre θ^* is important to achieve the best approximation for a wide range of θ values. In terms of the SLM model, using a small value for θ^* requires our numerical procedure to reconstruct a distribution that is sharp at the boundaries, which is hard in general for the numerical procedure to do accurately as we indicated previously. Very large values of θ^* create spikes, and lead to numerical problems in a similar way. Therefore a sensible choice of θ^* is a moderate value close to the range of θ that we are interested in. We have used $\theta^* = 1$ for most of our examples for the SLM model, and it gives a satisfactory outcome.

4.4.4 Summary

The performance of the analytic Maxent method at the centre of approximation will be identical to that of the numerical Maxent method. As we move away from the centre, the approximation worsens, similar to the way in which a Taylor series approximation worsens as we move away from the centre of the expansion. Adding more moments improves the range of approximation, so that the approximation can be trusted for a wider range of θ.

Our analytic Maxent approach offers an approximation to $\pi(p; \theta)$, as does the numerical Maxent method, but in addition to numerical approximation the analytic method offers a closed form density of the random variable in terms of the parameter θ. It achieves the analytic form by trading accuracy for tractability as we move away from the centre. Therefore it can implemented as a piecewise approximation for $\pi(p; \theta)$ if a grid of choices for the centre of approximation θ^* is considered. The analytic form may be useful for estimation of θ, because it enables us to construct a likelihood or piecewise likelihood function.

5
Discussion

In this chapter, we first introduce some possible generalisations of the analytic Maxent approach, namely: analytic Maxent solution for multivariate random variables; analytic Maxent solution with multiple parameters; and analytic Maxent solution with higher order derivative conditions. We outline how each of these generalisations can be performed, and point out some issues regarding them. In the subsequent section, we briefly discuss the adequacy of the diffusion approximation for Wright-Fisher type models. In the final section, we discuss the possibility of applying the Maxent method without using the diffusion approximation.

5.1 Generalisations of the Analytic Maxent Method

5.1.1 Analytic Maxent solution for multivariate $\boldsymbol{p}$

Let us consider a multivariate random variable $\boldsymbol{p} \in \Delta^n$ instead of a univariate variable p, where we assume for now that the underlying distribution for this $\boldsymbol{p}$ still depends only on a single parameter θ. Suppose that a sequence of analytic moments is known. The method of **Section 4.4** can be applied to derive the analytic Maxent distribution $\tilde{\pi}_d(\boldsymbol{p})$ centred at any value θ^*, where d is the maximum order of moments used. In the bivariate case, $\tilde{\pi}_d(\boldsymbol{p};\theta)$ has the following form similar to **Equation 4.52** :

$$\tilde{\pi}_d(\boldsymbol{p};\theta) = \exp\left[\sum_{j=0}^{d}\sum_{i_1=0}^{j}\sum_{i_2=0}^{j-i_1}\left\{\tilde{\lambda}_{(i_1,i_2)}(\theta) + c_{(i_1,i_2)}\right\} p_1^{i_1} p_2^{i_2}\right], \tag{5.1}$$

where the $\tilde{\lambda}_{(i_1,i_2)}(\theta)$s are determined in exactly the same way as in **Section 4.4**. The only difference arises during the numerical computation of $\lambda_{(i_1,i_2)}(\theta^*)$ from which the constants $c_{(i_1,i_2)}$ are determined as in **Equation 4.50** . As seen in **Equation 4.19** , a multiple integral instead of a univariate integral needs to be computed.

The analytic Maxent method in dimension $n > 1$ has the same problem that the numerical Maxent method has, namely the evaluation of a high dimensional integral. The constant terms c become harder to compute as the dimension n of the random variable $\boldsymbol{p}$ increases. Therefore the analytic Maxent method as well as the numerical Maxent method is most applicable for low dimensional random variables.

5.1.2 Analytic Maxent with multiple parameters

Suppose now that $p \in (0,1)$, but the distribution $\pi(p;\boldsymbol{\theta})$ depends on more than one parameter. There are now multiple derivatives for each moment, unlike what we had in **Equation 4.42** and **4.45**. For example, suppose that $\pi(p;\boldsymbol{\theta})$ depends on two parameters, θ_1 and θ_2; then we have the following equations instead of **Equation 4.42** :

$$\frac{\partial m_j}{\partial \theta_1}(\theta_1,\theta_2) = \frac{\partial}{\partial \theta_1}\left[\int_0^1 p^j \exp\left\{\sum_{i=0}^{n} \lambda_i(\theta_1,\theta_2)\, p^i\right\} dp\right] \tag{5.2}$$

$$\frac{\partial m_j}{\partial \theta_2}(\theta_1,\theta_2) = \frac{\partial}{\partial \theta_2}\left[\int_0^1 p^j \exp\left\{\sum_{i=0}^{n} \lambda_i(\theta_1,\theta_2)\, p^i\right\} dp\right], \tag{5.3}$$

for $j = 0, 1, 2, \ldots, n$.

Equation 5.2 leads to a system of linear equations in terms of the $\frac{\partial \lambda_i}{\partial \theta_1}(\theta_1,\theta_2)$ and **Equation 5.3** leads to a different system of linear equations in terms of the $\frac{\partial \lambda_i}{\partial \theta_2}(\theta_1,\theta_2)$, for $i = 0, 1, 2, \ldots, n$:

$$\frac{\partial m_j}{\partial \theta_1}(\theta_1,\theta_2) = \sum_{i=0}^{n} \frac{\partial \lambda_i}{\partial \theta_1}(\theta_1,\theta_2) m_{i+j}(\theta_1,\theta_2) \tag{5.4}$$

$$\frac{\partial m_j}{\partial \theta_2}(\theta_1,\theta_2) = \sum_{i=0}^{n} \frac{\partial \lambda_i}{\partial \theta_2}(\theta_1,\theta_2) m_{i+j}(\theta_1,\theta_2)\,, \tag{5.5}$$

for $j = 0, 1, 2, \ldots, n$.

The partial derivatives $\frac{\partial \lambda_i}{\partial \theta_1}(\theta_1,\theta_2)$ and $\frac{\partial \lambda_i}{\partial \theta_2}(\theta_1,\theta_2)$, for $i = 0, 1, 2, \ldots, n$, can be obtained by solving the two linear systems defined by **Equation 5.4** and **5.5**. Therefore, ideally $\lambda_i(\theta_1,\theta_2)$ can be recovered up to a constant c_i since both partial derivatives $\frac{\partial \lambda_i}{\partial \theta_1}(\theta_1,\theta_2)$ and $\frac{\partial \lambda_i}{\partial \theta_2}(\theta_1,\theta_2)$ are available.

However, in general, the two linear systems defined by **Equation 5.4** and **5.5** might not be consistent, in such a way that:

$$\frac{\partial}{\partial \theta_2}\left(\frac{\partial \lambda_i}{\partial \theta_1}\right)(\theta_1, \theta_2) \neq \frac{\partial}{\partial \theta_1}\left(\frac{\partial \lambda_i}{\partial \theta_2}\right)(\theta_1, \theta_2) \quad \text{for} \quad i = 0, 1, 2, \ldots, n\,, \tag{5.6}$$

where $\dfrac{\partial \lambda_i}{\partial \theta_1}(\theta_1, \theta_2)$ and $\dfrac{\partial \lambda_i}{\partial \theta_2}(\theta_1, \theta_2)$ are solutions of the two linear systems.

