INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Monte Carlo Methods for Inference in Population Genetic Models

Eric C. Anderson

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2001

Program Authorized to Offer Degree: Interdisciplinary Program in Quantitative Ecology and Resource Management

UMI Number: 3022805

UMI Microform 3022805

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Ann C. Anderson

Date 7/19/01

University of Washington

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Eric C. Anderson

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

Elizabeth Thompson

Elizabeth A. Thompson

Reading Committee:

Elizabeth Thompson

Elizabeth A. Thompson

Joseph Felsenstein

Joe Felsenstein

R S Waples

Robin S. Waples

Date: July 18, 2001

University of Washington

Abstract

Monte Carlo Methods for Inference in Population Genetic Models

by Eric C. Anderson

Chair of Supervisory Committee

Professor Elizabeth A. Thompson
Department of Statistics

This dissertation describes novel applications of Monte Carlo and Markov chain Monte Carlo (MCMC) techniques to statistical inference in problems from the field of conservation genetics. The inference problems are motivated by issues arising in the conservation and management of trout and salmon. The first half of the thesis deals with estimating effective population size and related quantities from temporally spaced samples of genetic data. A likelihood function for organisms with discrete generations is developed, based on the Wright-Fisher model and a hidden Markov chain formulation. Importance sampling methods for computing this likelihood are presented and applied to published data on *Drosophila*. Some modeling assumptions implicit in the use of the Wright-Fisher model are detailed, and a new model for genetic inheritance, based on a Pólya urn scheme, is presented and characterized. Methods are developed using this model for the MCMC estimation of the likelihood or posterior probability for the effective size of a population or for the ratio λ of the per-generation number of effective breeders to census breeders in a population. The Pólya urn model forms the basis for a probability model of allele frequency change conditional on λ in salmon populations. Specifying this model in terms of a large directed graph simplifies the application of a single-component Metropolis-Hastings algorithm for computing the posterior distribution of λ. The method is applied to genetic data simulated

upon the census counts of a threatened salmon population in Idaho, demonstrating that the method allows precise estimates of λ with such data.

The second half of the thesis focuses on approaches to inference within populations of recently-hybridized populations. First, I extend the methods of PRITCHARD *et al.* (2000) to allow inference of pure and admixed categories of individuals in structured populations. Finally, I develop methods based on explicit modeling of recent hybridization and evaluate the potential for using genetic data to distinguish F_1, F_2 and backcrossed hybrid categories among sympatric, hybridizing populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I owe thanks to a great many people who helped me in the preparation of this dissertation. I must first acknowledge the support, encouragement and criticism of my supervisory committee, starting with the mentoring given me by my committee chair, Elizabeth Thompson. I will be forever grateful that Elizabeth was willing to take me on as a student five years ago, despite my relative lack, at the time, of formal, college, mathematical training. In addition to profiting from Elizabeth's habit of being always available to her students, I also benefited tremendously from her mentoring style, knowledge of statistical genetics, and innate curiosity of statistical and genetic matters of all variety. I also owe special thanks to Joe Felsenstein. Were it not for his generosity in offering me an independent study in population genetics six years ago, I would never have found my way to this field that I find so rewarding. Robin Waples has been helpful throughout the dissertation process. In the early stages he suggested fruitful lines of research, in the middle stages he kindly provided me with useful datasets, and in the final stages he contributed many insightful comments and criticisms of the work in these pages. I met Matthew Stephens while he was a postdoctoral researcher at the University of Oxford. My visit to Oxford in October of 1999 inspired two of the chapters in this dissertation. Matthew, who has provided a great amount of technical help, is also an inspiring statistician to me, as well as a valued friend. Finally, Richard Fenske has been a model Graduate-School Representative. He was flexible in scheduling meetings, and it was a pleasure to get to know him during my years at the University.

I thank the QERM program for its sturdy and resilient existence through some of the trying times in the last four years. I have enjoyed the contact I have had with fellow QERM students. I particularly acknowledge my discussions with David Caccia on stochastic processes, Markov chain Monte Carlo, and perfect sampling methods, and I give thanks for the typically animated comments that Professor David Ford always had for me and others

when presenting at the QERM seminar series.

Elizabeth Thompson's students and postdocs in Statistics and Biostatistics were an invaluable resource for me, and I particularly thank Nicky Chapman (with whom I shared dissertation-completion-neurosis during spring quarter) and my esteemed office mate, Andrew George. I have had the pleasure of occupying an office in the Department of Statistics for the last three years. My exposure to the distinguished faculty and the dedicated students there was a wonderful experience. The Department never ceases to impress me, and I am honored to have been a part of it.

I regularly attended several seminar groups that enriched and broadened my experience at the University. I particularly thank the MathBio Czar, Professor Thomas Daniel in the Department of Zoology and Professors Garry Odell and Elizabeth Thompson for organizing and orchestrating the Mathematical Biology seminar in Zoology. I have also enjoyed the long-runnning Statistical Genetics seminar group in Statistics and Biostatistics, and, in my last year here, have had the pleasure of attending Adrian Raftery's Model-Based Clustering Working Group in Statistics.

I thank my collaborators on talks and papers in the last four years. I collaborated with Ellen Williamson at University of California, Berkeley on a paper that formed the material for Chapter 2 of this dissertation. In connection with that collaboration, I acknowledge the hospitality of Montgomery Slatkin and his lab group during my visits to Berkeley. Jonathan Pritchard was a collaborator for a talk I presented at the International Biometrics Conference in 2000. This work, which eventually grew into Chapter 5, started from several conversations during a brief visit to University of Oxford, where I was hosted by the lab of Peter Donnelly in the Department of Statistics. Most recently I have had the pleasure of teaming up with Paul Scheet for a short article in *Genetics*. Mark Beaumont and David Teel provided datasets that I use in Chapters 5 and 6.

During my years working on this dissertation I have been supported by National Science Foundation grant BIR–9807747 to Elizabeth Thompson, the National Science Foundation Mathematical Biology Training Grant #BIR–9256532 to Thomas Daniel and Garry Odell,

the Burroughs Wellcome Fund, Program in Mathematics and Molecular Biology, and a QERM Student Training Grant.

While in Seattle, outside the academic sphere, I have received support from a network of friends too numerous to thank individually. Through it all, always, I thank my mother, father, and sister for their unflagging love and support.

Finally, as my undergraduate training was in biology, with few mathematics courses, I have had to rely on the math background I gained as a high school student. The Mathematics Department at The Thacher School, and particularly Kurt Meyer, deserve special recognition. And I thank my dear friend and mentor of math and mountains, John Rosendahl.

Chapter 1

INTRODUCTION

The last two decades have witnessed remarkable advances in the fields of biotechnology and computation. As a result, biologists who study or manage natural populations of plants and animals today have access to numerous new tools. Of great value are techniques for assaying genetic variability within individuals, and within populations. With the advent of polymerase chain reaction (PCR) and the discovery of new, polymorphic, genetic markers, a number of questions about populations and the interactions of the individuals within them may be addressed using genetic data. The advances in computation have been equally significant. Today's computers are fast enough to allow numerically intensive statistical analyses to be run on desktop machines. This has led to significant development in the field of computational statistics, particularly in the use of Monte Carlo and Markov chain Monte Carlo (MCMC) techniques for computing likelihoods and posterior probabilities. This dissertation focuses on developing likelihood-based statistical procedures and the computational methods necessary to apply them to the analysis of genetic data in the context of questions of interest to biologists, conservation biologists and wildlife managers.

Each of Chapters 2 through 6 provides an almost self-contained account of the application of statistical theory and computational techniques to a practical problem in population genetics. Chapter 2 describes a hidden Markov model and an importance-sampling method for evaluating the likelihood for the parameter N_e—the genetically effective size—of a population from temporally-spaced genetic data. Chapter 3 considers the direct estimation of the ratio of the effective number of breeders to the census number of breeders; a parameter we call λ. This chapter describes problems with modeling genetic transmission in populations via the well-known Wright-Fisher model, and advances, instead, a model of genetic

inheritance derived from a Pólya urn scheme. Using a hidden Markov model with transitions based on this urn scheme, the final section of the chapter describes an MCMC method for computing the likelihood or posterior distribution for λ or N_e. Chapter 4 uses the urn scheme of the previous chapter to develop a likelihood model for λ in the case of salmon populations which, because of their interesting life histories, violate the discrete generation assumption of the standard Wright-Fisher model. This likelihood is used for Bayesian inference of λ in salmon populations. Once again, this relies on MCMC.

Chapters 5 and 6 deal with a different problem in population genetics—that of identifying the origin of individuals in mixed and admixed (*i.e.*, including some hybrid individuals) populations. Chapter 5 describes an extension to the work of PRITCHARD *et al.* (2000) allowing individuals in an admixed population to be of either pure or admixed genetic origin and uses an example with data from wildcats (*Felis sylvestris*) in Scotland. Chapter 6 develops a model for admixture that is more appropriate when hybrids have only been formed over a small number of generations. This method uses genetic data to compute the posterior probability that an individual in a sample is, for example, a pure individual, an F_1 or F_2 hybrid, or a backcross. The method is demonstrated on data from coastal cutthroat trout (*Oncorhynchus clarki*), steelhead trout (*O. mykiss*), and their hybrids. The final chapter, 7, includes brief concluding remarks and describes further work to be done in the areas treated by this dissertation.

This leaves the present chapter to briefly summarize how and why the main chapters of this dissertation are related and connected. I see three levels of connection. First, each practical application described confronts a problem in the study and management of natural (non-human) populations using genetic data that are increasingly available. In my case, each of these applications has been motivated by a particular problem which occurs in the study of salmon populations. Second, each chapter makes heavy use of Monte Carlo computational techniques. While Monte Carlo likelihood and MCMC have been used since the early nineties in genetic problems involving computations on pedigrees, and soon thereafter were applied to inference on evolutionary time scales, their application to population genetics problems has been more recent, and the chapters herein represent some of the first applications of MCMC to genetic inference at the population level. The

third point is more of a statistical curio than a unifying theme, but I find it interesting nonetheless—despite the fact that the problems addressed in Chapters 5 and 6 are quite different from those in Chapters 2 through 4, we encounter in all of them useful applications of hidden Markov models, mixture models, and urn models. In the remaining sections of this chapter, I will elaborate on the first two points, and for those unfamiliar with them, I provide a brief introduction to Monte Carlo techniques.

1.1 Genetic Data

Advances in biotechnology have revolutionized conservation biology and resource management. AVISE *et al.* (1995) review the genetic markers currently available to researchers, discuss the types of analyses those markers allow, and review applications in conservation genetics. The advent of Mendelian-inherited microsatellite markers (TAUTZ 1989; WRIGHT and BENTZEN 1994) has made informative genetic data increasingly available and inexpensive for such applications. Among other examples, DNA markers amplified from fin clips have been used in monitoring Pacific salmon (OLSEN *et al.* 1996), while hair samples have been used in studying bear (TABERLET *et al.* 1997), and chimpanzee (MORIN *et al.* 1993) populations. PCR-based technologies are especially appropriate for populations of conservation interest as sampling is non-destructive and/or non-invasive. It is thus possible to obtain data at multiple time points, and, since the markers are typically polymorphic, the data are informative in characterizing the population at each time point, and hence also in detecting and quantifying the gene frequency changes caused by small effective population size or genetic exchange with other populations.

With microsatellite markers and PCR, data may be extracted from archived tissues, giving the opportunity to obtain data from time points in a population's past. For example, museum-preserved skins from known populations of the pocket gopher provide genetic data on the populations at two time points (1950's, 1970's) which may be compared to current samples (Ellie Steinberg, UW Dept. of Zoology, pers. comm.). For some fish populations, the situation is even better. Some such populations have been the subject of long-term ecological research efforts with population size estimates available on a yearly basis, and

age composition inferred from fish scales. Genetic marker data may be obtained from these archived scales. Recently, MILLER and KAPUSCINSKI (1997) isolated DNA from northen pike scales collected from Lake Escanaba, WI. Using data from three years, 1961, '77, and '93, they estimated N_e from the temporal changes in allele frequencies over the two time intervals. In a similar, ongoing study, microsatellites from archived juvenile Keogh River (Vancouver Island) and Snow Creek (Washington State) steelhead scales provided ample material for amplifying microsatellite markers by PCR (ARDREN 1999). The use of archived fish tissues for population genetic studies is further described by NIELSEN *et al.* (1999).

1.2 Salmon Populations and Genetics

Pacific salmon (*Oncorhynchus* spp.) and their relatives, Atlantic salmon and the true trouts (*Salmo* spp.) and chars (*Salvelinus* spp.) in the family Salmonidae are an evolutionarily fascinating and commercially valuable group of fish. Both of these factors have contributed to the generation of an enormous amount of genetic data on these species. The five species of Pacific salmon native to the West Coast of North America, sockeye (*O. nerka*), chinook (*O. tshawytscha*), coho (*O. kisutch*), chum (*O. keta*), and pink (*O. gorbuscha*) exhibit an anadromous and semelparous life history; they hatch from eggs in fresh water, migrate to the ocean and mature to adulthood, then return to their natal fresh waters to spawn, and then die soon thereafter. A remarkable feature of these migrations is the salmon's homing ability. Returning adults typically spawn in the stream in which they were born. While this homing behavior is not always perfect, and straying (spawning in a non-natal stream) is known to occur, their homing does lead to a situation in which salmon populations represent relatively isolated, reproducing populations with low gene flow between populations. A comprehensive review of life histories in Pacific salmon may be found in the volume edited by GROOT and MARGOLIS (1991).

Atlantic salmon (*Salmo salar*), steelhead trout (*O. mykiss*), and coastal cutthroat trout (*O. clarki*) display anadromous forms and homing ability; however, they do not necessarily die after spawning. *O. mykiss* and *O. clarki* also exhibit non-anadromous or resident forms. Many of the chars (*Salvelinus* spp.) also exhibit resident and anadromous forms. There

has been considerably less research on chars than on Pacific salmon and the true trouts, but there is still considerable genetic information about them (see LEARY and ALLENDORF (1997) and TAYLOR *et al.* (2001) and references therein).

Commercial fishing for Pacific salmon is a mammoth industry on the West Coast. Sport fishing is also economically important to the region. These influences have had a siginficant impact on salmon population abundnace. Additionally, the nature of their life history makes salmon vulnerable to a number of different anthropogenic disturbances that affect the environments in which they live and through which they migrate. For example, hydropower operations, forestry, agriculture, road-building, and wetlands destruction all impact salmon at some some point in their life history. In response to dwindling population sizes and the wholesale destruction of spawning habitat for hydroelectric power generation, a great number of salmon hatcheries have been built to try to maintain or supplement salmon production on the West Coast. In recent years the aquaculture practice of net-pen rearing of Atlantic salmon on both coasts has increased dramatically. Hatchery and aquaculture practices have brought with them their own suite of impacts on wild salmon populations. In the lower 48 states of the United States, in particular, but also in Canada and Alaska, all of these impacts have resulted in extinction or crisis for a significant number of salmon populations (NEHLSEN *et al.* 1991).

The use of genetic markers in the management of salmonid stocks has a long and extensive history (RYMAN and UTTER 1987). More recently, with the application of the Endangered Species Act to populations of Pacific salmon, genetic characteristics of populations along with other traits are used in delineating the "Evolutionarily Significant Units" to which the Act is applied (WAPLES 1995). Consequently, there is a vast literature on salmon genetics and the use of genetic markers in the conservation of salmon. My purpose in the remainder of this section is not to attempt a comprehensive review of that literature (such reviews and case studies may be found in ALLENDORF and WAPLES (1996) and the articles co-published with UTTER (1999)) but rather to first give the reader some sense of the varieties of genetic data on salmon and then to highlight the uses of those data that are particularly relevant to this dissertation.

The first, and still the most widely-used, genetic markers for salmon are electrophoret-

ically detectable enzymes described and reviewed by UTTER *et al.* (1987). Today there are well over 60 commonly-used allozyme loci routinely used by the National Marine Fisheries Service in screening salmon populations for genetic variation and for other purposes. However, many newer, molecular markers have been discovered and used in salmon as well. GYLLENSTEIN and WILSON (1987) describe mitochondrial DNA (mtDNA) markers in the era before PCR. Also in the pre-PCR era, various probes were developed for hybridization to restriction-enzyme-digested total genomic DNA of salmon (DEVLIN *et al.* 1991). With the ability to amplify DNA by PCR, mtDNA markers were more easily prepared and other, new types markers became available. First were markers that were associated with known, functional genes like those coding for ribosomal DNA (PENDAS *et al.* 1995) or growth hormone (GROSS and NILSSON 1995). Soon thereafter, however, probes detected minisatellite loci in all salmon species (PRODOHL *et al.* 1995), and the method of Random Amplification of Polymorphic DNA (RAPD) was applied to salmon (ELO *et al.* 1997). Today, microsatellite markers (OLSEN *et al.* 1996) are widely available for salmon. Research into new markers for salmon continues, with the discovery of new short interspersed repeat (SINE) segments being a recent example (PEREZ *et al.* 1999).

Genetic markers have been used in at least several hundreds of studies of salmon populations. Some of these concern themselves primarily with reporting genetic variation across and within populations; however, many of them are directed toward answering specific questions or estimating particular quantities associated with the populations. Of particular relevance to this dissertation are studies involving genetic stock identification, the estimation of effective population size, and the detection of hybrids. In later chapters, I present novel computational methods for each these tasks.

Genetic stock identification refers to the use of multilocus genotypes (without knowledge of phase) to assign individuals sampled from a mixture of fish to one of the populations contributing to the mixture, and also to estimate the proportion of fish in the mixture from each of the source populations. The empirical studies employing these techniques are too numerous to list; however, WOOD *et al.* (1989) is a representative example using allozymes and other biological trait data, BEACHAM (1996) describes genetic stock identification using minisatellites and BEACHAM *et al.* (2000) and OLSEN *et al.* (2000) both provide represen-

tative cases using microsatellite markers.

Interestingly, the availability in the early 1980's of allozyme data from salmon populations and problems in salmon population management were the primary factors motivating the original development of statistical methodologies for these sorts of genetic mixtures. MILNER *et al.* (1981) present the basic framework and an EM-algorithm for maximum likelihood methods in genetic stock identification conditional on allele frequencies being known from the populations contributing to the mixture. MILLAR (1987) analyzes conditions of identifiability and shows that another method known as the "classification" method is a special case of the the likelihood approach. SMOUSE *et al.* (1990) present an EM-algorithm for inference in the case when the allele frequencies from the contributing stocks is not assumed known without error, and PELLA and MASUDA (2001) give a Bayesian version using MCMC to simulate from the posterior distributions of interest.

The estimation of effective size of salmon populations is another problem that has spurred the development and refinement of statistical methods. The work of WAPLES (1989), clarifying and generalizing previously developed F-statistic methods for estimating effective population size from temporal changes in allele frequencies, was motivated in large part by problems in salmon biology. In a series of papers, he explores genetic change over time (WAPLES and TEEL 1990; WAPLES 1990a) in salmon populations and tailors an F-statistic method for estimating effective size to the Pacific salmon life history (WAPLES 1990b). This method was later made more practical by an algorithm presented in TAJIMA (1992). These methods have been used to estimate the effective sizes of endangered salmon populations (WAPLES *et al.* 1993). Estimation of brown trout (*Salmo trutta*) effective sizes (JORDE and RYMAN 1996) motivated the development of another method for applying F-statistic estimators to temporal allele frequency data for the estimation of effective size in populations with overlapping generations (JORDE and RYMAN 1996).

A number of empirical studies have been conducted, estimating the effective size or the effective number of spawners in salmon populations. ARDREN (1999) uses microsatellites amplified from fish scales to try to estimate the effective size of two West Coast steelhead populations. KINCAID (1995) analyzes mating patterns and breeding history to try to estimate inbreeding effective size in hatchery populations of salmonids. HEDRICK *et al.*

(1995) discuss and estimate the effective size of winter run chinook salmon. TESSIER *et al.* (1997), using microsatellite and mtDNA markers, assess the effective population size of landlocked Atlantic salmon populations and the effect that supportive breeding programs have on the effective size of these populations.

The literature on the use of genetic methods for detecting hybrids between salmonid species is enormous and will not be exhaustively reviewed here. The studies using genetic markers include those that investigate naturally-occurring hybridization between sympatric species (*cf.* CAMPTON and UTTER 1985; TAYLOR *et al.* 2001; ELO *et al.* 1995) and numerous ones involving interbreeding between hatchery or farm-reared salmon and native populations of conspecifics (for example, CLIFFORD *et al.* 1998) or other species (JANSSON and OEST 1997).

1.3 Monte Carlo in Genetics

Many inference problems in statistical genetics involve complex stochastic models that include a great number of variables. Typically only a fraction of these variables can be directly observed. These variables summarize the observed genetic data. The remainder of the variables in the stochastic model are not directly observable and are referred to as *latent* variables. The likelihood function for such inference problems can be expressed as the sum over the latent variables of the joint probability of the observed data and the latent variables, conditional on the genetic parameters of interest. Often, however, the space of latent variables is huge and that sum is not directly computable.

Monte Carlo methods are stochastic integration techniques that are useful for approximating such intractable sums. I provide a brief introduction to Monte Carlo methods in Section 1.5, after establishing conventions for mathematical notation. Here I briefly review the recent history of Monte Carlo techniques, and, in particular, Markov chain Monte Carlo (MCMC) techniques in statistical genetics.

