UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
An Analysis of the Distribution of Genotypes for a Recent Model in Population Genetics

Permalink
https://escholarship.org/uc/item/23w4s1jc

Author
Brega, Moorea

Publication Date
2011

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

An Analysis of the Distribution of Genotypes for a Recent Model in Population Genetics

by

Moorea Brega

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kenneth Wachter, Chair
Professor Steven Evans
Professor John Wilmoth

Spring 2011

An Analysis of the Distribution of Genotypes for a Recent Model in Population Genetics

Copyright 2011
by
Moorea Brega

Abstract

An Analysis of the Distribution of Genotypes for a Recent Model in Population Genetics

by

Moorea Brega

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Kenneth Wachter, Chair

Recent genetic studies suggest that many age-related diseases may be attributed not to a single or small number of mutations, but rather to a large number of mutations, each of which is individually slightly deleterious. Following in the tradition of Kimura and Maruyama, we consider mutation accumulation in an infinite population with a large number of mutation types. We compare the distribution of genotypes under two extreme assumptions regarding genetic recombination: no recombination versus "free" recombination, in which recombination acts more rapidly than mutation and selection. Under a range of assumptions, including realistic mutation rates and demographic fitness measures, we find unexpected similarities in the predictions from the different models. While recombination predictably affects the level of mutant alleles present in the population, the overall shape of the genotype distribution under the two models is quite similar, as are the general behavior of demographic outcomes such as lifespan and hazard rates. Furthermore, the distribution of genotypes under the assumption of no recombination may be well approximated by a Poisson random measure. The qualitative similarities in genotype distributions and demographic characters under these extreme models of genetic recombination suggest that attempts to model recombination in a more realistic manner may not add much to our understanding when viewed from a demographic perspective.

The two models analyzed here, developed by Steinsaltz, Evans, and Wachter, are general enough to connect age-specific effects on demographic characters, such as mortality, to mechanisms of genetic change. While the 2005 model without recombination has a series solution, it cannot be directly evaluated except in the simplest of cases. Sampling from the distribution of genotypes in cases with a large number of mutation types is challenging. In this work we utilize a multiple-try Metropolis algorithm to sample from the distribution of genotypes for spaces containing up to 1000 different mutation types. We consider a variety of test cases, finding scenarios in which typical genotypes contain 100, 350 or even 850 mutations. Our success at accurately estimating genotype distributions and demographic outcomes under assumptions that produce such a large average number of mutations sug-

gests that this model could be utilized under more realistic scenarios, such as mutations associated with age-related disease.

"It is interesting to note that the uniqueness of individuals, which delights biologists so much, may be caused by 'littering' the organisms with defects and thus forming a unique pattern of individual damage." Gavrilov and Gavrilova [14]

To our unique patterns of individual damage.

Contents

List of Figures

List of Tables

Acknowledgments

There are many people who should be acknowledged for supporting me throughout this process. I would like to thank Allan Sly and Leon Barrett for their useful conversations regarding MCMC, which helped me to focus my own thoughts on the subject; Alejandro Cantarero, who suggested that I learn python and was a valuable resource (and friend); Bill Graner, who was encouraging and supportive throughout; my parents for their unwavering faith in me; and Elena Potapchuk for providing much needed perspective. I must also acknowledge the National Science Foundation and the Department of Statistics for their financial support during this process. Finally, I would like to thank my committee, Steve Evans, John Wilmoth, and particularly my advisor, Ken Wachter, who provided much valuable and interesting feedback throughout this long process.

Chapter 1

An Overview of Mathematical Models in Population Genetics and Demography

1.1 Introduction

In this work we study a recent model in population genetics proposed by David Steinsaltz, Steven Evans and Ken Wachter [30]. This model is interesting because its broad mathematical framework allows standard mutation-selection balance models, well-known in population genetics, to be connected to more sophisticated models of the effects of genetic mutations on fitness. Of particular interest is the case in which mutations have explicitly defined effects on demographic outcomes, such as hazard rates and survival probabilities. This flexibility makes the model particularly useful in studying various evolutionary theories of senescence, which posit that aging is due to the build-up over evolutionary time of mutations with age-specific effects.

While Steinsaltz, Evans and Wachter presented a series solution to their mutation-selection model in [30], it cannot be directly evaluated except in very simple cases. As a result, not much is known about the distribution of genotypes when demographic outcomes are used to measure genetic fitness. In order to characterize the distribution of genotypes in these circumstances, we utilize Markov chain Monte Carlo methods to sample from the distribution of genotypes. We find that the distribution of genotypes, with realistic mutation rates and demographic outcomes, can be well-approximated by a Poisson random measure. This result is rather surprising as it suggests that the process of mutation under this model is dominant over the process of selection, driving the distribution of genotypes toward a Poisson.

This result is also interesting because of its similarity to a second model proposed by Evans, Steinsaltz and Wachter in [11]. The second model essentially extends the original

mutation-selection model to include recombination. However, recombination is assumed to act on a much faster time scale than either mutation or selection. As a result, recombination ensures that all mutations are statistically independent and guarantees that the resulting distribution of genotypes is a Poisson random measure. Because these two models represents extremes – no recombination versus free recombination – the similarity in distribution of genotypes under the two models suggests that more realistic models of recombination may produce similar outcomes.

In the remainder of this chapter, we will review several evolutionary theories of senscence (§1.2), which guide and motivates this work. In §1.3 we review relevant models in population genetics, including classical mutation-selection balance equations, simple models of recombination and Haldane's principle, as well as discussing some recent models that use more flexible mathematical formulations. We end the chapter with a discussion of common demographic terms, in §1.4, as well as presenting demographic models of theories of senescence.

The mutation-selection model that is the focus of this work is presented in detail in §2.1. In §2.1.7 we discuss using demographic outcomes within the framework of the mutation-selection model. Because we will be comparing the distribution of genotypes for the Steinsaltz, Evans and Wachter (SEW) mutation-selection model to that in which recombination is present, we describe the free recombination model in §2.2, focusing in particular on using demographic outcomes.

Chapter 3 reviews the numerical methods utilized in this work. A simple method for numerically computing the solution to the Evans, Steinsaltz and Wachter (ESW) free recombination model is reviewed in §3.1. Several numeric approaches used to estimate the distribution of genotypes under the SEW mutation-selection model are detailed in §3.2. To illustrate the difficulty of numerically sampling from the distribution of genotypes we present several simple test cases in §3.3.

Chapter 4 contains the bulk of our results. In this chapter we focus on cases inspired by the evolutionary theories of senescence presented in §1.2. Specifically, we look at cases with realistic mutation rates and hundreds of possible types of genetic mutations, with effects ranging from large early-age effects to very small early-age effects. Most of these cases assume that the age-specific effects are described by gamma distributions. We focus on approximating the distribution of genotypes in these cases by a Poisson random measure in §4.1.1 and also explore the similarity of the distribution of genotypes under the SEW mutation-selection model and the ESW free recombination model. We briefly consider a less realistic model of age-specific mutation effects in §4.2. In this case, too, the Poisson approximation appears to be valid. Concluding comments are made in §4.3.

1.2 Evolutionary Theories of Senescence

In 1952, Peter Medawar argued that the force of natural selection decreases with adult age. Under this assumption, deleterious mutations with late-acting effects face less selective pressure than those with early-acting effects. The idea of age-dependent selection force formed the basis for several theories of the evolution of senescence, specifically the theories of mutation accumulation [25], antagonistic pleiotropy [38] and reinforcing (or positive) pleiotropy.

In mutation accumulation, proposed by Peter Medawar, deleterious mutations with large overall effects or with significant early-age effects (for example mutations that impact fertility) will face a high selective pressure and will eventually be weeded out of the population. Because the force of selection decreases with age, mutations that are either slightly deleterious or have late-acting effects will face less selective pressure and are more likely to be passed to the next generation. Under this theory, senescence is the result of an accumulation over many generations of a large number of slightly deleterious mutations or deleterious mutations with late-age (e.g. post-reproductive) effects.

Whereas mutation accumulation focuses on deleterious mutations, George C. Williams considered pleiotropic mutations, that is, mutations in which a single gene corresponds to several phenotypic traits. Introduced in 1957 [38], antagonistic pleiotropy is the theory that a single mutation may have both deleterious and beneficial effects. The force of selection on such a mutation would have to balance its deleterious and beneficial traits. In particular, Williams argued that selection may favor genes that are beneficial early in life even if they have deleterious effects later in life.

In reinforcing (or positive) pleiotropy, deleterious mutations are purely deleterious. However, this theory assumes that any mutation with late-age deleterious effects will also have deleterious effects at younger ages. If the early-age deleterious effects are sufficiently small, the mutation will not face a large selective pressure and mutation-selection balance will ensure that the mutation will not go to extinction. This theory may bypass some potential difficulties with the theory of mutation accumulation as an evolutionary explanation of senescence. In particular, it may avoid the issue of creating a "wall of death." Under the theory of mutation accumulation, mutations with only very late-age effects (i.e. mutations effecting survival at post reproductive ages) may face such low selective pressure that these mutations would accumulate uncontrollably. This could cause late-age mortality to spike and survival to drop to zero, creating a wall of death, an age after which survival is impossible (see [34]). The wall of death phenomenon in the context of the models considered in this work will be discussed briefly in Appendix C.2 and is also discussed in [35].

In all three of these theories for an evolutionary explanation of senescence, age-specific genetic effects are linked with demographic outcomes, such as mortality rates and fertility. As a result, any mathematical model incorporating these theories would need to connect mutation-selection models in population genetics to demographic characters. The models proposed by Evans, Steinsaltz and Wachter ([30] and [11]) are sufficiently general to allow demographic outcomes to be used as a constraint in the spread of deleterious mutations in

a population. These models will be discussed in detail in §2.1 and §2.2.

1.3 Mathematical models for mutation, selection and recombination

This section provides a brief tour through the history of mutation, selection and recombination models used in population genetics. Many of the classic models presented below, as well as quite a few others, are discussed in greater detail in the works of Bürger [6], Durrett[10], and Ewens [12].

1.3.1 Single Locus Models

We begin this review of models in population genetics with a discussion of single locus models. When focusing on a single locus, the genetics of the population can be influenced by mutation (one allele changing to another type of allele), selection (pressure encouraging the spread of beneficial genes and weeding out deleterious genes), and, in small populations, random mating (which can cause the genetics of the population to "drift"). Models will be presented in order of increasing complexity. First, we shall introduce the Hardy-Weinberg law in which random mating alone drives the genetics of the population. We will also briefly discuss the phenomenon of genetic drift. Next we will review classical model in which mutation alone or selection alone drives the change in population genetics. After reviewing how mutation alone and selection alone influence genetics, we turn to a classical mutation-selection balance equation. We also discuss a more complicated model of mutation-selection balance called Kingman's House of Cards. Finally, we end with a discussion of Haldane's principle.

Hardy-Weinberg

Perhaps the beginning of mathematical models in population genetics began with the Hardy-Weinberg law, which considers genetic variation under a Mendelian view of genetics and heredity. To understand this law, we consider a large, randomly-mating monoecious diploid population with discrete, non-overlaping generations. [1] This population consists of individuals with three possible genotypes, A_1A_1, A_1A_2 and A_2A_2, with A_1 and A_2 representing the only possible allelic types at the locus in question. We will denote the proportion of the initial population with these genotypes by X, $2Y$ and Z, respectively. Under random mating, the proportion of the population with these genotypes in the second generation is found using the following probabilities.

[1] In a monoecious population, every individual has both male and female sex organs. One can also consider a dioecious population, that is, a population that has two distinct sexes, as long as the initial genotype frequencies are the same for both sexes. See Bürger [6].

Parent Types	Mating Probability	Conditional Probability of Offspring		
		A_1A_1	A_1A_2	A_2A_2
$A_1A_1 \times A_1A_1$	X^2	1	0	0
$A_1A_1 \times A_1A_2$	$4XY$	$\frac{1}{2}$	$\frac{1}{2}$	0
$A_1A_1 \times A_2A_2$	$2XZ$	0	1	0
$A_1A_2 \times A_1A_2$	$4Y^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$A_1A_2 \times A_2A_2$	$4YZ$	0	$\frac{1}{2}$	$\frac{1}{2}$
$A_2A_2 \times A_2A_2$	Z^2	0	0	1

If we let X' denote the proportion of type A_1A_1 in the second generation (and similarly for $2Y'$ and Z') then we have

$$\begin{aligned} X' &= X^2 + 2XY + Y^2 = (X+Y)^2 \\ 2Y' &= 2XY + 2XZ + 2Y^2 + 2YZ = 2(X+Y)(Y+Z) \\ Z' &= Y^2 + 2YZ + Z^2 = (Y+Z)^2. \end{aligned}$$

This procedure can be repeated again to find the proportions for the third generation, X'', $2Y''$ and Z''. In doing so, we find that $X'' = X'$, $2Y'' = 2Y'$ and $Z'' = Z'$. That is, in a randomly mating population with neither selection nor mutation forces, the population will achieve stable frequencies in the second generation and maintain them for all subsequent generations. The significance of this law is to recognize that a population maintains its genetic variation over time. We note that the Hardy-Weinberg Law can also be extended to k alleles with the same result. A population in Hardy-Weinberg equilibrium is one in which genotypic frequencies obey the equations above (or their extension to k alleles).

Of course, the Hardy-Weinberg equation represents a highly simplified view of genetics. It assumes that all alleles are genetically neutral, conferring neither benefit nor detriment to the individual possessing that genetic material. It also assumes that alleles cannot change type via mutation. Both of these assumptions will be relaxed in the following sections. However, before we leave this simple model where mutation and selection have no influence over the genetics of the population we wish to discuss the importance of population size.

Genetic Drift

While this thesis will focus on large populations, it is important in any survey of mathematical population genetics to discuss finite population models and genetic drift. *Genetic drift* refers to the situation in which random mating alone causes changes in the proportion of the population with a certain allelic type. The earliest and simplest of model of genetic drift is the Wright-Fisher model, which considers a single locus with two possible alleles, A_1 and A_2. The population in question is a diploid population with a constant size, N, and discrete, non-overlapping generations labeled by n, where $n = 0, 1, 2, \ldots$. We let X_n denote

the number of copies of allele A_1 in the population in generation n. Because the population is diploid, there are $2N$ genes in each generation, so the proportion of A_1 alleles in generation n is given by $X_n/2N$.

The number of A_1 alleles in the next generation $(n+1)$ has a binomial distribution with $2N$ trials and probability of success $X_n/2N$. That is,

$$P(X_{n+1} = j | X_n = i) = \pi_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}.$$

The X_n form a Markov chain with transition probabilities π_{ij}. It is easy to show that the expected number of copies of allele A_1 is constant over time, that is,

$$\mathbb{E}[X_{n+1}] = \mathbb{E}[X_n] = \cdots = \mathbb{E}[X_0].$$

However, while the average frequency may be constant in time, it is important to note that the states 0 and $2N$ are absorbing states of the Markov chain. Without mutation to reintroduce a lost allele, once one of the allele types is lost to the population, it is lost for every subsequent generation. Eventually, variation in the population will be lost and every member of the population will have the same genotype. That is, all genes in the population will be allele A_1 or all genes will be allele A_2. In a small population, random mating is an important factor in determining the genetic make-up of the group. For example, genetic drift can drive beneficial mutations to extinction or deleterious genes to fixation. The remaining models considered in this work will assume that populations are large enough that genetic drift can safely be ignored.

Mutation

In the basic Wright-Fisher model, once an allele is lost to a generation, it is lost to the population forever. By introducing mutation, we consider the possibility that mutation may change the allelic type of an offspring, either to an allele already present in the population or to an entirely new allele. In this section we will present a basic model of mutation in which there are a finite number of possible allelic types at the locus in question. The second scenario, in which mutation events always produce new allelic types, is used by Kingman in his mutation-selection balance model called Kingman's House of Cards, that will be presented in a later section.

In a simple model where mutation alone acts on the genetics of a population, we assume that all alleles are neutral, that is, no allele produces an advantage or disadvantage to the individual carrying it. We assume that the population is large enough that we can safely ignore genetic drift. For simplicity we further assume that this population has discrete, non-overlapping generations. Suppose that at a gene locus there are k possible allelic types, $A_1, \ldots A_k$, with the frequency of type A_i given by p_i. For $i \neq j$, the probability that an allele of type A_i produces an offspring of type A_j is given by the mutation rate, μ_{ij}. We adopt the

mathematical convention that $\mu_{ii} = 0$ for $i = 1, 2, \ldots k$. That is, a mutation event always produces an allelic type that is different from the parent type. In the next generation the frequency of type A_i is determined by those individuals of type A_i in the current generation that do not experience mutation events and those individuals in the population of a different type that produce offspring of type A_i through mutation. Using p_i' to denote the frequency of A_i in the next generation, we have

$$p_i' = p_i \left(1 - \sum_j \mu_{ij}\right) + \sum_j \mu_{ji} p_j$$

It is easy to show that in such a population there exists a unique equilibrium for allele frequencies if all the mutation rates are positive and that the frequencies converge to this equilibrium at a geometric rate.

Selection

Previously we considered alleles that are genetically neutral and assumed that mutation was the driving force behind the frequencies of allelic types in a population. Now we want to consider alleles that have different fitnesses, that is, they may confer some benefit or be detrimental to the individual carrying said allele. The term "fitness" in this context encompasses fertility (in terms of the average number of offspring) and viability (survival of the offspring to reproductive maturity). We assume that mating is random and that the population is large enough to ignore genetic drift. Generations are discrete and non-overlapping. We begin with the case of a haploid population.

As before, we focus on a single locus with k possible alleles, $A_1, \ldots A_k$, with the frequency of type A_i given by p_i. We let W_i denote the fitness associated with allelic type A_i. Without mutation, the relative frequency of type A_i in the next generation depends only on the fitness of individuals of that type relative to the rest of the population. Using p_i' to denote the allele frequency in the next generation, we have

$$p_i' = p_i \frac{W_i}{\bar{W}}$$

where $\bar{W}$ is the average population fitness, $\bar{W} = \sum p_i W_i$. Using n to denote the generation number, $n = 0, 1, \ldots$, the frequency of allele A_i in generation n is given by

$$p_i(n) = p_i(0) \frac{(W_i)^n}{\sum_j p_j(0)\,(W_j)^n}.$$

If any allele has a higher fitness than all other alleles, say A_1, then it is clear that for $i \neq 1$, $(W_i/W_1)^n \to 0$ as n goes to infinity. As long as the initial frequency of the fittest allele A_1 is positive ($p_1(0) > 0$), the fittest allele will go to fixation in the population and all other alleles will be lost. That is, $p_1(n) \to 1$ as n goes to infinity.

More interesting dynamics can occur in a diploid population. We now use W_{ij} to denote the fitness of an individual with genotype A_iA_j. Note that $W_{ij} = W_{ji}$ because both correspond to the unordered genotype A_iA_j. Let W_i denote the marginal fitness of allelic type A_i. Unlike the haploid case where the fitness of each type is fixed, the marginal fitnesses for the diploid population depend on the relative frequencies for each allelic type,

$$W_i = \sum_j W_{ij} p_j.$$

If we let P_{ij} be the frequency of individuals of type A_iA_j, then the frequency in the next generation is given by

$$P'_{ij} = \frac{W_{ij} P_{ij}}{\bar{W}} = \frac{W_{ij} p_i p_j}{\bar{W}}$$

where

$$\bar{W} = \sum_{i,j} W_{ij} P_{ij} = \sum_{ij} W_{ij} p_i p_j = \sum_i W_i p_i.$$

The frequency of allele A_i after selection is

$$p'_i = p_i \frac{W_i}{\bar{W}}.$$

At this point we should note that the haploid model can be considered as a special case of the diploid model when the fitnesses are multiplicative, that is, $W_{ij} = v_i v_j$ for some constants v_i. Then,

$$W_i = \sum_j W_{ij} p_j = \sum_j v_i v_j p_j = v_i \bar{v}$$

and

$$\bar{W} = \sum_{ij} W_{ij} p_i p_j = \sum_{ij} v_i v_j p_i p_j = \bar{v}^2.$$

Thus, the frequency in the next generation is

$$p'_i = p_i \frac{W_i}{\bar{W}} = p_i \frac{v_i \bar{v}}{\bar{v}^2} = p_i \frac{v_i}{\bar{v}}$$

which is simply the haploid model.

We will now focus on the simplest diploid case: a single locus with only two allelic types, A_1 and A_2. We let p denote the frequency of A_1 and note that $1 - p$ is the frequency of A_2. While there are several convenient mathematical notations to use for the fitness of genotypes in this model (see, for example, [12]), we choose to follow the notation used by Bürger in [6].

Because the frequencies of allelic types depend on the ratio between the fitness of type A_i and the mean fitness, the frequency is unchanged if we multiply the fitnesses by a constant value. As a result, we will use relative rather than absolute fitness for convenience. In all future references, the term "fitness" can be read as "relative fitness."

Let $W_{11} = 1$, $W_{12} = 1 - hs$ and $W_{22} = 1 - s$ be the relative fitnesses of the genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. The parameter h is called the degree of dominance and $s > 0$ is the selection coefficient. The allele A_1 is called dominant if $h = 0$, partially dominant if $0 < h < 1/2$, recessive if $h = 1$ and partially recessive if $1/2 < h < 1$. In the case where $h = 1/2$ there is no dominance by either allele.

The marginal fitnesses and mean fitnesses are computed as before and we find that

$$\begin{aligned} W_1 &= W_{11}p + W_{12}(1-p) = 1 - hs + hsp \\ W_2 &= W_{12}p + W_{22}(1-p) = 1 - s + s(1-h)p \\ \bar{W} &= 1 - s + 2s(1-h)p - s(1-2h)p^2. \end{aligned}$$

The change in frequency of A_1 is given by

$$\Delta p = p' - p = p\frac{W_1}{\bar{W}} - p = \frac{p(1-p)}{\bar{W}}s(1-h-(1-2h)p) = \frac{p(1-p)}{2\bar{W}}\frac{d\bar{W}}{dp}.$$

Clearly $p = 0$ and $p = 1$ are equilibrium points of the system (in which there is only one type of allele present in the population). There can exist at most one further solution, if there is some $0 < p < 1$ for which $1 - h - (1 - 2h)p = 0$. If such an equilibrium exists, it is

$$p^* = \frac{1-h}{1-2h}.$$

There are three cases to consider for this equilibrium point:

- $0 \leq h \leq 1$: In this case the homozygote A_1A_1 has higher fitness than either the heterozygote or the homozygote A_2A_2. The mean fitness is an increasing function of p for $p \in [0, 1]$. As a result, $\Delta p > 0$ for all $p \in (0, 1)$ and the only equilibria are at $p = 0$ and $p = 1$. Assuming $p(0) \neq 0$, we will have $p(n) \to 1$ as n goes to infinity. That is, the allele A_1 becomes fixed in the population and A_2 is lost. The rate at which A_1 goes to fixation depends on whether A_1 is dominant, intermediate or recessive. The intuition that an initially rare advantageous dominant allele will go to fixation faster than an initially rare recessive allele is supported by the mathematical description of this model.

- $h < 0$: This case is called overdominance because the heterozygote in this scenario has an advantage over both homozygotes. Here, $\bar{W}(p)$ is a concave function of p with a local maximum at the point p^*. Because $\Delta p < 0$ for $p < p^*$ and $\Delta p > 0$ for $p > p^*$,

the equilibrium is stable. Furthermore, because p^* is a globally asymptotically stable point, p converges to p^* monotonically. Unlike the first case, this scenario allows for a polymorphic population.

- $h > 1$: This case is called underdominance because the heterozygote is less fit than both the homozygotes. Here, the mean fitness function $\bar{W}(p)$ has a local minimum at the equilibrium point p^*. As a result, this equilibrium is unstable. The limit of $p(n)$ then depends on the initial state. Specifically, if $p(0) < p^*$ then $p(n) \to 0$ as n goes to infinity while if $p(0) > p^*$ we have $p(n) \to 1$ as n goes to infinity. In both cases one allelic type is lost and the population becomes monomorphic.

Mutation-Selection Balance

Now that we have reviewed basic mathematical models in which mutation alone or selection alone drives the genetics of a population, we want to model a scenario in which both mutation and selection are present. As we saw in the haploid models, selection can drive all but the fittest allele to extinction. Mutation can balance the force of selection by continually reintroducing less fit alleles to the population. As before, we suppose that the population is haploid and has discrete, non-overlaping generations. For this simple model, we assume that selection precedes mutation, that is, individuals are first selected and then reproduce, at which point their offspring may experience a mutation. As with the mutation-only model discussed previously, we assume there are k possible allelic types, labeled $A_1, \dots A_k$. A mutation event results in an allele changing from type A_i to type A_j where $i \neq j$.

The frequency of individuals of type A_i in the next generation is determined by type A_i individuals who survive to reproduce offspring of the same type and individuals of a different type who survive but produce offspring of type A_i,

$$p_i' = \frac{p_i W_i}{\bar{W}} + \sum_j \left(\frac{p_j W_j \mu_{ji}}{\bar{W}} - \frac{p_i W_i \mu_{ij}}{\bar{W}} \right). \tag{1.1}$$

This series of equations can be written more conveniently in matrix form. The mutation matrix $\tilde{U} = (\tilde{u}_{ij})$ contains the mutation rates,

$$\tilde{u}_{ij} = \begin{cases} 1 - \sum_k \mu_{ik} & i = j \\ \mu_{ji} & i \neq j \end{cases}.$$

Notice that $\tilde{u}_{ij}$ represents the chance that there is no mutation event, resulting in an offspring of the same type as the parent. The mutation-selection matrix $C = (c_{ij})$ is then given by

$$c_{ij} = \tilde{u}_{ij} W_j.$$

The frequency of each allelic type in the next generation is governed by the mutation-selection equation (1.1), which we now write in terms of the mutation-selection matrix C,

$$\mathbf{p}' = \frac{1}{\bar{c}} C\mathbf{p}$$

where $\bar{c} = \sum_i (C\mathbf{p})_i = \bar{W}$ and the frequencies are constrained to sum to one, $\sum_i p_i = 1$. Under certain conditions (specifically that C is a primitive matrix), this equation has a unique equilibrium which is globally asymptotically stable. Furthermore, this equilibrium produces a fully polymorphic population, $p_i > 0$ for all i.

In a diploid population, the life cycle again begins with selection. After selection individuals produce (haploid) germ cells, that may undergo mutation. These haploid cells then combine to form zygotes for the next generation. Allele frequencies are measured for zygotes before selection occurs. Equation (1.1) continues to describe the mutation-selection dynamics for the germ cells in a diploid population excepting that W_i now represents the *marginal* fitness of allele A_i.

The dynamics of the mutation-selection equation in the diploid case can be quite complicated. There is an obvious extension of the haploid result in the case where the fitnesses of diploid individuals are multiplicative. The simplification where $\mu_{ij} = \mu_j$ for $i \neq j$, indicating that the mutation rate does not depend on the type of the parent, is called the house-of-cards model. One variation of the House of Cards model will be described in more detail in the following section.

Kingman's House of Cards

Kingman's House of Cards model [21] considers the time evolution for genetic fitness of a large haploid population with discrete, non-overlapping generations. Mutations occur at a single locus and have a continuum of possible effects on fitness. Although the dynamics of the model appear simple, the limiting distributions of fitness for the population may be quite different, depending on the moments of the fitness of mutant individuals.

Let p_n denote the distribution of fitness of the n^{th} generation. Because fitness is measured relative to the general population, the fitness distribution is modeled as a measure on the unit interval, with 0 indicating the least fit individuals and 1 indicating the most fit. For this model, fitness indicates an individual's ability to reproduce offspring that reach maturity. If selection alone were acting upon the population, the distribution of fitness in the next generation would be skewed relative to the current population's fitness distribution, reflecting the fact that individuals who are more fit will have more surviving offspring. Because the population is haploid, an offspring is genetically identical to the parent. Then,

$$p_{n+1}(dx) = \frac{x p_n(dx)}{w_n} \quad \text{where} \quad w_n = \int x p_n(dx).$$

After selection, an individual experiences a mutation from the type of the parent with probability β. The fitness of an individual who has experienced a mutation event is dis-

tributed according to q, where q is a measure on the unit interval. With both mutation and selection acting on the population, the fitness in the next generation is given by

$$p_{n+1}(dx) = (1-\beta)\frac{xp_n(dx)}{w_n} + \beta q(dx).$$

The limiting distribution for the fitness of this population depends on the sum of the moments of q. If the sum of the moments of q are strictly greater than β^{-1}, the limiting distribution is the fitness distribution for individuals with mutant alleles, q, skewed toward individuals with higher fitness

$$p(dx) = \frac{\beta s q(dx)}{s-(1-\beta)x}$$

where $s > 1-\beta$.[2] This case is called "democratic." If the sum of the moments of q are less than or equal to β^{-1}, the limiting distribution is given by

$$p(dx) = \frac{\beta q(dx)}{1-x} + \left(1 - \int \frac{\beta q(dy)}{1-y}\right)\delta_1(dx)$$

where $\delta_1(dx)$ indicates a point-mass at $x = 1$. This second scenario leads to two different situations, the meritocratic case and the aristocratic case. Both cases have the same form for the limiting distribution, although the interpretation for the two cases is quite different. The particular criteria for determining which of the two cases is applicable depends in a slightly complicated way on the moments of q[3]. However, the difference in the genetic composition of the population for the two cases is quite simple to understand. In the meritocratic case, the point-mass at $x = 1$ consists of a group of mutant individuals whose fitness is higher than all other individuals in the population. In the aristocracy case, on the other hand, the upper bound of the fitness in the initial population was higher than the upper bound of support for the mutant fitness distribution. As a result, there were individuals in the initial population with fitness higher than any possible mutant. The point-mass at $x = 1$ represents the non-mutant descendants of those inherently more-fit individuals.

Continuous Time Models

All of the models presented thus far assume that the population in question has discrete, non-overlapping generations. It is sometimes convenient, however, to allow overlapping generations and, thus, use a continuous rather than discrete time model. Unlike discrete time models, where we often assume that selection precedes mutation in acting on every individual in the population, continuous time models often decouple selection and mutation.

[2]The parameter s satisfies the equation $\int \frac{\beta x q(dx)}{s-(1-\beta)x} = 1$.

[3]Specifically, it depends on the limit of the quotient of terms in a renewal sequence dependent on the moments of q.

Specifically, if you consider a small period of time, Δt, the chance of both selection and mutation acting on an individual in that time is of order $(\Delta t)^2$ and thus negligible. The familiar mutation-selection equation (1.1) now becomes

$$\frac{dp_i}{dt} = p_i(r_i - \bar{r}) + \sum_j (p_j \mu_{ji} - p_i \mu_{ij}). \tag{1.2}$$

The parameter r_i is the marginal Malthusian fitness of allele A_i. The marignal Malthusian fitness is the intrinsic growth rate of the subpopulation with allele A_i. It is related to the fitness W by $W_i = e^{r_i}$.

Haldane's Principle

In a 1937 paper Haldane analyzed the decrease in mean fitness of a population due to recurrent deleterious mutations. Studying large populations (so that genetic drift may safely be ignored), Haldane noticed that "the loss of fitness to the species depends entirely on the mutation rate and not at all on the effect of the gene upon the fitness of the individual carrying it, provided this is large enough to keep the gene rare" [17]. This observation is known as Haldane's principle.

The genetic load of the population is a common mathematical tool for measuring the loss of fitness to a population. Specifically, the genetic load of the population is (usually) defined as the relative difference between the fitness of the best genotype $W_{\max}$ and the average fitness of the population $\bar{W}$. For discrete time models (those with discrete generations), the genetic load is given by

$$L = \frac{W_{\max} - \bar{W}}{W_{\max}}.$$

In continuous time models the genetic fitness is the difference between the Malthusian fitness (the intrinsic growth rate) of the subpopulation with the best genotype $m_{\max}$ and the average Malthusian fitness for the entire population $\bar{m}$,

$$L = m_{\max} - \bar{m}.$$

Bürger in [6] provides a nice analysis of genetic load for several different mutation-selection models, which will be summarized here. We begin by considering the classical mutation-selection model with one site and finitely many alleles. For discrete generations the model is formulated in terms of W_i, the fitness of allele A_i; $p_i(n)$, the relative frequency of allele A_i at time n; and μ_{ij}, the mutation rate from allele A_i to A_j (which models the probability that an individual of type A_i will produce an offspring of type A_j). The change in allele frequency over time is given by the system of dynamical equations

$$p_i' = \frac{p_i W_i}{\bar{W}} + \sum_j \left(\frac{p_j W_j \mu_{ji}}{\bar{W}} - \frac{p_i W_i \mu_{ij}}{\bar{W}} \right).$$

The continuous time model is similar to its discrete time counterpart excepting the use of the Malthusian fitness m_i in place of the fitness W_i for allele A_i.

Bürger [6] shows that under many different mutation schemes, the genetic load for classical mutation-selection models is, to first order, equal to the total mutation rate from the fittest allele to less fit alleles. This result holds as long as the equilibrium point of the system without mutation is regular and externally stable. Furthermore, if the difference in fitness between genotypes is of $O(s)$ then the error in the approximation is of order $O(\mu^2/s)$, where μ is the total mutation rate and s is the selection coefficient. These results hold for both haploid and diploid populations and for discrete or continuous time (see Chapter 3 of [6] for details).

Bürger also discusses the genetic load for a more general formulation of the mutation-selection model. This model, which employs a broader mathematical framework than those mutation-selection models discussed thus far, will be presented in the following section. However, because of the flexibility in mathematically describing mutations and genotypes, the general mutation-selection model includes as special cases many classical mutation-selection models, such as the continuum of alleles model. In the continuum of alleles model, there is a continuum of possible allelic types rather than a finite number. Bürger is able to show using the more flexible model that Haldane's principle can be extended to the continuum of alleles model as follows. To first order in μ: For any number of discrete optimal types, the equilibrium mean fitness is independent of the fitness function and the mutation distribution; the mutational load is $L = \mu$.

1.3.2 Multiple Loci Models

The classical mutation and selection models reviewed so far all focus on alleles at a single locus. However, the evolutionary theories of senescence discussed in §1.2 focus on the accumulation of mutations along lineages. To incorporate mutation accumulation into models of population genetics, it will be necessary to consider mutations occurring at many different loci. This section will focus on model of mutation, selection and recombination when alleles can occur at multiple loci.

Because selection acts on the fitness of individuals in a population, it is necessary to describe the interaction of mutations at different loci to determine the fitness of an individual. There are two primary categories of interactions between mutations at different loci: non-epistatic and epistatic. Non-epistatic fitness means that there is no interaction between loci. In this case, the fitness of an individual is additive; it can be expressed as a sum of single-mutation fitness coefficients. In cases where the fitness is epistatic, the model must specify how the loci interact to create the overall fitness of the individual.

Bürger's Mutation-Selection Model

As previously mentioned, Bürger presents a more general formulation of the mutation-selection model in [6]. In this more general model, $\mathcal{X}$ is a locally compact space representing possible mutations. The Borel set $Y \subset \mathcal{X}$ represents a genotype whose relative frequency in the population at time t is given by $P_t(Y)$. Following selection, the relative frequency of individuals of type Y, $P_t^s(Y)$, is given by

$$P_t^s(Y) = \frac{1}{\bar{W}(t)} \int_Y W(y) P_t(dy).$$

where $\bar{W}(t)$ represents the mean fitness at time t. Under this more general representation of allelic types (a locally compact set versus a countable set) the mean fitness is computed by

$$\bar{W}(t) = \int_{\mathcal{X}} W(x) P_t(dx).$$

However, while the representation of alleles and genotypes is more general than the models considered previously, the selection-only equation is a very natural extension to the case with a single locus and a finite, discrete number of alleles presented in §1.3.1.

Similarly, the mutation-only equation presented in §1.3.1 has a natural extension when considering a more general representation for possible mutations. When mutation alone acts on the genetics of a population, the change in relative frequency of genotype Y is due to three events. First, there are individuals do not have genotype Y but who produce offspring of genotype Y due to a mutation event. In this case, some allele x mutates to an allele $y \in Y$ resulting in an individual with the genotype Y. Second, there are individuals with genotype Y who produce offspring of type Y. This event occurs when there is no mutation event between the generations. Third, there are individuals with genotype Y who produce offspring with a different genotype. In this last case, some allele $y \in Y$ mutated to another allele so that the resulting genotype is not Y.

These events can be expressed mathematically if we introduce some notation. Let $\mu(x)$ denote the mutation probability of allelic type x. That is, $\mu(x)$ represents the probability that an allele of type x mutates to an allele of another type. Because $\mu(x)$ is a probability we must have $0 \leq \mu(x) \leq 1$. The term $1 - \mu(x)$, then, is the probability that there is no mutation event of type x. The mutation kernel $u(x, y)$ is the probability of a mutation from allelic type x to type y conditioned on a mutation event occurring. As a result, the product $\mu(x)u(x, y)$ represents the fraction of type y individuals that mutated from type x. Finally, in order to integrate over the alleles contained in genotype Y we need to define a σ-finite measure λ. The relative frequency of individuals with genotype Y after mutation, P_t^μ, is given by

$$P_t^\mu(Y) = \int_Y (1 - \mu(y))\, P_t(dy) + \int_Y \left(\int_{\mathcal{X}} \mu(x) u(x, y) P_t(dx) \right) \lambda(dy).$$

We can put these two processes together by assuming that mutation follows selection. In that case, the time evolution of the frequency measure P_t is given by

$$\bar{W}P_{t+1}(Y) = \int_Y (1-\mu(y))W(y)P_t(dy) + \int_Y \left(\int_{\mathcal{X}} W(x)\mu(x)u(x,y)P_t(dx) \right) \lambda(dy).$$

This more general mutation-selection formulation includes the continuum of alleles model, as previously mentioned, the stepwise-mutation model, and Kingman's House of Cards model, among others.

Recombination

All of the models presented so far assume that the genetics of the population, represented by the frequency of allelic types or by the distribution of fitness in the population, are determined by mutation, selection or genetic drift. In studying alleles at multiple loci, recombination becomes a factor as well. Before discussing more complicated models that incorporate recombination as well as mutation and selection, we wish to present a simple model in which recombination alone drives the genetics of the population.

Consider the simple case of a diploid population in which mutations occur at two loci, called site A and site B. At each loci there are several possible alleles, labeled A_i and B_i, respectively. Without genetic recombination, a population which initially has only gametes of the form A_iB_i will never produce gametes A_iB_j where $i \neq j$. That is, without recombination, the two loci are completely statistically dependent; knowing the type at locus A tells you the type at locus B. When two (or more) loci are statistically dependent, we say they are in linkage disequilibrium.

Let P_{ij} denote the frequency of gamete type A_iB_j. We let p_i be the frequency of allele A_i, so that $p_i = \sum_j P_{ij}$ and similarly define q_j to be the frequency of allele B_j, with $q_j = \sum_i P_{ij}$. Linkage equilibrium occurs when the loci are statistically independent, so that for every i and j we have

$$P_{ij} = p_i q_j.$$

Linkage disequilibrium is usually measured by the differences D_{ij}, where

$$D_{ij} = P_{ij} - p_i q_j.$$

The parameter r denotes the recombination rate, which is the probability of a cross-over recombination event. In general, r depends on the distance between the loci along the chromosome, with more distant loci having a higher recombination rate than closer loci. The parameter r satisfies $0 \leq r \leq 1/2$, where $r = 0$ is the case where the loci are completely linked and can be treated as a single locus and $r = 1/2$ indicates that the loci are completely

unlinked. We can now determine the gamete frequencies in the next generation under the assumption of random mating in the population.

If no cross-over event occurs, a parent of type A_iB_j/A_kB_l will produce a gamete of type A_iB_j or A_kB_l. Because the probability of no recombination event is $1-r$, each of these gametes is produced with probability $\frac{1}{2}(1-r)$. If a cross-over event does occur, which happens with probability r, the possible gametes produced are A_iB_l and A_kB_j, each of which has probability $\frac{1}{2}r$. As a result,

$$P'_{ij} = (1-r)P_{ij} + rp_iq_j.$$

Notice that

$$p'_i = \sum_j P'_{ij} = \sum_j (1-r)P_{ij} + \sum_j rp_iq_j = (1-r)p_i + rp_i = p_i$$

and similarly, $q'_j = q_j$. Thus, in a randomly mating diploid population with recombination, *gamete* frequencies may change but *allele* frequencies remain constant.

The linkage disequilibria in the next generation are given by

$$D'_{ij} = P'_{ij} - p_iq_j = (1-r)P_{ij} + rp_iq_j - p_iq_j = (1-r)D_{ij}.$$

Clearly, then, the linkage disequilibria in generation n depend only on the recombination rate r and the linkage disequilibria in the initial population,

$$D_{ij}(n) = (1-r)^n D_{ij}(0).$$

For any non-zero recombination rate, as n goes to infinity, the linkage disequilibria go to zero geometrically with rate $1-r$. Any model with multiple loci that incorporates recombination will include a discussion of how quickly the statical dependence between loci is broken through the process of recombination. In models with more than two loci, it is necessary to analyze the linkage disequilibria between all groups of at least two loci.

Mutation Counting

Kimura and Maruyama [20] were among the first to consider the mutation counting model, in which an individual's fitness depends only on the number of mutant alleles in the genome. Mathematically it convenient to think of the mutations as copies of one type of mutation because they all have the same effect on fitness, even though from a biological perspective these mutations may be different because they occur at different locations in the genome.

In [20] Kimura and Maruyama consider a very large population with epistatic fitness, that is, mutant alleles at different loci interact in a non-additive way when determining fitness. Specifically, the authors assume that fitness is a quadratic function of the number of mutant alleles i,

$$w_i = 1 - h_1 i - h_2 i^2,$$

where w_i is the fitness of an individual with i mutations and h_1 and h_2 are non-negative constants. To avoid negative fitness, Kimura and Maruyama assume $w_i = 0$ for $i \geq n$, where n is the smallest integer for which the quadratic function is negative. The authors study the effect of quadratic fitness on mutation load, which is the same as the genetic load of the population, discussed previously in §1.3.1. Because the highest possible fitness in the population is 1, the authors define mutational load as the difference between 1 and the average fitness of the population. The mutational load is approximated for three different populations: a randomly mating diploid population with free recombination among the genes, a randomly mating population with no cross-over events, and a haploid asexually reproducing population. The authors find that quadratic fitness has very different effects on mutational load for the three populations.

When the population undergoes free recombination, the number of mutations per genome is Poisson distributed with rate λ, where λ is the average number of mutations per genotype before selection. When the gene frequency is low, the change in gene frequency over time may be estimated by

$$\frac{dp}{dt} = u - hp$$

where p is the gene frequency, u is the mutation rate and h is the average selection coefficient against the mutant gene. The average selection coefficient is given by

$$-h = \frac{\sum_i f_i(w_{i+1} - w_i)}{\sum_i f_i w_i}$$

where f_i is the frequency of individuals with i mutations before selection. Using the quadratic fitness function the average selection coefficient can be approximated by

$$h = \frac{h_1 + h_2 + 2h_2\lambda}{1 - (h_1 + h_2)\lambda - h_2\lambda^2}.$$

The mutational load can then be approximated by the expression

$$L = (h_1 + h_2)\lambda + h_2\lambda^2.$$

The approximation is reasonable as long as the average selection coefficient h is much larger than the mutation rate u. The authors find that under the free recombination assumption, quadratic gene interactions reduce the mutational load by about a half relative to the mutational load for non-epistatic mutations as long as $|h_1| < h_2$. That is, $L \approx \sum u$ (where $\sum u$ is the number of new mutations produced per gamete per generation) when h_1 is small but $L \approx 2\sum u$ when $h_2 = 0$, the case with no epistasis.

The authors also consider two other populations. The first is a diploid population with only one pair of chromosomes in which no cross-over events occur. Assuming random mating, the frequency of diploid individuals with i mutations is found by expanding $\left(\sum_j g_j\right)^2$ where g_j is the frequency of chromosomes with j mutations before selection. The relative selective value, v_i, of chromosomes with i mutations is given by

$$v_i = \sum_j w_{i+j} g_j = 1 - h_1(i + \mu_1') - h_2(i^2 + 2i\mu_1' + \mu_2')$$

where μ_1' and μ_2' are the first and second moments of the distribution of mutations among chromosomes,

$$\mu_1' = \sum_j j g_j \qquad \text{and } \mu_2' = \sum_j j^2 g_j.$$

This population undergoes selection, which works as usual to skew the distribution of genotypes according to the selective value of the chromosome,

$$g_i \to \frac{g_i v_i}{\bar{v}}$$

where $\bar{v}$ is the mean selective value. Mutation follows selection. For simplicity the authors assume that a fixed proportion M of the chromosomes with i mutations gain an additional mutation rather than modeling the number of new mutations by a Poisson distribution. Of course the approximation is reasonable when the mean number of new mutations is small. The change in frequency of chromosomes with i mutations is then given by

$$g_i' = \frac{g_i v_i (1 - M)}{\bar{v}} + \frac{g_{i-1} v_{i-1} M}{\bar{v}}$$

for $i \geq 1$. The authors find that under this model the average number of mutant genes per individual is $\lambda = 2\mu_1'$. The mutational load is given by

$$L = 1 - \bar{w} = 2h_1\mu_1' + 2h_2(\mu_1' + \mu_2').$$

The last population that the authors consider is an asexually reproducing population with no recombination. For this model the authors assume that the number of new mutations follows a Poisson distribution with mean $2M$. Under this assumption, the frequency of individuals with i mutations in the next generation is given by

$$f_i' = \sum_{j=1}^{i} \frac{w_{i-j} f_{i-j}}{\bar{w}} \frac{(2M)^j}{j!} e^{-2M}.$$

At equilibrium the frequency of individuals with i mutations is constant. In particular, it must hold that $f_0' = f_0$. As a result, at equilibrium

$$f_0' = \frac{w_0 f_0}{\bar{w}} e^{-2M} \implies \bar{w} = w_0 e^{-2M}.$$

Because the mutational load is the difference between the fittest genotype (that with no mutations) and the average fitness of the population, it holds that

$$L = 1 - e^{-2M} \approx 2M.$$

This is exactly the mutational load expected with non-epistatic gene interactions.

Barton and Turelli

Barton and Turelli [3] consider a model for analyzing selection and recombination on many loci. This paper is an extension of the authors' previous work on selection on polygenic characters [32]. The model presented in [3] is general enough to model a haploid organism which experiences both viability and sexual selection or a diploid organism that experiences only viability selection. Because our interest at this time is on models of natural selection, we will summarize their model with respect to a diploid organism undergoing viability selection. We note that the notation and equations are the same for the haploid organism, although the interpretations of the variables presented are different.

The model assumes that natural selection works via sex-independent viability selection and that mating among diploids is random. The loci in this model are assumed to be autosomal, meaning that they are not on sex chromosomes. Let X_i denote the state of locus i so that $\mathbf{X} = (X_1, X_2, \cdots X_n)$ represents the genotype of a haploid (germ) cell. In the diallelic case X_i could be an indicator function for one of the two possible alleles. In a case with multiple possible alleles at a single locus, X_i could be a vector where '1' indicates the allele is present and '0' indicates that it is not.

For a diploid organism $\mathbf{X}$ will denote the maternally inherited genotype and $\mathbf{X}^*$ will denote the paternally inherited genotype. In generation t the frequency of the newly formed diploid organism $(\mathbf{X}, \mathbf{X}^*) = (\mathbf{x}, \mathbf{x}^*)$ is given by $f(\mathbf{x}, \mathbf{x}^*) = f(\mathbf{x})f(\mathbf{x}^*)$. $W(\mathbf{x}, \mathbf{x}^*)$ is a fitness measure that determines the viability of the diploid with genotype $(\mathbf{x}, \mathbf{x}^*)$. For the haploid model $W(\mathbf{x}, \mathbf{x}^*)$ can also include sexual selection, fertility interactions between haploids and nonrandom mating. The fraction of offspring produced by $(\mathbf{x}, \mathbf{x}^*)$ is given by $\mathbb{E}[W(\mathbf{X}, \mathbf{X}^*)]$ where the expectation is taken with respect to the zygote frequencies produced by random mating.

The response to selection is determined by the selection coefficients $a_{U,\emptyset}$, $a_{\emptyset,V}$ and $a_{U,V}$ where U is a non-empty collection of possibly repeated indices for the maternally inherited genotype and V is similarly defined for the paternally inherited genotype. These factors are defined by the relation

$$\frac{W}{\bar{W}} = 1 + \sum_U a_{U,\emptyset}(\zeta_U - C_U) + \sum_V a_{\emptyset,V}(\zeta_V^* - C_V) + \sum_{U,V} a_{U,V}(\zeta_U - C_U)(\zeta_V^* - C_V).$$

where

$$\begin{aligned}\zeta_i &= X_i - \mathbb{E}[X_i] = X_i - m_i = X_i - \sum_{\mathbf{x}} f(\mathbf{x}) x_i \\ \zeta_U &= \prod_{i \in U} \zeta_i \text{ and} \\ C_U &= \mathbb{E}[\zeta_U].\end{aligned}$$

The terms ζ_V^* and C_V are defined similarly for the paternally inherited genotype.

Assume selection proceeds recombination. Let $\cdot'$ denote the quantity $\cdot$ after selection and let Δ_s denote the change due to selection. The change in diploid genotype frequency due to selection is given by

$$\begin{aligned}\Delta_s f(\mathbf{x}, \mathbf{x}^*) =& f(\mathbf{x}) f(\mathbf{x}^*) \frac{W(\mathbf{x}, \mathbf{x}^*) - \bar{W}}{\bar{W}} \\ =& f(\mathbf{x}) f(\mathbf{x}^*) \Bigg(\sum_U a_{U,\emptyset} (\zeta_U - C_U) + \sum_V a_{\emptyset,V} (\zeta_V^* - C_V) \\ &+ \sum_{U,V} a_{U,V} (\zeta_U - C_U)(\zeta_V^* - C_V) \Bigg).\end{aligned}$$

The change in mean frequency for locus i is given by

$$\begin{aligned}\Delta_s m_{i,\emptyset} &= \sum_U a_{U,\emptyset} C_{U+i} \\ \Delta_s m_{\emptyset,i} &= \sum_V a_{\emptyset,V} C_{V+i}.\end{aligned}$$

The moments after selection, $C'_{S,T}$, are measured relative to the original means m_i and is given by

$$\Delta_s C_{S,T} \equiv C'_{S,T} - C_{S,T} = \sum_{\mathbf{x},\mathbf{x}^*} \zeta_S \zeta_T^* \Delta_s f(\mathbf{x}, \mathbf{x}^*).$$

Let $\cdot''$ denote the quantity $\cdot$ after both selection and recombination. After recombination, the diploid organism undergoes meiosis and random mating to produce the next generation, $t+1$. Because the diploid population mates randomly it suffices at this point to consider allele frequencies and means for the haploid cells only.

As the simple recombination-only model illustrates, recombination simply shuffles existing genotypes; it does not affect allele frequency. Thus, the mean of X_i for new haploids in the next generation is the average of the frequencies for maternally and paternally inherited genotypes after selection

$$m_i'' = \frac{m'_{i,\emptyset} + m'_{\emptyset,i}}{2}.$$

Denote by $r_{S,T}$ the frequency of recombination events that partition the loci into the sets S and T. Let r_N be the total frequency of recombination rates that break up the set of loci N, so that for non-empty S and T we have

$$r_N = \sum_{S+T=N} r_{S,T}.$$

Then

$$C''_N = \sum_{S+T=N} r_{S,T} \left(\frac{C'_{S,T} + C'_{T,S}}{2} \right) + (1 - r_N) \left(\frac{C'_{N,\emptyset} + C'_{\emptyset,N}}{2} \right).$$

As before, the moments C''_N given above are relative to the original means m_i before selection. Using the expressions for the means after selection and recombination, it is possible to find an expression for the change in central moments between successive generations, ΔC_N, given by

$$\begin{aligned} \Delta C_N &= \sum_{\mathbf{x}} \prod_{i \in N} (x_i - m_i'')(f(\mathbf{x}) + \Delta f(\mathbf{x})) - C_N \\ &= \sum_{\mathbf{x}} \prod_{i \in N} (\zeta_i - \Delta m_i)(f(\mathbf{x}) + \Delta f(\mathbf{x})) - C_N \end{aligned}$$

where the change is means between successive generations is given by

$$\Delta m_i = \frac{\Delta_s m_{i,\emptyset} + \Delta_s m_{\emptyset,i}}{2}$$

and $\Delta f(\mathbf{x})$ denotes the change in the distribution of haploid genotypes between successive generations.

The authors explore their model further using specific examples, including natural selection on diploids with Gaussian stabilizing selection on an additive polygenic trait and sexual selection due to female choice. They also provide a method of computing the linkage disequilibria and discuss approximations for the linkage disequilibria under the assumption of weak selection.

Baake Models

Recent models for mutation, selection and recombination that employ more general mathematical frameworks are those by Ellen Baake [1] (for mutation and single cross-over recombination) and Michael and Ellen Baake [2] (which extends the methods of the earlier model to consider mutation, cross-over recombination and selection). Here, we will focus on the extended model that includes recombination, mutation and selection. In order to provide explicit solutions, the authors assume that fitness is additive, that is, total fitness is a sum of independent fitness contributions from each site.

The authors consider an infinitely large, haploid population and assume that there exists a linearly ordered set of $n+1$ sites or loci at which genetic mutations may occur. Each site i has an associated set of possible mutations, denoted by X_i. This space of mutations is referred to as an alphabet when the space contains only a finite number of alleles. In this model, the space X_i can be either a finite set or a more general set such as a compact subset of $\mathbb{R}$. For mathematical convenience, the allele space for site i, X_i, is assumed to be locally compact. The authors also assume that the genotype space X has a product structure so that $X = X_0 \times X_1 \times \cdots \times X_n$. The set of positive regular Borel measures on X is denoted by $\mathcal{M}_+(X)$ and $\mathcal{P}(X)$ denotes the set of probability measures on X.

The authors define three operators acting on $\mathcal{M}_+(X)$: Φ_{mut}, the mutation operator, Φ_{rec}, the recombination operator, and Φ_{sel}, the selection operator. We begin with a discussion of the mutation operator. The mutation rate matrix is denoted by Q where $Q_{k,l}$ is the rate at which genotype l mutates to genotype k. This matrix has non-negative entries everywhere except the diagonal and is, in fact, a Markov generator. With the product structure assumed above for the genotype space, the mutation generator Q may be written as the sum of Q_i where Q_i is the tensor product of a rate matrix at site i and identity matrices at all other sites. If each of the X_i is finite, containing M_i alleles, then

$$Q_i = \mathbf{1}_{M_0} \otimes \mathbf{1}_{M_1} \cdots \otimes \mathbf{1}_{M_{i-1}} \otimes q_i \otimes \mathbf{1}_{M_{i+1}} \otimes \mathbf{1}_{M_n}$$

where q_i the rate matrix for site i of dimension M_i. For a measure $\omega \in \mathcal{M}_+(X)$, the pure-mutation equation given by

$$\dot{\omega} = \Phi_{\text{mut}}(\omega) := Q\omega = \left(\sum_{i=0}^{n} Q_i \right) \omega.$$

When the initial condition is a probability measure, $\omega_0 \in \mathcal{P}(X)$, the pure-mutation equation has the unique solution

$$\omega_t = \exp\left(t \sum_{i=0}^{n} Q_i \right) \omega_0.$$

In other words, given an initial probability distribution on the space of genotypes, the above

equation describes the probability distribution at time t assuming that the population is subject to mutation alone.

The process of recombination breaks the link between two adjacent sites in a randomly chosen genotype and combines each fragment with the complementary fragment from another randomly chosen genotype that was broken at the same location. Let L denote the set of all links between sites (or loci), with the link between sites i and $i+1$ denoted by the half-integer $i + \frac{1}{2}$. Let $\pi_i : X \to X_i$ be the projection of the genotype onto site i. The mapping $\pi_{i.}\omega$ is defined by $\pi_{i.}\omega(E) = \omega(\pi_i^{-1}(E))$ for any Borel $E \subset X_i$. That is, the mapping $\pi_{i.}\omega$ is the ω-measure of all genotypes whose i^{th} site has a mutation in the set E. The recombination operator is defined in terms of the recombinator R_α where

$$R_\alpha := \frac{1}{||\omega||}((\pi_{<\alpha.}\omega) \otimes (\pi_{>\alpha.}\omega)) \text{ for } \omega \in \mathcal{M}_+(X) \text{ and } \alpha \in L.$$

The abbreviation $\pi_{<\alpha}$ is used to denote $\pi_{\{1,\ldots\lfloor\alpha\rfloor\}}$, the projection onto sites $1 \ldots \lfloor\alpha\rfloor$.

To derive the pure-recombination equation, one can first consider the recombination process applied to a finite population. In the finite population scenario each individual has independent Poisson clocks associated with each link in their genome. The parameter of the Poisson clock at link α is denoted by ϱ_α. Link α breaks when the Poisson clock rings, indicating a recombination event. Taking the limit as the population size goes to infinity gives us the pure-recombination differential equation,

$$\dot{\omega} = \Phi_{\text{rec}}(\omega) := \sum_{\alpha \in L} \varrho_\alpha (R_\alpha - \mathbf{1})(\omega).$$

With initial condition $\omega_0 \in \mathcal{M}_+(X)$, the pure-recombination equation has the unique solution

$$\omega_t = \sum_{G \subset L} a_G(t) R_G(\omega_0)$$

where

$$a_G(t) = \exp\left(-\sum_{\alpha \in \bar{G}} \varrho_\alpha t\right) \cdot \prod_{\beta \in G} (1 - \exp(-\varrho_\beta t))$$

and R_G is the composite recombinator

$$R_G := \prod_{\alpha \in G} R_\alpha.$$

The final process considered is selection. Let P be a bounded linear operator that maps the space of signed finite regular Borel measures on the space of genotypes to itself, $P : \mathcal{M}(X) \to \mathcal{M}(X)$. Furthermore, suppose that P satisfies the following property: For $\nu \in$

$\mathcal{M}_+(X)$ and E a Borel set with $\nu-$measure 0, $\nu(E) = 0$, we have $(P(\nu))(E) \geq 0$. The pure-selection equation is then given by

$$\dot{\omega} = \Phi_{\text{sel}}(\omega) := P\omega - \frac{P\omega(X)}{||\omega||}\omega$$

where $\Phi_{\text{sel}}(\omega) := 0$ for $\omega = 0$. This equation is clearly motivated by the standard selection-only model discussed in §1.3.1 and the interpretation of the linear operator P in this case is much the same as in the standard selection-only model. That is, P describes the genetic fitness of different genotypes. With initial condition $\omega_0 \in \mathcal{M}_+(X)$, the pure-selection equation has a unique solution given by

$$\omega_t = \frac{||\omega_0||}{||\eta_t||}\eta_t$$

where

$$\eta_t = \exp(tP)\omega_0.$$

Putting together the processes of selection, mutation and recombination leads to the nonlinear differential equation

$$\dot{\omega} = \Phi_{\text{mut}}(\omega) + \Phi_{\text{rec}}(\omega) + \Phi_{\text{sel}}(\omega)$$

where ω is a measure on the space of genotypes. Assuming that both the mutation generator, Q, and linear operator P from the selection operator have product structures, there exists a unique solution to the mutation-recombination-selection process. Setting $S = Q + P = \sum_{i=1}^{n} S_i$ and

$$\eta_t = \exp(tS) \sum_{G \subset L} a_G(t) R_G(\eta_0)$$

the mutation-recombination-selection equation has the solution

$$\omega_t = \frac{||\omega_0||}{||\eta_t||}\eta_t$$

where $\omega_0 = \eta_0 \in \mathcal{M}_+^{\otimes}$ is the initial condition.