A example of this can be provided by considering the SLS model with unequal mutation rates (θ_1 and θ_2) in the absence of selection ($s_1 = s_2 = 0$). Using the corresponding analytic moments in **Equation 4.3** , and applying **Equation 5.4** and **Equation 5.5** with the first 2 moments, we have the following $\dfrac{\partial \lambda_i}{\partial \theta_1}$s and $\dfrac{\partial \lambda_i}{\partial \theta_2}$s:

$$\begin{pmatrix} \frac{\partial \lambda_0}{\partial \theta_1} \\ \frac{\partial \lambda_1}{\partial \theta_1} \\ \frac{\partial \lambda_2}{\partial \theta_1} \end{pmatrix} = \begin{pmatrix} -\frac{(\theta_1+\theta_2+1)(6\theta_1+6\theta_2+1)}{(\theta_1+\theta_2)(2\theta_1+2\theta_2+1)} \\ \frac{4(\theta_1+\theta_2+1)}{\theta_1} \\ -\frac{(\theta_1+\theta_2+1)(2\theta_1+2\theta_2+3)}{\theta_1(2\theta_1+1)} \end{pmatrix} \tag{5.7}$$

$$\begin{pmatrix} \frac{\partial \lambda_0}{\partial \theta_2} \\ \frac{\partial \lambda_1}{\partial \theta_2} \\ \frac{\partial \lambda_2}{\partial \theta_2} \end{pmatrix} = \begin{pmatrix} -\frac{\theta_1(\theta_1+\theta_2+1)\left(4\theta_1^2-4\theta_1\theta_2-8\theta_2^2-2\theta_2-1\right)}{(\theta_1+\theta_2)(2\theta_1+2\theta_2+1)(2\theta_2+1)\theta_2} \\ \frac{2(\theta_1+\theta_2+1)(2\theta_1+1-2\theta_2)}{(2\theta_2+1)\theta_2} \\ -\frac{(\theta_1+\theta_2+1)(2\theta_1+2\theta_2+3)}{(2\theta_2+1)\theta_2} \end{pmatrix}. \tag{5.8}$$

Notice that if we differentiate **Equation 5.7** with respect to θ_2 and differentiate **Equation 5.8** with respect to θ_1, none of the resulting second order partial derivatives matches, so they all display the inconsistency in **Equation 5.6** .

This inconsistency leads to multiple analytic Maxent solutions, two solutions $\tilde{\pi}_n^{\theta_2}(p; \theta_1, \theta_2)$ and $\tilde{\pi}_n^{\theta_1}(p; \theta_1, \theta_2)$ for the above example, at a single centre of approximation $\boldsymbol{\theta}^*$. In general, we would hope that $\dfrac{\partial}{\partial\theta_2}\left(\dfrac{\partial\lambda_i}{\partial\theta_1}\right)$ and $\dfrac{\partial}{\partial\theta_1}\left(\dfrac{\partial\lambda_i}{\partial\theta_2}\right)$ converge to the same function as the number of moments n increases, and so the two distinct analytic Maxent solutions would converge to the same $\pi(p;\boldsymbol{\theta})$ as the number of moments n increases. However, a detailed investigation of this problem and providing a proof of this exceeds our current goals.

5.1.3 Analytic Maxent with higher order derivatives

We finally return to the case of univariate p and univariate θ. The analytic Maxent distribution, $\tilde{\pi}_n\,(p;\theta)$, satisfies all of the moment constraints and the first order derivative conditions in **Equation 4.42** by formulation. For completeness, let us consider higher order derivative conditions for the analytic Maxent distribution. For example, we wish to consider whether $\tilde{\pi}_n\,(p;\theta)$ satisfies the following second order derivative conditions,

$$m''_j\,(\theta) = \frac{\partial^2}{\partial\theta^2}\left[\int_0^1 p^j\tilde{\pi}_n\,(p;\theta)\,dp\right] \quad \text{for} \quad j = 0,1,2\ldots,n\,, \tag{5.9}$$

where $m''_j(\theta)$ denotes the second derivative of the known analytic expression $m_j(\theta)$ with respect to θ.

Higher order derivatives are iterative, thus **Equation 5.9** is equivalent to

$$\frac{\partial}{\partial\theta}\left(\frac{\partial m_j}{\partial\theta}\right) = \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\left[\int_0^1 p^j \exp\left\{\sum_{i=0}^{n}\lambda_i\,(\theta)\,p^i\right\} dp\right]\right] \quad \text{for} \quad j = 0,1,2\ldots,n\,. \tag{5.10}$$

Equation 5.10 is true if:

$$\frac{\partial m_j}{\partial \theta} = \frac{\partial}{\partial \theta}\left[\int_0^1 p^j \exp\left\{\sum_{i=0}^{n} \lambda_i\left(\theta\right) p^i\right\} dp\right] \quad \text{for} \quad j = 0, 1, 2 \ldots, n\,, \tag{5.11}$$

which are exactly the first order derivative conditions we impose on the analytic Maxent distribution $\tilde{\pi}_n\left(p;\theta\right)$ in **Equation 4.42** .

Thus, the analytic Maxent distribution $\tilde{\pi}_n\left(p;\theta\right)$ satisfies the second order derivative conditions and indeed any higher order derivative conditions. This can be proved using mathematical induction. Therefore higher order derivative conditions are redundant for determining the analytic Maxent distribution $\tilde{\pi}_n\left(p;\theta\right)$.

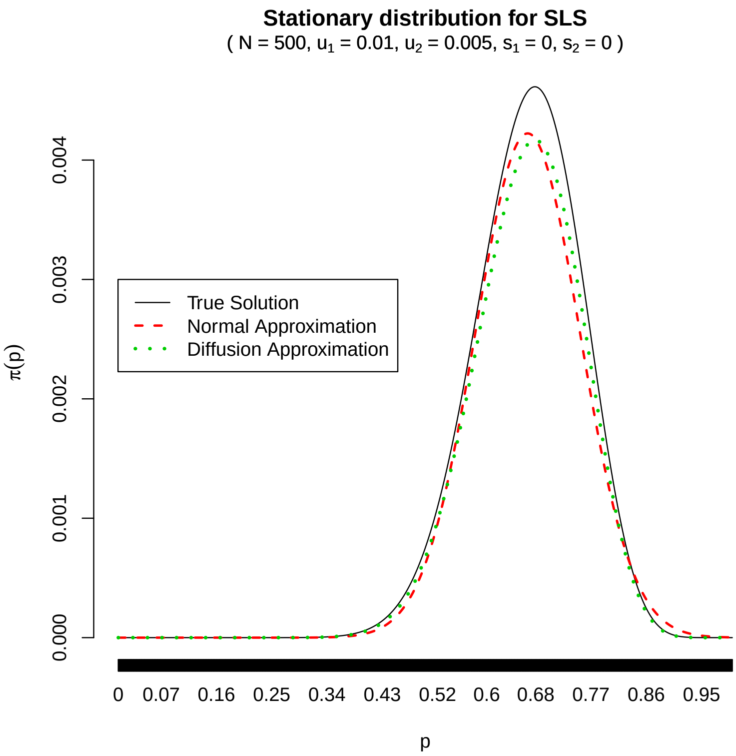
5.2 Adequacy of the Diffusion Approximation

We used the diffusion approximation to recover the necessary moments for both our numerical Maxent procedures and our analytic Maxent procedures. We chose to use the diffusion approximation mainly because it provides tractable analytic moments at steady state. However, this adds an extra layer of approximation error in addition to the approximation of using the maximum entropy procedures to recover a distribution from its moments.

The performance of the diffusion approximation as a method for capturing the behaviour of population genetics models was studied some decades ago. **Ewens (1963c)** gave a numerical study, while **Watterson (1962)** and **Ewens (1965)** gave some mathematical justifications for using the diffusion approximation for various models. However, a recent study by **Parsons, Quince and Plotkin (2010)** showed that the diffusion approximation can lead to erroneous and even contradictory results in some scenarios. Their study was done using microbes, and involved violation of many assumptions. We think the poor performance seen in their study was due to the violation of assumptions rather than inherent problems with the diffusion approximation.