Four years ago, as I was starting the research reported in this dissertation, MCMC methods had been used extensively in the analysis of data on extended or complex pedigrees. In segregation and linkage analysis of trait and genetic marker data observed on members

of a known pedigree, MCMC is used to sample over the space of latent variables which may be defined as genotypes (GUO and THOMPSON 1992) or meiosis indicators (THOMPSON 1994) or both (LANGE and MATTHYSSE 1989). Much work had been done in this field to improve the mixing of chains in the space of latent variables. One example in the area of inference of ancestral types on a large complex pedigree structure is the simulated tempering method of (GEYER and THOMPSON 1995). To improve mixing and ensure irreducibility, methods were developed in which multiple components of the latent variables are updated simultaneously. In pedigree analysis, such methods include use of a block-updating Gibbs sampler (JANSS *et al.* 1995), a whole-meiosis Gibbs sampler (THOMPSON and HEATH 1999), and a whole-locus Gibbs sampler (KONG 1991). Brand new MCMC methods, at the time, like reversible-jump MCMC (GREEN 1995) were quickly adopted to analyze more complex model spaces in genetic analysis. HEATH (1997) applies reversible jump methods to detect and locate multiple quantitative trait loci (QTL) from trait and genome-scan data, where the number of QTL is not prespecified and thus the dimension of the model varies within a single MCMC run.

MCMC methods had also been used in analyses of inference of relationship among individuals from genetic data. PAINTER (1997) develops methods for estimation of sibship structure, sampling directly over the space of alternative sibship structures, using data on microsatellite markers. GEYER *et al.* (1993) use a Metropolis-Hastings MCMC method to construct a Monte Carlo likelihood function for relationship parameters among a group of individual California condors, on whom there are multilocus DNA fingerprint data. At the other extreme of the evolutionary time scale, MCMC methods had also been used in phylogenetic analyses, to estimate evolutionary parameters, such as the product of mutation rate and effective size (KUHNER *et al.* 1995), or the rate of increase of populations (KUHNER *et al.* 1997). In these analyses, the latent variable is the structure and inter-coalescence times of the ancestral coalescent (KINGMAN 1982) of a sample of DNA sequences, and is sampled using a Metropolis-Hastings algorithm. NEWTON *et al.* (1997) propose an alternative specification of the coalescent structure, leading to an MCMC sampler which can make large changes in ancestral topology in a single MCMC step. In some cases, this specification may provide a better mixing sampler. Other Monte Carlo likelihood methods had also been

used for coalescent models in the context of estimation of growth rates (GRIFFITHS and TAVARÉ 1994) and recombination rates (GRIFFITHS and MARJORAM 1996), and inference of mutation models (NIELSEN 1997).

Since that time, the use of MCMC methods in statistical genetics has grown dramatically, due in large part to the growing acceptance of Bayesian techniques and the close association between Bayesian computation and MCMC. STEPHENS (2001) reviews recent MCMC and importance sampling methods for likelihood and Bayesian inference using coalescent models, and the numerous new MCMC approaches for inference of phylogenies are reviewed by HUELSENBECK and BOLLBACK (2001). MCMC methods have also enjoyed further refinements and applications in the detection of quantitative trait loci from data on outbred pedigrees (HOESCHELE 2001) and in the inference of breeding values in animal breeding (GIANOLA 2001).

While MCMC methods have been used on pedigrees and on coalescents for some time, only recently have they been used in conservation genetics problems at the intervening population time-scale where a pedigree structure is not available and the data, either observed or latent, are allele frequencies in specified populations. The methods developed in this thesis are applications of MCMC to problems relevant in conservation genetics. Other Monte Carlo approaches relevant to conservation genetics on short time scales have been developed in the last several years and include the detection and characterization of recent bottlenecks using genetic data (BEAUMONT 1999), the estimation of effective population size (KITADA *et al.* 2000), the analysis of population structure and admixture (PRITCHARD *et al.* 2000), and genetic stock identification in fisheries management (PELLA and MASUDA 2001).

1.4 Notational Conventions

Statistical genetics is rich with complex data arrangements and, therefore, multiply subscripted variables. To avoid confusion, I adopt the following conventions for mathematical notation, and apply them throughout the chapters of the dissertation: scalar random variables are either uppercase, italicized, Roman letters or lowercase Greek letters; for example X or θ. Greek letters are also used for quantities considered to be parameters in a frequentist

setting. For the Roman letters, vector-valued random variables are written in slant-bold, with their components being written as scalars with a single subscript: $\boldsymbol{X} = (X_1, \ldots, X_n)$. Random variables which are collections of vectors are denoted by Roman bold, $\mathbf{X}$, with the i^{th} vector element within it denoted by $\boldsymbol{X}_i$ and scalar components by double subscripts separated by a comma, *i.e.*, $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$ with $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,n})$. Some collections of vectors require more than one subscript to denote each vector. These are still denoted by Roman bold characters with the obvious extensions; for example $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$, $\mathbf{X}_t = (\boldsymbol{X}_{t,1}, \ldots, \boldsymbol{X}_{t,m})$, and $\boldsymbol{X}_{t,j} = (X_{t,j,1}, \ldots, X_{t,j,n})$. The shading and subscripting conventions for Greek letter random variables or parameters are similar, with uppercase bold substituted for collections of vectors: scalar θ, vector $\boldsymbol{\theta}$, collection of vectors (and beyond) $\boldsymbol{\Theta}$.

$P(\cdot)$ denotes probability mass and density functions alike. $P(X)$ is used as a name for the distribution of the variable X and also to denote the marginal probability mass or density at a realized value of X, *i.e.*, $P(X)$ also serves as a shorthand for $P(X = x)$. Only in cases where confusion would be likely or where the distinction is of central importance shall I make a notational distinction between the realized value of a random variable and the random variable itself. In such cases, as in the following section on Monte Carlo, I use the lowercase version of the random variable to denote the realized value. When dealing with a sequence of realized values of a random variable X generated as a sample to be used for Monte Carlo, I denote the i^{th} such member of the sample by $X^{(i)}$. The notation $P(\cdot|\cdot)$ is used for conditional probability mass and density functions. Hence $P(X|Y)$ is the conditional probability of X given Y. When multiple variables are involved, commas are used between them; $P(Y, Z|W, X)$ is the conditional joint probability of Y and Z given W and X. Sometimes the probability depends on a parameter, and that dependence is more conveniently expressed by subscripting the P. For example, $P_\theta(X|Y)$. The expected value of a random variable X, is written $\mathbb{E}(X)$, and the conditional expectation of X given Y is written $\mathbb{E}(X|Y)$. If the expection of X is taken with reference to a distribution indexed by a particular parameter, say θ_0, it may be written as $\mathbb{E}_{\theta_0}$. The variance of X is denoted $\text{Var}(X)$.

Occasionally, such as in the following section or while denoting proposal distributions

for Metropolis Hastings sampling, I shall use upper or lowercase italic letters to refer to probability mass or density functions of random variables.

1.5 Monte Carlo techniques

This section provides elementary background on the Monte Carlo method, importance sampling, Rao-Blackwellization, and Markov chain Monte Carlo.

1.5.1 History and definition of Monte Carlo

The term "Monte Carlo" was apparently used by Stanislaw Ulam and John von Neumann as a Los Alamos code word for the stochastic simulations they applied to building hydrogen bombs following World War II. Shortly after the War, and coinciding with the debut of the ENIAC computer in 1947, von Neumann and Ulam suggested that the ENIAC would be useful for applying "statistical sampling" approximations to solving the problem of neutron diffusion in fissionable material. Their methods, involving the laws of chance, performed well, and soon thereafter were aptly named by Nick Metropolis after Monte Carlo, the international gaming destination. The moniker stuck and soon after the War a wide range of difficult problems yielded to the new techniques (METROPOLIS 1987). Despite the widespread use of the methods, and numerous descriptions of them in articles and monographs, it is virtually impossible to find a succint definition of "Monte Carlo method" in the literature. Perhaps this is owing to the intuitive nature of the topic which spawns many definitions by way of specific examples. Some authors prefer to use the term "stochastic simulation" for almost everything, reserving "Monte Carlo" only for Monte Carlo integration and Monte Carlo tests (RIPLEY 1987). Others seem less concerned about blurring the distinction between simulation studies and Monte Carlo methods. Be that as it may, I adopt the following terse definition:

> Monte Carlo is, in essence, the approximation of an expectation by the sample mean of a function of simulated random variables.

We will find that this definition is broad enough to cover everything that has been called Monte Carlo, and yet makes clear the fundamental feature of Monte Carlo in very familiar

terms: Monte Carlo is about invoking laws of large numbers to approximate expectations. This applies when the simulated variables are independent of one another, and may apply when they are correlated with one another, for example if they are states visited by an ergodic Markov chain. The Monte Carlo method is useful precisely because very many quantities of interest may be expressed as expectations.

While most Monte Carlo simulations are done by computer today, there were several applications of Monte Carlo methods using coin-flipping, card-drawing, or needle-tossing (rather than computer-generated pseudo-random numbers) centuries ago—long before the name Monte Carlo arose.

In more mathematical terms: Consider a random variable X (though depicted as a scalar, all of the following extends to multidimensional random variables) having probability mass function or probability density function $f_X(x)$ which is greater than zero on a set of values $\mathcal{X}$. Then the expected value of a function g of X is

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) \tag{1.1}$$

if X is discrete, and

$$\mathbb{E}[g(X)] = \int_{x \in \mathcal{X}} g(x) f_X(x) dx \tag{1.2}$$

if X is continuous. If we were to take an n-sample of X's, $(x^{(1)}, \ldots, x^{(n)})$[1], and we computed the mean of $g(x)$ over the sample, then we would have the Monte Carlo *estimate*

$$\widetilde{g_n}(x) = \frac{1}{n} \sum_{i=1}^{n} g(x^{(i)})$$

of $\mathbb{E}[g(X)]$. We could, alternatively, speak of the random variable

$$\widetilde{g_n}(X) = \frac{1}{n} \sum_{i=1}^{n} g(X^{(i)})$$

which we call the Monte Carlo *estimator* of $\mathbb{E}[g(X)]$.

---

[1] In this section, the distinction between the random variable X and its realized value x is crucial, so a notational distinction is made between the two—capital X and $X^{(i)}$ refer to random variables while lowercase x and $x^{(i)}$ refer to realized values of the random variable X.

If $\mathbb{E}[g(X)]$, exists, then the weak law of large numbers tells us that for any arbitrarily small ϵ

$$\lim_{n\to\infty} P\left(|\widetilde{g_n}(X) - \mathbb{E}[g(X)]| \geq \epsilon\right) = 0.$$

This tells us that as n gets large, there is small probability that $\widetilde{g_n}(X)$ deviates more than a tiny bit from $\mathbb{E}[g(X)]$. For our purposes, the strong law of large numbers says much the same thing—the important part being that so long as n is large enough, $\widetilde{g_n}(x)$ arising from a Monte Carlo experiment shall be as close to $\mathbb{E}[g(X)]$ as desired. This extends to samples from Markov chains via the weak law of large numbers for the number of passages through a recurrent state in an ergodic Markov chain (see FELLER 1957).

It should also be clear that $\widetilde{g_n}(X)$ is unbiased for $\mathbb{E}[g(X)]$:

$$\mathbb{E}[\widetilde{g_n}(X)] = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} g(X^{(i)})\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[g(X^{(i)})] = \mathbb{E}[g(X)].$$

These properties of random samples become useful when one realizes that very many quantities of interest may be cast as expectations. Most importantly for applications in this dissertation, it is possible to express all probabilities and summations as expectations:

Probabilities: Let Y be a random variable. The probability that Y takes on some value in a set $\mathcal{A}$ can be expressed as an expectation using the indicator function:

$$P(Y \in \mathcal{A}) = \mathbb{E}[\mathcal{I}\{Y \in \mathcal{A}\}] \tag{1.3}$$

where $\mathcal{I}\{Y \in \mathcal{A}\}$ is the indicator function that takes the value 1 when $Y \in \mathcal{A}$ and 0 when $Y \notin \mathcal{A}$.

Summations: Any sum, even the sum of a collection of deterministic variables may be represented as an expectation. For example the sum of a function $q(b)$ over the values of b in a finite set $\mathcal{B}$, may be written as an expectation of a random variable, say W, which takes values in $\mathcal{B}$. In the simplest form, W could take any value in $\mathcal{B}$ with equal probability p, and the sum could be cast as the expectation

$$\sum_{b\in B} q(b) = \frac{1}{p}\sum_{b\in B} q(b)p = \frac{1}{p}\mathbb{E}[q(W)].$$

The immediate consequence of this is that all probabilities and summations can be approximated by the Monte Carlo method. And further, there is no restriction that says W above must have a uniform distribution. This was just for easy illustration. We will explore this point more while considering importance sampling.

The variety of quantities that may be estimated by Monte Carlo is great. In addition to probabilities and summations, it is also possible to estimate integrals, probability distributions, and variances, *etc.* In each of these cases, the fundamental feature is the same—the quantity of interest may be expressed as an expectation which is then approximated by Monte Carlo. For example, to approximate a probability distribution, (1.3) may be applied K times to K different sets $\mathcal{A}_k$, $k = 1, \ldots, K$, which, taken together, give a histogram representation of the distribution. Or, to estimate the variance of a random variable Y, that variance may be expressed as an expectation (namely the expected value of $(Y - \mathbb{E}Y)^2$) and estimated via Monte Carlo accordingly.

Many problems in statistical genetics provide examples where a quantity of interest is a summation. In such cases the probability $P(Y)$ of an observed event Y must be computed as the sum over very many latent variables X of the joint probability $P(Y, X)$. In such a case, Y is typically fixed, *i.e.*, we have observed $Y = y$, and we are interested in $P(Y = y)$, but we can't observe the values of the latent variables which may take values in the space $\mathcal{X}$. Though it follows from the laws of probability that

$$P(Y = y) = \sum_{x \in \mathcal{X}} P(Y = y, X = x),$$

quite often $\mathcal{X}$ is such a large space (contains so many elements) that it is impossible to compute the sum. Application of the law of conditional probability, however, gives

$$P(Y = y) = \sum_{x \in \mathcal{X}} P(Y = y, X = x) = \sum_{x \in \mathcal{X}} P(Y = y | X = x) P(X = x). \tag{1.4}$$

The term following the last equals sign is the sum over all x of a function of x (namely, $P(Y = y|X = x)$), weighted by the marginal probabilities $P(X = x)$. Clearly this is an expectation, and therefore may be approximated by Monte Carlo, giving us

$$P(Y = y) \approx \frac{1}{n} \sum_{i=1}^{n} P(Y = y | X = x^{(i)}) \tag{1.5}$$

where $x^{(i)}$ is the i^{th} realization from the marginal distribution of X.

Unfortunately, (1.5) would probably provide a very poor Monte Carlo estimate. Though it is typically easy to formulate a quantity as an expectation and to propose a "naive" Monte Carlo estimator, it is quite another thing to have the Monte Carlo estimator actually provide good estimates in a reasonable amount of computer time. For most problems, a number of Monte Carlo estimators may be proposed; however some Monte Carlo estimators are clearly better than others. Typically, a "better" Monte Carlo estimator has smaller variance (for the same amount of computational effort) than its competitors. The variance of a Monte Carlo estimator is easily defined. Going back to our original notation, we have the random variable $\widetilde{g_n}(X)$, a Monte Carlo estimator of $\mathbb{E}(g(X))$. Like all random variables, we may compute its variance (if it exists) by the standard formulas:

$$\text{Var}(\widetilde{g_n}(X)) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} g(X^{(i)})\right) = \frac{\text{Var}(g(X))}{n} = \frac{1}{n}\sum_{x\in\mathcal{X}}[g(x) - \mathbb{E}(g(X))]^2 f_X(x) \quad (1.6)$$

if X is discrete, and

$$\text{Var}(\widetilde{g_n}(X)) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} g(X^{(i)})\right) = \frac{\text{Var}(g(X))}{n} = \frac{1}{n}\int_{x\in\mathcal{X}}[g(x) - \mathbb{E}(g(X))]^2 f_X(x)dx \quad (1.7)$$

if X is continuous. For the rest of this section, we will do everything in terms of integrals over continuous variables, but it applies equally well to sums over discrete random variables.

There are many ways to reduce the variance of Monte Carlo estimators. Of these "variance-reduction" techniques, the one called "importance sampling" is particularly useful. I include a short section on it here, as it will be used in Chapter 2.

1.5.2 Importance sampling

Importance sampling (HAMMERSLEY and HANDSCOMB 1964) is the art of choosing a good distribution from which to simulate one's random variables. It involves multiplying the integrand by 1 (usually dressed up in a "tricky fashion") to yield an expectation of a quantity that varies less than the original integrand over the region of integration. For example, let $h(X)$ be a density for the random variable X which takes values only in $\mathcal{A}$ so

that $\int_{x\in A} h(x)dx = 1$. Then $\frac{h(x)}{h(x)}$ is a "tricky way" to write 1, and so it follows that

$$\int_{x\in\mathcal{A}} g(x)dx = \int_{x\in\mathcal{A}} g(x)\frac{h(x)}{h(x)}dx = \int_{x\in\mathcal{A}} \frac{g(x)}{h(x)}h(x)dx = \mathbb{E}_h\left(\frac{g(X)}{h(X)}\right), \tag{1.8}$$

so long as $h(x) \neq 0$ for any $x \in \mathcal{A}$ for which $g(x) \neq 0$, and where $\mathbb{E}_h$ denotes the expectation with respect to the density h. This gives a Monte Carlo estimator:

$$\widetilde{g_n^h}(X) = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X^{(i)})}{h(X^{(i)})} \quad \text{where} \quad X^{(i)} \sim h(X). \tag{1.9}$$

Using (1.7) and the Cauchy-Schwarz inequality, it can be shown that $\mathrm{Var}(\widetilde{g_n^h}(X))$ is minimized when $h(x) \propto |g(x)|$ (see RUBINSTEIN 1981, p. 123). If we restrict our attention to what for most of our purposes is the relevant case,[2] that is, $g(x) \geq 0\ \forall x \in \mathcal{A}$, then it is immediately apparent that the choice of the density $h(x)$ which minimizes Monte Carlo variance is proportional to $g(x)$, *i.e.*, if $\alpha h(x) = g(x)$ where α is some constant of proportionality, then clearly we have $g(x)/h(x) = \alpha\ \forall\{x : h(x) > 0\}$ so $\mathbb{E}(g(X)/h(X)) = \alpha$ and hence the Monte Carlo variance would be zero by (1.7).

This seems wonderful—to obtain a Monte Carlo estimator with zero variance we could use (1.9), choosing our density h proportional to the function g. The absurdity of this wishful thinking is that the ability to simulate *independent* random variables from $h(x)$, or the ability to compute the density $h(x)$, itself, implies that the normalizing constant of the distribution is computable, which in turn would imply that the original integral involving $g(x)$ is computable and there would hence be no reason to do Monte Carlo at all! Ultimately, however, it makes clear that a good importance sampling function (as h is called) will be one that is as close as possible to being proportional to $g(x)$

In summary, a good importance sampling function $h(x)$ has the following properties:

1. $h(x) > 0$ whenever $g(x) \neq 0$, (this is required for (1.8) to hold)

2. $h(x)$ should be close to being proportional to $|g(x)|$

3. it should be easy to simulate values from $h(x)$

[2]This is typically the relevant case because we are interested in non-negative quantities like probabilities.

4. it should be easy to compute the density $h(x)$ for any value x that you might realize.

Fulfilling this wish-list in high dimensional space (where Monte Carlo techniques are most useful) is often a tall task—it is the main difficulty addressed in Chapter 2.

Note also that $g(x)$ is any arbitrary function, so it certainly includes the integrand of a standard expectation. For example, with $X \sim f_X$ we might be interested in $\mathbb{E}(r(X))$ for some function r so we could use

$$\mathbb{E}(r(X)) = \int r(x) f_X(x) dx = \int \frac{r(x) f_X(x)}{h(x)} h(x) = \mathbb{E}_h \left(\frac{r(x) f_X(x)}{h(x)} \right)$$

and approximate that by Monte Carlo, simulating values $x^{(1)}, \ldots, x^{(n)}$ from a distribution $h(x)$ that is close to proportional to $r(x) f_X(x)$.

Going back to the sum over latent variables problem often encountered in statistical genetics, importance sampling gives us a way to improve upon (1.5). From (1.4) it is clear that the optimal importance sampling function would be the conditional distribution of X given Y, *i.e.*,

$$P(Y = y) = \sum_{x \in \mathcal{X}} P(Y = y, X = x) = \sum_{x \in \mathcal{X}} \frac{P(Y = y, X = x)}{P(X|Y = y)} P(X|Y = y).$$

Note that the right side is a conditional expectation of a function of X. As before $P(X|Y)$ is not computable. So one must turn to finding some other distribution, say $P^*(X)$, that is close to $P(X|Y)$ but which is more easily sampled from and computed.

It is worth noting that, for making likelihood inference of a quantity, it is sufficient to be able to determine the likelihood function up to an unkown normalizing constant, *i.e.*, it suffices to calculate $\psi P(Y = y)$ where ψ is unknown. If the value of the normalizing constant ψ is not required, then the unnormalized probability $\psi P(Y = y)$ may be obtained by importance sampling using an importance sampling function $h(x)$ which is also known only up to a normalizing constant. Estimating likelihood ratios in this way has been described by THOMPSON and GUO (1991).

A common pitfall of importance sampling: As a final word on importance sampling, it should be pointed out that *the tails of the distributions matter*! While $h(x)$ might be

roughly the same shape as $g(x)$, serious difficulties arise if $h(x)$ gets small much faster than $g(x)$ out in the tails. In such a case, though it is improbable (by definition) that you will realize a value $x^{(i)}$ from the far tails of $h(x)$, if you do, then your Monte Carlo estimate will be very large—$g(x^{(i)})/h(x^{(i)})$ for such an improbable $x^{(i)}$ may be orders of magnitude larger than the typical values $g(x)/h(x)$ that you see. On the other hand, if no or few such values $x^{(i)}$ are realized from the far tails of $h(x)$, then the Monte Carlo estimate tends to be too small, which is equally undesirable. Such cases make importance sampling difficult, and underscore the importance of choosing a good importance sampling function $h(x)$.