The authors also provide closed-form expressions for the linkage disequilibria and their evolution in time.

1.4 Demographic Models

The models reviewed above lie in the realm of population genetics – they describe the change in allele or gamete frequencies of a population over time with specific forces (such as

selection, mutation and recombination) acting on individuals in the population. While the fitness of the individual incorporates both fertility of the parent type and viability of the offspring, these models do not explicitly include demographic features, such as lifespan and how fertility changes with age. Because we are interested in studying how the frequencies of mutations with age-specific fitness effects (on either fertility or survival or both) change over time, we must explicitly include demographic outcomes in a population genetics model. Our primary interest in this work is characterizing how deleterious mutations affect lifespan. Individuals whose genomes contain mutations that reduce lifespan may produce, on average, fewer offspring than individuals with a longer lifespan. This will result in a smaller proportion of the population in the following generation who have those deleterious mutations in their genomes. To model this process mathematically it is necessary to first introduce some demographic terms and concepts. We will also present some recent models that incorporate both demographic characters and evolutionary theories of senescence.

1.4.1 Demographic Terms and Background

We begin our discussion of demography by introducing terminology used in models of population growth. The growth of a population is measured by the net reproduction ratio, *NRR*, which is the ratio of the size of the next population to the current population size. The *NRR* is determined by finding the expected number of female offspring borne by a randomly chosen woman in the population,

$$NRR = \int f_x l_x dx.$$

The fertility function, f_x, represents the chance that a woman produces a female offspring (a daughter) in the infinitesimal age range x to $x + dx$. The survivorship function, l_x, describes the chance that a randomly chosen woman from the population survives to age x. Survivorship functions, which are computed in practice as the proportion of the population that survives to age x, refer to cohorts of individuals, that is, individuals born at the same time.

The survivorship function is related to the hazard function, or force of mortality, by the equation

$$l_x = \exp\left(-\int_0^x h_a da\right) = \exp(-H_x)$$

where h_a is the hazard function and H_x is called the cumulative hazard function. The cumulative hazard function is the sum of hazard rates up to age x. Notice that the hazard function can be recovered from the survivorship by taking the negative of the derivative of log survivorship,

$$h_x = -\frac{d}{dx}\log(l_x) = -\frac{1}{l_x}\frac{dl_x}{dx}.$$

1.4.2 Gompertz

The Gompertz model for hazard functions was proposed by Benjamin Gompertz in 1825 to model hazard rates in adult populations. In this model, the hazard function is exponential,

$$h_x = \alpha \exp(\beta x)$$

where α is the initial hazard and β determines how quickly mortality increases with age. In the Gompertz model survivorship is given by

$$l_x = l_0 \exp\left(-\frac{\alpha}{\beta}\left(e^{\beta x} - 1\right)\right)$$

where l_0 is the initial survivorship. Closely related, the Gompertz-Makeham model assumes that the hazard function has both a constant (non-age dependent) and non-constant (age dependent) term,

$$h_x = \lambda + \alpha \exp(\beta x).$$

In practice, the Gompertz model for hazard rates is not valid for younger adult ages or for the oldest ages in the population. Among younger adults, deaths tend to have definitive causes and events like accidents rather than illness drive mortality. For middle aged and older adults (say those 35 to 85 years [39]), Gompertz hazards rate provide reasonable fits to human mortality data [36]. This could be because causes of death become less definitive in older ages, with adults suffering from a variety of ailments and illnesses, all of which may contribute to declining health and eventual death. However, among the oldest segment of the population, such as centenarians and supercentenarians, hazards rates appear to flatten. A discussion of this trend can be found in [34].

Ultimately, the goal of any model that incorporates both elements of population genetics, such as mutation, selection and recombination, and demographic characters, such as lifespan and fertility, is to reproduce realistic population lifespan functions. That is, there should be some conditions under which the expected lifespan from the model follows the general trend observed in adult human populations. Specifically, there should be a range of mid-adult ages for which hazard rates are approximately exponential but where hazard rates plateau or even decrease at the oldest human ages.

1.4.3 Charlesworth

Brian Charlesworth considered the problem of including demographic characters in a mutation-selection model in [7]. This model, which has been very influential to experimentalists in

the field of evolutionary senescence incorporates age-specific fitness effects for deleterious mutations to a mutation-selection equation. Mutations in this model have age-specific effects on survival and fecundity which, in turn, effect the net reproduction of individuals with those mutations in their genome. Assuming that the population size is constant, the decrease in reproduction for these individuals can be used as a measure of fitness to assess the equilibrium proportion of the population with certain genotypes. Charlesworth explored this model further in [8], presenting specific models for mutation age-effects. The notation below is taken from [8].

Charlesworth's model is a continuous time model for a large, diploid, randomly mating population in which wild-type alleles mutate to deleterious alleles at a large number of widely separated sites. An individual with no mutations has a mortality rate at age y of $\mu(y)$ and a reproductive rate of $m(y)$. Mutations are assumed to be heterozygous (non-recessive) and are characterized by an effect age x. Charlesworth presents two possible models for mutation effects. The window effect model assumes that a mutation increases mortality over a limited range or window of ages. In this model, x represents the average age of effect, so that the mutation increases mortality over the ages $(x - \epsilon/2, x + \epsilon/2)$ for some ϵ. The cumulative effect model, on the other hand, assumes that the mutation increases mortality at all ages after the age x. In this model x can be thought of as the age of onset or the age of activation for the mutation.

Because all mutations are assumed to be deleterious, a copy of mutation with age effect x increases mortality. This increase in mortality at age y caused by mutation x is denoted by $\delta\mu(y, x)$. Mutations can also reduce fertility. The reduction in fecundity at age y by mutation x is denoted by $\delta m(y, x)$. $S(y)$ represents sensitivity in the net fitness, denoted by w, to a decrease in mortality at age y. This is determined by the partial derivative

$$S(y) = -\frac{\partial w}{\partial \mu(y)}.$$

Similarly, the sensitivity in net fitness due to an increase in fecundity at age y is denoted by $S'(y)$ and computed by

$$S'(y) = \frac{\partial w}{\partial m(y)}.$$

Fitness is measured by the intrinsic growth rate of the subpopulation with a specific genotype to the growth rate of the subpopulation with the wild-type genotype. Mutations in this model are assumed to have very small effects on mortality and reproduction so that the reduction in net fitness due to a mutation with age effect x can be approximated by integrating over the effects on fitness only for those ages at which the mutation is active,

$$\delta w(x) \approx \int \left(S(y)\delta\mu(y, x) + S'(y)\delta m(y, x)\right) dy.$$

Charlesworth assumes that the balance between the force of selection and the rate at which mutations enter the population at a given locus results in a small equilibrium frequency for that mutation. Under those assumptions (see p. 125-126 in [7]), the equilibrium frequency of non-recessive mutations with effect age x at a given locus is the ratio of the mutation rate at that locus over the reduction in net fitness, $\delta w(x)$. The equilibrium total number of mutations per diploid individual with effect ages in the range x to $x + dx$ is found by summing over all loci,

$$n(x)dx \approx \frac{\nu(x)dx}{\delta w(x)}$$

where $\nu(x)$ in the rate of new mutations acting at age x per individual per age. For simplicity, Charlesworth assumes that $\nu(x)$ is independent of the age x, so that if U is the total diploid mutation rate and d is the longest age to which any individual can survive, $\nu(x) = U/d$. The net increase in mortality at age z relative to the mortality of an individual with the wild-type genotype can be approximated by

$$\Delta\mu(z) \approx \nu \int \frac{\delta\mu(z,x)dx}{\int \left(S(y)\delta\mu(y,x) + S'(y)\delta m(y,x)\right) dy}.$$

A similar formula can be found for the net decline in fertility at age z by replacing $\delta\mu(z,x)$ with $\delta m(z,x)$. Charlesworth also provides an approximation to the additive genetic covariance between the mortality rate at age z and at age z', denoted by $C_A(\mu(x), \mu(z'))$.

In [8] Charlesworth assumes that reproduction begins at age b and continues at a constant rate until the end of life. Under this model, γ represents the extrinsic rate of mortality, which is assumed to have a larger effect than genetic mutations on mortality. In this scenario, $S(y) = \exp(-\gamma(y - b))$ for $y > b$ and $S(y) = 1$ for $y \leq b$. Using the window model and assuming that the effect of the mutation x on mortality at age y depends only on the difference $|y - x|$, Charlesworth finds that the increase in mortality at age z is exponential after a lag of size ϵ from the age of reproduction,

$$\Delta\mu(z) = \nu \exp(\gamma(z - b))$$

for $z > b + \epsilon$. The net mean rate of mortality is approximately given by the Gompertz-Makeham form

$$\mu(z) = \gamma + \nu + \nu \exp(\gamma(z - b)).$$

Charlesworth also considers the cumulative effect model for mutations under the assumption of a constant extrinsic mortality rate and a constant fertility rate beyond the age of maturity. In this case the effect of the mutation on mortality depends on $y - x$ for ages $y > x$. One possible model suggested is $\mu(y,x) = (\delta\mu)\left(1 - \mathrm{e}^{-k(y-x)}\right)$. Under the cumulative effect model, the net increase in mortality is different from the Gompertz-Makeham form for

ages near the start of reproduction, age b, but closely approaches the Gompertz-Makeham form for late ages.

In an attempt to produce mortality rates that plateau at the oldest ages, Charlesworth also considered a mixed model in which mutations have both age-specific and non-age-specific effects. In this model, the change in mortality rate at age z does approach a limit as $z \to \infty$. As a result, the mean mortality approaches a limit as age goes to infinity rather than continuing to increase, as was the case with previous models where mutations only have age-specific effects.

1.4.4 Gavrilov and Gavrilova

Gavrilov and Gavrilova consider a very different explanation for aging in [14] by applying reliability theory to biological organisms. Reliability theory models the failure time of a component or system of components and has been used to estimate the lifetime of mechanical systems. Gavrilov and Gavrilova show that applying reliability theory to a biological system (organism) can result in the well-known Gompertz form for the mortality rate and can also reproduce the leveling of hazard curves observed in human (and other) populations.

Under reliability theory, a system can be composed of one or more vital components. The system fails when any of the vital components fail. Much of the terminology and ideas of reliability theory are already used in demography. In particular, the survival function $S(x)$ denotes the probability that a component or system fails after time x. The failure rate $\lambda(x)$, also called the hazard rate, has the familiar form

$$\lambda(x) = -\frac{1}{S(x)}\frac{dS(x)}{dx}.$$

In demography the hazard rate is also called the force of mortality. It will be denoted by $\mu(x)$ in the following models of biological systems.

A system can be classified as either aging or non-aging according to its failure rate. When the failure rate is constant over time, $\lambda(x) = \lambda$, the system is non-aging and the lifespan distribution follows the exponential form, $S(x) = S_0 \exp(-\lambda x)$. For biological populations, this lifespan distribution is found in wild populations with high extrinsic mortality rates (that is, deaths caused by accident, disease, predation, etc.). A system that fails more often over time or one in which the failure rate increases with time is called an aging system. In demography, the Gompertz-Makeham form is an example of a failure rate with both a non-aging (constant) term and an aging (non-constant) term. In a biological system, the aging component of the hazard rate would be due mainly to age-related diseases, such as certain cancers and heart disease in human beings.

It is well known in reliability theory that the limiting distribution for the lifespan of a system follows either a Gompertz distribution or a Weibull distribution. That is, when the failure of the system is determined by the failure of the first component, as the number of components grow, the distribution of lifespan for the system converges to one of two

distributions. The Weibull distribution is most commonly observed in mechanical systems whereas the Gompertz distribution is most commonly seen in biological systems.

From a reliability perspective, aging in biological systems could be modeled as an accumulation of damage to the system over time. In a simple organism each component is unique and vital and any damage to a single component will result in the organism's death. Such an organism does not experience aging because damage to the system cannot accumulate over time. For more complex organisms, systems are filled with redundancies that allow a component to be damaged without causing the immediate death of the organism. In this model aging is a natural consequence of having redundant systems that increase the organism's reliability and lifespan. Mortality plateaus can also be explained by redundant systems. In particular, it is possible for enough damage to accumulate to overcome all the redundancies in the system, meaning that the organism cannot sustain any further damage to any component without dying. Similar to the simple organism, survival now depends on the failure of a single, vital component. This would result in a leveling of mortality rates.

Gavrilov and Gavrilova consider three models for biological organisms of varying complexity and show that the models produce different hazard rates. We begin by discussing the simplest of the three models. A system is composed of initially functional elements that are arranged into m blocks. A block fails when every element in the block fails. All of the blocks are assumed to have the same number of elements, denoted by n and all elements are assumed to have a constant rate of failure, called k. Under this simple model a block behaves as an aging system even though all the components of the block are non-aging. In other words, even though the components have constant failure rates, the failure rate for the block is not constant over time. Specifically, the failure rate for a *block*, μ_b, initially follows a power law, $\mu_b(x) \approx nk^n x^{n-1}$, where x denotes time, for small x ($x \ll 1/k$), but eventually approaches the limit $\mu_b(x) \approx k$ when $x \gg 1/k$. If the blocks are arranged in parallel, so that the *system* fails when one of the blocks fails, the failure rate for the system, μ_s, exhibits the same general behavior: it follows a power law for $x \ll 1/k$, $\mu_s(x) \approx mnk^n x^{n-1}$ but eventually plateaus to mk for $x \gg 1/k$.

A slightly more complicated model, which may better describe biological systems, assumes that most components are initially non-functional. Let q denote the probability that a component is initially functional and assume that q is small. Then, the number of initially functional elements in a block follows a truncated Poisson distribution with $\lambda = nq$ and normalizing constant c.[4] Under this model the failure rate for the system follows a Gompertz law for large n and small x, $\mu_s(x) \approx R\exp(\alpha x)$, where $R = cm\lambda k\mathrm{e}^{-k}$ and $\alpha = \lambda k$. For large x, $x \gg 1/k$, failure rate plateaus to $\mu_s(x) \approx mk$.

In the most general model, the probability that an element is initially functional is q where $0 < q \leq 1$. Because q is no longer restricted to be small, the distribution of the number of initially functional elements in a block is a truncated binomial with parameters n and q.[5]

[4]The distribution is truncated on the left because each block requires at least one functional element for the entire system to function and on the right because blocks are assumed to have exactly n elements.

[5]As with the previous case, the distribution is truncated on the left because the system needs at least

Under this model the system follows a binomial law for mortality $\mu_s(x) \approx cmn(qk)^n(x_0 + x)^{n-1}$ where $x_0 = (1-q)/qk$ represents the time it would take for an ideal system (with all elements initially functional) to accumulate the number of defects seen in the non-ideal system (see [13] for more details). When $q < 1$ there is an initial period (whose length depends on q) where the hazard rate increases approximately exponentially. Depending on the value of q, the hazard rate can also follow the binomial law for mortality and the Weibull law (power law) before approaching the upper limit of mk.

1.5 Discussion

The evolutionary theories of senescence reviewed in §1.2 all utilize the idea that the force of natural selection decreases with adult age and posit that genetic mutations have age-specific fitness effects. This suggests that incorporating evolutionary theories of senescence in a mathematical model of mutation and selection requires that genetic fitness be a function of both the genotype and the age of the individual. Many of the mathematical population genetics models reviewed in §1.3.1 and §1.3.2 assume that fitness is a constant factor. Different models use different approaches to represent the number of types of mutations and even the number of loci at which mutations occur. For example, the classical mutation-selection model assumes that there is a single locus with a finite number of different mutations. Kingman's House of Cards model also assumes a single locus but allows uncountably many allele types. In both cases, however, fitness is described by a single number.

Of course, single locus models, such as the classical mutation-selection model and Kingman's House of Cards model, are unsuited to studying evolutionary theories of senescence precisely because they focus on alleles at a single locus. Recent work by geneticists suggests that aging and age-related disease may be caused by many mutations occurring throughout the genome. While there are some age-related diseases, such as age-related macular degeneration, in which a small number of alleles explain most of the genetic risk of the illness, most age-related maladies appear to have many associated alleles, each of which produces a very small increase in the risk of disease when considered individually [22]. The finding that many loci appear to be correlated with common age-related diseases means that we must consider a mathematical model that allows for multiple loci. Furthermore, because these alleles have been studied individually, it is currently unknown if there is some interaction between alleles that explains more of the genetic disease risk or if there are many more loci, each with small individual risk, that have yet to be discovered. As a result, any mathematical model would need to be flexible enough in its formulation to allow for either epistatic or non-epistatic fitness.

In §1.3.2 we reviewed several models that employ abstract spaces to model multiple loci, each of which may have any number of possible alleles. Of particular interest are the models of Bürger, Barton and Turelli, and Baake and Baake. While none of these models explicitly

---

one functioning element per block to function.

defines fitness as having an age-specific component, the models may be general enough to extend the definition of fitness to include age. On the other hand, the Baake and Baake selection-mutation-recombination model explicitly assumes that genetic fitness is additive in order to provide closed form solutions. If we wish to model epistatic fitness, the Baake and Baake model may not be easy to analyze. Barton and Turelli provide a general model for natural selection and recombination but do not include genetic mutation in their model. Without mutation to reintroduce deleterious mutations to the population, there can be no mutation accumulation.

Charlesworth, on the other hand, approaches the problem from a demographic perspective. His model includes both mutation and selection and has explicit age-specific effects on both fertility and mortality. As with the continuous time mutation-selection models, fitness in Charlesworth's model is measured by the change in growth rate of the subpopulation with a specific genotype to the growth rate of the subpopulation with the fittest genotype (in this model, the null genotype). However, as we reviewed in §1.4.1, the growth of the population is determined by the net reproduction ratio, which is a nonlinear function of mortality. Charlesworth's approach uses a linear approximation to estimate the change in *NRR* due to a genetic mutation. While Charlesworth's model is clearly a step in the right direction, even reproducing realistic lifespan functions with the type of mortality plateaus observed in human populations, it leaves room for improvement.

In the next chapter we will review a model of mutation and selection proposed by David Steinsaltz, Steven Evans and Ken Wachter [30]. It employs an abstract mathematical framework along the lines of Bürger, Barton and Turelli, and Baake and Baake. Like Charlesworth, the model can incorporate age-specific effects on mortality. Unlike Charlesworth, however, the model explores the full, nonlinear effects of mutations on genetic fitness. We will also present two functions that will be used to describe age-specific effects of mutations on mortality. One such function, a gamma function, was inspired by the model of Gavrilov and Gavrilova [37]. Gavrilov and Gavrilova take a very different approach to modeling senescence, assuming that biological systems have built-in redundancies. Ageing in their model is a direct result of vital systems having redundant components, meaning that systems can accumulate a certain amount of damage before causing death. While this work will not focus on reliability models of senescence, Wachter, Steinsaltz and Evans discuss the possibility of incorporating such models into the theory of mutation accumulation in [37].

Chapter 2

SEW Models

The goal of this chapter is to review the mutation-selection model proposed by David Steinsaltz, Steven Evans and Ken Wachter in [30] and discuss in detail the open questions regarding the Steinsaltz, Evans and Wachter (SEW) model that this work will address. Like the models proposed by Bürger (see §1.3.1 or Chapter IV of [6]) and Michael and Ellen Baake (see §1.3.2 or [2]), this model employs an abstract but versatile mathematical framework for representing mutations and genotypes. The SEW mutation-selection model allows mutations to occur at (possibly) infinitely many locations and is flexible enough to model mutations with age-specific effects on survival along the lines of Charlesworth [8] (see §1.4.3). Throughout the remainder of this work the SEW mutation-selection model may be referred to as the SEW model, the mutation-selection model or the no recombination model. At the end of the chapter we will also review an extension of the SEW model that includes genetic recombination.

2.1 SEW Mutation-Selection Model

The no recombination model assumes an infinite population of haploid individuals. Mutation and selection occur continuously in time, with mutation accumulating among lineages. As a result, there is no mutation back to the ancestral wild-type genome. The model further assumes that all mutations are deleterious, meaning that the ancestral wild-type genotype is the fittest possible genotype in the population.

The next section will focus on the mathematical framework, discussing in detail how mutations and genotypes are represented mathematically in the SEW model. We will then review the derivation of the model, focusing on the expressions that describe the influence of mutation alone or selection alone on the genetics of the population under this mathematical framework. The processes of mutation and selection are then combined to produce the full SEW model. We will also present the SEW model in the context of the evolutionary theories of senescence, reviewing how age-specific mutation effects on mortality can be represented

in this framework.

2.1.1 Mathematical Framework

As previously mentioned, the SEW model has a very general mathematical framework. Potential mutations are represented as elements of a complete and separable metric space, denoted by $\mathcal{M}$. Each mutation is associated with a description of the fitness effects of that mutant allele. For our purposes, this description will be a function indicating age-specific increases in mortality. In the SEW model mutant alleles are not tied to specific sites or locations in the genome. Two mutations occurring at different locations in the genome but having the same effect (e.g. the same age-specific effects on mortality) are modeled as two copies of the same mutation *type*. As a result, a genotype has the form $\sum \delta_{m_i}$, where δ is the delta function and the $m_i \in \mathcal{M}$ are not necessarily distinct. The space of all possible genotypes is denoted by $\mathcal{G}$. Mathematically, $\mathcal{G}$ is represented by the space of integer-valued boundedly finite Borel measures on $\mathcal{M}$.

An individual randomly chosen from the population at time t has genotype g with probability $P_t(g)$. P_t is a probability measure on the measures in the genotype space $\mathcal{G}$. That is, P_t is the distribution of a random measure. Because P_t is the distribution of a random measure, it has an associated intensity measure, which we denote by ρ_t. Mutations arise in the population from $B \subset \mathcal{M}$ at the rate $\nu(B)$. The mutation rate ν is a boundedly finite Borel measure on the space of mutation types $\mathcal{M}$.

The genetic fitness of an individual depends on the selective cost of the individual's genotype, denoted by $S(g)$. As with other continuous time mutation-selection models (see §1.3.1 for an overview of continuous time models in population genetics), the fitness of a genotype is measured by the growth of the subpopulation with that genotype. That growth rate depends on the selective cost. This will be discussed in more detail when we review the mechanism of selection. Mathematically, S is represented by a continuous function that maps the space of genotypes to the positive reals, $S : \mathcal{G} \to \mathbb{R}^+$, and vanishes only on the null genotype. Here, the null genotype refers to the genotype without any deleterious mutations. The null genotype represents the ancestral wild-type selectively neutral genotype. As a result, the null genotype has the highest fitness of any possible genotype. As mentioned previously, all mutations are deleterious, meaning that adding mutations to a genotype always increases selective cost, $S(g + g') \geq S(g)$ for any genotypes g and g'.

2.1.2 Mutation

Suppose genetic mutation is the only force acting on the genotypes of a haploid population. Additionally, assume that back-mutations (from a mutant allele back to a wild-type allele) are not allowed. Under these assumptions, the genotype of an individual will be identical to the genotype of the parent except possibly for additional mutations in the child's genome. Any change in the proportion of the population with genotype g will be due to individuals

of type g who produce offspring of a different genotype due to a mutation event and to adults of type g' who produce offspring of type g due to a mutation event. Mutations are introduced to the population over time according to a Poisson process with rate $\nu \times \lambda$ where λ is the Lebesgue measure. Because mutation is a Poisson process, the probability of two or more mutation events occurring in an individual in time Δt is at least order $(\Delta t)^2$, which can safely be ignored.

To represent the mechanism of mutation within the mathematical framework discussed above, we employ a P_t-integrable test function that maps the space of genotypes to the reals, $\Phi : \mathcal{G} \to \mathbb{R}$. The term $P_t\Phi$ denotes the P_t-expected value of the test function Φ over the space of potential genotypes,

$$P_t\Phi = \int_{\mathcal{G}} \Phi(g) P_t(dg).$$

The change in P_t-expected value of Φ over time due to mutation alone acting on the genetics of the population is described by the expression

$$\frac{d}{dt} P_t\Phi = P_t \left(\int_{\mathcal{M}} \left(\Phi(\cdot + \delta_m) - \Phi(\cdot) \right) \nu(dm) \right). \tag{2.1}$$

Although the expression describing the change in population genetics due to mutation alone is rather abstract under this general mathematical framework, it can, with an appropriate choice of the space of mutations $\mathcal{M}$ reduce to a classical mutation-only model. In particular, suppose there are a finite number of possible mutations, so that $\mathcal{M} = \{m_1, m_2 \dots m_n\}$. In this case, a genotype may be represented as $g = \sum n_i e_i$, where n_i is the number of copies of mutation i and e_i is the i^{th} coordinate vector. If we let the test function Φ be the indicator function for genotype g, $\Phi(g) = \mathbf{1}_g$, then equation (2.1) simplifies to

$$\frac{d}{dt} P_t(g) = \sum_{i=1}^{N} [P_t(g - e_i) - P_t(g)] \nu(m_i).$$

This is precisely the classical mutation model in which a genotype can contain countably many copies of each of n different mutation types, discussed by Bürger in Section III.1.2 of [6].

It is useful to note that the mutation-only model has a familiar solution in some cases. We will begin, however, by discussing the general solution before proceeding to the special case. Let Π be a Poisson random measure on $\mathcal{M} \times \mathbb{R}^+$ with intensity $\nu \otimes \lambda$ where λ is the Lebesgue measure.[1] The measure Π represents the process of introducing mutations to the

[1]Because Poisson random measures will be discussed throughout this chapter and the next, it may be prudent to review their defining properties. If Π is a Poisson random measure on the space $\mathcal{M}$ with intensity measure ρ then $\Pi(B)$ is a Poisson random variable with rate $\rho(B)$ for $B \subset \mathcal{M}$. In other words, the number of events from the set B has a Poisson distribution. If the subsets B and D are disjoint, $B \cap D = \emptyset$, then $\Pi(B)$ and $\Pi(D)$ are independent Poisson random variables. More details can be found in a standard probability

population over time according to the mutation rate ν. Let Z_t be a finite integer-valued random variable on $\mathcal{M}$ defined by

$$Z_t := \int_{\mathcal{M}\times[0,t]} \delta_m \Pi(d(m,u)).$$

Z_t represents a genotype at time t. Then, the distribution of genotypes at time t is given by $P_t = \mathbb{E}[\Phi(W + Z_t)]$ where W is a random measure on $\mathcal{M}$ with distribution P_0 that is independent of Z_t. In the special case where the initial random measure P_0 is Poisson, the measure at time t is also a Poisson random measure. Because P_t is also a Poisson random measure in the case where P_0 is Poisson, we can characterize the distribution at time t by its intensity measure at time t, ρ_t. The intensity measure at time t is given by

$$\rho_t(B) = \rho_0(B) + t\nu(B)$$

for $B \subset \mathcal{M}$ where ρ_0 is the intensity of the initial random measure P_0.

2.1.3 Selection

Now we will consider the scenario in which selection is the only force acting on genotypes in the population. Because the selective cost of a genotype affects the growth rate of the subpopulation with that genotype, the change in genotype frequency over time is determined by the growth of the subpopulation relative to the growth of the entire population,

$$\frac{d}{dt}P_t(dg) = \frac{d}{dh}\frac{e^{-S(g)h}P_t(dg)}{\int_{\mathcal{G}} e^{-S(g')h}P_t(dg')} = (P_tS - S(g))\,P_t(dg). \tag{2.2}$$

For a general test function, $\Phi : \mathcal{G} \to \mathbb{R}$, this becomes

$$\frac{d}{dt}P_t\Phi = -\int_{\mathcal{G}} \Phi(g)\left[S(g) - \int_{\mathcal{G}} S(g')P_t(dg')\right]P_t(dg). \tag{2.3}$$

As with the mutation-only model, we can connect this selection-only model with those presented in §1.3.1 and §1.3.2 by using an indicator function as our test function, $\Phi(g) = \mathbf{1}_{g=g'}$. With the indicator as our test function we have $P_t\Phi = \int_{\mathcal{G}} \mathbf{1}_{g=g'}P_t(dg) = P_t(dg')$ and, thus,

$$\frac{d}{dt}P_t(dg') = -S(g')P_t(dg') + P_t(dg')\int_{\mathcal{G}} S(g'')P_t(dg'').$$

This reduces exactly to equation (2.2). Recall from §1.3.1 that for continuous time models, the fitness is measured by the Malthusian parameter, r_i, which is the intrinsic growth rate

text such as Kallenberg [19].

for the subpopulation with allele A_i. The classical continuous time pure selection equation is

$$\dot{p}_i = p_i(r_i - \bar{r})$$

where $\bar{r}$ is the mean fitness of the population. The more general selection-only model described by equations (2.2) (and more abstractly by equation (2.3)) is clearly analogous to the classical selection-only model. While r_i represents the rate of (exponential) growth in the population with allele A_i, the selective cost $S(g)$ indicates the decrease in the size of the subpopulation with genotype g relative to the growth of the subpopulation with all wild-type alleles.

When the selective cost function is non-epistatic, so that the cost of a genotype is the sum of the costs of each mutation in the genotype, this general selection-only model can have a familiar solution. In particular, if the initial distribution P_0 is a Poisson measure then the distribution of genotypes at time t will also be a Poisson random measure whose intensity at time t satisfies

$$\rho_t(dm) = \rho_0(dm') - \int_0^t \left(S(\delta_m) - \int_{\mathcal{M}} S(\delta_{m'})\rho_s(dm') \right) \rho_s(dm)ds.$$

When the selective cost function is epistatic, or non-additive, the distribution of P_t will not generally be a Poisson random measure, even if P_0 is a Poisson random measure.

2.1.4 Mutation and Selection

In the combined mutation-selection model, the chance of a mutation and a selection event occurring in a time interval of length Δt is of order $(\Delta t)^2$ and, thus, negligible. As a result, the change in genotype distribution, P_t, over time due to both mutation and selection acting on genotypes in the population is simply the sum of the contribution from mutation acting alone and the contribution from selection acting alone,

$$\begin{aligned}\frac{d}{dt}P_t\Phi =& P_t\left(\int_{\mathcal{M}} \left[\Phi(\cdot + \delta_m) - \Phi(\cdot)\right] \nu(dm)\right) - \int_{\mathcal{G}} \Phi(g)\left[S(g) - \int_{\mathcal{G}} S(g')P_t(dg')\right] P_t(dg) \\ =& P_t\left(\int_{\mathcal{M}} \left(\Phi(\cdot + \delta_m) - \Phi(\cdot)\right) \nu(dm)\right) - P_t(\Phi S) + (P_t\Phi)(P_tS). \end{aligned} \tag{2.4}$$

As before, it is easiest to see the similarity between this model and the mutation-selection models presented previously in §1.3.1 and §1.3.2 when the test function Φ is an indicator function, $\Phi(g) = \mathbf{1}_{g=g'}$. Using an indicator function as the test function, equation (2.4) becomes

$$\frac{d}{dt}P_t(g') = \int_{\mathcal{M}} P_t(g' - \delta_m)\nu(dm) - P_t(g')\int_{\mathcal{M}} \nu(dm) - S(g')P_t(dg') + P_t(dg')\int_{\mathcal{G}} S(g'')P_t(dg'').$$

This equation is clearly analogous to the continuous time mutation-selection model given by equation (1.2), discussed in §1.3.1, as well as Bürger's more general mutation-selection model reviewed in §1.3.2.

2.1.5 Solution to the SEW Mutation-Selection Model

In [30] Steinsaltz, Evans and Wachter show that equation (2.4) has the following solution. Let Π denote a Poisson random measure on $\mathcal{M} \times \mathbb{R}^+$ with intensity measure $\nu \times$ Lebesgue. Then, define

$$X_t := X_0 + \int_{\mathcal{M}\times[0,t]} \delta_m d\Pi(m,u)$$

where X_0 is a random measure with distribution P_0 that is independent of Π. Suppose that there is a positive T such that

$$\mathbb{E}\left[\exp\left(-\int_0^t S(X_u)du\right)S(X_t)\right] < \infty$$

for all $t \in [0,T)$. Then

$$P_t\Phi = \frac{\mathbb{E}\left[\exp\left(-\int_0^t S(X_u)du\right)\Phi(X_t)\right]}{\mathbb{E}\exp\left(-\int_0^t S(X_u)du\right)} \tag{2.5}$$

is the solution to equation (2.4) on $[0,T)$.

More useful for our purposes is the following series expansion, whose proof can also be found in [30]. For the series expansion, we let $P_t\Phi = \tilde{P}_t\Phi/\tilde{P}_t\mathbf{1}$ where $\tilde{P}_t\Phi = \sum_n \tilde{P}_t J_n\Phi$. J_n restricts X_t to the event where there are exactly n mutations in the genome by time t. Let $\tau(1), \tau(2), \cdots \tau(n)$ denote the arrival times of the first n mutations that are laid down according to the Poisson random measure Π. We define $Y_i = X_{\tau(i)}$. Then, we have

$$\tilde{P}_t J_n\Phi = \nu(\mathcal{M})^n e^{-\nu(\mathcal{M})t}\mathbb{E}\left[\frac{H_{t,n}\Phi(Y_n)}{S(Y_1)\cdots S(Y_n)}\right] \tag{2.6}$$

where

$$H_{t,n} = \mathbb{P}\left(\sum Z_j/S(Y_j) < t \middle| Y_1, \ldots Y_n\right)$$

for independent, identically distributed exponential rate one random variables $Z_1, Z_2 \cdots$. If $\sum \nu(\mathcal{M})^n\mathbb{E}[(S(Y_1)\cdots S(Y_n))^{-1}]$ is finite then P_t converges in distribution as t goes to infinity. If the sum is infinite, $P_t J_n$ goes to zero for all n.

2.1.6 Mutation Counting Model

A simple, concrete example of the SEW mutation-selection model may be illuminating. Suppose that every mutation, regardless of location in the genome, has the same effect on selective cost so that they can be considered as copies of the same mutation. Because selective cost determines genetic fitness in the SEW mutation-selection model, this scenario is the familiar set-up for the mutation counting model discussed in §1.3.2. Using the mathematical framework for the SEW model, the mutation space in this scenario contains only one type of mutation, $\mathcal{M} = \{m\}$. A genotype g in this case is simply an integer representing the number of mutations in the genome.

To find the expression for the change in the proportion of the population with n mutations in their genome, we let $\Phi = \mathbf{1}_{g=n}$. With this choice of indicator function as the test function Φ, we have

$$\Phi(g + \delta_m) = \mathbf{1}_{g+\delta_m=n} = \begin{cases} 1 & \text{when } g = n-1 \\ 0 & \text{otherwise} \end{cases}.$$

The SEW mutation-selection model given by equation (2.4) then reduces to

$$\frac{dP_t(n)}{dt} = \nu P_t(n-1) - \nu P_t(n) - S(n)P_t(n) + P_t(n)\sum_j S(j)P_t(j).$$

This simple example is instructive for several reasons. In the first place, it shows that the general, abstract mathematical framework employed in formulating the SEW model is sufficiently flexible to describe classical scenarios that are well-known and widely studied in population genetics. As hinted earlier, the scenario with finitely many mutations can also be modeled by choosing an appropriate form for $\mathcal{M}$. In the second place, the solution to the SEW model reduces to a particularly simple form when there is only one type of mutation. The mutation counting formulation of the SEW model will be useful later when we test several methods for sampling from the distribution of genotypes, as it represents one of the few scenarios in which we can easily and accurately estimate the true distribution directly from the series solution, equation (2.6). We shall now present this solution.

Solution to the Mutation Counting Model

If the population has reached equilibrium so that the genetic make-up of the population is not changing over time, then $\frac{d}{dt}P_t = 0$. At equilibrium, then, the mutation counting form of the SEW model becomes

$$0 = \nu P(n-1) - \nu P(n) - S(n)P(n) + P(n)\sum_j S(j)P(j).$$

Solving this equation recursively we find that

$$P(n) = \frac{\nu P(n-1)}{S(n)} = \frac{\nu^n P(0)}{S(1)\cdots S(n)}.$$

Alternately, we can use the series solution given by equation (2.6). Using $\Phi = \mathbf{1}_{g=n}$ we have $\tilde{P} J_n \Phi = \nu(\mathcal{M})^{-n}$ and

$$P(n) = \frac{\frac{\nu^n}{S(1)\cdots S(n)}}{1 + \sum_j \frac{\nu^j}{S(1)\cdots S(j)}} = \frac{\nu^n P(0)}{S(1)\cdots S(n)}.$$

The second equality comes from that the fact that

$$P(0) = \frac{1}{1 + \sum_j \frac{\nu^j}{S(1)\cdots S(j)}},$$

which can be easily verified by noting that $\sum_n P(n) = 1$.

Solution to the Mutation Counting Model with Non-Epistatic Selective Cost Functions

The solution to the mutation counting model can be even further simplified when the selective cost is assumed to be additive. In the additive or non-epistatic case we have $S(n) = nS(1)$. Then,

$$1 = \sum_{n=0}^{\infty} P(n) = \sum_{n=0}^{\infty} \frac{\nu^n P(0)}{\prod_{i=1}^n S(i)} = \sum_{n=0}^{\infty} \frac{\nu^n P(0)}{\prod_{i=1}^n iS(1)} = P(0) \sum_{n=0}^{\infty} \frac{\nu^n}{n! S(1)^n}.$$

Recall that

$$\sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{\nu}{S(1)}\right)^n = \exp\left(\frac{\nu}{S(1)}\right).$$

Then, $P(0) = \exp\left(-\frac{\nu}{S(1)}\right)$ and

$$P(n) = \exp\left(-\frac{\nu}{S(1)}\right) \frac{1}{n!} \left(\frac{\nu}{S(1)}\right)^n.$$

That is, the stationary distribution in this case is Poisson distributed with rate $-\nu/S(1)$.

Although we will not consider non-epistatic cost functions in this work, we include this case to emphasize the connection between the SEW model and classical mutation-selection models.

2.1.7 Demographic Example

The case that we are interested in exploring in this work is that of demographic selective cost functions. In particular we will model the cost associated with a genotype by assuming that mutations produce an increase in the hazard function relative to the hazard rate for individuals with the null or wild-type genotype. By increasing the hazard rate or force of mortality, the genotype causes reduced survivorship relative to individuals with selectively neutral (wild-type) genotypes. This, in turn, produces a decrease in the growth of the subpopulation with that genotype relative to the growth of the subpopulation with the null genotype. The selective cost of a genotype will be measured by the difference in growth, or the net reproduction ratio, of these two subpopulations.

We follow the common practice of assuming that there is a background (extrinsic) hazard rate which is constant for all ages beyond the age of maturity and is zero before the age of maturity. The age of maturity, α, represents the earliest age of fertility. Using the notation introduced in §1.4.1 we let $h_a(g)$ denote the hazard function at age a of genotype g. The null or wild-type genotype represents the fittest possible genotype in the population and will be denoted by $g = 0$. These individuals are only subject to the background hazard rate, which will be denoted by $h_a(0) \equiv \lambda$ for ages $a > \alpha$. For individuals with the null genotype, the cumulative hazard function, the sum of the hazard rates up to age x, is given by $H_x(0) = \lambda(x - \alpha)$.

A single copy of mutation m increases the cumulative hazard function for an individual by the amount $\eta(m)\kappa(m, x)$. If the genotype is represented by $g = \sum \delta_m$ then the cumulative hazard function for an individual with genotype g is

$$\begin{aligned} H_x(g) &= \lambda(x - \alpha) + \sum_{m \in g} \eta(m)\kappa(m, x) \\ &= H_x(0) + \sum_{m \in g} \eta(m)\kappa(m, x). \end{aligned}$$

The parameter η represents the size of the effect of mutation m. In this work we will assume that the mutation effect is comparable to the background hazard rate λ. Although this assumption is not necessary we employ it to model the theory of senescence that posits that aging is due to an accumulation of slightly deleterious mutations over evolutionary time. The function $\kappa(m, x)$ is the cumulative mutation profile for mutation m at age x. This function describes the age-specific effects of the mutation on the cumulative hazard rate. We will consider several types of cumulative mutation profiles in this work. For example, inspired by Charlesworth's "window effect" model [8], where a mutation causes an increase in mortality rate in a window of ages, we consider point-mass mutations. A point mass mutation is the limit of the "window effect" model as the window of effect ages goes to a single age, m. Because the mutation increases the hazard rate at a single point, the effect on the cumulative hazard function will be modeled by a step function. In this case, the mutation has no effect

on the cumulative hazard function at ages before m and a constant effect on the cumulative hazard at ages above m. We will also discuss cases where mutations have gamma profiles and $\kappa(m, x)$ is the cumulative distribution function of a gamma. For this model, we assume that all mutations have the same gamma rate parameter but different shape parameters. Because the mean of a gamma distribution is the ratio of the shape parameter to the rate parameter, mutations in this case will have different mean effect ages.

As mentioned previously, the selective cost of a genotype will be measured by the difference in growth rates between the subpopulation with that genotype and the subpopulation with the null genotype. The growth of the subpopulation with genotype g is measured by the net reproduction ratio, which depends on the lifespan function and the age-specific fertility rate. Following the same convention used above we denote the baseline survivorship function by $l_x(0)$. The baseline survivorship (or lifespan) function represents the probability that an individual with the null genotype survives to age x. The survivorship is related to the cumulative hazard rate by the follow expression,

$$\begin{aligned} l_x(0) &= \exp(-H_x(0)) \\ &= \exp(-\lambda(x-\alpha)). \end{aligned}$$

An individual with genotype $g = \sum \delta_m$ has the survivorship function $l_x(g)$, which is given by

$$\begin{aligned} l_x(g) &= \exp(-H_x(g)) \\ &= l_x(0) \exp\left(-\sum_{m \in g} \eta(m)\kappa(m, x) \right). \end{aligned}$$

The age-specific fertility rate is denoted by f_x. In this work we assume that mutations increase hazard rates but do not effect fertility. Future work could include creating mutation effect models that depress fertility as well as increasing hazard rates. To further simplify the model we assume that the rate of fertility is constant between the ages of α and β and zero for all other ages. Throughout this work we will set the youngest age of fertility to 15 and the oldest age of fertility to 50 in an attempt to align this simple model with the ages of fertility in humans. We note, however, that this simple model of fertility does not fit the pattern of human reproduction, in which fertility rates peak shortly after reproductive maturity and decline thereafter [7].

Introduced in §1.4.1, the net reproduction ratio (*NRR*) measures the ratio of the size of the next population to the size of the current population. The net reproduction ratio for the subpopulation with the null genotype is found by taking the product of the age-specific fertility and the lifespan function for individuals with the null genotype and integrating over all ages,

$$NRR(0) = \int_0^\infty f_x l_x(0) dx.$$

Similarly, the *NRR* for the subpopulation with genotype g is given by

$$NRR(g) = \int_0^\infty f_x l_x(g) dx.$$

With a demographic selective cost function, the fitness of a genotype is measured by the difference in net reproduction ratios between the subpopulation with genotype g and the subpopulation with the optimal genotype,

$$S(g) = \int_0^\infty f_x l_x(0) dx - \int_0^\infty f_x l_x(g) dx.$$

Because the population consists of individuals with varying hazards (depending on their genotypes), the *NRR* for the entire population depends on the expectation of the survivorship function over genotypes present in the population,

$$NRR = \int_0^\infty f_x \mathbb{E} l_x(G) dx.$$

The difference in the *NRR* produced by a single additional copy of mutation m is the expected marginal selective cost,

$$\begin{aligned}
\mathbb{E}[S(G + \delta_m) - S(G)] &= \int_\alpha^\infty f_x \left(1 - e^{-\eta(x)\kappa(m,x)}\right) \mathbb{E} l_x(G) dx \\
&= \int_\alpha^\infty f_x \left(1 - e^{-\eta(x)\kappa(m,x)}\right) l_x(0) \mathbb{E}\left[\exp\left(-\sum_{m \in G} \eta(m)\kappa(m,x)\right)\right] dx.
\end{aligned}$$

For simplicity we assume that the population is stationary, meaning that the population is not growing over time. In other words, the overall size of the population is constant even though subpopulations may experience growth or decline depending on their genetic fitness. Under the assumption of stationarity, the population *NRR* is equal to one. In practice it will be necessary to rescale the fertility rate to ensure that the population is, indeed, stationary.

2.2 ESW Free Recombination Model

Evans, Steinsaltz and Wachter extended their mutation-selection model reviewed in §2.1.4 to include recombination in [11]. Recall that recombination is the process of creating the genotype of a new individual via a random combination of the genotypes of the parents. Previous approaches, such as those of Barton and Turelli [3], Ellen Baake [1], and Michael and

Ellen Baake [2], model recombination by looking at either a single cross-over event or multiple cross-over events between two randomly chosen genotypes in the population. In studying cross-over events, it is necessary to have some representation for the physical locations of the loci and specifically, their distances from each other on a chromosome. This is because in a cross-over event, a break occurs at some location along the chromosome and the resulting two fragments combine with corresponding fragments from a homologous chromosome. For such models it is necessary to determine how quickly the linkage disequilibria goes to zero, that is, how quickly the loci become statistically independent.

In the free recombination model of Evans, Steinsaltz and Wachter, it is unnecessary to have a physical representation of the loci because the loci are statistically independent. As with the SEW mutation-selection model, mutations can be modeled by their age-specific effects without reference to their location on the chromosome. Furthermore, because the loci are statistically independent, a new genotype is simply a random assortment of alleles present in the population relative to their frequencies in the population at that time. We note, however, that the derivation of the free recombination model does not assume that loci are independent. Rather, the independence of loci is a consequence of the assumptions used to create the model.

Although technical details of the derivation of the ESW free recombination model will not be repeated here, we will provide a brief overview. The free recombination model can be derived by adding recombination to the SEW mutation-selection model under the assumption that selection and mutation act on much slower time scales than recombination and that no part of the genome is immune to recombination. The authors begin by considering a discrete time analog of the SEW mutation-selection model. Genotypes undergo selection followed by mutation followed by recombination. Recombination is modeled in a manner similar to that of Barton and Turelli. Specifically, the set $R \subset \mathcal{M}$ denotes the collection of sites that segregate together during a recombination event. A new genotype is formed by $g(\cdot \cap R) + g'(\cdot \cap R^c)$ where g and g' are the genotypes of two individuals chosen uniformly at random from the population. Taking the limit of the discrete time model as the time between generations goes to zero assuming that recombination acts on a faster time scale than both mutation and selection results in the free recombination model.

In adding recombination, the distribution of genotypes under the ESW free recombination model becomes a Poisson random measure. Because a Poisson random measure is completely characterized by its intensity measure, the free recombination model can be expressed as the change in the intensity measure ρ_t over time,

$$\frac{d\rho_t(dm)}{dt} = \nu(dm) - \rho_t(dm)\mathbb{E}_{\rho_t}\left[S(G + \delta_m) - S(G)\right], \tag{2.7}$$

where G is a genotype randomly chosen from the population at time t. To be more precise, G is a Poisson random measure on the space of mutations $\mathcal{M}$ with intensity measure ρ_t. The intensity at time t can be found by integrating over time,

$$\rho_t(dm) = \rho_0(dm) + t\nu(dm) - \int_0^t \mathbb{E}[S(G+\delta_m) - S(G)]\rho_s(dm)ds. \tag{2.8}$$

Notice that the first part of this solution is the solution to the mutation-only model discussed in §2.1.2.

2.2.1 Formal Description of the ESW Free Recombination Model

We now review the more technical description of the model and its solution. Let $\mathcal{H}^+$ denote the space of finite nonnegative measures on $\mathcal{M}$. That is, $\mathcal{H}^+$ is the space of intensities for Poisson random genotypes. Define $F : \mathcal{M} \times \mathcal{H}^+$ by

$$F_\pi(x) := \mathbb{E}[S(X^\pi + \delta_x) - S(X^\pi)]$$

for $x \in \mathcal{M}$ and $\pi \in \mathcal{H}^+$. Define the operator $D : \mathcal{H}^+ \to \mathcal{H}^+$ by

$$\frac{d(D\pi)}{d\pi}(m) := F_\pi(m)$$

meaning, for any bounded $f : \mathcal{M} \to \mathbb{R}$, we have

$$\int_{\mathcal{M}} f(x)d(D\pi)(x) = \int_{\mathcal{M}} f(x)F_\pi(x)d\pi(x).$$

Assuming that the selective cost function satisfies a Lipschitz condition (see [11] for details), then for any $\rho_0 \in \mathcal{H}^+$,

$$\rho_t = \rho_0 + t\nu - \int_0^t D\rho_s ds$$

is the intensity measure for the Poisson random measure with distribution P_t and this solution is unique.

2.2.2 Demographic Example

Because this work will focus on demographic selective cost functions, we will briefly discuss the ESW free recombination model with the demographic selective cost function introduced in §2.1.7. Recall that for the demographic selective cost function

$$S(g+\delta_m) - S(g) = \int_0^\infty \left(1 - e^{-\eta(m)\kappa(m,x)}\right) f_x l_x(g)dx.$$

Replacing the specific genotype g with a random genotype G and taking the P_t-expected value over all genotypes in the population at time t we have

$$\mathbb{E}_{\rho_t}[S(G+\delta_m) - S(G)] = \int_0^\infty \left(1 - e^{-\eta(m)\kappa(m,x)}\right) f_x \mathbb{E}_{\rho_t}[l_x(G)]dx.$$

Plugging in the demographic selective cost function to equation (2.8), we find that the intensity at time t is given by

$$\rho_t(dm) = \rho_0(dm) + t\nu(dm) - \int_0^t \rho_s(dm) \int_0^\infty \left(1 - e^{-\eta(m)\kappa(m,x)}\right) f_x \mathbb{E}_{\rho_s}[l_x(G)]dxds. \quad (2.9)$$

The aggregate survivorship function for the population is the expectation of the survivorship function taken over all possible genotypes, $\mathbb{E}l_x(G)$. From §2.1.7 we know that

$$\mathbb{E}_{\rho_s}[l_x(G)] = l_x(0)\mathbb{E}_{\rho_s}\left[\exp\left(-\sum_{m\in G}\eta(m)\kappa(m,x)\right)\right].$$

Because the number of copies of each mutation type is Possion distributed in the free recombination model, the expected population survival function has a particularly simple form,

$$\mathbb{E}_{\rho_s}[l_x(G)] = l_x(0)\exp\left(-\int_{\mathcal{M}}\left(1 - \mathrm{e}^{-\eta(m')\kappa(m',x)}\right)\rho_t(dm')\right).$$

Putting everything together we have

$$\begin{aligned}\rho_t(dm) &= \rho_0(dm) + t\nu(dm) \\ &\quad - \int_0^t \rho_s(dm)\left[\int_0^\infty \left(1 - e^{-\eta(m)\kappa(m,x)}\right) f_x l_x(0)\right. \\ &\quad \left.\times \exp\left(-\int_{\mathcal{M}}\left(1 - \mathrm{e}^{-\eta(m')\kappa(m',x)}\right)\rho_t(dm')\right)dx\right]ds. \end{aligned} \quad (2.10)$$

2.3 Discussion

The two model presented in this chapter represent two extremes in modeling genetics: either the genome undergoes no recombination (the SEW mutation-selection model) or recombination is occurring continuously and on a much faster time scale than either mutation or selection (the free recombination model). Both assumptions are unrealistic for modeling much of our DNA. However, studying these two extremes may provide an idea of the spectrum of possible outcomes under these related models. Specifically, we are interested in determining how much demographic outcomes (such as mortality curves and hazard rates) depend on the assumptions of the two models. While we know that the solution to the mutation-selection model is not a Poisson random measure in general, it would be useful to

know if the solution can be well-approximated by a Poisson random measure, which has many convenient properties (such as, in this application, independence in the number of copies of different types of mutations). Determining that both models produce similar demographic outcomes would give us some confidence in applying our results to real-world populations that experience less extreme forms of genetic recombination. Furthermore, if the solution to the mutation-selection model is "close" to that of the free recombination model then the free recombination solution could be used to approximate mutation distributions in cases with more realistic recombination rates. This is important because the free recombination solution is much easier to estimate than the solution for the SEW mutation-selection model.

Although we previously presented a series solution to the SEW model (equation (2.6)), this solution cannot be evaluated directly except in some cases with very simple mutation spaces, such as the mutation counting model of §2.1.6. In particular, the term

$$\mathbb{E}\left[\frac{H_{t,n}\Phi(Y_n)}{S(Y_1)\cdots S(Y_n)}\right]$$

requires evaluating the selective cost function for every genotype with n mutations and every possible *ordering* of mutations over time that would result in that genotype. In addition to the difficulty of evaluating each term of the series, it is also necessary to determine the normalizing constant $\tilde{P}_t\mathbf{1}$.

By contrast, the solution to the free recombination model with demographic selective cost functions can be found using standard methods for numerically solving equation (2.10). Furthermore, recent results have shown that, assuming that the population initially contains only those individuals with wild-type alleles (that is, all individuals have the null genotype), then equation (2.10) has a unique, stable equilibrium point. This means that the intensity function converges to a unique measure. While we also know that the distribution of genotypes under the SEW model converges to a unique solution when the population initially contains only wild-type individuals, the limiting distribution for the mutation-selection case is, as mentioned previously, much harder to evaluate. Comparing the outcomes of the two models, then, will require accurately estimating the series solution for the SEW model. We will discuss details of the numerics in the next chapter.

Chapter 3

Numerical Methods

In this chapter we present the numerical methods used to estimate the solutions to the ESW free recombination model and the SEW mutation-selection model reviewed in the previous chapter. We also provide results for several cases with small mutation spaces to illustrate some of the difficulties of estimating the limiting distribution in the model without recombination. Because the equilibrium solution to the free recombination model is much easier to estimate than that for the SEW mutation-selection model, we begin with our discussion with the free recombination model.

3.1 Shortcut Method for the Free Recombination Model

As mentioned previously, it is possible to estimate the limiting distribution for the free recombination model by numerically solving the system given by (2.9) and integrating over a long period of time. In practice, one would continue to update the intensity ρ_n until it does not appear to change much between subsequent time intervals. However, a much simpler "shortcut" method has been suggested by Steve Evans. At equilibrium, $\frac{d}{dt}\rho_t = 0$, so the equilibrium intensity ρ must satisfy

$$0 = \nu(m) - \rho(m)\mathbb{E}_\rho[S(G+\delta_m) - S(G)].$$

Solving for ρ produces the following update scheme,

$$\rho_{n+1}(m) = \frac{\nu(m)}{\mathbb{E}_{\rho_n}[S(G+\delta_m) - S(G)]}.$$

For demographic selective cost functions, the shortcut method is given by

$$\rho_{n+1}(m) = \frac{\nu(m)}{\int_0^\infty \left(1 - e^{-\eta\kappa(m,x)}\right) f_x l_x(0) e^{-\int_{m'\in\mathcal{M}} \rho_n(m')(1-\exp(-\eta\kappa(m',x)))\mathrm{d}m'} \mathrm{d}x}. \tag{3.1}$$

Evans has shown that the shortcut method converges to the minimal solution when starting with $\rho \equiv 0$.

In the special case where mutations have point-mass profiles (see §2.1.7) the integration can be computed exactly and does not need to be numerically approximated. The update for the shortcut algorithm in this case has a reasonably simple form. First, though, we wish to remind the reader that a point-mass mutation is one that increases the hazard rate at a single age, which we shall call the age of onset, m. Such a mutation increases the cumulative hazard function by a constant value after age m. The cumulative mutation profile κ, which describes the age-specific increase in the cumulative hazard function, for a point-mass profile is modeled by a Heaviside function, $\kappa(m, x) = H(x - m)$. With this mutation profile we have

$$1 - \mathrm{e}^{-\eta H(x-m)} = \begin{cases} 1 - \mathrm{e}^{-\eta} & \text{if } x \geq m \\ 0 & \text{if } x < m \end{cases}.$$

We have assumed above that η, the size of the mutation effect, is the same for all mutation types. Although this assumption is not necessary, we will use it in all cases discussed in this work.

Suppose the mutation space contains k point-mass profile mutations with mutation i having age of onset m_i, where $m_i < m_{i+1}$. Fertility is constant between the age of maturity, α, and the oldest age of reproduction, β. Let $h_n(m_i) = (1 - \exp(-\eta))\rho_n(m_i)$ and recall that $l_x(0) = \exp(-\lambda(x - \alpha))$, where λ is the background or extrinsic hazard rate. Then,

$$\begin{aligned}
h_{n+1}(m_1) =& \nu(m_1) \left(\frac{f}{\lambda} \left[\left(\mathrm{e}^{-\lambda(m_1-\alpha)} - \mathrm{e}^{-\lambda(m_2-\alpha)} \right) \mathrm{e}^{-h_n(m_1)} \right.\right. \\
&+ \left(\mathrm{e}^{-\lambda(m_2-\alpha)} - \mathrm{e}^{-\lambda(m_3-\alpha)} \right) \mathrm{e}^{-h_n(m_1)-h_n(m_2)} \\
&\left.\left. + \cdots + \left(\mathrm{e}^{-\lambda(m_k-\alpha)} - \mathrm{e}^{-\lambda(\beta-\alpha)} \right) \mathrm{e}^{-h_n(m_1)-h_n(m_2)-\cdots-h_n(m_k)} \right] \right)^{-1} \\
h_{n+1}(m_2) =& \nu(m_2) \left(\frac{f}{\lambda} \left[\left(\mathrm{e}^{-\lambda(m_2-\alpha)} - \mathrm{e}^{-\lambda(m_3-\alpha)} \right) \mathrm{e}^{-h_n(m_1)-h_n(m_2)} \right.\right. \\
&\left.\left. + \cdots + \left(\mathrm{e}^{-\lambda(m_k-\alpha)} - \mathrm{e}^{-\lambda(\beta-\alpha)} \right) \mathrm{e}^{-h_n(m_1)-h_n(m_2)-\cdots-h_n(m_k)} \right] \right)^{-1} \\
&\vdots \\
h_{n+1}(m_k) =& \frac{\nu(m_k)}{\frac{f}{\lambda} \left[\left(\mathrm{e}^{-\lambda(m_k-\alpha)} - \mathrm{e}^{-\lambda(\beta-\alpha)} \right) \mathrm{e}^{-h_n(m_1)-h_n(m_2)-\cdots-h_n(m_k)} \right]}.
\end{aligned}$$

For general mutation profiles, however, it is necessary to numerically approximate the integrals in equation (3.1). In the implementation of Evans' shortcut method used in this work, the integrals are approximated using Simpson's rule. Simpson's rule requires an even

number of intervals to evaluate the integral. As a result, we choose appropriate step sizes in age, Dx, to ensure an even number of intervals. We must also ensure that the net reproduction ratio for the population is one, meaning that the population size is fixed over time. We do so by computing the *NRR* after each update of ρ and rescaling the fertility by the factor $1/NRR$. The final rate of fertility from the shortcut method is saved and will be used as the fertility rate for the SEW mutation-selection model tests. In many cases the fertility rate for the free recombination model proves to be very close to the fertility rate needed to ensure a stationary population under the SEW model. The shortcut update is run until the L^2 norm of the difference $\rho_{n+1} - \rho_n$ is smaller than a given tolerance. In all test cases this tolerance was 10^{-6}.

3.2 Numerical Approaches for the Mutation-Selection Model

Because of the difficulty of estimating the equilibrium distribution under the SEW mutation-selection model, we consider three different numerical approaches to this problem. The first and simplest method we employ will act as a baseline for cases with very small mutation spaces. This algorithm is referred to as the naive algorithm. The remaining two approaches both utilize Markov chain Monte Carlo methods.