The adequacy of the diffusion approximation can be judged using the investigation of **Watterson (1962)** and **Moran (1962)**, who point out that the limiting argument of the diffusion approximation is very sensitive to the rate with which various population parameters change as $N \to \infty$. Therefore the performance of the diffusion approximation depends on the so-called balance of the relevant population parameters.

Figure 5.1: Performance of the diffusion approximation with $N = 500$.



Using the numerical method of finding the stationary distribution for the original discrete process, which we used in **Subsection 2.2.2**, we can compare the true stationary distribution generated by the diffusion approximation against the true stationary distribution of the original discrete process, for models where both of these are available. We find that there are cases where the diffusion approximation can achieve a reasonable level of accuracy for N as small as 500. In **Figure 5.1**, the diffusion approximation represents **Equation 3.39** which is the analytic solution for the diffusion equation of the SLS model at steady state. However, N much bigger than 500 is needed in general. For a further discussion on this topic, and an error estimation of the diffusion approximation, see **Ethier and Norman (1977)**, who examine the diffusion approximation for the Wright-Fisher model in mathematical detail.

Because we used moments generated by the diffusion approximation, the performance of our Maxent computation of the stationary distribution as an approximation of the original discrete process can only be as good as the diffusion approximation itself. The performance of the diffusion approximation for describing a real population, in turn, can only be as good as the original discrete model. If many assumptions of the Wright-Fisher model are violated, as in the case pointed out by **Parsons, Quince and Plotkin (2010)**, our end product can be a very poor description of reality.

5.3 Maxent without the Diffusion Approximation

We are aware of numerical methods for solving a partial differential equation such as the Kolmogorov equation. Numerical stationary distributions under the diffusion approximation could therefore be obtained from the solution of the partial differential equation. For example, **Boitard and Loisel (2007)** propose a numerical method based on finite differences for solving the Kolmogorov equation corresponding to a two-locus Wright-Fisher model under the diffusion approximation.

However, our approaches of both numerical and analytic Maxent are more general, in the sense that we do not need to rely on the diffusion approximation. The numerical and analytic Maxent methods can be applied to other sources of moments. For example, the numerical Maxent method could incorporate numerical moments from accurate coalescent simulations to reconstruct the underlying distribution. The analytic Maxent method can work with any source of analytic moments, such as the analytic moments of the original discrete process gained in **Subsection 2.2.3** for the SLS model. In fact, we have already shown an example of this under the guise of the normal approximation in **Subsection 2.2.3** for the SLS model. It can be shown that applying Maxent using the first two moments only, namely mean and variance, leads to a normal distribution for $\tilde{\pi}_2(p)$. Therefore, the normal approximation is in fact a special case of Maxent using only the first two moments. In **Figure 5.1** , the normal approximation that uses the moments of the original discrete process, not moments from the diffusion approximation, is plotted along with the diffusion approximation and the numerical solution of the discrete process.

All our examples of the numerical and analytic Maxent methods have been shown in terms of reconstructing a stationary distribution of a genetic process. However, the same methods can also be used to incorporate numerical and analytic moments from any other distribution. The methods can be applied in any situation where an approximation of an underlying distribution is needed, given a sequence of its moments.

6

Conclusions and Future work

6.1 Conclusions

For a range of genetic models, we have demonstrated that the analytic stationary moments can be obtained without first finding the stationary distribution, using the diffusion approximation. In **Section 4.1**, we illustrated this procedure for the SLS model without selection, the SLM model with two alleles, and the TLD model. For the two-locus model with linkage (TLD), we derived the analytic stationary moments in terms of three gametic frequencies (p_1, p_2 and p_3), instead of using two allelic frequencies (p and q) and the coefficient of linkage disequilibrium D.

In **Section 4.2**, we introduced the maximum entropy principle to reconstruct distributions in population genetics from their numerical or analytic moments. In **Section 4.3**, we successfully reconstructed the stationary distribution for the two-locus model with linkage (TLD) using its numerical moments under the diffusion approximation. Using the stationary distribution, we computed the expected linkage disequilibrium $\mathbb{E}\left(r^2\right)$ and the variance $\mathbb{V}\left(r^2\right)$ for a range of scaled mutation rates θ and scaled recombination rates ρ.

Our computation of $\mathbb{E}\left(r^2\right)$ agrees with the recent computation of $\mathbb{E}\left(r^2\right)$ by **Song and Song (2007)**, except for very small θ for which the density is very spiked and hard to reconstruct. Our computation of $\mathbb{V}\left(r^2\right)$ shows that it is a strictly decreasing function of ρ, but it is more complicated as a function of θ. It initially increases as θ increases from zero, but starts to decrease after reaching a certain maximum point. This maximum point of θ after which $\mathbb{V}\left(r^2\right)$ starts to decrease depends on the ρ value of the underlying system.

In **Section 4.4**, we proposed a general method of reconstructing a continuous density function from its analytic moments. This method offers a new approach for solving problems where the likelihood function is not available but the corresponding analytic moments are. This opens the possibility of estimating the parameter of an underlying population where only data and theoretical moments are available but not the likelihood function. The diallelic SLM model is used to show that this method is accurate and simple to apply.

6.2 Future Work

Following the investigations described in this thesis, a number of further projects could be undertaken. There are two main goals: firstly, to further improve the stability and efficiency of our Maxent approach, and secondly to apply our method to a wider range of problems in population genetics and other fields. The first two subsections discuss the first goal, and the final subsection discusses the second goal.

6.2.1 Multivariate Chebyshev polynomials

We have programmed the numerical Maxent approach in R and Fortran. It is both fast and accurate in univariate cases. However, its efficiency decreases quickly as the dimension of the random variable $\boldsymbol{p}$ increases. It takes more than a day to reconstruct the trivariate distribution for the TLD model on a supercomputer with 40 CPUs for just one value of (θ, ρ).

We have identified that the slowness is a result of an inefficient usage of moments. Currently, the available power moments are put into multivariate Chebyshev polynomials, which are constructed using tensor products of the shifted univariate Chebyshev polynomials as in **Equation 4.28** , and the multivariate distribution is constructed using these multivariate Chebyshev polynomials. This procedure is described in **Section 4.3.2**.

The multivariate Chebyshev polynomials constructed as tensor products of univariate Chebyshev polynomials are optimal for problems on a cubic region,

$$\{(p_1, p_2, p_3, \ldots, p_n) \in \mathbb{R}^n : 0 \le p_i \le 1 \quad \text{for} \quad i = 1, 2, \ldots, n\} \ .$$

In population genetics, we are often interested in frequencies, which lead to problems defined on standard simplexes,

$$\left\{ (p_1, p_2, p_3, \ldots, p_n) \in \mathbb{R}^n : \sum_{i=1}^{n} p_i \leq 1 \quad \text{and} \quad p_i \geq 0 \quad \text{for} \quad i = 1, 2, \ldots, n \right\}.$$

Using multivariate Chebyshev polynomials constructed specifically for standard simplexes would greatly improve the performance of our method. However, to the best of our knowledge, multivariate Chebyshev polynomials defined on standard simplexes are not readily available. Very recently **Ryland and Munthe-Kaas (2011)** developed a method for constructing multivariate Chebyshev polynomials on a triangular domain. This new development might be what we need, and is worthwhile investigating. Also see **Farouki, Goodman and Sauer (2003)** for a recent discussion on orthogonal polynomials for a simplex domain. A successful adoption of multivariate Chebyshev polynomials defined on standard simplexes would allow us to achieve the same level of accuracy with far fewer moments, hence greatly alleviate the numerical difficulties in multivariate cases.