1.5.3 Rao-Blackwellized estimators

Especially in the Bayesian analysis of complex stochastic models, one may be interested in approximating the marginal posterior probability distributions of many different quantities of interest. In performing Monte Carlo in such models, many different variables are simulated, and by using or combining those simulated variables in different ways, one can derive different Monte Carlo estimators for the same quantity. Some of these estimators will be preferable to others. A variance reduction technique named "Rao-Blackwellization" (LIU *et al.* 1994) provides a guideline for determing which of the possible Monte Carlo estimators for a quantity should be used. I describe Rao-Blackwellization here in the context in which it is employed throughout the dissertation—in the Monte Carlo estimation of probabilities.

Intuitively, the principle of Rao-Blackwellization can be understood as follows: let W be a random variable describing the probability that a variable, say λ, falls in a set $\mathcal{L}$. Then, assuming we could simulate values of W, we could estimate $P(\lambda \in \mathcal{L})$ by the Monte Carlo estimate, $\frac{1}{n}\sum_{i=1}^{n} w^{(i)}$. Each $w^{(i)}$ will be a value between zero and one. An alternative estimate could be proposed: $\frac{1}{n}\sum_{i=1}^{n} a^{(i)}$, where each $a^{(i)}$ takes the value 0 if a uniform real number on $(0,1)$, say $U^{(i)}$ is less than $w^{(i)}$, and the value 1 if $U^{(i)} > w^{(i)}$. Doing so would be naive, because some information is lost in condensing the real-valued $w^{(i)}$'s into the integer-valued $a^{(i)}$'s. Nonetheless, this sort of loss of information is routinely carried out by people doing complicated Monte Carlo studies. Rao-Blackwellization, in the limited context I will use it in the dissertation, is the name given to improving upon a Monte Carlo estimate of

the form $\frac{1}{n}\sum_{i=1}^{n} a^{(i)}$ by using, instead, a Monte Carlo estimate of the form $\frac{1}{n}\sum_{i=1}^{n} w^{(i)}$. The general context in which it arises in the following chapters is described below.

Let us suppose that we have a probability model with three variables (which, again, may be multidimensional), X, Y and λ, having the joint distribution $P(X, Y, \lambda)$. Suppose that the value of Y is known, and the quantity that we desire to know is $P(\lambda \in \mathcal{L}|Y)$, the conditional probability, given Y, that λ is in some set $\mathcal{L}$. This quantity may be written as

$$\begin{aligned} P(\lambda \in \mathcal{L}|Y) &= \sum_{x \in \mathcal{X}} P(\lambda \in \mathcal{L}, X = x|Y) \\ &= \sum_{x \in \mathcal{X}} P(\lambda \in \mathcal{L}|Y, X = x) P(X = x|Y) \\ &= \mathbb{E}[P(\lambda \in \mathcal{L}|Y, X)|Y]. \end{aligned} \tag{1.10}$$

Thus, a Monte Carlo estimator of $P(\lambda \in \mathcal{L}|Y)$ may be obtained as

$$P(\lambda \in \mathcal{L}|Y) \approx \frac{1}{n}\sum_{i=1}^{n} P(\lambda \in \mathcal{L}|Y, X^{(i)}) \tag{1.11}$$

where $X^{(i)}$ is simulated from $P(X|Y)$. This would, in fact, be the Rao-Blackwellized estimator for $P(\lambda|Y)$. Nonetheless, its use is sometimes overlooked, for the following reason: in many cases such as this, it is impossible to obtain samples $X^{(i)}$ "directly" from $P(X|Y)$. Rather, only samples $(X^{(i)}, \lambda^{(i)})$, $i = 1, \ldots, n$, are available, and it is thus tempting to apply (1.3) (rather than using (1.11)) to obtain the Monte Carlo estimator

$$P(\lambda \in \mathcal{L}|Y) \approx \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\{\lambda^{(i)} \in \mathcal{L}\}. \tag{1.12}$$

This is often a poor choice, however, because, for independent samples $(\lambda^{(i)}, X^{(i)}, Y^{(i)})$, the variance of (1.12) can be shown to always be greater than or equal to the variance of the Monte Carlo estimator of (1.11) (GELFAND and SMITH 1990). This may be proved using the Rao-Blackwell theorem (hence the name) applied to the case where $P(\lambda \in \mathcal{L}|Y)$ is the "parameter" to be estimated, (1.12) is the unbiased estimator based on an insufficient statistic for the parameter, and (1.11) is the Rao-Blackwellized version of (1.12).

When the samples $(\lambda^{(i)}, X^{(i)}, Y^{(i)})$ are dependent across i, as is the case with Markov chain Monte Carlo, the superiority of the estimator (1.11) is more difficult to establish.

LIU *et al.* (1994) prove the superiority of (1.11) for estimating expectations of functions of either λ or X, alone, with Markov chains having certain properties. However, they also present an example in which a Rao-Blackwellized estimator has higher variance than the simple estimator. Nonetheless, experience suggests that in most cases involving the MCMC estimation of probabilities, the Rao-Blackwellized estimator will perform better, especially as regards estimating probabilities in the tails of the distribution.

Quite often, the quantity $P(\lambda \in \mathcal{L}|Y, X^{(i)})$ must be computed anyway during the process of realizing a simulated pair of variables $(X^{(i)}, \lambda^{(i)})$. In such cases the variance reduction of (1.11) comes "for free." The practitioner of Monte Carlo should always be on the lookout for opportunities to Rao-Blackwellize Monte Carlo estimators.

1.5.4 Markov chain Monte Carlo—using dependent samples of X

The "Monte Carlo" part of Markov chain Monte Carlo is essentially identical to regular Monte Carlo (except assessing convergence, *etc.*). The main difference is that the elements of the Monte Carlo sample $(X^{(1)}, \ldots, X^{(n)})$ are not independent of one another. Rather they are sampled as states visited by an ergodic Markov chain. The main novelty is the need for some method to construct a Markov chain with the appropriate limit distribution. Almost all techniques for doing so are variants of the Metropolis-Hastings sampler (HASTINGS 1970). This method provides a way of constructing a time-reversible Markov chain with limit distribution π by satisfying, at each transition of the chain, the detailed balance condition implied by π. This may be achieved when π is known only up to a normalizing constant. Details of this may be found in numerous texts.

Chapter 2

IMPORTANCE SAMPLING AND MONTE CARLO LIKELIHOOD FOR N_E

2.1 Introduction

Reductions in population size can lead to inbreeding which increases the probability of population extinction in typically outbreeding species (FRANKHAM 1995). Reductions in population size also lead to a loss of genetic diversity which may reduce a population's ability to adapt to changing conditions (SOULÉ 1986). To predict the risk to a population from these types of genetic factors, biologists are often interested in knowing the effective population size, N_e. An effective size is defined by comparison to an ideal population model, the Wright-Fisher model. The Wright-Fisher model assumes discrete, non-overlapping generations of constant size, and it assumes that the gametes which unite to form adults in one generation are randomly sampled with replacement from the previous generation. The variance effective size of a natural population is the size of a Wright-Fisher population which would experience a comparable increase in variance of gene frequency over time. The inbreeding effective size is defined similarly, but is based on the increase in gene identity by descent over time.

It is possible to estimate the variance effective size from observed changes in allele frequencies in a population over time. Moment-based estimators using F-statistics have been developed for this purpose (KRIMBAS and TSAKAS 1971; NEI and TAJIMA 1981; POLLAK 1983; WAPLES 1989; JORDE and RYMAN 1995). Recently, WILLIAMSON and SLATKIN (1999) described a method to estimate N_e by the method of maximum likelihood. To find the maximum likelihood estimate $\widehat{N_e}$ of N_e, given allele frequencies observed in samples taken from a population at different times, one models the population underlying the samples as a Wright-Fisher population. $\widehat{N_e}$ is then the size of that underlying, ideal population for which the observed data are most probable. In simulation studies WILLIAMSON and

SLATKIN (1999) showed that the maximum likelihood estimator outperformed the moment-based estimators, and they also demonstrated how a likelihood approach may be extended to estimate paramaters in more complex population models.

This likelihood method has been restricted to data on diallelic loci, because, with data on multiallelic loci, evaluating the likelihood for N_e exactly is computationally intractable. In this chapter, I describe this problem as one of inference from a hidden Markov chain (BAUM *et al.* 1970), and describe an algorithm for importance sampling which makes it possible to compute the likelihood by Monte Carlo. Much of this work was pursued in collaboration with Dr. Ellen Williamson when she was a postdoctoral researcher in Montgomery Slatkin's group at University of California, Berkeley. It was previously presented in ANDERSON *et al.* (2000), and the treatment here follows very closely from that.

2.2 Formulation of the Model and Monte Carlo

2.2.1 The model

The data are random genetic samples collected at different generations. The first sample is collected at generation 0 and the last sample at generation T. Any samples drawn at intervening generations may be evenly or irregularly spaced in time. For notational simplicity, we assume for now that individuals are genotyped at a single locus, though we describe later the extension to multiple, independently-segregating loci. The data include K different allelic types, indexed by $k = 1, \ldots, K$. The allele frequencies observed in samples taken from different generations will differ due to genetic drift and sampling variation.

Let $\boldsymbol{Y}_t = (Y_{t,1}, \ldots, Y_{t,K})$ be the counts of the K different allelic types in the sample at generation t, and let S_t denote the number of diploid individuals in the sample. We assume that the samples were taken from a Wright-Fisher population of size N_e, and denote the unobserved population allele counts at generation t by $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,K})$, with $\sum_{k=1}^{K} X_{t,k} = 2N_e$. By the formulation of the Wright-Fisher model, the $\boldsymbol{X}_t$ form a Markov chain in time, with transitions defined by multinomial probabilities depending on N_e,

$$P_{N_e}(\boldsymbol{X}_t|\boldsymbol{X}_0, \ldots, \boldsymbol{X}_{t-1}) = P_{N_e}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) = (2N_e)! \prod_{k=1}^{K} \frac{[X_{t-1,k}/(2N_e)]^{X_{t,k}}}{X_{t,k}!}. \qquad (2.1)$$

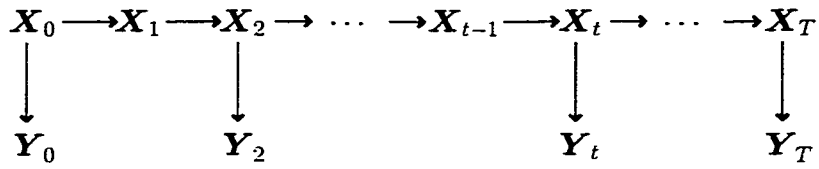


Figure 2.1: A graph showing the dependence between components of $\mathbf{X}$ and $\mathbf{Y}$. The $\boldsymbol{Y}_t$'s are observations of a hidden Markov chain. The graph shown represents a situation where samples were taken at generations 0, 2, t, and T, and no samples were taken at generations 1 and $t-1$.

The genetic sample at a time t is assumed to be drawn with replacement from the copies of alleles present in the population at time t, and sampled individuals are assumed to still be able to reproduce. This is equivalent to drawing the sample $\boldsymbol{Y}_t$ from a very large gamete pool produced by the population at time t: Sampling Plan I of WAPLES (1989). This type of sampling applies to many organisms, especially those species with high fecundity that may be sampled as juveniles (in which case the juveniles are assumed to carry a representative sample of alleles from the adults, and the number of juveniles is very large, so sampling without replacement from the juveniles is like sampling with replacement from the adults) or those that may be sampled as adults in populations having census sizes considerably larger than their effective sizes (WAPLES 1989). The sample allele counts $\boldsymbol{Y}_t$, given the latent variable $\boldsymbol{X}_t$, are conditionally independent of all the other variables and follow the multinomial distribution depending on the parameter N_e, the sample size S_t, and $\boldsymbol{X}_t$:

$$P_{N_e}(\boldsymbol{Y}_t|\boldsymbol{X}_t) = (2S_t)! \prod_{k=1}^{K} \frac{[X_{t,k}/(2N_e)]^{Y_{t,k}}}{Y_{t,k}!} \tag{2.2}$$

when $S_t > 0$. If there is no sample taken from the population at generation t, then $S_t \equiv 0$, and we define $P_{N_e}(\boldsymbol{Y}_t|\boldsymbol{X}_t) \equiv 1$.

Such a system forms a hidden Markov chain with the dependence structure shown in the directed graph of Figure 2.1. The allele counts in the population when the first sample is drawn, $\boldsymbol{X}_0$, are nuisance parameters. To avoid having to estimate the nuisance parameter $\boldsymbol{X}_0$ we consider an integrated likelihood by assuming a prior distribution, $P_{N_e}(\boldsymbol{X}_0)$, on $\boldsymbol{X}_0$ and integrating over that prior. We use a uniform prior on the set of components of $\boldsymbol{X}_0$

satisfying $\sum_{k=1}^{K} X_{0,k} = 2N_e$. It would be possible to use a different prior based on either theoretical considerations (WRIGHT 1938; WRIGHT 1952) or empirical evidence as to typical frequencies of alleles in different types of locus systems. However, in numerical results, the effect of different priors on the integrated likelihood are negligible (Ellen Williamson, University of California, Berkeley, pers. comm.) This is expected—so long as the allele count priors are relatively non-informative (as they should be) the information in the first sample, $\boldsymbol{Y}_0$, will always be much greater than the information in the prior; thus the influence of the prior is minimal.

The likelihood for N_e is the probability of the data $\mathbf{Y} = (\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_T)$ given the parameter N_e. The probability of $\mathbf{Y}$ is the sum of the joint probability of $\mathbf{Y}$ and the latent variables $\mathbf{X} = (\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T)$ over the space of all $\mathbf{X}$

$$\begin{aligned} P_{N_e}(\mathbf{Y}) &= \sum_{\mathbf{X}} P_{N_e}(\mathbf{Y}, \mathbf{X}) \qquad (2.3) \\ &= \sum_{\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T} \left(P_{N_e}(\boldsymbol{X}_0) P_{N_e}(\boldsymbol{Y}_0 | \boldsymbol{X}_0) \prod_{t=1}^{T} P_{N_e}(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}) P_{N_e}(\boldsymbol{Y}_t | \boldsymbol{X}_t) \right). \end{aligned}$$

For the case of $K = 2$ and N_e small the likelihood in (2.4) may be computed exactly. WILLIAMSON and SLATKIN (1999) effected the summation in (2.4) in terms of multiplication of transition probability matrices. The dimension of the square matrices is $(N_e - 1)!/[(N_e - K)!(K-1)!]$ which increases rapidly with N_e and K. The hidden Markov form of the system allows a more efficient direct computation of the likelihood using the algorithm of BAUM (1972). Nonetheless, exact evaluation for multiple alleles would still require prohibitively large amounts of computation and storage. An alternative is to estimate $P_{N_e}(\mathbf{Y})$ by Monte Carlo.

2.2.2 Monte Carlo evaluation

For likelihood inference, we must evaluate $P_{N_e}(\mathbf{Y})$ for a number of different values of N_e. Expressing this probability as an expectation with respect to the distribution of $\mathbf{X}$ gives

$$P_{N_e}(\mathbf{Y}) = \sum_{\mathbf{X}} P_{N_e}(\mathbf{Y}, \mathbf{X}) = \sum_{\mathbf{X}} P_{N_e}(\mathbf{Y}|\mathbf{X}) P_{N_e}(\mathbf{X}) = \mathbb{E}_{N_e}\Big(P_{N_e}(\mathbf{Y}|\mathbf{X}) \Big). \qquad (2.4)$$

In this form the expectation would be taken over the marginal probabilities of $\mathbf{X}$, and it could be estimated by Monte Carlo as

$$P_{N_e}(\mathbf{Y}) \approx \frac{1}{m}\sum_{i=1}^{m} P_{N_e}(\mathbf{Y}|\mathbf{X}^{(i)}) \tag{2.5}$$

for large m, with $\mathbf{X}^{(i)}$ being the i^{th} realization from the marginal distribution of $\mathbf{X}$. Such a naive scheme fails, however, because $P_{N_e}(\mathbf{Y}|\mathbf{X}^{(i)})$ varies greatly over the values of $\mathbf{X}$ realized from their marginal distribution, resulting in enormous Monte Carlo variance.

Instead, we pursue a more efficient Monte Carlo approximation by using importance sampling (Section 1.5.2). We express $P_{N_e}(\mathbf{Y})$ as an expectation with respect to a different distribution $P^*_{N_e}(\mathbf{X})$ such that $P^*_{N_e}(\mathbf{X}) > 0$ for all $\mathbf{X}$ such that $P_{N_e}(\mathbf{Y},\mathbf{X}) > 0$. Thus we have:

$$P_{N_e}(\mathbf{Y}) = \sum_{\mathbf{X}} \frac{P_{N_e}(\mathbf{Y},\mathbf{X})}{P^*_{N_e}(\mathbf{X})} P^*_{N_e}(\mathbf{X}) = \mathbb{E}^*_{N_e}\left(\frac{P_{N_e}(\mathbf{Y},\mathbf{X})}{P^*_{N_e}(\mathbf{X})}\right) \tag{2.6}$$

where $\mathbb{E}^*_{N_e}$ indicates that the expectation is over the space of $\mathbf{X}$ weighted by the distribution $P^*_{N_e}(\mathbf{X})$. The expectation (2.6) may be estimated by Monte Carlo, giving

$$P_{N_e}(\mathbf{Y}) \approx \tilde{P}_{N_e}(\mathbf{Y}) = \frac{1}{m}\sum_{i=1}^{m} \frac{P_{N_e}(\mathbf{Y},\mathbf{X}^{(i)})}{P^*_{N_e}(\mathbf{X}^{(i)})} \tag{2.7}$$

for large m where $\mathbf{X}^{(i)}$ is the i^{th} realization of $\mathbf{X}$ drawn from $P^*_{N_e}(\mathbf{X})$. The Monte Carlo variance of $\tilde{P}_{N_e}(\mathbf{Y})$ is made small when $P_{N_e}(\mathbf{Y},\mathbf{X})/P^*_{N_e}(\mathbf{X})$ varies little across the possible values of $\mathbf{X}$, and would be minimized if $P^*_{N_e}(\mathbf{X})$ were exactly proportional to $P_{N_e}(\mathbf{Y},\mathbf{X})$. Such a distribution of $\mathbf{X}$ would, by definition, be the conditional distribution $P_{N_e}(\mathbf{X}|\mathbf{Y})$. Unfortunately, for the same reasons that $P_{N_e}(\mathbf{Y})$ cannot be computed exactly, it is infeasible to compute $P_{N_e}(\mathbf{X}|\mathbf{Y})$. Nonetheless, the Monte Carlo variance of $\tilde{P}_{N_e}(\mathbf{Y})$ will be reduced to the extent that $P^*_{N_e}(\mathbf{X})$ resembles $P_{N_e}(\mathbf{X}|\mathbf{Y})$. The next subsection describes a method for rapid simulation of $\mathbf{X}^{(i)}$'s from a distribution $P^*_{N_e}(\mathbf{X})$ which is close to $P_{N_e}(\mathbf{X}|\mathbf{Y})$. As is required for the importance sampling, it is also possible to compute $P^*_{N_e}(\mathbf{X}^{(i)})$ quickly for each $\mathbf{X}^{(i)}$ generated.

2.2.3 Sampling from $P^*_{N_e}(\mathbf{X})$ by a forward-backward method

BAUM *et al.* (1970) describe computations applicable to hidden Markov chains that may be adapted for the purpose of efficiently realizing latent variables, such as $\mathbf{X} = (\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T)$,

from their exact conditional distribution given the observed variables, $\mathbf{Y} = (\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_T)$. This "forward-backward" algorithm first employs a forward step in which the conditional probability distributions of each $\boldsymbol{X}_t$, given the observed variables up to and including $\boldsymbol{Y}_t$, are recursively computed and stored using the relation

$$P(\boldsymbol{X}_t|\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t) \propto \sum_{\boldsymbol{X}_{t-1}} P(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_{t-1}) P(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) P(\boldsymbol{Y}_t|\boldsymbol{X}_t), \tag{2.8}$$

which is normalized by the sum of that quantity over all the values of $\mathbf{X}_t$. The last such conditional distribution computed is $P(\boldsymbol{X}_T|\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_T)$. The backward step begins with simulating a value $\boldsymbol{X}_T^{(i)}$ from this distribution (where, as before, the superscript (i) indicates a realized value of a random variable). One then proceeds backward, realizing $\boldsymbol{X}_{T-1}^{(i)}$ from its conditional distribution given all of the observed variables, $\mathbf{Y}$, and $\boldsymbol{X}_T^{(i)}$. In similar fashion, one realizes $\boldsymbol{X}_{T-2}^{(i)}$ and so forth back to $\boldsymbol{X}_0^{(i)}$. In this backward phase, each $\boldsymbol{X}_t^{(i)}$ is simulated from its conditional distribution given all the data $\mathbf{Y}$ and all of the components of $\mathbf{X}$ which have been realized so far. That is, $\boldsymbol{X}_t^{(i)}$ is drawn from

$$P(\boldsymbol{X}_t|\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_T, \boldsymbol{X}_{t+1}^{(i)}, \ldots, \boldsymbol{X}_T^{(i)}). \tag{2.9}$$

Because of the conditional independence implied by the hidden Markov chain structure, (2.9) reduces to $P(\boldsymbol{X}_t|\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t, \boldsymbol{X}_{t+1}^{(i)})$ which may be computed using the distributions stored during the forward step by the relation

$$P(\boldsymbol{X}_t|\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t, \boldsymbol{X}_{t+1}^{(i)}) \propto P(\boldsymbol{X}_t|\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t) P(\boldsymbol{X}_{t+1}^{(i)}|\boldsymbol{X}_t). \tag{2.10}$$

At the end of the backward step, it is clear that the resulting realization, $(\boldsymbol{X}_0^{(i)}, \ldots, \boldsymbol{X}_T^{(i)})$, is from the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$.