3.2.1 Naive Algorithm

A naive approach to approximating the equilibrium distribution of mutations, $P(m)$, is to model the Poisson process laying down the mutations (over evolutionary time) according to $\nu \times$ Lebesgue. In particular, fix a largest number of possible mutation events, M, and a number of trials, N. For each trial, record the order in which the mutations arrive, $m^{(1)}, m^{(2)}, \cdots m^{(M)}$. Then, for trial j, let the $Y_i^{(j)}$ represent the genotype with i mutations, (see §2.1.5 for a description of the solution to the SEW model),

$$\begin{aligned} Y_1^{(j)} &= m^{(1)} \\ Y_2^{(j)} &= m^{(1)} + m^{(2)} \\ &\vdots \\ Y_M^{(j)} &= m^{(1)} + m^{(2)} + \cdots + m^{(M)}. \end{aligned}$$

For $n \leq M$, we can approximate the expected value by

$$\mathbb{E}\left[\frac{\Phi(Y_n)}{S(Y_1)\cdots S(Y_n)}\right] \approx \frac{1}{N}\sum_{j=1}^{N}\frac{\Phi\left(Y_n^{(j)}\right)}{S\left(Y_1^{(j)}\right)\cdots S\left(Y_n^{(j)}\right)}.$$

The probability that a genotype contains exactly n mutations, where $n \leq M$, is estimated by

$$\begin{aligned} P\mathbf{1}_{|g|=n} &= \frac{\sum_n \tilde{P}J_n\mathbf{1}_{|g|=n}}{\sum_n \tilde{P}\mathbf{1}} \\ &\approx \frac{\sum_{n=1}^{M}\nu(\mathcal{M})^n\frac{1}{N}\sum_{j=1}^{N}\frac{1}{S\left(Y_1^{(j)}\right)\cdots S\left(Y_n^{(j)}\right)}}{1+\sum_{n=1}^{M}\nu(\mathcal{M})^n\frac{1}{N}\sum_{j=1}^{N}\frac{1}{S\left(Y_1^{(j)}\right)\cdots S\left(Y_n^{(j)}\right)}}. \end{aligned}$$

Although we have used the variables N and M in the above description of the naive algorithm, we will henceforth use NumEv as the largest number of possible mutations (formerly M) and NumTrials as the number of trials or number of genotypes generated by the algorithm (formerly N). These more specific names will help to avoid confusion in later sections when we will be using all three numerical approaches to estimate the distribution under the SEW model.

The naive method is the most intuitive approach and the simplest to implement. As we shall see, it works well for cases with a small number of possible mutations where the mutation rate and fertility are low enough that genotypes with high probability all contain a small number of mutations. The approach is less useful as the space of possible mutations ($\mathcal{M}$) gets larger or in cases with very late-acting mutations and high mutation rates. In particular, by setting the largest possible number of mutation events, NumEv, the user is restricting the algorithm to consider only those genotypes containing NumEv or fewer mutations. If the user chooses a value of NumEv that is too low, the algorithm may be stuck exploring a subset of $\mathcal{G}$ (the space of all possible genotypes) that has low probability. The results of such experiments may be wildly misleading, causing local maxima in an otherwise low probability space to appear more likely than they actually are.

In addition, in spaces with a large number of mutations, many of the possible genotypes may be highly unlikely. With a fixed number of trials it may happen that the most likely genotypes (specifically, the most likely time-ordered genotypes) won't be among the genotypes generated randomly by the Poisson process. Again, this can lead to an inaccurate picture of the distribution of genotypes. Taking a cue from Bayesians, who have long needed algorithms to sample from distributions where the normalizing factor is unknown or difficult to compute, we instead turn our attention to Markov chain Monte Carlo methods [15]. In particular, we will focus on the Metropolis-Hastings algorithm. See [26] for the original Metropolis algorithm and [18] for Hastings' generalization.

3.2.2 Metropolis-Hastings Algorithm

The goal of the Metropolis-Hastings algorithm is to draw samples from a distribution $\pi(x)$. A proposal distribution $Q(x'|x^t)$, which depends on the current state of the chain, x^t, is used to generate a new proposed state x'. This proposed state is either accepted or rejected using the rule

$$x^{t+1} = \begin{cases} x' & \text{with probability } \alpha = \min\left\{\frac{\pi(x')Q(x^t|x')}{\pi(x^t)Q(x'|x^t)}, 1\right\} \\ x^t & \text{with probability } 1-\alpha \end{cases}.$$

The original Metropolis algorithm required that the proposal distribution Q be symmetric. Hastings' generalization of the Metropolis algorithm removed that requirement. Because the distribution π appears only in a proportion, one need only know a function that dominates the distribution. The algorithm is most efficient when $Q(x'|x^t) \approx \pi(x')$.

With the series solution to the SEW model, it is necessary to compute the normalizing constant $\tilde{P}\mathbf{1}$ in order to determine the distribution of genotypes. Without the normalizing constant, the probability that a randomly chosen individual has genotype g is proportional to the numerator $\tilde{P}\mathbf{1}_g$,

$$P(g) \propto \nu(\mathcal{M})^n \mathbb{E}\left[\frac{\mathbf{1}(Y_n = g)}{S(Y_1)\cdots S(Y_n)}\right].$$

The difficulty in using the Metropolis-Hastings algorithm to sample from the distribution of genotypes is in computing this expectation. For example, if $g = \sum_k n_k\delta_k$ and $|g| = N$ then there are $\binom{N}{n_1, \ldots n_m}$ ways in which the mutations could have arrived (according to the Poisson process that is laying down mutations according to $\nu \times$ Lebesgue). Thus, computation of the expectation involves $\binom{N}{n_1, \ldots n_m}$ terms. These computations are untenable in practice. A possible alternative is to keep track of the order in which mutations arrive so that

$$P(g) = \sum_{\vec{g} \in G} P(\vec{g})$$

where $G = \{\vec{g}\} = \{\text{all orderings of the mutations in } g\}$. In keeping track of the order, however, the space over which the random walk has to crawl is much larger, since each possible genotype now has many different representations.

Details of the Implementation

The implementation of the Metropolis-Hastings algorithm assumes that the mutation space $\mathcal{M}$ contains a finite number of mutations, M. This assumption is not terribly restrictive for

the cases that we consider since we are dealing with easily discretized spaces.

Initialization

The user chooses an initial genotype. This genotype is represented as a vector that describes the number of copies of each type of mutation present in the genome,

$$g = \mathbf{n} = (n_1, n_2, \dots n_M)$$

where n_i indicates the number of copies of mutation type i. Let $N = |g| = \sum_i n_i$ be the number of mutations in the unordered genotype g. The algorithm takes this initial genotype and creates an ordered genotype, $\vec{g}$,

$$\vec{g} = (m^{(1)}, m^{(2)}, \dots m^{(N)})$$

where $m^{(i)}$ denotes the i^{th} mutation laid down by the underlying Poisson process. The ordered genotype contains the same number of mutations as g but describes the order in which these mutations arrived. The probability of the ordered genotype is proportional to

$$\begin{aligned} P(\vec{g}) \propto & \frac{\nu(\mathcal{M})^N}{S(m^{(1)})S(m^{(1)}m^{(2)})\cdots S(m^{(1)}m^{(2)}\cdots m^{(N)})} P_{\Pi}\left(m^{(1)}m^{(2)}\cdots m^{(N)}\right) \\ = & \frac{\nu(\mathcal{M})^N}{S(m^{(1)})S(m^{(1)}m^{(2)})\cdots S(m^{(1)}m^{(2)}\cdots m^{(N)})} \frac{\nu\left(m^{(1)}\right)\cdots\nu\left(m^{(N)}\right)}{\nu(\mathcal{M})^N}. \end{aligned}$$

Proposed Step

The proposed ordered genotype is generated in two steps. First, a new (unordered) genotype, g', is proposed. For each mutation type, a proposed number of copies of the mutation is drawn according to a double-sided discrete exponential distribution centered about the number of copies contained in the current genotype. The double exponential is restricted to nonnegative numbers.

$$\begin{cases} n_i' \leftarrow n_i - k & \text{for } k = 0, 1, \dots n_i \text{ with probability } \frac{1-\exp(-\lambda(n_i+1))}{2-\exp(-\lambda(n_i+1))} \frac{(1-\exp(-\lambda))\exp(-k\lambda)}{1-\exp(-\lambda(-n_i+1))} \\ n_i' \leftarrow n_i + k & \text{for } k = 0, 1, \dots \text{ with probability } \frac{1}{2-\exp(-\lambda(n_i+1))}(1-\exp(-\lambda))\exp(-k\lambda). \end{cases}$$

This implementation uses the same the exponential parameter λ for each mutation type. Repeating this process for each type of mutation produces the proposed genotype, $g' = (n_1', n_2', \dots n_M')$.

The second step is to randomly generate an ordered genotype $\vec{g}'$ corresponding to the proposed genotype g',

$$\vec{g}' = \left(m^{(1)'}, m^{(2)'}, \ldots m^{(N')'}\right)$$

where N' is the total number of mutations in the proposed genotype, $N' = \sum_i n_i'$. Notice that the proposed genotype and the current genotype may have a different number of mutations.

Accepting or Rejecting the Proposed Step

In order to compute α, the probability of accepting the proposed move, we must calculate the ratio of the probabilities of the ordered genotypes, denoted by a_1,

$$a_1 = \frac{\pi(x')}{\pi(x^t)} = \frac{P(\vec{g}')}{P(\vec{g})},$$

and the ratio of the jump (or proposal) probabilities, denoted by a_2,

$$a_2 = \frac{Q(x^t|x')}{Q(x'|x^t)} = \frac{Q(\vec{g}|\vec{g}')}{Q(\vec{g}'|\vec{g})}.$$

The probability of accepting the move is then

$$\alpha = \min\{a_1 a_2, 1\}.$$

The ratio of the probabilities of the ordered genotypes is given by

$$a_1 = \frac{P(\vec{g}')}{P(\vec{g})} = \frac{\frac{\nu(\mathcal{M})^{N'}}{S\left(m^{(1)'}\right)S\left(m^{(1)'}m^{(2)'}\right)\cdots S\left(m^{(1)'}m^{(2)'}\cdots m^{(N')'}\right)}\frac{\nu\left(m^{(1)'}\right)\cdots\nu\left(m^{(N')'}\right)}{\nu(\mathcal{M})^{N'}}}{\frac{\nu(\mathcal{M})^{N}}{S(m^{(1)})S(m^{(1)}m^{(2)})\cdots S(m^{(1)}m^{(2)}\cdots m^{(N)})}\frac{\nu\left(m^{(1)}\right)\cdots\nu\left(m^{(N)}\right)}{\nu(\mathcal{M})^{N}}}.$$

We compute the probability of jumping to the ordered genotype in two steps. First, we compute the probability of jumping to the unordered genotype. Because each mutation type is updated independently we have

$$Q(g'|g) = \prod_{i=1}^{M} q(n_i'|n_i).$$

Notice that if $n_i' = n_i + k$ then

$$q(n_i'|n_i) = \frac{1}{2 - \exp(-\lambda(n_i+1))}(1-\exp(-\lambda))\exp(-k\lambda)$$

and $n_i = n_i' - k$ so

$$q(n_i|n_i') = \frac{1-\exp(-\lambda(n_i'+1))}{2-\exp(-\lambda(n_i'+1))}\frac{(1-\exp(-\lambda))\exp(-k\lambda)}{1-\exp(-\lambda(-n_i'+1))}.$$

Then,

$$\frac{q(n_i|n_i')}{q(n_i'|n_i)} = \frac{2-\exp(-\lambda(n_i+1))}{2-\exp(-\lambda(n_i'+1))}.$$

Clearly the same result holds for the case where $n_i' = n_i - k$.

There are

$$\binom{N}{n_1 n_2 \dots n_M}$$

possible ordered genotypes corresponding to the unordered genotype g and

$$\binom{N'}{n_1' n_2' \dots n_M'}$$

possible ordered genotypes corresponding to the unordered genotype g'. Because the ordered genotypes are generated randomly, we have

$$Q(\vec{g}'|\vec{g}) = \frac{1}{\binom{N'}{n_1' n_2' \dots n_M'}} \prod_{i=1}^{M} q(n_i'|n_i)$$

and

$$Q(\vec{g}|\vec{g}') = \frac{1}{\binom{N}{n_1 n_2 \dots n_M}} \prod_{i=1}^{M} q(n_i|n_i').$$

The ratio is

$$a_2 = \frac{Q(\vec{g}|\vec{g}')}{Q(\vec{g}'|\vec{g})} = \frac{N'!}{N!} \prod_{i=1}^{M} \frac{2-\exp(-\lambda(n_i+1))}{2-\exp(-\lambda(n_i'+1))} \frac{n_i!}{n_i'!}.$$

In practice there have been several difficulties with this approach. As with the naive algorithm, this approach has proven to be somewhat problematic in cases with either a large number of mutations or with mutations that have late-acting effects coupled with high mutation rates. It appears that the main issue is that ordered genotypes are created randomly. Although the number of each type of mutation is determined according to a double exponential distribution centered around the number of copies of that mutation type in the current genotype, the time-ordered genotype is a random shuffling of the unordered genotype. This can result in proposed steps that are highly unlikely, especially when the mutation space is large.

Consider the simple case where mutations have point-mass profiles. Suppose there are only two possible mutation types and these two types have ages of onset m_1 and m_2 with

$m_1 < m_2$. It is clear that of all the possible orderings of an unordered genotype g, the one with mutations in *descending* order will be most likely. Intuitively this is because the mutation with later-age effects (in this case, m_2) has less of an impact on lifespan and, consequently, less of an impact on the number of offspring produced. Over time, as more mutations accumulate in the population, individuals who receive an additional copy of m_2 will have a longer lifespan (and more offspring) than individuals who instead receive a copy of m_1. For example, of all genotypes with exactly two mutations, the ordered genotype (m_2, m_2) will be most likely, followed by (m_2, m_1), then (m_1, m_2) and finally (m_1, m_1). By randomly generating the order, we have an equal chance of generating (m_2, m_1) and (m_1, m_2) (for example) even though we know individuals with the genotype (m_2, m_1) are more likely to survive and reproduce. Exactly how much more likely that order is depends on a number of factors (including the ages of onset). One can imagine how much more problematic this issue may be when there are hundreds of mutations in the genotype.

A secondary (but somewhat related) issue is the fact that each step of the algorithm involves updating the number of copies of every type of mutation. In practice it appears that the chain can get stuck, accepting virtually no moves because most orders of the genotype with one additional copy of an early-acting mutation will be highly unlikely relative to the current state. This is particularly an issue when one mutation has significantly later effects than the other mutations. For the same reason, using the same exponential rate parameter for all mutation types is also unwise. Mutations with late-age effects will occur more frequently in the genome and with more variation in the number of copies than mutations with early-age effects. All of these issues will be illustrated by several concrete examples in §3.3.

3.2.3 Multiple-Try Metropolis Algorithm

In an attempt to partially side-step the issues present in the Metropolis-Hastings approach, we consider a third approach to sampling from the genotype distribution under the SEW model. This last approach uses the multiple-try Metropolis algorithm [23], which proceeds as follows:

- At time t, k independent states are chosen from the proposal distribution $Q(x^{(t)}, \cdot)$, where $x^{(t)}$ is the current state of the chain.
- For each proposed move, y_j, we compute the weight,

  $$w(y_j, x^{(t)}) = P(x^{(t)})Q(x^{(t)}, y)\lambda(x^{(t)}, y)$$

  where λ is a symmetric, non-negative function. For the implementation used in this work we set $\lambda = 1$.
- Choose $y \in \{y_1, y_2 \dots y_k\}$ according to the weights $w(y_1, x^{(t)}), \dots w(y_k, x^{(t)})$.
- Draw $k-1$ independent states, $x_1, \dots x_{k-1}$, from $Q(y, \cdot)$, setting $x_k = x^{(t)}$.

- Compute the weights associated with these states, $w(x_1, y) \dots w(x_k, y)$.
- Accept the proposed move to state y with probability

$$\min\left(1, \frac{w(y_1, x^{(t)}) + \cdots + w(y_k, x^{(t)})}{w(x_1, y) + \cdots + w(x_k, y)}\right).$$

As with the Metropolis-Hastings algorithm, in the implementation of the multiple-try Metropolis algorithm used in this work, the chain crawls over the space of ordered genotypes. This approach eliminates the need to compute the expectation over all possible orders in the $\tilde{P}\mathbf{1}_{g=n}$ term. In what follows, all genotypes should be understood to be time-ordered. That is, the genotype $\vec{g} = (m_i, m_j)$ indicates that the mutation event producing m_i occurred before the mutation event that produced m_j.

A proposed state differs from the current state by no more than one mutation. Specifically, given the current (ordered) genotype, we can choose one of three actions:

- Add a mutation anywhere along the genotype (that is, we can add another mutation event at any point in time).
- Remove a mutation (erase one of the mutation events that occurred).
- Change the type of one mutation (keep the mutation event but change the mutation from type i to type j).

If the current ordered genotype contains N mutations, then a mutation is added to it with probability $\frac{N+1}{2N+1}$. A mutation is deleted with probability $p\frac{N}{2N+1}$ and a mutation changes type with probability $(1-p)\frac{N}{2N+1}$. In the case with only one type of mutation, p is set to 1.

As a simple example, suppose that the current genotype is $\vec{g} = (m)$. In this example, $N = 1$ and we have the following four options for the proposed ordered genotype $\vec{g}'$,

$$\vec{g}' = \begin{cases} (m', m) & \text{with probability } \frac{1}{3}\frac{\nu(m')}{\nu(\mathcal{M})} \\ (m, m') & \text{with probability } \frac{1}{3}\frac{\nu(m')}{\nu(\mathcal{M})} \\ () & \text{with probability } \frac{1}{3}p \\ (m') & \text{with probability } \frac{1}{3}(1-p)\frac{\nu(m')}{\nu(\mathcal{M})} \end{cases}$$

where $m' \in \mathcal{M}$ is any mutation in the space of mutation types. The genotype $\vec{g} = ()$ denotes the null or wild-type genome. Please note that in the tests that follow, the probability of deleting a chosen mutation will be referred to as DelP, rather than p.

This approach has an advantage over the Metropolis-Hastings algorithm in cases where the mutation space is large. Because the algorithm produces several proposed steps and then chooses one of them according to their weights, it is more likely to accept the proposed move than in the Metropolis-Hastings algorithm when dealing with a large mutation space.

In practice it also appears less likely for the chain to get stuck, probably due to the different method of generating ordered genotypes. However, because the proposed genotypes differ from the current genotype by at most one mutation, this method can require thousands of iterations to locate highly likely genotypes, especially when the mutation space is large.

3.3 Illustrative Test Cases

Our ultimate goal in this work is to characterize the distribution of genotypes under the SEW mutation-selection model for mutation spaces that have hundreds or thousands of different mutation types. In particular we will eventually focus on mutations with profiles that are described by gamma distributions. Such profiles are of interest to us because they allow us to easily model both mutations that have large early-age effects and mutations that have small early-age effects but much larger late-age effects. Before considering these large mutation spaces, however, we will present results for small mutation spaces, consisting of at most four distinct types of mutations.

The purpose of these smaller cases is two-fold. First, it allows us to verify that the implementations of algorithms presented in §3.2-§3.2.3 are functioning properly and can be used to sample from the target distribution (in this case, the distribution of genotypes). Second, it will illustrate why some of the methods described above are ill-suited to sampling from the space of genotypes when the mutation space is large. In particular we find that the naive algorithm and the Metropolis-Hastings algorithm fail for even small mutation spaces when the mutations have gamma profiles.

We begin by focusing on a mutation space with a single type of mutation. We chose to start with this case, which we discussed previously in §2.1.6 as the mutation counting model, because the series solution to the SEW model for a mutation space with a single type of mutation is much simpler than it is in cases with more than one type of mutation. To make the series solution even more tractable we will use a highly stylized form for the age-specific effects of the mutation on the hazard rate. The mutation profile used, called a point-mass profile, concentrates the effect of the mutation to a single age, resulting in an increase in hazard rate at a single point. Although this type of mutation profile is highly unrealistic and thus not useful in attempting to model mutations in the real world, its simplicity will allow us to accurately estimate the true distribution of genotypes.

We next consider a mutation space containing two types of mutations, each with a point-mass profile. This example is slightly more complicated than the mutation counting model because with two mutation types the series solution becomes intractable. As a result, we can only estimate the distribution of genotypes using the three numerical methods presented previously, namely the naive (N) algorithm, the Metropolis-Hasting (MH) algorithm and the multiple-try Metropolis (MTM) algorithm. However, while we cannot compute the true distribution exactly, agreement between the output of the three algorithms strengthens our belief that the algorithms are performing correctly.

Our last test case involves a mutation space with four types of mutations. Unlike the previous two cases, the mutation profiles in this case are described by gamma distributions where all four mutations have the same rate parameter but each has a different shape parameter. The shape parameters are chosen so that the age-specific effects of each mutation are quite different. One mutation has an early mean-age effect, indicating that much of the effect of the mutation is concentrated during early reproductive years. A second mutation has a mean-age effect in middle age, indicating moderate early-age effects. The third mutation has a late mean-age effect and the fourth mutation has a very late mean-age effect. These last two mutations have small to very small early-age effects. This test case is primarily used to illustrate the difficulty of sampling from the distribution of genotypes under the SEW model when the mutation space contains more than a small number of mutations.

3.3.1 Mutation Counting Model with Point-Mass Profiles

In the first scenario we consider, the mutation space $\mathcal{M}$ consists of a single mutation with a point-mass profile. As mentioned previously, for the point-mass profile case, the effect of the mutation on the hazard rate is concentrated at a single age, m. The effect on the cumulative hazard rate, then, is modeled by a Heaviside step function with step at the age of onset m. Recall from §2.1.6 that when the mutation space consists of a single type of mutation, the series solution to the ESW free recombination model simplifies considerably to

$$P(n) = \frac{\nu(\mathcal{M})^n \frac{1}{S(1)\cdots S(n)}}{1 + \sum_j \nu(\mathcal{M})^j \frac{1}{S(1)\cdots S(j)}}. \tag{3.2}$$

When the mutation has a point-mass profile and the fertility rate is constant from the age of maturity, α, to the oldest age of reproduction, β, and zero otherwise, the selective cost function can be evaluated exactly and has the simple form

$$\begin{aligned} S(1) =& (1 - e^{-\eta}) \frac{f}{\lambda} \left(e^{-\lambda(m-\alpha)} - e^{-\lambda(\beta-\alpha)} \right) \\ S(2) =& (1 - e^{-2\eta}) \frac{f}{\lambda} \left(e^{-\lambda(m-\alpha)} - e^{-\lambda(\beta-\alpha)} \right) \\ & \vdots \\ S(n) =& (1 - e^{-n\eta}) \frac{f}{\lambda} \left(e^{-\lambda(m-\alpha)} - e^{-\lambda(\beta-\alpha)} \right). \end{aligned}$$

The product of the selective cost functions is then given by

$$S(1)S(2)\cdots S(n) = \left(\frac{f}{\lambda} \right)^n \left(e^{-\lambda(m-\alpha)} - e^{-\lambda(\beta-\alpha)} \right)^n (1 - e^{-\eta})(1 - e^{-2\eta}) \cdots (1 - e^{-n\eta}).$$

Plugging this product into equation (3.2), we find that the probability that a randomly chosen individual has a genotype with exactly n mutations is given by the expression

$$P(n) = \frac{\frac{\nu(\mathcal{M})^n}{\left(\frac{f}{\lambda}\right)^n \left(e^{-\lambda(m-\alpha)} - e^{-\lambda(\beta-\alpha)}\right)^n \prod_{k=1}^{n} 1-\exp(-k\eta)}}{1 + \sum_j \frac{\nu(\mathcal{M})^j}{\left(\frac{f}{\lambda}\right)^j \left(e^{-\lambda(m-\alpha)} - e^{-\lambda(\beta-\alpha)}\right)^j \prod_{k=1}^{j} 1-\exp(-k\eta)}}. \tag{3.3}$$

We compared the true distribution of genotypes in the single point-mass profile case, given by equation (3.3), to the estimates obtained from the naive algorithm, the MH algorithm and the MTM algorithm for three test cases. The three test cases discussed here have ages of onset 25, 35 and 45 years, respectively. Demographic parameters, such as the background hazard rate λ and the mutation rate ν, as well as algorithm parameters for the three numerical approaches, are listed in Table 3.1. The term "Parameter" listed in the Metropolis-Hastings (MH) row refers to the exponential parameter that determines the number of copies of each mutation type in the proposed genotype (details can be found in §3.2.2). The term "Burn" refers to an initial number of steps that were discarded before samples were collected for analysis. Questions of whether or not the Markov chain has converged to the target distribution are considered in Appendix A.1 and will not be repeated here. As will be typical for all the cases discussed in this work, the MH and MTM algorithms were initialized with the null genotype.

The fertility rate for each case was set equal to the fertility rate that ensures a stationary population under the ESW free recombination model. The shortcut method was used to solve the free recombination model under the same mutation space and with the same parameters. The number of iterations before convergence of the shortcut method and the resulting fertility rate are listed in Table 3.2.

Histograms of the number of mutations per genotype resulting from the three algorithms as well as the distribution given by equation (3.3) are shown in Figure 3.1 (age of onset 25 years), Figure 3.2 (age of onset 35 years) and Figure 3.3 (age of onset 45 years). A quick visual analysis suggests that in all three cases, the empirical distributions are similar to the true distribution. For a more concrete measure, we consider the L^∞ distance between the true distribution and the output from the three algorithms, which is provided in Table 3.4. The table shows that the naive algorithm is more accurate than the MH and MTM algorithms in the cases with younger ages of onset (25 and 35). In the last case (with an age of onset of 45 years) the three algorithms perform similarly. The difference between the naive algorithm and the true distribution is largely caused by the fact that the naive algorithm estimates the selective cost function by numerically integrating rather than using the closed-form expression presented above.

Because we are ultimately interested in demographic outcomes, Figure 3.4 shows the expected population survival function, $\mathbb{E}[l_x(G)]$, computed using the actual probability distribution, and the estimates from the naive algorithm, the MH algorithm and the MTM algorithm. For all three cases, the survival functions estimated from the three algorithms

Table 3.1: Parameters for test cases with a single point-mass mutation.

All Tests	η	λ	α	β	Dx
	0.1	0.05	15	50	0.1
Trial	$\nu(\mathcal{M})$	Mutation Age			
1	0.3	25			
2	0.1	35			
3	0.02	45			
NH	NumEv	NumTrials			
	100	10000			
MH	Burn	Samples	Parameter		
	10000	100000	0.5		
MTM	Burn	Samples	DelP	Kmax	
	10000	100000	0.75	5	

Table 3.2: Results from the free recombination model for the single point-mass mutation cases.

Trial	Mutation Age	Iterations	Fertility	Rho
1	25	131	0.095884234	12.74311634
2	35	58	0.070267910	9.34025338
3	45	26	0.062151901	6.04213278

Table 3.3: Proportion of the proposed moves accepted by the MH and MTM algorithms for the three single point-mass profile cases.

		Acceptance Rate	
Trial	Mutation Age	MH Algorithm	MTM Algorithm
1	25	0.82088	0.89839
2	35	0.82129	0.90091
3	45	0.81340	0.89664

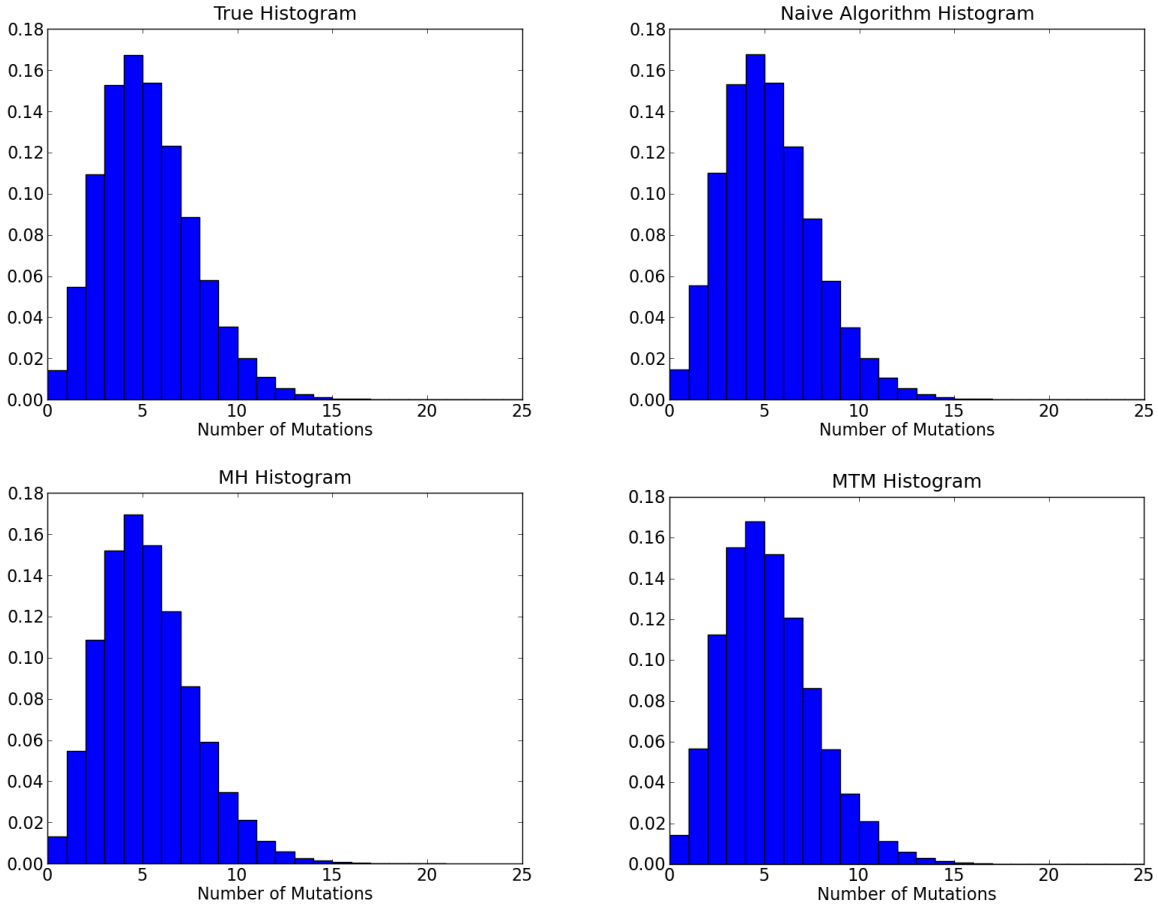

Figure 3.1: Histograms for the single point-mass case with age of onset 25 years. True probability distribution (top left), naive algorithm (top right), MH algorithm (bottom left) and MTM algorithm (bottom right).

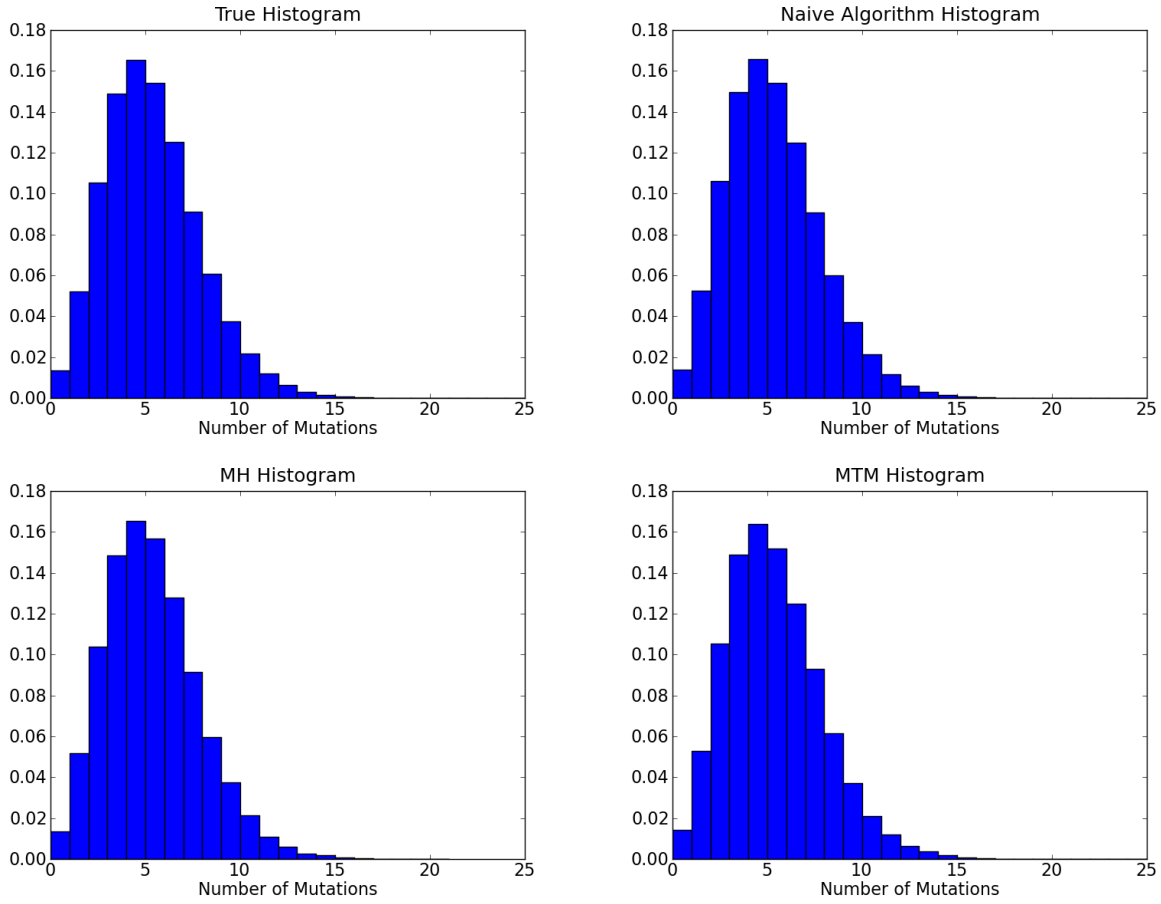

Figure 3.2: Histograms for the single point-mass case with age of onset 35. True probability distribution (top left), naive algorithm (top right), MH algorithm (bottom left) and MTM algorithm (bottom right).

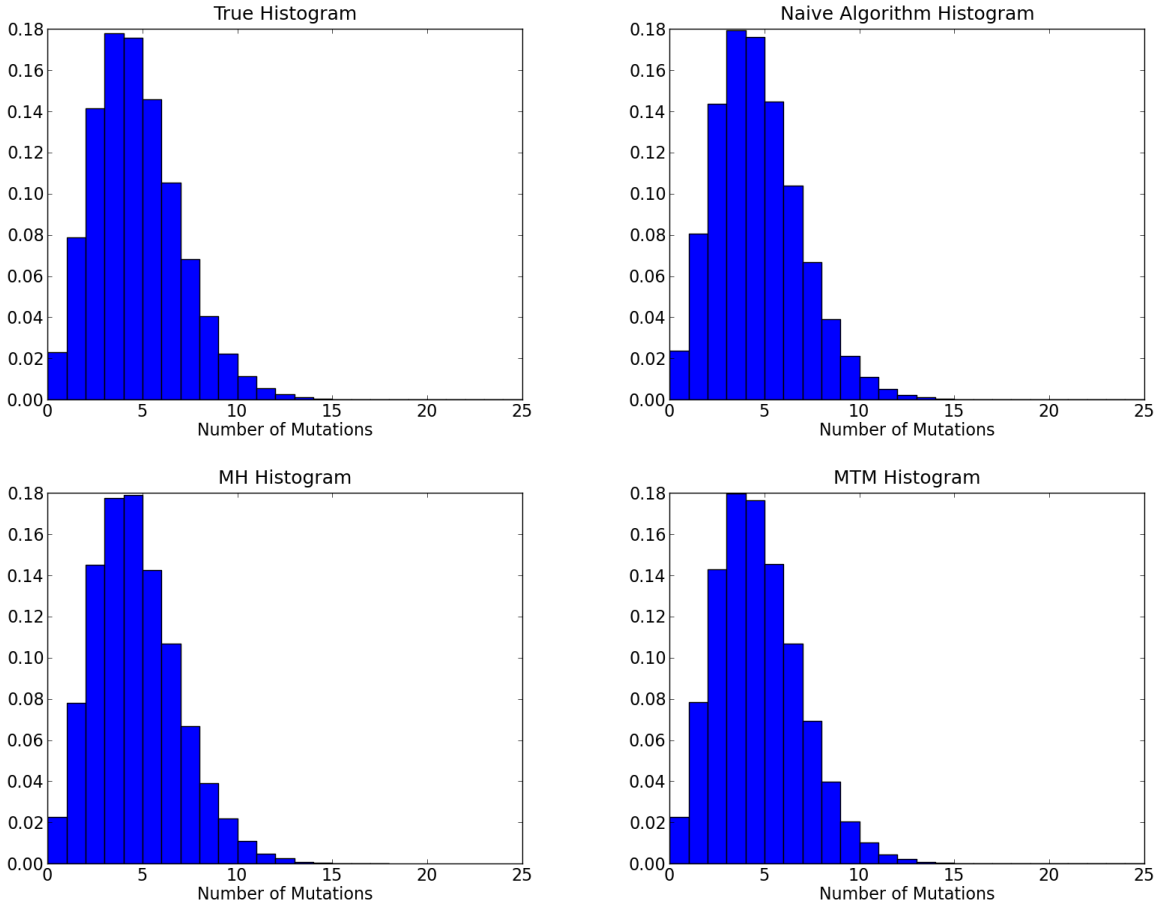

Figure 3.3: Histograms for the single point-mass case with age of onset 45. True probability distribution (top left), naive algorithm (top right), MH algorithm (bottom left) and MTM algorithm (bottom right).

are essentially the same, with L^2 distances between the actual survival function and the approximations on the order of 10^{-3} (see Table 3.5 for details).

Table 3.4: Distance (L^∞) between the actual probability distribution and the empirical distribution from the naive algorithm (N), the MH algorithm and the MTM algorithm for the three single point-mass profile cases.

Trial	Actual & N	Actual & MH	Actual & MTM
1	0.0007096	0.00226183	0.00294207
2	0.0009512	0.00262767	0.00241233
3	0.00232584	0.00370885	0.00178069

Table 3.5: Distance between $\mathbb{E}[l_x(G)]$ using the actual probabilities and the output from the algorithms.

	Actual & N		Actual & MH		Actual & MTM	
Trial	$\|\|\cdot\|\|_2$	$\|\|\cdot\|\|_\infty$	$\|\|\cdot\|\|_2$	$\|\|\cdot\|\|_\infty$	$\|\|\cdot\|\|_2$	$\|\|\cdot\|\|_\infty$
1	0.00161847	0.00051097	0.00165686	0.00052309	0.00384565	0.00121412
2	0.00134495	4.24646e-04	0.00030519	9.63588e-05	0.00022359	7.05952e-05
3	0.00172254	0.00054396	0.00102528	0.00032377	0.00103437	0.00032664

3.3.2 Mutation Space with Two Point-Mass Profile Mutations

Unlike the cases with a single mutation type, the true probability distribution for spaces with two mutations cannot be directly computed. As a result, output from the three algorithms (N, MH and MTM) were compared only against one another. The results for two such test cases are reproduced below. The first case has mutations with ages of onset 20 years and 30 years; the second case has mutations with ages of onset 20 years and 40 years. The parameters used in these two cases are listed in Table 3.6. The fertility rates, set to ensure no population growth under the free recombination model, are listed in Table 3.7.

It is worth taking a minute to call attention to the acceptance ratios for the MH and MTM algorithms. In the cases with a single point-mass profile mutation, the acceptance ratio (the ratio of accepted moves to proposed moves) was about 82% for the MH algorithm and 90% for the MTM algorithm (see Table 3.3). With two point-mass profile mutations, the acceptance ratios for both algorithms have dropped, with the MH algorithm accepting only 42-54% of the proposed steps and the MTM accepting almost 73% of the proposed steps. Note that the MH algorithm has a higher acceptance ratio for the case where both

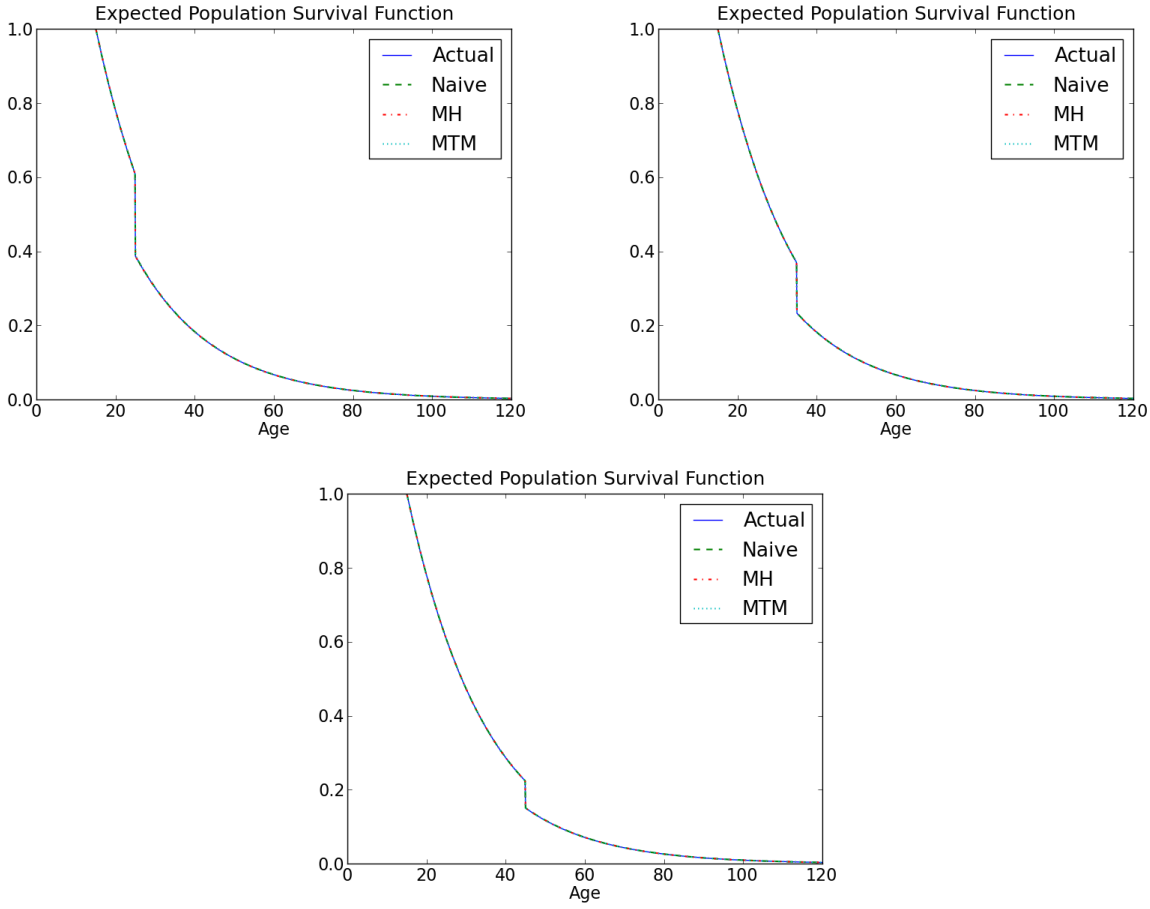

Figure 3.4: Expected population survival for the single point-mass case with age of onset 25 (top left), 35 (top right), and 45 (bottom).

mutations have early to mid-range ages of onset and a lower acceptance ratio for the case where one mutation has a much later age of onset (40 years) than the other mutation (20 years). The MTM algorithm, by contrast, has essentially the same acceptance ratio for both test cases. This difference becomes even more pronounced when considering larger mutation spaces, as we will do in the next section.

Table 3.6: Parameters for test cases with two point-mass mutations.

All Tests	η	λ	α	β	Dx
	0.1	0.05	15	50	0.1
Trial Number	$\nu(\mathcal{M})$	Mutation Age			
4	0.2	(20,30)			
5	0.1	(20,40)			
NH	NumEv	NumTrials			
	1000	10000			
MH	Burn	Samples	Parameter		
	10000	100000	0.5		
MTM	Burn	Samples	DelP	Kmax	
	10000	100000	0.75	5	

The plots in Figure 3.5 show the histograms for the total number of mutations per genotype in the case with ages of onset $m_1 = 20$ years and $m_2 = 30$ years. The marginal distributions for the two mutation types were also computed using samples from each algorithm and are shown in Figure 3.6. The plots on the left side of Figure 3.6 show the marginal distributions for the number of copies of m_1 per genotype while the plots on the right show the marginal distributions for the number of copies of m_2. The three algorithms produce similar histograms for the total number of mutations, as well as similar marginal histograms. The L^∞ distance between the distributions for the total number of mutations

Table 3.7: Output for the free recombination model for the cases with two point-mass mutations.

Trial Number	Mutation Age	Iterations	Fertility	Rho
4	(20,30)	13	0.0755627	(1.57387, 3.91628)
5	(20,40)	34	0.0683494	(0.751241, 7.31567)

Table 3.8: Acceptance rates for proposed moves in the Metropolis-Hasting algorithm and the multiple-try Metropolis algorithm.

Trial Number	MH Algorithm	MTM Algorithm
4	0.54146	0.72870
5	0.42910	0.72864

per genotype between the three algorithms are listed in Table 3.9. The table confirms that the three approaches yield very similar results. Figure 3.7 (left) shows the expected population survival function computed from the output of the three algorithms. Again, the visual observation that the expected survival functions are nearly identical is supported by computing the distances between the functions, shown in Table 3.10.

Figure 3.8 shows the histograms for the total number of mutations in the case where the ages are onset are $m_1 = 20$ years and $m_2 = 40$ years. The marginal distributions are shown in Figure 3.9. As with the previous case, the three algorithms all produce similar results, a finding that is confirmed by comparing the L^∞ distance between the distributions (Table 3.9).

In both of the cases considered here, the empirical distribution estimated by the naive algorithm and that from the MTM samples appear closest of the three comparisons. Although we don't know the true distribution of genotypes in the cases with two mutations, we expect the naive algorithm to provide a reasonable estimate because the mutation space is small and the mutations have early- to mid-reproductive ages of onset. This ensures that the selective pressure against the mutations is fairly strong, which, in turn, means that typical genotypes will contain few copies of each type of mutation. The method utilized by the naive algorithm – fixing a largest possible genotype (in this case containing at most 1000 mutations, which is much larger than the typical genotype) and choosing a random order of arrival for those mutations – suggests that genotypes with high probability will be encountered often in these cases. As a result, we expect the naive algorithm to perform pretty well on these test cases. Close agreement between the naive algorithm and the algorithms using Markov chains to sample from the distribution of genotypes is encouraging.

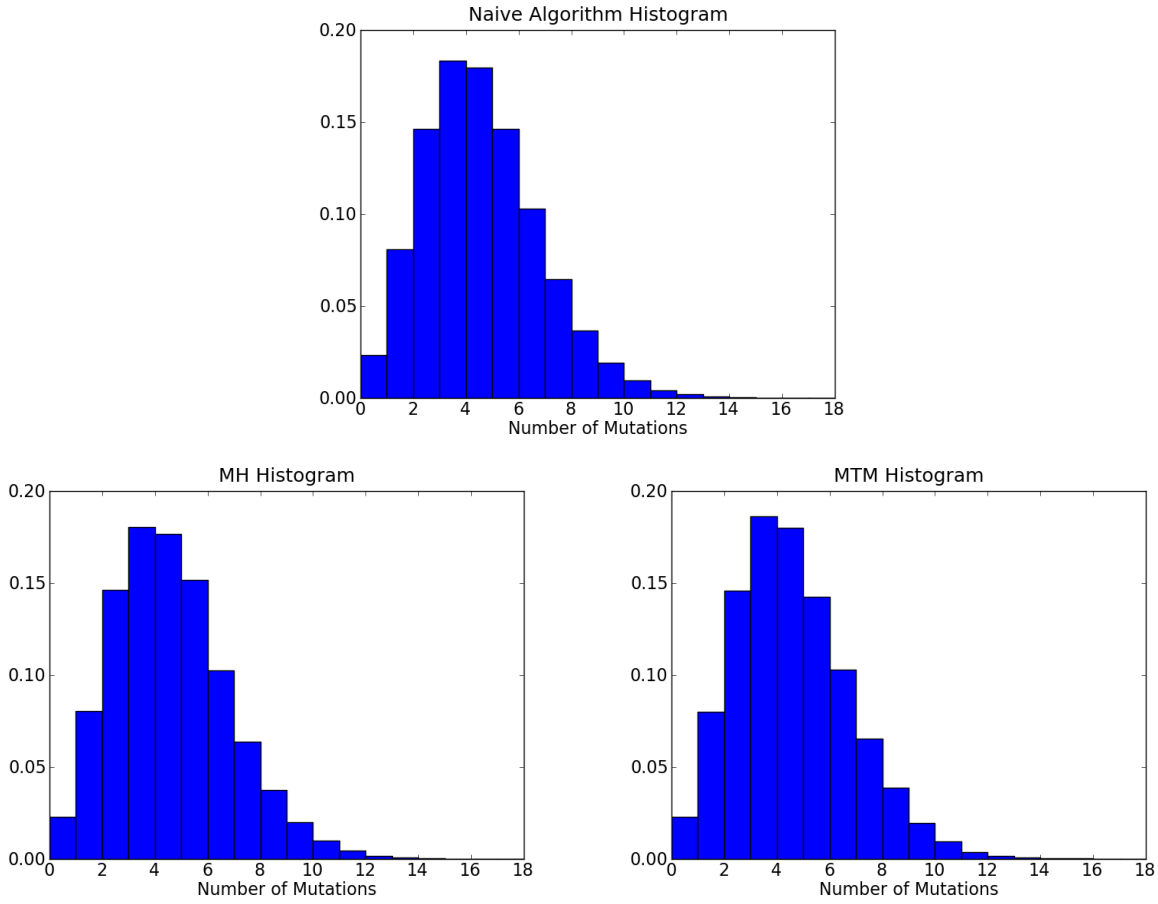

Figure 3.5: Histograms for the two point-mass mutations case with ages of onset $m_1 = 20$ years and $m_2 = 30$ years. Naive algorithm (top), MH algorithm (bottom left) and MTM algorithm (bottom right).

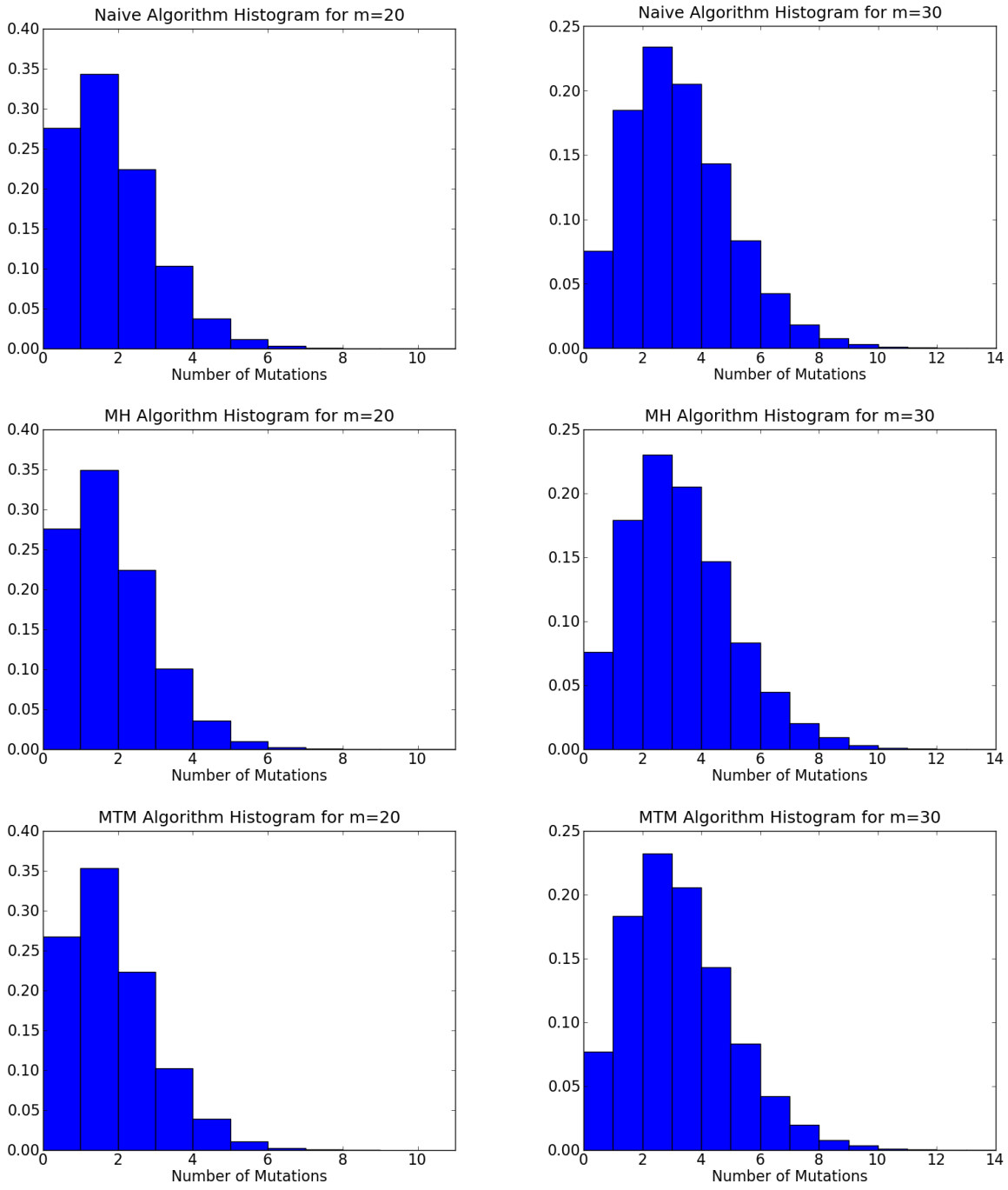

Figure 3.6: Marginal distribution histograms for the two point-mass mutations case with ages of onset $m_1 = 20$ years (left column) and $m_2 = 30$ years (right column) for the naive algorithm (top row), the MH algorithm (middle row) and the MTM algorithm (bottom row).

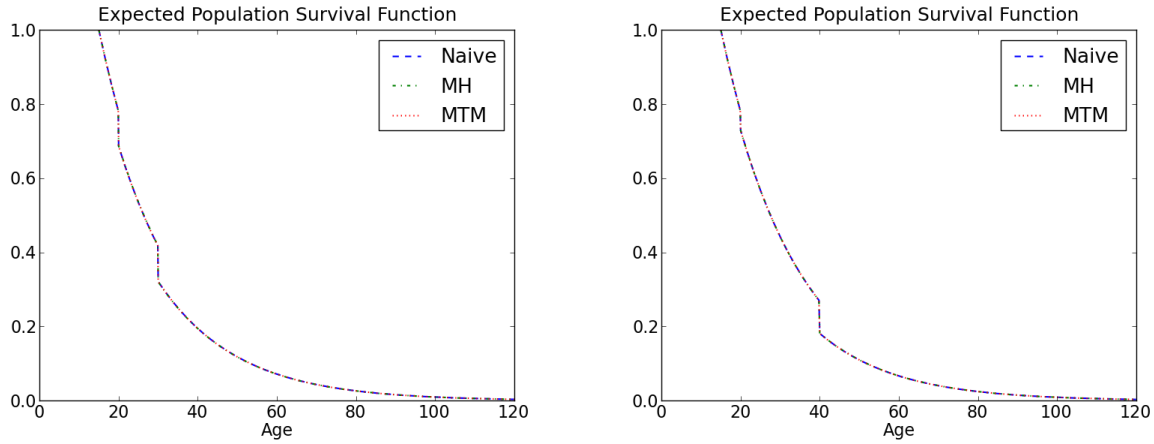

Figure 3.7: Expected population survival function for the two point-mass mutations case with ages of onset $m_1 = 20$ years and $m_2 = 30$ years (left) and $m_1 = 20$ years and $m_2 = 40$ years (right).

Table 3.9: Distance (L^∞) between the distributions of the total number of mutations obtained from the three algorithms.

Trial	N & MH	N & MTM	MH & MTM
4	0.00583465	0.00368195	0.00857
5	0.00538978	0.00289	0.00416

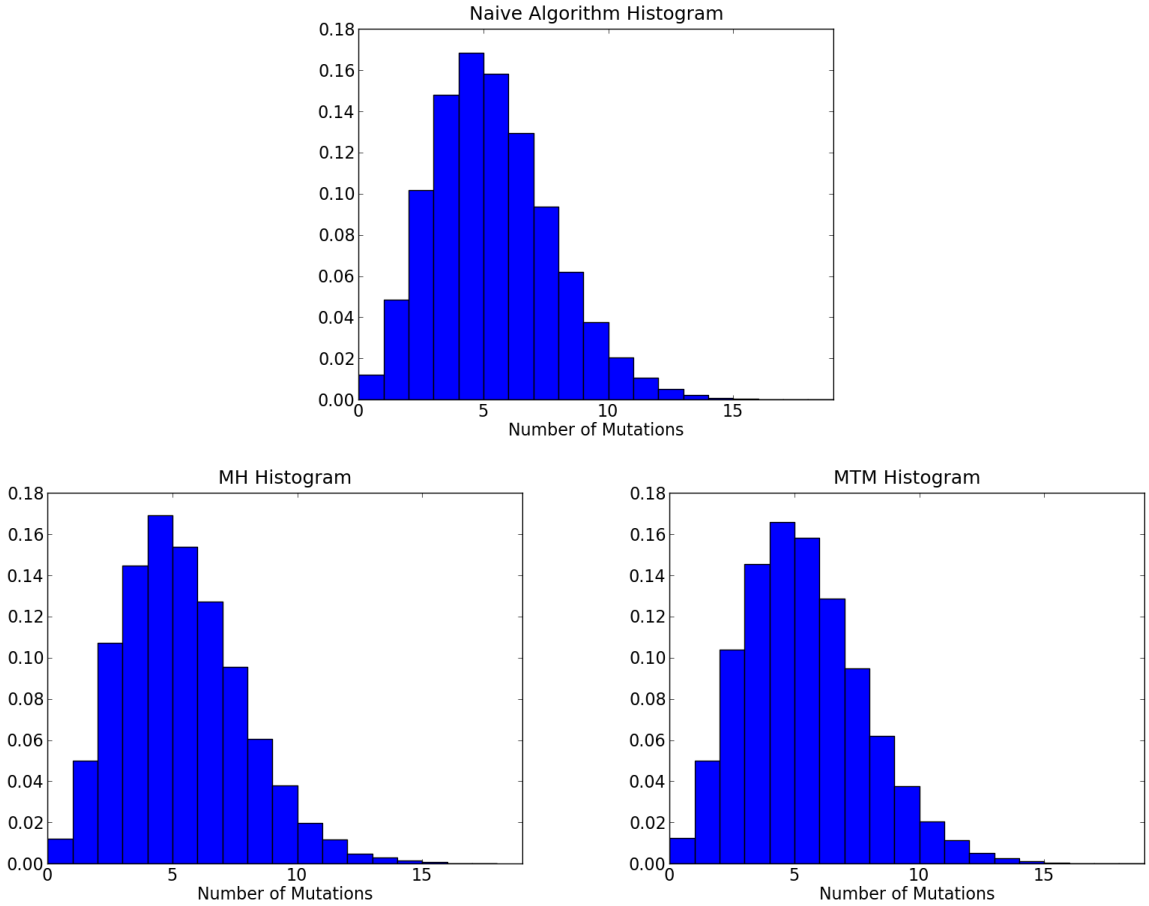

Figure 3.8: Histograms for the two point-mass mutations case with ages of onset $m_1 = 20$ years and $m_2 = 40$ years. Naive algorithm (top), MH algorithm (bottom left) and MTM algorithm (bottom right).