6.2.2 Cubature and modern computing methods

Computing a high dimensional integral using Gaussian cubature many times is the second source of numerical difficulty that we experience in the multivariate cases. It demands a huge memory and 40 CPUs to evaluate the integral in **Equation 4.19** using a parallel computing array to bring our method for the 3-dimensional TLD model to a acceptable speed. An investigation of methods of adaptive numerical integration might lead to a more efficient evaluation of the integral; see **Genz and Cools (2003)** for an example of such adaptive algorithms.

Numerical integration using quadrature is a highly parallelisable computing problem.

We have implemented our method in Fortran in a traditional parallel environment using CPUs. A very recent advance in computing is the use of accelerators with different computer architectures from the CPU, and this has offered new possibilities in the past 5 years. See **Gillan, Steinke, Bock, Borchert, Spence and Scott (2012)** for a further discussion of modern computing methods for high dimensional numerical integrations.

Even a small improvement in the evaluation of the integral will lead to a big improvement in terms of the overall efficiency of our method, because the integral is evaluated many times during the optimisation.

6.2.3 Coalescent simulations and Maxent

Traditionally, physicists and engineers use the numerical Maxent approach to reconstruct distributions from numerical sequences of sample moments that are available from experimental data. See **Miller and Liu (2002)** for a convergence result for the Maxent approach when sample moments are used. However, most scientific fields do not have the data quality and quantity that is available for these problems in physics and engineering.

In population genetics, we might not have sufficient data quality and quantity to use sample moments from data. However, coalescent simulations have become widely acceptable, and the sample moments from coalescent simulations could provide a valid source of information for the numerical Maxent procedure to operate with. This would return an explicit density function of the distribution underlying the coalescent simulations. Therefore the performance of the numerical Maxent method with sample moments from coalescent simulations is worthwhile investigating.

A

Appendix A

A.1 Conditional Expectations for the SLM Model

For the SLM model, the conditional expectations are symmetric. Expanding the conditional expectations using Taylor series, we have the terms in **Table A.1**, where u is the mutation probability, k is the number of distinct allelic types, and p_1, p_2, $\ldots$, p_{k-1} are the diffusion variables for allele types A_1, A_2, $\ldots$, A_{k-1} at locus A.

Table A.1: Taylor series expansion for the conditional expectations of the SLM diffusion process.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left(\delta p_i\right) &= \left(1-kp_i\right)u+R_2\left(u;\boldsymbol{p}\right)\\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left\{\left(\delta p_i\right)^2\right\} &= \frac{p_i\left(1-p_i\right)}{2N}+\frac{u}{2N}\left(kp_i-1\right)\left(2p_i-1\right)+R_2\left(u;\boldsymbol{p}\right)\\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left(\delta p_i\delta p_j\right) &= -\frac{p_ip_j}{2N}+\frac{u}{2N}\left\{p_i\left(kp_j-1\right)+p_j\left(kp_i-1\right)\right\}+R_2\left(u;\boldsymbol{p}\right)
\end{aligned}$$

A.2 Conditional Expectations for the TLD Model

For the TLD model, the conditional expectations are not entirely symmetric. Expanding the conditional expectations using Taylor series, we have the first order terms in **Table A.2** and the second order terms in **Table A.3**, where u is the mutation probability, C is the recombination fraction between the two loci, and p_1, p_2, p_3 are the diffusion variables for the possible gamete types 1, 2 and 3.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left(\delta p_1\right) &= -2up_1+up_2+up_3-CD+R_2\left(u,C;\boldsymbol{p}\right)\\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left(\delta p_2\right) &= u-3up_2-up_3+CD+R_2\left(u,C;\boldsymbol{p}\right)\\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left(\delta p_3\right) &= u-up_2-3up_3+CD+R_2\left(u,C;\boldsymbol{p}\right)
\end{aligned}$$

Table A.2: Taylor series expansion for the conditional expectations of the first order terms of the TLD diffusion process.

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left\{(\delta p_1)^2\right\} &= \frac{p_1(1-p_1)}{2N} + \frac{u(-2p_1+p_2+p_3)(1-2p_1)}{2N} \\
&\quad + \frac{C(2p_1-1)D}{2N} + R_2(u,C;\boldsymbol{p}) \\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left\{(\delta p_2)^2\right\} &= \frac{p_2(1-p_2)}{2N} + \frac{u(1-3p_2-p_3)(1-2p_2)}{2N} \\
&\quad + \frac{C(2p_1-1)D}{2N} + R_2(u,C;\boldsymbol{p}) \\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}\left\{(\delta p_3)^2\right\} &= \frac{p_3(1-p_3)}{2N} + \frac{u(1-p_2-3p_3)(1-2p_3)}{2N} \\
&\quad + \frac{C(2p_3-1)D}{2N} + R_2(u,C;\boldsymbol{p}) \\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}(\delta p_1\delta p_2) &= -\frac{p_1p_2}{2N} + \frac{-u\{p_1(1-3p_2-p_3)+p_2(-2p_1+p_2+p_3)\}}{2N} \\
&\quad + \frac{C(p_1+p_2)D}{2N} + R_2(u,C;\boldsymbol{p}) \\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}(\delta p_1\delta p_3) &= -\frac{p_1p_3}{2N} + \frac{-u\{p_1(1-p_2-3p_3)+p_3(-2p_1+p_2+p_3)\}}{2N} \\
&\quad + \frac{C(p_1+p_3)D}{2N} + R_2(u,C;\boldsymbol{p}) \\
\mathbb{E}_{\boldsymbol{\delta p}|\boldsymbol{p}}(\delta p_2\delta p_3) &= -\frac{p_2p_3}{2N} + \frac{-u\{p_2(1-p_2-3p_3)+p_3(1-3p_2-p_3)\}}{2N} \\
&\quad + \frac{C(p_2+p_3)D}{2N} + R_2(u,C;\boldsymbol{p})
\end{aligned}
$$

Table A.3: Taylor series expansion for the conditional expectations of the second order terms of the TLD diffusion process.

B

Appendix B

Table B.1: Some TLD stationary moments under the diffusion approximation.

Degree	Expectations at steady state for the TLD model under the diffusion approximation
1	$\mathbb{E}(p_1) = \mathbb{E}(p_2) = \mathbb{E}(p_3) = \frac{1}{4}$
2	$\mathbb{E}\left(p_1^2\right) = \mathbb{E}\left(p_2^2\right) = \mathbb{E}\left(p_3^2\right) = \frac{1}{4\Lambda}\left(4096\theta^4 + 1536\theta^3\rho + 128\theta^2\rho^2 + 4352\theta^3 + 1216\rho\theta^2 + 64\rho^2\theta + 1568\theta^2 + 304\rho\theta + 8\rho^2 + 216\theta + 26\rho + 9\right)$
	$\mathbb{E}(p_1p_2) = \mathbb{E}(p_1p_3) = \frac{\theta}{\Lambda}\left(1024\theta^3 + 384\rho\theta^2 + 32\rho^2\theta + 704\theta^2 + 192\rho\theta + 8\rho^2 + 144\theta + 26\rho + 9\right)$
	$\mathbb{E}(p_2p_3) = \frac{8\theta^2}{\Lambda}\left(128\theta^2 + 48\rho\theta + 4\rho^2 + 72\theta + 14\rho + 9\right)$
3	$\mathbb{E}\left(p_1^3\right) = \mathbb{E}\left(p_2^3\right) = \mathbb{E}\left(p_3^3\right) = \frac{1}{4\Lambda}\left(8\rho^2 + 1024\theta^4 + 180\theta + 26\rho + 920\theta^2 + 200\rho\theta + 1728\theta^3 + 512\rho\theta^2 + 32\rho^2\theta + 32\theta^2\rho^2 + 384\rho\theta^3 + 9\right)$
	$\mathbb{E}\left(p_1^2p_2\right) = \mathbb{E}\left(p_1p_2^2\right) = \mathbb{E}\left(p_1^2p_3\right) = \mathbb{E}\left(p_1p_3^2\right) = \frac{\theta}{2\Lambda}\left(512\theta^3 + 480\theta^2 + 192\rho\theta^2 + 132\theta + 144\rho\theta + 16\rho^2\theta + 9 + 26\rho + 8\rho^2\right)$
	$\mathbb{E}\left(p_2^2p_3\right) = \mathbb{E}\left(p_2p_3^2\right) = \frac{2\theta^2}{\Lambda}\left\{(2\rho + 5 + 16\theta)(2\rho + 3 + 8\theta)\right\}$
	$\mathbb{E}(p_1p_2p_3) = \frac{2\theta^2}{\Lambda}\left(128\theta^2 + 3 + 56\theta + 12\rho + 4\rho^2 + 48\rho\theta\right)$