An approximation for multiple alleles: In our application, with multiple alleles at a locus, since there are so many possible states that each $\boldsymbol{X}_t$ may take, the above procedure is computationally infeasible. However, we make use of the BAUM *et al.* (1970) algorithm in spirit, employing two alterations to make it feasible to simulate from $P_{N_e}^*(\mathbf{X})$ and to compute its value. It should be noted that although the method described below involves a series of "approximations" by which $P_{N_e}^*(\mathbf{X})$ differs from $P_{N_e}(\mathbf{X}|\mathbf{Y})$, the final sampling and

computation of $P^*_{N_e}(\mathbf{X})$, is exactly from the distribution $P^*_{N_e}(\mathbf{X})$ as constructed, so its use in (2.7) gives a true Monte Carlo estimate.

The first approximation is to perform the forward-backward cycle separately for each allele. To describe this, we introduce some more notation. Denote by $\boldsymbol{X}_{(k)}$ the vector $(X_{0,k}, \ldots, X_{T,k})$ of latent counts of the k^{th} allele from time $t = 0$ to $t = T$. Similarly we define $\boldsymbol{Y}_{(k)} = (Y_{0,k}, \ldots, Y_{T,k})$. To do the forward-backward cycle separately over alleles we first focus on allele 1, simulating $\boldsymbol{X}^{(i)}_{(1)}$ by the forwards-backwards mechanism as if the data were on a diallelic locus with observed counts $\boldsymbol{Y}_{(1)}$ from samples of size $S_0, \ldots, S_T$ through time. Once we have realized $\boldsymbol{X}^{(i)}_{(1)}$ we update the sizes of the population and the sample. Thus we define the updated population size vector $2\boldsymbol{N}^*_{(2)} = (2N_e - X^{(i)}_{0,1}, \ldots, 2N_e - X^{(i)}_{T,1})$ and an updated sample size vector $2\boldsymbol{S}^*_{(2)} = (2S^*_{0,2}, \ldots, 2S^*_{T,2}) = (2S_0 - Y_{0,1}, \ldots, 2S_T - Y_{T,1})$, in effect removing the first allelic type from the remainder of the data and the population. We then use the forward-backward mechanism again to simulate $\boldsymbol{X}^{(i)}_{(2)}$, as though the data were counts $\boldsymbol{Y}_{(2)}$ from a diallelic locus drawn from a population with sizes that change over time, $\boldsymbol{N}^*_{(2)}$, and sample sizes $\boldsymbol{S}^*_{(2)}$. This continues sequentially over alleles updating population sizes and sample sizes as above: $2\boldsymbol{N}^*_{(k)} \leftarrow (2\boldsymbol{N}^*_{(k-1)} - \boldsymbol{X}^{(i)}_{(k-1)})$ and $2\boldsymbol{S}^*_{(k)} \leftarrow (2\boldsymbol{S}^*_{(k-1)} - \boldsymbol{Y}^{(i)}_{(k-1)})$, until $\boldsymbol{X}_{(K-1)}$ has been realized, which also determines that $\boldsymbol{X}_{(K)} \leftarrow (2\boldsymbol{N}^*_{(K-1)} - \boldsymbol{X}^{(i)}_{(K-1)})$. (Here and later the notation $A \leftarrow B$ means "the value B is assigned to the variable A.") At the end one has obtained a realized value $\mathbf{X}^{(i)}$ which may be used in (2.7).

$P^*_{N_e}(\mathbf{X})$ using a continuous approximation: Though realizing alleles sequentially, as above, greatly reduces the number of terms required to use (2.8) and (2.10), the method would still involve a prohibitive amount of summation over binomial probabilities. Thus, we construct $P^*_{N_e}(\mathbf{X})$ employing a normal approximation to binomial probabilities which replaces all such sums by analytically tractable integrals. Recall that if $W \sim \text{Binomial}(n, p)$, then the transformed variable $\sin^{-1}(W/n)^{1/2}$ is approximately normally distributed with variance $1/(4n)$. Notice that this quantity does not depend on p. We use this transformation to define the quantities $\phi_{t,k} = \sin^{-1}[Y_{t,k}/(2S^*_{t,k})]^{1/2}$ when $S^*_{t,k} > 0$, and $\theta_{t,k} = \sin^{-1}[X_{t,k}/(2N^*_{t,k})]^{1/2}$. By realizing the continuous values $\theta^{(i)}_{t,k}$ in a forward-backward frame-

work within a continuous setting, the computational demands are greatly reduced. And then, by transforming each $\theta_{t,k}^{(i)}$ back into the appropriate, discrete $X_{t,k}^{(i)}$ we have a way to realize $\mathbf{X}^{(i)}$ from $P_{N_e}^*(\mathbf{X})$ and to compute the probability $P_{N_e}^*(\mathbf{X}^{(i)})$. The details of this procedure are given in Section 2.3. We use it to compute the Monte Carlo estimate $\tilde{P}_{N_e}(\mathbf{Y})$ using (2.7).

2.2.4 Monte Carlo variance and multiple loci

The quantity $\tilde{P}_{N_e}(\mathbf{Y})$ is only an estimate of the true value $P_{N_e}(\mathbf{Y})$. By the Central Limit Theorem, for large m, $\tilde{P}_{N_e}(\mathbf{Y})$ will be approximately normally distributed (HAMMERSLEY and HANDSCOMB 1964) with mean $P_{N_e}(\mathbf{Y})$ and a variance which may be approximated without bias by the quantity

$$\widehat{\mathrm{Var}}(\tilde{P}_{N_e}(\mathbf{Y})) = \frac{1}{m(m-1)} \sum_{i=0}^{m} \left(\frac{P_{N_e}(\mathbf{Y}, \mathbf{X}^{(i)})}{P_{N_e}^*(\mathbf{X}^{(i)})} - \tilde{P}_{N_e}(\mathbf{Y}) \right)^2 . \tag{2.11}$$

These facts may be used to obtain a confidence interval estimate around $\tilde{P}_{N_e}(\mathbf{Y})$ for each value of N_e investigated.

The ability to estimate N_e with adequate precision requires data from many loci. The extension to data on J independently-segregating loci, indexed by $j = 1, \ldots, J$, is straightforward: each locus is treated separately, and the estimated likelihoods from each locus are multiplied together.

Thus, let $\tilde{P}_{N_e,j}(\mathbf{Y})$ be the Monte Carlo likelihood estimate from the data on the j^{th} locus. The Monte Carlo likelihood estimate using all the loci is then

$$\tilde{P}_{N_e}^J(\mathbf{Y}) = \prod_{j=1}^{J} \tilde{P}_{N_e,j}(\mathbf{Y}). \tag{2.12}$$

This provides a more efficient Monte Carlo estimator than another unbiased estimator for $\tilde{P}_{N_e}^J(\mathbf{Y})$ that one might consider:

$$\frac{1}{n} \sum_{i=1}^{n} \left(\prod_{j=1}^{J} \frac{P_{N_e,j}(\mathbf{Y}, \mathbf{X}^{(i)})}{P_j^*(\mathbf{X}^{(i)})} \right).$$

A proof of this is given in Appendix A. For (2.12) to hold, the initial allele counts must have independent prior distributions, $P_{N_e}(\boldsymbol{X}_0)$. Additionally, implicit in (2.12) is the assumption

that the loci used are in linkage equilibrium at $t = 0$, and they remain in equilibrium over the interval between samples. While even unlinked loci will exhibit random departures from linkage equilibrium due to finite population size (HILL 1981; BARTLEY *et al.* 1992), these random departures from linkage equilibrium should not greatly influence the accuracy of (2.12).[1] $\tilde{P}^J_{N_e}(\mathbf{Y})$ will also have an approximately normal distribution. An unbiased estimator for its Monte Carlo variance is

$$\widehat{\mathrm{Var}}(\tilde{P}^J_{N_e}(\mathbf{Y})) = \prod_{j=1}^{J}\Big(\tilde{P}_{N_e,j}(\mathbf{Y})\Big)^2 - \prod_{j=1}^{J}\Big([\tilde{P}_{N_e,j}(\mathbf{Y})]^2 - \widehat{\mathrm{Var}}(\tilde{P}_{N_e,j}(\mathbf{Y}))\Big). \tag{2.13}$$

which can be derived following the variance of a product of J independent random variables, W_j, $j = 1, \ldots J$:

$$\begin{aligned}
\mathrm{Var}(\textstyle\prod W_j) &= \mathbb{E}([\textstyle\prod W_j]^2) - [\mathbb{E}(\textstyle\prod W_j)]^2 \quad \text{(definition of variance)} \\
&= \mathbb{E}([\textstyle\prod W_j^2]) - [\mathbb{E}(\textstyle\prod W_j)]^2 \quad \text{(powers distribute over products)} \\
&= \textstyle\prod \mathbb{E}(W_j^2) - \prod[\mathbb{E}(W_j)]^2 \quad \text{(independence of the } W_j\text{)} \\
&= \textstyle\prod \mathbb{E}(W_j^2) - \prod\Big(\mathbb{E}(W_j^2) - \mathrm{Var}(W_j)\Big) \quad \text{(definition of variance).}
\end{aligned} \tag{2.14}$$

Denoting $\tilde{P}_{N_e,j}(\mathbf{Y})$ in (2.13) by W_j and taking the expectation gives the same result, verifying that the expression in (2.13) is unbiased for $\mathrm{Var}(\tilde{P}^J_{N_e}(\mathbf{Y}))$. This can be used to compute a confidence interval estimate around $\tilde{P}^J_{N_e}(\mathbf{Y})$.

When displaying the Monte Carlo likelihood curve it is preferable to plot the log-likelihood values, $\log \tilde{P}^J_{N_e}(\mathbf{Y})$, for different values of N_e. In this case, the endpoints of the confidence intervals may be similarly log-transformed.

2.3 *Details of Using $\theta_{t,k}$ and $\phi_{t,k}$ in a Continuous Setting*

The many details of the continuous approximation used to simulate from $P^*_{N_e}(\mathbf{X})$ are described in the following five sub-sections. In summary, the approximation works as follows:

[1] These random departures from linkage equilibrium have been used by HILL (1981) and BARTLEY *et al.* (1992) to estimate a population's effective size from the observed gametic disequilibrium in a single genetic sample. An interesting line of future research might be to derive a probability model that modeled both the temporal changes between samples and the gametic disequilibrium within each sample, and thus use both of those sources of information to estimate the effective size.

the process of allele frequency drift in a Wright-Fisher model is approximated by Brownian motion of the angularly transformed allele freqencies. Like the allele counts of $\mathbf{X}$, these transformed variates are related through time in a hidden Markov chain but with normally-distributed transition densities (due to the Brownian motion approximation). Therefore, the forward-backward method may be applied to these transformed variates, so that realizations of them may be made conditional on the data. This is described in Sections 2.3.1 and 2.3.2.

Of course, for the importance sampling, we require simulated values of $\mathbf{X}$, and not simulated values of a transformation of $\mathbf{X}$. Therefore, the angularly transformed, simulated variables must be transformed *back* into $\mathbf{X}$'s before they are useful. This is a difficult task because, in the process of back-transforming, one must be certain to consider the existence of all the alleles in the population, and because the Brownian motion, though unbounded, is an approximation to a stochastic process with boundaries. The procedure for back-transforming the simulated variables is described in Section 2.3.4.

Computing $P^*_{N_e}(\mathbf{X}^{(i)})$, after having simulated $\mathbf{X}^{(i)}$ by this method, requires that one consider the many possible values of the transformed, continuous variates that would have led to the same $\mathbf{X}^{(i)}$. The method for computing $P^*_{N_e}(\mathbf{X}^{(i)})$ is described in Sections 2.3.3 and 2.3.5.

We define the random variables $\phi_{t,k} = \sin^{-1}[Y_{t,k}/(2S^*_{t,k})]^{1/2}$ when $S^*_{t,k} > 0$, and $\theta_{t,k} = \sin^{-1}[X_{t,k}/(2N^*_{t,k})]^{1/2}$. These quantities have an approximate normal distribution which is independent of their means. We use them in our construction of the importance sampling function $P^*_{N_e}(\mathbf{X})$. Below we will concentrate on their use for realizing $\boldsymbol{X}^{(i)}_{(k)}$, keeping in mind that if $k > 1$ then we will have already realized $\boldsymbol{X}^{(i)}_{(k-1)}$, and will be using the updated population and sample sizes $N^*_{(k)}$ and $S^*_{(k)}$. If $k = 1$ then $N^*_{(1)}$ and $S^*_{(1)}$ are defined to be N_e and S, respectively.

2.3.1 The forward step

Following CAVALLI-SFORZA and EDWARDS (1967), if $\theta_{t-1,k}$ is normally ($\mathcal{N}$) distributed with mean μ_{t-1} and variance σ^2_{t-1} then, after a generation of genetic drift in a population

of $N^*_{t,k}$ diploids, $\theta_{t,k}$ has an approximate normal distributions with mean μ_{t-1} and variance $\sigma^2_t = \sigma^2_{t-1} + 1/(8N^*_{t,k})$. If there are data $Y_{t,k}$ from a sample of size $S_{t,k}$ at time t, then $\phi_{t,k}$ has an approximate normal distribution with mean $\theta_{t,k}$ and variance $1/(8S^*_{t,k})$, so, given that $\theta_{t,k} \sim \mathcal{N}(\mu_t, \sigma^2_t)$, the conditional distribution of $\theta_{t,k}$ given $\phi_{t,k}$ is also normal. These relations form the basis of a continuous approximation for doing the forward step. For the purpose of realizing $\mathbf{X}$ we assume that the uniform prior on $\boldsymbol{X}_0$ is equivalent to a diffuse prior on $\theta_{0,k}$. Therefore $\theta_{0,k}|\phi_{0,k} \sim \mathcal{N}(\mu_0, \sigma^2_0)$ with $\mu_0 = \phi_{0,k}$ and $\sigma^2_0 = 1/(8S^*_{0,k})$. With that as a starting point, we work iteratively forward in time assigning values

$$\mu_t \longleftarrow \mu_{t-1} \tag{2.15}$$

$$\sigma^2_t \longleftarrow \sigma^2_{t-1} + 1/(8N^*_{t,k}) \tag{2.16}$$

if $S^*_{t,k} = 0$. If $S^*_{t,k} > 0$, however, then one first computes μ_t and σ^2 as in (2.15) and (2.16), but then further updates the values to reflect the information in the sample at time t:

$$\mu_t \longleftarrow \frac{\mu_t/(8S^*_{t,k}) + \sigma^2_t\phi_{t,k}}{1/(8S^*_{t,k}) + \sigma^2_t} \tag{2.17}$$

$$\sigma^2_t \longleftarrow \frac{\sigma^2_t/(8S^*_{t,k})}{1/(8S^*_{t,k}) + \sigma^2_t}. \tag{2.18}$$

This is analogous to computing a posterior distribution from a normal prior and normal data (see, for example, GELMAN *et al.* 1996, p. 43).

Carrying this out until $t = T$ gives values for the mean and variance of $\theta_{T,k}$ given $\phi_{0,k}, \ldots, \phi_{T,k}$, assuming they follow a normal distribution. In fact, for each t, it gives us the parameters for the normal distribution of $\theta_{t,k}$ conditional on $\phi_{r,k}$ for all $r \leq t$. We are thus in a position to realize $\theta^{(i)}_{t,k}$'s in the backward step and transform those $\theta^{(i)}_{t,k}$'s back into the $X^{(i)}_{t,k}$'s that we need.

2.3.2 The backward step

The backward step is more complicated than the forward step, because after realizing each value of $\theta^{(i)}_{t,k}$ we must transform it into the discrete value $X^{(i)}_{t,k}$ that we require. This transformation process requires some extra bookkeeping to ensure that we do not waste time

realizing $\mathbf{X}^{(i)}$'s which are incompatible with the data. This bookkeeping is taken care of by the map $\mathcal{M}$ described in Section 2.3.4. We first realize the value $\theta^{(i)}_{T,k}$ from a $\mathcal{N}(\mu_T, \sigma^2_T)$ distribution. Then we transform that to the realization $X^{(i)}_{T,k}$ by a many-to-one map $\mathcal{M}$ which has two effects: the first is that of folding and translating the distribution of $\theta_{T,k}$ so that it is bounded between 0 and $\pi/2$, mapping $\theta^{(i)}_{T,k} \in (-\infty, \infty)$ to a value $\theta^*_{T,k} \in [0, \pi/2]$. The second is transforming that $\theta^*_{T,k}$ into the appropriate value $X^{(i)}_{T,k}$ (see Section 2.3.4).

Working backward, each $\theta^{(i)}_{t,k}$, for $t = T-1$ down to $t = 0$, is realized from a $\mathcal{N}(\mu_t, \sigma^2_t)$ distribution and then transformed into the corresponding $\theta^*_{t,k}$ and $X^{(i)}_{t,k}$ by $\mathcal{M}$. In keeping with (2.10), before any $\theta^{(i)}_{t,k}$ is realized, μ_t and σ^2_t must be appropriately updated, based on the values of μ_t and σ^2_t stored during the forward step and the realized value $\theta^{(i)}_{t+1,k}$. This involves making the assignments

$$\mu_t \longleftarrow \frac{\mu_t/(8N^*_{t+1}) + \sigma^2_t \theta^*_{t+1,k}}{1/(8N^*_{t+1}) + \sigma^2_t} \tag{2.19}$$

$$\sigma^2_t \longleftarrow \frac{\sigma^2_t/(8N^*_{t+1})}{1/(8N^*_{t+1}) + \sigma^2_t} \tag{2.20}$$

in the order as written.

2.3.3 Computing the probability $P^*_{N_e}(\mathbf{X}^{(i)})$

By carrying out the forward-backward steps above on the first allele, the realization $\boldsymbol{X}^{(i)}_{(1)}$ is obtained. Then, $\boldsymbol{N}^*_{(2)}$ and $\boldsymbol{S}^*_{(2)}$ are computed, and used in the forward-backward steps to obtain $\boldsymbol{X}^{(i)}_{(2)}$. Executing these steps for all the alleles yields the realization $\mathbf{X}^{(i)}$ which is used in (2.7). $P_{N_e}(\mathbf{Y}, \mathbf{X}^{(i)})$ in (2.7) is easily computed using the expansion shown between the large parentheses in (2.4).

It remains only to compute $P^*_{N_e}(\mathbf{X}^{(i)})$, which can be done by recording the probability of realizing each component $X^{(i)}_{t,k}$. Though this probability depends on the values of μ_t, σ^2_t, $N^*_{(k)}$ and several bookkeeping variables, we denote it here simply by $\mathcal{Q}(X^{(i)}_{t,k})$. (The function $\mathcal{Q}$ is described in Section 2.3.5). So long as the realization of $\boldsymbol{X}^{(i)}_{(k)}$ over alleles occurs in the same order over k ($k = 1, 2, \ldots, K$) for each i, then

$$P^*_{N_e}(\mathbf{X}^{(i)}) = \prod_{k=1}^{K} \prod_{t=0}^{T} \mathcal{Q}(X^{(i)}_{t,k}). \tag{2.21}$$

2.3.4 Details of $\mathcal{M}$

The fact that we are realizing $\boldsymbol{X}_{(k)}^{(i)}$'s one allele at a time requires that we do some extra bookeeping to keep our importance sampling scheme efficient. Primarily we must avoid realizing $\mathbf{X}^{(i)}$'s for which $P_{N_e}(\mathbf{Y}, \mathbf{X}^{(i)}) = 0$. Potential problems arise because by the method we use to realize values from $P_{N_e}^*(\mathbf{X})$, $X_{t,k}$ may only take values between 0 and $2N_{t,k}^*$, inclusive. If $2N_{t,k}^* = 0$ at any value of t, then for any $s > t$, $X_{s,k}^{(i)}$ must also be 0. To avoid situations in which this leads to $P_{N_e}(\mathbf{Y}, \mathbf{X}^{(i)})$ being zero (like when $X_{t,k}^{(i)} = 0$ and $Y_{t,k} > 0$) we introduce the following scheme and additional notation:

$$\begin{aligned}
\delta_{t,k} &= \begin{cases} 1 & \text{if } X_{t,k}^{(i)} = 0 \text{ implies } P_{N_e}(\mathbf{Y}, \mathbf{X}^{(i)}) = 0 \\ 0 & \text{otherwise} \end{cases} \\
\gamma_{t,k} &= \min_{r<t} 2N_{r,k}^* \\
\kappa_{t,k} &= \text{the number of allelic subscripts } \ell : k < \ell \leq K \text{ such that} \\
&\qquad Y_{r,\ell} > 0 \text{ for at least one } r \geq t
\end{aligned} \tag{2.22}$$

Knowing the above quantities, we can define the function $\mathcal{M}$. In the remainder of this section and in the following one we drop the t and k subscripts for clarity.