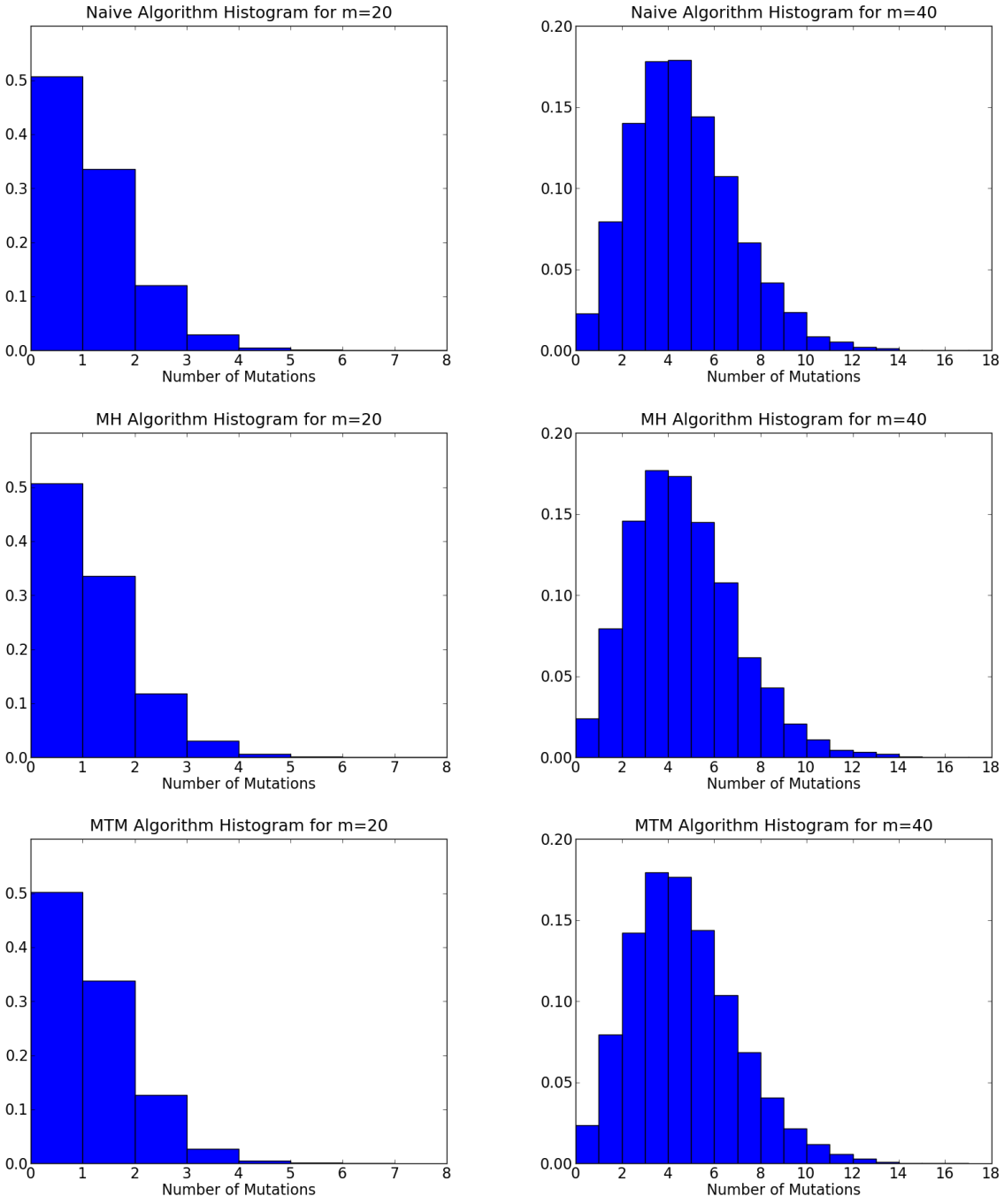

Figure 3.9: Marginal distribution histograms for the two point-mass case with ages of onset $m_1 = 20$ years (left column) and $m_2 = 40$ years (right column) for the naive algorithm (top row), the MH algorithm (middle row) and the MTM algorithm (bottom row).

Table 3.10: Distance between the expected survival functions for the three algorithms.

	N & MH		N & MTM		MH & MTM	
Trial	$\|\|\cdot\|\|_2$	$\|\|\cdot\|\|_\infty$	$\|\|\cdot\|\|_2$	$\|\|\cdot\|\|_\infty$	$\|\|\cdot\|\|_2$	$\|\|\cdot\|\|_\infty$
4	0.00316648	0.00102678	0.00121233	0.00035107	0.00361067	0.00137785
5	0.00070704	0.00017743	0.00087384	0.00029289	0.00070317	0.00019434

3.3.3 Mutation Space with Four Gamma Profile Mutations

The previous cases presented in this chapter consist of spaces where mutations have point-mass profiles. Essentially, these mutations have an age of onset or an age of activation before which the mutation is benign. The simplicity of this model of mutation effect on hazard rates allowed us to accurately estimate the true probability distribution in the mutation counting case ($\mathcal{M} = \{m\}$). However, it is the simplicity of this profile that makes it biologically unrealistic for modeling real mutations. Perhaps a more realistic model for the age-specific effects of genetic mutations on the hazard rate assumes that a mutation has non-zero effects at every age. In this section we will model mutation effects on the hazard rate with a gamma distribution.

Each mutation in the space has the same gamma rate parameter of 0.05 but different shape parameters. The shape parameters used are 1.125, 2.25, 4.125 and 6.00. These parameters where chosen in line with tests from [37], which discusses the behavior of the ESW free recombination model. Figure 3.10 shows the cumulative gamma profiles for this case. Recall that the cumulative mutation profile adds to the baseline cumulative hazard function when an individual has a copy of that mutation in their genome. For gamma profiles, mutations with smaller shape parameters have a larger effect on the cumulative hazard function at younger ages than do those with larger shape parameters. Compare, for example, a shape parameter of 6.00 and 1.125. The mutation with shape parameter 6.00 has a very small effect on the cumulative hazard function over reproductive ages, increasing it by less than $.01\eta$ over the entire range of reproductive ages, 15 to 50 years. Indeed, the mean age-effect on hazard rate for the mutation with shape parameter 6.00 is 120 years. The mutation with shape parameter 1.125, on the other hand, has quite large effects even at younger reproductive ages. For example, the increase in the cumulative hazard function due to a single copy of a mutation with shape parameter 1.125 is nearly 0.5η by age 30 and nearly 0.8η by the age of 50. Because the force of selection decreases with age, we expect the typical genotype to contain substantially more mutations with large shape parameters than mutations with smaller shape parameters.

Table 3.11 lists the parameters for the test case with four gamma mutations. The MH algorithm and the MTM algorithm were both run twice on the same mutation space with the same demographic parameters (such as background hazard rate and mutation rate) but with different algorithm parameters. Trial 1 refers to the first set of algorithm parameters; Trial 2 refers to the second set of parameters. For the MH algorithm, the exponential parameter (used in choosing the number of copies of each mutation type in the proposed genotype) was set to 0.5 for the first trial and 0.7 for the second. For the MTM algorithm, the probability of deleting a chosen mutation in the current genotype ("DelP") was 0.75 for the first trial and 1.0 for the second. Details of the algorithm parameters can be found in §3.2.2 (for the description of the MH algorithm) and §3.2.3 (for the description of the MTM algorithm).

As before, the fertility rate used in the SEW model was set to the fertility rate under the free recombination model applied to the same parameters and mutation space. The output

from the free recombination model for the four gamma mutations case, including the fertility rate and the intensity measure, can be found in Table 3.12. Notice that the intensity measure for the free recombination model, which describes the mean number of each mutation type in a genotype, follows the expected pattern. That is, on average there are substantially more copies of the mutations with larger shape parameters than there are copies of mutations with smaller shape parameters. Indeed, the expected number of copies of the fourth mutation type (with shape parameter 6.00) is much larger than for any of the other mutation types (190 vs fewer than 14 for the other three combined). While the question of the similarity of the distribution of genotypes under the two models will not be addressed until Chapter 4, this huge difference in the expected number of copies of different mutation types is something that appears to cause difficulties for both the naive and MH algorithms, as we shall now see.

Table 3.11: Parameters for the test case with four gamma mutations.

All Tests	η	λ	α	β	Dx	$\nu(\mathcal{M})$	Gamma rate
	0.1	0.05	15	50	0.1	0.05	0.05
NH	NumEv	NumTrials					
	1000	10000					
MH Trial	Burn	Samples	Parameter				
1	10000	100000	0.5				
2	10000	100000	0.7				
MTM Trial	Burn	Samples	DelP	Kmax			
1	10000	100000	0.75	5			
2	10000	100000	1.0	5			

Figure 3.11 shows the histograms for the total number of mutations estimated from the three algorithms. The axes on the five histograms are the same to enable a direct comparison of the results. The most obvious observation is that the three algorithms (naive, top row;

Table 3.12: Results from the free recombination shortcut algorithm run with the parameters listed in Table 3.11.

Iterations	Fertility	Gamma shape parameter	Rho
11	0.06378240	1.125	0.362802
		2.25	1.17842
		4.125	12.1020
		6.00	190.108

Table 3.13: Ratio of proposed moves to accepted moves for the MH and MTM algorithms under both sets of parameters.

Trial Number	MH Algorithm	MTM Algorithm
1	0.00289	0.58550
2	0.00085	0.57889

Table 3.14: Net reproduction ratio (*NRR*) computed using the results of the algorithms for the four gamma mutations case.

Algorithm	Trial	*NRR*
Naive	–	1.00397
MH	1	1.03015
MH	2	1.03969
MTM	1	1.00139
MTM	2	1.00397

MH, center row; MTM, bottom row) all produce vastly different histograms. Indeed, even the two trials with the MH algorithm produce very different histograms from each other. For example, using the naive algorithm, the mean number of mutations per genotype is 14.042 with a variance of 5.725. For the MH algorithm, Trial 1, the mean is 153.182 mutations with a variance of 83.192; for Trial 2, the mean is much smaller, 91.987 mutations with a variance of 9.431 mutations. For the MTM algorithm, the mean is 170.106 mutations with a variance of 181.294 for Trial 1 and the mean is 174.466 with a variance of 181.584 mutations for Trial 2. Unsurprisingly, the expected population survival functions, shown in Figure 3.12 are also noticeably different. Given the relatively close agreement between the distributions generated by the three algorithms in the test cases with one or two point-mass mutations, this discrepancy seems unlikely to be caused by a simple programming error. Rather, it is probably caused by issues inherent in the different approaches.

To delve into this issue more thoroughly, we will consider several ordered genotypes and compute the associated factors $\tilde{P}(\vec{g})$ given by the expression

$$\begin{aligned}\tilde{P}\mathbf{1}_{\vec{g}} =& \frac{\nu(\mathcal{M})^N}{S(m^{(1)})S(m^{(1)}m^{(2)})\cdots S(m^{(1)}m^{(2)}\cdots m^{(N)})} P_{\Pi}\left(m^{(1)}m^{(2)}\cdots m^{(N)}\right) \\ =& \frac{\nu(\mathcal{M})^N}{S(m^{(1)})S(m^{(1)}m^{(2)})\cdots S(m^{(1)}m^{(2)}\cdots m^{(N)})} \frac{\nu\left(m^{(1)}\right)\cdots\nu\left(m^{(N)}\right)}{\nu(\mathcal{M})^N}.\end{aligned}$$

The $\tilde{P}(\vec{g})$ are not probabilities as they have not been normalized. However, the probability of the ordered genotype is *proportional* to $\tilde{P}(\vec{g})$. Table 3.15 lists several ordered genotypes,

ranging from those containing a single mutation to several genotypes with 180 mutations. Computations were done in python using 128-bit floating point precision. The mutation m_1 (or simply 1) refers to the mutation with gamma shape parameter 1.125, m_2 (or 2) to shape parameter 2.25, m_3 (or 3) to 4.125, and m_4 (or 4) to 6.00. As expected, when considering genotypes containing a single mutation, the unnormalized probability $\tilde{P}(g)$ is largest for the genotype with a single copy of m_4 and smallest for the genotype with a single copy of m_1. This result is perfectly in line with our understanding of the force of natural selection decreasing with age. Less selective pressure against mutation type m_4 translates to a better chance of having a copy of mutation m_4. What is more interesting is to note the difference in scale in the $\tilde{P}(\vec{g})$ for these four genotypes. Specifically, we find that the unnormalized probabilities for these four genotypes are all within two orders of magnitude, ranging from 0.3 to 76.9. While this may seem like a large difference, the ordered genotypes containing 180 mutations tell us otherwise.

In addition to the four genotypes with a single mutation, Table 3.15 lists a randomly generated ordered genotype with twelve mutations, a randomly generated genotype with 180 mutations, the same genotype with mutations in descending order, the genotype with 180 copies of m_4, the genotype with 179 copies of m_4 followed by one copy of m_1 and the genotype with one copy of m_1 followed by 179 copies of m_4. These cases with 180 mutations show us that the difference in scale between the unnormalized probabilities can be hundreds of orders of magnitude. We have chosen to focus on genotypes containing 180 mutations because the results from the MTM trials, if they are to be believed, suggest that typical genotypes will contain roughly 160-185 mutations.

Knowing that the force of selection decreases with age tells us that the most likely genotype with 180 mutations will be the genotype with 180 copies of m_4. The unnormalized probability of that ordered genotype is roughly 7.4×10^{65}. Although not quite as likely, the ordered genotype with 179 copies of m_4 followed by a single copy of m_1 is also very likely, with an unnormalized probability of 2.1×10^{65}. However, by moving the copy of the mutation m_1 to the beginning of the time-ordered genotype, the unnormalized probability drops drastically to 2.3×10^{-99}. Of course we have chosen to look at these genotypes knowing that the selective pressure against one is much larger than the selective pressure against the other. As a result, these are cases in which we already know that the ordered genotype will be highly likely (or unlikely). The randomly generated genotype with 180 mutations gives us a better sense of a "typical" ordered genotype drawn according to $\nu(\mathcal{M})$, which in this case and all other cases considered in this work is uniformly distributed across all mutations in the space. This randomly generated genotype, however, is the least likely ordered genotype considered in the table. It is significantly less likely than the same genotype in descending order ($\tilde{P} = 6.1 \times 10^{-272}$, vs 5.2×10^{-94}) as well as the randomly generated genotype with only twelve mutations.

Given that the naive algorithm generates genotypes randomly (in this test it randomly generates 1000 mutation events, 10000 times), it is entirely reasonable to believe that the algorithm is simply not generating those few cases that are most likely. The discrepancy

Table 3.15: Several ordered genotypes $\vec{g}$ and $\tilde{P}(\vec{g})$ for each.

Genotype	$\tilde{P}(\vec{g})$
()	1
(1)	0.331528
(2)	1.03786
(3)	10.1387
(4)	76.8756
(2, 2, 1, 2, 4, 2, 4, 3, 2, 3, 2, 2)	2.96721e-09
$(4,2,4,4,1,1,1,2,2,3,3,3,3,4,2,4,4,1,3,4,4,3,2,2,3,3,2,3,1,4,$ $4,4,3,4,4,2,1,1,1,4,3,3,4,4,4,3,2,1,3,2,2,3,2,2,2,3,2,2,1,2,$ $1,3,2,1,2,3,3,4,2,1,3,3,1,3,1,4,3,3,3,1,4,2,2,4,1,2,4,3,2,1,$ $4,1,1,3,4,1,1,4,1,1,3,3,2,3,3,2,3,3,3,4,3,2,4,2,4,4,3,3,3,4,$ $2,3,4,1,4,3,2,1,4,3,3,1,1,1,1,1,1,2,4,4,2,4,3,3,2,4,2,3,3,1,$ $1,3,4,3,1,4,2,4,2,2,2,1,2,3,4,4,3,4,1,3,4,3,2,3,1,4,1,2,3,3)$	6.12663e-272
$(4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,$ $4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,$ $3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,$ $3,3,3,3,3,3,3,3,3,3,3,3,3,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,$ $2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,$ $1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)$	5.22111e-94
(180 copies of m_4)	7.42234e+65
(179 copies of m_4, m_1)	2.05789e+65
(m_1, 179 copies of m_4)	2.34470e-99

seen in the MH algorithm output is due to a similar problem. A new genotype is proposed by selecting the number of copies of each mutation type (from a discrete double exponential distribution centered on the current number of copies of that mutation) and then *randomly* ordering those mutations. Randomly ordering the proposal genotypes appears to be resulting in a large number of proposed steps that are unlikely, and thus rejected. This is evident by the extremely low acceptance ratios for the algorithm. In the first trial, the acceptance rate was about 0.3% and it was even lower (about 0.01%) for the second trial.

A second problem is that the number of copies of each type of mutation is updated in every proposal and that the distribution from which the number is chosen has the same rate parameter for all mutation types. Because we expect more copies of mutation m_4 than m_1 it would make more sense to propose bigger changes in the number of copies of m_4 than in the number of copies of m_1. These problems could possibly be alleviated by changing how the proposed states are chosen. Specifically, updating only one mutation type at a time (that is, one mutation type per step) would probably increase the acceptance ratio. Similarly, allowing different rates for the exponential distribution governing the number of copies of each mutation type would probably also increase the number of proposed genotypes accepted. These ideas were deemed to have too little virtue to be worth the time needed to implement them, however, as the overall acceptance rate would probably still be low due to the fact that some mutations (such as m_1, m_2, and m_3 in this case) may have a very small presence in the population. Even after tuning the additional parameters needed to ensure reasonably sized steps for each mutation type, the issue of randomly ordering the proposed genotypes would remain problematic.

It is worth noting that the discrepancy between the output of the three algorithms when $\mathcal{M}$ consists of four gamma mutations is not a result of the fact that the mutation profiles changed from point-mass to gamma distributions. To ensure that the discrepancy is due to having a larger mutation space rather than different mutation profiles, we ran all three algorithms on a smaller mutation space containing gamma mutations. For this test, the mutation space had only two types of gamma mutations with the same rate parameter of 0.05 and shape parameters 1.125 and 2.25. The three algorithms all produce similar histograms for the total number of mutations (see Figure 3.13).

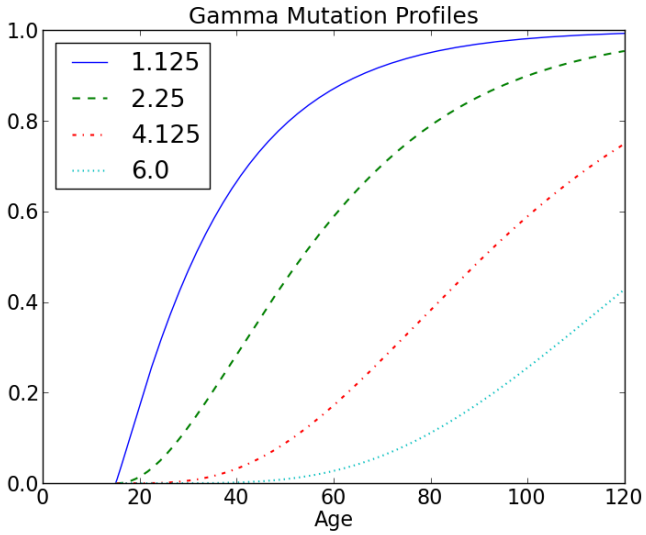

Figure 3.10: Gamma mutation profiles for the four gamma mutations case. All four mutations have the same rate parameter of 0.05 but different shape parameters.

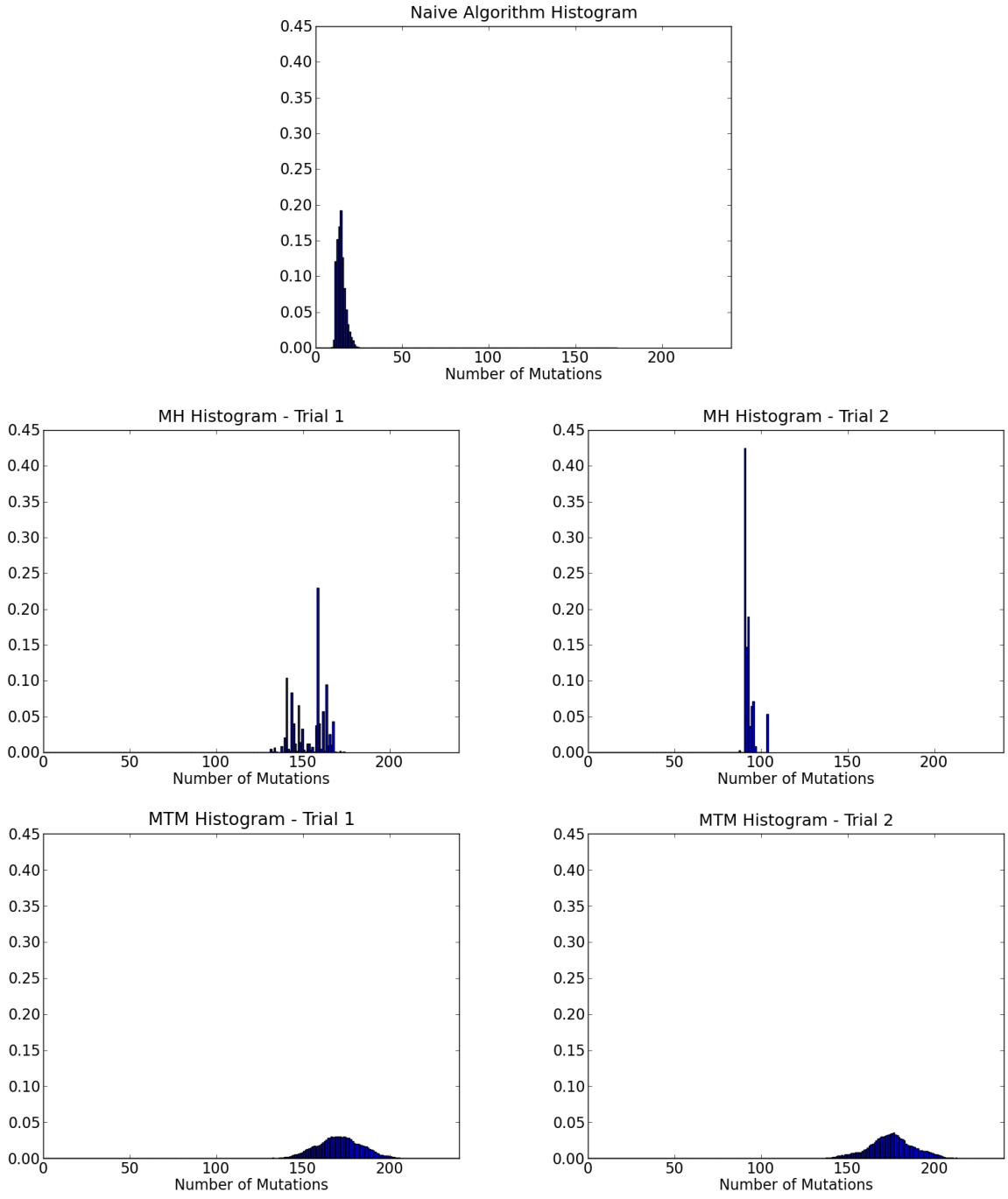

Figure 3.11: Histograms for the four gamma mutations case. Naive algorithm (top), MH Trial 1 (second row, left), MH Trial 2 (second row, right), MTM Trial 1 (bottom row, left), MTM Trial 2 (bottom row, right).

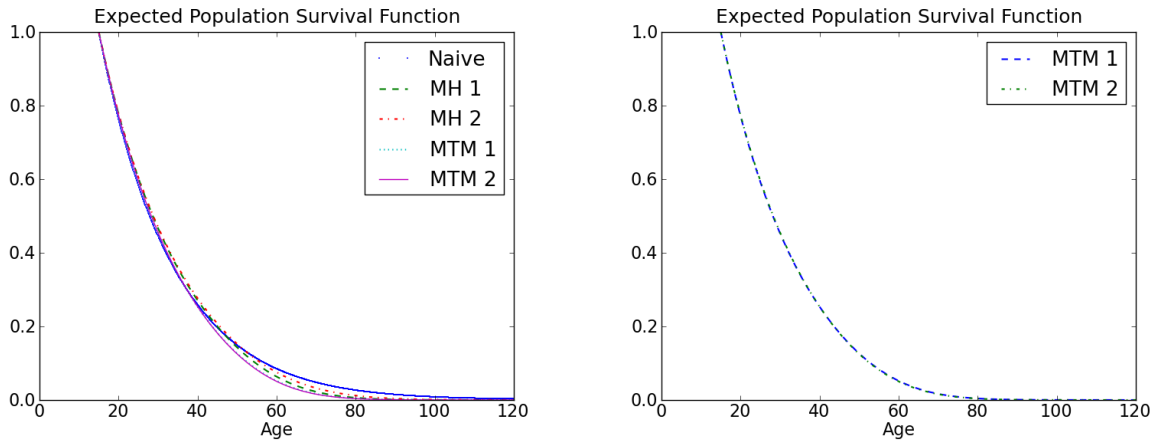

Figure 3.12: Expected population survival for the four gamma mutations case. The plot on the left shows the expected survival from all five cases considered originally (naive algorithm, MH and MTM). The plot on the right reproduces only the expected survival from the two trials of the MTM algorithm.

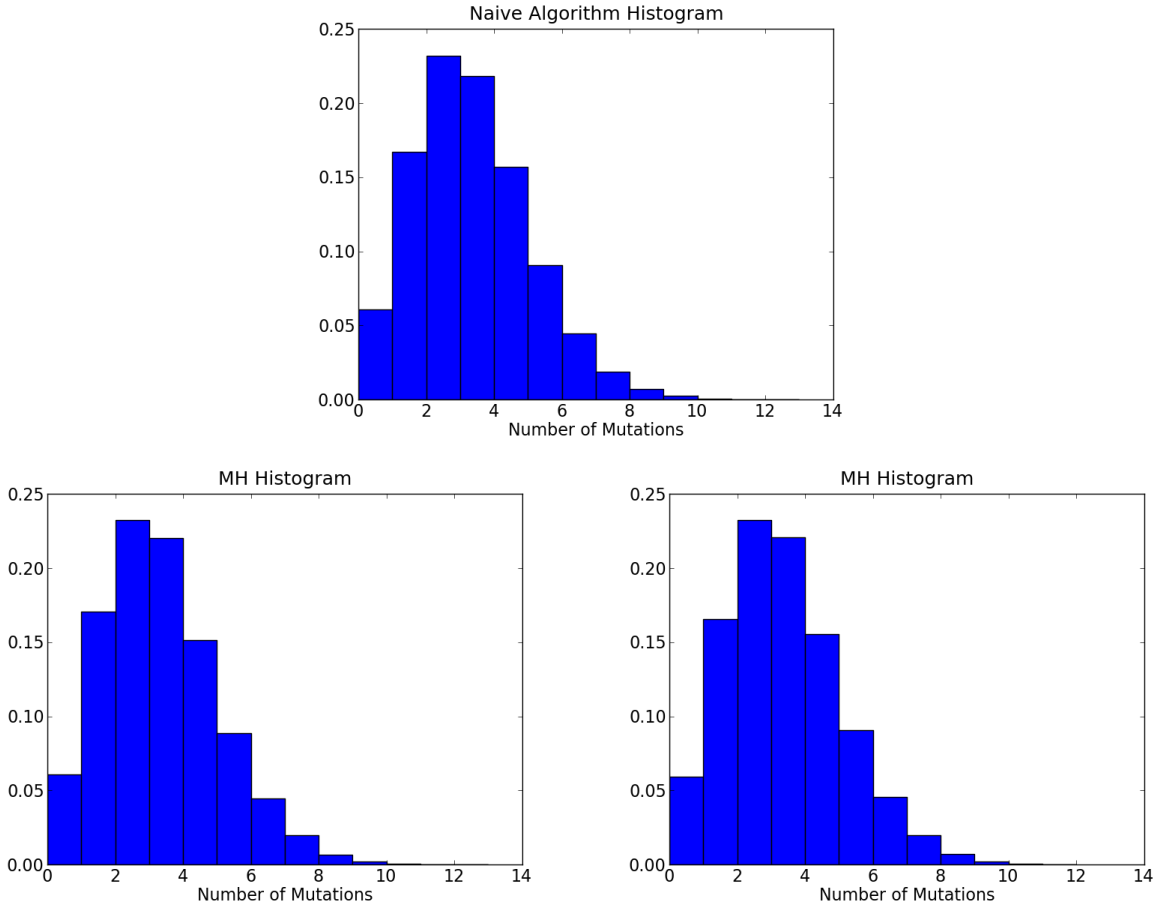

Figure 3.13: Histograms for the two gamma mutations case.

3.3.4 Comments

The ultimate goal of this work is to shed light on the distribution of genotypes under the SEW model. Specifically, we are interested in applications to senescence, which current research suggests may be caused by a large number of slightly deleterious mutations. However, before we move on to consider cases with a large number of possible mutations, it is worthwhile to spend a moment reflecting on the simple, small mutation space cases considered so far.

In the first place, it is clear that the MH algorithm, implemented as described in §3.2.2, has an acceptance ratio that is too low to be useful in the larger cases we will consider next. Similarly, the naive algorithm has also been shown to fail to generate likely genotypes even in the relatively small case of four possible mutations. While it is the most exact method for estimating the distribution of genotypes for very small mutation spaces (consisting of only one or two types of mutations), it will be terribly inefficient and misleading when applied to larger spaces. As a result, all further tests will rely on the MTM algorithm, which had acceptance rates of 58% (compared to less than 1% for the MH algorithm) for the case with four gamma mutations.

In the second place, it is worth mentioning that the number of mutations per genotype under the SEW model in these small test cases can be very different from the average genotype genotype length under the free recombination model. Table 3.16 shows the estimated mean and variance for the three single point-mass mutation cases discussed in §3.3.1. Comparing the mean number of mutations per genotype from Table 3.16 to ρ (the mean number of mutations under the free recombination model) in Table 3.1, it is clear that there is a greater accumulation of mutations under the free recombination model than under the SEW mutation-selection model, which has no recombination. For example, with an age of onset of 25 and a total mutation rate of 0.3, the average genotype under the free recombination model will have 12.7 mutations to the 4.7 copies when there is no recombination. Similarly, when the age of onset of the mutation is 45 and the total mutation rate is 0.02, the typical genotype under free recombination will have 6 mutations compared to 4.2 copies if there is no recombination.

Similarly, we can compare the mean number of mutations per genotype under the two models for the cases with two point-mass mutations, discussed in §3.3.2. Table 3.17 shows the estimated mean and variance for the total number of mutations in the two point-mass mutations cases while Table 3.18 show the mean and variance for each type of mutation. In the case with ages of onset $m_1 = 20$ and $m_2 = 30$, under free recombination, the average genotype would have 1.57 copies of m_1 and 3.9 copies of m_2 (see Table 3.6 for ρ for these cases). With no recombination, a typical genotype has fewer copies of each type of mutation: 1.3 copies of m_1 and 2.75 copies of m_2. When the ages of onset for the two mutations are $m_1 = 20$ and $m_2 = 40$, the difference between the mean number of mutations between the two models is more pronounced. Under the free recombination model, a genotype contains an average of 0.75 copies of m_1 and 7.3 copies of m_2. With no recombination, however, a typical genotype has roughly 0.7 copies of m_1 but only 4.17 copies of m_2.

Finally, we wish to emphasize that many of the histograms for the number of mutations per genotype in the previous sections *look* approximately Poisson. We can make this statement more concrete by analyzing the differences between the means and the variances for the various test cases considered so far. Table 3.16 provides the mean and variance in the number of mutations per genotype for the single point-mass mutation cases and Tables 3.17 and 3.18 provide the means and variances in the two point-mass mutations cases. If the distribution of genotypes is a Poisson random measure (or can be well approximated by a Poisson), then the mean number of mutations per genotype should be approximately equal to the variance in the number of mutations per genotype. For the cases with two mutations, we should also find that the average number of each type of mutation should be roughly the same as the variance in the number of that type of mutation per genotype.

To test the difference between the means and variances, we use the Poisson dispersion test, also called the variance test. Under the null hypothesis that the data X_i come from a Poisson distribution, the test statistic

$$D = \sum_i \frac{(X_i - \bar{X})^2}{\bar{X}} = \frac{(n-1)S^2}{\bar{X}}$$

has a chi-squared distribution with $n - 1$ degrees of freedom. Table 3.19 show the test statistics and p-values for the dispersion test applied to each case. For the cases involving mutation spaces with two mutations, we test the total number of mutations per genotype and the number for each type of mutation. A small p-value indicates that we reject the null hypothesis that the data comes from a Poisson distribution. The test statistics and p-values reported use all the samples collected after the burn-in period even though these samples are not independent.

With the large number of samples (10,000 samples for the naive algorithm and 100,000 samples for the MH and MTM algorithms) even small differences between the mean and variance can be statistically significant. However, while the difference between the mean and variance may be statistically significant, it may not be *practically* different. Take, for example, the two point-mass mutations case with ages of onset $m_1 = 20$ and $m_2 = 30$. The average number of copies of m_1 is 1.3 while the variance is about 1.4. For m_2, the average is 2.76 with a variance of 3.2. Because we are interested in demographic outcomes, such as the expected population survival function and the hazard rate, these differences, while statistically significant, may not have much impact on the outcomes we wish to measure. If this observation holds for the larger test cases considered next, it would suggest that although the distribution of genotypes under the SEW model is not Poisson, it may be approximately Poisson. Or rather, it may be reasonably approximated by a Poisson in light of demographic outcomes, such as lifespan.

Table 3.16: Estimated mean and variance for the number of mutations per genotype in the single point-mass mutation cases.

	Actual		Naive Algorithm	
Trial	Mean	Var	Mean	Var
1	4.77669	6.12061	4.76244	6.09767
2	4.86328	6.26069	4.84358	6.22872
3	4.19290	5.20710	4.15403	5.14815
	MH Algorithm		MTM Algorithm	
Trial	Mean	Var	Mean	Var
1	4.79212	6.17943	4.74764	6.18059
2	4.85405	6.14073	4.86386	6.33499
3	4.1665	5.09810	4.16261	5.00479

Table 3.17: Estimated mean and variance for the total number of mutations in the two point-mass mutations cases.

	Naive Algorithm		MH Algorithm		MTM Algorithm	
Trial	Mean	Var	Mean	Var	Mean	Var
4	4.09013	4.85732	4.11066	4.91203	4.0984	4.86112
5	4.86469	5.84422	4.86484	6.04809	4.87237	5.96136

Table 3.18: Estimated mean and variance for the each mutation type in the two point-mass mutations cases.

		Naive Algorithm		MH Algorithm		MTM Algorithm	
Trial	MutAge	Mean	Var	Mean	Var	Mean	Var
4	20	1.33538	1.43392	1.31847	1.39301	1.33911	1.40317
4	30	2.75475	3.16186	2.79219	3.24809	2.75929	3.20347
5	20	0.695436	0.723934	0.69868	0.742106	0.69908	0.715327
5	40	4.16925	5.01342	4.16616	5.22987	4.17329	5.15158

Table 3.19: Poisson dispersion test results for the single point-mass mutation cases (discussed in §3.3.1) and two point-mass mutations cases (discussed in §3.3.2).

		Naive Algorithm		MH Algorithm		MTM Algorithm	
Trial	MutAge	χ^2	p-value	χ^2	p-value	χ^2	p-value
1	25	12812	0	128948	0	130181	0
2	35	12872	0	126506	0	130244	0
3	45	12417	0	122358	0	120230	0
4	(20,30)	11874	0	119493	0	118608	0
4	20	10736	1.70e-07	105652	0	104783	0
4	30	11476	0	116326	0	116096	0
5	(20,40)	12012	0	124321	0	122349	0
5	20	10408	2.09e-03	106214	0	102323	1.24e-07
5	40	12023	0	125530	0	123440	0

Chapter 4

Large Mutation Spaces

The purpose of this chapter is twofold. First, we wish to characterize the distribution of genotypes under the SEW model in cases with a large numbers of mutations. In particular, we want to determine how similar or dissimilar this distribution is to a Poisson random measure and whether a Poisson approximation can be used in estimating demographic outcomes such as expected survival and population hazard rates. Second, we will explore how similar the distribution of genotypes and demographic outcomes are under the SEW mutation-selection and ESW free recombination models. To this end, we begin by considering large spaces of mutations with gamma profiles.

4.1 Mutations with Gamma Profiles

Each mutation in the space $\mathcal{M}$ has the same rate parameter of 0.05 but different shape parameters. In the cases considered below, the shape parameters range from 1.0 to ξ, where ξ is between 5 and 7. In all of these cases the mutation space is composed of 1000 mutations with equally-spaced shape parameters. These parameters were chosen in line with tests from [37], which discusses the behavior of the free recombination model with gamma mutations.

Table 4.1 lists the parameters used in four cases involving gamma mutations. The background hazard rate, λ, and the mutation effect size, η, are the same for all four cases and were set to biologically reasonable values. The fertility rate, f_x, is assumed to be constant from the age of maturity, $\alpha = 15$ years, to the oldest age of reproduction, $\beta = 50$ years. This rate was determined using the shortcut algorithm for the free recombination model applied to the same mutation space and with the same background hazard rate and mutation rate. The number of iterations before convergence of the shortcut algorithm and the resulting fertility rates for the four cases are shown in Table 4.2.

As we mentioned in the previous chapter, the fertility rate is tuned at each iteration of the shortcut algorithm to ensure that the net reproduction ratio (NRR) is equal to one.

Using the fertility rate from the free recombination model does not, however, ensure that the net reproduction ratio will be equal to one under the SEW mutation-selection model. The expected survival function under the mutation-selection model is estimated by averaging over the survival function for each sampled genotype obtained by the MTM algorithm. To ensure that the *NRR* is equal to one, we would need to rescale the fertility by the factor $1/NRR$. The approximation to the *NRR* for the four cases under the SEW model are shown in Table 4.3. In all four cases the *NRR* is close to 1.0, with the smallest *NRR* of 0.978 belonging to Case 3 and the largest *NRR* of 1.06 belonging to Case 2. This indicates that the fertility rate for the free recombination model with the same parameters and mutation space is quite similar to the fertility rate necessary to ensure a stationary population under the SEW model.

Three of the four cases use the same value for DelP, the probability of deleting a chosen mutation from the current genotype (see §3.2.3 for details regarding the multiple-try Metropolis algorithm). The choice of DelP does have an effect on how quickly the chain traverses the space of possible genotypes. For example, we must have DelP > 0; otherwise, the proposed genotype could never contain fewer mutations than the current genotype and the associated Markov chain would not be reversible. With a very small value for DelP, DelP ≈ 0, the proposed genotypes would almost always have at least as many mutations as the current genotype. In this case, it could be possible for the chain to become trapped in extremely low-probability genotypes containing a large number of mutations. On the other end of the spectrum, when the deletion probability is one, the proposed genotype will contain exactly one more or one fewer mutation than the current genotype; the algorithm will never propose a genotype that has the same number of mutations. This means that for the chain to move from the genotype $g = m_1$ to the genotype $g = m_2$ the chain must take at least two steps. For example, the chain could move from $m_1 \to () \to m_2$ (where () represents the null or wild-type genotype). Alternatively, the chain could move from $m_1 \to m_1 + m_2 \to m_2$.

The values for DelP used in all of the large mutation space experiments presented here are 0.375 and 0.5. These values were chosen because they produced reasonable acceptance ratios, 30% to 38%, for Case 4, in which $\xi = 5$. Eight independent chains were started in the null state (a wild-type genotype) using different values for DelP (0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875 and 1.0). All eight chains produced similar histograms for the total number of mutations, as well as similar estimates for the average number of mutations per genotype. These plots are included in Appendix B. Smaller values for DelP produced higher acceptance rates than values near 1.0. Notice that theses values for the deletion probability are much lower than those used in the test cases with four or fewer mutations. The low number of mutation types (as well as the generally small number of mutations in a typical genotype) in those test cases suggested that adding and deleting mutations in a genotype would allow the chain to traverse the space faster than changing the mutation type. In cases with very large mutation spaces (and large numbers of mutations in a typical genotype), there are many more possible genotypes with the same total number of mutations. This makes changing mutation types very appealing. As a result, for large mutation spaces, it is

necessary to choose a value for DelP that balances exploration of genotypes with the same total number of mutations against moving to genotypes with a different total number of mutations.

The acceptance ratios for the MTM runs are also provided in Table 4.3. The rate of acceptance varies among the four cases with a low of 34% for Case 3 and a high of 44% for Case 1. In collecting samples for the four cases, each chain was started in the state representing the null genotype. The chains in the four cases were run for a different number of steps before collecting the samples. This is due to the fact that some of the chains appeared to converge to the genotype distribution more quickly than other chains. In particular, the cases with larger shape parameters (such as Case 3, where $\xi = 7$) appeared to require longer burn-in periods and more samples overall than the cases with smaller shape parameters. A discussion of the MCMC convergence diagnostics applied to the data for these four cases can be found in Appendix A.1.

Table 4.1: Parameters for the test cases with 1000 gamma mutations with shape parameters from 1.0 to ξ (inclusive).

All Tests	η	λ	α	β	Dx	Gamma rate
	0.1	0.05	15	50	0.5	0.05
MTM Case	ξ	$\nu(\mathcal{M})$	Burn	Samples	DelP	Kmax
1	6	0.15	150000	1000000	0.375	5
2	5.5	0.17	150000	750000	0.375	5
3	7	0.12	250000	2250000	0.5	5
4	5	0.12	150000	550000	0.375	5

Table 4.2: Output from shortcut algorithm for the free recombination model. The four test cases considered have mutation spaces with 1000 gamma mutations with shape parameters from 1.0 to ξ (inclusive).

Case	ξ	Fertility	Iterations
1	6	0.07330193	25
2	5.5	0.07537674	26
3	7	0.07036571	25
4	5	0.069345238	13

Figure 4.1 shows the histograms for the total number of mutations per genotype for the four cases. The histogram corresponding to Case 4 is the most symmetric of the four histograms. Case 4 has both the smallest mutation rate and the shortest interval of shape

Table 4.3: Output from the MTM algorithm for the SEW model under test cases with 1000 gamma mutations with shape parameters from 1.0 to ξ (inclusive).

Case	ξ	Acceptance Rate	*NRR*
1	6.0	0.442601	1.01993
2	5.5	0.398316	1.06021
3	7.0	0.3438732	0.978127
4	5.0	0.382936	1.02591

parameters for the four test cases, with shape parameters ranging from 1.0 to 5.0. Because shape parameters near 1.0 correspond to mutations that have larger early-age effects and because the strength of selection decreases with age, we expect more of the mutations in Case 4 to face heavy selection. In the other cases considered here, the shape parameters have a larger range, meaning that a larger proportion of the possible mutations have small early-age effects and face less selective pressure. Case 4 also has the lowest number of mutations in a typical genotype. This is due both to the low mutation rate and the fact that it has the highest proportion of mutations with large early-age effects on the cumulative hazard. For example, from Table 4.4 we see that less than 1/3 of the mutations in an "average" genotype for Case 4 have shape parameters in the range 1.0 to 4.0.

Case 2 provides a nice comparison to Case 4 because the range of shape parameters in Case 2, 1.0 to 5.5, is similar to the range in Case 4. However, while these cases have similar shape parameter ranges, the overall mutation rate for Case 4 is much smaller than it is for Case 2: 0.12 for Case 4 whereas it is 0.17 for Case 2. Although the histogram for Case 2 is generally symmetric, it is less symmetric than the histogram for Case 4. Common genotypes under Case 2 also have substantially more mutations, from 100 to 180, than do common genotypes under Case 4, which contain 40 to 90 mutations.

Cases 1 and 3 have the largest range of shape parameters. They also have the least symmetric histograms for the total number of mutations. However, although the histograms in Cases 1 and 3 are not as symmetric as those for Cases 2 and 4, they are approximately symmetric. Genotypes in these cases also contain more mutations than in Cases 2 and 4. For example, common genotypes in Case 3 contain roughly 250 to 450 mutations. For Case 1, common genotypes contain 140 to 220 mutations. Although the number of mutations per genotype in Case 1 is similar to the number of mutations per genotype in Case 2, it is important to remember that the mutation rate in Case 1 is lower than the mutation rate in Case 2. That means that there are more mutations being introduced into the population in Case 2 than in Case 1. However, because the shape parameter range in Case 2 is smaller than in Case 1, more of the possible mutations in Case 2 face higher selective pressure than do mutations in Case 1. This results in a smaller number of mutations per genotype for Case 2, even though it has a higher mutation rate.

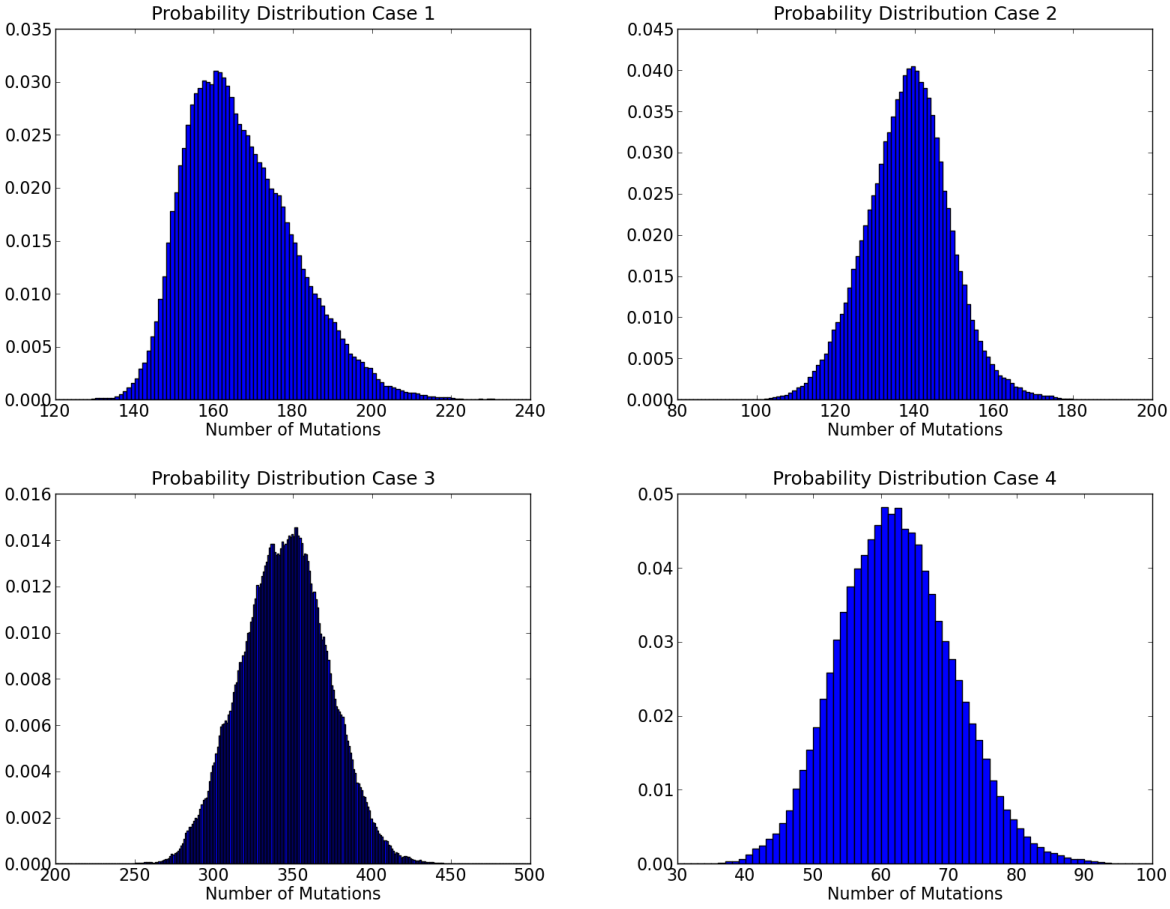

Figure 4.1: Histograms for the total number of mutations per genotype for the four cases with gamma mutations with shape parameters from 1.0 to ξ.

4.1.1 Similarity to a Poisson Random Measure

In addition to producing the most symmetric histogram of the four cases, Case 4 may also be the most similar to a Poisson random measure. For example, consider Table 4.4, which lists the mean and variance in the total number of mutations per genotype for each of the four cases. We note that the mean and variance in the total number of mutations per genotype are not that close; the variance in the total number of mutations per genotype is 70.4, which is about 14% higher than the mean of 61.8. However, when we compare the means and variances in the number of mutations with shape parameters in a smaller range, say $1.5 < m \leq 2.0$ or $3.5 < m \leq 4.0$, we see that the difference is much smaller. However, in only one of the seven subintervals listed, does the marginal mean and variance actually pass the dispersion test. This means that, in general, the difference between the mean and the variance cannot be explained by chance.

Although the difference between the mean and variance in each subinterval of shape parameters is statistically significant, the actual difference tends to be quite small. In all but the last interval, where the shape parameters range from 4.5 to 5.0, the variance is within about 3% of the mean and in many of these subintervals it is within 1%. The generally close agreement between the average and the variance in the number of mutations within a subinterval of shape parameters is most evident when considering the dispersion statistics (listed in the χ^2 column) and their associated p-values. Because there are the same number of observations in each shape interval within a given case, we can directly compare the dispersion statistics and the p-values for different shape intervals within a given case. Looking at Case 4, we find the smallest p-value is associated with the largest shape parameters, $4.5 < m \leq 5.0$. The table suggests that, for Case 4 at least, the number of mutations per genotype may be approximately Poisson distributed within small shape parameter intervals and that this approximation is best when the shape parameters are less than 4.5. The distribution of the number of mutations seems less similar to a Poisson distribution when the shape parameters are above 4.5.

The trend where the mean and variance are most similar for small shape parameters is not readily apparent in the other cases. Consider, for example, Case 2. The closest agreement between the mean and the variance is in the subinterval of shape parameters with $4.5 < m \leq 5.0$, where the observed difference is not statistically significant. In all other subintervals, the difference is statistically significant, with p-values that are essentially 0. Furthermore, in almost every subinterval (excepting $1.0 \leq m \leq 1.5$ and $4.5 < m \leq 5.0$), the variance is 12-20% higher than the mean in that interval. Similarly, for Case 1, the variances are almost all within 5-15% of the mean, with no clear pattern. However, while it is not obvious from the p-values, we do find the same pattern in Case 3. That is, the variance appears to be closer to the mean (relative to the size of the mean) for smaller shape parameters than for larger shape parameters. For the subintervals with shape parameters larger than 5.0, the variance is 17-40% larger than the mean, whereas it is within about 6% for subintervals with smaller shape parameters.

The sometimes large difference observed between the mean and variance in the subintervals of shape parameters presented in Table 4.4 could be due to the difficulty of sampling from the genotype distribution under the SEW model. Although several diagnostic criteria were applied to each case (see Appendix A.1 for details), it is possible that even the large number of steps used to generate these data is inadequate to sufficiently explore the space of possible genotypes when there are so many different types of mutations. However, when we compare the data in Table 4.4 to means and variances estimated using fewer samples, we find that although the means and variances are different, the differences are generally small. An example is included in Appendix A.1. This suggests that more samples would not substantially change the conclusions that we will draw from the data presented here.

As noted previously, the distribution of the number of mutations per genotype may be approximately Poisson even though it is not actually Poisson. Indeed, the similarity between the means and the variances in the subintervals considered above suggest that a Poisson approximation may be reasonable. In order to get a better sense of how similar to a Poisson random measure the genotype distribution is under the SEW model, we turn to a finer partition of the mutation space. Rather than looking at intervals of mutation types, we now look at each mutation individually. If the distribution of genotypes were in fact a Poisson random measure, then the number of copies of two different mutations, say m_1 and m_2, would be independent and Poisson distributed with means $\rho(m_1)$ and $\rho(m_2)$, respectively. Furthermore, the number of copies of m_1 or m_2 would be Poisson distributed with mean $\rho(m_1) + \rho(m_2)$. As a result, we can compare the sum of the marginal variances (the variance in the number of copies of m_i for each i) to the variance in the cumulative sum of the mutations. That is, we compare the sum of the variances,

$$SV_k = \sum_{i=1}^{k} \text{Var}(\text{number of copies of } m_i)$$

to the cumulative variance,

$$CV_k = \text{Var}(\text{number of mutations } m \in \{m_1, \ldots m_k\}).$$

For a Poisson random measure, of course, these two numbers should be the same.

Figures 4.2 and 4.3 show the sum of the variances, the cumulative variance and the cumulative sum of the means for each of the four cases. As expected from the data provided in the table, the sum of the marginal variances is very similar to the sum of the marginal means. The cumulative variance, however, is quite different from the sum of the variances, particularly for large shape parameters. The difference between the cumulative variance and the sum of the variances indicates that there are large correlations between the number of different mutation types. This observation is not terribly surprising because the SEW mutation-selection model does not include genetic recombination, which breaks statistical dependencies between loci in the genome. As expected, the cumulative variance and the sum of the variances are closest in Case 4. In that case, the cumulative variance is quite close

Table 4.4: Estimated means and variances for Cases 1-4. The marginal means and variances correspond to the number of mutations per genotype with shape parameters (m) in the given intervals.

	Case 1				Case 2			
	Mean	Variance	χ^2	p-value	Mean	Variance	χ^2	p-value
$1.0 \leq m \leq \xi$	166.334	190.679	1146360	0	137.787	113.454	617552	0
$1.0 \leq m \leq 1.5$	0.942218	1.02505	1087912	0	0.753552	0.795316	791565	6.62e-244
$1.5 < m \leq 2.0$	1.51251	1.59733	1056075	0	1.29558	1.53196	886837	0
$2.0 < m \leq 2.5$	2.40218	2.53109	1053664	1.21e-304	2.14219	2.56360	897538	0
$2.5 < m \leq 3.0$	3.92265	3.98447	1015758	7.34e-29	3.57661	4.24591	890350	0
$3.0 < m \leq 3.5$	6.32708	6.94444	1097573	0	6.27615	7.22210	863040	0
$3.5 < m \leq 4.0$	10.2278	11.0687	1082214	0	10.8454	12.2204	845086	0
$4.0 < m \leq 4.5$	16.9770	17.9433	1056916	0	18.9146	21.8231	865325	0
$4.5 < m \leq 5.0$	27.0789	27.1953	1004295	1.20e-3	34.7338	34.7074	749428	3.21e-001
$5.0 < m \leq 5.5$	41.3440	45.5066	1100682	0	59.2490	66.4545	841209	0
$5.5 < m \leq 6.0$	55.5995	64.2969	1156426	0	–	–	–	–
	Case 3				Case 4			
	Mean	Variance	χ^2	p-value	Mean	Variance	χ^2	p-value
$1.0 \leq m \leq \xi$	344.318	757.616	4950758	0	61.7655	70.3822	626727	0
$1.0 \leq m \leq 1.5$	1.00961	1.02729	2289401	1.89e-76	0.483436	0.474935	540327	8.84e-21
$1.5 < m \leq 2.0$	1.59522	1.65655	2336505	0	0.805187	0.812326	554875	1.77e-6
$2.0 < m \leq 2.5$	2.30087	2.38823	2335428	0	1.42565	1.44976	559298	5.81e-19
$2.5 < m \leq 3.0$	3.68019	3.77548	2308257	1.45e-163	2.48581	2.46550	545506	8.76e-6
$3.0 < m \leq 3.5$	6.09614	6.37484	2352866	0	4.27732	4.16385	535408	4.69e-45
$3.5 < m \leq 4.0$	9.36255	9.37602	2253236	6.36e-2	7.92369	7.92560	550132	4.49e-1
$4.0 < m \leq 4.5$	14.9570	15.0115	2258202	5.61e-5	14.9752	14.8780	546425	3.21e-4
$4.5 < m \leq 5.0$	23.3678	25.1011	2416892	0	29.3892	32.4575	607419	0
$5.0 < m \leq 5.5$	37.0537	45.2533	2747896	0	–	–	–	–
$5.5 < m \leq 6.0$	57.2252	72.8376	2863853	0	–	–	–	–
$6.0 < m \leq 6.5$	81.2535	104.864	2903798	0	–	–	–	–
$6.5 < m \leq 7.0$	106.416	149.596	3162961	0	–	–	–	–

to the sum of the variances until around $m = 4.5$, at which point the cumulative variance increases more rapidly than the sum of the variances.

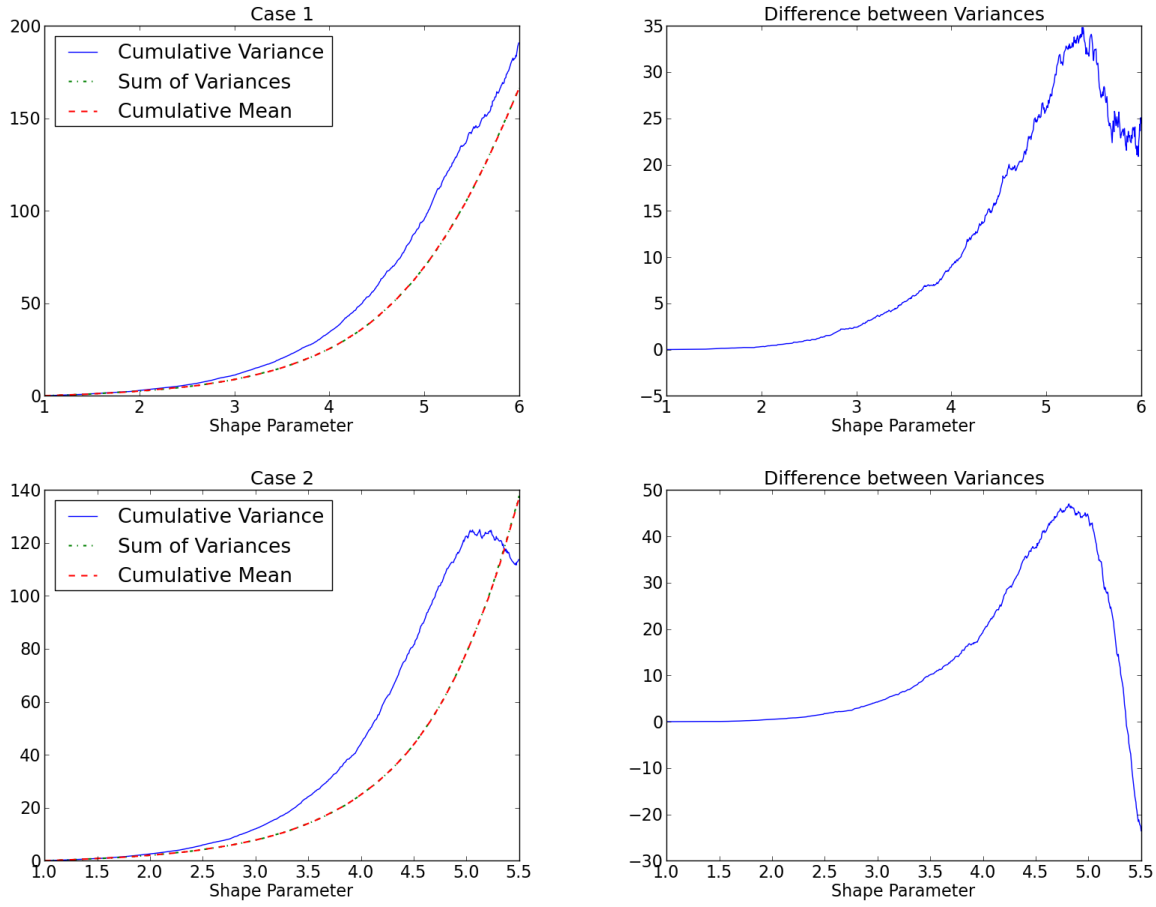

Figure 4.2: The plots on the left show the sum of the variances, the cumulative variance and the cumulative mean for Case 1 (top row) and Case 2 (bottom row). The plots on the right show the difference in the variances for Case 1 (top row) and Case 2 (bottom row).