where $\Lambda = (8\theta + 1)\left(2048\theta^3 + 768\rho\theta^2 + 64\rho^2\theta + 1280\theta^2 + 304\rho\theta + 8\rho^2 + 216\theta + 26\rho + 9\right)$

C

Appendix C

C.1 Entropy

The term entropy was first introduced in the field of thermodynamics more than a hundred years ago. It has since penetrated many disciplines, but takes various meanings in different fields. The idea of entropy flourished after the development of the maximum entropy principle about fifty years ago. The maximum entropy principle is a powerful tool for recovering an unknown distribution from limited information. In this section we first use an intuitive example to explain what entropy is and why it is of interest to us. In the subsequent subsection, we discuss Shannon's entropy and its special properties. We then give a mathematical justification of the maximum entropy principle.

C.1.1 Information, uncertainty, and probability

Entropy is a measure of organisation at a molecular level in physics, and it is often envisaged as a measure of probabilistic uncertainty in statistics. The latter is the idea we shall use, often referred to as information entropy.

For a stochastic system, information entropy is a measure of the average missing information regarding the predictability of the future of the system. A large information entropy is attached to a relatively unpredictable system, as it means we need a relatively large amount of information before removing all the uncertainty in the system. A deterministic system has zero information entropy because its future outcomes are predictable in the absence of any further information. Uncertainty and information are opposite sides of the same coin; having more of one means having less of the other, and therefore entropy is often regarded as a measure of uncertainty.

A probability distribution can also be considered as a description of uncertainty, so it is reasonable to consider that every probability distribution has an entropy level attached to it. Therefore, choosing a probability distribution to model a stochastic system inevitably assigns a certain entropy level to the system.

To give an intuitive reason as to why we require the concept of information entropy, let us imagine a forgetful politician promoting his policies in five cities before an election. Every day his limo takes him to the city of his choice. Being a forgetful politician, he cannot remember the cities he has visited previously, and so he chooses the next city impulsively and unpredictably. For simplicity, let us imagine that he chooses the next city independently from his previous choices, so he may end up making numerous visits to the same city.

If no information (data or prior knowledge) is available regarding the politician's long-term subconscious preferences, we should not be eager to make a guess as to where he will end up next. However, if we are forced to say something about the likelihood of his future choices, the best we can say without being unsound is that he is equally likely to visit each city.

Mathematically the equally-likely model can be summarised as shown in **Table C.1**.

Table C.1: Discrete Uniform distribution for the Politician example.

	City	A	B	C	D	E
	x	1	2	3	4	5
Equally-likely Model	$Pr\,(X = x)$	0.2	0.2	0.2	0.2	0.2

If we are forced to choose a solution to a problem about which we have little or no information, it is natural to choose the most conservative option. That is, if all solutions appear to be equally correct with respect to the information we possess, our choice should be the one that is the least radical or extreme. This is the rationale behind choosing the *Equally-likely Model* in **Table C.1**. Any other choice of distribution, with unequal probabilities, would imply that we somehow have additional insight into the politician's choices. This principle was initially considered by Jacob Bernoulli and Pierre Simon Laplace, and later called *the principle of indifference.*

Unfortunately, the principle of indifference is almost as far as intuition can carry us. Being provided with further information, it often becomes unclear which of the possible distributions is the most conservative choice. Let us consider a second scenario. Imagine that we have been given the long term average of the politician's visiting preferences regarding the five cities.

Mathematically, suppose we have the knowledge

$$\mathbb{E}(X) = 2.5\,.$$

This provides a constraint and gives us additional information about the politician's distribution. This constraint narrows down our choices, but it does not present us with a unique distribution. For example, both *Model 1* and *Model 2* in **Table C.2** satisfy the constraint in the second scenario.

Table C.2: Two of many possible distributions for the Politician example.

	City	A	B	C	D	E	$\mathbb{E}(X)$
	x	1	2	3	4	5	
Model 1	$Pr(X = x)$	0.325	0.2	0.2	0.2	0.075	2.5
Model 2	$Pr(X = x)$	0.1	0.6	0.1	0.1	0.1	2.5

In order to choose the most conservative model for the second scenario, we need to understand the relationship between uncertainty and probability distributions in terms of our problem. If there were fewer cities, then the level of uncertainty would be lower. Clearly, there would be no uncertainty if there was only one city for the politician to choose from. Secondly, if he showed a strong preference for particular cities then there would also be a lower degree of uncertainty. Visiting the same city 99 times in 100 days is hardly unpredictable. Hence, intuitively, the level of uncertainty depends on the size of the sample space and on how the probabilities are allocated. Distributions with a larger support or a greater dispersion are associated with greater uncertainty. A sensible measure of uncertainty should reflect these, as well as other, basic properties. Variance is a measure of dispersion or variation, and in some sense variance is a measure of uncertainty as well.

In the next subsection, we will see that certain properties of information entropy make it a better measure of uncertainty than other quantities such as variance.

Returning to the politician's second scenario, *Model 1* and *Model 2* in table C.2 have the same support, but it can be seen in Figure C.1 that *Model 2* is more informative and has less uncertainty.

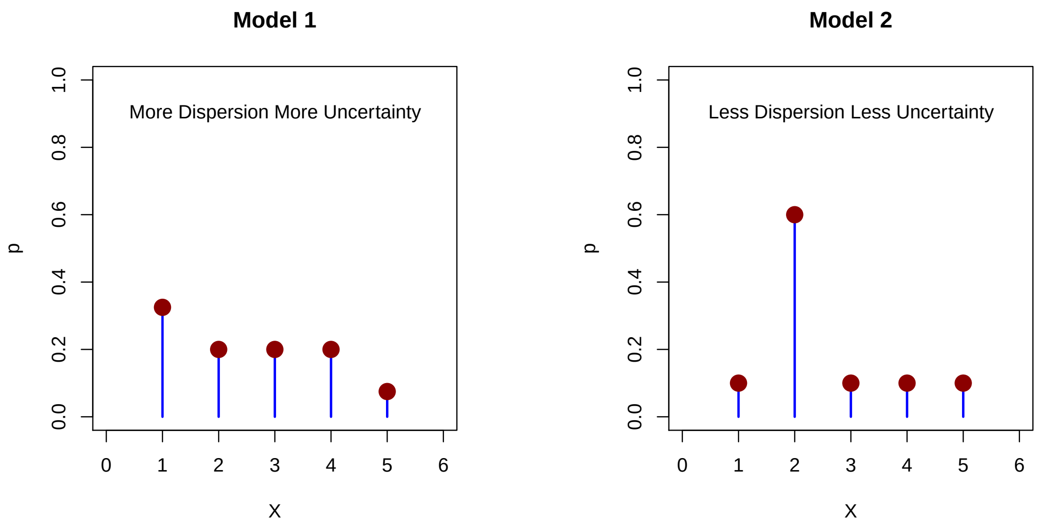


Figure C.1: Different distribution shapes, hence different levels of dispersion and entropy.