At its heart, $\mathcal{M}$ is a function that takes an angularly transformed allele frequency, θ, and transforms that, first to an allele frequency, and then to an allele count. The first difficulty comes from the bookkeeping described in the preceding paragraph. The second difficulty is that it is possible θ is not in the interval $[0, \pi/2]$. This is taken care of by converting a value θ outside of $[0, \pi/2]$ to a value θ^* which is in the interval $[0, \pi/2]$. This is done by reflecting and translating the value of θ until it is in the interval $[0, \pi/2]$, and it is done in such a way that, given the density for θ on $(-\infty, \infty)$, it is relatively easy to determine the probability that θ^* (the reflected and translated version) is within a given interval inside $[0, \pi/2]$. More precisely, this is described in the following: with N^* and γ positive integers, $\delta \in \{0, 1\}$, and $\kappa \in \{0, 1, \ldots, \min(2N^* - \delta, \gamma - \delta)\}$, let $\mathcal{M}(\theta; N^*, \delta, \gamma, \kappa) : \mathbb{R}^1 \to \{\delta, \ldots, \min(2N^* - \delta, \gamma - \delta)\} \times [0, \pi/2]$ be the many-to-one map that takes a realization of $\theta \in (-\infty, \infty)$ to the ordered pair (X, θ^*) where X is an integer such that $\delta \leq X \leq \min(2N^* - \delta, \gamma - \delta)$, and θ^* is a real

number between 0 and $\pi/2$, inclusive. $\mathcal{M}$ may be described by the following pseudocode. We first define the quantities $L = \sin^{-1}(.5/(2N^*))^{1/2}$ and

$$H = \begin{cases} \sin^{-1}[(2N^* - \kappa + .5)/(2N^*)]^{1/2} & , \ \kappa \geq 1 \\ \sin^{-1}[(2N^* - .5)/(2N^*)]^{1/2} & , \ \kappa = 0. \end{cases}$$

Then,

if $(\delta = 2N^* - \kappa$ **or** $\delta = \gamma - \kappa)$ **then** $\theta^* \longleftarrow 0$

else if $(L \leq \theta < H)$ **then** $\theta^* \longleftarrow \theta$

else if $(\theta < L)$

and if $(\delta = 0)$ **then** $\theta^* \longleftarrow \theta$

else if $(\delta = 1)$ **then** $\theta_{[L]} \longleftarrow 2L - \theta$ (*this is reflection around* $\theta = L$), and then

if $(L \leq \theta_{[L]} < H)$ **then** $\theta^* \longleftarrow \theta_{[L]}$

else we know $\theta_{[L]} \geq H$, and we consider the sequence $\theta_{[i]} = i(L - H) + \theta_{[L]}$, $i = 1, 2, \ldots$, and we assign $\theta^* \longleftarrow \theta_{[i^*]}$ where i^* is the least i such that $L \leq \theta_{[i]} < H$. (*The sequence* $\theta_{[i]}$ *represents successive translation leftward*).

else if $(\theta \geq H)$

and if $(\kappa = 0)$ **then** $\theta^* \longleftarrow \pi/2$

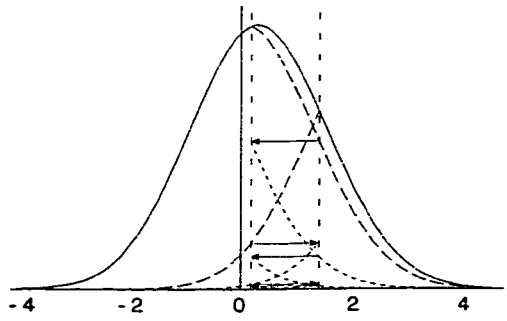
else if $(\kappa \geq 1)$ **then** $\theta_{[H]} \longleftarrow 2H - \theta$ (*this is reflection around* $\theta = H$), and then

if $(L \leq \theta_{[H]} < H)$ **then** $\theta^* \longleftarrow \theta_{[H]}$

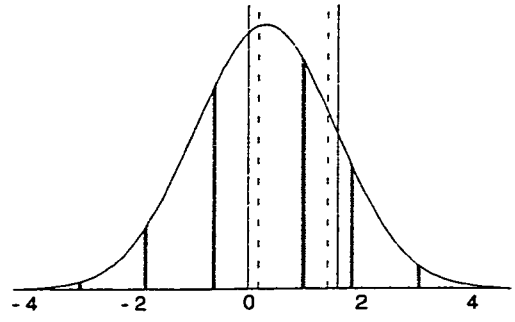
else we know $\theta_{[H]} < L$ and we consider the sequence $\theta_{[j]} = j(H - L) + \theta_{[H]}$, $j = 1, 2, \ldots$, and we assign $\theta^* \longleftarrow \theta_{[j^*]}$ where j^* is the least j such that $L \leq \theta_{[j]} < H$. (*The sequence* $\theta_{[j]}$ *represents successive translation rightward*).

finally we use θ^*, making the assignment $X \longleftarrow \lfloor 2N^* \sin^2 \theta^* + .5 \rfloor$

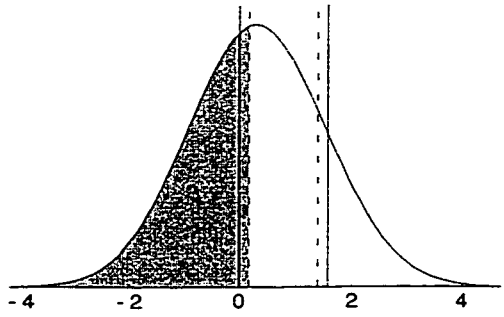
where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. The reflections and translations are depicted graphically in Figure 2.2(a).



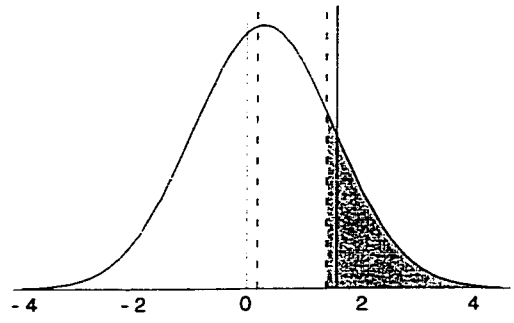
(a) Reflections and Translations



(b) $\delta = 1,\ \kappa = 1,\ X^{(i)} = 13$



(c) $\delta = 0,\ X^{(i)} = 0$



(d) $\kappa = 0,\ X^{(i)} = 2N^* = 20$

Figure 2.2: Figures representing $\mathcal{M}$ and $\mathcal{Q}$ for $N^* = 10$. The normal curve is the density for θ. (a) Reflections and translations as described in Section 2.3.4. Long-dashed lines represent the curve after reflection through L or H, while the short-dashed lines represent the reflected curve after one or more successive translations. (b) If $\min\{2N^* - \kappa, \gamma - \kappa\} > \delta$, $\delta = 1$, and $\kappa = 1$ then $X^{(i)}$ is constrained to be in $\{1 \ldots, 2N^* - 1\}$. The shaded regions correspond to those values of θ for which $X^{(i)} = 13$ by $\mathcal{M}$. The total shaded area is equal to $\mathcal{Q}_{\mu,\sigma^2}(X = 13; 10, 1, \gamma, 1)$. (c) If $\delta = 0$ then $X^{(i)}$ may take the value zero. The shaded area shows $\mathcal{Q}_{\mu,\sigma^2}(X = 0; 10, 0, \gamma, \kappa)$. (d) If $\kappa = 0$ then X may take the value $2N^*$. The shaded area shows $\mathcal{Q}_{\mu,\sigma^2}(X = 0; 10, \delta, \gamma, 0)$.

2.3.5 *The probability* $\mathcal{Q}_{\mu,\sigma^2}(X = X^{(i)}; N^*, \delta, \gamma, \kappa)$ *of realizing* $X = X^{(i)}$

If $\theta \sim \mathcal{N}(\mu, \sigma^2)$, and $(X, \theta^*) = \mathcal{M}(\theta; N^*, \delta, \gamma, \kappa)$, then we denote the marginal probability that $X = X^{(i)}$ by $\mathcal{Q}_{\mu,\sigma^2}(X = X^{(i)}; N^*, \delta, \gamma, \kappa)$. The value of $\mathcal{Q}_{\mu,\sigma^2}(X = X^{(i)}; N^*, \delta, \gamma, \kappa)$ can be expressed using the notation from the above section. First, $\mathcal{Q}_{\mu,\sigma^2}(X = X^{(i)}; N^*, \delta, \gamma, \kappa) = 0$ if $X^{(i)} < \delta$ or $X^{(i)} > 2N^* - \kappa$, though such values of $X^{(i)}$ should never occur from $\mathcal{M}$ anyway. Second, there are cases when $\mathcal{M}$ constrains $X^{(i)}$ to be either 0 or 1 with probability one. Hence if $\min\{2N^* - \kappa, \gamma - \kappa\} = \delta$ then $\mathcal{Q}_{\mu,\sigma^2}(X = \delta; N^*, \delta, \gamma, \kappa) = 1$. If, on the other hand, $\min\{2N^* - \kappa, \gamma - \kappa\} > \delta$, then for $X^{(i)} = 0$ and $X^{(i)} = 2N^*$ we have

$$\begin{aligned} \mathcal{Q}_{\mu,\sigma^2}(X = 0; N^*, 0, \gamma, \kappa) &= P(-\infty < \theta < L) \\ \mathcal{Q}_{\mu,\sigma^2}(X = 2N^*; N^*, \delta, \gamma, 0) &= P(H \le \theta < \infty) \end{aligned}$$

while for $0 < X^{(i)} < \min(2N^* - \kappa, \gamma - \kappa)$ we define $a = \sin^{-1}[(X^{(i)} - .5)/(2N^*)]^{1/2}$ and $b = \sin^{-1}[(X^{(i)} + .5)/(2N^*)]^{1/2}$, and have

$$\begin{aligned} \mathcal{Q}_{\mu,\sigma^2}(X = X^{(i)}; N^*, \delta, \gamma, \kappa) &= P(a \le \theta < b) \\ &\quad + \mathcal{I}\{\delta = 1\} P(a \le \theta_{[L]} < b) + \mathcal{I}\{\kappa > 0\} P(a \le \theta_{[H]} < b) \\ &\quad + \mathcal{I}\{\delta = 1\} \sum_{i=1}^{\infty} P(a \le \theta_{[i]} < b) + \mathcal{I}\{\kappa > 0\} \sum_{j=1}^{\infty} P(a \le \theta_{[j]} < b) \end{aligned} \tag{2.23}$$

where $\mathcal{I}\{\cdot\}$ is the indicator function (taking the value 1 if the statement in braces is true, and 0 otherwise) and $P(a \le \theta < b)$ is the probability that a $\mathcal{N}(\mu, \sigma^2)$ random variable is between a and b, namely $\int_a^b (2\pi\sigma^2)^{-1/2} \exp\{[-(\theta - \mu)^2]/(2\sigma^2)\} d\theta$. We compute this probability be numerical integration in our programs. In practice, the infinite sums are approximated by summing the first several terms of the series, until the contribution of the next term is very small (*e.g.*, $< .0000001$). Values of $\mathcal{Q}$ for different values of δ and κ appear as shaded regions in Figure 2.2(b–d).

This folding and translating might seem to be a very involved process, but it is computationally much faster than realizing θ from a truncated normal distribution and computing the probability of $X^{(i)}$ when θ is from such a distribution.

2.4 *Simulated and Real Datasets*

The method is demonstrated by computing log-likelihood curves for N_e from three different datasets. First, to verify that the method gives correct results we apply it to a simple simulated dataset (dataset 1) for which it is possible to compute the likelihood exactly. This dataset consists of simulated samples of 100 diploid organisms typed at 20 diallelic loci at generations 0, 6, and 12, sampled from a Wright-Fisher population of 25 diploid individuals. This sort of scenario, in which the samples include more individuals than the effective size of the population, would occur if juvenile samples of a highly fecund species were taken from a population of small effective size. For each locus, the initial allele frequency in the population at time zero was an independently drawn uniform real number between 0 and 1. The log-likelihood for N_e given these data was estimated for values between 10 and 52, in steps of 2, using $m = 20{,}000$ realizations of $\mathbf{X}$ from $P^*_{N_e}(\mathbf{X})$ for each locus and each N_e.

A second simulated dataset (dataset 2) shows how the method performs with multiallelic markers taken from a Wright-Fisher population. The dataset includes three samples of 100 diploids for 12 five-allele loci at generations 0, 4, and 8 from a population of 50 diploids. The allele frequencies at each locus in generation 0 for these simulations were independently drawn from a uniform Dirichlet density with five components. For this dataset, the log-likelihood was computed for values of N_e between 20 and 100 in increments of four using $m = 50{,}000$ realizations of $\mathbf{X}$ for each locus and each value of N_e.

Both of these datasets include only one simulated set of data. Because of the computational time required to compute a likelihood for each dataset having loci with multiple alleles, it is not possible to find the maximum likelihood estimate for N_e from a great number of replicate, simulated datasets. Therefore, the following analyses on the simulated datasets do not serve to assess the bias or the variance of the maximum likelihood estimator for N_e, but are meant solely to demonstrate that the Monte Carlo importance sampling method is capable of accurate approximation of the likelihood curve. An assessment of the bias and variance of the maximum likelihood estimator for N_e was carried out by WILLIAMSON and SLATKIN (1999) for data on diallelic loci.

Finally, the method has been applied to data on a population of *Drosophila* reported in

BEGON *et al.* (1980). These data were analyzed using F-statistics by BEGON *et al.* (1980) as well as by POLLAK (1983). They observed allele frequencies in three samples at each of nine enzyme loci. The first two samples were taken a little more than one year apart, and the third sample was taken some eight months later. Though the natural populations do not have discrete generations, they have been modeled previously by BEGON *et al.* and POLLAK as populations with discrete generations. Because of the different growth rates of flies during different seasons, seven generations separate the first two samples, while only two generations separate the second two samples (BEGON *et al.* 1980). The sample sizes for all loci were the same, with larger sample sizes taken in the later sampling periods. The sample sizes were $S_0 = 190$, $S_7 = 250$, and $S_9 = 335$ flies. POLLAK (1983) notes that since BEGON *et al.* (1980) sampled adult flies, their sampling scheme is closer to what is known in the literature as Sampling Scheme II than it is to Sampling Scheme I. However, as discussed by WAPLES (1989) the probability models underlying the two different sampling schemes are very similar when the actual size of the population is much larger than the effective size of the population. This is the case with these *Drosophila*. BEGON *et al.* (1980) report census sizes in the tens of thousands of flies, while the estimated N_e is orders of magnitude smaller. Because of this, it is still reasonable to analyze the data using the likelihood method developed here.

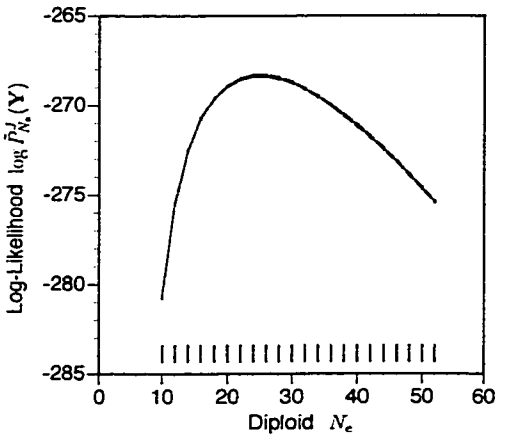
The data appear as allele frequencies in Table 1 of BEGON *et al.* (1980). Unfortunately the allele frequencies at the *Pgm* locus are misreported there and fail to sum to one. Thus only the remaining eight loci were used. Of these eight, three had three alleles, two had four alleles, two had five alleles and one had six alleles. We evaluated $\bar{P}^J_{N_e}(\mathbf{Y})$ at values of N_e between 200 and 1200 in increments of 50, with two more points ($N_e = 425$ and $N_e = 475$) included near the peak of the likelihood curve. For each point we used $m = 500{,}000$ realizations of $\mathbf{X}$.

2.5 Results

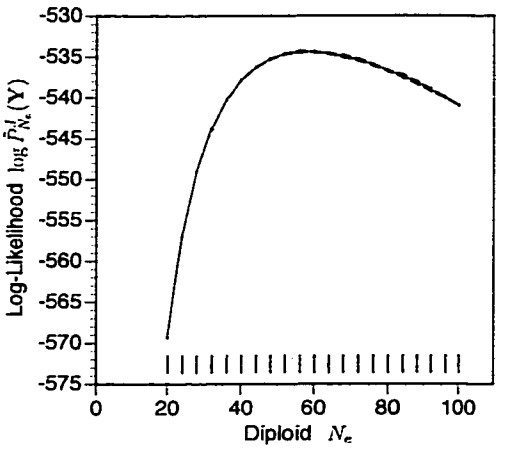
For each of the datasets, we were able to use our importance-sampling method to compute a log-likelihood curve. Using a program written in C, the runs for datasets 1 and 2 each

took about 10 hours on a laptop computer with a 266 Mhz G3 (Macintosh) processor. The log-likelihood curves from datasets 1 and 2 appear as solid lines in Figure 2.3. The estimated 90% confidence intervals around each value of $\log \tilde{P}^J_{N_e}(\mathbf{Y})$ appear as two dashed lines bordering the log-likelihood curve. Despite the fact that few Monte Carlo replicates (m = 20,000 and 50,000) were used, the Monte Carlo variance is minimal, as indicated by the fact that the dotted lines practically lie on top of the estimated log-likelihood curve. In both cases, the true values of N_e (25 and 50, respectively) are well within two units of log-likelihood from the maximum likelihood estimates which may be read from the graph as 24 and 56. Since Dataset 1 consists only of diallelic loci, it is possible to compute the exact log-likelihood curve. This exact curve has been plotted as a dotted line in Figure 2.3(a). It is impossible to distinguish the exact curve because the Monte Carlo estimate is very accurate in this case.

The log-likelihood curve computed for the data of BEGON *et al.* (1980) is shown in Figure 2.4. It took about 54 hours on a desktop computer with a 450 Mhz G4 (Macintosh) processor to produce the results. As before, the 90% confidence intervals around the Monte Carlo estimates appear as dotted lines. With this dataset, even with m = 500,000 realizations of $\mathbf{X}$, the Monte Carlo variance is not negligible. It is, however, small enough that reliable inferences may be made from the log-likelihood curve. The maximum likelihood estimate of N_e is 500. Using the values of N_e at which the log-likelihood has decreased two units from its maximum gives an estimate of a 95% confidence interval for the true N_e. These points are 250 and 975. By contrast, POLLAK (1983), using an F-statistic method, estimated N_e to be 251 with a standard error of 115. Recomputing Pollak's estimator, excluding the *Pgm* locus (as done in the likelihood analysis), gives the F-statistic estimate of 268 for N_e. The discrepancy between the maximum likelihood estimate and the F-statistic estimate is discussed in the next section. The present results are not comparable to the N_e estimated by BEGON *et al.* (1980) because, at the time, those authors were unable to make a single estimate of N_e using the samples at all three time points.



(a) Dataset 1



(b) Dataset 2

Figure 2.3: Log-likelihood curves estimated by Monte Carlo from datasets 1 and 2. The values of N_e at which the likelihood was computed are indicated by vertical lines above the horizontal axis in each figure. The log-likelihood values are connected by a solid line. Vertical bars intersecting the solid line indicate 90% confidence intervals around $\log \tilde{P}^J_{N_e}(\mathbf{Y})$ computed using the Monte Carlo variance estimate (2.13). The endpoints of the confidence intervals are connected by dashed lines. These features are difficult to see because the confidence intervals around the Monte Carlo estimate of the log-likelihoods are very small. (In other words, the Monte Carlo estimate of the log-likelihood is very good for these simulated datasets.) (a) Dataset 1 is simulated data from 20 diallelic loci. (b) Dataset 2 is simulated data from 12 loci with five alleles each.

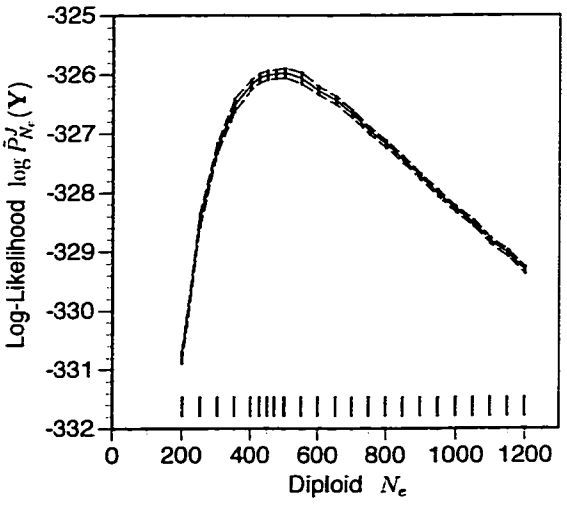


Figure 2.4: Log-likelihood curve from the data of BEGON *et al.* (1980) estimated by Monte Carlo. The format of the plot is as for Figure 2.3.

2.6 Discussion

As discussed in WILLIAMSON and SLATKIN (1999), the maximum likelihood estimator of N_e is less biased and has lower variance than the F-statistic estimator of N_e. In addition, WILLIAMSON and SLATKIN (1999) show that formulating the problem in a likelihood framework allows them to use explicit stochastic models for growing or shrinking populations. Until now, it was impractical to compute the likelihood for N_e using all the data when loci with more than two alleles were available. While it has been suggested that one may bin low-frequency alleles together to turn multiallelic loci into apparently diallelic loci and then apply exact likelihood calculation methods to such reduced data, this invariably throws away some information. Furthermore, different binning strategies lead to different results. Allowing full use of the data, the Monte Carlo likelihood procedure described here is a preferable way to analyze temporal data on multiallelic loci. The method is suitable for multiallelic loci such as the microsatellite markers becoming available in a wide variety of species.

Monte Carlo methods use realizations of random variables to estimate an expectation

by a sample average. There are a number of ways one can express the likelihood of N_e as an expectation, and then estimate it by Monte Carlo, but few of those schemes will be successful, because most will have high Monte Carlo variance. We attempted several different schemes before settling on the importance sampling method presented here. Although these less sophisticated Monte Carlo estimators produced reasonable estimates in very small problems, when applied to data involving loci with many alleles these methods failed to converge reliably, even after many days of computation (unpublished data).

The importance sampling method is successful because the importance sampling function, $P^*_{N_e}(\mathbf{X})$, closely resembles $P_{N_e}(\mathbf{X}|\mathbf{Y})$, the conditional probability of $\mathbf{X}$ given $\mathbf{Y}$. This is achieved by recognizing the hidden Markov chain structure of the problem and using the forward-backward algorithm of BAUM *et al.* (1970). Doing so gives a Monte Carlo estimator with demonstrably small Monte Carlo variance. Though the computational demands of this procedure are non-trivial, the reduction in Monte Carlo variance obtained makes it worthwhile. Nonetheless, it may be possible to improve the estimates by making additional changes to $P^*_{N_e}(\mathbf{X})$ so that it more closely resembles $P_{N_e}(\mathbf{X}, \mathbf{Y})$, especially in the tails of the distribution. This would further reduce the Monte Carlo variance.