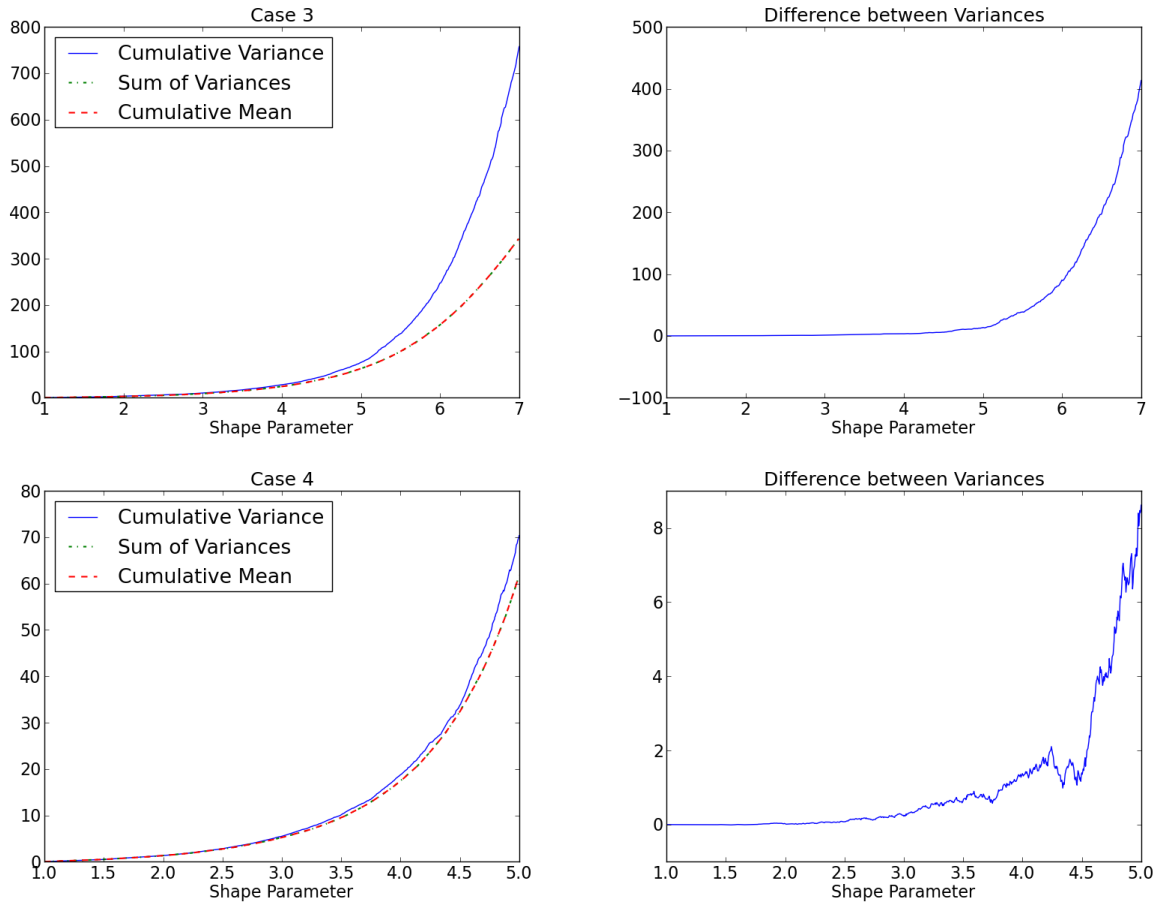

Figure 4.3: The plots on the left show the sum of the variances, the cumulative variance and the cumulative mean for Case 3 (top row) and Case 4 (bottom row). The plots on the right show the difference in the variances for Case 3 (top row) and Case 4 (bottom row).

4.1.2 Approximations using a Poisson Random Measure

The data provided in Table 4.4 suggest that while the distribution of mutations is not a Poisson random measure, it may be possible to approximate it by a Poisson random measure. To this end we estimated the average number of each type of mutation per genotype for the four gamma cases. Figure 4.4 shows the marginal mean number of mutations per genotype for Case 1 (top row) and Case 2 (bottom row), while Figure 4.5 shows the marginal mean number of mutations for Case 3 (top row) and Case 4 (bottom row). The intensity measure ρ, which corresponds to the mean number of each type of mutation under the free recombination model, is also shown for each case.

The most striking element of these plots is that the marginal means data in all four cases appear to be approximately exponential in shape parameter. The appearance of this exponential behavior under the SEW model is of interest because it also appears in the free recombination model for large mutation spaces where mutations have gamma profiles. This behavior was first noted by Wachter, Steinsaltz and Evans in [37]. We present exponential approximations to ρ for the four gamma cases in Appendix C.

Because all four of the large gamma cases show a clear, exponential pattern in the marginal means plot, the means were fitted with an exponential curve using the python package scikits.statsmodels. The model

$$\log(\text{Marginal Mean}) = \alpha\,\text{Shape Parameter} + \beta$$

was fitted using OLS. Table 4.5 shows the resulting parameters.

In the four gamma cases considered, all of the 1000 possible mutations had non-zero means, as shown in Table 4.5 in the "Number of observations" column. The fitted values of α range from 0.881 to 1.18, with larger values of α corresponding to smaller values of ξ. Similarly, the fitted values of β range from -5.53 to -7.17, with -7.17 corresponding to the smallest value of ξ, $\xi = 5$, and -5.53 corresponding to the largest, $\xi = 7.0$. The exponential approximations are overlaid (dotted line) on the marginal means data in Figures 4.4 and 4.5. The exponential fit appears to be best in Cases 2 and 4, which are the cases with the shortest ranges for mutation shape parameters. In Cases 1 and 3, the approximation is good for small shape parameters but overestimates the mean for large shape parameters. In Case 1, the approximation is too high for $m > 5.5$, whereas in Case 3, it is too high for $m > 6.5$.

Because of the strong exponential pattern and the vaguely Poisson behavior observed in the marginal distributions, it is natural to estimate the expected population survival function and the expected population hazard rate under the SEW model by assuming that the resulting distribution is a Poisson random measure whose intensity function is the exponential curve fitted to the marginal means data. Figure 4.6 shows the expected population survival function calculated directly from the samples as well as the Poisson approximation. In general there is a close agreement between the population survival function estimated from the MTM samples and from the Poisson approximation. Table 4.6 shows the ages associated with various survival probabilities computed from the expected population survival function.

The survival function estimated directly from the MTM samples (labeled "Empirical" in the table) and the survival function from the Poisson approximation (labeled "Approx." in the table) produce very similar results. For example, in Case 1, the survival probability 0.5 corresponds to the age 27.0 under both the empirically determined survival function and the Poisson approximation. That means that a randomly chosen individual in the population has a 50% chance of surviving to age 27. Of the probabilities listed in the table, the empirically determined survival function and the Poisson approximation differ by at most half a year. For a different view of the similarity between the empirical survival function and the Poisson approximation, consider Table 4.7, which shows the L^2 distance and the L^∞ distance between these survival functions for Cases 1–4. In all four cases, the L^2 norm of the difference in survival functions is around 0.006. The L^∞ distance ranges from 0.0012 to 0.0017 for the four cases.

Table 4.5: Coefficients from using scikits.statsmodels.OLS to fit the model log(Marginal Means) = α Shape Parameter + β.

Case	ξ	Number of observations	Coefficients	
1	6.0	1000	α	0.940180
			β	-5.85360
2	5.5	1000	α	1.10918
			β	-6.48347
3	7.0	1000	α	0.880654
			β	-5.52569
4	5.0	1000	α	1.18243
			β	-7.17319

The expected population hazard functions for the four cases are shown in Figure 4.7. Recall that the hazard function is the negative rate of change of the log survival function. As expected from the close agreement between the empirical survival function and the survival function using a Poisson approximation, the hazard functions are also in close agreement. In most of the cases, the biggest difference between the empirical hazard function and the Poisson approximation occur at very late ages. The population hazard function under the free recombination model is also shown. In all four cases, the three hazard rates are similar for ages near the age of maturity (age 15) and become more dissimilar at later ages. In particular, the hazard rates under the free recombination model are much higher than the hazard rates under the SEW model. The hazard rates under the two models are most similar in Case 4.

Figure 4.6 shows, and Table 4.7 confirms, that a Poisson approximation to the distribution of genotypes under the SEW model can be used in these cases when the outcome that we wish to estimate is the expected population survival function. The approximation is quite good even in Cases 1 and 3, where the exponential approximation overestimated the average

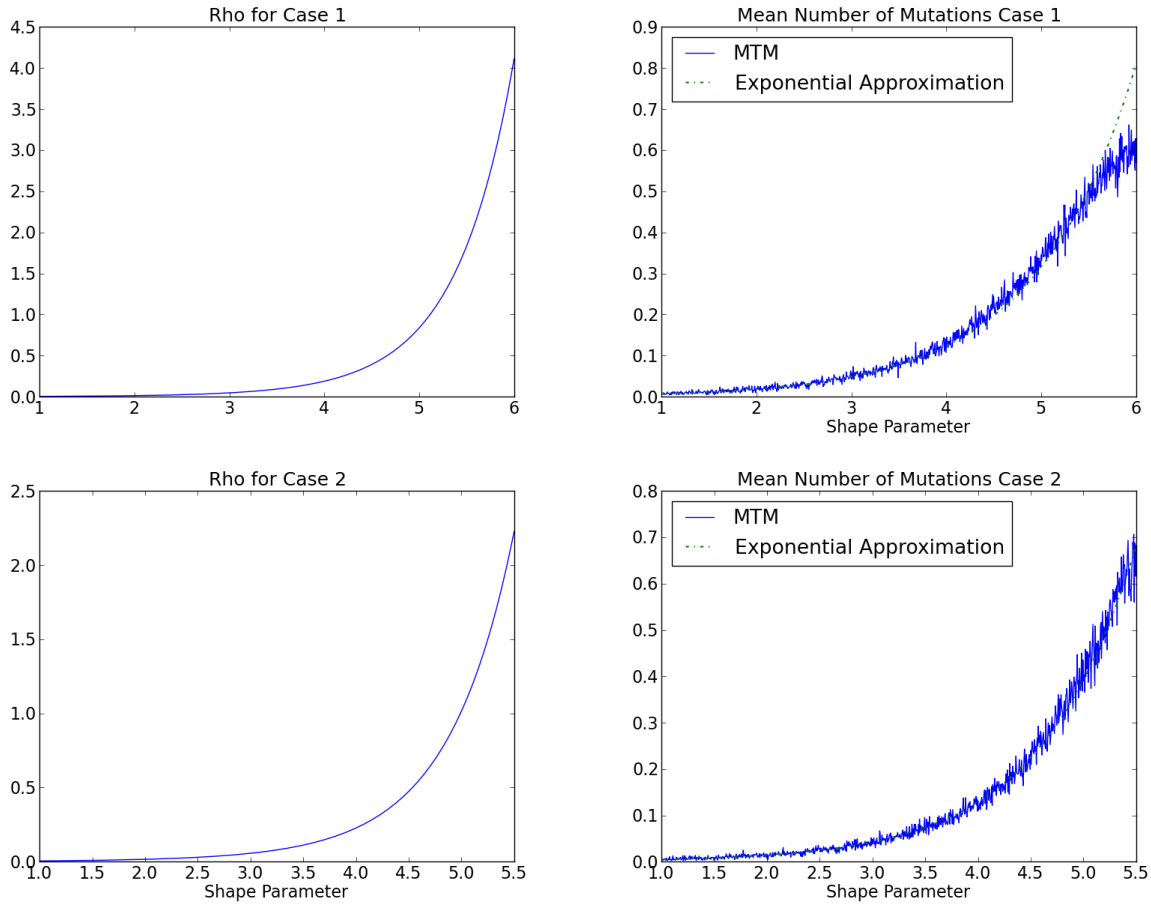

Figure 4.4: The figures on the left show the intensity measure ρ under the free recombination model for Case 1 (top row) and Case 2 (bottom row). The figures on the right show the empirical mean number of mutations, as well as the exponential approximation to the marginal means data, for Case 1 (top row) and Case 2 (bottom row).

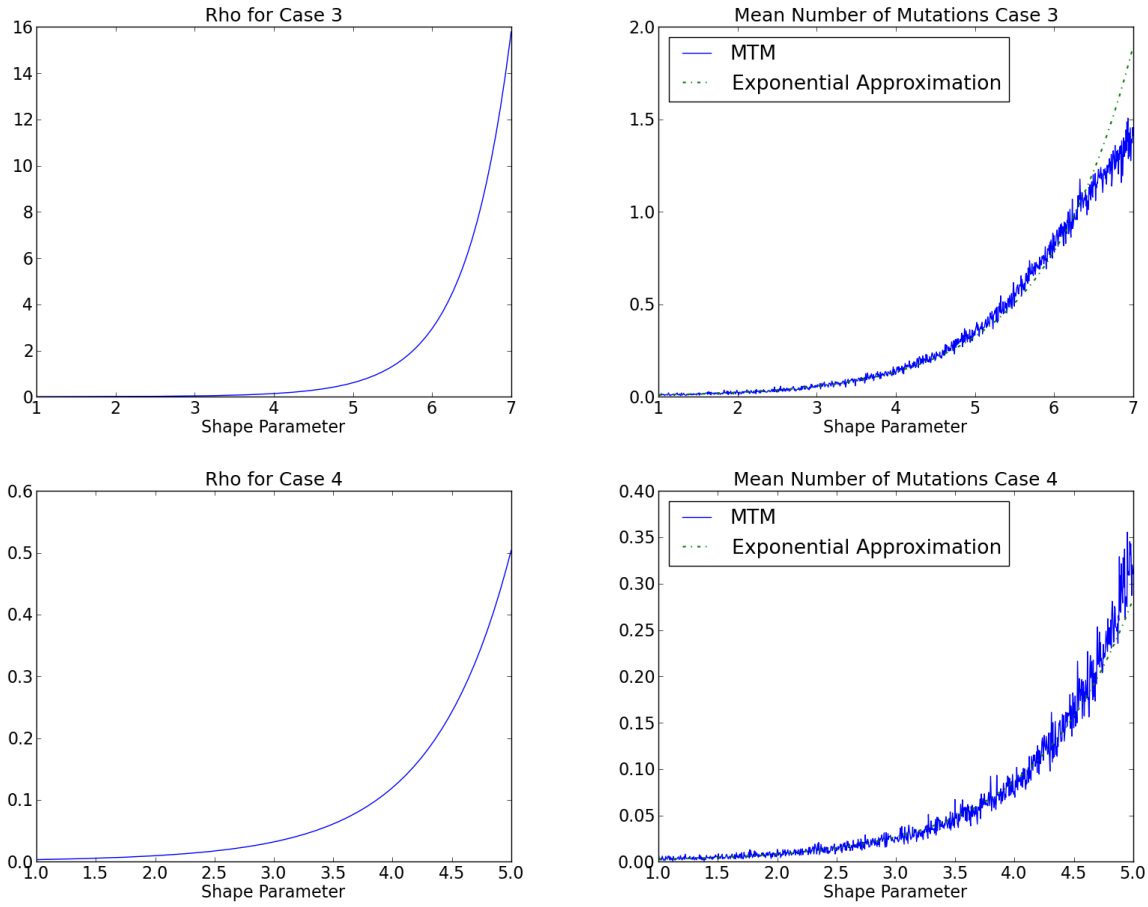

Figure 4.5: The figures on the left show the intensity measure ρ under the free recombination model for Case 3 (top row) and Case 4 (bottom row). The figures on the right show the empirical mean number of mutations, as well as the exponential approximation to the marginal means data, for Case 3 (top row) and Case 4 (bottom row).

Table 4.6: Survival probabilities computed from the expected population survival functions from the samples obtained by the MTM algorithm ("Empirical"), the Poisson approximation with the exponential function fit to the marginal means in place of the intensity function ("Approx."), and free recombination model ("ESW").

	Case 1			Case 2		
Survival Probability	Empirical	Approx.	ESW	Empirical	Approx.	ESW
0.5	27.0	27.0	27.0	27.0	27.0	27.0
0.4	30.0	30.0	30.5	30.5	30.5	30.0
0.3	34.0	34.0	34.0	34.5	34.5	33.0
0.2	39.0	39.0	38.0	39.0	39.0	37.0
0.1	46.0	46.0	43.5	46.0	46.0	42.5
0.05	52.0	52.0	48.0	52.0	52.0	47.0
0.01	63.5	63.5	55.5	64.0	64.0	55.5
0.005	67.5	68.0	58.0	68.0	68.5	58.5
0.001	77.0	77.0	63.5	78.0	78.0	64.5
	Case 3			Case 4		
Survival Probability	Empirical	Approx.	ESW	Empirical	Approx.	ESW
0.5	27.0	27.0	28.0	27.5	27.5	27.5
0.4	30.0	30.0	31.0	31.0	31.5	31.0
0.3	34.0	34.0	35.0	35.5	35.5	35.0
0.2	39.0	39.0	39.0	41.0	41.0	40.0
0.1	46.0	46.0	44.5	49.0	49.5	47.5
0.05	51.5	51.5	49.0	56.5	57.0	54.0
0.01	62.0	62.0	55.5	72.0	72.5	67.0
0.005	65.5	66.0	57.5	78.5	79.0	72.0
0.001	73.5	73.5	62.0	93.5	94.0	84.0

Table 4.7: Distance between the expected population survival function estimated directly from the MTM samples and from the Poisson approximation.

Case	$\|\| \cdot \|\|_2$	$\|\| \cdot \|\|_\infty$
1	0.00649023	0.00147418
2	0.00659143	0.00168235
3	0.00533242	0.00127027
4	0.00698008	0.00135615

number of mutations with large shape parameters. To understand why this is the case, it helps to recall that when the distribution of genotypes is a Poisson random measure, the expected population survival function can be computed by

$$\mathbb{E}[l_x(G)] = l_x(0) \exp\left(- \int_{\mathcal{M}} \left(1 - \mathrm{e}^{-\eta(m')\kappa(m',x)}\right) \rho(m') dm' \right).$$

Using the MTM algorithm, we assume that the chain has run for long enough to forget its initial state and that the resulting samples are taken from the true distribution of genotypes. When estimating the expected population survival function from the MTM samples directly, we use

$$\mathbb{E}[l_x(G)] \approx \frac{1}{\text{sample size}} \sum_g l_x(0) \exp\left(- \sum_{m \in g} \eta(m)\kappa(m,x) \right),$$

where the sum is over the genotypes contained in the sample.

For mutations with large shape parameters, the mutation profile $\kappa(m, x)$ is small over the ages of fertility. For example, consider a gamma mutation with shape parameter 6.0. This mutation profile was plotted against age in Figure 3.10. Over the ages of fertility, ages 15 to 50, a single copy of this mutation increases the cumulative hazard by at most 0.00913η. Using $\eta = 0.1$ (which we used in all the large gamma mutations cases presented here), a single additional copy of this mutation decreases survival by at most 9.129e-04 over the ages of fertility. Using a Poisson approximation to the distribution of genotypes and assuming that $\rho(m = 6.0) = 0.8$, a value chosen to be in line with the empirically determined mean number of copies of this mutation for Cases 1 and 3, we find that this mutation decreases expected survival by at most 7.300e-04 over the ages of fertility. If we increase $\rho(m = 6.0)$ to 1.0, simulating the overestimate of the empirical mean when using an exponential function, the expected population survival decreases by at most 9.125e-04 due to the presence of this mutation in the genome.

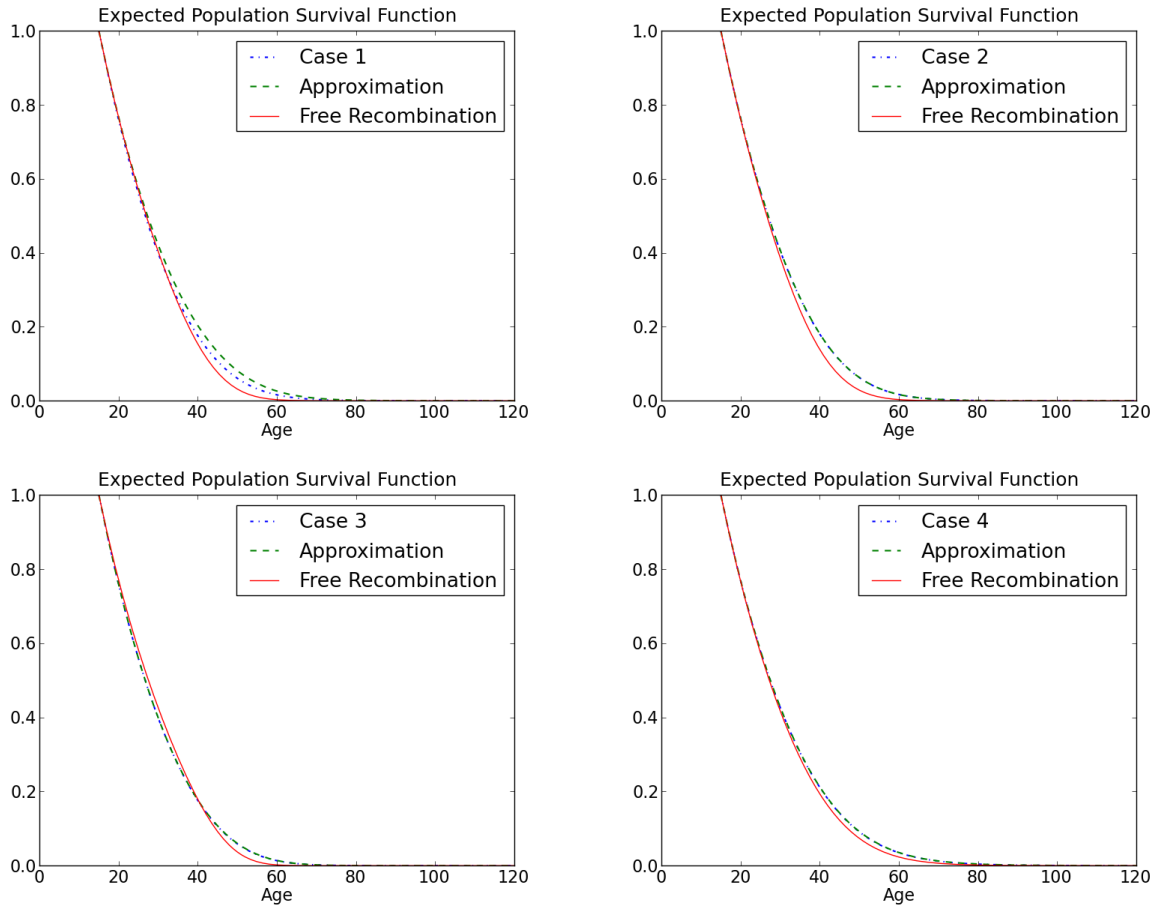

Figure 4.6: Expected population survival functions for the 1000 gamma mutations cases. Each plot shows the expected survival function under the SEW model estimated from the MTM samples and the Poisson approximation, as well as the survival function under the free recombination model.

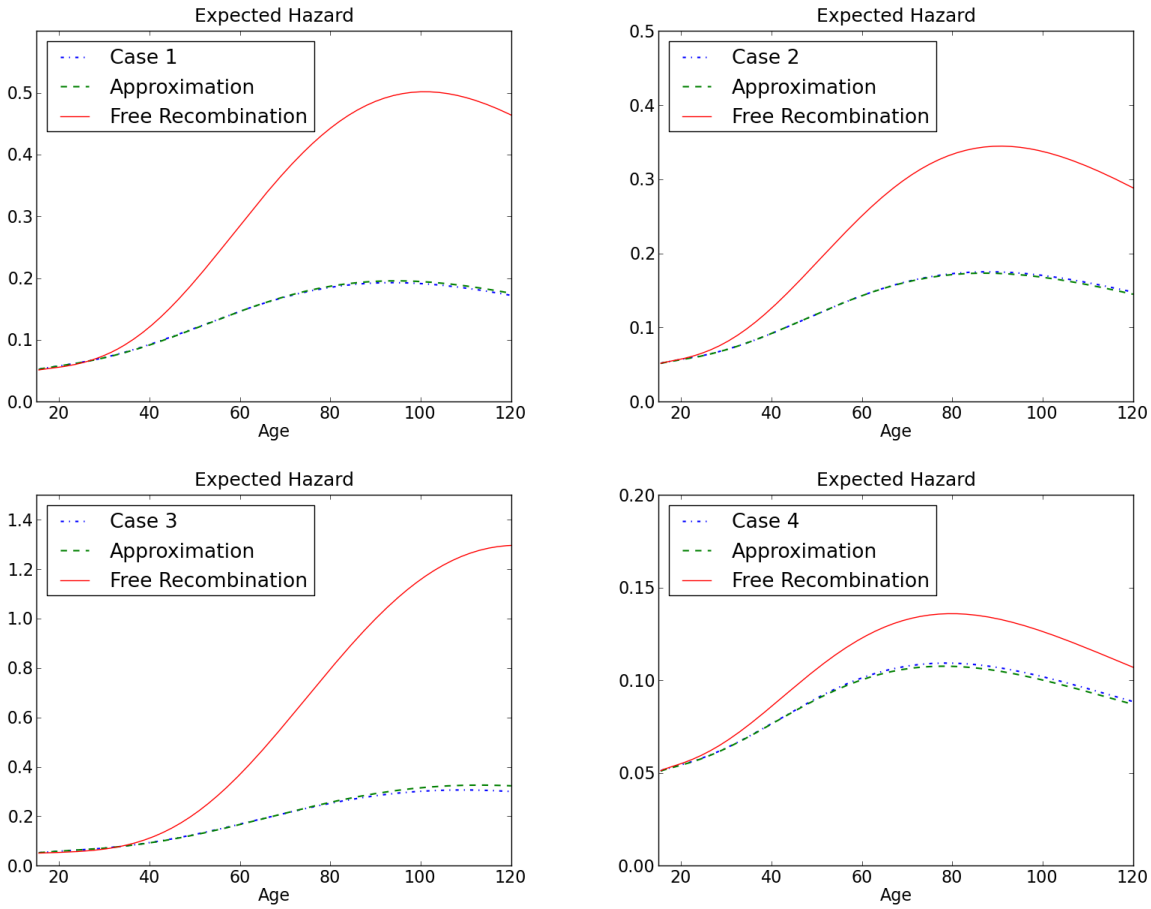

Figure 4.7: Expected hazard function for the 1000 gamma mutations cases. Each plot shows the expected hazard rate estimated from the MTM samples and the Poisson approximation, as well as the hazard rate under the free recombination model.

4.1.3 Similarity to the Free Recombination Model

It is interesting that the distribution of genotypes under the mutation-selection (no recombination) model and the free recombination model in the large gamma mutations cases described above exhibit an approximately exponential relationship between the gamma shape parameter for each mutation and the mean number of copies of each mutation. However, the intensity under the free recombination model is larger, overall, than the marginal means under the SEW model. This indicates that there are more mutations per genotype, on average, when there is free recombination than when there is no recombination.

Consider, for example, Case 1. Under free recombination, the average genotype contains 523.7 mutations, as opposed to 166.3 mutations with no recombination. For small shape parameters, the free recombination model actually has slightly fewer copies of these mutations on average than the no recombination model. For example, under the free recombination model, a typical genotype has roughly 0.57 mutations with shape parameters in $1.0 \leq m \leq 1.5$, 1.0 mutations in $1.5 < m \leq 2.0$, 1.84 mutations in $2.0 < m \leq 2.5$ and 3.48 mutations in $2.5 < m \leq 3.0$. Compare these to the average number of copies of mutations under the SEW model, found in Table 4.4. For large shape parameters, say $m > 4.5$, the average free recombination model genotype has at least twice as many mutations in a given interval as the average genotype under the mutation-selection model. For example, under the free recombination model, there are roughly 58.72 mutations with shape parameters in $4.5 < m \leq 5.0$, whereas there are 27.08 copies in under the SEW model; there are 127.26 copies of mutations in $5.0 < m \leq 5.5$ in the free recombination model genotype but only 41.34 under the SEW model; there are 282.74 copies of mutations in $5.5 < m \leq 6.0$ in the free recombination genotype but only 55.60 in the SEW genotype.

Similarly, in Case 3, the average genotype under the free recombination model has fewer mutations for $m \leq 4.0$ than the average genotype under the SEW model. In fact, the average free recombination genotype can have as many as 60% fewer mutations in a given subinterval for shape parameters near 1.0. For large shape parameters, $m > 5.0$, the free recombination genotype has at least twice as many mutations in the subintervals considered as the average genotype under the SEW model. The average number of mutations per genotype under the free recombination model is also much larger than the average for the no recombination model, with 1584.9 mutations under the free recombination model and 346.8 mutations under the SEW model. Of the four cases considered here, the outcomes for the two models in this case are the least similar.

In Cases 2 and 4, on the other hand, the average genotype under the free recombination model has more mutations in nearly every subinterval considered. The average number of mutations under the two models is fairly similar for small shape parameters, with the average mutation number under the free recombination model at most 20% higher than the SEW model mutation average for $m \leq 2.5$. However, the differences can be quite large for large shape parameters. For Case 2, the free recombination genotype contains twice as many mutations, on average, with shape parameters in the range $4.5 < m \leq 5.5$ as the average

genotype under the SEW model. The average number of mutations in the subintervals considered were most similar between the two models for Case 4. In Case 4, the average genotype for the free recombination model contained, at most, 55% more mutations in a given shape interval than is present in the average genotype under the SEW model. In addition, the average number of mutations per genotype in the two models was also most similar in Case 4. Under the free recombination model, the average genotype contains 88.92 mutations whereas it has only 61.77 mutations under the SEW model. As previously mentioned, the mutation spaces in Cases 2 and 4 also have the shortest ranges of shape parameters, with shape parameters ranging from 1.0 to 5.0 (in Case 4) and from 1.0 to 5.5 (in Case 2). Cases 1 and 3, having larger ranges of shape parameters, also have the smallest proportion of mutations with large early-age effects. For example, in Case 1, there are 100 mutation types with shape parameters in the range $1.0 \leq m \leq 1.5$, whereas in Case 3 there are only 84.

The average genotype under the free recombination model, having more mutations than the corresponding genotype under the SEW model, produces an expected population survival function that decreases to zero more rapidly than the expected population survival function with no recombination (see Figure 4.6). With similar number of mutations with small shape parameters under the two models, the difference is largely due to the fact that under the free recombination model, genotypes contain many more mutations, on average, with large shape parameters than do genotypes under the no recombination model. The difference between the expected population survival function is most similar in Case 4, as measured by the L^2 and L^∞ distances between the functions in this case, see Table 4.8.

Table 4.8: Distance between the expected population survival functions under the SEW model and the free recombination model.

Case	$\|\| \cdot \|\|_2$	$\|\| \cdot \|\|_\infty$
1	0.126034	0.0308901
2	0.184985	0.0427490
3	0.123913	0.0276150
4	0.0936692	0.0186994

The similarity between the expected population survival functions under the two models is important for constraining more realistic models of recombination. The two models considered in this work represent extremes in modeling recombination. In the free recombination model, recombination happens continuously, so that a genotype is composed of mutations drawn in proportion to their frequency in the population. The SEW mutation-selection model, by contrast, has no recombination. Under the SEW model, the genotype of an individual is the same as the genotype of the parent excepting perhaps an additional mutation. Mutations under this model simply accumulate along lineages with no way of shedding mutations from genotypes over time. This model may be useful when considering

mitochondrial DNA in humans. However, in much of our DNA, we experience both mutation and recombination, albiet at slower scales than modeled here. As a result, a more realistic model of recombination can be nicely sandwiched between these two extremes.

From the perspective of more realistically modeling recombination in human populations, it is important to note that both models produce similar average numbers of mutations for mutations with large early-age effects. This results in a fairly close agreement between the expected population survival functions for the two models for early-reproductive ages, approximately ages 15 to 40. For later ages, over age 40, the free recombination model produces a lower survival rate than the SEW model. The similarity in the models for younger ages is even more apparent when considering the population hazard rates (see Figure 4.7). The two models produce fairly similar hazard rates for young ages, roughly ages 15 to 30 or 35, before the hazards diverge. The free recombination model, producing genotypes with a larger number of mutations, on average, than the SEW model, also has much higher hazard rates for later ages. In particular, the hazard rates for the two models generally experience their greatest difference between ages 75 and 120.

4.1.4 Additional Gamma Cases

In the four large gamma mutations cases considered so far, the distribution of genotypes appears to be approximately Poisson distributed. However, all four of these cases involve mutation spaces with both early mean-age effects and late mean-age effects. One may reasonably ask if these results hold when the mutation space consists of only those mutations with late mean-age effects or only those mutations with early mean-age effects. We shall briefly consider both cases in this section.

As with the previous four cases, the mutation spaces in the following two cases each contain 1000 gamma mutations with the rate parameter 0.05. Both cases have the same overall mutation rate of 0.12. In Case 5, shape parameters range from 3.5 to 7.0, representing a space in which all mutations are primarily late-acting. Case 6 represents the scenario in which all mutations are primarily early-acting, with shape parameters ranging from 1.0 to 3.5. Recalling that the mean for a gamma distribution is the shape parameter divided by the rate parameter, we see that the mutations in Case 5 have mean-age effects in the range of 70 years to 140 years, while the mutations in Case 6 have mean-age effects ranging from 20 years to 70 years. The other parameters for these cases, such as background hazard rate and mutation effect size, are the same as in the previous four cases and are listed in Table 4.9. The fertility rate was set to the fertility rate needed to ensure a stationary population under the free recombination model on the same mutation space with the same parameters, see Table 4.10.

As one would expect, the histograms for the number of mutations per genotype in these two cases, shown in Figure 4.8, are quite different. Because the mean-age effects for mutations in Case 5 occur significantly after the latest age of reproduction, here set to 50 years, the selective pressure against such mutations is quite low. The incredibly low level of selective

Table 4.9: Parameters for additional gamma test cases.

All Tests	η	λ	α	β	Dx	Gamma rate	Kmax
	0.1	0.05	15	50	0.5	0.05	5
MTM Case	$\nu(\mathcal{M})$	Burn	Samples	DelP	ξ_0	ξ_1	
5	0.12	300000	1150000	0.5	3.5	7.0	
6	0.12	100000	400000	0.5	1	3.5	

Table 4.10: Output from shortcut algorithm for the free recombination model. The test case considered have mutation spaces with 1000 gamma profile mutations.

Case	ξ_0	ξ_1	Fertility	Iterations
5	3.5	7.0	0.074995506	96
6	1.0	3.5	0.06888683	9

pressure allows many copies of these mutations to build in the genome over evolutionary time. As a result, typical genotypes in Case 5 contain many more mutations than in any case considered so far (see Figure 4.8, left). Specifically, typical genotypes under the mutation-selection model in Case 5 generally contain 750 to 950 mutations, with an average of 847 mutations (see Table 4.12 for the mean and the variance in the number of mutations per genotype). By contrast, three-fifths of all the mutations in the Case 6 mutation space have mean-age effects during reproductive years. As a result, the selective pressure against most of the mutations in Case 6 is quite high, leading to genotypes with many fewer mutations. The histogram of the number of mutations per genotype for Case 6, Figure 4.8, right, shows that genotypes typically have been 5 and 30 mutations, with an average of about 14 mutations. The number of mutations per genotype in Case 6 is much lower than in the other four large gamma mutations cases considered previously. These two additional gamma test cases, then, provide a nice bookend to the cases considered in the previous section.

Table 4.11: Output from the MTM algorithm for the SEW model under test cases with 1000 gamma profile mutations.

Case	ξ_0	ξ_1	Acceptance Rate	*NRR*
5	3.5	7.0	0.343654	1.12458
6	1.0	3.5	0.318394	1.01730

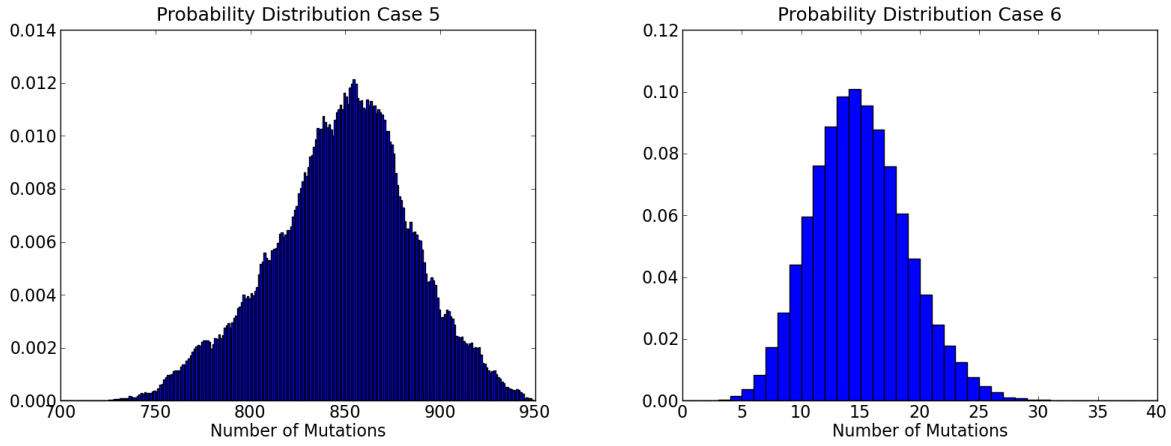

Figure 4.8: Histograms for the total number of mutations per genotype for the additional test cases with gamma mutations.

Poisson Approximation

Although Cases 5 and 6 represent extremes among the large gamma mutations cases, with a large number of mutations in typical genotypes in Case 5 and a fairly small number of mutations per genotype in Case 6, we find the histograms for the number of mutations per genotype are reasonably symmetric in both cases, suggesting that they may be approximately Poisson distributed. As with the previous four large gamma mutations cases, we compare the mean number of mutations to the variance in the number of mutations per genotype for both cases using the Poisson dispersion test. Unsurprisingly, we find that the differences are all statistically significant, indicating that chance cannot explain the observed differences between the means and the variances. However, we also find that the differences, while significant, are generally small. That the second case (Case 6) would produce a symmetric histogram and have means and variances that are generally similar is not that surprising. The large early and mid-range age effects guarantee that selection will be high, which is similar to Case 4.

However, the observation that the means and the variances are similar in these two cases with extreme gamma mutation spaces suggests a more general result. The previous four examples suggested that a Poisson approximation may be reasonable in cases with realistic mutation rates, demographic selective cost functions and where the mutation spaces have the following characteristics: the mutation space is large; the mutation profiles are smooth and non-zero for all adult ages; and the mutation space contains mutations with early mean-age, middle mean-age, and late mean-age effects. The results from Cases 5 and 6 suggest that a Poisson approximation may also be reasonable in cases where all mutations have generally late or very late mean-age effects (such as the mutation space in Case 5) and in cases where

all mutations have early to middle mean-age effects (such as the mutation space in Case 6).

Table 4.12: Estimated means and variances for Cases 5 and 6. The marginal means and variances correspond to the number of mutations per genotype with shape parameters (m) in the given intervals.

	Case 5			
	Mean	Variance	χ^2	p-value
$3.5 \leq m \leq 7.0$	847.346	1360.42	1846334	0
$3.5 \leq m \leq 4.0$	12.2210	14.2776	1343520	0
$4.0 \leq m \leq 4.5$	23.1624	22.2137	1102897	5.58e-218
$4.5 \leq m \leq 5.0$	41.8042	44.9539	1236643	0
$5.0 \leq m \leq 5.5$	74.7267	73.9202	1137587	1.07e-016
$5.5 \leq m \leq 6.0$	132.987	138.757	1199897	3.99e-231
$6.0 \leq m \leq 6.5$	225.032	288.289	1473264	0
$6.5 \leq m \leq 7.0$	337.412	387.746	1321554	0
	Case 6			
	Mean	Variance	χ^2	p-value
$1.0 \leq m \leq 3.5$	14.4236	16.0241	444386	0
$1.0 \leq m \leq 1.5$	0.76413	0.755297	395375	1.06e-7
$1.5 < m \leq 2.0$	1.29440	1.31930	407693	6.22e-18
$2.0 < m \leq 2.5$	2.13082	2.27458	426985	8.18e-192
$2.5 < m \leq 3.0$	3.68605	3.75472	407451	6.12e-17
$3.0 < m \leq 3.5$	6.54823	6.59680	402966	4.67e-4

Figure 4.9, which shows the cumulative variance (the variance in the number of mutations with shape parameters in ξ_0 to k, where $k \in [\xi_0, \xi_1]$) and the sum of the variances (the cumulative sum of the marginal variances). In general, the cumulative variance is approximately the same as the sum of the variances for smaller shape parameters, but can be much larger than the sum of the variances for higher shape parameters. The difference appears to be much smaller in Case 6 than in Case 5. This result is not surprising when we recall the close agreement between the sum of the variances and the cumulative variance in Case 4, which has the same mutation rate but a shorter range of shape parameters.

To determine how well the distribution of genotypes is approximated by a Poisson, we turn our attention to the empirical mean number of each mutation type for the two cases, shown in Figure 4.10. We have included the intensity measure ρ under the free recombination for comparision. The first thing to notice is that the intensity function in both cases is still approximately exponentially distributed, as are the marginal means data. As with the previous cases, we fitted the marginal means data to an exponential function. The fitted coefficients are listed in Table 4.13. The pattern observed previously appears to hold

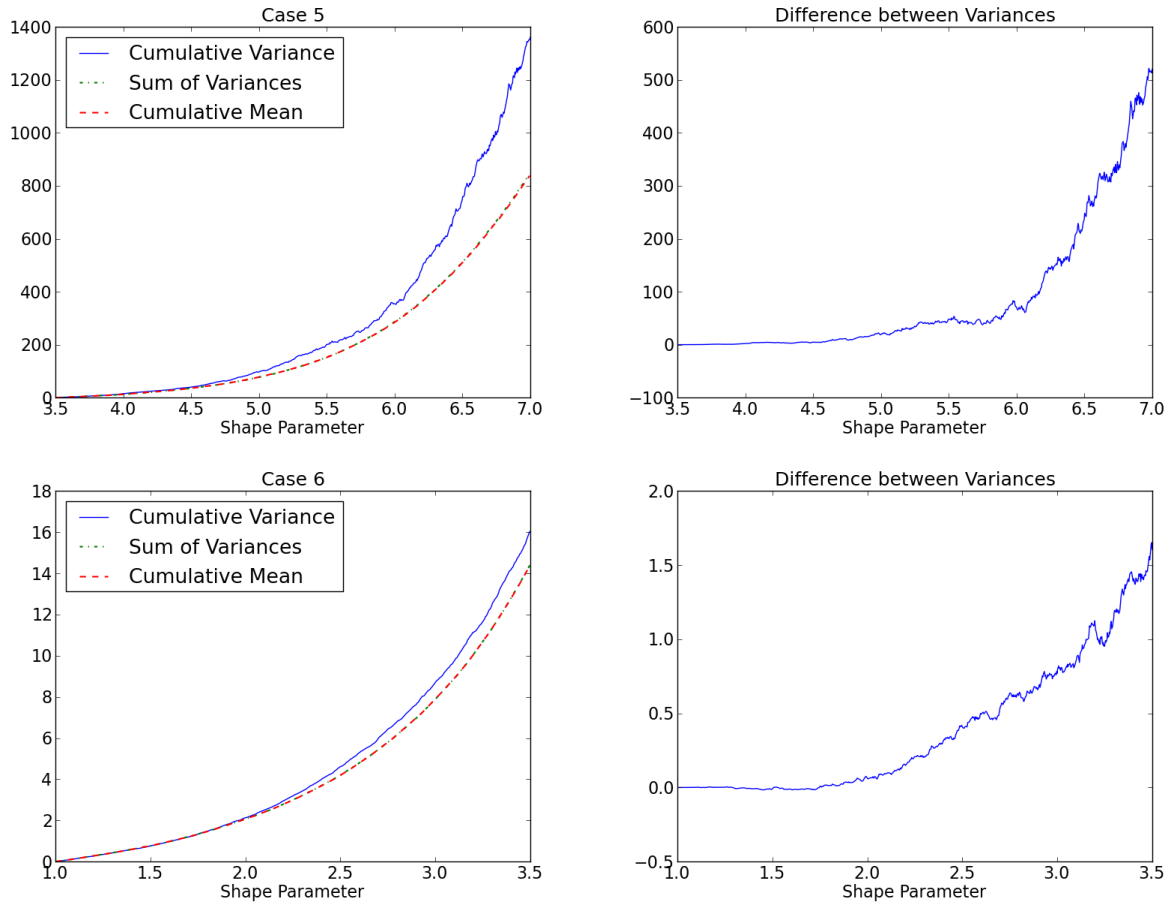

Figure 4.9: The plots on the left show the sum of the variances, the cumulative variance and the cumulative mean for Case 5 (top row) and Case 6 (bottom row). The plots on the right show the difference in the variances for Case 5 (top row) and Case 6 (bottom row).

here as well. The empirically determined marginal means data for Case 6, in which the average number of mutations per genotype is small, follows the exponential curve more closely than in Case 5. As we observed with Cases 1 and 3, the exponential approximation fails at later ages. It is somewhat unclear whether this result is caused by an insufficient number of samples in those cases. Cases 1, 3 and 5 represent cases in which we expect average genotypes to contain large numbers of mutations and it is possible that more samples may reduce this trend. It is also possible that there is some type of natural depression in the number of mutations with the oldest mean-age effects. Because mutations are not statistically independent under the SEW model, it is possible that the presence of mutations with earlier age effects could be pulling down the number of copies of mutations with the latest age effects. Over evolutionary time, individuals that experience mutation events with large early-age or mid-age effects were much less likely to survive and reproduce than those individuals with mutations that primarily act on older ages. Without recombination to split genotypes, having a mutation that substantially depresses lifespan would guarantee that all future generations of this genetic line are severely disadvantaged. Given this fact, it is surprising that the distribution of genotypes can be approximated by a Poisson distribution, in which mutation types are statistically independent. As with the previous cases, it seems that this approximation holds best for mutations with smaller shape parameters.

Table 4.13: Coefficients from using scikits.statsmodels.OLS to fit the model log(Marginal Means) $= \alpha$ Shape Parameter $+ \beta$.

Case	m	Number of observations	Coefficients	
5	[3.5,7.0]	1000	α	1.12963
			β	-6.65665
6	[1.0,3.5]	1000	α	1.07445
			β	-6.96019

Although the exponential approximation is better in Case 6 than in Case 5, in both cases the fit is reasonable. As a result, we use the exponential approximation to the marginal means data in place of the intensity measure ρ in approximating the distribution of genotypes by a Poisson random measure. The approximation is quite good in both cases when our interest lies in estimating the expected survival function, shown in Figure 4.11, or the expected hazard rate, shown in Figure 4.12. We can confirm this visual observation by computing the difference between the expected survival function estimated directly from the MTM samples and the Poisson approximation, which is around 0.01 under the L^2 norm for Case 5 and 0.006 for Case 6, as we can see from Table 4.14.

We can also confirm this observation by comparing survival probabilities estimated from the MTM samples directly and from the Poisson approximation. These data are provided in Table 4.15. In Case 6, the approximation and the empirical survival functions are nearly identical. However, we point out that the accuracy of the table is limited by the choice of

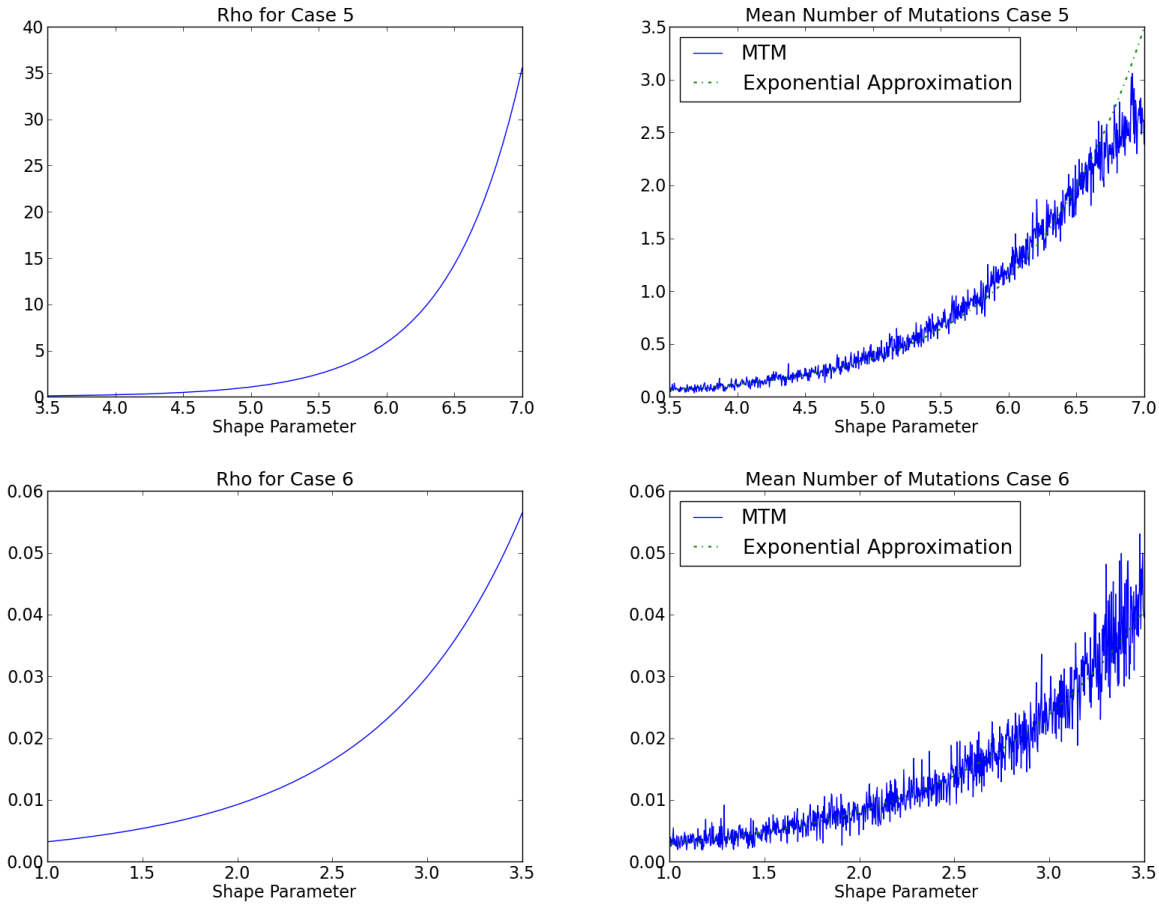

Figure 4.10: The figures on the left show the intensity measure ρ under the free recombination model for Case 5 (top row) and Case 6 (bottom row). The figures on the right show the empirical mean number of mutations, as will as the exponential approximation to the marginal means data, for Case 5 (top row) and Case 6 (bottom row).

discretization in age, in this case 0.5 years. Notice that for Case 6 we have not included the age to which at most 0.001 of the population survives. This is because for all three expected survival functions, the empirical, the approximation and the survival function under free recombination, more than 0.001 of the population survives to the oldest age considered, 120 years. The Poisson approximation in Case 5 is not quite as good as it is in Case 6, although the difference in estimated age for the various survival probabilities listed is small, no more than 0.5 years.

Table 4.14: Distance between the expected population survival functions estimated directly from the MTM samples and from the Poisson approximation.

Case	$\|\| \cdot \|\|_2$	$\|\| \cdot \|\|_\infty$
5	0.00519794	0.00136816
6	0.00570866	0.00106093

Finally, we wish to draw attention to the fact that the expected survival function under the SEW model and the expected survival function under the free recombination model are quite close in Case 6. Indeed, the hazard rates under the two models in Case 6 are only slightly different, with the largest differences occurring around middle age (40 to 60 years). At middle age, the hazard rate under the free recombination model is slightly larger than the hazard rate under the no recombination model. This suggests that with a mutation space in which mutations face high selective pressure (resulting from their large reproductive-age effects), the two models produce very similar demographic outcomes. Case 5, by contrast, produces the largest differences between expected survival function and hazard rates yet observed. Indeed, the expected survival function approaches zero much faster in Case 5 than in any of the other gamma test cases, with only 0.001 of the population surviving to age 52 years. This drop in survival probability occurs at a much age than in Case 3, where 0.001 of the population survived to age 62 years. This observation is also confirmed by looking at the L^2 distance between the expected survival functions under the two models. Listed in Table 4.16, the distance is 0.37, which is much higher than we observed for Cases 1-4, where the distances ranged from 0.09 (Case 4) to 0.18 (Case 2).

Table 4.15: Survival probabilities computed from the expected population survival functions from the samples obtained by the MTM algorithm ("Empirical"), the Poisson approximation with the exponential function fit to the marginal means in place of the intensity function ("Approx."), and free recombination model ("ESW").

	Case 5			Case 6		
Survival Probability	Empirical	Approx.	ESW	Empirical	Approx.	ESW
0.5	28.5	28.5	28.0	27.5	27.5	27.0
0.4	32.0	32.0	31.0	31.0	31.0	30.5
0.3	36.0	36.5	34.0	35.5	35.5	35.5
0.2	41.0	41.0	37.0	41.5	41.5	40.5
0.1	46.5	46.5	40.5	51.5	51.5	50
0.05	51.0	51.0	43.0	62.0	62.0	60
0.01	58.5	59.0	47.5	87.0	87.5	83.5
0.005	61.0	61.5	49.0	99.0	99.0	94.5
0.001	66.5	66.5	52.0	¿ 120	¿ 120	¿ 120

Table 4.16: Distance between for the SEW model and the free recombination model.

Case	$\|\| \cdot \|\|_2$	$\|\| \cdot \|\|_\infty$
5	0.378329	0.105509
6	0.0583749	0.0105569

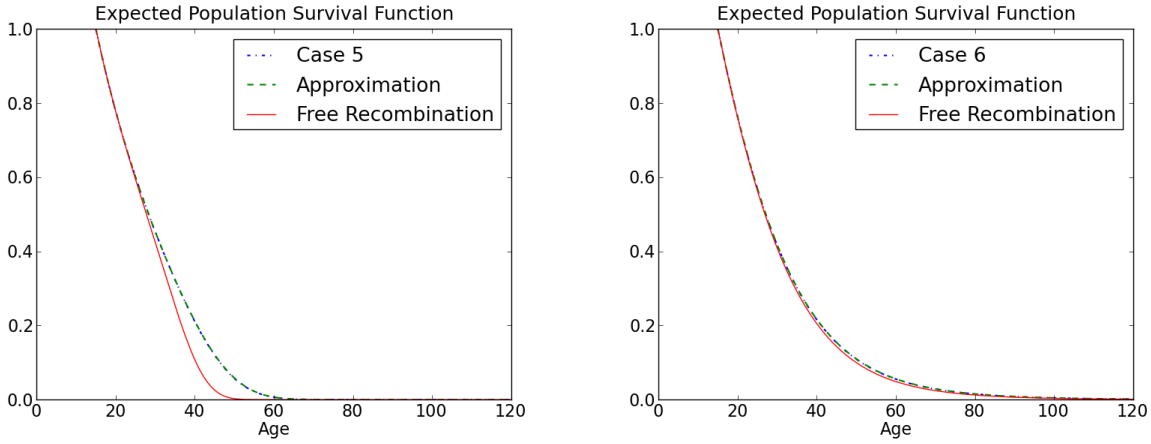

Figure 4.11: Expected population survival functions for the additional 1000 gamma mutations cases. Each plot shows the expected survival function under the SEW model estimated from the MTM samples and the Poisson approximation, as well as the survival function under the free recombination model.

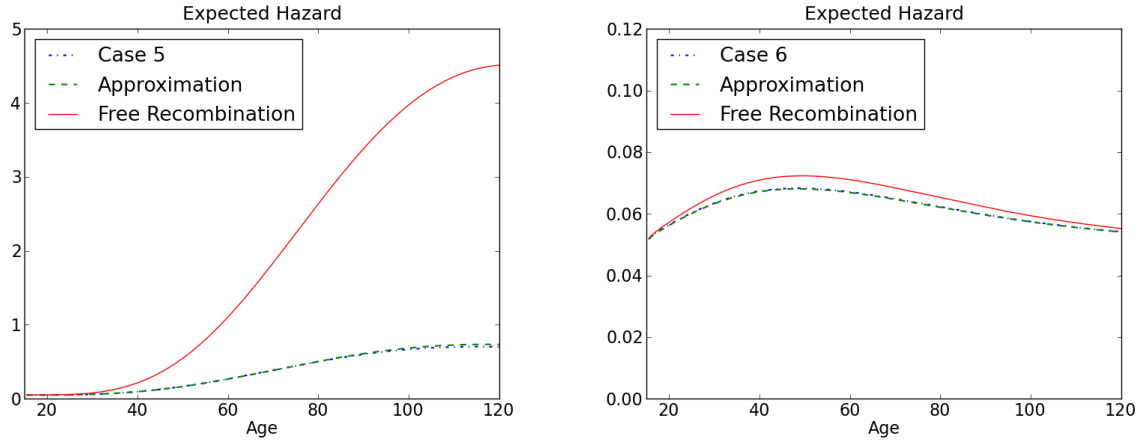

Figure 4.12: Expected hazard function for the additional 1000 gamma mutations cases. Each plot shows the expected hazard rate estimated from the MTM samples and the Poisson approximation, as well as the hazard rate under the free recombination model.

4.2 Mutations with Modified Point-Mass Profiles

The previous gamma mutations cases have suggested that while the distribution of genotypes is not Poisson, it can be well approximated by a Poisson random measure if the goal is measuring expected survival or population hazard rate. We have also discovered that those demographic outcomes are often similar to the outcomes under the free recombination model. This has been demonstrated with demographic selective cost functions and realistic mutation rates (0.12–0.17) for several scenarios involving gamma profile mutations, namely, mutation spaces containing mutations with both early mean-age effects and late mean-age effects (Cases 1–4), a mutation space with primarily late-age effects (Case 5) and a mutation space with primarily early-age effects (Case 6). One might ask, however, if these results apply to other types of mutation profiles or if the gamma profile case is, in some way, unique. To this end we present one last test case.

The final test case uses highly stylized mutation profiles that would not be used in practice to describe realistic age-specific mutation effects. These profiles will be referred to as "modified point-mass" profiles because they are a slight modification of the point-mass profile used in §3.3. Recall that under the point-mass model for mutation profiles, mutations concentrate their effects on the hazard rate at a single age, called the age of onset and denoted by m. This results in a step increase in cumulative hazard starting at age m. Although such a model is unrealistic from a biological perspective, its simplicity proved useful in testing the MCMC algorithms applied to this problem. Because of its extreme simplicity and stylized nature, the point-mass profile also provides a nice contrast to the more realistic gamma profiles. Unfortunately, it has been shown in [35] that the solution to the free recombination model unravels in the case with constant fertility, constant mutation rate ν above the age of maturity and point-mass profiles with ages of onset m where $\mathcal{M} = [\alpha, \infty]$. This result also holds when fertility is constant only in a range of ages and zero otherwise, and when the age of onset for point-mass mutations are restricted to this range of fertile ages.