Adopting a model with an excessively low level of uncertainty is equivalent to claiming information or knowledge that we do not possess. In terms of our second scenario, once we have used all of the available information by imposing the constraint $\mathbb{E}(X) = 2.5$, we have no information left to justify choosing a distribution other than that with the maximum level of uncertainty.

If information entropy is adopted as a measure of uncertainty, then the most conservative model in terms of information or uncertainty is the maximum entropy solution. It is intuitively clear that the *Equally-likely Model* in **Table C.1** should be the maximum entropy solution in the first scenario, however it is not clear which model should be the maximum entropy solution for the second scenario. To determine the maximum entropy solution, we need to define information entropy mathematically.

C.1.2 Shannon's entropy

Information entropy or Shannon's entropy is an uncertainty measure introduced by **Shannon and Weaver (1948)** for discrete distributions. Useful properties of Shannon's entropy make it a widely accepted uncertainty measure.

Let $H[f_X]$ denote Shannon's entropy for a discrete random variable X, which has probability mass function f_X. The Shannon's entropy is defined as a functional of the probability mass function; it is *not* a function of the random variable X.

Let $f_X(x_i) = p_i$ for $i = 0, 1, 2, \ldots, n$, where $x_0, x_1, x_2, \ldots, x_n$ are all the values that X takes with nonzero probability. If this notation is used, then Shannon's entropy $H[f_X]$ can be viewed as a real-valued function $h(p_0, p_1, p_2, \ldots, p_n)$. Specifically,

$$h : \Delta^n \to \mathbb{R}\,,$$

where Δ^n denotes the standard n-simplex,

$$\Delta^n = \left\{ (p_0, p_1, p_2, \ldots, p_n) \in \mathbb{R}^{n+1} : \sum_{i=0}^{n} p_i = 1 \quad \text{and} \quad p_i \geq 0 \quad \text{for} \quad i = 0, 1, \ldots, n \right\}.$$

With this notation, Shannon's entropy has the form

$$
\begin{aligned}
H\left[f_X\right] &= h\left(p_0, p_1, p_2, \ldots, p_n\right) \\
&= -K \sum_{i=0}^{n} p_i \log_b p_i \,,
\end{aligned}
\tag{C.1}
$$

where K and b are positive constants, which merely determine the choice of a unit of measure. For the sake of a simpler notation, let us set $K = 1$ and b to be Euler's number $e = 2.71828\ldots$. These two specifications do not affect any properties of Shannon's entropy that are of interest to us. Hence,

$$
\begin{aligned}
H\left[f_X\right] &= h\left(p_0, p_1, p_2, \ldots, p_n\right) \\
&= -\sum_{i=0}^{n} p_i \ln p_i \,.
\end{aligned}
$$

Shannon and Weaver (1948) derived this measure by demanding three reasonable properties, and they also proved that this is the only measure that possesses all three properties. See also **Shannon (2001)** for a reprinted version with corrections. The three properties are listed in **Table C.3**.

The continuity property in **Table C.3** is clearly needed for any sensible uncertainty measure. The monotonic increasing property in **Table C.3** can be understood in terms of the politician example: more cities create greater uncertainty if all cities have an equal chance of being visited.

The property of strong additivity in **Table C.3** requires additional explanation. Every joint distribution can be decomposed into a marginal distribution and a conditional distribution. The rationale behind strong additivity is that the level of uncertainty for

Table C.3: Three important properties of Shannon's Entropy

1. Continuity :

   $h(p_0, p_1, p_2, \ldots, p_n)$ is a continuous function of $p_0, p_1, p_2, \ldots, p_n$.

2. Monotonic increasing :

   If f_X^n is a discrete uniform function on $n+1$ distinct support points,

   then $H\left[f_X^n\right]$ increases monotonically as n increases.

3. Strong additivity :

   Given discrete random variables Q and W such that,

   $$f_{Q,W}(q, w) = f_Q(q) f_{W|Q}(w; q) \, ,$$

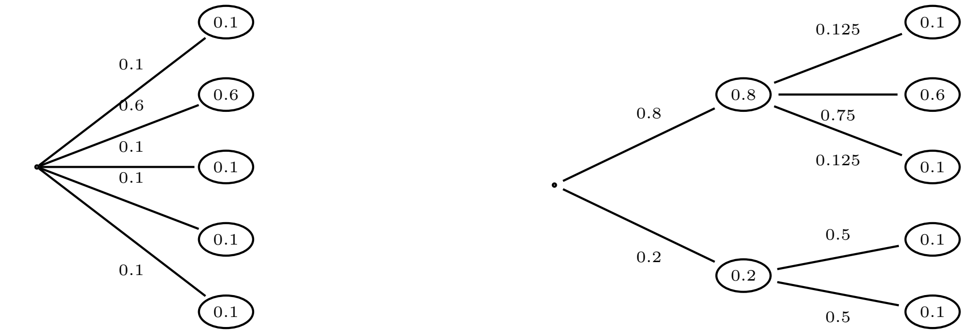
   then $H\left[f_{Q,W}\right] = H\left[f_Q\right] + \mathbb{E}_Q\left(H\left[f_{W|Q}\right]\right)$,

   where $f_{Q,W}$ is the joint probability mass function of Q and W,
   f_Q is the marginal probability mass function of Q,
   $f_{W|Q}$ is the conditional probability mass function of W given Q.

the joint distribution must be the sum of the uncertainty contribution from the marginal distribution and the weighted uncertainty contribution from the conditional distribution. The uncertainty should be neither increased nor decreased by the decomposition.

Taking *Model 2* in **Table C.2** as an example, a possible decomposition is illustrated as a tree diagram on the right in **Figure C.2** , while the original *Model 2* is on the left. Intuitively, **Figure C.2** represents the scenario that the politician has two decisions to make instead of just one. He first chooses the province of the city that he wants to visit. Then he chooses an individual city within the selected province. The decision is decomposed into two decisions, but the outcomes and probabilities are the same as those under the original model.

Figure C.2: Decomposition of the politician's choice.



Before Decomposition After Decomposition

Hence the two models in **Figure C.2** are essentially the same model represented in two different ways, so they should be associated with the same degree of uncertainty. Any other decompsition should also give the same uncertainty. In terms of *Model 2* and Figure C.2, strong additivity means :

$$\begin{aligned}
&\overbrace{H\left[0.1, 0.6, 0.1, 0.1, 0.1\right]}^{\text{Total Uncertainty}} \\
&\qquad = \underbrace{H\left[0.8, 0.2\right]}_{\substack{\text{Contribution}\\ \text{from the marginal}}} \\
&\qquad + \\
&\qquad \underbrace{0.8H\left[0.125, 0.75, 0.125\right] + 0.2H\left[0.5, 0.5\right]}_{\substack{\text{Contribution}\\ \text{from the conditional}}}
\end{aligned} \tag{C.2}$$

Shannon's entropy also has other desirable properties as an entropy measure, such as non-negativity and symmetry. See **Kapur and Kesavan (1992)** for a brief summary, and **Aczél, Forte and Ng (1974)**, **Aczél and Daróczy (1975)**, **Mathai and Rathie (1975)** for more detailed discussion of its properties.