It should be pointed out that while many Monte Carlo problems involving high dimensional random variables like $\mathbf{X}$ make use of Markov Chain Monte Carlo (MCMC) methods, the present method does not. In MCMC, successive realizations are correlated. Yet in this method each $\mathbf{X}^{(i)}$ realized from the distribution $P^*_{N_e}(\mathbf{X})$ is independent of all the other realized values. In subsequent chapters of the dissertation, however, the nature of the problems becomes more complex, and developing a similar importance sampling scheme would be very difficult and complex—even moreso than here.

It is interesting that the maximum likelihood estimate differs so much from the estimate given by POLLAK (1983) for the same data. There are differences between the two estimation methods that must account for the discrepancy. The most notable differences occur when combining information from multiple samples in time. Consider the fact that a better estimate of N_e may be made from two samples taken many generations apart than from two samples separated by fewer generations. Likewise two large samples will yield a better estimate than two small samples. When there are many samples, the relative infor-

mation content in different inter-sample intervals will depend on the relative sample sizes and the number of generations between the samples. By its nature, the maximum likelihood approach will appropriately weight information from different intervals. In contrast, POLLAK's F-statistic, F_{K_r}, neither includes terms for sample size nor interval length between samples, and $\hat{N}_{K_r}$, his estimate of N_e based on F_{K_r}, includes a term only for the number of generations between the first and the last sample and is invariant to permutations of the sample sizes at different times. Since the data from BEGON *et al.* (1980) span sampling intervals of different lengths and include different sample sizes at different times, differences between our results and those in POLLAK (1983) are not unexpected.

The Monte Carlo variance of our estimate of the likelihood given the data from a natural population of *Drosophila* was higher than the variance associated with our estimates from simulated data. Although a good estimate was achieved after sufficient computation, it may still be that data generated under a model that differs from the Wright-Fisher model present difficulty for the Monte Carlo likelihood method. For example, it may be that the effective size of the natural *Drosophila* population was different during the two different sampling intervals.

If desired, one could extend the likelihood framework to allow for N_e changing over time. For example, if the estimated census size of the population were available and was known to change over time it would be more sensible to estimate directly the ratio, λ, of the effective size of the population to the census size of the population. This ratio would be more useful for the purposes of modeling genetic change in the population than a single estimate of N_e over the entire time period between the first and last samples. The forward-backward approach implemented here could easily be modified to accommodate estimating this parameter, λ. It would also be possible to extend the approach here to estimate likelihoods from explicit stochastic models of populations of organisms with more complex life-histories; for example, overlapping generations or age-structured populations. However, these topics are taken up using MCMC in the following two chapters, where we will see that it is also preferable to propose a different model for genetic transmission from one generation to the next.

2.7 *Extensions and Caveats*

Since the time this importance sampling method was first developed, I have experimented with two refinements. The first involves a logistic approximation to the normal density. In this modified version, each $\theta_{t,k}$ is simulated from a logistic distribution (rather than a normal distribution) with the appropriate mean and variance. Since the cumulative distribution function of a logistic random variable is available in closed form, it is no longer necessary to perform the numerical integration required for handling the normal density in (2.23). On the BEGON *et al.* (1980) dataset, this results in a four-fold decrease in running time per iteration of the algorithm; however, the logistic distribution is a poorer approximation to the binomial distribution, and roughly four times as many replicates are required to achieve the same Monte Carlo variance as achieved with the normal distribution. It appears that little is to be gained by using the logistic approximation.

A second refinement I have experimented with entails alleles which appear in the early samples, but not in the later ones. Because of the way the forward step is carried out, the probability of realizing no copies of such an allele at time T, or any time before, will never exceed 1/2. It is therefore improbable to realize an $\mathbf{X}^{(i)}$ in which such an allele vanishes at an early time from the population. It is not improbable, however, that such an allele would be lost quickly from the population under the Wright-Fisher model. This is precisely the situation in which importance sampling may encounter problems—the realized value $\mathbf{X}^{(i)}$ is improbable under $P^*_{N_e}$, but the value of $P_{N_e}(\mathbf{Y}, \mathbf{X}^{(i)})$ is not commensurately tiny. It is possible to modify the forward step so it treats alleles which are lost from later samples more sensibly.

While the above suggestions might improve the importance sampling function to some degree, after further investigations with the original method I have some general reservations about the technique presented here. Particularly troubling is the fact that the probability of realizing a particular $\mathbf{X}^{(i)}$ depends on the order of the different alleles (*i.e.*, the order of the allele labels $k = 1, \ldots, K$. Since only a single ordering of the alleles is used here, it may be the case that the importance sampling distribution, $P^*_{N_e}(\mathbf{X})$ is not as close an approximation to $P_{N_e}(\mathbf{X}|\mathbf{Y})$ as we would like. This unfortunately, is likely to be a greater

problem in loci with more alleles, because there are more possible ways to order all those alleles. Fortunately, the Markov chain Monte Carlo methods of the next two chapters provide an alternative to the importance sampling method.

Chapter 3

λ AND A PÓLYA URN MODEL FOR GENETIC INHERITANCE

3.1 Introduction

In the previous chapter, we saw the Wright-Fisher model for genetic inheritance in a randomly-mating population. This is one of the simplest stochastic models possible for the transmission of genes between generations without a specified pedigree structure. Under the Wright-Fisher model, the marginal distribution of the number of offspring of an individual is binomial. Its use in determining the effective size of a natural population involves a considerable degree of abstraction—the many complicated interactions, stochastic events, and fitness considerations that lead to extra-binomial variance in family size or a certain rate of increase in inbreeding are presumed to be adequately accounted for by considering that the natural population can be modeled as a Wright-Fisher population of a particular size.

Explicitly modeling the many factors influencing inter-generation genetic dynamics in a natural population would be very challenging. Fortunately, for many natural populations, the Wright-Fisher population of size N_e provides an admirable approximation. However, when data are available on the census size, C, of a population, biologists are often interested in estimating the ratio of effective size to census size, N_e/C, which we shall denote λ. Pursuing likelihood or Bayesian inference for λ while strictly adhering to the Wright-Fisher model requires a somewhat inelegant interpretation of population size and leads to difficulties in applying Markov chain Monte Carlo methods to the problem. Namely, the probability of the unobserved allele counts in the population is not well-defined for different values of λ under the Wright-Fisher model. For this reason, I propose a new model for genetic inheritance based on an urn sampling scheme with stochastic replacements. Given the census size of a breeding population, this model has a single parameter, s, which can change the

effective size of the population, without altering the number of individuals in that population. This provides several advantages in the inference problems pursued here, simplifying the interpretation of the modeled process and the computation necessary for estimating λ. Results from this model, however, may be easily translated back to an interpretation from within the familiar Wright-Fisher perspective.

In this chapter I will briefly describe the Wright-Fisher based model for estimating λ, and a single-site-updating scheme for a Markov chain Monte Carlo (MCMC) algorithm suitable for approximating likelihood ratios or posterior probabilities for λ. I will then illustrate the shortcomings of the Wright-Fisher model in this context. This motivates the development in Section 3.3 of the urn model. I devote Section 3.4 to alternative interpretations and developments of the urn model which provide a wider range of biological interpretations for it. In Section 3.5, I derive expressions for the inbreeding and variance effective sizes of the urn-model population, investigate probabilities of allele fixation in the urn model, and show the relationship between census sizes C_t through time, the parameter of the urn model, s, and the parameter λ. In Section 3.6, I argue that in terms of allele fixation probabililities the Wright-Fisher model corresponds to possibly unrealistic assumptions about the distribution of family sizes. Finally, I briefly describe the use of the urn model to develop a simple MCMC scheme for Bayesian inference of λ. I apply this to the data of BEGON *et al.* (1980) from the preceding chapter. Elaborations on this MCMC method involving different sampling times and schemes, and different life-histories, are presented in the following chapter.

3.2 Estimating the Ratio of Effective to Census Population Size

In this section, I consider the scenario in which census sizes of breeding adults have been recorded each generation over some time period in a semelparous population with discrete generations and genetic samples are taken with replacement at some intervals of time either from those breeding adults or from the juveniles descended from them (other sampling models will be entertained in the following chapter). Much of this treatment is adapted from ANDERSON and THOMPSON (1999).

At generation t, C_t diploid individuals reproduce, giving rise to C_{t+1} individuals at

generation $t+1$. We take genetic samples of size $S_0, \ldots, S_T$ (assume $S_0 > 0$, $S_T > 0$) individuals, and find counts of the K different allelic types at a locus, $\mathbf{Y} = (\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_T)$ where $\boldsymbol{Y}_t = (Y_{t,1}, \ldots, Y_{t,K})$. For estimating λ the population at t is modelled as $\lfloor \lambda C_t \rfloor$ ideally-reproducing (*i.e.*, via a Wright-Fisher inheritance model) adults, where $\lfloor x \rfloor$ is the largest integer $\leq x$. Underlying the data are latent allele counts $\mathbf{X} = (\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T)$ with $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,K})$. $\{\boldsymbol{X}_t,\ t \geq 0\}$ is a first-order Markov chain with transition probabilities $P_\lambda(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t)$ being multinomial with cell probabilities $\boldsymbol{X}_t/\lfloor 2\lambda C_t \rfloor$ and number of trials $\lfloor 2\lambda C_{t+1} \rfloor$. The genetic data at time t are assumed to be samples from the gamete pool produced by the $\lfloor \lambda C_t \rfloor$ adults. This assumption is equivalent to sampling adults with replacement, and closely approximates the sampling of juveniles, when the number of juveniles is large. Hence, for $S_t > 0$, $P_\lambda(\boldsymbol{Y}_t|\mathbf{X}) = P_\lambda(\boldsymbol{Y}_t|\boldsymbol{X}_t)$ is multinomial with parameters $\boldsymbol{X}_t/\lfloor 2\lambda C_t \rfloor$ and $2S_t$. For $S_t = 0$, $P_\lambda(\boldsymbol{Y}_t|\boldsymbol{X}_t) \equiv 1$. Summing out the nuisance parameters $\boldsymbol{X}_0$ over a uniform prior $P_\lambda(\boldsymbol{X}_0)$ gives the likelihood

$$\begin{aligned} L(\lambda) = P_\lambda(\mathbf{Y}) &= \sum_{\mathbf{X}} P_\lambda(\mathbf{Y}, \mathbf{X}) \qquad (3.1) \\ &= \sum_{\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T} P_\lambda(\boldsymbol{X}_0) P_\lambda(\boldsymbol{Y}_0|\boldsymbol{X}_0) \prod_{t=1}^{T} P_\lambda(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) P_\lambda(\boldsymbol{Y}_t|\boldsymbol{X}_t). \end{aligned}$$

With $K = 2$, the sum over $\mathbf{X}$ may be evaluated exactly. With larger K, however, the huge space of possible $\boldsymbol{X}_t$'s makes this infeasible.

3.2.1 MCMC likelihood for λ

To obtain an efficient Monte Carlo estimate of $L(\lambda)$, one may consider the likelihood ratios $L(\lambda)/L(\lambda_0)$, (THOMPSON and GUO 1991; GEYER and THOMPSON 1992)

$$\frac{L(\lambda)}{L(\lambda_0)} = \frac{P_\lambda(\mathbf{Y})}{P_{\lambda_0}(\mathbf{Y})} = \sum_{\mathbf{X}} \frac{P_\lambda(\mathbf{Y}, \mathbf{X})}{P_{\lambda_0}(\mathbf{Y}, \mathbf{X})} P_{\lambda_0}(\mathbf{Y}|\mathbf{X}) = E_{\lambda_0}\left(\frac{P_\lambda(\mathbf{Y}, \mathbf{X})}{P_{\lambda_0}(\mathbf{Y}, \mathbf{X})} \middle| \mathbf{Y} \right) \qquad (3.2)$$

which may be estimated by $\frac{1}{m}\sum_{i=1}^{m} P_\lambda(\mathbf{Y}, \mathbf{X}^{(i)})/P_{\lambda_0}(\mathbf{Y}, \mathbf{X}^{(i)})$ where each $\mathbf{X}^{(i)}$ is realized from $P_{\lambda_0}(\mathbf{X}|\mathbf{Y})$. This is an efficient Monte Carlo estimator of the likelihood ratio provided λ is near to λ_0. Independent samples of $\mathbf{X}$ are not available because $P_{\lambda_0}(\mathbf{X}|\mathbf{Y})$ is known only up to scale. Instead, values of $\mathbf{X}^{(i)}$ can be realized from a Markov chain with

limit distribution $P_{\lambda_0}(\mathbf{X}|\mathbf{Y})$ using a component-wise Metropolis-Hastings algorithm (Hastings 1970): Start with initial values of $\mathbf{X}$; Select a pair $(X_{t,k}, X_{t,\ell})$, $k \neq \ell$ at random from $\mathbf{X}$; Propose updating the pair to $(X^*_{t,k}, X^*_{t,\ell}) = (X_{t,k} - w, X_{t,\ell} + w)$, where w is a random integer drawn with probability $q(w|X_{t,k}, X_{t,\ell})$; accept the proposal with probability $\min\{1, [q(-w|X^*_{t.k}, X^*_{t,\ell})P_{\lambda_0}(\mathbf{Y}, \mathbf{X}^*)]/[q(w|X_{t,k}, X_{t,\ell})P_{\lambda_0}(\mathbf{Y}, \mathbf{X})]\}$. After initial updates for burn-in, $\mathbf{X}^{(i)}$'s are sampled as the state of $\mathbf{X}$ at a spacing of u updates.

When estimating a curve for $L(\lambda)$, the range of λ's of interest may be large. In such a case it does not suffice to realize $\mathbf{X}^{(i)}$'s under a single λ_0. Instead, one must sample from several chains, each indexed by a different λ_0, $\lambda_0 \in \Lambda$. GEYER (1994) describes a reverse logistic regression method for reweighting the samples from each chain and estimating the whole likelihood surface.

3.2.2 A Bayesian approach

The MCMC sampler described above is useful for Bayesian inference of λ as well. Suppose that we are interested in computing the posterior probabilities for each $\lambda \in \Lambda$. Let $P(\lambda)$ be the prior distribution for λ assigning probability mass to each of the discrete points $\lambda \in \Lambda$. We shall assume for brevity and simplicity that $\sum_{\lambda\in\Lambda} P(\lambda) = 1$. It is possible to also propose changes to λ in the MCMC sampler, and thereby sample from the posterior distribution of λ by the following scheme: given the current state of the chain, $(\mathbf{X}, \lambda)$, propose a new value λ^* with probability $h(\lambda^*|\lambda)$ from the proposal distribution $h(\cdot|\lambda)$. Then, accept the proposal with probability

$$\min\left\{1, \frac{h(\lambda|\lambda^*)}{h(\lambda^*|\lambda)}\frac{P(\lambda^*)}{P(\lambda)}\frac{P_{\lambda^*}(\mathbf{Y}, \mathbf{X})}{P_{\lambda}(\mathbf{Y}, \mathbf{X})}\right\}$$

The successive values $\lambda^{(i)}$ generated in this way are a dependent sample from $P(\lambda|\mathbf{Y})$ and may accordingly be used to estimate that posterior distribution.

3.2.3 Shortcomings of the Wright-Fisher model in this case

The first difficulty in using the Wright-Fisher model to formulate the likelihood given in (3.1) is the dependence of the sampling terms $P_\lambda(\boldsymbol{Y}_t|\boldsymbol{X}_t)$ on λ. The parameter λ is supposed to affect the transitions $P_\lambda(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t)$, so it is inelegant to also have the dependence appear in

$P_\lambda(\boldsymbol{Y}_t|\boldsymbol{X}_t)$. This dependence results from the fact that a given allele count $X_{t,k}$ corresponds to a different allele *frequency* depending on the value of λ. Since the likelihood is the sum over all $\mathbf{X}$, this dependence has little real effect, but its presence is enough to cause people to take pause[1].

The undesireable effect of this dependence becomes even more clear when one considers the possibility of sampling adults without replacement *before* reproduction—Sampling Plan II of NEI and TAJIMA (1981), also described in WAPLES (1989). Under such sampling, in the context of a population with very high variance in family size, it is possible that the effective size estimated for a population could be smaller than the number of adults sampled without replacement from the population modeled as a Wright-Fisher population. While this presents no problem conceptually—it is quite possible that the effective size of a population may be smaller than the number of individuals sampled from it—it exposes how difficult it would be to correctly model Sampling Scheme II within the MCMC scheme of the previous chapter. Trying to explicitly model sampling of adults without replacement in the above model becomes difficult because $\boldsymbol{X}_t$ is affected by the sampling process, and the degree to which it is affected depends not on $\lfloor \lambda C_t \rfloor$ but on C_t itself. Suffice it to say that pursuing the explicit modeling of Sampling Scheme II within the context of the Wright-Fisher model of genetic inheritance, while possible, would not be straightforward.

A further shortcoming of the Wright-Fisher model is even more practically problematic—different values of λ imply different sizes of the state space of the latent variable $\mathbf{X}$, and this makes the computation of $L(\lambda)/L(\lambda_0)$ or of $P(\lambda|\mathbf{Y})$ more difficult and less efficient. Consider first the computation of $L(\lambda)/L(\lambda_0)$ using the identity (3.2). This identity only holds when the support of $\mathbf{X}$ under λ_0 is equal to or contains the support of $\mathbf{X}$ under λ. This only transpires when $\lambda_0 > \lambda$, since if $\lambda > \lambda_0$, it is possible that a value of $\mathbf{X}$ for which $P_\lambda(\mathbf{Y}, \mathbf{X}) > 0$ might include a t for which $\sum_{k=1}^{K-1} X_{t,k} > \lfloor 2\lambda_0 C_t \rfloor$, and hence $P_{\lambda_0}(\mathbf{Y}, \mathbf{X})$ would be zero. As a consequence, the importance sampling algorithm implied by (3.2) can be employed only for $\lambda_0 > \lambda$. This is inefficient because it is not possible to use the realizations generated at values of $\lambda_0 < \lambda$ to estimate $L(\lambda)/L(\lambda_0)$. It should be pointed out

[1]The audience at the 2000 PMMB short course in Berkeley apparently found this unwanted dependence to be worrying when Elizabeth presented this work.

that a similar problem would occur with the importance sampling method of the previous chapter if one tried to combine the realizations from $P^*_{N_e}(\mathbf{X})$ with realizations from another distribution, say $P^*_{N_{e0}}(\mathbf{X})$ ($N_{e0} < N_e$) to estimate the likelihood at N_e.

Additionally, it is not immediately clear how best to compute $P_\lambda(\mathbf{Y}, \mathbf{X}^{(i)})/P_{\lambda_0}(\mathbf{Y}, \mathbf{X}^{(i)})$ even when $\lambda_0 > \lambda$. It is easy to compute $P_{\lambda_0}(\mathbf{Y}, \mathbf{X}^{(i)})$, but not so for $P_\lambda(\mathbf{Y}, \mathbf{X}^{(i)})$ when $\lambda \neq \lambda_0$. One approach we have used is to "let the last allele at each locus take up the slack." In other words, when $\mathbf{X}^{(i)}$ has been simulated from a chain under λ_0, then $P_\lambda(\mathbf{Y}, \mathbf{X}^{(i)})$ is computed by replacing each component, $\boldsymbol{X}_t^{(i)}$, of $\mathbf{X}^{(i)}$ with a new vector differing in the number of alleles subscripted by K. That is, the new vector is $(X^{(i)}_{t,1}, \ldots, X^{(i)}_{t,K-1}, X^{(i)}_{t,K} - \lfloor 2\lambda_0 C_t \rfloor + \lfloor 2\lambda C_t \rfloor)$. While this sort of scheme seems to work reasonably well in practice, the quantity $P_\lambda(\mathbf{Y}, \mathbf{X}^{(i)})/P_{\lambda_0}(\mathbf{Y}, \mathbf{X}^{(i)})$ computed in this way is not invariant to permutations of the order of alleles, and so it is unattractive. Since the ratio $P_{\lambda^*}(\mathbf{Y}, \mathbf{X})/P_\lambda(\mathbf{Y}, \mathbf{X})$ must also be computed for the Bayesian approach to estimating λ described above, this problem also plagues the Bayesian approach using the Wright-Fisher model.

In short, to compute the likelihood or the posterior distribution for λ using MCMC, it is desirable to be able to propose a reasonable model under which changes in the value of λ do not change the size of the state space of $\mathbf{X}$. Such a model is derived in the next section.

3.3 The Urn Model

From the preceding section, it should be clear that we desire a model of genetic inheritance that allows a population of effective size N_e, *but census size* C, to give rise to a new generation with genetic dynamics (increase in allele frequency variance, inbreeding, *etc.*) characteristic of a Wright-Fisher population of size N_e. The important, and more difficult part, is that we require doing this *without modeling the population as having any size other than* C. If our primary concern was that of matching the increase in allele frequency, we would seek a genetic inheritance model with the same number of individuals that we observe in the census, but with an increased variance in progeny allele frequency. An obvious way to do this would be to allow allele frequency in the following generation to follow a scaled beta-binomial distribution, rather than a scaled binomial distribution, because the beta-binomial

is a classical discrete distribution with extra-binomial variance. While such a distribution is convenient from the perspective of allele frequency, it is also the case that it arises from a system of mating that we can specify in terms of a sampling process on individual gene copies. In this section I will describe an urn sampling scheme from which it arises. In Section 3.4 I will describe another genesis of the same distribution.

Let us consider intergenerational sampling from generation 1, with C_1 reproductive adults, to their offspring who become C_2 reproductive adults. Let there be K alleles, indexed by i, with numbers of each at time 1 given by n_i ($\sum_{i=1}^{K} n_i = 2C_1$) and at time 2 by x_i ($\sum_{i=1}^{K} x_i = 2C_2$). By our urn scheme, the alleles in generation 2 are drawn by the following scheme: a gene copy is drawn at random from those present in generation 1, and a copy of it is placed in generation 2. Then the original gene copy is returned to generation 1, along with s new copies of it. This defines a multivariate Pólya-Eggenberger urn scheme (JOHNSON and KOTZ 1977), where the genes of different allelic type may be regarded as balls of different color. By this scheme, $\boldsymbol{X} = (X_1, \ldots, X_K)$, conditional on the n_i, follows the compound multinomial Dirichlet distribution, which is a multivariate generalization of the well-known beta-binomial distribution (JOHNSON *et al.* 1997).