Large spaces of point-mass mutations, then, are not a good test case. However, it was speculated by Wachter, Evans and Steinsaltz in [35] that attaching an initial small effect at young ages to those mutations whose main effect is concentrated at older ages would prevent the solution to the free recombination model from unraveling. We have chosen to follow that suggestion and call the resulting mutation profile a modified point-mass profile. With the modified point-mass profile, the mutation profile is modeled as a double step function, with a small step of size $\delta = 0.001$ at the age of maturity α, and a second step of size $1 - \delta$ at the age of onset m. The test case considered here contains 351 modified point-mass mutations with ages of onset ranging from $\alpha = 15$ years to $\beta = 50$ years in step sizes of 0.1 years. The background hazard rate for this test was set to the standard 0.05 and the size of the mutation effect was 0.1. Other parameters used in this test, such as the burn-in period for the MTM algorithm and the number of samples collected, are listed in Table 4.17. As with the gamma mutations cases discussed previously, the fertility rate was set to the fertility rate that ensures a stationary population under the free recombination model.

The fertility rate under the free recombination model was 0.088 (Table 4.18). Unlike the cases with gamma mutations, the fertility rate under the free recombination model for the modified point-mass mutations is too high to ensure that a stationary population under the SEW model. The net reproduction ratio estimated from the MTM samples using this fertility rate was 1.29, much higher than the 1.0 which ensures a stationary population. As a result, the fertility would need to be rescaled by 1/1.29, resulting in a fertility rate of 0.068544. This suggests that the output from the free recombination model and the SEW model in this case may be more different than in the gamma cases.

Table 4.17: Parameters for the test case with 351 modified point-mass mutations.

Parameters	η	λ	α	β	Dx	δ	$\nu(\mathcal{M})$
	0.1	0.05	15	50	0.1	0.001	0.17
MTM Parameters	Burn	Samples	DelP	Kmax			
	150000	600000	0.375	5			

Indeed, compare the intensity measure for the free recombination model with modified point-mass mutations, shown in Figure 4.13, left, to the plot of the mean number of each mutation type under the SEW model, shown on the right. Unlike the gamma cases considered previously, the intensity measure and the marginal means data in this case have different shapes. Specifically, the intensity function is sigmoid in nature (see Appendix C.3 for details) while the marginal mean number of mutations under the SEW model is neither sigmoid nor exponential. The log of the marginal means data, shown in Figure 4.14 (left), does appear to be approximately exponential and was, thus, fitted by an exponential curve,

$$\log(\text{Marginal Mean}) = a + \exp(bx + c).$$

The exponential curve was fitted using curve_fit from scipy.optimize. The resulting parameters were $a = -5.67638$, $b = 0.0767402$ and $c = -2.22687$. The fitted exponential is plotted in a dashed line on the same figure. The exponential approximation to the log means is best for smaller ages of onset. In particular, the log of the marginal means increases much more rapidly for late ages of onset (above 45 years) than does the exponential approximation. The plot on the right in Figure 4.14 shows the marginal means data for the SEW model (solid line) and the double exponential approximation (dashed line). Again, the fit is best for small ages of onset, with the approximation increasing much less dramatically at late ages than the actual means.

With the intensity measure and the marginal means data exhibiting different patterns in this case, one may wonder if the Poisson approximation, which proved to be quite good in the gamma mutations cases, might not hold in this case. Surprisingly, the Poisson approximation also holds for modified point-mass profiles. Table 4.20 shows the mean and variance in the

Table 4.18: Output from shortcut algorithm for the free recombination model when the mutation space contains 351 modified point-mass mutations.

Fertility	Iterations
0.08846245	211

Table 4.19: Output from the MTM algorithm for the SEW model when the mutation space contains 351 modified point-mass mutations.

Acceptance Rate	*NRR*
0.341496	1.29062

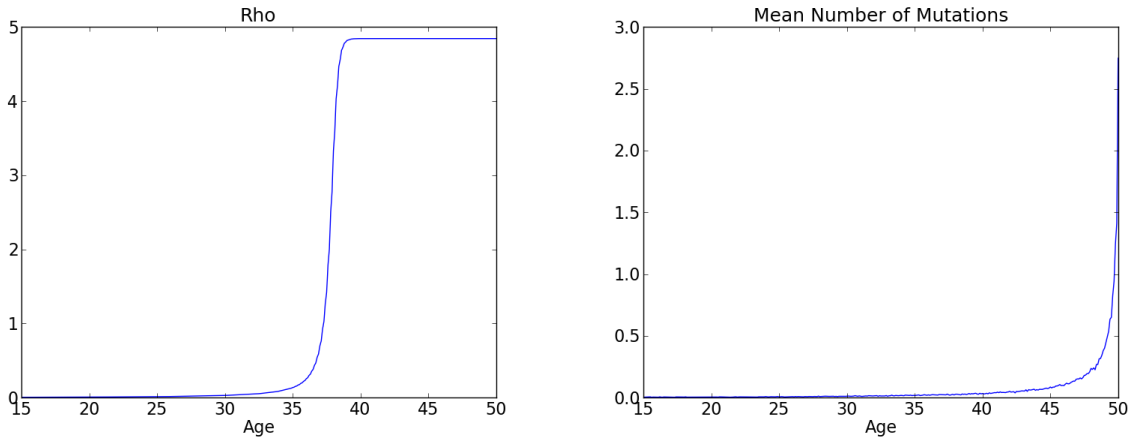

Figure 4.13: The plot on the left shows the intensity measure ρ for the free recombination model in the modified point-mass mutations case. The plot on right is the empirical mean number of mutations under the SEW model.

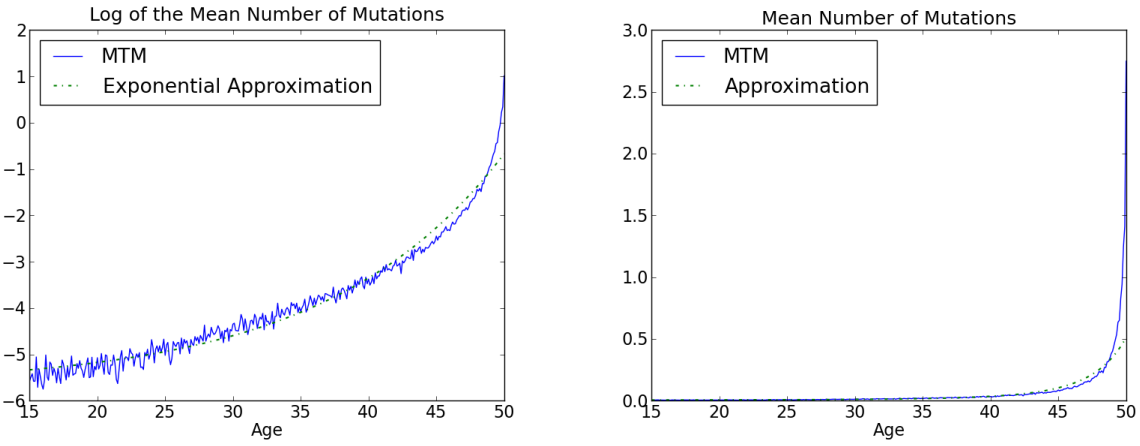

Figure 4.14: The log of the mean number of each type of mutation (solid line) fitted with an exponential curve (dashed line) are shown on the left. The marginal means data (solid line) and the double exponential approximation (dashed line) are shown on the right.

total number of mutations per genotype as well as the mean and variance in the number of mutations with age of onset in subintervals of size 5 years. In all cases, the means and the variances are reasonably close. As with the gamma cases, the differences between the means and variances are nearly all (excepting one subinterval) statistically significant. However, the differences are small enough that a Poisson approximation is actually quite good, as we shall demonstrate below.

The double exponential fitted to the marginal means data was used for the intensity function of the Poisson approximation to the genotype distribution under the SEW model. While the double exponential approximation does not accurately reflect the behavior of the means data for late ages of onset, this discrepancy results in a negligible difference when considering the expected population survival function (Figure 4.15, left) or the expected hazard rate (right). The difference between the population survival functions estimated directly from the MTM samples and the Poisson approximation is quite small, with an L^2 distance of 0.0370246 and an L^∞ distance of 0.00973396.

This observation is confirmed by comparing the youngest age at which the probability of survival is at most p, shown in Table 4.21. The ages are very close for the empirical expected population survival function and the approximation for most of the probabilities tested. The differences become more pronounced when dealing with small probabilities of survival (less than 1%) with the Poisson approximation based on the double exponential tending to overestimate the age.

While the Poisson approximation in this case is quite good, it is clear that the difference between the demographic outcomes for free recombination model and the SEW model are

Table 4.20: Estimated means and variances for the test case with modified point-mass mutations. The marginal means and variances correspond to the number of mutations per genotype with age of onset (m) in the given intervals.

	Mean	Variance	χ^2	p-value
$15 \leq m \leq 50$	22.6355	28.5054	755590	0
$15 \leq m \leq 20$	0.240712	0.248565	619574	3.01e-70
$20 \leq m \leq 25$	0.313238	0.313918	601300	1.17e-1
$25 \leq m \leq 30$	0.473408	0.483884	613276	1.20e-33
$30 \leq m \leq 35$	0.754100	0.774506	616235	3.73e-49
$35 \leq m \leq 40$	1.26960	1.34050	633504	1.71e-198
$40 \leq m \leq 45$	2.62594	2.87120	656038	0
$45 \leq m \leq 50$	16.9585	20.4001	721762	0

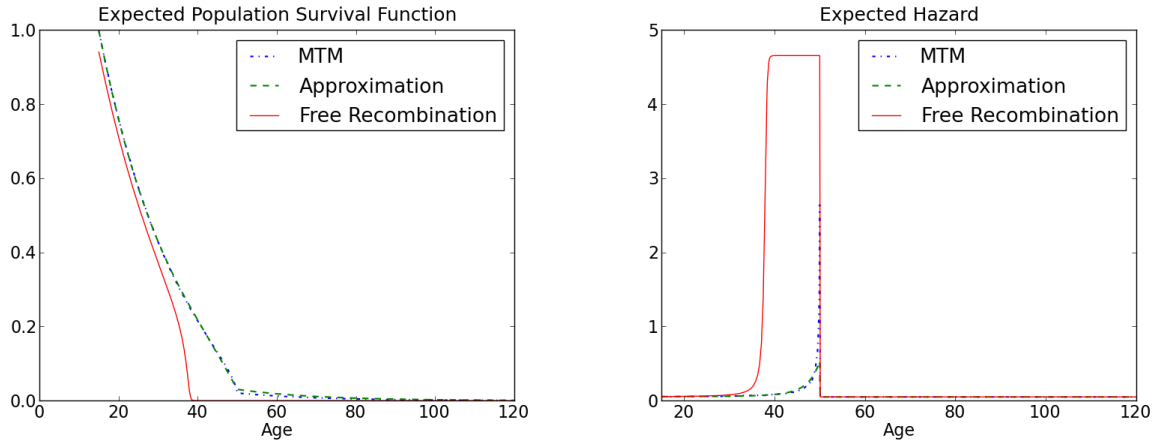

Figure 4.15: The figure on the left shows the expected population survival function under the SEW model estimated from the MTM samples and the Poisson approximation, as well as the survival function under the free recombination model. The figure on the right shows the expected hazard rates.

Table 4.21: Survival probabilities computed from the expected population survival functions from the samples obtained by the MTM algorithm ("Empirical"), the Poisson approximation with the exponential function fit to the log marginal means in place of the intensity function ("Approx."), and free recombination model ("ESW").

Survival Probability	Empirical	Approx.	ESW
0.5	27.4	27.4	25.9
0.4	31.1	31.2	29.1
0.3	35.5	35.7	32.5
0.2	40.9	41.0	35.4
0.1	46.8	46.4	37.1
0.05	49.3	49.0	37.7
0.01	64.6	72.3	38.2
0.005	78.5	86.2	38.4
0.001	110.7	118.4	38.8

more different for modified point-mass profiles than they were for the gamma profiles. In particular, the expected population survival function for the two models is only similar for younger ages, less than 35 or so. The survival probabilities go to zero much faster under the free recombination model than under the SEW model – for example, only 0.1% of the population survives to age 38.8 under the free recombination model whereas 0.1% of the population survives to age 110.7 under the SEW model.

4.3 Conclusion

This work primarily explores two questions. First, how similar is the distribution of genotypes under the SEW model to a Poisson random measure? Second, how similar are the distributions of genotypes under the two models, free recombination and no recombination?

4.3.1 Similarity to a Poisson Random Measure

With a low mutation rate, the distribution of genotypes under the SEW model can be approximated by a Poisson random measure. Of more interest are cases with realistic mutation rates. The four large gamma mutation profile cases discussed in §4.1 are all examples with mutation rates similar to those experienced by human beings. These four cases, while very specific, have been instrumental in delving into the first question. They show that while the distribution of genotypes is not Poisson, it can be well approximated by a Poisson random measure in some cases when the ultimate goal is to model demographic outcomes. In particular, we find that using a Poisson random measure whose intensity is estimated from the empirically observed mean number of each type of mutation produces demographic outcomes (such as lifespan) that are quite similar to the outcomes determined from the data directly.

By looking at two additional gamma cases, we find that the approximation holds even when the mutation space is primarily composed of mutations with mean-age effects in reproductive years, and when the mutation space is primarily composed of mutations with late mean-age effects. We also find that the Poisson approximation is not limited to mutations with gamma profiles. This was determined by considering a large mutation space with modified point-mass profiles. Although modified point-mass profiles are highly unrealistic models for the age-specific effects of real-world mutations, it is encouraging to know that these results hold for both the smooth gamma profiles and the highly stylized modified point-mass profiles.

However, the similarity between the empirically determined means and variances is not limited to the large mutation space cases considered in this chapter. Indeed, the similarity was first encountered when considering small mutation spaces, in particular those with only one or two types of mutations with point-mass profiles (see §3.3.4 and Tables 3.16, 3.17 and 3.18 for more details). In all of these cases, while the distribution is not a Poisson random measure and the difference between the means and the variances in the number of mutations are statistically significant, the variances are generally within about 25% of the means, with the variance usually larger than the mean. In these cases, too, the distribution can be approximated by a Poisson random measure when the desired outcome is expected population survival.

It is important to emphasize how very unexpected this result actually is. It is well known in population genetics models that because recombination breaks genotypes at random loci, the process eventually produces statistical independence between loci. Most models that include recombination take some pains to explicitly characterize how quickly statistical inde-

pendence occurs. In the ESW free recombination model discussed is this work, recombination occurs so quickly that loci are statistically independent, producing a Poisson distribution of mutations.

It is common in population genetics to model mutation by assuming that a Poisson process is responsible for introducing mutations to a population over evolutionary time. If mutation alone were acting on the genetics of the population, the distribution of genotypes would be Poisson. Selection, however, forces the system away from a Poisson distribution. As we discussed in Chapter 2, except in very special cases (such as non-epistatic or additive selective cost functions), the solution to the SEW mutation-selection model is not Poisson distributed.

It is therefore, very interesting and unexpected to find that while the distribution of genotypes without recombination is not Poisson, it can be reasonably approximated by a Poisson random measure. The approximation is not perfect, of course – the mutations under the SEW model are statistically dependent, which appears to cause a depression in the number of mutations with the oldest-age effects when the mutation space also contains mutations with large early reproductive age effects. However, the approximation is perfectly serviceable when the goal is to model expected survival or population hazard rates.

4.3.2 Similarity to the Free Recombination Model

We can make several general observations regarding the similarity of the distribution of genotypes under the two models, free recombination and no recombination. In the first place, we find that the mutation space itself is quite important in determining the similarity between the outcomes of the two models. In particular we find that, for mutations with gamma profiles, the outcomes of the two models are closer for mutation spaces in which a large proportion of the mutations face heavy selective pressure (resulting from those mutations having large early or mid-reproductive age effects). As the fraction of mutations with large late or very late-age effects grows, the statistical dependence between mutations under the SEW model appears to depress the number of copies of the mutations with the smallest effects. Of course, in all cases, the average number of mutations per genotype is higher under the free recombination model than under the SEW model. However, because much of the difference is concentrated at mutations with late-age effects, when there are few survivors in the population (due to the extrinsic hazard rate), the difference in mutation numbers causes only moderate differences in lifespan. More importantly, however, we find substantial qualitative similarities between the outcomes from the two models. For example, with gamma mutations, both models produce mean numbers of mutations types that are roughly exponential in shape parameter.

The two models are probably most different in the case of mutations with modified point-mass profiles. In all of the test cases involving the gamma mutations, the intensity measure ρ and the marginal means data had the same general shape (roughly exponential) even though the marginal means data was usually smaller than the intensity measure. In the case with

modified point-mass profiles, however, the marginal means data and the intensity measure had very different shapes. This translated into quite noticeable differences between lifespan and hazard rates under the two models. However, because this type of mutation profile is highly stylized and not realistic for real-world mutations, this difference may not be terribly important.

4.3.3 Senescence

In our tests we considered two models for age-specific mutation effects, representing two different theories of the evolution of senescence, mutation accumulation and positive pleiotropy. The simplicity of point-mass mutations makes them simple models for studying mutation accumulation. With the point-mass profile mutations, all mutations are deleterious and those with young ages of onset face more selective pressure and are less common in the population than mutations with late ages of onset. This is apparent even in the small mutation space cases we tested. For example, with two point-mass profile mutations, there are fewer copies of the mutation with age of onset 20 than the mutation with a later age of onset. When the two possible mutations both had early to middle ages of onset, $m_1 = 20$ and $m_2 = 30$, there were about twice as many copies of m_2 (the later acting mutation) in an average genotype, than copies of m_1 (the early acting mutation). Specifically, there were about 2.75 copies of $m_2 = 30$ to the 1.33 copies of $m_1 = 20$. When the second mutation had later age effects, such as the case where $m_1 = 20$ and $m_2 = 40$, there were almost six times as many copies of the mutation with the later age of onset, on average, as there were copies of the mutation with an early age of onset, 4.17 copies of $m_2 = 40$ to 0.70 copies of $m_1 = 20$. In the second case, copies of the mutation with a later age of onset are building over evolutionary time because the selective pressure on the second mutation is so much less than the selective pressure on the mutation with early-age effects.

One of the problems with using point-mass profiles, however, is that it can result in a wall of death unless the mutation space is small. Mutations with gamma profiles, on the other hand, are in line with the theory of reinforcing pleiotropy. In this case, all mutations are deleterious and have nonzero effects at every adult age. That is, even mutations that have a large late-age effects will have some (perhaps quite small) deleterious effect at early ages, too. In this model as well, mutations with primarily late-age effects were more common, on average, than mutations with large early-age effects.

In studying senescence, the hazard rates for the cases with gamma mutations are quite interesting. The Gompertz model, introduced in 1825 and standard in the field of demography, posits that the rate of mortality increase is roughly exponential after the age of maturity. Although this model fits human and some laboratory animal data (such as *Drosophila*, see [33]) for most adult ages, it fails to hold for the oldest ages, where mortality rates appear to flatten. Charlesworth, who also considered adding age-specific mutation effects to population genetics models, was able to reproduce the familiar Gompertz-Makeham form for mortality and was also able to reproduce the flattening that appears at the oldest ages by adding a

non-age-specific effect for each mutation (see §1.4.3 for a brief description of his model and see [7] and [8] for more details). However, Charlesworth's model relies on a linear approximation to estimate the effects of the mutations on genetic fitness. The two models considered in this work do not rely on a linear approximation but, instead, consider the full nonlinear effects of mutations on fitness. It was previously observed that a flattening or even decrease in hazard rate can occur for gamma profile mutations under the free recombination model (see [37]). However, this work has shown the same behavior in hazard rates under the no recombination model as well.

We note that modified point-mass profiles did not produce realistic hazard rates. We did not expect them to do so. Rather, this case was used to test the limits of the Poisson approximation by utilizing much less realistic models for mutation effects. However, it is encouraging that with gamma profiles for mutation effects, we are able to reproduce a range of plateaus by simply changing the range of gamma shape parameters. By restricting shape parameters to smaller values, we can produce hazard rates that decrease for the oldest ages. For example, in Case 6, in which shape parameters range from 1.0 to 3.5, the hazard rate increases until around middle age and then decreases to nearly its original level by the age of 120 years. Allowing shape parameters to take on larger values can produce little to no leveling in hazard rates. In Case 5, where the shape parameters range from 3.5 to 7.0, the hazard rates increased steadily until very late ages (around 100) and then appeared to plateau. However, the hazard rates in this case did not decrease, even in the oldest ages. This range of behaviors, from hazard rates that increase and then decrease to hazard rates that increase and then plateau, suggests that the gamma model for mutation effects is very flexible and could be used in practice.

Bibliography

[1] Ellen Baake. Mutation and recombination with tight linkage. *Journal of Mathematical Biology*, 42(5):455–488, May 2001.

[2] Michael Baake and Ellen Baake. An exactly solved model for mutation, recombination and selection. *Canadian Journal of Mathematics*, 55(1):3–41, 2003.

[3] N. H. Barton and Michael Turelli. Natural and sexual selection on many loci. *Genetics*, 127(1):229–255, January 1991.

[4] Stephen P. Brooks. Quantitative convergence assessment for markov chain monte carlo via cusums. *Statistics and Computing*, 8(3):267–274, August 1998.

[5] Stephen P. Brooks and Gareth O. Roberts. Convergence assessment techniques for markov chain monte carlo. *Statistics and Computing*, 8(4):319–335, December 1998.

[6] Reinhard Bürger. *The Mathematical theory of selection, recombination, and mutation*. Mathematical and Computational Biology. John Wiley and Sons, Chichester, New York, 2000.

[7] Brian Charlesworth. *Evolution in age-structured populations*. Cambridge Studies in Mathematical Biology. Cambridge University Press, second edition, 1994.

[8] Brian Charlesworth. Patterns of age-specific means and genetic variances of mortality rates predicted by the mutation-accumulation theory of ageing. *Journal of Theoretical Biology*, 210(1):47–65, May 2001.

[9] Mary Kathryn Cowles and Bradley P. Carlin. Markov chaine monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, June 1996.

[10] Rick Durrett. *Probability models for DNA Sequence Evolution*. Springer-Verlag, New York, 2002.

[11] Steven N. Evans, David Steinsaltz, and Kenneth W. Wachter. A mutation-selection model for general genotypes with recombination, September 2006.

[12] Warren J. Ewens. *Mathematical Population Genetics I: Theoretical Introduction*. Springer-Verlag, New York, second edition, 2004.

[13] Leonid A. Gavrilov and Natalia S. Gavrilova. *The Biology of Life Span: A Quantitative Approach*. Harwood Academic Publishers, Chur, Switzerland, 1991.

[14] Leonid A. Gavrilov and Natalia S. Gavrilova. The reliability theory of aging and longevity. *Journal of Theoretical Biology*, 213(4):527–545, December 2001.

[15] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, second edition, 2004.

[16] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, November 1992.

[17] J. B. S. Haldane. The effect of variation of fitness. *The American Naturalist*, 71(735):337–349, 1937.

[18] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[19] Olav Kallenberg. *Foundations of Modern Probability*. Probability and its Applications. Springer-Verlag, New York, second edition, 2002.

[20] Motoo Kimura and Takeo Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54(6):1337–1351, December 1966.

[21] J. F. C. Kingman. A simple model for the balance between selection and mutation. *Journal of Applied Probability*, 15(1):1–12, March 1978.

[22] Peter Kraft and David J. Hunter. Genetic risk prediction – are we there yet? *New England Journal of Medicine*, 360(17):1701–1703, April 2009.

[23] Jun S. Liu, Faming Liang, and Wing Hung Wong. The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134, March 2000.

[24] Steven N. MacEachern and L. Mark Berliner. Subsampling the gibbs sampler. *The American Statistician*, 48(3):188–190, August 1994.

[25] Peter B. Medawar. *An Unsolved Problem of Biology*. H.K. Lewis & Co., London, 1952.

[26] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[27] Anand Patil, David Huard, and Christopher J. Fonnesbeck. Pymc: Bayesian stochastic modelling in python. *Journal of Statistical Software*, 35(4):1–81, July 2010.

[28] Adrian E. Raftery and Steven M. Lewis. How many iterations in the gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 763–773. Oxford University Press, 1992.

[29] Skipper Seabold, Josef Perktold, and Jonathan Taylor. Statsmodels' project page, 2010.

[30] David Steinsaltz, Steven N. Evans, and Kenneth W. Wachter. A generalized model of mutation-selection balance with applications to aging. *Advances in Applied Mathematics*, 35(1):16–33, July 2005.

[31] Shripad Tuljapurkar. The evolution of senescence. In Kenneth W. Wachter and Caleb E. Finch, editors, *Between Zeus and the Salmon: The Biodemography of Longevity*, chapter 4, pages 65–77. National Academy Press, 1997.

[32] Michael Turelli and N. H. Barton. Dynamics of polygenic characters under selection. *Theoretical Population Biology*, 38(1):1–57, August 1990.

[33] James W. Vaupel, James R. Carey, Kaare Christensen, et al. Biodemographic trajectories of longevity. *Science*, 280(5365):855–860, May 1998.

[34] Kenneth W. Wachter. Hazard curves and lifespan prospects. In James R. Carey and Shripad Tuljapurkar, editors, *Life Span: Evolutionary, Ecological and demographic perspectives*, volume 29 of *Population and Development Review*, pages 270–291. Population Council, 2003.

[35] Kenneth W. Wachter, Steven N. Evans, and David Steinsaltz. The age-specific force of natural selection and walls of death. Submitted, June 2008.

[36] Kenneth W. Wachter and Caleb E. Finch, editors. *Between Zeus and the Salmon: The Biodemography of Longevity*. National Academy Press, 1997.

[37] Kenneth W. Wachter, David R. Steinsaltz, and Steven N. Evans. Vital rates from the action of mutation accumulation. *Journal of Population Ageing*, 2(1):5–22, 2009.

[38] George C. Williams. Pleiotropy, natural selection, and the evolution of senescence. *Evolution*, 11:398–411, 1957.

[39] Anatoli I. Yashin, Ivan A. Iachine, and Alexander S. Begun. Mortality modeling: A review. *Mathematical Population Studies: An International Journal of Mathematical Demography*, 8(4):305–332, October 2000.

[40] Bin Yu and Per Mykland. Looking at markov samplers through cusum path plots: a simple diagnostic idea. *Statistics and Computing*, 8(3):275–286, August 1998.

Appendix A

MCMC Convergence

A.1 Convergence

In this work, we are interested in determining how similar the distribution of genotypes under the SEW model is to a Poisson random measure. Attempting to answer a question so very vague naturally requires one to first refine the question. "How similar is similar enough?" and "Similar under what measure?" are two natural refinements. "Is the distribution of the total number of mutations per genotype Poisson/approximately Possion?" and "Are the number of copies of different mutation types independent?" also immediately spring to mind. However, given the flexibility of the no recombination and free recombination models and the ultimate goal of determining demographic outcomes such as lifespan and the expected hazard rate for the population modeled, we can further refine our questions. Even finding that the distribution of genotypes is not exactly Poisson may not make much practical difference in the demographic outcomes we measure. Regardless, it should be clear that whatever avenue we choose to refine our central question, much relies on our ability to accurately sample from the distribution of genotypes under the SEW mutation-selection.

A series solution to the distribution of genotypes under the SEW model was stated in equation (2.6). Given the difficulty of directly estimating the solution using the series expansion (which involves computing expectations over the order of arrival of mutations for each genotype), it is necessary to use some other method to estimate the distribution. We have chosen to use the multiple-try Metropolis algorithm, which defines a Markov chain with the appropriate limiting distribution that crawls over the space of time-ordered genotypes. The sampled genotypes are, in turn, used to estimate a variety of outcomes. For example, the sampled genotypes are used to estimate the distribution of the total number of mutations, the average number of mutations per genotype, the average number of copies of each mutation type, and the expected population lifespan, to name a few.

Except for methods of "perfect" sampling (where one samples exactly from the target distribution), samples obtained via Markov chain Monte Carlo methods contain error. The

central idea behind MCMC is to construct a chain whose stationary distribution is the distribution from which you wish to sample and to run this chain for long enough that it has forgotten its starting state and is close to its stationary distribution. Naturally, we would like to have some way of checking whether that has actually happened. There is a great deal of literature regarding this topic and we will not attempt to reproduce it here. Rather, we will briefly discuss a small number of *diagnostic* methods that are applied to samples from the chain in question (see [9] and [5] for summaries of a number of methods, including some that are not applicable to the problem at hand). From the literature it is clear that there is uncertainty within the community as to how useful these techniques actually are and we have been cautious not to rely too heavily on the output of any one diagnostic.

A.1.1 One chain or multiple?

Even a cursory search for MCMC convergence diagnostics will show a division between the camps that suggest using multiple chains, start at different locations in the parameter space (possibly based on an over-dispersed estimate of the sampling distribution) and a single chain, run for a very long time. In this work, we have chosen to consider several methods from the second camp: a single chain run for a very long time. In part this is because the MTM algorithm, as we have implemented it, takes very small steps (producing a genotype that is at most one mutation different from the genotype of the current iteration). Taking small steps keeps the acceptance rate fairly high and has proven to be easy to implement. Unfortunately, it also means that the chain explores the space of possible genotypes very slowly. As a result, a single long run of the chain will have a better chance of covering the space of genotypes than several chains with much shorter runs.

A.1.2 Geweke

Geweke's diagnostic uses a single long run of the Markov chain and compares the mean of some function g of the output from the beginning of the run to the mean from the end of the run. Although the algorithm is generally described in terms of its application to a Gibbs sampler, the method can be applied to any MCMC output (see [9]). As such, the description of the algorithm provided here will not assume use of a Gibbs sampler.

Applying the function g to the state of the Markov chain at each step creates a time series, $g(X_i)$. The expectation $\mathbb{E}[g(X)]$ can be estimated by averaging over n steps of the chain,

$$\bar{g}_n = \frac{\sum_{i=1}^{n} g(X_i)}{n}.$$

This estimator has an asymptotic variance given by $S_g(0)/n$ where $S_g(\omega)$ is the spectral density for the time series $g(X_i)$. If the Markov chain has converged to the stationary

distribution then the estimate of the mean from the beginning of the chain should not be significantly different from the estimate of the mean from the end of the chain.

The estimate $\bar{g}_{n_A}$ is based on the first n_A samples and has asymptotic variance $S_g^A(0)$, while the estimate $\bar{g}_B$ uses the last n_B samples and has asymptotic variance $S_g^B(0)$. We assume that $n_A + n_B < n$, where n is the number of samples. Geweke suggested setting the parameters n_A and n_B are to be $0.1n$ and $0.5n$, respectively. Assuming that the chain is stationary and the ratios n_A/n and n_B/n are fixed, the distribution of the test statistic

$$Z_n = \frac{\bar{g}_{n_A} - \bar{g}_{n_B}}{\sqrt{S_g^A(0)/n_A + S_g^B(0)/n_B}}$$

approaches a standard normal as $n \to \infty$. Thus, for large enough samples, this statistic can be used to test the null hypothesis that the chain has converged by time n.

For this work there are several obvious advantages of Geweke's method as a diagnostic tool. First, the method can be applied to any MCMC output, not just output from a Gibbs sampler. Second, it uses a single chain as opposed to methods (such as the popular method by Gelman and Rubin [16]) that require multiple, independent chains. Third, it is common enough to be included in many standard MCMC packages, including PyMC, a Markov chain Monte Carlo package for python (see [27]). Unfortunately for us, Geweke's method can be used to show that the chain *has not converged* but cannot be used to show that it has converged. Knowing that the chain has not converged (meaning that the difference in means is statistically significant) is useful because it indicates that the chain must be run longer. However, failing to reject the null hypothesis does not give us much information. It could indicate that the chain has converged or it could indicate that the chain is mixing slowly and is stuck in one region of the space.

A.1.3 CUSUM Plots

Yu and Mykland [40] propose a graphical method for assessing convergence from a single chain of length n. The algorithm requires that another method be used to determine the burn-in period, n_0, for the chain before taking samples. The samples are used to create a CUSUM (or partial sum) plot of some summary statistic of the output. Let $\hat{\mu}$ be the estimate of the mean of the summary statistic T, based on the samples after the burn-in period:

$$\hat{\mu} = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} T(X_i).$$

The CUSUM is the sum of the differences between the summary statistic for each sample and the mean $\hat{\mu}$,

$$\hat{S}_t = \sum_{\tau=n_0+1}^{t} (T(X_\tau) - \hat{\mu})$$

for $t = n_0 + 1, \ldots n$. The $\hat{S}_t$ are plotted against t to create the CUSUM plot.

The smoothness of the CUSUM plot indicates whether or not the chain is mixing slowly. The smoother the plot, the slower the mixing of the chain. Yu and Mykland suggested using iid samples from a Normal distribution with mean and variance set equal to the sample mean and sample variance as a benchmark. The iid samples, according to Yu and Mykland, approximate the "ideal" CUSUM path for an iid sequence from the limiting distribution. As a result, similar amounts of smoothness and wandering away from zero between the two paths indicates good mixing.

This method can be made slightly less qualitative (since graphical interpretations of "similarity" and "hairiness" between the paths are quite subjective) by defining the "hairiness" of the path by the following process (see [4]). First, compute

$$d_t = \begin{cases} 1 & \text{if } S_{t-1} > S_t \text{ and } S_t < S_{t+1} \\ & \text{or } S_{t-1} < S_t \text{ and } S_t > S_{t+1} \\ 0 & \text{else} \end{cases} \tag{A.1}$$

for $t = n_0 + 1 \ldots n - 1$. Then, the average of the d_t

$$D_{n_0,n} = \frac{1}{n - n_0 - 1} \sum_{t=n_0+1}^{n-1} d_t \tag{A.2}$$

is a number between 0 and 1, where 0 indicates that the plot is completely smooth and 1 indicates that the plot is "hairy." For large samples, $D_{n_0,n}$ will be approximately normal with mean $\frac{1}{2}$ and variance $\frac{1}{4(n-n_0-1)}$, allowing us to test for non-convergence of the chain. A value of $D_{n_0,n}$ outside of $\frac{1}{2} \pm Z_{\alpha/2}\sqrt{\frac{1}{4(n-n_0-1)}}$ indicates non-convergence.

However, as Brooks points out in [4], this test rests on the assumptions that the samples $T(X_i)$ are iid and distributed symmetrically about the mean. Because MCMC samplers generally produce dependent samples, the first assumption will certainly be violated in practice. To attempt to remove the dependence on the first assumption, Brooks notes that the chain can be made approximately independent by thinning the samples, a dubious solution at best. An alternative to thinning that may be used for sticky chains (where the parameter may remain in the same state for several iterations) is to use a different definition for d_t, given by

$$d_t = \begin{cases} 1 & \text{if } S_{t-1} > S_t \text{ and } S_t < S_{t+1} \\ & \text{or } S_{t-1} < S_t \text{ and } S_t > S_{t+1} \\ & \text{or } S_{t-1} < S_t, S_{t+k} > S_t \text{ and } S_t = S_{t+1} = \cdots S_{t+k} \text{ for } k \in \mathbb{Z} \\ & \text{or } S_{t-1} > S_t, S_{t+k} < S_t \text{ and } S_t = S_{t+1} = \cdots S_{t+k} \text{ for } k \in \mathbb{Z} \\ \frac{1}{2} & \text{if } S_{t-1} = S_t = S_{t+1} \\ 0 & \text{else} \end{cases} \tag{A.3}$$

Brooks does not recommend using the alternative definition due to its added computational costs. For the chains we wish to analyze, it is often the case that the chain will move to a different genotype with the same total number of mutations. If we define $T(X)$ to be the total number of mutations per genotype then we expect there to be many instances in which $T(X_t) = T(X_{t+1})$ Unfortunately, while it may be true that $T(X_t) = T(X_{t+1})$, it will not generally be true that $S_t = S_{t+1}$. As a result, the expanded definition of d_t will probably not be worth the additional computation time or power.

The second assumption may be addressed by using the empirically estimated median rather than the mean in computing $\hat{S}_t$, which would ensure that $\mathbb{P}(d_t = 1) = 1/2$. If the mean is used in computing $\hat{S}_t$, $\mathbb{P}(d_t = 1)$ can be computed using the one-step transition density for the Markov chain (see [4] for details). In this case the burn-in period can also be estimated by the following procedure: Given N observations, begin by setting $n = N/20$ and $n_0 = n/2$. Compute $D_{n_0,n}$ to determine if the chain has converged after n/2 iterations. Now repeat the process with the first $2n$ iterations, computing $D_{n,2n}$. Continue the process until all N observations have been used, resulting in the sequence $D_{n/2,n}, D_{n,2n}, \cdots D_{10n,N}$. Plot the sequence to determine if the D appear to approach a single value. If $D_{t/2,t}$ marks the point at which the D generally converge to a single value, then the burn-in period can be estimated by $n_0 = t/2$.

Yu and Mykland provide a nice diagnostic tool for the problem at hand because it can be applied to the output of any MCMC algorithm and because it requires only a single chain. However, it suffers from some of the same problems as Geweke's diagnostic, namely that it cannot be used to show convergence and that it requires another method to be used first to determine the length of the burn-in period. Further, for the chains we wish to analyze, the output may be highly correlated, making the assumption that $\mathbb{P}(d_t = 1) = 1/2$ unlikely to hold. It will also be difficult to calculate $\mathbb{P}(d_t = 1)$.

A.1.4 Raftery and Lewis

The Raftery-Lewis diagnostic [28] assumes that one is interested in computing the $100 \cdot q^{\text{th}}$ quantile of a given parameter θ to within a certain accuracy. Although the diagnostic was first described in terms of the Gibbs sampler, it can applied to any MCMC output.

We denote the $100 \cdot q^{\text{th}}$ quantile by θ_q so that $P(\theta \leq \theta_q) = q$. To estimate the quantile

θ_q, one can simply order n samples of the parameter θ and take the $n \cdot 100 \cdot q^{\text{th}}$ sample as the estimate $\hat{\theta}_q$. As the number of samples grows, the sample distribution converges to the true distribution and the estimated probability $P(\theta \leq \hat{\theta}_q) = \hat{P}_q$ converges to q. This diagnostic allows the user to specify how accurately they wish to estimate $\hat{P}_q$ by setting the parameters r and s. For example, if $q = 0.025$, $r = 0.005$ and $s = 0.95$ then the diagnostic will determine how many samples are required to estimate the 2.5th percentile to within ± 0.005 with probability 95%.

After $\hat{\theta}_q$ is estimated from the samples, a binary sequence $\{Z_t\}$ is defined as follows from the samples: $Z_t = 1$ if $\theta^t \leq \hat{\theta}_q$ and 0 otherwise. The binary sequence is not itself a Markov chain but is approximately a Markov chain if sufficiently thinned. Once the thinning parameter k has been determined, the diagnostic then determines the number of iterations, M, that can be discarded for the burn-in and, finally, the number of additional iterations, N, needed to obtain the desired accuracy. The additional iterations will be thinned by the factor k.

Because one of our primary goals is to characterize the distribution of the total number of mutations per genotype, the Raftery-Lewis diagnostic may be useful. However, the method focuses on convergence for a single quantile rather than convergence of the chain in general. Unfortunately, different quantiles may require different values of k, M and N to achieve the desired accuracy in estimation. When estimating multiple quantiles, Raftery and Lewis suggest running the diagnostic on each quantile separately and choosing the maximum values of k, M and N. In doing so we assure that for each quantile, the estimate will be with r of the true value with probability s. That is, if we test Q quantiles, we expect at least sQ of the estimates to be within r of their true values. Furthermore, the practice of thinning the collected samples is questionable. While thinning the samples reduces autocorrelation, simple subsampling methods (such as retaining every k^{th} sample) reduce the accuracy of the estimated parameter (see [24]).

A.2 Results

As an illustrative example of applying these convergence diagnostics to the output of the MTM algorithm, we will focus on the output from one of the large gamma cases. In particular, we will be applying these methods to the case with 1000 gamma profile mutations with shape parameters ranging from 1.0 to 7.0 (referred to as Case 3 in §4.1). We have chosen to focus on this case because we expect more mutations per genotype on average for this case than for the other three large gamma cases (where the largest shape parameter is 5.0, 5.5 or 6.0). Mutations with larger shape parameters have smaller age effects over the range of fertility (ages 15-50) and thus have much less selective pressure on them than mutations with shape parameters close to 1.0. With less selective pressure, we expect more copies of these mutations in a typical genotype. Because all tests in §4.1 are started in the null genotype and because every proposed genotype contains at most one more mutation than the

current genotype, it will take longer for the chain to explore genotypes with a large number of mutations. As a result, we expect the Markov chain to converge to the limiting distribution more slowly in Case 3 than in the other cases considered. We will provide results of the convergence diagnostics applied to the other large gamma cases in subsequent sections. However, we will not provide the same detailed discussion of the various diagnostics applied to those cases.

A.2.1 Case 3

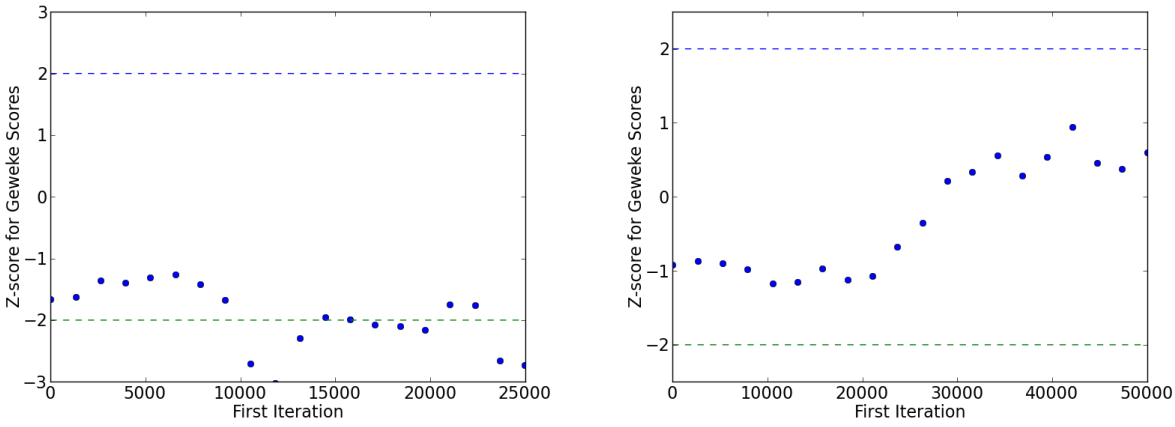

Figure A.1: Geweke scores for the first 50,000 iterations (left) and the first 100,000 iterations (right). Each point is the Geweke score comparing the first 10% of the subchain to the last 50% of the subchain. The score is plotted against the first iteration used in the subchain.

We begin by applying Geweke's diagnostic. Recall that the output from the MTM algorithm, denoted by X_i in §A.1.2, consists of genotypes. Each (unordered) genotype is represented by a vector whose i^{th} element is the number of copies of mutation type i, $X = (n_1, n_2, \ldots n_M)$. We apply Geweke's diagnostic to the total number of mutations per genotype. Using the notation from §A.1.2 we have $g(X) = |X| = \sum_i n_i$. Figure A.1 shows Geweke scores for output from Case 3. These plots were generated using pymc.geweke with each plot containing the Geweke scores for 20 subchains. Each point is the Geweke score comparing the first 10% of the subchain to the last 50% of the subchain. The score is plotted against the "first iteration," which marks the beginning of the subchain used to generate the score. The plot on the left uses the first 50,000 iterations of the MTM output while the plot on the right uses the first 100,000 iterations.

Several of the scores on the left-hand plot are clearly outside of two standard deviations of zero, indicating that the chain has not converged. Recall that the diagnostic assumes

that the chain has reached its stationary distribution and tests the null hypothesis that the difference in means between the first and last part of the chain are zero against the alternative that they are not 0. As a result, we clearly need a longer burn-in than 50,000 iterations. In the plot on the right hand side, all the Geweke scores are within two standard deviations of zero. Although we cannot reject the null hypothesis that the chain has converged, we cannot say that the chain *has* converged.

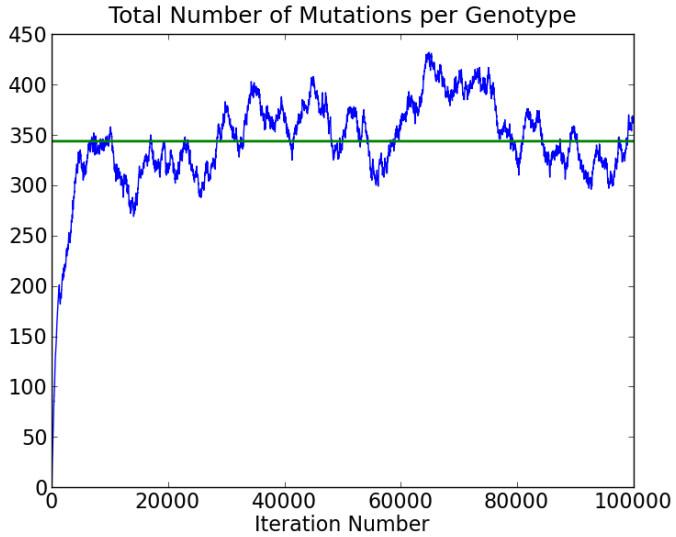

Figure A.2: Total number of mutations per genotype plotted against iteration number.

Yu and Mykland's CUSUM plot for analyzing convergence of the chain requires another method to first determine the length of the burn-in period. The Geweke's scores for the first 50,000 iterations suggest that the chain has not converged in that time. However, when considering the first 100,000 iterations, we cannot reject the hypothesis that the chain has converged. This suggests that we may want to discard several tens of thousands of iterations for the burn-in. To get a better idea of how quickly the chain is traversing the space of genotypes, and how many iterations we may want to discard for the burn-in, we next plot the total number of mutations in a genotype against the iteration number.

Figure A.2 shows the number of mutations per genotype for the first 100,000 iterations of one run of the MTM algorithm when the chain was started in the null genotype. The number of mutations per genotype increases fairly steadily from its starting point of 0 to around 300 mutations per genotype by 5,000 iterations. Between 5,000 and 28,000 iterations the chain explores genotypes containing 275 to 350 mutations. For the next 30,000 iterations the chain generally explores genotypes with containing even more mutations, roughly 300 to 400 mutations per genotype. In the next 20,000 iterations, the chain explores still larger

genotypes, these containing between 350 and 425 mutations. Then, in the final 20,000 iterations shown here, the chain returns to genotypes containing 300 to 350 mutations.

This plot provides several insights to the movement of the chain. First, it suggests that at least the first 10,000 iterations should be discarded as burn-in. It also shows the chain is mixing fairly slowly. This is not terribly surprisingly because at each step the proposed genotype will differ from the current genotype by at most one mutation. That means that it will take the chain many iterations to move from genotypes with 300 mutations to genotypes with 400 mutations. The plot also shows that there are long periods where the number of mutations per genotype is generally increasing (such as between iterations 56,000 to 65,000) or generally decreasing (such as between iterations 75,000 to 91,000). Although output from MCMC algorithms are generally correlated, this plot also suggests that the samples will have high autocorrelation even over long lag periods. This is confirmed by Figure A.3. The autocorrelation was plotted after discarding the first 10,000 iterations. High autocorrelation over a long lag suggests that the chain is not mixing very well.

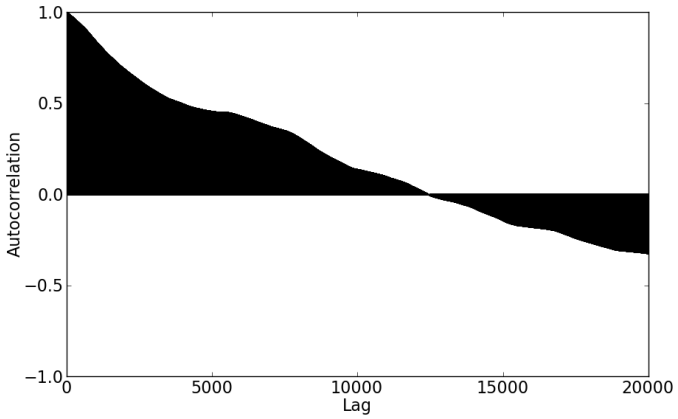

Figure A.3: Autocorrelation after discarding the first 10,000 iterations.

We next generated CUSUM plots for the output. As with the Geweke's scores, we used the total number of mutations per genotype as the summary statistic for the output from the MTM algorithm. We generated two CUSUM plots, both shown in Figure A.4. In the plot on the left hand side the first 10,000 iterations were discarded as burn-in and the plot was generated with the next 40,000 iterations. In the plot on the right the first 10,000 iterations were discarded and the plot was generated with the next 90,000 iterations. The smoothness of the CUSUM plot for the MTM samples provides further evidence that the chain is mixing slowly. Each plot also shows a "benchmark" CUSUM path, generated by an iid sequence from a Normal distribution with the same mean and variance as the MTM output. In both plots, the benchmark path stays much closer to 0 than the CUSUM path for the MTM data.

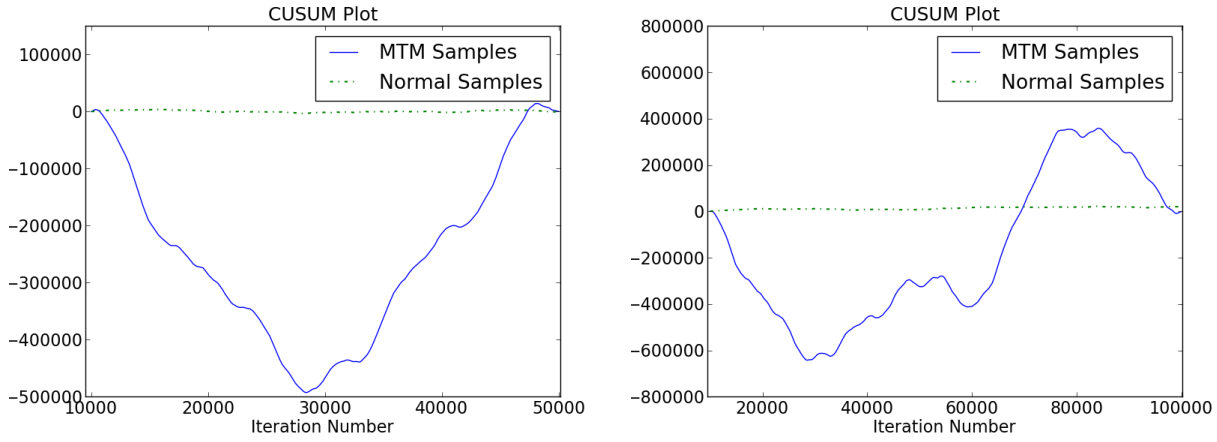

Figure A.4: CUSUM plot after discarding the first 10,000 iterations as burn-in. The plot on the left is for the next 40,000 iterations. The plot on the right is for the 90,000 iterations after the burn-in period.

Naively testing the hairiness of each plot, we find that $D_{n_0,n_1} = 0.002$ when $n_0 = 10,000$ is the burn-in period and $n_1 = 50,000$. This value of D falls quite far outside of the range $(0.4951, 0.5049)$, indicating that the chain has not converged. For the second case, $D_{n_0,n_2} = 0.0028$ when $n_0 = 10,000$ and $n_2 = 100,000$. Again, this value is outside of the range $(0.49673, 0.50327)$, indicating non-convergence of the chain. However, the degree to which the output is autocorrelated clearly breaks the assumption of independence required for this test, making the results uninformative.

We next followed the suggestion of Brooks [4] and compute a sequence of $D_{n_0,n}$ to determine if the D statistics appear to approach a single value. For this test we let $N = 500,000$, $n = N/20 = 25,000$ and $n_0 = n/2$. Following the procedure described in A.1.3, we obtained a sequence of D statistics which are plotted in Figure A.5 against n. The sequence of D statistics appears to oscillate between 0.00388 and 0.0046 for n larger than 250,000. However, given the large number of mutations used to calculate these statistics, and the fact that the D statistics are very small in general, this does not appear to be evidence of convergence.

Of course, the total number of mutations per genotype is only one statistic that we could use to assess convergence. We are also interested in the number of mutations falling within various ranges, such as the number of mutations per genotype with shape parameter in 1.0 to 1.5. Categorizing mutations by shape parameters in ranges of size 0.5 (so $1.0 \leq m \leq 1.5$, $1.5 \leq m \leq 2.0$, etc), we performed the same procedure for each subset of mutations. The D statistics for subsets of mutations with shape parameters less than 3.0 are shown in Figure A.6; mutations with shape parameters between 3.0 and 5.0 are shown in Figure A.7; mutations with shape parameters greater than 5.0 are shown in Figure A.8. As with the D

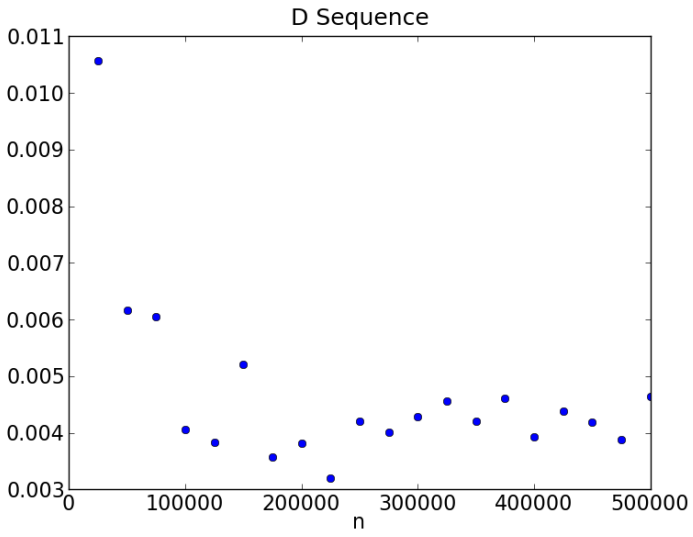

Figure A.5: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$.

statistics for the total number of mutations per genotype, the D statistics in these marginal cases are very small. Even in cases where the D statistics may visually appear to be oscillating around a single value for large enough n, say $n > 250,000$, the D still oscillate by as much as 25% of their lowest value over that range. For example, in Figure A.6, bottom left, where $2.0 \leq m \leq 2.5$, the D statistics appear to oscillate between 0.009 and 0.0011 for $n > 200,000$. However, this difference of 0.002 is fairly large considering the size of the D statistics in this range. Similarly, in Figure A.8 top right, where $5.5 \leq m \leq 6.0$, the D statistics range from 0.0029 to 0.0033 for $n > 275,000$. The difference of 0.0004 is still about 14% of the smallest D statistic over this range.

One of the primary problems in assessing convergence with the CUSUM approach is that the chain may have the same *total* number of mutations for several iterations, even if it has accepted moves to other states. That is, there are many genotypes with the same number of mutations. In addition, the chain only takes small steps, changing a genotype by at most one mutation each iteration. This means that the chain will have fairly long excursions above the empirical mean number of mutations per genotype, as well as below it. This is most evident when looking at the total number of mutations over 500,000 iterations, shown in Figure A.9. The empirical mean number of mutations computed from all 500,000 iterations is also plotted. The CUSUM diagnostic tells us that the chain is mixing very slowly and will require many iterations to satisfactorily explore the genotype space.

The final diagnostic tool that we apply to the output for Case 3 is the Raftery-Lewis diagnostic. We tested nine quantiles, $q \in \{0.1, 0.2, \cdots 0.9\}$, each with r = 0.05 and s = 0.95, see §A.1.4 for detailed descriptions of these parameters. The diagnostics were applied

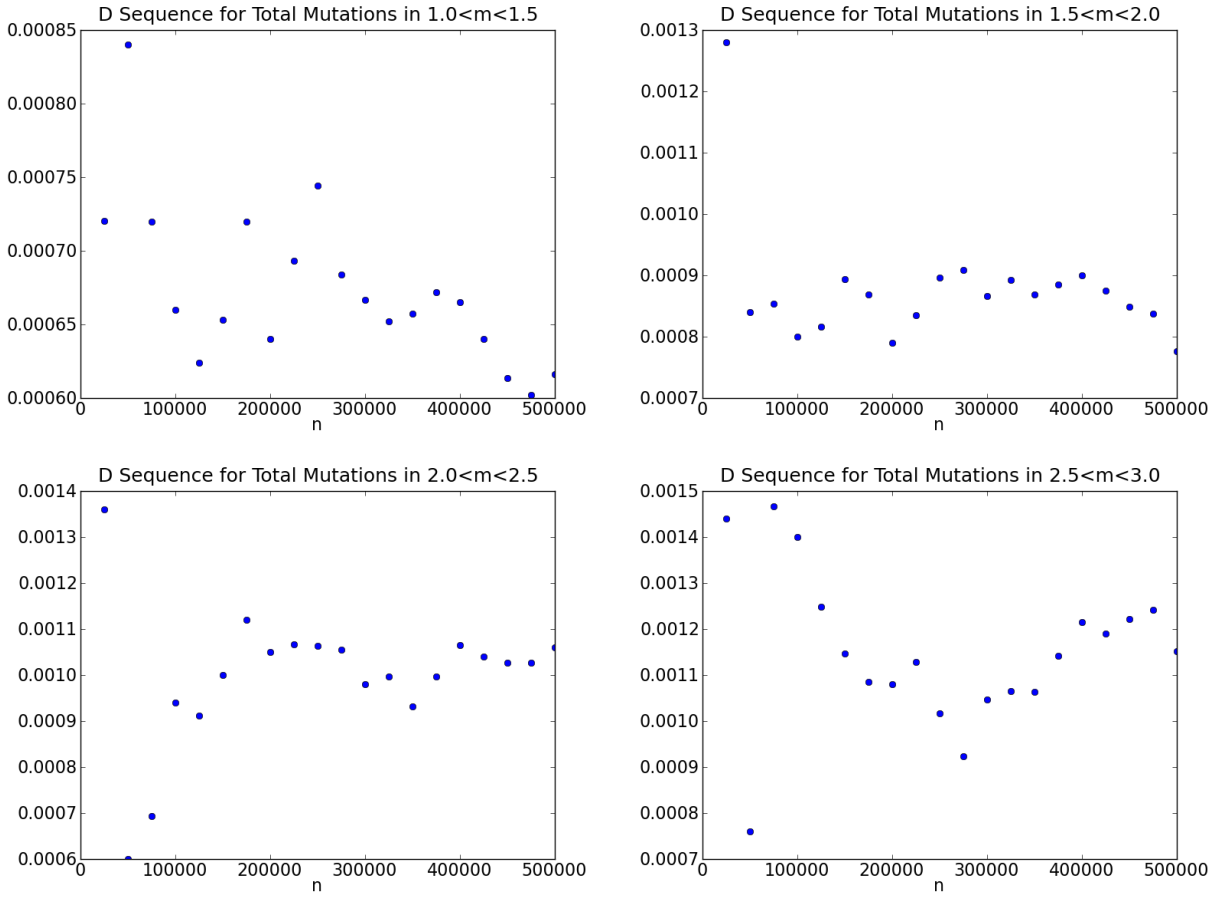

Figure A.6: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the number of mutations in the following intervals, $1.0 \leq m \leq 1.5$ (top left), $1.5 \leq m \leq 2.0$ (top right), $2.0 \leq m \leq 2.5$ (bottom left), $2.5 \leq m \leq 3.0$ (bottom right).