C.1.3 Differential entropy

So far we have considered entropy measures for discrete distributions. Shannon's entropy can be generalised to continuous intervals or continuous regions by replacing the summation in **Equation C.1** with an integral. The continuous counterpart for Shannon's entropy is called either continuous entropy or differential entropy:

$$H[f_X] = -\int_A f_X(x) \ln f_X(x)\, dx\,, \tag{C.3}$$

where f_X is the density function of X and A is the entire support of X.

Although the differential entropy defined in **Equation C.3** seems to be a natural extension of Shannon's entropy, it lacks some desirable properties that Shannon's entropy holds. An example is that, in some cases, it can take negative values. Consider the uniform distribution on the interval $[a, b]$:

$$f_X(x) = \frac{1}{b-a} \quad \text{for} \quad a < x < b\,.$$

The differential entropy defined in **Equation C.3** for the uniform distribution evaluates to $\ln(b-a)$, which is negative if $(b-a) < 1$.

An entropy measure that can be negative is hard to interpret, because uncertainty would seem to be non-negative. However, it does make sense for the difference between two distinct uncertainties to be negative. We interpret the differential entropy of a distribution to be the relative entropy between the distribution under consideration and the uniform distribution on the same support.

The uniform distribution for any support is considered to be the most conservative model in terms of entropy for that support. A distribution that is close to the uniform distribution in terms of a statistical distance is more conservative than a distribution that is further away from the uniform distribution. Minimising the distance from the uniform distribution is equivalent to finding the most conservative model in terms of entropy.

Let us select the Kullback-Leibler divergence by **Kullback and Leibler (1951)** for our measure of statistical distance, and suppose we have two probability density functions, f_X and g_X. Then the distance between f_X and g_X is

$$\delta\left(f_X, g_X\right) = \int_A f_X(x) \ln \left\{ \frac{f_X(x)}{g_X(x)} \right\} dx\,, \tag{C.4}$$

where $\delta\left(f_X, g_X\right)$ is the Kullback-Leibler divergence between f_X and g_X.

If g_X is the uniform distribution and A is the interval $[a, b]$, then the Kullback-Leibler divergence $\delta\left(f_X, g_X\right)$ gives the distance of f_X from the uniform distribution. In this case, the Kullback-Leibler divergence $\delta\left(f_X, g_X\right)$ turns out to be

$$\delta\left(f_X, g_X\right) = \ln(b-a) + \int_a^b f_X(x) \ln f_X(x)\, dx\,. \tag{C.5}$$

The differential entropy defined by **Equation C.3** and the Kullback-Leibler divergence in **Equation C.5** differ only by a negative sign and an additive constant. Therefore, differential entropy can be thought of as a special case of Kullback-Leibler divergence. The theory behind Kullback-Leibler divergence helps us to justify the use of differential entropy. Maximising the differential entropy is equivalent to minimising the Kullback-Leibler divergence from the uniform distribution on the same support.

The discussion above gives our justification for finding the most conservative model by solving the following variational problem:

$$\begin{aligned} &\text{Maximise} && I\left[f_X\right] = -\int_A f_X(x) \ln f_X(x)\, dx \,, \\ &\text{Subject to} && m_i = \int_A x^i f_X(x)\, dx \quad \text{for} \quad i = 0,\, 1 \ldots,\, n \,, \end{aligned} \tag{C.6}$$

where the m_is are constants with respect to X.

The solution to **Equation C.6** is the density function f_X with the largest differential entropy that satisfies all of the constraints. This criterion is formally known as the maximum entropy principle. See **Jaynes (1982)** and **Cover and Thomas (2004)** for further detailed discussion of the use of differential entropy and the maximum entropy principle.

D

Appendix D

D.1 Chebyshev Polynomials of the First Kind

D.1.1 Definition

The Chebyshev polynomials of the first kind are defined as the following:

$$T_n(x) = \frac{\left(x - \sqrt{x^2 - 1}\right)^2 + \left(x + \sqrt{x^2 - 1}\right)^2}{2}. \tag{D.1}$$

They can also be defined using trigonometric functions,

$$T_n(x) = \cos\left(n \arccos x\right) \ . \tag{D.2}$$

D.1.2 Recurrence relation

The Chebyshev polynomials of the first kind satisfy the following recurrence relation,

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x) \quad \text{for} \quad x \in [-1, 1] \ . \end{aligned} \tag{D.3}$$

D.1.3 Orthogonality

The Chebyshev polynomials of the first kind T_n form a sequence of orthogonal polynomials with respect to the weight function,

$$\frac{1}{\sqrt{1-x^2}} \,, \tag{D.4}$$

on the interval $[-1, 1]$, that is,

$$\int_0^1 \frac{T_i(x)T_j(x)}{\sqrt{1-x^2}}\,dx = \begin{cases} 0 & \text{for} \quad i \neq j\,, \\ \pi & \text{for} \quad i = j = 0\,, \\ \dfrac{\pi}{2} & \text{for} \quad i = j \neq 0\,, \end{cases} \tag{D.5}$$

where π is the transcendental number $3.141592653589793\ldots$.

Bibliography

Abramov, R. (2009). The multidimensional moment-constrained maximum entropy problem: A BFGS algorithm with constraint scaling. *Journal of Computational Physics* **228**, 96–108.

Abramowitz, M. and Stegun, I. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. Dover publications.

Aczél, J. and Daróczy, Z. (1975). *On measures of information and their characterizations*. Academic Press.

Aczél, J., Forte, B. and Ng, C. (1974). Why the Shannon and Hartley entropies are "natural". *Advances in Applied Probability* **6**, 131–146.

Amindavar, H. and Ritcey, J. (1994). Padé approximations of probability density functions. *Aerospace and Electronic Systems, IEEE Transactions on* **30**, 416–424.

Bandyopadhyay, K., Bhattacharya, A., Biswas, P. and Drabold, D. (2005). Maximum entropy and the problem of moments: A stable algorithm. *Physical Review E* **71**, 057701.

Barthelmann, V., Novak, E. and Ritter, K. (2000). High dimensional polynomial interpolation on sparse grids. *Advances in Computational Mathematics* **12**, 273–288.

Bhaskar, A. and Song, Y. (2011). Closed-form asymptotic sampling distributions under the coalescent with recombination for an arbitrary number of loci. *Arxiv preprint arXiv:1107.4700* .

Biswas, P. and Bhattacharya, A. (2010). Function reconstruction as a classical moment problem: a maximum entropy approach. *Journal of Physics A: Mathematical and Theoretical* **43**, 405003.

Boitard, S. and Loisel, P. (2007). Probability distribution of haplotype frequencies under the two-locus Wright-Fisher model by diffusion approximation. *Theoretical Population Biology* **71**, 380–391.

Borwein, J. and Lewis, A. (1991). On the convergence of moment problems. *Transactions of the American Mathematical Society* **325**, 249–271.

Cherry, J. and Wakeley, J. (2003). A diffusion approximation for selection and drift in a subdivided population. *Genetics* **163**, 421–428.

Clason, C. and von Winckel, G. (2011). A general spectral method for the numerical simulation of one-dimensional interacting fermions. Elsevier.

Cover, T. and Thomas, J. (2004). *Elements of information theory*. Wiley Online Library.

Crow, J. and Kimura, M. (1970). *An introduction to population genetics theory*. Harper & Row, New York.

Etheridge, A. and Lemaire, S. (2011). Diffusion approximation of a multilocus model with assortative mating. *Arxiv preprint arXiv:1101.5485* .

Ethier, S. (1979). A limit theorem for two-locus diffusion models in population genetics. *Journal of Applied Probability* **16**, 402–408.

Ethier, S. and Nagylaki, T. (1980). Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Advances in Applied Probability* **12**, 14–49.