The compound multinomial distribution with $2C_2$ trials arises as the marginal distribution of $\boldsymbol{X}$ from the hierarchy

$$\begin{aligned} \boldsymbol{Q} &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ \boldsymbol{X}|\boldsymbol{Q} &\sim \text{Mult}_K(2C_2, \boldsymbol{Q}), \end{aligned}$$

and has the probability mass function

$$P(\boldsymbol{X}|2C_2; \alpha_1, \ldots, \alpha_K) = \frac{(2C_2)!\,\Gamma(\alpha_\bullet)}{\Gamma(2C_2 + \alpha_\bullet)} \prod_{i=1}^{K} \left(\frac{\Gamma(X_i + \alpha_i)}{X_i!\,\Gamma(\alpha_i)} \right) \tag{3.3}$$

where $\Gamma(y) = \int_0^\infty t^{y-1} e^{-t} dt$ is the gamma function, $\alpha_\bullet = \sum_{i=1}^{K} \alpha_i$, and, as before $\sum_{i=1}^{K} x_i = 2C_2$. For the urn sampling scheme described above, $\boldsymbol{X}$ has p.m.f (3.3) with $\alpha_i = n_i/s$, $i = 1, \ldots, K$.

Conceptually, it is apparent that s plays an important role in determining the variance in family size in this urn-model for genetic inheritance. If $s = 0$ we have sampling with

replacement—the fundamental feature of the Wright-Fisher model. However, if $s > 0$ then each time a gene copy in generation 1 is selected to produce a daughter in generation 2, the probability is increased that that particular gene copy will have more daughters (and the probability is increased that other individuals will have fewer daughters). So, when $s > 0$ the variance in offspring number is increased (and the effective size of the population is thus decreased). The situation of $s < 0$ corresponds to an effective size which is larger than the census size of the population, as is sometimes achieved by a prescribed system of breeding in captive populations. $s = -1$ is hypergeometic sampling of gene copies into the next generation. In an urn-model population of constant size, this would correspond to every gene copy being copied exactly one time into the next generation. In the remainder of this chapter, however, we will concentrate on cases of $s \geq 0$. Though this model was developed, conceptually, in terms of stochastic replacements that were in integer units of gene copies, there is no mathematical restriction that s be an integer; non-integer values for s are permissible.

It is possible to derive standard results about the increase of allele frequency variance and the probability of identity by descent from a generation of genetic sampling via this urn model. This will be done in Section 3.5. These calculations are simplified by the fact that the marginal distribution of the i^{th} component in the compound multinomial distribution has a beta-binomial distribution with parameters $2C_2$, α_1 and $\alpha_\bullet - \alpha_i$. Thus it has p.m.f.

$$\begin{aligned} P(X_i|2C_2;\alpha_1,\ldots,\alpha_K) &= P(X_i|2C_2;\alpha_i,\alpha_\bullet-\alpha_i) \\ &= \frac{(2C_2)!\,\Gamma(\alpha_\bullet)}{\Gamma(2C_2+\alpha_\bullet)} \times \frac{\Gamma(X_i+\alpha_i)}{X_i!\,\Gamma(\alpha_i)} \times \\ &\quad \frac{\Gamma(2C_2-X_i+\alpha_\bullet-\alpha_i)}{(2C_2-X_i)!\,\Gamma(\alpha_\bullet-\alpha_i)} \end{aligned} \tag{3.4}$$

and first two central moments

$$\mathrm{E}X_i = 2C_2\frac{\alpha_i}{\alpha_\bullet} \tag{3.5}$$

$$\mathrm{Var}(X_i) = 2C_2\left(\frac{2C_2+\alpha_\bullet}{1+\alpha_\bullet}\right)\left(\frac{\alpha_i}{\alpha_\bullet}\right)\left(\frac{\alpha_\bullet-\alpha_i}{\alpha_\bullet}\right). \tag{3.6}$$

Any two components of a compound multinomial distribution are correlated, having covari-

ance

$$\mathrm{Cov}(X_i, X_j) = 2C_2\left(\frac{2C_2 + \alpha_\bullet}{1 + \alpha_\bullet}\right)\left(\frac{\alpha_i}{\alpha_\bullet}\right)\left(\frac{\alpha_j}{\alpha_\bullet}\right).$$

Derivation of the preceding results may be found in JOHNSON *et al.* (1997, pg. 81,82), where the authors also point out that the variance-covariance matrix for a compound multinomial random variable can be written as the product of $(2C_2 + \alpha_\bullet)/(1 + \alpha_\bullet)$ and the variance covariance matrix for a multinomial random variable with $2C_2$ trials and cell probabilities of $\alpha_i/\alpha_\bullet$, $i = 1, \ldots, K$.

3.4 Other Interpretations of the Urn Model

The urn model described above is a special case of more general classes of genetic inheritance models. Recognizing this provides us with a better idea of why this model might be suitable from a biological perspective, and will help in the analysis of fixation probabilities later in the chapter.

First, this urn model is a special case of a conditional branching process model. In such a model, each gene in a population independently produces a random number k of offspring, with k following the same distribution for each gene. The final result, in the following generation, however, is made conditional upon the population size at that time being C_2 diploids. By this conditioning, the numbers of offspring of each gene are no longer independent, but they are still exchangeable. This type of model was introduced by (MORAN and WATTERSON 1958), and studied in great detail by KARLIN and MCGREGOR (1965). The urn model corresponds to a conditional branching process model in which offspring number has the negative binomial distribution, a versatile distribution which has been previously employed to model the distribution of family sizes (RAO *et al.* 1973).

The negative binomial probability mass function may be parameterized in terms of α and β, shape and scale parameters, respectively, analogous to the parameters of a gamma distribution. The pmf of a NegBin(α, β) rv is

$$P(X|\alpha, \beta) = \left(\frac{\Gamma(X + \alpha)}{\Gamma(X + 1)\Gamma(\alpha)}\right)\left(\frac{1}{\beta + 1}\right)^{\alpha}\left(\frac{\beta}{\beta + 1}\right)^{X}.$$

Further, if X_1 and X_2 have negative binomial distributions with common scale β and shape

parameters α_1 and α_2, then their sum has a NegBin$(\alpha_1 + \alpha_2, \beta)$ distribution. This, and the fact that the ratio of a negative binomial variable with a sum of itself and other independent negative binomial variables with common scale is a beta-binomial random variable, makes it easy to verify that the conditional branching process model with negative binomial offspring distribution is the same as the urn model.

Another general class of models, which I call "two-stage" models may be formulated as follows:

1. Let there be C_1 diploids in the current generation

2. Form a gamete (or juvenile) pool in which the i^{th} gene copy $(i = 1, \ldots, 2C_1)$ is represented by H_i copies of itself, where H_i is a random number drawn from some distribution P_H, and where P_H is the same for all i.

3. Sample $2C_2$ gametes with replacement from the gamete pool to form the next generation.

This model will behave differently according to the type of distribution and the parameters specified for P_H.

When P_H is chosen to be a gamma distribution, Gamma(α, β), the two-stage model is identical to the urn model described in the previous section. This is so because the ratio of a gamma random variable to a sum of itself and other independent gamma random variables with the same scale parameter is a beta random variable. This relation generalizes to the multivariate case: if $Y_1, \ldots, Y_K$ follow independent gamma distributions, each with their own shape parameter α_i, but all with the same scale parameter, then

$$\boldsymbol{Q} = \left(\frac{Y_1}{Y_1 + \cdots + Y_K} \, , \, \cdots \, , \, \frac{Y_K}{Y_1 + \cdots + Y_K} \right)$$

is a Dirichlet random vector with parameters $(\alpha_1, \ldots, \alpha_K)$. Knowing this, it is straightforward to verify that the urn model with parameter s is identical to the two-stage model with P_H being a gamma distribution with shape parameter $\alpha = 1/s$, and arbitrary scale.

The gamma distribution might seem a poor choice for the number of gametes produced by a gene, since this will yield non-integer values for the number of gametes. However, the

mathematical tractability achieved by using a continuous distribution for gamete number is substantial. Further, the gamma distribution is a continuous analogue which closely resembles the negative binomial distribution. Obviously, the continuous approximation will be better when the scale is larger, as will be the case with organisms that produce many offspring. Nonetheless, even with fairly small offspring number, the continuous approximation is good.

If we allow that the urn model is a good approximation to the two-stage model with negative binomial gamete numbers, then we have yet another interpretation of the model owing to the genesis of the negative binomial as a gamma-weighted mixture of Poisson random variables. That is, if $\phi \sim \text{Gamma}(\alpha, \beta)$ and $H \sim \text{Poisson}(\phi)$, then, marginally $H \sim \text{NegBin}(\alpha, \beta)$. Therefore, another biologically reasonable interpretation of the urn model is that the number of offspring copies of a gene surviving to childhood follows a Poisson distribution; however, there is heterogeneity in the population, so that the expected number of offspring surviving to childhood varies randomly across the pool of parental gene copies, following a gamma distribution. Children are then chosen randomly (and with replacement) to survive to adulthood. So long as the number of juveniles is large, sampling with replacement is a reasonable approximation to the actual hypergeometric sampling that would actually transpire.

3.5 Comparison to the Wright-Fisher Model

For different values of C_1, C_2, and s, how does this model compare to the Wright Fisher model? In this section we investigate some fundamental quantities in the urn model: variance in offspring number, probability of identity by descent, and allele frequency variance. From these calculations, it is possible to derive the inbreeding and variance effective sizes of ideal populations reproducing via the urn model. I also investigate the probability of allele fixation in these models, and finally describe how one may relate the stochastic replacement quantity s in a population of changing census size over time, to the quantity λ.

3.5.1 *Offspring number*

The expected number of offspring copies of a parental gene may be computed simply from (3.5). We consider an urn model in which one ball (gene copy) is white and the other $2C_1 - 1$ balls are black. The expected number of offspring genes of a single gene copy is then the expected number of white balls obtained when $2C_2$ total balls are drawn in our urn sampling scheme. In this case $n_i = 1$, so $\alpha_i = 1/s$ and $\alpha_\bullet = 2C_1/s$. So the expected number of offspring genes is

$$\mathbb{E}(\#\text{ of offspring of a single gene}) = 2C_2 \cdot \frac{1/s}{2C_1/s} = \frac{C_2}{C_1}. \tag{3.7}$$

Note that this is the same expectation that one would get for Wright-Fisher sampling (s=0) in a population of changing size.

The variance in offspring number is obtained by substituting the proper expressions into (3.6), giving

$$\begin{aligned}\text{Var}(\#\text{ of offspring of a single gene}) &= \frac{2C_2\left(2C_2 + \frac{2C_1}{s}\right)\left(\frac{1}{s}\right)\left(\frac{2C_1-1}{s}\right)}{\left(1+\frac{2C_1}{s}\right)\left(\frac{2C_1}{s}\right)^2}. \\ &= \frac{C_2}{C_1} \cdot \frac{2C_1 + 2C_2 s - 1 - sC_2/C_1}{2C_1 + s} \end{aligned} \tag{3.8}$$

It is instructive to notice that in the case of $C_1 = C_2 = C$ this reduces to

$$\text{Variance of offspring number} = \left(1 - \frac{1}{2C}\right)\left(\frac{s+1}{1+\frac{s}{2C}}\right)$$

which is precisely the variance in offspring number in a Wright-Fisher population, inflated by the factor $2C(s+1)/(2C+s)$. Thus, if $s = 0$, we obtain the Wright-Fisher variance in offspring number, as we ought to.

3.5.2 *Identity by descent*

We may similarly determine the probability that a randomly chosen pair of gene copies drawn from the $2C_2$ gene copies of generation 2 are identical by descent (IBD), *i.e.*, they are both copies of a single gene copy in generation 1. Let the random variable X_j denote the

number of offspring of a single, specific, gene copy, say the j^{th} gene copy, in generation 1. Conditional on $X_j = x_j$, the probability that a pair drawn from generation 2 is IBD *and* that both members of the pair are copies of the j^{th} gene in generation 1 is $(\frac{x_j}{2C_2})(\frac{x_j-1}{2C_2-1})$. Therefore, the probability that a pair is IBD and are copies of the j^{th} gene may be written as

$$P(\text{IBD } \textit{and} \text{ copies of gene } j) = \sum_{x_j=1}^{2C_2} \left(\frac{x_j}{2C_2}\right)\left(\frac{x_j-1}{2C_2-1}\right) P(X_j = x_j),$$

where $P(X_j = x_j)$ is, as given in (3.4), the beta-binomial probability of gene offspring number. This may be rewritten as

$$\begin{aligned} &= \left(\frac{1}{2C_2(2C_2-1)}\right) \sum_{x_j=1}^{2C_2} (x_j^2 - x_j) P(X_j = x_j) \\ &= \left(\frac{1}{2C_2(2C_2-1)}\right)\Big(\text{Var}(X_j) + (\mathbb{E}X_j)^2 - \mathbb{E}X_j\Big) \end{aligned} \tag{3.9}$$

which, upon substituting expressions (3.7) and (3.8) for the mean and the variance, and factoring out a factor of C_2/C_1 simplifies to

$$\begin{aligned} &= \frac{1}{2C_1(2C_2-1)}\left(\frac{2C_1 + 2C_2 s - 1 - s\frac{C_2}{C_1}}{2C_1+s} + \frac{C_2}{C_1} - 1\right) \\ &= \frac{1}{2C_1(2C_2-1)}\left(\frac{(2C_2 - \frac{C_2}{C_1} - 1)s - 1}{2C_1+s} + \frac{C_2}{C_1}\right). \end{aligned} \tag{3.10}$$

Equation 3.10 gives the probability that a randomly drawn pair is IBD *and* both are copies of a particular gene j. Since the events of being copies of a particular gene j and copies of a gene k are disjoint for $j \neq k$, the probability of IBD may be found by summing (3.10) over j from 1 to $2C_1$, and so

$$P(\text{IBD}) = \frac{1}{2C_2-1}\left(\frac{(2C_2 - \frac{C_2}{C_1} - 1)s - 1}{2C_1+s} + \frac{C_2}{C_1}\right). \tag{3.11}$$

It is once again instructive to consider the case $s = 0$, in which the above expression simplifies to $1/(2C_1)$. This does not depend at all on C_2, which is true with Wright-Fisher sampling. With respect to inbreeding in the Wright-Fisher model, it does not matter how much a population has grown over the past generation; what counts is how small the population was in the parental generation. The same is not true in the urn model with $s \neq 0$. This

results from the fact that the relative effect of s depends on how many gene copies are initially in the urn *and* how many are drawn from it.

If the urn-model population were of constant size C ($C_1 = C_2 = C$), then (3.11) reduces to

$$P(\text{IBD}) = \frac{s+1}{2C+s}.$$

Thus, for an urn-model population of constant size C to have an inbreeding effective size of N_e, s would have to be chosen so that

$$s = \frac{2C - 2N_e}{2N_e - 1}. \tag{3.12}$$

3.5.3 Allele frequency variance

We can consider an allele of type "i" which is present in n_i copies out of the $2C_1$ gene copies of generation 1. The number of copies of allelic type i in the $2C_2$ gene copies of the following generation is then the random variable X_i. Substituting the appropriate quantities for $\alpha_\bullet$ and α_i into (3.6) gives

$$\text{Var}(X_i) = \frac{2C_2\left(2C_2 + \frac{2C_1}{s}\right)\left(\frac{n_i}{s}\right)\left(\frac{2C_1 - n_i}{s}\right)}{\left(\frac{2C_1}{s}\right)^2\left(1 + \frac{2C_1}{s}\right)} \tag{3.13}$$

$$= \frac{2C_2}{4C_1^2} \cdot n_i(2C_1 - n_i)\frac{2C_1/s + 2C_2}{2C_1/s + 1}. \tag{3.14}$$

Hence the corresponding allele *frequency*, $X_i/(2C_2)$ will have variance which is $1/(4C_2^2)$ of the expression in (3.14). Denoting $n_i/(2C_1)$ by p, we may write

$$\text{Var}\left(\frac{X_i}{2C_2}\right) = \frac{p(1-p)}{2C_2}\left(\frac{2C_1 + 2C_2 s}{2C_1 + s}\right). \tag{3.15}$$

Once again, with $s = 0$ we have the binomial variance of the Wright-Fisher model. Also, it is clear from (3.15) that for $s > 0$ the allele frequency variance is inflated over that of a Wright-Fisher population of size C_2. The magnitude of the effect of s depends on the relative and absolute sizes of C_1 and C_2. With $C_1 = C_2 = C$ (3.15) reduces to

$$\text{Var}\left(\frac{X_i}{2C}\right) = p(1-p)\frac{1+s}{2C+s}.$$

This means that for an urn-model population of constant size C to have a variance effective size of N_e, s would be

$$s = \frac{2C - 2N_e}{2N_e - 1}, \tag{3.16}$$

the same as (3.12).

3.5.4 Probability of allele fixation in one generation

So far, we have seen that both the inbreeding and variance effective sizes of an urn-model population of constant size C are N_e when s is chosen to be $(2C - 2N_e)/(2N_e - 1)$. Given this, it is tempting to imagine that the eigenvalue effective size (EWENS 1979) in an urn model population of size C will be the same as that in a Wright-Fisher population of size N_e when $s = (2C - 2N_e)/(2N_e - 1)$. The eigenvalue effective size is the size of a Wright-Fisher population with the same largest non-unit eigenvalue of its transition probability matrix, and it determines the rate of loss of rare alleles over time in the population. I have not yet computed the eigenvalue effective size of this urn model. Such an analysis should not be difficult, however, as this urn model is an exchangeable model in the sense of EWENS (1979, p. 77), and so the general theory of CANNINGS (1974) for computing eigenvalues as the expected value of products of offspring numbers should hold. I have not yet pursued this, as I am primarily interested in the application of this urn model to inference problems. Nonetheless, future work along this line might permit some analytical results of the model to be obtained directly from the earlier work of KARLIN and MCGREGOR (1965).

However, I have performed some related numerical investigations. Rather than determining the eigenvalue effective size of the population, directly, I have calculated, for the Wright-Fisher model and the urn model, the probability that an allele is lost or becomes fixed *in a single generation.* This quantity will be related to the eigenvalue effective size, and is simpler to compute. These numerical investigations (using the probabilities of the zero class computed from Equation 3.4 and from the binomial distribution) show that the probability of allele fixation in a single generation of reproduction in an urn model population of census size C and variance effective size $N_e < C$ is always smaller than the probability of allele fixation in a single generation in Wright-Fisher population of size N_e. An example

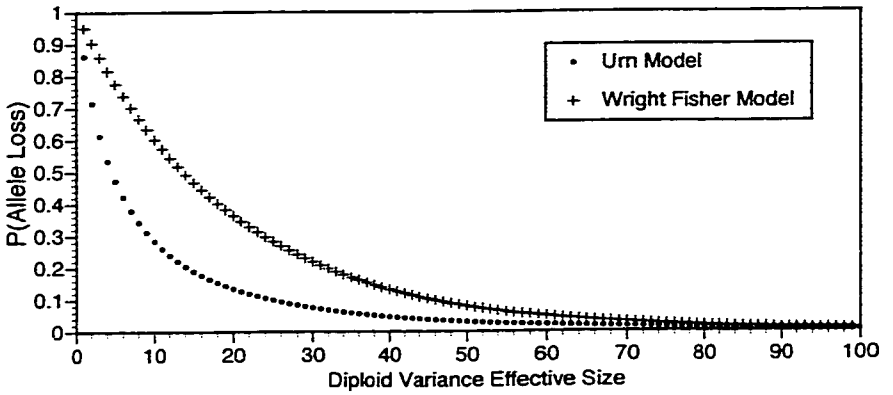


Figure 3.1: Probability that an allele at frequency .025 is lost from a population in a single generation of reproduction. Crosses show the single-generation fixation probabilitities for a Wright-Fisher population of size given on the horizontal axis. Filled circles show the single-generation fixation probabilities for an urn model population with census size $C = 100$ and variance effective size given on the horizontal axis. As is apparent, rare alleles are more likely to be lost in a single generation from a Wright-Fisher population of size N_e than from an urn model population of variance effective size N_e.

of this is given in Figure 3.1 which shows the probability of allele loss for different values of the variance effective size. Dots are plotted for the probability that an allele found in five copies in an urn-model population of size 100 diploids is lost in the next generation. Plotted on the same graph are crosses showing the probability that an allele at frequency $5/200 = .025$ is lost in one generation of Wright-Fisher sampling in a population of the same variance effective sizes. It is apparent from the figure that the single-generation fixation probabilities are much higher in the Wright-Fisher model. For example, with $\lambda = .3$ ($N_e = 30$ in the graph), it is three times more probable that zero copies of an allele will appear in the following generation in the Wright-Fisher population than in the urn-model population.

It might seem unfortunate that the urn-model population does not correspond closely to the Wright-Fisher population in terms of single-generation allele fixation probabilities. I argue, however, that this results from a deficiency of the Wright-Fisher model, rather than an "inaccuracy" in the urn model I have proposed. To some extent the discrepancy must be due to the fact that in a Wright-Fisher population of size N_e, there are only $2N_e + 1$ possible

values for the number of copies of any one allele, as opposed to $2C+1$ possible values in an urn model population. The implication of this is that in an urn-model population, alleles may persist at lower frequencies than would be allowed in a Wright-Fisher population. Since almost all natural populations have census sizes which are larger than their effective sizes, alleles in those natural populations may survive at lower frequencies than would be allowed in a Wright-Fisher population of comparable variance or inbreeding effective size. In other words, modeling a population by a Wright-Fisher model of comparable variance effective size will, in almost all cases, lead one to overestimate the probability that an allele is lost from the population. I devote Section 3.6 to this topic.