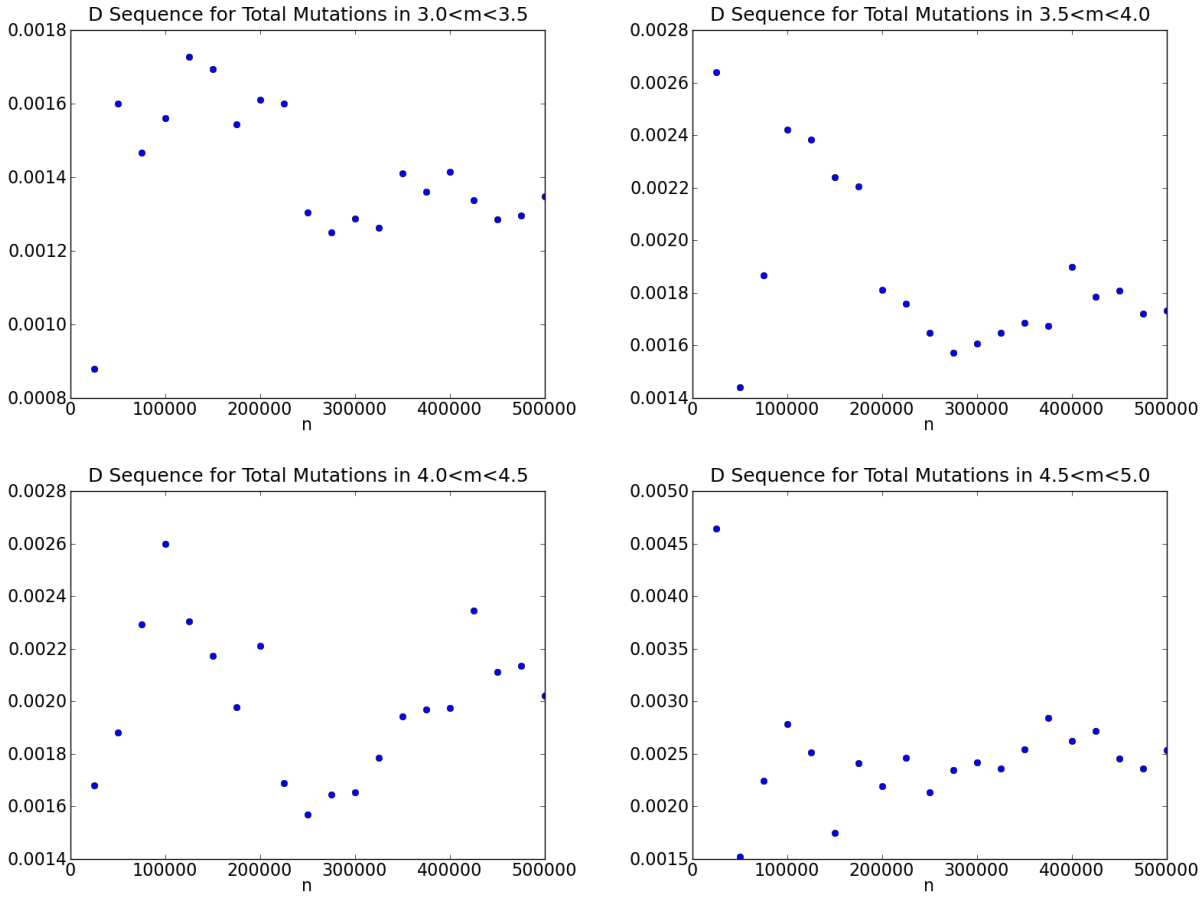

Figure A.7: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the number of mutations in the following intervals, $3.0 \leq m \leq 3.5$ (top left), $3.5 \leq m \leq 4.0$ (top right), $4.0 \leq m \leq 4.5$ (bottom left), $4.5 \leq m \leq 5.0$ (bottom right).

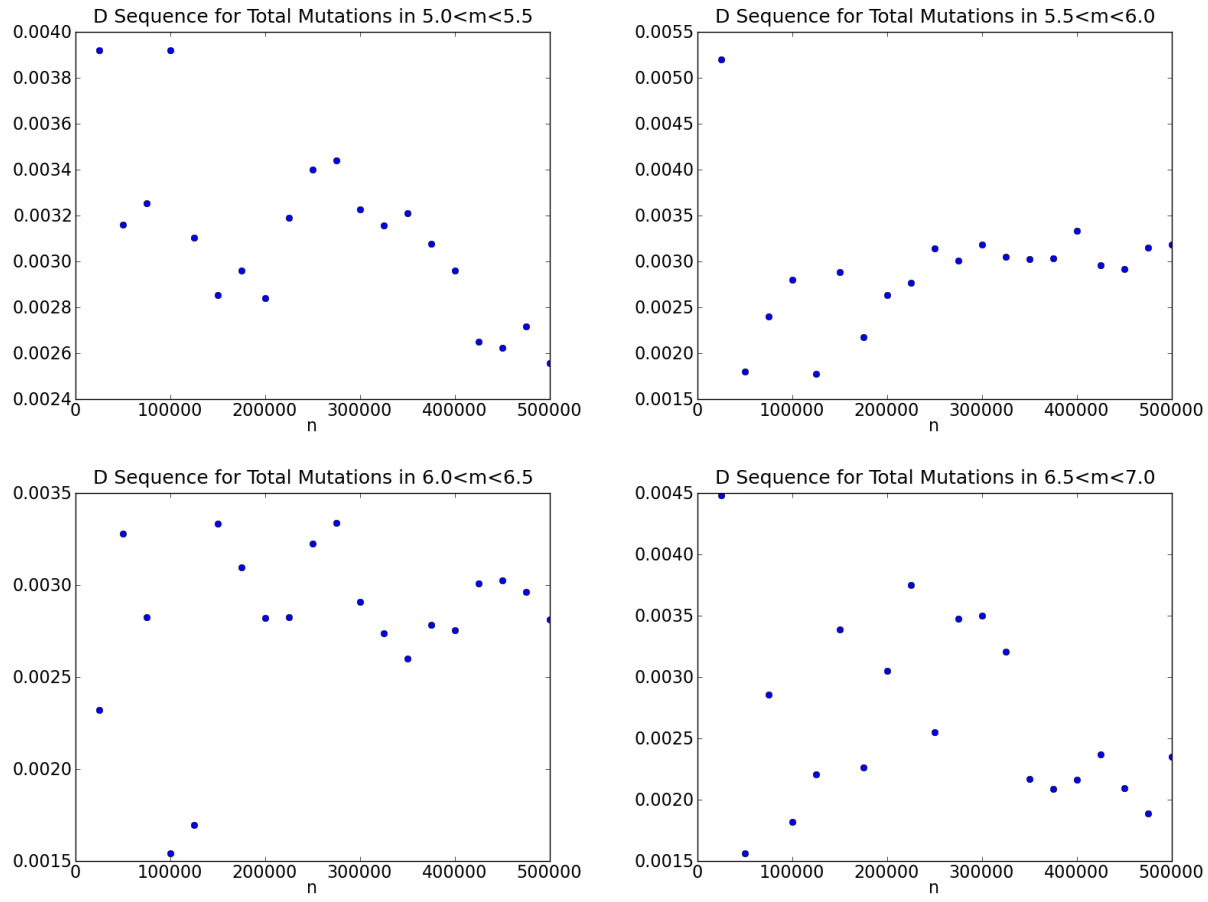

Figure A.8: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the number of mutations in the following intervals, $5.0 \leq m \leq 5.5$ (top left), $5.5 \leq m \leq 6.0$ (top right), $6.0 \leq m \leq 6.5$ (bottom left) and $6.5 \leq m \leq 7.0$ (bottom right).

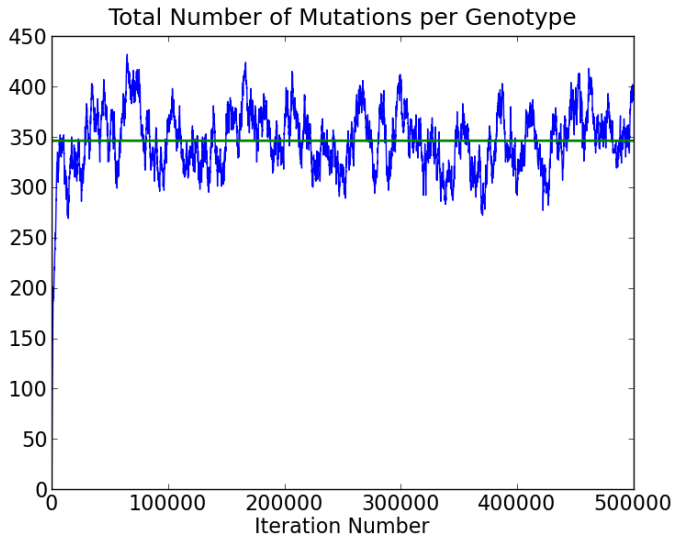

Figure A.9: Total number of mutations per genotype plotted against iteration number. The empirical mean number of mutations per genotype (shown in red) was computed using all 500,000 iterations.

to two subsets of a run of the MTM algorithm: the first 50,000 iterations and the first 100,000 iterations. The results are shown in Table A.1. Although the estimates of M, the length of the burn-in period, and N, the additional number of mutations needed to obtain the desired accuracy, are different for each quantile (and for the two subsets of data), the results are reasonably consistent. For example, when using only the first 50,000 iterations, the diagnostic suggests a burn-in period of 2000-9600 iterations, while the suggested burn-in period is 3000-12000 when the diagnostic is applied to the first 100,000 iterations. The additional number of iterations needed ranges from a low of about 80,000 to almost 100,000 when applied to the first 50,000 iterations. When the diagnostic is applied to the first 100,000 iterations, the additional number of iterations ranges from about 120,000 to 1.4 million. Although the upper range in the number of mutations needed is quite high (1.4 million), most quantiles were estimated to require about 600,000 iterations after the burn-in to estimate the given quantile. However, this number assumes that one will thin the data by the factor k, something we have elected not to do.

Overall, the diagnostics applied to data from the MTM algorithm run with Case 3 parameters suggests that the chain crawls over the space of genotypes fairly slowly. As a result, a large number of iterations will be needed to ensure confidence in the conclusions drawn from the data. We also note that when the same diagnostics were applied to data from the other cases considered in §4.1, the results were consistent with what has been presented here. That is, Geweke failed to indicate a lack of convergence, the CUSUM plots suggested

Table A.1: Results from applying the Raftery-Lewis diagnostic to the first 50,000 iterations (left hand side) and 100,000 iterations (right hand side) of the MTM output for Case 3. Detailed descriptions of the parameters can be found in §A.1.4. In all cases $r = 0.05$ and $s = 0.95$; q is the quantile to be estimated, k is the thinning factor required for an independence chain, M in the length of the burn-in period and N is the number of additional iterations needed after the burn-in.

50k Iterations				100k Iterations			
q	k	M	N	q	k	M	N
0.1	1414	6384	221664	0.1	1247	4860	186570
0.2	1410	2856	206592	0.2	1425	4095	290430
0.3	1152	4180	395580	0.3	1616	5576	539806
0.4	1476	5950	674100	0.4	1911	4625	524882
0.5	1289	3185	389697	0.5	1905	6120	755856
0.6	1638	4553	523682	0.6	2441	11931	1386210
0.7	1526	9630	954360	0.7	2153	10519	1045620
0.8	1067	3200	232832	0.8	2629	7900	571328
0.9	718	2064	80352	0.9	1024	3219	126799

that the chain was mixing slowly, the output was highly correlated and the Raftery-Lewis diagnostic suggested a large number of iterations.

A.2.2 Case 1

We will now briefly present the convergence diagnostics applied to Case 1, in which the mutation space consists of 1000 gamma profile mutations with shape parameters ranging from 1.0 to 6.0.

Figure A.10 shows the Geweke scores for 20 subchains. The plot on the left uses the first 50,000 iterations while the plot on the right uses the first 100,000 iterations. In each plot, the empirical mean number of mutations from the first 10% of the subchain is compared to the mean number of mutations in the last 50% of the subchain. Each score is plotted against the first iteration of the subchain. In both plots, all scores lie in the range ± 2, suggesting that the chain fails to show a lack of convergence. As with the previously discussed case (Case 3), the chain moves very slowly from the perspective of the total number of mutations per genotype. This is because many different genotypes will have the same total number of mutations. As a result, the Geweke diagnostic may not detect that the chain has not converged.

Figure A.11 shows the total number of mutations per genotype plotted against iteration number. The plot on the left shows the first 100,000 iterations and the plot on the right shows the first 600,000 iterations of the chain. The chain was started in the null genotype.

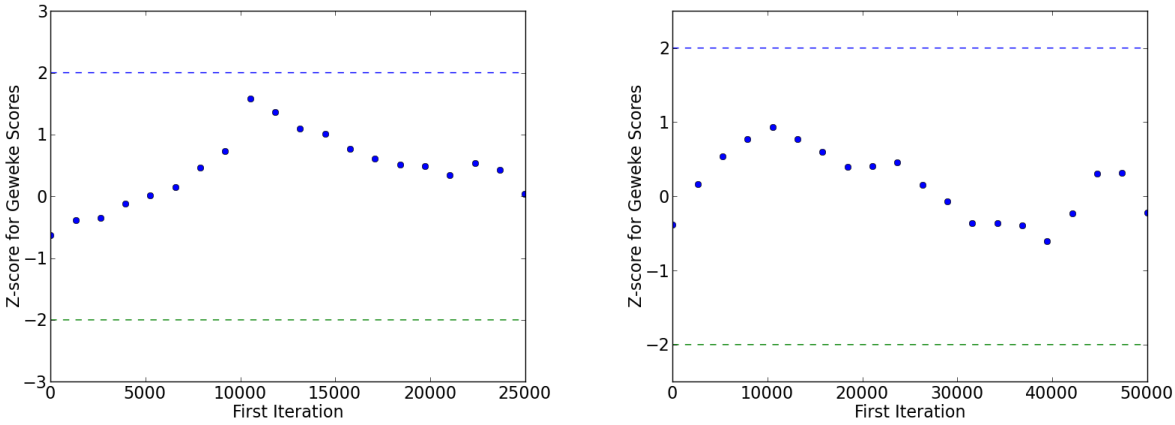

Figure A.10: Geweke scores for the first 50,000 iterations (left) and the first 100,000 iterations (right) of the MTM output for Case 1. Each point is the Geweke score comparing the first 10% of the subchain to the last 50% of the subchain. The score is plotted against the first iteration used in the subchain.

From the plot on the left we see that the total number of mutations per genotype increases from 0 to the range 150-200 mutations fairly quickly. Looking more closely we find that the chain first reaches 100 mutations in iteration 607. The plot on the right shows that the chain primarily stays in the 150-200 mutation range over the 600,000 iterations with brief excursions to genotypes with more than 200 mutations. The horizontal line in the plot on the right indicates the empirical mean number of mutations per genotype over the entire 600,000 iterations.

Figure A.12 shows the autocorrelation in the chain after discarding the first 10,000 iterations for burn-in. Notice that the chain, while still highly correlated, is much less correlated than the output for Case 3. In particular, the output for Case 3 had a consistently high correlation (0.5 or higher) for lags of at least the first 5,000 iterations (after discarding 10,000 iterations as the burn-in period). In this case the correlation drops to around 0.25 after 1,000 or so iterations.

Figure A.13 shows the CUSUM plots for the first 50,000 iterations (left) and the first 100,000 iterations (right). In both cases the first 10,000 iterations were discarded as burn-in. As with Case 3, the CUSUM plots show that the chain is making very long excursions away from the empirically determined mean.

Figure A.14 shows the sequence of $D_{n/2,n}$ plotted against n for $n \in \{30000, 60000, \cdots 600000\}$. While the sequence of D statistics do not approach 0.5, they are consistently in the range 0.0097 to 0.0103 for $n \geq 270000$. Figures A.15-A.17 show the D statistics for the total number of mutations within a given range of shape parameters. In most, but not all, of

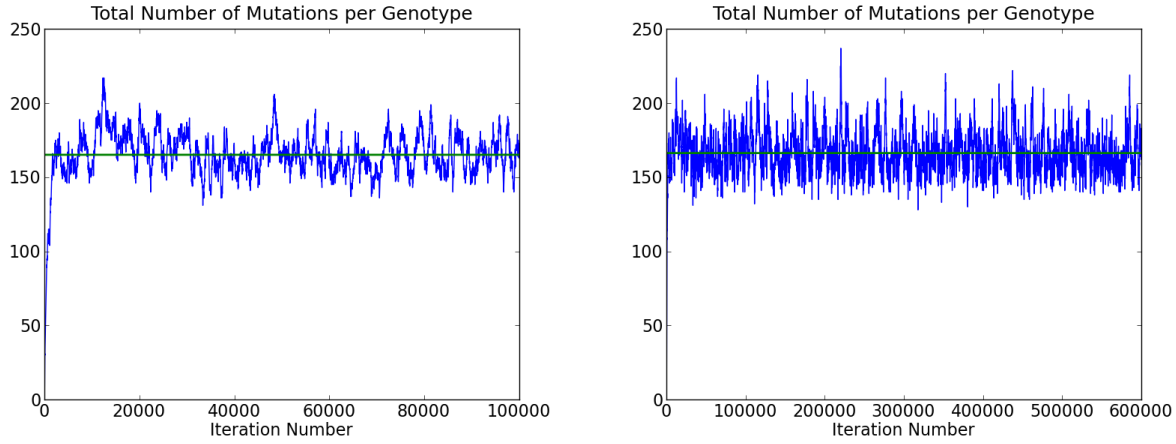

Figure A.11: Total number of mutations per genotype plotted against iteration number.

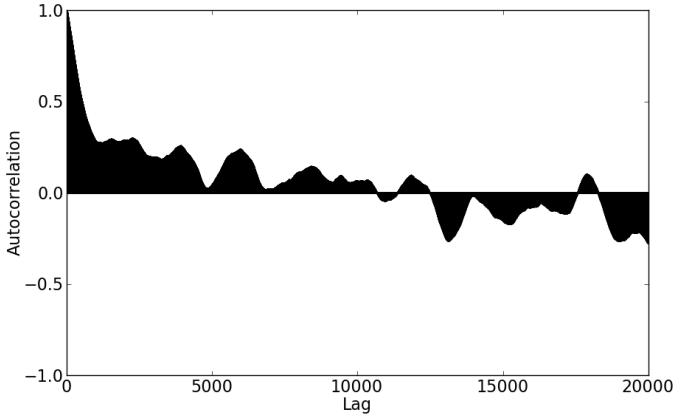

Figure A.12: Autocorrelation of the MTM output for Case 1 after discarding the first 10,000 iterations.

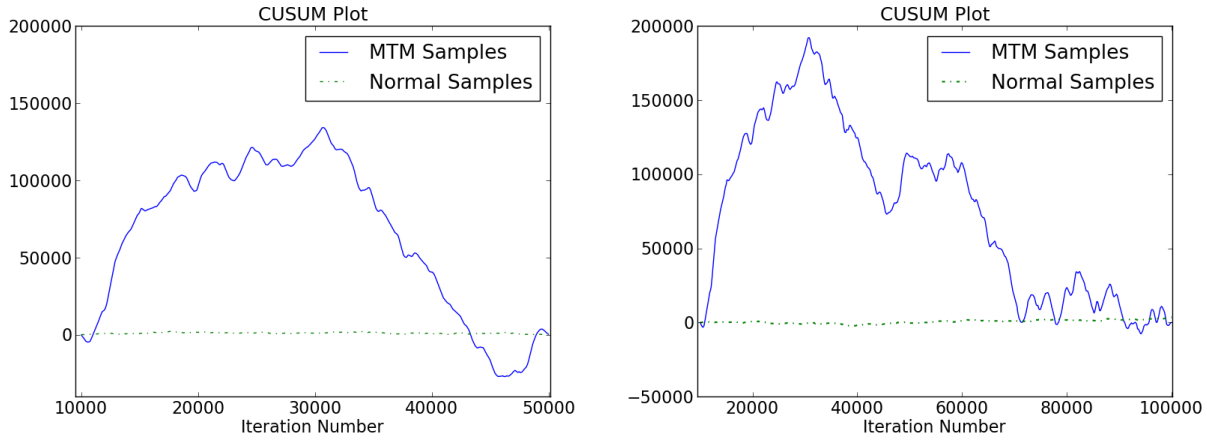

Figure A.13: CUSUM plot for the first 50,000 iterations (left) and first 100,000 iterations (right) after discarding the first 10,000 iterations as burn-in.

these plots, the D statistics are within an interval of width 0.0005 or smaller for sufficiently large n. These plots suggest that burn-in periods of 150,000 or 200,000 iterations may be appropriate.

Table A.2 shows the results of applying the Raftery-Lewis diagnostic to the first 50,000 iterations (left hand side of the table) and to the first 100,000 iterations (right hand side of the table). In all cases, $r = 0.05$ and $s = 0.95$. The quantiles 0.1 to 0.9 in steps of size 0.1 were all tested. The results from the Raftery-Lewis diagnostics were fairly consistent whether using 50,000 iterations or 100,000 iterations. The longest estimated burn-in period was just over 2000 iterations and the longest run for additional samples was just under 250,000.

Finally, Figure A.18 shows the Geweke scores for iterations 150,000 to 250,000 (left) and for iterations 200,000 to 300,000 (right). Both plots fail to show a lack of convergence over these iterations. From the above analysis we conclude that a burn-in period of 150,000 iterations should be sufficient and we should collect at least 350,000 samples.

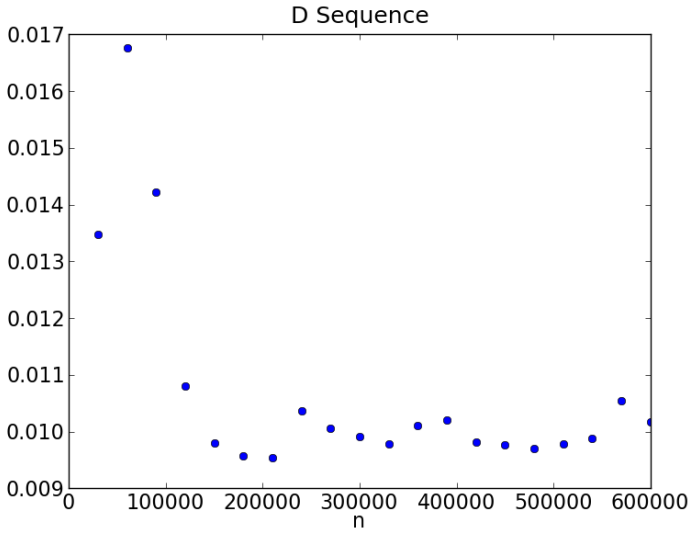

Figure A.14: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{30000, 60000, \cdots 600000\}$.

Table A.2: Results from applying the Raftery-Lewis diagnostic to the first 50,000 iterations (left hand side) and 100,000 iterations (right hand side) of the MTM output for Case 1. Detailed descriptions of the parameters can be found in §A.1.4. In all cases $r = 0.05$ and $s = 0.95$; q is the quantile to be estimated, k is the thinning factor required for an independence chain, M in the length of the burn-in period and N is the number of additional iterations needed after the burn-in.

50k Iterations				100k Iterations			
q	k	M	N	q	k	M	N
0.1	339	968	42504	0.1	349	901	37577
0.2	339	1200	88300	0.2	373	1100	87150
0.3	420	1344	132992	0.3	497	1496	150128
0.4	498	1820	212870	0.4	550	1564	183192
0.5	563	2178	265056	0.5	516	2071	244378
0.6	572	1620	183360	0.6	612	2046	231384
0.7	590	2160	200880	0.7	591	2346	227700
0.8	539	1232	86768	0.8	539	2096	139515
0.9	397	1395	55935	0.9	424	1428	51663

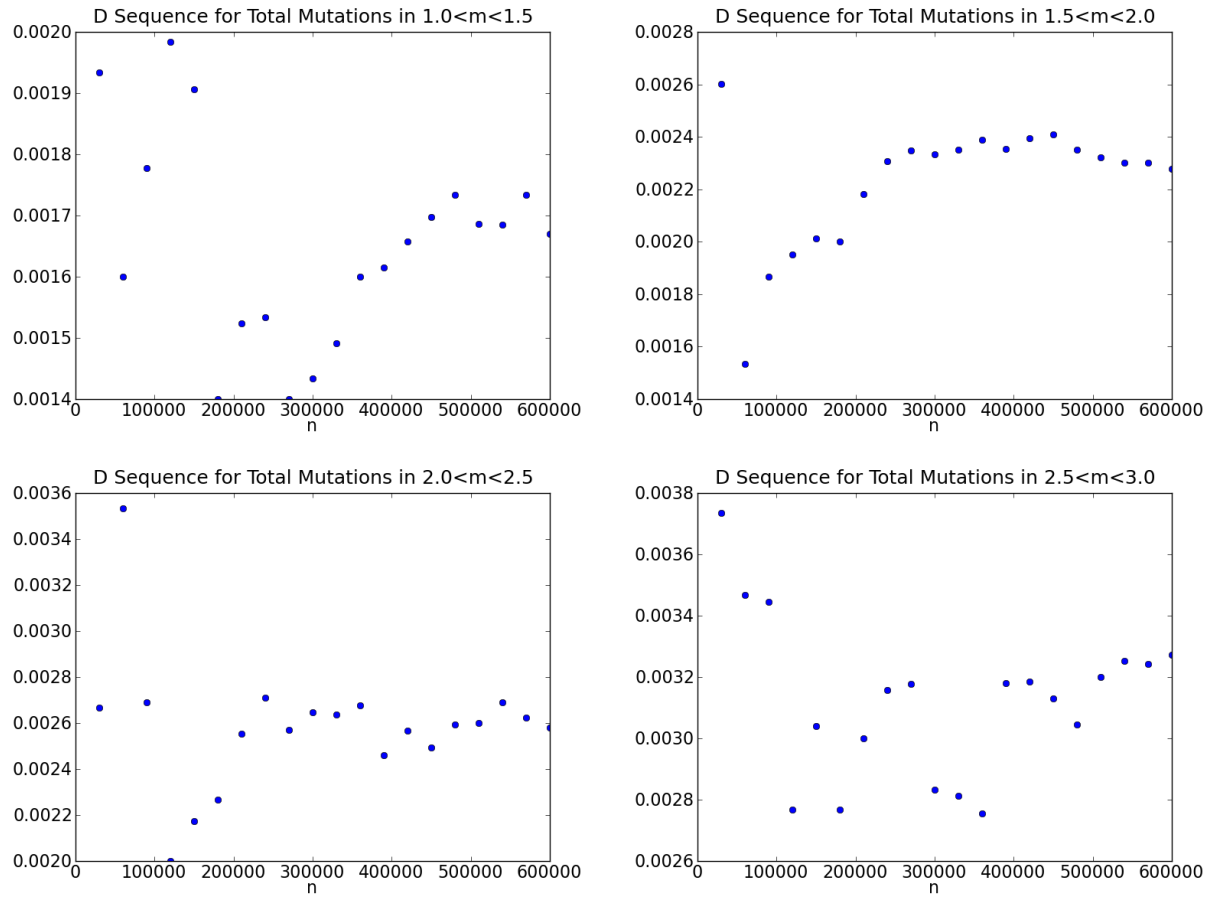

Figure A.15: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{30000, 60000, \cdots 600000\}$. The D are calculated from the total number of mutations in the following intervals, $1.0 \leq m \leq 1.5$ (top left), $1.5 \leq m \leq 2.0$ (top right), $2.0 \leq m \leq 2.5$ (bottom left), $2.5 \leq m \leq 3.0$ (bottom right).

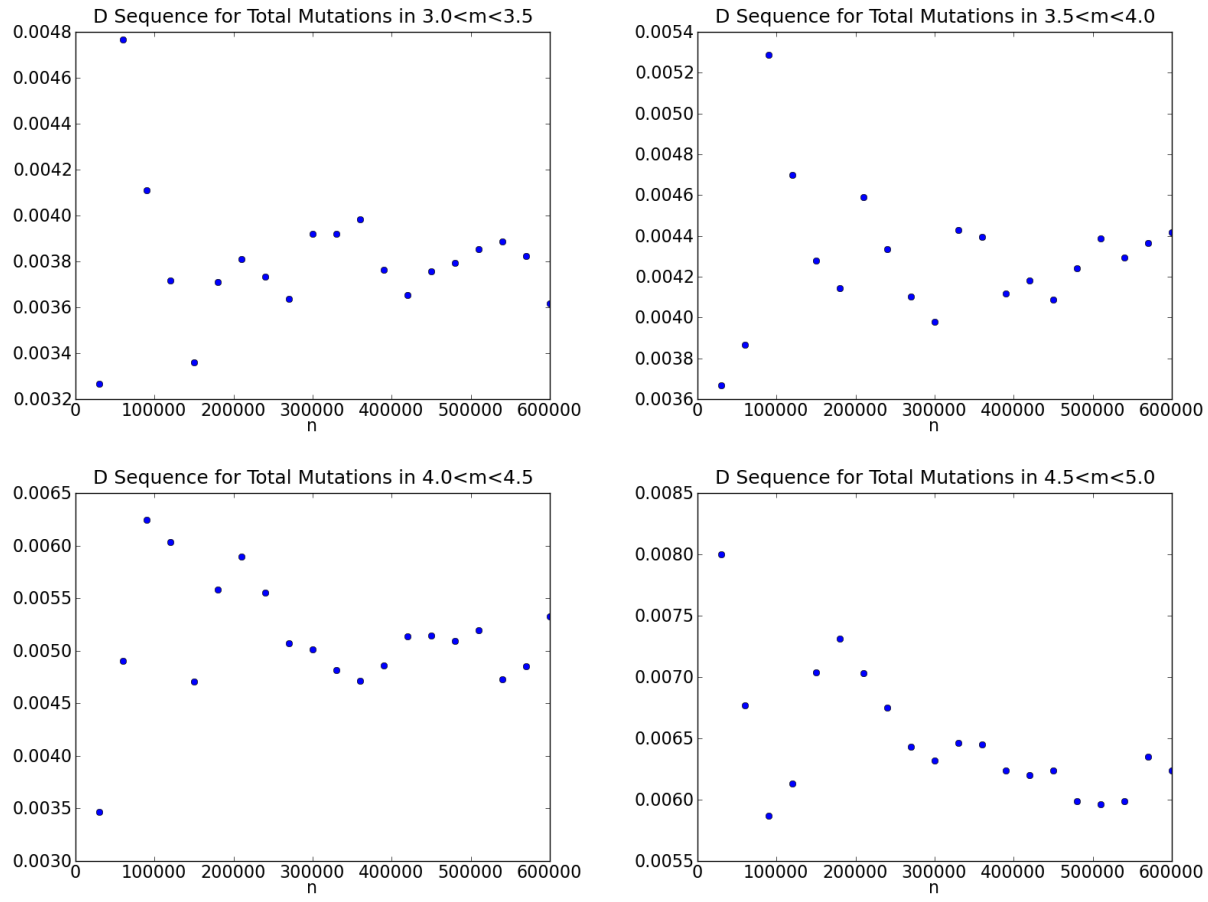

Figure A.16: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the total number of mutations in the following intervals, $3.0 \leq m \leq 3.5$ (top left) and $3.5 \leq m \leq 4.0$ (top right) $4.0 \leq m \leq 4.5$ (bottom left), $4.5 \leq m \leq 5.0$ (bottom right).

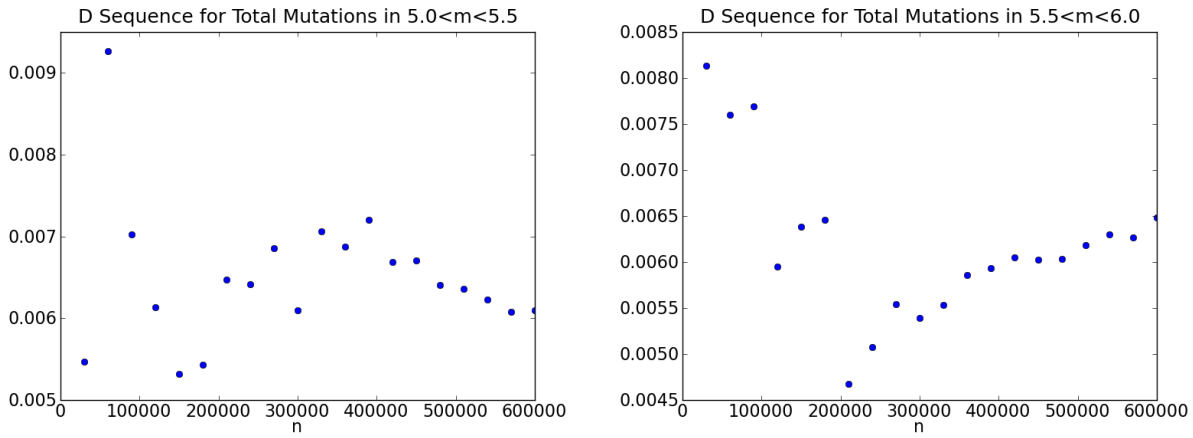

Figure A.17: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the total number of mutations in the following intervals, $5.0 \leq m \leq 5.5$ (left), $5.5 \leq m \leq 6.0$ (right).

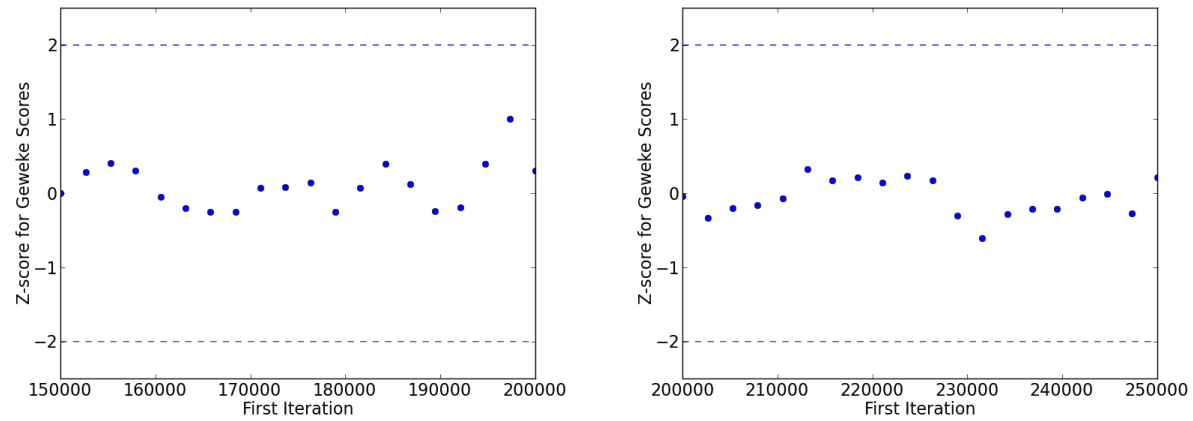

Figure A.18: Geweke scores for iterations 150,000 to 250,000 (left) and for iterations 200,000 to 300,000 (right). Each point is the Geweke score comparing the first 10% of the subchain to the last 50% of the subchain. The score is plotted against the first iteration used in the subchain.

A.2.3 Case 2

In Case 2 of the large gamma-profile mutation test cases, the shape parameters for the mutation profiles range from 1.0 to 5.5. This is the second shortest range of shape parameters tested (the shortest being Case 4). This case also has the highest mutation rate.

Figure A.19 shows the Geweke scores for 20 subchains. The plot on the left uses the first 50,000 iterations while the plot on the right uses the first 100,000 iterations. The scores for both plots are with ±2, indicating a failure to show lack of convergence.

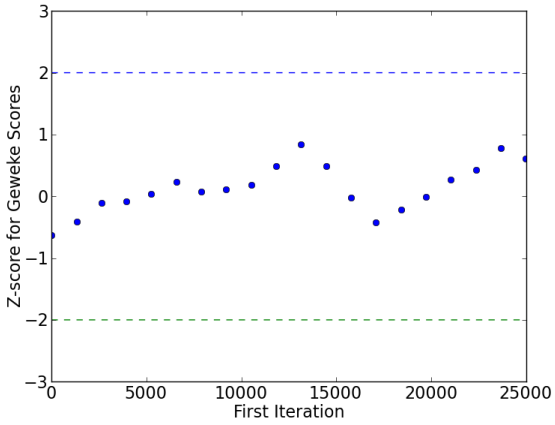

Figure A.19: Geweke scores for the first 50,000 iterations (left) and the first 100,000 iterations (right) of the MTM output for Case 2. Each point is the Geweke score comparing the first 10% of the subchain to the last 50% of the subchain. The score is plotted against the first iteration used in the subchain.

Figure A.20 shows the total number of mutations per genotype plotted against iteration number. The plot on the left shows the first 100,000 iterations and the plot on the right shows the first 500,000 iterations of the chain. The chain was started in the null genotype. As with the previous case, the total number of mutations per genotype rapid increases from 0 (the starting state) to the range 120 to 160 mutations. In this case, the chain first reaches 120 mutations on iteration 1233. From the plot on the right it is clear that the chain primarily explores genotypes with 120-160 mutations, making brief excursions to genotypes with significantly fewer (~100) or significantly more (~180) mutations. The horizontal line in the plot on the right is the empirical mean number of mutations per genotype calculated using all 500,000 iterations.

Figure A.21 shows the autocorrelation of the MTM output for Case 2 after discarding the first 10,000 iterations. The output for Case 2 is significantly less autocorrelated than the output for Case 3. For example, the correlation is less than 0.25 for lags larger than about

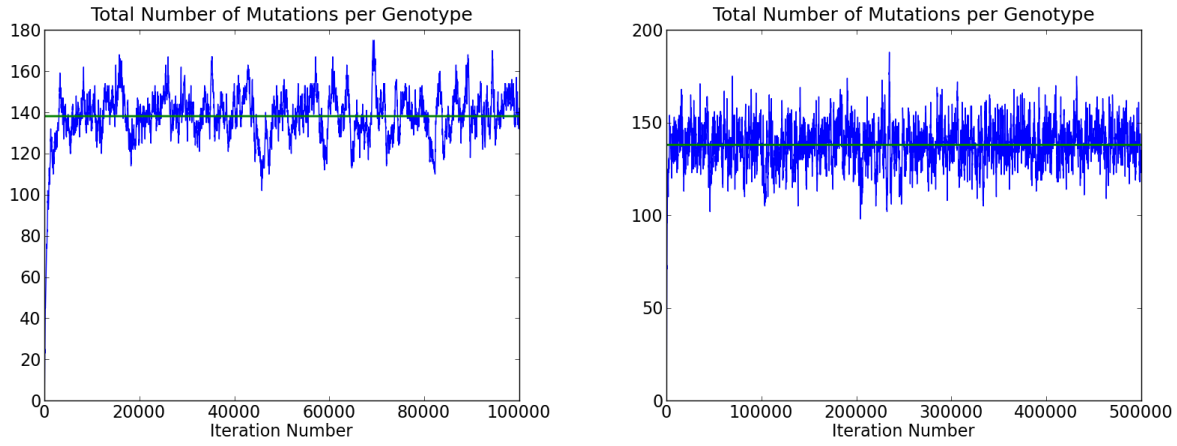

Figure A.20: Total number of mutations per genotype plotted against iteration number.

1,000 iterations for Case 2 whereas in Case 3 the autocorrelation is above 0.5 for lags of up to 5,000 iterations.

Figure A.22 shows the CUSUM plots for the first 50,000 iterations (left) and the first 100,000 iterations (right), discarding the first 10,000 iterations as burn-in. As with Case 3, the CUSUM plots show that the chain is making very long excursions away from the empirically determined mean. However, the excursions in this case appear to be of a shorter duration than in either Case 1 or Case 3.

Figure A.23 shows the sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. Again, the sequence of D statistics do not approach 0.5. The D statistics are fairly flat around $n = 300{,}000$ but then consistently increase starting around $n = 400,000$. Figures A.24-A.26 show the D statistics for the total number of mutations within a given range of shape parameters. In almost all of these plots, the D statistics appear to be within 15% of one another for large values of n. Again, these plots suggest that burn-in periods of 150,000 to 200,000 may be appropriate.

Table A.3 shows the results of applying the Raftery-Lewis diagnostic to the first 50,000 iterations (left hand side of the table) and to the first 100,000 iterations (right hand side of the table). In all cases, $r = 0.05$ and $s = 0.95$. The quantiles 0.1 to 0.9 in steps of size 0.1 were all tested. The Raftery-Lewis estimates for the burn-in and thinning parameter were larger when using 100,000 iterations. As a result, the estimate for the number of additional iterations needed was also longer when using 100,000 iterations rather than 50,000 iterations. The longest estimated burn-in period was just under 1700 iterations and the longest run for additional iterations was just under 150,000.

Finally, Figure A.27 shows the Geweke scores for iterations 150,000 to 250,000 (left) and for iterations 200,000 to 300,000 (right). Both plots fail to show a lack of convergence

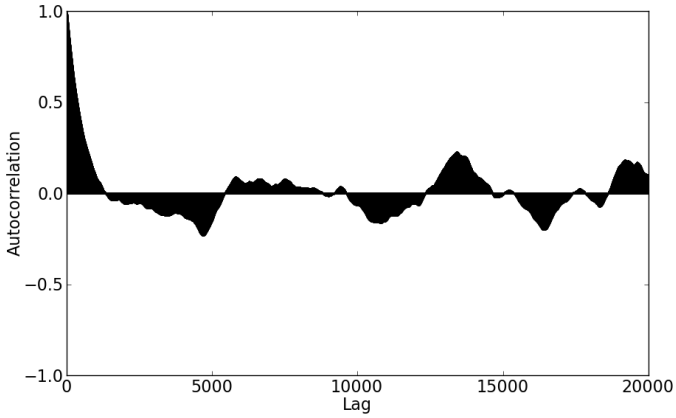

Figure A.21: Autocorrelation of the MTM output for Case 2 after discarding the first 10,000 iterations.

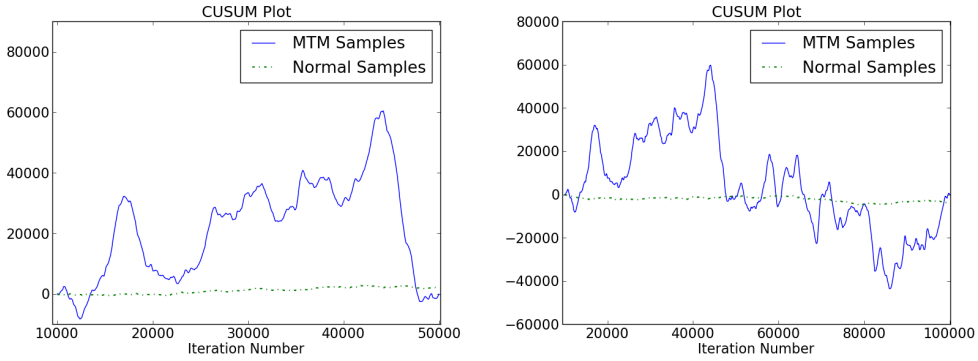

Figure A.22: CUSUM plot for the first 50,000 iterations (left) and first 100,000 iterations (right).

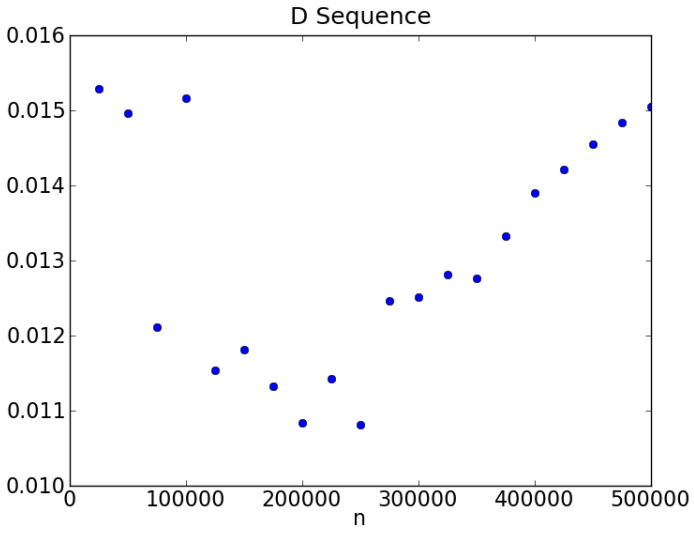

Figure A.23: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$.

Table A.3: Results from applying the Raftery-Lewis diagnostic to the first 50,000 iterations (left hand side) and 100,000 iterations (right hand side) of the MTM output for Case 2. Detailed descriptions of the parameters can be found in §A.1.4. In all cases $r = 0.05$ and $s = 0.95$; q is the quantile to be estimated, k is the thinning factor required for an independence chain, M in the length of the burn-in period and N is the number of additional iterations needed after the burn-in.

50k Iterations				100k Iterations			
q	k	M	N	q	k	M	N
0.1	479	1332	53983	0.1	546	1674	72261
0.2	407	1209	91962	0.2	557	1596	115836
0.3	497	1134	116298	0.3	587	1513	149609
0.4	380	1105	122395	0.4	392	1275	148125
0.5	344	936	111345	0.5	459	1197	145908
0.6	302	918	100764	0.6	457	1280	147360
0.7	325	960	90560	0.7	409	1330	129010
0.8	325	1080	65438	0.8	378	1078	70609
0.9	274	780	29991	0.9	323	960	31860

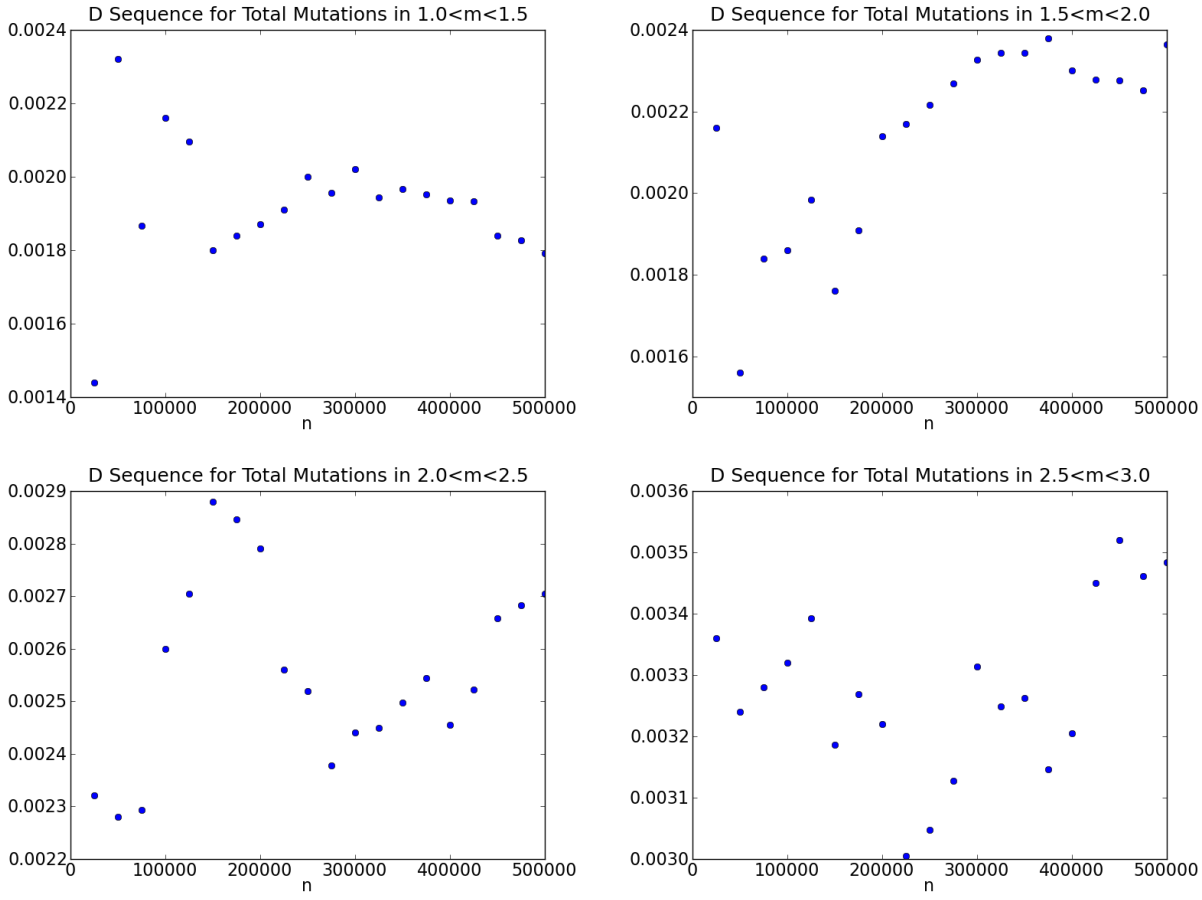

Figure A.24: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the total number of mutations in the following intervals, $1.0 \leq m \leq 1.5$ (top left), $1.5 \leq m \leq 2.0$ (top right), $2.0 \leq m \leq 2.5$ (bottom left), $2.5 \leq m \leq 3.0$ (bottom right).

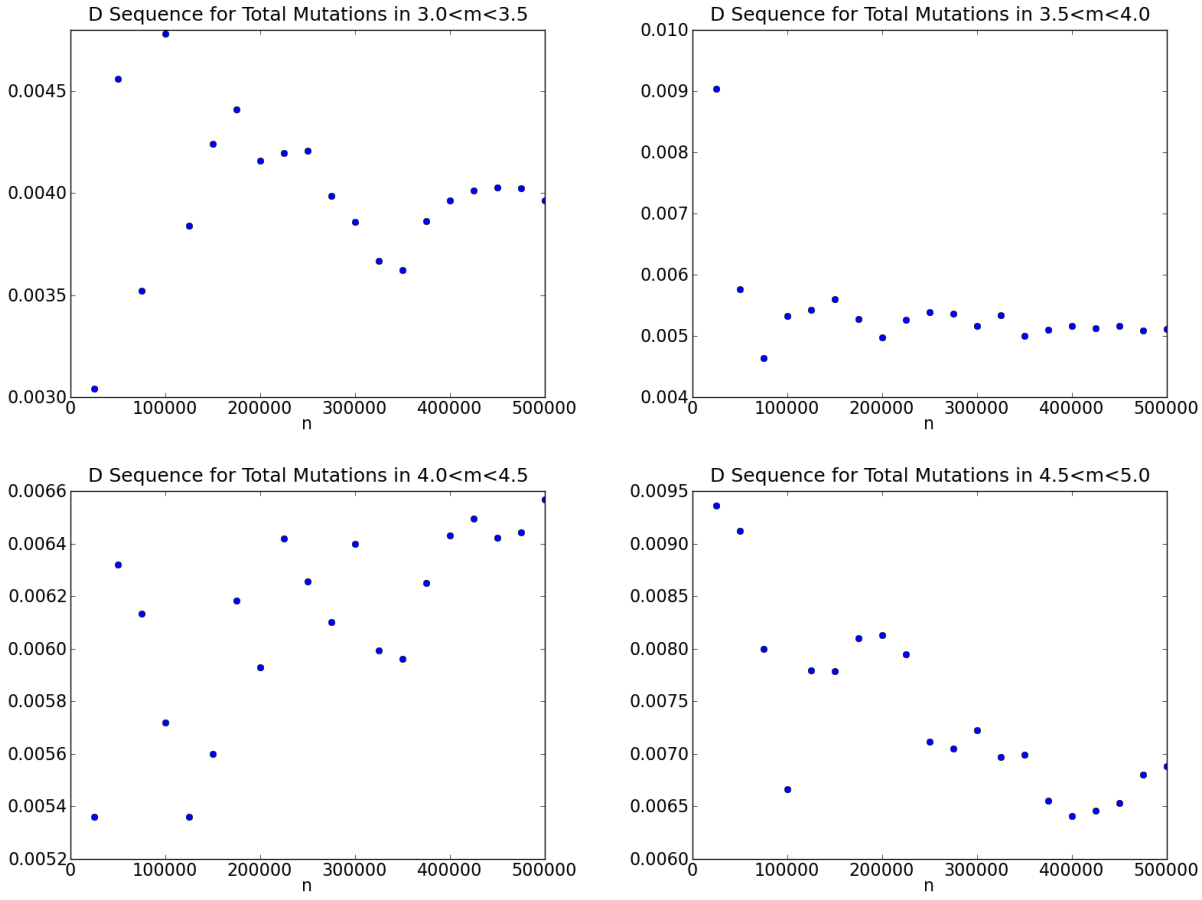

Figure A.25: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the total number of mutations in the following intervals, $3.0 \leq m \leq 3.5$ (top left), $3.5 \leq m \leq 4.0$ (top right), $4.0 \leq m \leq 4.5$ (bottom left) and $4.5 \leq m \leq 5.0$ (bottom right).

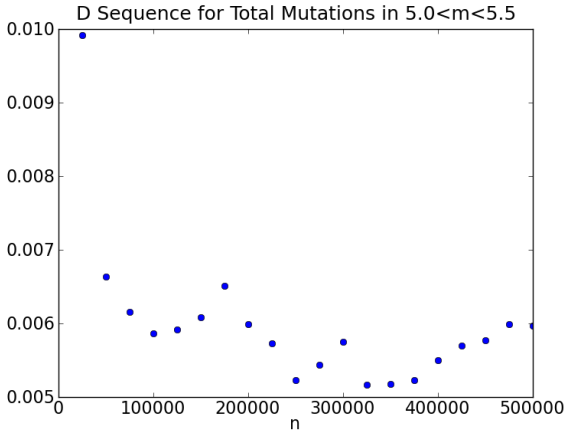

Figure A.26: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the total number of mutations in the interval $5.0 \leq m \leq 5.5$ (bottom row).

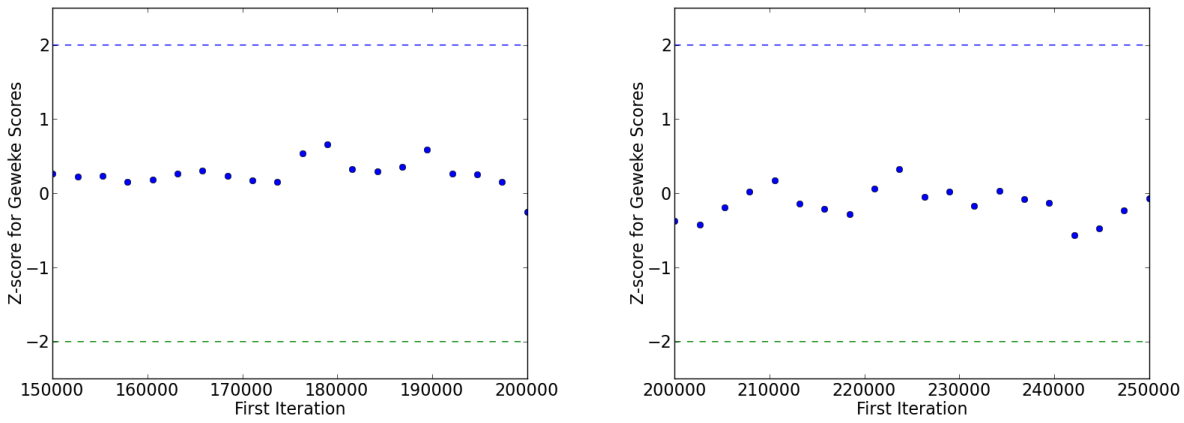

Figure A.27: Geweke scores for iterations 150,000 to 250,000 (left) and for iterations 200,000 to 300,000 (right). Each point is the Geweke score comparing the first 10% of the subchain to the last 50% of the subchain. The score is plotted against the first iteration used in the subchain.

over these iterations. From the above analysis we conclude that a burn-in period of 150,000 iterations should be sufficient and we should collect at least 250,000 samples.

A.2.4 Case 4

Finally we consider Case 4, in which the mutation space consists of 1000 gamma profile mutations with shape parameters ranging from 1.0 to 5.0. This case has the shortest range of shape parameters and also one of the lowest mutation rates.

Figure A.28 shows the Geweke scores for 20 subchains. The plot on the left uses the first 50,000 iterations while the plot on the right uses the first 100,000 iterations. In both plots, all scores lie well within the range ± 2, suggesting that the chain fails to show a lack of convergence.

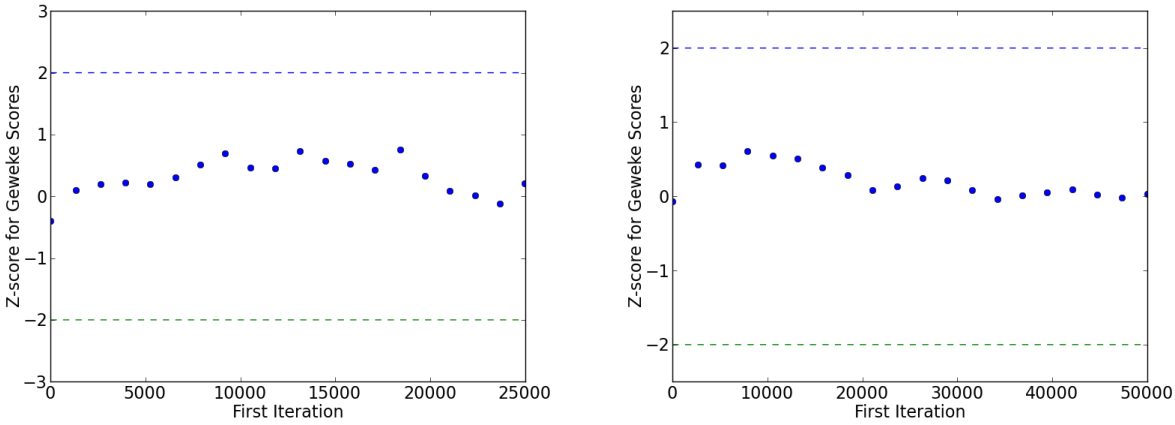

Figure A.28: Geweke scores for the first 50,000 iterations (left) and the first 100,000 iterations (right) of the MTM output for Case 4. Each point is the Geweke score comparing the first 10% of the subchain to the last 50% of the subchain. The score is plotted against the first iteration used in the subchain.

Figure A.29 shows the total number of mutations per genotype plotted against iteration number. The plot on the left shows the first 100,000 iterations and the plot on the right shows the first 500,000 iterations of the chain. The chain was started in the null genotype. As with the other cases, the chain rapid moves from genotypes with no mutations to genotypes with 50-80 mutations. The chain first hits a genotype with 60 mutations at iteration 1864. The plot on the right shows that the chain primarily stays in the region of genotypes with roughly 45-80 mutations, although it periodically explores genotypes with more than 80 or fewer than 40 mutations.

Figure A.30 shows the autocorrelation in the output of the MTM algorithm for Case 4 after discarding the first 10,000 iterations as the burn-in period. The output for this chain shows the least autocorrelation of any of the cases considered so far.