Ethier, S. and Nagylaki, T. (1988). Diffusion approximations of Markov chains with two time scales and applications to population genetics, ii. *Advances in Applied Probability* **20**, 525–545.

Ethier, S. and Nagylaki, T. (1989). Diffusion approximations of the two-locus Wright-Fisher model. *Journal of Mathematical Biology* **27**, 17–28.

Ethier, S. and Norman, M. (1977). Error estimate for the diffusion approximation of the Wright–Fisher model. *Proceedings of the National Academy of Sciences* **74**, 5096–5098.

Ewens, W. (1963a). The diffusion equation and a pseudo-distribution in genetics. *Journal of the Royal Statistical Society. Series B (Methodological)* **25**, 405–412.

Ewens, W. (1963b). The mean time for absorption in a process of genetic type. *J. Austral. Math. Soc* **3**, 375–383.

Ewens, W. (1963c). Numerical results and diffusion approximations in a genetic process. *Biometrika* **50**, 241–249.

Ewens, W. (1964a). The maintenance of alleles by mutation. *Genetics* **50**, 891–898.

Ewens, W. (1964b). The pseudo-transient distribution and its uses in genetics. *Journal of Applied Probability* **1**, 141–156.

Ewens, W. (1965). The adequacy of the diffusion approximation to certain distributions in genetics. *Biometrics* **21**, 386–394.

Ewens, W. (1969). *Population genetics*. Methuen Publishing.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Ewens, W. (2004). *Mathematical population genetics: theoretical introduction*, volume 1. Springer Verlag.

Farouki, R., Goodman, T. and Sauer, T. (2003). Construction of orthogonal bases for polynomials in Bernstein form on triangular and simplex domains. *Computer Aided Geometric Design* **20**, 209–230.

Fisher, R. (1930). The genetical theory of natural selection.

Fisher, R. et al. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **42**, 321–341.

Fisher, R. et al. (1930). The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh* **50**, 205–220.

Fletcher, R. (1987). *Practical methods of optimization, volume 1*. Wiley.

Freund, R. and Zha, H. (1993). A look-ahead algorithm for the solution of general Hankel systems. *Numerische Mathematik* **64**, 295–321.

Genz, A. and Cools, R. (2003). An adaptive numerical cubature algorithm for simplices. *ACM Transactions on Mathematical Software (TOMS)* **29**, 297–308.

Geyer, C. (2008). *trust: Trust Region Optimization*. R package version 2.41-2.8.

Gillan, C., Steinke, T., Bock, J., Borchert, S., Spence, I. and Scott, N. (2012). Comparing the implementation of two-dimensional numerical quadrature on GPU, FPGA and ClearSpeed systems to study electron scattering by atoms. *Concurrency and Computation: Practice and Experience* **24**, 84–95.

Griffiths, R. (1979). A transition density expansion for a multi-allele diffusion model. *Advances in Applied Probability* **11**, 310–325.

Griffiths, R. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology* **17**, 37–50.

Griffiths, R. (1981). Transient distribution of the number of segregating sites in a neutral infinite-sites model with no recombination. *Journal of Applied Probability* **18**, 42–51.

Hildebrand, F. (1987). *Introduction to numerical analysis*. Dover Pubns.

Jaynes, E. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE* **70**, 939–952.

Jenkins, P. and Song, Y. (2009). Closed-form two-locus sampling distributions: accuracy and universality. *Genetics* **183**, 10871103.

Jenkins, P. and Song, Y. (2011). Padé approximants and exact two-locus sampling distributions. *Arxiv preprint arXiv:1107.3897* .

Kapur, J. and Kesavan, H. (1992). *Entropy optimization principles with applications*. Academic Pr.

Kendall, M., Stuart, A. and Ord, K. (1991). *Advanced Theory of Statistics: Classical Inference and Relationship*, volume 2. Oxford University Press.

Kimura, M. (1955a). Random genetic drift in multi-allelic locus. *Evolution* **9**, 419–435.

Kimura, M. (1956). Random genetic drift in a tri-allelic locus; exact solution with a continuous model. *Biometrics* **12**, 57–66.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719.

Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability* **1**, 177–232.

Kimura, M. and Crow, J. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Kimura, M. and Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**, 763–771.

Kimura, M. et al. (1955b). Stochastic processes and distribution of gene frequencies under natural selection. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 20.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

Littler, R. and Good, A. (1978). Fixation times and probabilities for an independent loci model in genetics. *Theoretical Population Biology* **14**, 204–214.

Mathai, A. and Rathie, P. (1975). *Basic concepts in Information Theory and statistics: Axiomatic foundations and applications*. Wiley Eastern.

Mead, L. and Papanicolaou, N. (1984). Maximum entropy in the problem of moments. *Journal of Mathematical Physics* **25**, 2404–2417.

Miller, D. and Liu, W. (2002). On the recovery of joint distributions from limited information. *Journal of Econometrics* **107**, 259–274.

Moran, P. (1962). *The statistical processes of evolutionary theory*, volume 506. Clarendon Press Oxford:.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia Univ Pr.

Nocedal, J. and Wright, S. (1999). *Numerical optimization*. Springer verlag.

Ohta, T. and Kimura, M. (1969a). Linkage disequilibrium due to random genetic drift. *Genet. Res* **13**, 47–55.

Ohta, T. and Kimura, M. (1969b). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63**, 229–238.

Ohta, T. and KIMURA, M. (1970). Development of associative overdominance through linkage disequilibrium in finite populations. *Genet. Res* **16**, 165–177.

Ohta, T. and Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* **68**, 571–580.

Parsons, T., Quince, C. and Plotkin, J. (2010). Some consequences of demographic stochasticity in population genetics. *Genetics* **185**, 1345–1354.

Poland, D. (2000). Maximum-entropy calculation of energy distributions. *The Journal of Chemical Physics* **112**, 6554–6562.

Russell, J. and Fewster, R. (2009). Evaluation of the linkage disequilibrium method for estimating effective population size. *Modeling Demographic Processes in Marked Populations* **3**, 291–320.

Ryland, B. and Munthe-Kaas, H. (2011). On multivariate chebyshev polynomials and spectral approximations on triangles. Springer.

Shannon, C. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* **5**, 3–55.

Shannon, C. and Weaver, W. (1948). A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423.

Shohat, J. and Tamarkin, J. (1943). *The problem of moments.*

Silver, R. and Röder, H. (1997). Calculation of densities of states and spectral functions by Chebyshev recursion and maximum entropy. *Physical Review E* **56**, 4822–4829.

Song, Y. and Song, J. (2007). Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. *Theoretical Population Biology* **71**, 49–60.

Tagliani, A. (1999). Hausdorff moment problem and maximum entropy: a unified approach. *Applied Mathematics and Computation* **105**, 291–305.

Watterson, G. (1962). Some theoretical aspects of diffusion theory in population genetics. *The Annals of Mathematical Statistics* **33**, 939–957.

Watterson, G. (1996). Motoo Kimura's use of diffusion theory in population genetics. *Theoretical Population Biology* **49**, 154–188.

Wheeler, J., Prais, M. and Blumstein, C. (1974). Analysis of spectral densities using modified moments. *Physical Review B* **10**, 2429–2447.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics* **16**, 97–159.

Wright, S. (1937). The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences of the United States of America* **23**, 307–320.

Wright, S. (1945). The differential equation of the distribution of gene frequencies. *Proceedings of the National Academy of Sciences of the United States of America* **31**, 382–389.

Wright, S. (1949). Adaptation and selection. *Genetics, paleontology and evolution* , 365–389.

Wright, S. (1968). *Evolution and the genetics of populations: A treatise in four volumes*. Univ. of Chicago Press.

Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics* **115**, 347–354.