Figure A.31 shows the CUSUM plots for the first 50,000 iterations (left) and the first 100,000 iterations (right), discarding the first 10,000 iterations as burn-in. As with the other

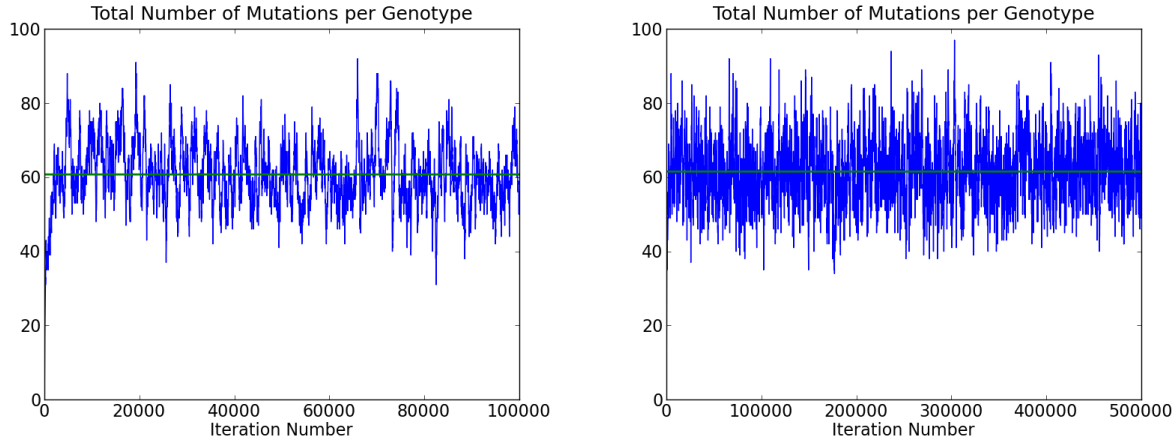

Figure A.29: Total number of mutations per genotype plotted against iteration number.

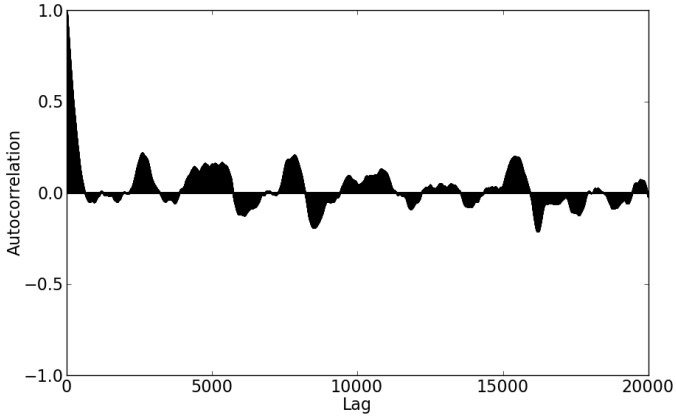

Figure A.30: Autocorrelation of the MTM output for Case 4 after discarding the first 10,000 iterations.

cases, the CUSUM plots show that the chain is making very long excursions away from the empirically determined mean. The excursions in this case appear to be on par with those of Case 2 and are generally shorter than those of Cases 1 and 3.

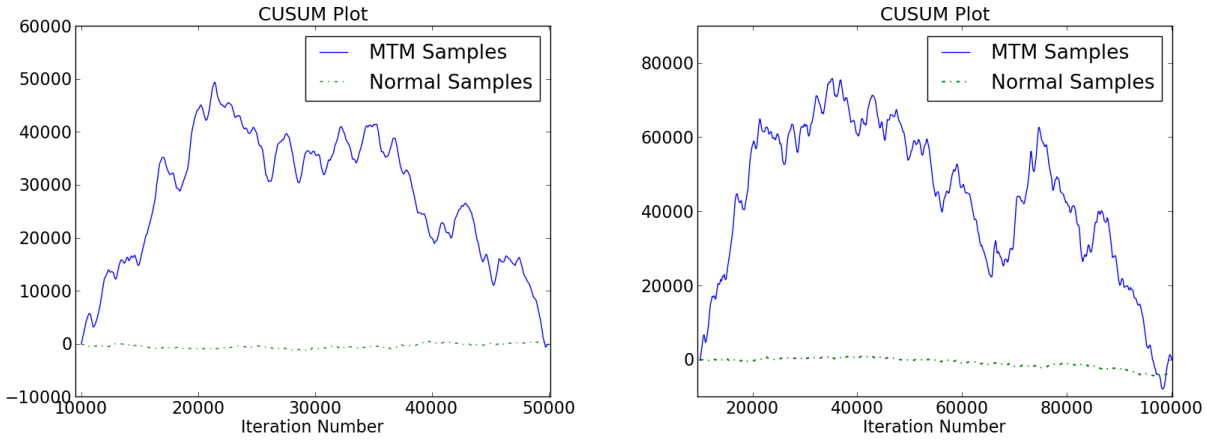

Figure A.31: CUSUM plot for the first 50,000 iterations (left) and first 100,000 iterations (right).

Figure A.32 shows the sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D statistics are fairly flat starting around $n = 200{,}000$, with an absolute difference between the largest D and the smallest D in that range of 0.00065. This corresponds to a relative difference of about 4% (relative to the smallest D statistic for $n \geq 200,000$). Figures A.33 and A.34 show the D statistics for the total number of mutations within a given range of shape parameters. In almost all of these plots, the D statistics appear to be within 10% of one another for large values of n. These plots suggest that a burn-in period of 150,000 iterations may be appropriate.

Table A.4 shows the results of applying the Raftery-Lewis diagnostic to the first 50,000 iterations (left hand side of the table) and to the first 100,000 iterations (right hand side of the table). In all cases, $r = 0.05$ and $s = 0.95$. The quantiles 0.1 to 0.9 in steps of size 0.1 were all tested. The Raftery-Lewis estimates for the burn-in and thinning parameter were fairly consistent for most quantiles when using either 50,000 or 100,000 iterations. The longest estimated burn-in period was just over 1100 iterations and the longest run for additional iterations was just under 125,000.

Finally, Figure A.35 shows the Geweke scores for iterations 150,000 to 250,000 (left) and for iterations 200,000 to 300,000 (right). Both plots fail to show a lack of convergence over these iterations. From the above analysis we conclude that a burn-in period of 150,000 iterations should be sufficient and we should collect at least 200,000 samples.

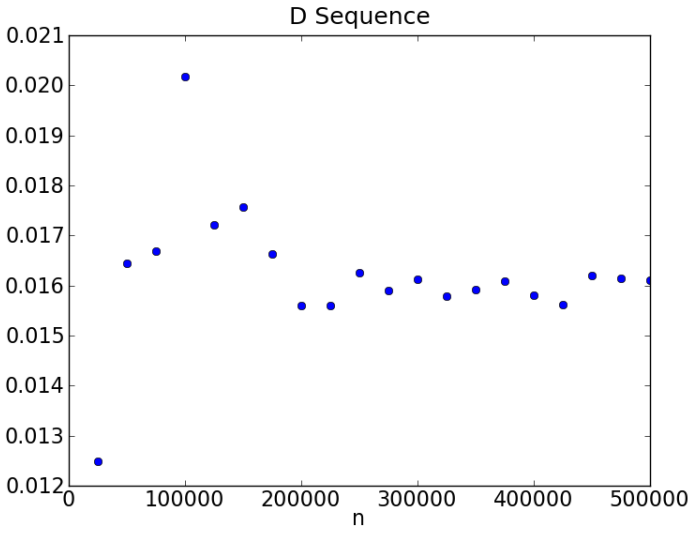

Figure A.32: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$.

Table A.4: Results from applying the Raftery-Lewis diagnostic to the first 50,000 iterations (left hand side) and 100,000 iterations (right hand side) of the MTM output for Case 4. Detailed descriptions of the parameters can be found in §A.1.4. In all cases $r = 0.05$ and $s = 0.95$; q is the quantile to be estimated, k is the thinning factor required for an independence chain, M in the length of the burn-in period and N is the number of additional iterations needed after the burn-in.

50k Iterations				100k Iterations			
q	k	M	N	q	k	M	N
0.1	265	589	24304	0.1	272	602	29541
0.2	241	1000	80250	0.2	311	845	64155
0.3	258	816	80631	0.3	309	966	93564
0.4	285	880	101200	0.4	311	897	103638
0.5	328	1008	123792	0.5	329	966	114471
0.6	353	945	102935	0.6	378	1120	131520
0.7	351	1160	108286	0.7	400	1122	105666
0.8	300	864	58800	0.8	328	910	63140
0.9	181	621	21816	0.9	269	935	32890

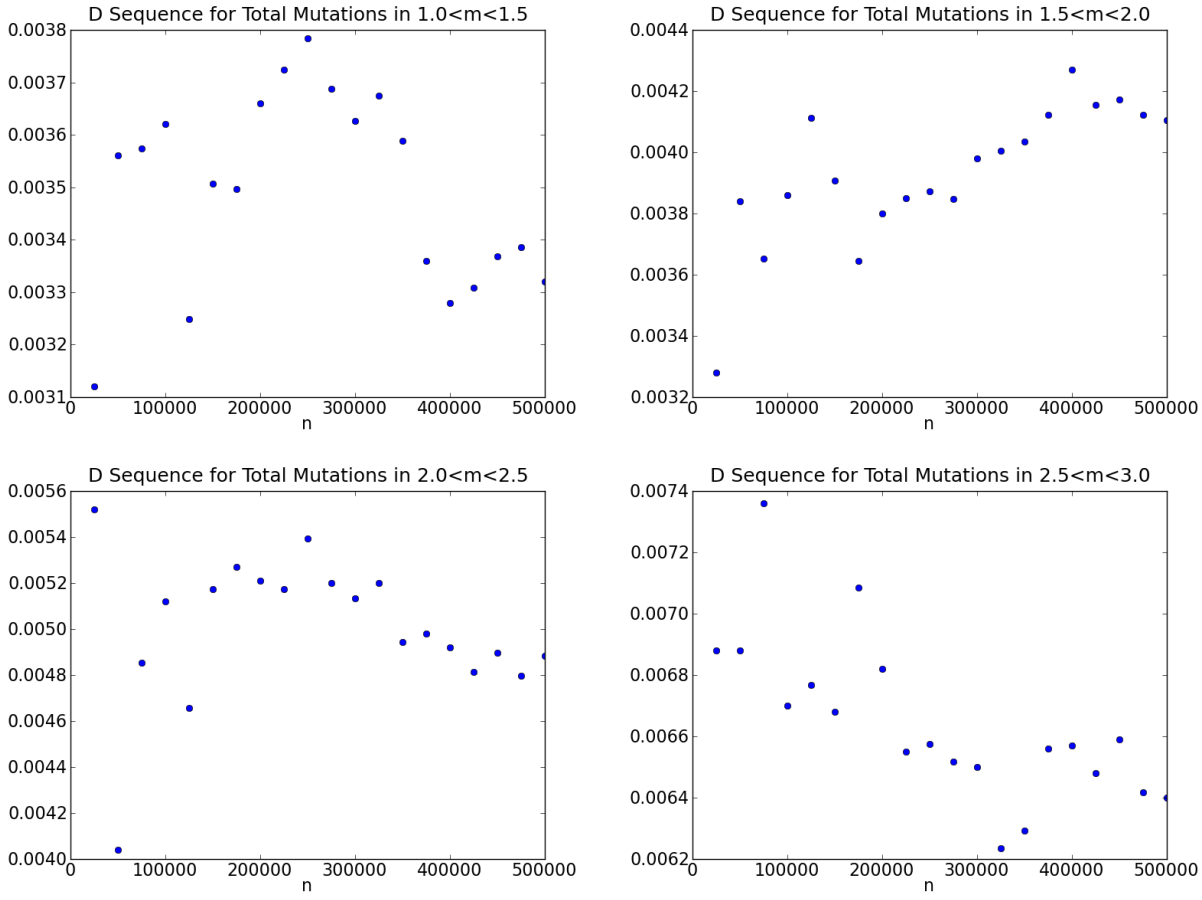

Figure A.33: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the total number of mutations in the following intervals, $1.0 \leq m \leq 1.5$ (top left), $1.5 \leq m \leq 2.0$ (top right), $2.0 \leq m \leq 2.5$ (bottom left), $2.5 \leq m \leq 3.0$ (bottom right).

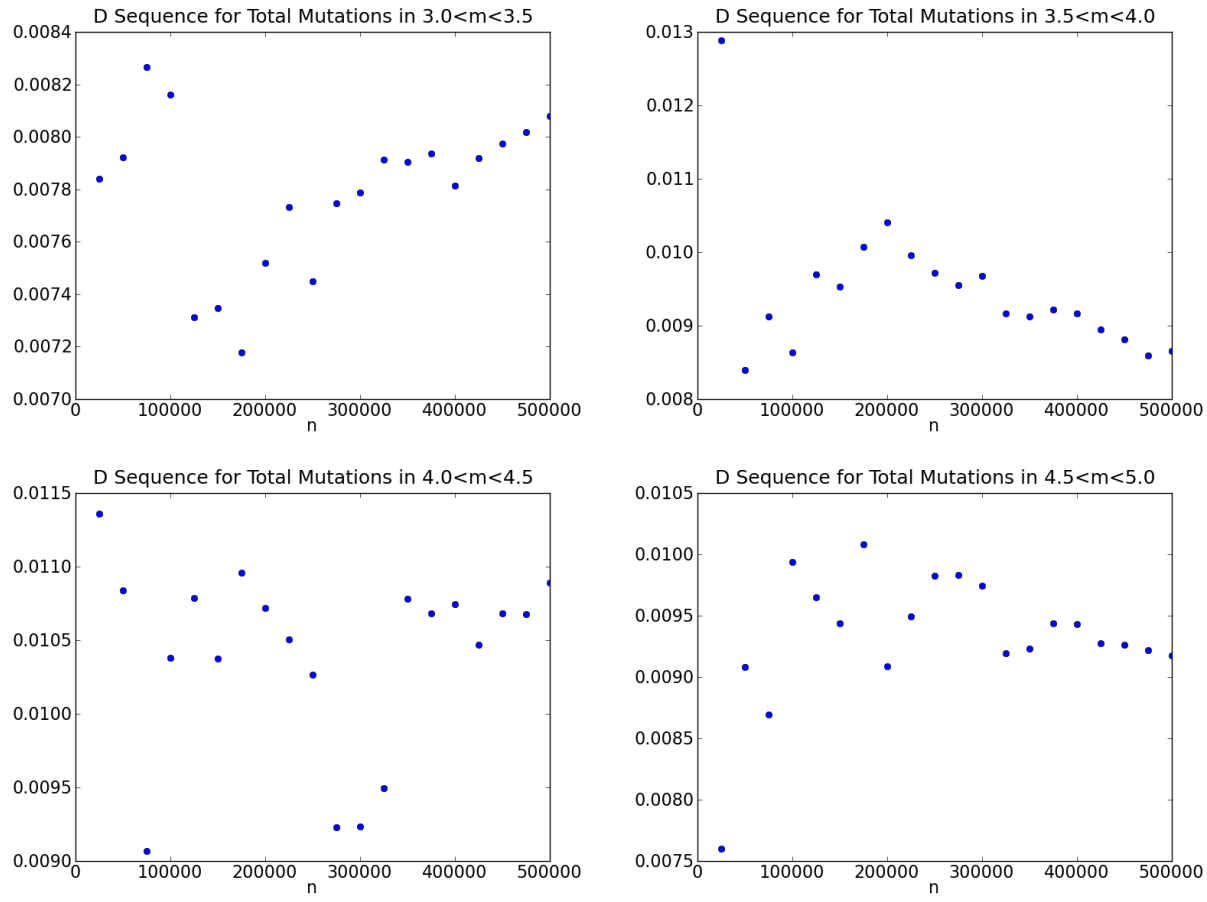

Figure A.34: Sequence of $D_{n/2,n}$ plotted against n for $n \in \{25000, 50000, \cdots 500000\}$. The D are calculated from the total number of mutations in the following intervals, $3.0 \leq m \leq 3.5$ (top left) and $3.5 \leq m \leq 4.0$ (top right). $4.0 \leq m \leq 4.5$ (bottom left), $4.5 \leq m \leq 5.0$ (bottom right).

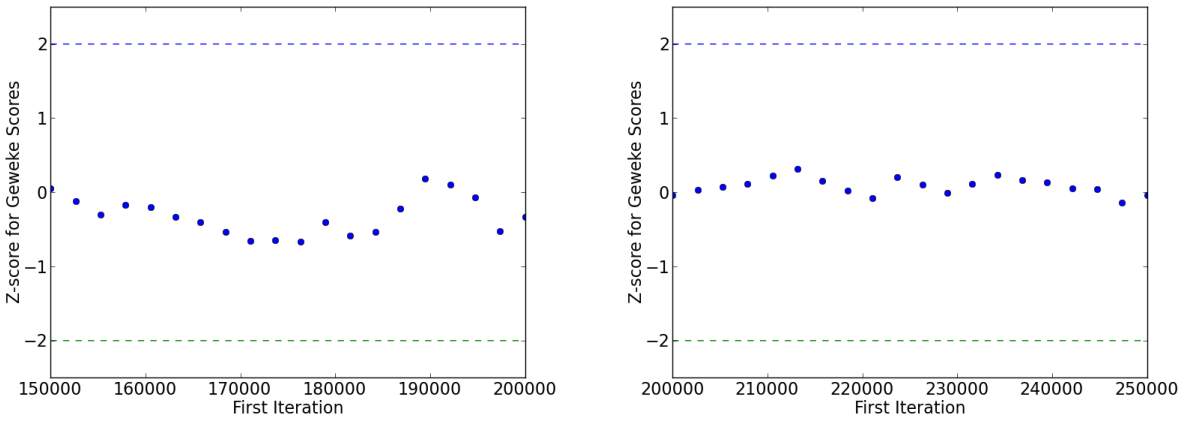

Figure A.35: Geweke scores for iterations 150,000 to 250,000 (left) and for iterations 200,000 to 300,000 (right). Each point is the Geweke score comparing the first 10% of the subchain to the last 50% of the subchain. The score is plotted against the first iteration used in the subchain.

A.3 Additional Iterations

Of course, all diagnostic tests for MCMC should be taken with a grain of salt. The burn-in periods and sample sizes discussed in the previous sections are, while guided by the diagnostic tools, somewhat arbitrary. There always remains the question of whether or not additional iterations may substantially change the outcome. In this work in particular, where the chain crawls over the space of ordered genotypes, there are so many highly unlikely orderings of a genotype that we do not expect the chain to fully explore the space the genotypes. However, we do want to have some faith that the chain has discovered the region of the space of ordered genotypes which is most likely and that it has not gotten stuck in local maximums.

One method that we can employ is to let the chain run for much longer. We will briefly discuss the effects of additional iterations on two cases, Case 1 and Case 3. We begin with Case 3. Figure A.36 shows several histograms for Case 3. In all four histograms the first 250,000 iterations were discarded as the burn-in. The plot on the top left shows the histogram computed using 750,000 samples; the top right uses 1.25 million samples; the bottom left uses 1.75 million samples and the bottom right, reproduced from §4.1 uses 2.25 million samples. The histograms generated with additional samples are certainly smoother than the histogram with only 750,000 samples. Despite this, the means and variances estimated from the four cases are reasonably similar, with the main difference being a general reduction in variance. Means and variances corresponding to the four histograms are listed in Table A.5.

Case 3, of course, is one of the more difficult cases we consider and appears to require

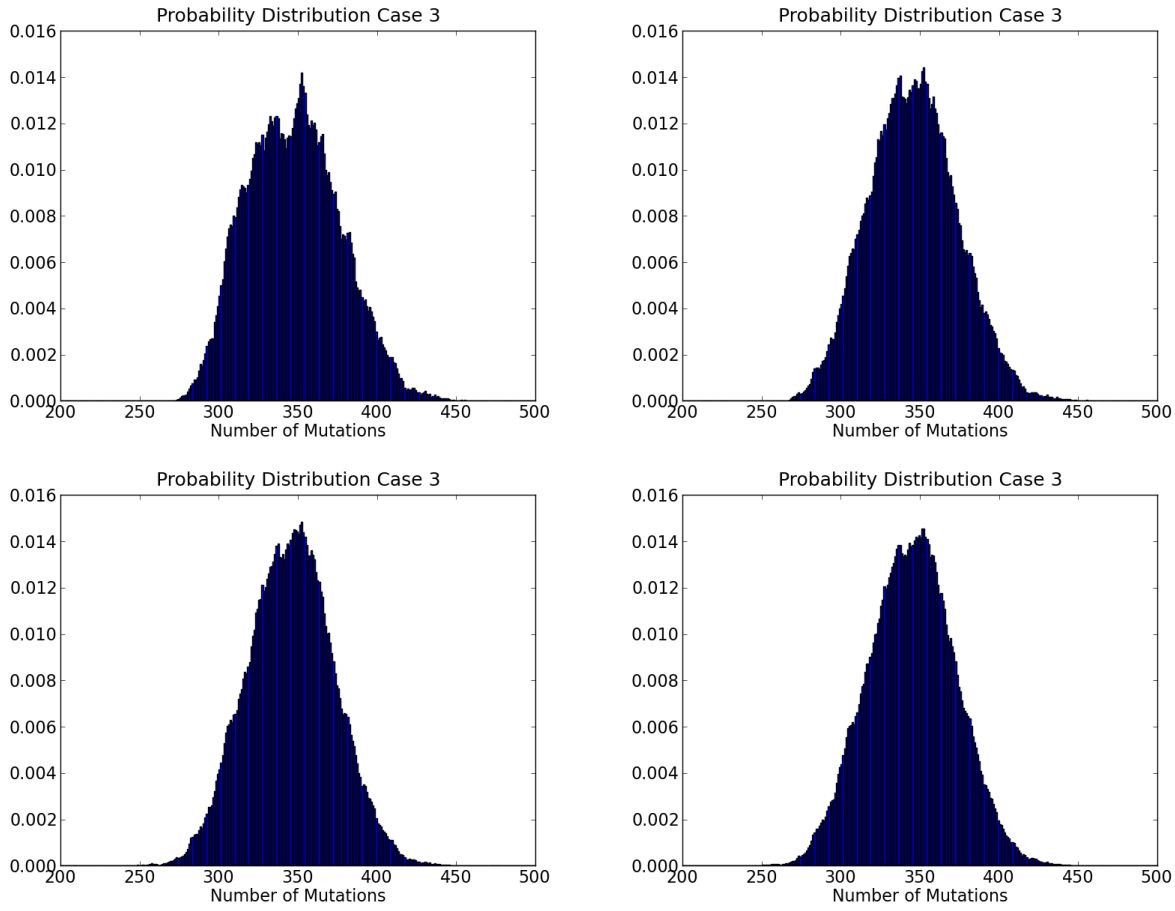

Figure A.36: Histograms for Case 3. In all four plots the burn-in period is 250,000 iterations. The plot on the top left uses 750,000 samples, the plot on the top right is generated from 1.25 million samples, the plot on the bottom left uses 1.75 million samples and the plot on the bottom uses 2.25 million samples.

Table A.5: Estimated mean and variance for Case 3 using 750,000 samples, 1.25 million samples, 1.75 million samples and 2.25 million samples.

	750,000 Samples		1.25 million Samples		1.75 million Samples		2.25 million Samples	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
$1.0 \leq m \leq \xi$	346.756	850.589	345.047	777.201	345.014	744.989	344.318	757.616
$1.0 \leq m \leq 1.5$	0.949715	0.956017	0.963459	0.969611	0.994706	0.991544	1.00961	1.02729
$1.5 < m \leq 2.0$	1.67586	1.70109	1.62731	1.66610	1.63867	1.64831	1.59522	1.65655
$2.0 < m \leq 2.5$	2.26294	2.14632	2.30665	2.36135	2.31133	2.41566	2.30087	2.38823
$2.5 < m \leq 3.0$	3.79382	4.05943	3.69092	3.87022	3.65380	3.68177	3.68019	3.77548
$3.0 < m \leq 3.5$	6.15368	6.76874	6.19487	6.59036	6.14134	6.50148	6.09614	6.37484
$3.5 < m \leq 4.0$	9.76680	9.77157	9.52729	9.07214	9.33726	9.03490	9.36255	9.37602
$4.0 < m \leq 4.5$	15.1695	14.3397	15.0472	14.9859	14.9671	14.8910	14.9570	15.0115
$4.5 < m \leq 5.0$	23.3334	23.5936	23.2624	24.6859	23.3101	24.4356	23.3678	25.1011
$5.0 < m \leq 5.5$	37.3839	47.9219	37.1020	44.2892	37.2135	43.7551	37.0537	45.2533
$5.5 < m \leq 6.0$	57.3715	72.6855	57.4060	73.7718	57.3623	70.6010	57.2252	72.8376
$6.0 < m \leq 6.5$	81.8047	114.368	81.5014	104.913	81.5131	106.737	81.2535	104.864
$6.5 < m \leq 7.0$	107.090	165.924	106.417	153.018	106.570	149.296	106.416	149.596

many more iterations than Cases 1, 2 or 4. As a result, we will now consider the effect of additional iterations on Case 1. The results shown here are also more typical of Cases 2 and 4. Figure A.37 shows several histograms for Case 1. All three histograms discard the first 150,000 iterations as the burn-in. The plot on the top left in generated using 150,000 samples, the plot on the top right uses 350,000 samples and the plot on the bottom uses 750,000 samples. Again, while the histogram generated with the most samples is certainly smoother than the other two histograms, the three histograms are not terribly different. The means for the number of mutations per genotype are also reasonably close for all three sample sizes, as seen in Table A.6. For example, using 750,000 samples, the mean number of mutations per genotype is 166.35 and the variance in the number is 193.22. With only 350,000 samples, the mean number of mutations per genotype is 166.6 but the variance is larger, 214.9. In general, running the chain for longer reduces variance but does not radically change the means.

Table A.6: Estimated mean and variance for Case 1 using 350,000 samples, 750,000 samples, and 1 million samples.

	350,000 Samples		750,000 Samples		1,000,000 Samples	
	Mean	Variance	Mean	Variance	Mean	Variance
$1.0 \leq m \leq \xi$	166.598	214.889	166.354	193.224	166.334	190.679
$1.0 \leq m \leq 1.5$	0.959089	1.09235	0.934101	1.01544	0.942218	1.02505
$1.5 < m \leq 2.0$	1.50962	1.54981	1.48184	1.52674	1.51251	1.59733
$2.0 < m \leq 2.5$	2.44802	2.43936	2.43595	2.57175	2.40218	2.53109
$2.5 < m \leq 3.0$	3.91711	4.67557	3.93335	4.16092	3.92265	3.98447
$3.0 < m \leq 3.5$	6.42722	7.20124	6.40836	6.88547	6.32708	6.94444
$3.5 < m \leq 4.0$	10.0595	10.8118	10.2112	11.2320	10.2278	11.0687
$4.0 < m \leq 4.5$	16.9484	18.8845	16.9794	18.5256	16.9770	17.9433
$4.5 < m \leq 5.0$	27.0122	26.0140	26.9847	26.5366	27.0789	27.1953
$5.0 < m \leq 5.5$	41.3653	46.2364	41.3794	45.4949	41.3440	45.5066
$5.5 < m \leq 6.0$	55.9517	67.8250	55.6058	64.0315	55.5995	64.2969

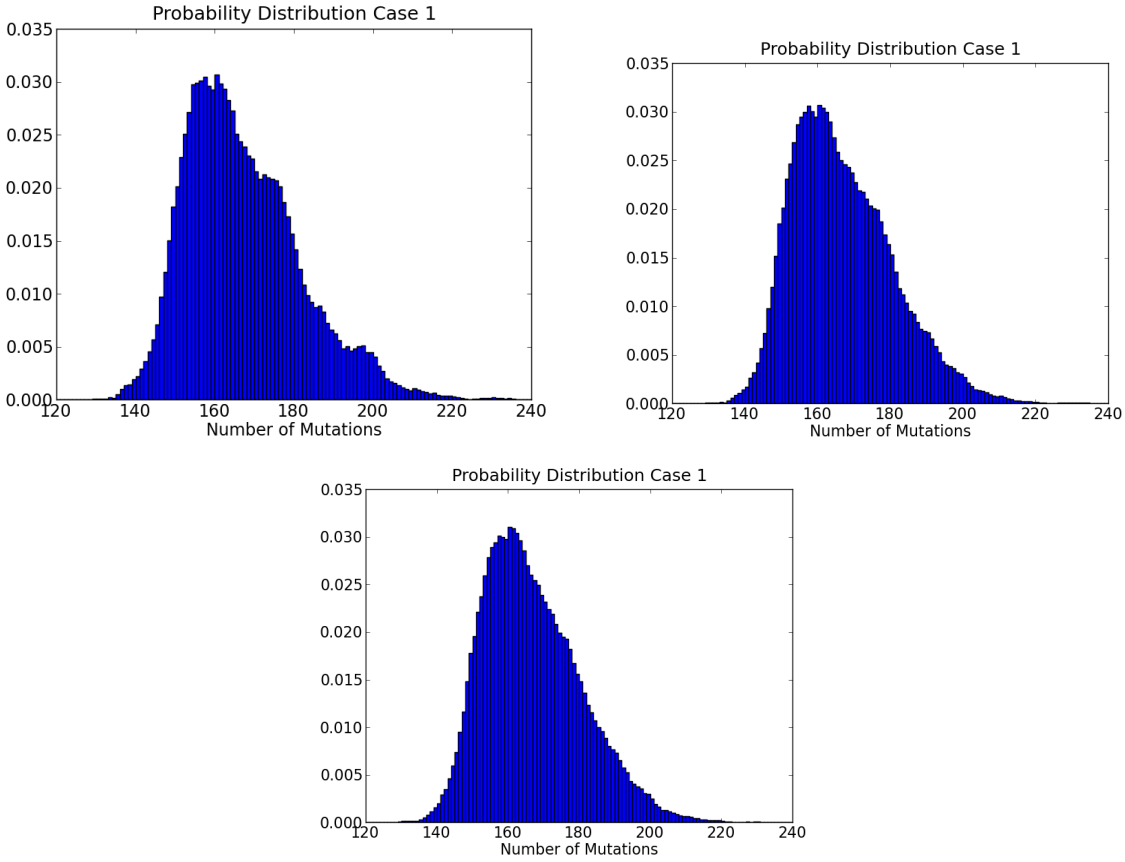

Figure A.37: Histograms for Case 1. In all three plots the burn-in period is 150,000 iterations. The plot on the top left uses 350,000 samples, the plot on the top right is generated from 750,000 samples, the plot on the bottom left uses 1 million samples.

Appendix B

Deletion Probability for the Multiple-Try Metropolis Algorithm

To determine a reasonable value for DelP, the probability that a chosen mutation is deleted (the chosen mutation changes type with probability 1 - DelP), we ran the following eight trials. For each of these trials, labeled 1-8, the mutation space contained 1000 gamma mutation profiles with the same rate parameter of 0.05 and shape parameters ranging from 1.0 to 5.0. In each trial the chain was run with a burn-in period of 50,000 steps and an additional 100,000 steps were collected as samples. The deletion probabilities tested were 0.125 (Trial 1), 0.25 (Trial 2), 0.375 (Trial 3), 0.5 (Trial 4), 0.625 (Trial 5), 0.75 (Trial 6), 0.875 (Trial 7) and 1.0 (Trial 8). All other parameters for the eight test cases were the same and can be found in Table B.1. The fertility rate, determined by the short-cut algorithm for the ESW free recombination model, was 0.069345238.

The histograms for the total number of mutations for Trials 1-8 are shown in Figures B.1 and B.2. All eight trials produced histograms with the total number of mutations ranging from around 40 mutations per genotype to around 90. The average number of mutations per genotype for the eight trials were also similar: 61.7 mutations for Trial 1, 63.8 for Trial 2, 60.7 for Trial 3, 60.9 for Trial 4, 61.7 for Trial 5, 60.9 for Trial 6, 62.9 for Trial 7 and 61.1 for Trial 8. However, while the histograms for the eight trials are generally similar, the acceptance rates for proposed steps were not. Shown in Table B.2, the acceptance rates were highest in the first two trials (about 56% for Trial 1 and 48% for Trial 2) and lowest for the last trial (6% for Trial 8).

Because a high acceptance rate means we move around the genotype space more frequently, we ideally would like to choose a value for DelP that produces a high acceptance rate. For the eight trials considered here, smaller values for DelP produced higher acceptance rates than values near 1.0. However, small values for DelP also mean that proposed genotypes will infrequently have fewer mutations than the current genotype. Because we start the Markov chain in the null genotype, having a small value for DelP is initially useful. In that case, proposed genotypes will generally contain at least as many mutations as the

current genotype, allowing the chain to move more quickly to genotypes with many mutations. However, a small value for DelP may be less useful once the chain nears genotypes that are highly likely. As a result we have chosen to use either DelP = 0.375 or DelP = 0.5 for all of the large mutation space tests.

Table B.1: Parameters for the eight test cases with 1000 gamma profile mutations with shape parameters from 1.0 to 5.0 (inclusive).

All Tests	η	λ	α	β	Dx	$\nu(\mathcal{M})$	Gamma rate	ξ
	0.1	0.05	15	50	0.5	0.12	0.05	5
Trial	Kmax	DelP	Burn	Samples				
1	5	0.125	50000	100000				
2	5	0.25	50000	100000				
3	5	0.375	50000	100000				
4	5	0.5	50000	100000				
5	5	0.625	50000	100000				
6	5	0.75	50000	100000				
7	5	0.875	50000	100000				
8	5	1.0	50000	100000				

Table B.2: Output from the MTM algorithm for the eight test cases with 1000 gamma profile mutations with shape parameters from 1.0 to 5.0 (inclusive).

Trial	DelP	Acceptance Ratio
1	0.125	0.56530
2	0.25	0.48007
3	0.375	0.38214
4	0.5	0.29093
5	0.625	0.21528
6	0.75	0.15113
7	0.875	0.10173
8	1.0	0.06407

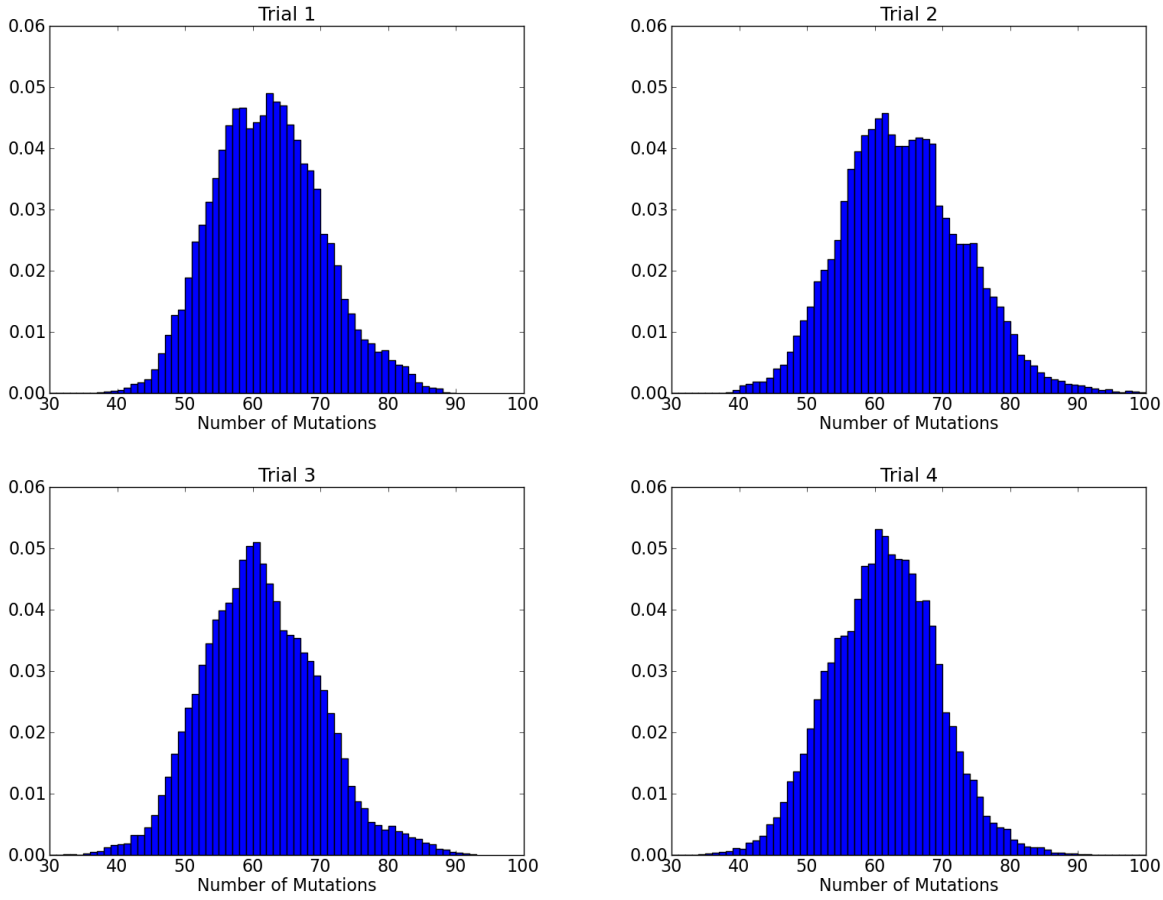

Figure B.1: Histograms for the total number of mutations per genotype for the test cases with 1000 gamma profile mutations with shape parameters from 1.0 to 5.0. The top rows shows the histograms for Trial 1 (DelP = 0.125), left, and Trial 2 (DelP = 0.25), right. The bottom row displays the histograms for Trial 3 (DelP = 0.375), left, and Trial 4 (DelP = 0.5), right.

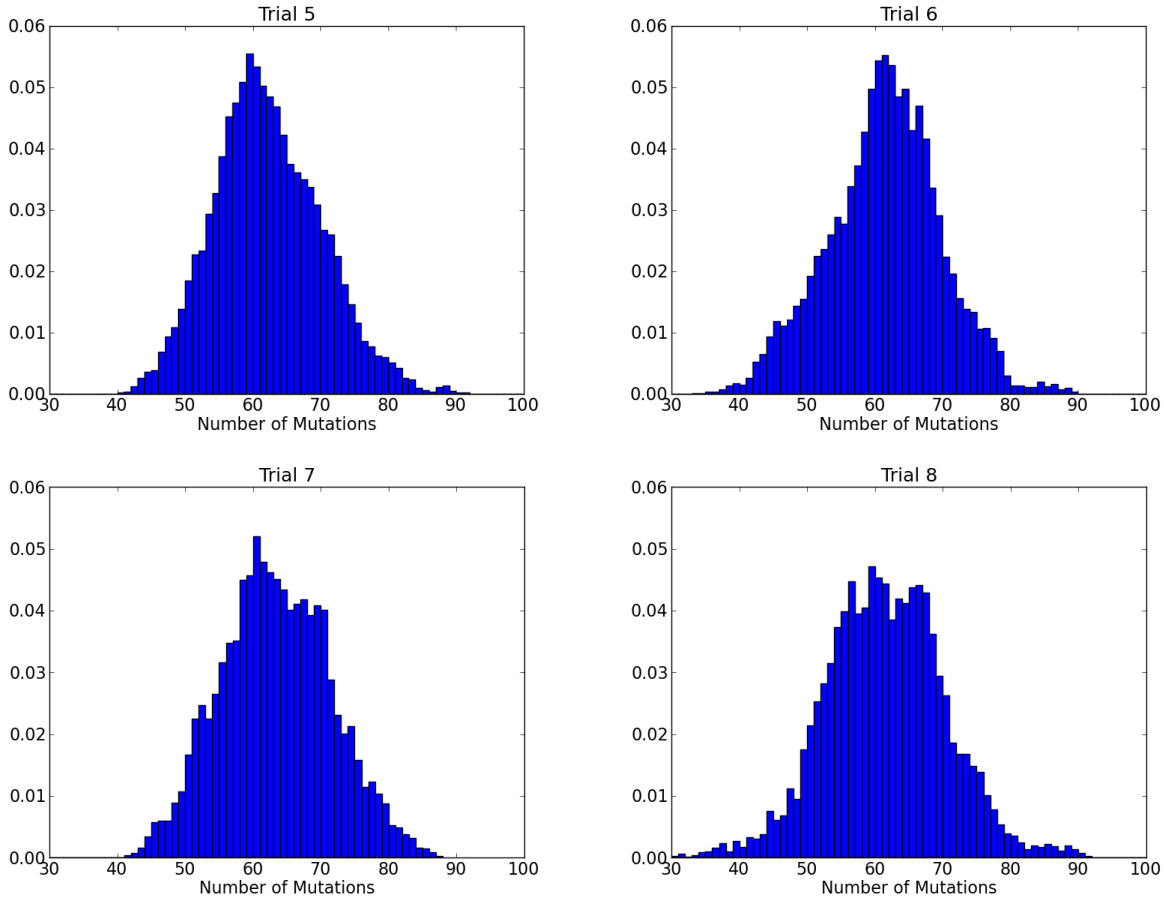

Figure B.2: Histograms for the total number of mutations per genotype for the test cases with 1000 gamma profile mutations with shape parameters from 1.0 to 5.0. The top rows shows the histograms for Trial 1 (DelP = 0.625), left, and Trial 2 (DelP = 0.75), right. The bottom row displays the histograms for Trial 3 (DelP = 0.875), left, and Trial 4 (DelP = 1.0), right.

Appendix C

The Nature of ρ in the ESW Free Recombination Model

C.1 Exponential Behavior

Recall that the solution to the ESW free recombination model is a Poisson random measure with intensity measure ρ. This means that for mutation type $m \in \mathcal{M}$, the number of copies of m in a genotype will be Poisson distributed with mean $\rho(m)$. As mentioned in §4.1.2, the intensity measure ρ appears to be approximately exponential for cases with large mutation spaces where mutations have gamma profiles. In particular, we are referring to Cases 1-4, in which the mutation space contains 1000 gamma profile mutations with the same gamma rate parameter of 0.05 but different gamma shape parameters. The shape parameters range from 1.0 to ξ, inclusive, where ξ is between 5.0 and 7.0.

To verify that the intensity measures for these four cases are approximately exponential, each intensity measure was fitted with an exponential curve. This was accomplished using the python package scikits.statsmodels [29] to fit the linear model

$$\log(\rho) = \alpha \text{Shape Parameter} + \beta$$

using OLS. The coefficients α and β from the fitted curves are shown in Table C.1. Figure C.1 shows the intensity measure ρ from the free recombination model and the exponential approximation for the four cases. In all four cases the intensity measure is larger than the approximation for both small values (near 1) and large values (near ξ) of the shape parameter but is smaller than the approximation for middle values of the shape parameter. Here, the terms "large" and "middle" are relative to ξ, the maximum shape parameter for the case.

The plots of absolute and relative difference between ρ and its exponential approximation for all four cases can be found in Figures C.2 and C.3, respectively. In all four cases the largest absolute difference occurs near the top of the range of possible shape parameters for the case. As a particular example, consider Case 3, where the largest possible shape parameter is 7.0.

In Case 3, the absolute difference between ρ and its exponential approximation is small (at most 0.07) for shape parameters smaller than 5.5. For shape parameters ranging from 6.0 to 7.0, the absolute difference grows from 0.17 to 4.4. The relative difference, however, is large (between 0.2 and 0.3) for both small shape parameters (1.0-1.3) and large shape parameters (between 6.7 and 7.0). The same general pattern is displayed in the other cases as well. For example, consider Case 4, where the largest possible shape parameter is 5.0. In Case 4, the absolute difference is small (less than 0.008) for shape parameters less than 4.4 but grows from 0.02 when the shape parameter is 4.6 to 0.07 when the shape parameter is 5.0. The relative difference is largest (between 0.1 and 0.15) for shape parameters in the range 1.0 to 1.25 and 4.75 to 5.0.

Table C.1: Coefficients determined using scikits.statsmodels.OLS to fit the model $\log(\rho) = \alpha$Shape Parameter $+ \beta$.

Case	ξ	Number of observations	Coefficients	
1	6.0	1000	α	1.37664
			β	-7.07234
2	5.5	1000	α	1.36028
			β	-6.86402
3	7.0	1000	α	1.41517
			β	-7.47105
4	5.0	1000	α	1.26068
			β	-7.14568

C.2 Unraveling

It is natural to wonder if the approximately exponential behavior observed in the intensity function ρ for large mutation spaces with gamma profile mutations is also observed with different types of mutation profiles. Large spaces of point-mass profile mutations are not a good candidate because they can lead to an unraveling of the solution to the free recombination model. Unraveling occurs when mutations that have only very late-acting effects build up over time due to the low selective pressure against them. With the number of such mutations tending to infinity, hazard rates spike at the oldest ages, producing a Wall of Death, that is, an age after which no one survives. The Wall of Death at a late age leads to a reduction in selective pressure against mutations with infinitesimally younger late-age effects. This reduction in selective pressure, in turn, allows infinitesimally younger late-age mutations to accumulative, producing a spike in hazard rates at this infinitesimally younger old-age. With a Wall of Death at an infinitesimally younger age, selective pressure is reduced for mutations with slightly younger age effects. And so on. This process can continue until the Wall of

Death reaches the age of maturity. In this case lifespan approaches what Tuljapurkar [31] calls the "salmon limit": an individual grows to maturity, reproduces and instantly dies. Wachter, Evans and Steinsaltz discuss this phenomenon in more detail in [35]. In particular, they show that unraveling occurs when the mutation space consists of point-mass profiles with ages of onset m at all reproductive ages, $\mathcal{M} = [\alpha, \infty]$ for any constant mutation rate ν. However, this also holds when fertility is constant over a finite range of ages, α to β.

C.3 Sigmoid Behavior

A candidate suggested by Wachter, Evans and Steinsaltz [35] is the modified point-mass profile. With a modified point-mass profile, each mutation has a small initial cost. The mutation profile is modeled as a double step function, with a step of size δ at the age of maturity α, and a second step of size $1-\delta$ at age m, the age of onset. The test case considered here is a mutation space with 351 modified point-mass profiles was ages of onset ranging from $\alpha = 15$ to $\beta = 50$ in step sizes of 0.1 years. The background hazard rate for this test was set to the standard 0.05 and the size of the mutation effect was 0.1. The parameter δ was 0.001.

Figure C.4 (left) shows ρ (solid line) for the modified point-mass profile case. Unlike the cases with gamma profile mutations, ρ in this case does not exponentially increase. Rather, it follows a sigmoid curve: the intensity function is fairly flat, $\rho \approx 0$, until around age 35, then it rapidly increases to $\rho \approx 4.8$ by age 40 and remains fairly constant at 4.8 for ages of onset above 40. The intensity measure was fitted by the curve

$$\hat{\rho}(m) = \frac{k}{1 + \exp(-t_0 m - t_1)}$$

using curve_fit from scipy.optimize. The algorithm was not seeded with initial values for the parameters. The coefficients returned by the curve_fit function were $k = 4.85527$, $t_0 = 2.88929$ and $t_1 = -109.039$. The approximation using these coefficients is plotted with a dotted line. In general the approximation is quite close to the actual intensity ρ, although the intensity function is slightly larger and increases more gradually than the approximation over the ages 30 to 35. This observation is verified by considering the absolute difference between ρ and its sigmoid approximation, shown in Figure C.4, right.

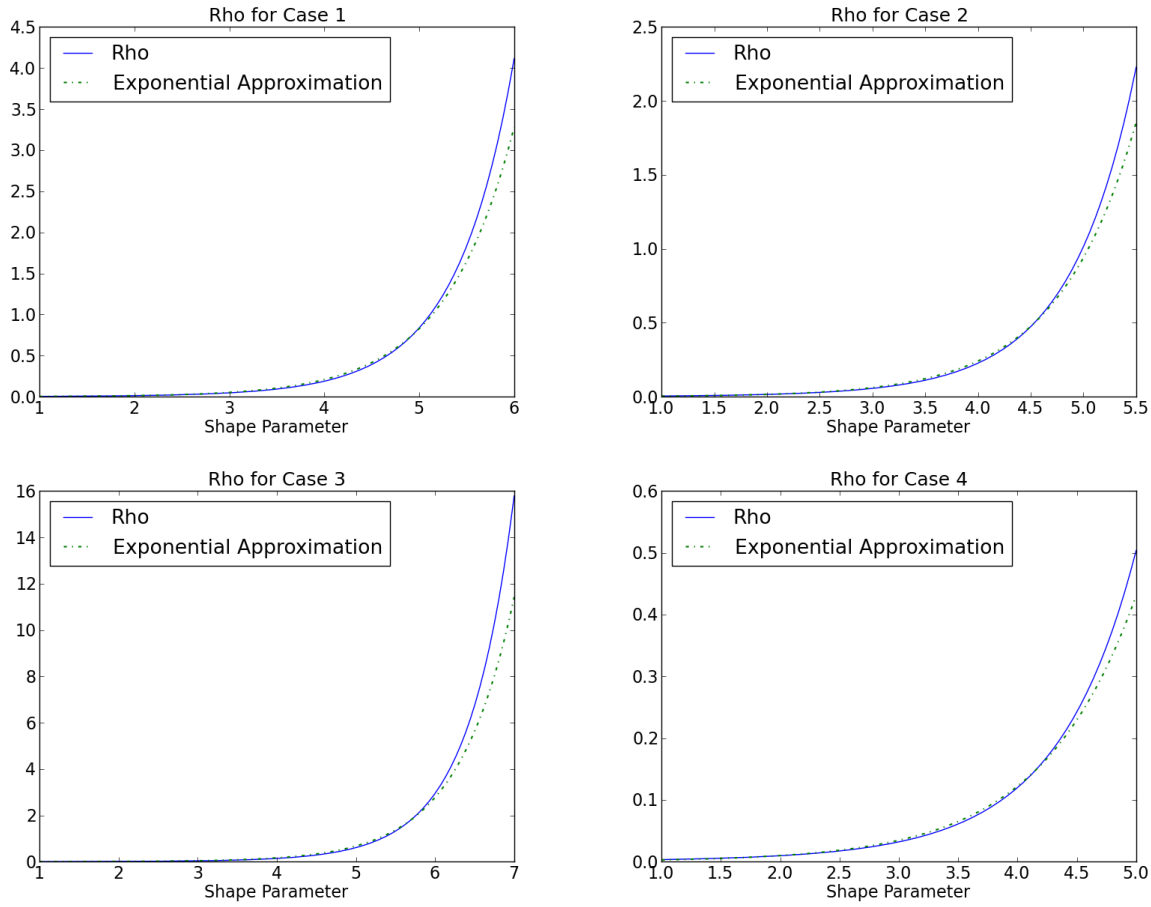

Figure C.1: Intensity measure ρ (solid line) and the exponential approximation to ρ (dotted line) for the free recombination model in Case 1 (top left), Case 2 (top right), Case 3 (bottom left) and Case 4 (bottom right). The intensity measure for each case was fitted with an exponential curve whose parameters are listed in Table C.1.

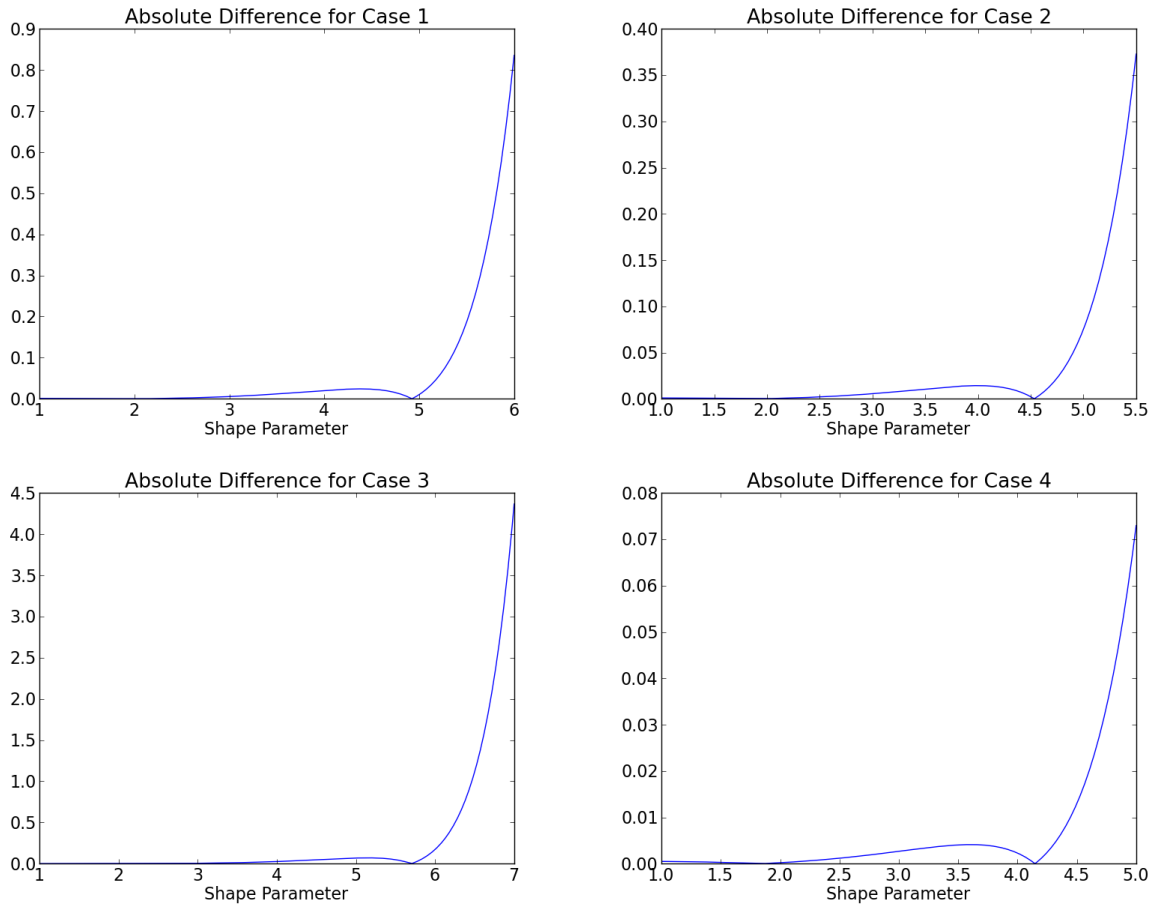

Figure C.2: Absolute difference between the intensity measure ρ and the exponential approximation to ρ for the free recombination model in Case 1 (top left), Case 2 (top right), Case 3 (bottom left) and Case 4 (bottom right). The intensity measure for each case was fitted with an exponential curve whose parameters are listed in Table C.1.

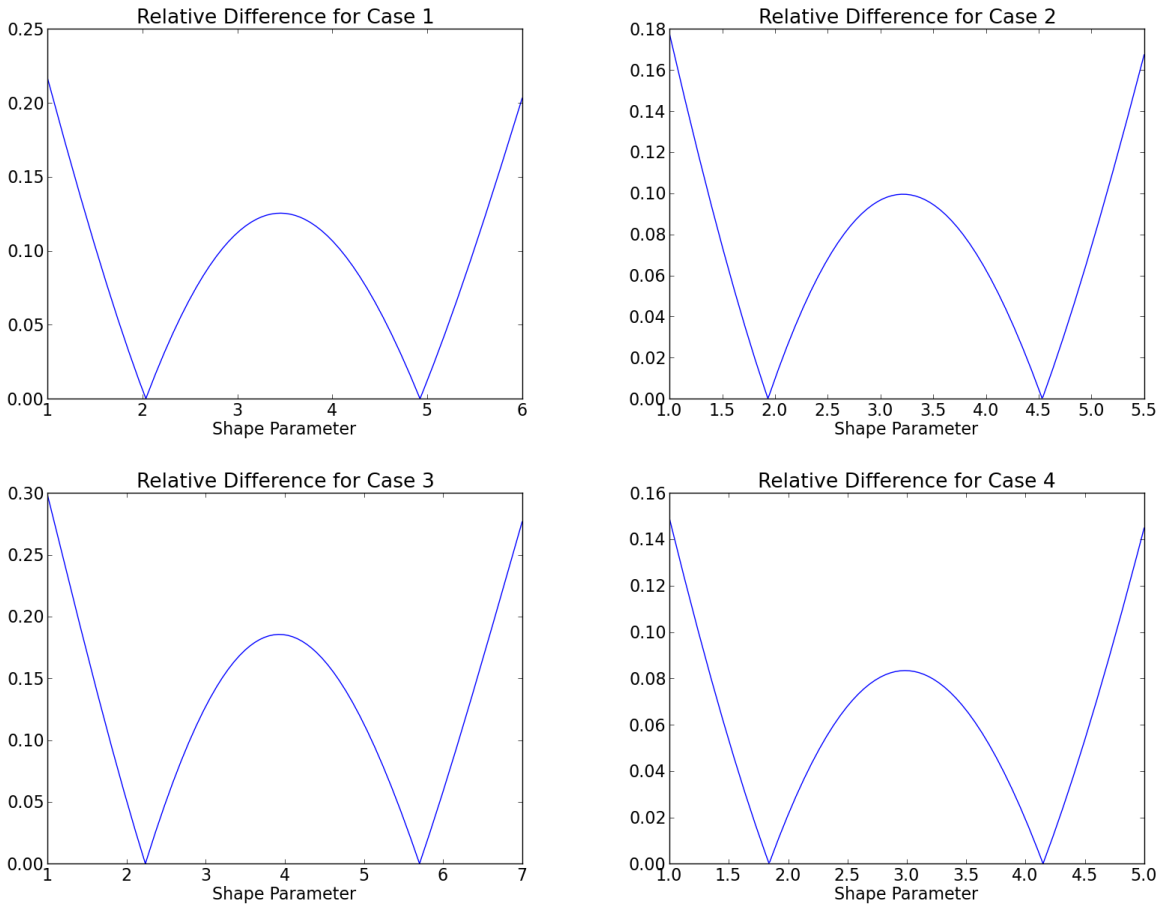

Figure C.3: Relative difference between the intensity measure ρ and the exponential approximation to ρ for the free recombination model in Case 1 (top left), Case 2 (top right), Case 3 (bottom left) and Case 4 (bottom right). The intensity measure for each case was fitted with an exponential curve whose parameters are listed in Table C.1.

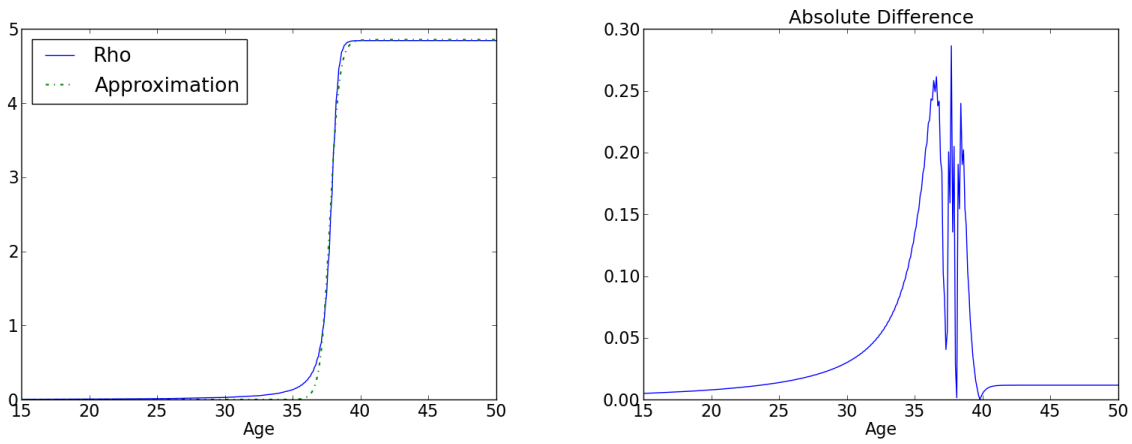

Figure C.4: The plot on the left shows the intensity measure ρ (solid line) and the sigmoid approximation to ρ (dotted line) for the free recombination model with modified point-mass profiles. The plot on the right shows the absolute difference between the ρ and the approximation.