3.5.5 Comparable λ

If census sizes are known, $C_0, \ldots, C_T$, and we want to find an urn model that is comparable to one in which the reproduction is like a Wright-Fisher population with effective sizes $\lfloor \lambda C_0 \rfloor, \ldots, \lfloor \lambda C_T \rfloor$, then we can define stochastic replacements s_t for each pool of C_t adults, $t = 1, \ldots, T$. In this, we shall make our main concern that of matching the increase in allele frequency variance. In a generation of Wright-Fisher sampling from a gamete pool with allele frequency p and drawing $\lfloor 2\lambda C_t \rfloor$ gene copies, the variance will be $p(1-p)/\lfloor 2\lambda C_t \rfloor$. Setting this equal to the allele frequency variance derived for the urn model (3.15), and changing C_1 and C_2 to C_{t-1} and C_t, respectively, gives

$$\frac{p(1-p)}{\lfloor 2\lambda C_t \rfloor} = \frac{p(1-p)}{2C_t}\left(\frac{2C_{t-1}+2C_t s}{2C_{t-1}+s}\right).$$

Then, disregarding the floor function, which resulted from a discretization imposed by adherence to the Wright-Fisher model, this may be solved for s, giving

$$s_t = \frac{2C_{t-1}(1-\lambda)}{2\lambda C_t - 1} \quad , \quad t = 1, \ldots, T. \tag{3.17}$$

This will form the basis for making inference about λ using the urn model.

3.6 Allele Fixation in Population-Genetic Models

In 3.5.4 we saw that an urn model with a given variance effective size has a much smaller probability of allele loss in one generation than the corresponding Wright-Fisher model.

The purpose of this section is to demonstrate that the probability of fixation in a single generation of reproduction in a Wright-Fisher model of size N_e is probably too high for a natural population of variance effective size N_e. This is done indirectly, by showing that the two-stage model (Section 3.4), even under an extreme assumption about P_H, exhibits single-generation fixation probabilities for a given variance effective size that are lower than those in the Wright-Fisher model.

The argument proceeds as follows: let P_H belong to $\mathcal{C}$, the class of distributions with compact support on non-negative values. This corresponds to individuals having some maximum number $O_{\max}$ of offspring that may survive to the gamete stage in the two-stage model. The urn model can approximately fit into this scheme if we define the scale parameter of the corresponding gamma distribution to be small enough that the probability that offspring number is greater than $O_{\max}$ is very small. Within this class, one may have discrete or continuous, unimodal or multimodal distributions, *etc.* Through computer simulations, I have found that, for a given variance effective size, the $P_H \in \mathcal{C}$ that gives rise to the largest probability of allele fixation in a single generation of reproduction seems to be the scaled Bernoulli distribution—with probability p an individual has $O_{\max}$ offspring, and with probability $1-p$ it has zero offspring. (Of course, the value of $O_{\max}$ is irrelevant in this case and may be set to unity without loss of generality.) I have not yet tried seriously to prove that $P_H \sim$ Bernoulli gives the maximum single-generation fixation probability for a two-stage model of given variance effective size, but a proof should be possible. The probability of fixation in a Wright-Fisher model of size N_e however, is still always greater than that in the two-stage model with $P_H \sim$ Bernoulli(p) for given variance effective size (which depends on p). An example of this appears in Figure 3.2, which is similar to Figure 3.1, except that it includes a series of dots for the two-stage model with Bernoulli offspring distribution.

Figure 3.1 shows us that the probability of allele loss in a single generation in a Wright-Fisher model *exceeds* that of a two stage model with Bernoulli offspring number distribution for all variance effective sizes. This suggests that using a Wright-Fisher model of size N_{e_v} to calculate the probability of allele extinction in one generation in a natural population of census size C and estimated variance effective size N_{e_v} will likely overestimate that probability. This is apparent, because the single-generaton fixation probability calculated

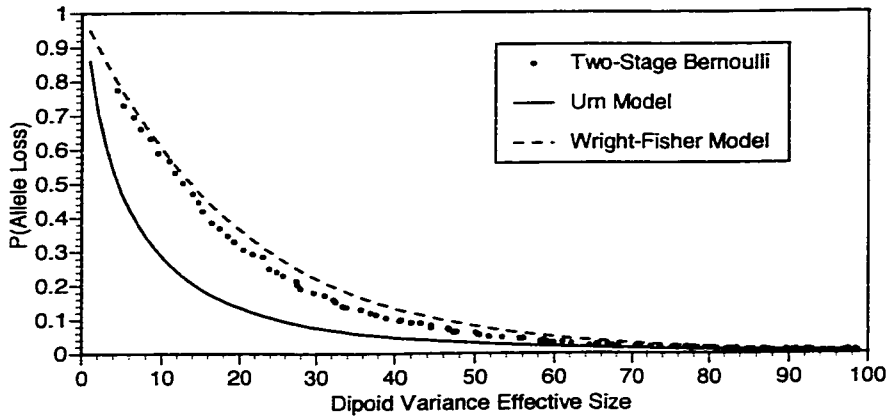


Figure 3.2: Probabilities that an allele at frequency .025 is lost in one generation of reproduction as a function of variance effective size under three different models of genetic inheritance. The dashed line is for a Wright-Fisher model of size N_e. The solid line is for an urn model of size 100, with s set to obtain the corresponding variance effective size. The dots are simulation results using a two-stage model in a population of size 100 adults with $P_H \sim$ Bernoulli(p). The parameter p increases from .04 to 1.0, from left to right in the figure—higher values of p give rise to higher variance effective sizes.

from a Wright-Fisher model is greater than that for any two-stage model of census size C. The degree to which the single-generation probability of allele fixation computed from a Wright-Fisher model of size N_{e_v} overestimates the true fixation probability depends on the actual distribution of offspring numbers in the natural population. If the natural population is one in which a few individuals have many offspring, and the rest have zero, then the Wright-Fisher model will not be grossly inaccurate. If, however, the population is one in which the distribution of offspring number is not bimodal (for example if the distribution of offspring number is shaped like the negative binomial distribution), then the Wright-Fisher model of size N_{e_v} will poorly represent that natural population in terms of the rate of loss of rare alleles. In such a case, the urn model will provide a much more faithful representation. While the former scenario may be appropriate for organisms in which families all live or die together on the basis of some environmental factors, it seems unlikely that it would apply well to many species.

Finally, these concerns are important from an inference perspective. In making likelihood

inference about effective size, the information in fixation probability is used (since one makes full use of the entire distribution of allele counts—not just their variance). As a result, for most natural populations, using a likelihood derived from the Wright-Fisher model will yield an estimate of effective size which is larger than the estimate that would be obtained under a likelihood model based on the urn-sampling scheme using an estimate of the census size of the population (as described in the following section). Such an effect, however, is quite small, as we see below.

3.7 MCMC for Bayesian Estimation Under this Urn Model

The urn model allows easier implementation of the sampler described in Section 3.2.2 for Bayesian inference of λ. The formulation still follows directly from Sections 3.2.1 and 3.2.2 as presented, but with the definition of $P_\lambda(\boldsymbol{X}_t, \boldsymbol{X}_{t-1})$ made in terms of transition probabilities dictated by the urn model with a stochastic replacement corresponding to a particular value of λ and the observed census sizes. Thus, for a particular value of λ and the census sizes C_{t-1} and C_t at times $t-1$ and t, $P_\lambda(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})$ is written in terms of s_t, following 3.3:

$$\begin{aligned} P_\lambda(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) &= P(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}, s_t, C_t, C_{t-1}) \\ &= P(\boldsymbol{X}_t|2C_t; \alpha_1, \ldots, \alpha_K) \\ &= \frac{(2C_t)!\,\Gamma(\alpha_\bullet)}{\Gamma(2C_t+\alpha_\bullet)} \prod_{i=1}^{K} \left(\frac{\Gamma(X_{t,i}+\alpha_i)}{X_{t,i}!\,\Gamma(\alpha_i)} \right) \end{aligned} \tag{3.18}$$

where $\alpha_i = X_{t-1,i}/s_t$, $\alpha_\bullet = \sum_{i=1}^{K} \alpha_i$, and s_t is computed from Equation 3.17. I will briefly describe a simple implementation of an MCMC sampler here. This is a special case of the sampler described in the following chapter where a more comprehensive treatment is given.

I implemented such a sampler to compute the posterior probability for values of $\lambda \in \Lambda = \{\lambda_{\min}, \lambda_1, \ldots, \lambda_n, \lambda_{\max}\}$ using a simple random walk proposal distribution for λ (*e.g.*, $h(\lambda^*|\lambda) = 1/2$ for $\lambda^* = \lambda_{i-1}$ or $\lambda^* = \lambda_{i+1}$, $\lambda \notin \{\lambda_{\min}, \lambda_{\max}\}$; $h(\lambda^* = \lambda_1|\lambda_{\min}) = 1$; and $h(\lambda^* = \lambda_n|\lambda_{\max}) = 1$), and a discrete uniform proposal distribution $q(w|X_{t,k}, X_{t,\ell})$ between the integers $-L$ and L, inclusive, where $L = 3 + \sqrt{\min\{X_{t,k}, X_{t,\ell}\}}$. I proposed E updates to random components of $\mathbf{X}$ for each update proposed to λ, where E was chosen so that

each component of $\mathbf{X}$ was expected to receive one proposal for each proposal to change λ. The proportion of time during the MCMC run in which $\lambda = \lambda_i$ estimates the posterior probability that $\lambda = \lambda_i$.[2]

I used this sampler to compute a posterior distribution for λ given the data of BEGON *et al.* (1980). They reported census sizes of tens of thousands of flies. I thus set $C_t = 10,000$ and computed a posterior for $\lambda \in \Lambda$ assuming a uniform prior on the values in Λ. The points in Λ were chosen to correspond to the values of N_e for which log-likelihood values were computed in Chapter 2. Starting values for the $\mathbf{X}$ were obtained using a realization from the forward-backward sampler of Chapter 2. This made burn-in essentially unnecessary. The sampler was run for 100,000 updates of λ, with the collection interval $u = 1$. This took 12 hours on a laptop computer with a 266 Mhz G3 (Macintosh) processor. Since this is a case with a one-dimensional parameter with uniform prior, I was able to convert the posterior distribution for λ into a log-likelihood curve for λ and hence for N_e. Thus, I was able to compare it to the log-likelihood curve computed in Chapter 2. The result appears in Figure 3.3.

The two curves are nearly identical. However, under the urn model the log-likelihood is not as low for small values of N_e as for the importance sampling method under the Wright-Fisher model. This slight difference may arise from the different probabilities that the two models assign to the event of allele fixation as described in Section 3.6.

The curve obtained by MCMC required less computational time than that obtained by importance sampling. It was also significantly simpler to implement the MCMC scheme than the importance sampling scheme of Chapter 2. The MCMC scheme and sampling models here are as simple as possible. In the following chapter they will both be extended to handle a wider range of scenarios.

---

[2]This uses the Monte Carlo estimator for probabilities (1.3). In the following chapter I present a more sophisticated method for updating λ that will also permit a Rao-Blackwellized estimator (*i.e.*, Equation 1.11) for the posterior distribution of λ.

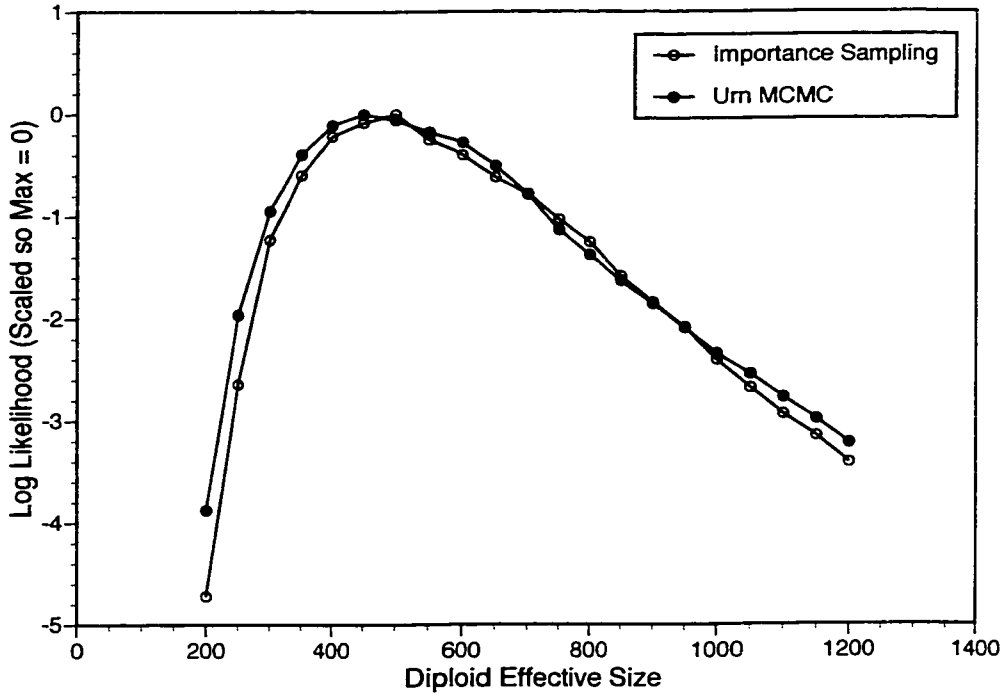


Figure 3.3: Log-likelihood (scaled so that the maximum is zero) for N_e given the data of BEGON *et al.* (1980). Open circles show the result from the importance sampling scheme of Chapter 2. Filled circles show the log-likelihood curve computed by an MCMC scheme using the urn model of the present chapter and an assumed census size of 10,000 flies each generation. The two methods give comparable results, but the MCMC scheme requires less computer time. The curve for the MCMC scheme is shifted slightly to the left relative to the curve from importance sampling. This may be due to the differences in fixation probabilities between the Wright-Fisher model used in the importance sampling scheme, and the urn model used for the MCMC approach.

3.8 Discussion

In this chapter, I have described and characterized an alternative to the Wright-Fisher model for genetic inheritance. This alternative, which I call the "urn model" carries particular benefits for estimating λ, the per-generation ratio of effective breeders to the census number of breeders, when the census number is known. The urn model allows the formulation of a sensible likelihood for λ and it makes MCMC calculation of that likelihood possible. The likelihood was derived under three assumptions that merit particular attention here. The first is the assumption that λ is constant over time. The remaining two concern questions of the census size: "What ages or stages of individuals should be counted in the census size of the population?" and "What if census sizes are not known without error, but are, rather, themselves estimates with some uncertainty?"

3.8.1 Constancy of λ

In the likelihood for λ developed in this chapter, the assumption is made that λ is constant over time. In practice, this assumption will likely be violated. In natural populations, one would expect that λ could be influenced by time-varying factors like population size or climatic conditions. In managed or controlled populations it would be natural for λ to change over time due to changes in harvest regimes or breeding practices or pesticide use. The methods I have developed could be extended to allow the value of λ to vary over the different inter-sample intervals. For example, $\lambda_{[0,t_1]}$ could represent the value of λ that applied to the generations between the sample at time 0 and the sample at time t_1 when the next sample is drawn from the population, and $\lambda_{[t_1,t_2]}$ would apply to the interval between the sample at time t_1 and the next sample in time, and so forth. Then, these specific $\lambda_{[\cdot,\cdot]}$'s could all be estimated as separate parameters. Some precision will be lost because, though all the $\lambda_{[\cdot,\cdot]}$'s would be estimated jointly, only a fraction of all the data will apply directly to each $\lambda_{[\cdot,\cdot]}$. Since λ may also be affected by the census size of the population (for example, if there are limited nesting sites, then a generation descended from a large group of parents might be expected to have a smaller λ than a generation descended from only a few parents who did not have to compete for nest sites), one could also formulate a model in which λ in

any generation, t, is a function of C_t or C_{t-1}. In the Bayesian setting it would be possible to design a reversible jump MCMC (GREEN 1995) sampler for comparing different models which correspond to different assumptions about how λ varies over time or as a function of census size.

It is also not unlikely that λ would vary each generation *within* an inter-sample interval; however, it would be folly to try to model this with a separate λ for each generation because, within a single inter-sample interval, separate λ's for each generation would not be practically identifiable. For the case of λ varying from generation to generation within the inter-sample interval, it does not seem straightforward to derive an approximate expression for what the maximum likelihood estimator for $\lambda_{[\cdot,\cdot]}$ estimates. It is not, for example, the harmonic mean of the λ values each generation, except in the case of constant census size.

When $\lambda_{[\cdot,\cdot]}$'s for all the inter-sample intervals are constrained in the model to be equal to an overall λ as I have presented the model earlier in the chapter, another question arises: "How is the information from different intervals weighted and combined to arrive at a single overall estimate of λ?" The likelihood-based or Bayesian estimate of λ will be influenced most by the intervals in which census sizes are small. This occurs because the amount of information about λ in the data, relative to the noise from the random process of drawing genetic samples, increases when the census size descreases. This observation does suggest that in future analyses estimating λ in populations with census sizes that fluctuate greatly, it would be prudent to carry out an alternative analysis under a model in which intervals containing very small census sizes had a $\lambda_{[\cdot,\cdot]}$ which was a separate parameter from the $\lambda_{[\cdot,\cdot]}$'s in intervals with much larger census sizes.

The observation also exposes a generic difficulty in estimating N_e from multiple samples in time. The first formal method described for doing so, that of POLLAK (1983), assumes that N_e is constant over the entire period from the first to the last sample. This is likely to be untrue, especially if the census size fluctuates greatly over time. The likelihood analyses of the BEGON *et al.* (1980) dataset presented in this chapter and the preceding one are similar in that they assume a constant N_e or C_t from $t = 0$ to T. In all of the above cases, asking the question of how to weight information from different intervals shows us that esimating a single N_e over a period of time which includes more than two sampling episodes

is not a particularly well-defined problem. When census sizes are available, then estimating λ, instead of N_e provides an alternative.

3.8.2 How do we define census sizes?

In populations with discrete generations and in which the reproductively active adults may be readily counted and sampled separately from the rest of the population, it is quite clear what the census size C_t should be—the number of actively reproducing adults at generation t. For example, with pink salmon (*Oncorhynchus gorbuscha*), C_t should count the number of spawning adults at time t. In other species, in which the distinction betwen mature adult and developing youngster is not so clear, it is also less clear what quantity (*i.e.*, total number of animals, number of females, *etc.*) should be chosen to be represented by C_t. Still, a number of sensible choices could be used, depending on the life-history features of the species under study. For example, in a population with discrete generations in which only the individuals above a certain age are reproductively active, C_t should count the number of such individuals at time t—it should not be the whole population size—adults and youngsters together. Likewise, the genetic samples should be drawn from the reproductively active individuals or their immediate offspring.

Many organisms do not have discrete generations, of course. In such cases, the choice of which population census quantities to define as C_t and which segments of the population to sample should be made within the context of a stochastic model that is faithful to the life history of the organisms under study. Only within the context of such a life-history based model, will the meaning of λ be clear. Chapter 4 describes the elaboration of the urn model of this chapter to the life history pattern of Pacific salmon that mature at different ages.

There is a somewhat lively literature and debate involving the estimation of the ratio N_e/N, where N_e in the numerator represents a "long term" effective size of a population over time and N in the denominator represents some sort of "long term" census population size (NUNNEY 1995; HUSBAND and BARRETT 1995; VUCETICH *et al.* 1997). The goal in this series of papers is quite different from mine here. Under the perspective of the above authors, fluctuations in population size and overlapping generations are lumped together

with variance in offspring number as inseparable factors which influence the effective size of a population. In contrast, the Pólya urn model here allows one to define the parameter λ to measure the degree to which genetic change is influenced by variation in offspring number *separately* from fluctuations in population size (and, in the next chapter, separately from the effect of overlapping generations in the population).

It should be clear why being able to estimate λ is advantageous: it is often possible to directly observe, and therefore account for, fluctuating population size and overlapping year classes. It is harder to observe λ, and in fact, in organisms with high juvenile mortality, it must be estimated using genetic data. Once an estimate of λ has been made, however, it can be used to predict genetic change in a population given patterns of fluctuating population size or overlapping generations observed in the future. This sort of "predictive analysis", applied to census data collected in the future, but using an estimated λ from genetic and census data collected in the past, is not available if one computes "long-term" N_e/N as other authors have pursued.

3.8.3 Census sizes estimated with error

Thoughout this chapter (and the next) I assume that the census sizes C_t are known without error. While census sizes of some organisms can be determined quite accurately, they are seldom known with certainty. For some populations, in fact, census estimates may be very imprecise. It would be worthwhile to include this uncertainty in census sizes into the estimation procedure for λ. I have not pursued that here, but leave it as an open problem.

An *ad hoc* approach to propagating the uncertainty in census size estimates to the estimates of λ would probably be worth investigating, since treating the problem fully from the likelihood or Bayesian perspective would be very challenging, computationally. For example, if the true, unknown census size values were modeled as latent variables, then in an MCMC scheme, the Hasting's ratio involved in proposing changes to those latent census sizes would depend not only on the census size data, but also on the genetic data at all the loci. Not only that, but changes to the latent census size would also change the size of the space of the latent $\mathbf{X}$ variables. Designing an MCMC sampler that mixed well in this

context would, I believe, be exceedingly difficult.

Chapter 4

λ AND OVERLAPPING GENERATIONS

4.1 Introduction

The reproduction of natural populations is not always well-characterized by a model with discrete generations. In particular, of the species of Pacific salmon, only the pink salmon, *Oncorhynchus gorbuscha*, has a discrete-generation life history; all pink salmon, within their native range, mature at two years of age. The other species of Pacific salmon mature, reproduce, and senesce at a variety of ages. For example, a spawning collection of chinook salmon might consist of three-, four-, five-, and six-year old fish. Each different year class has descended from a different collection of reproducing parents.

These factors complicate the estimation of effective size in salmon populations—the Wright-Fisher model simply does not describe their life history very well. In simulation studies, however, WAPLES and TEEL (1990) and WAPLES (1990a) show that many quantities of interest, such as allele frequency variance, rate of loss of heterozygosity, and the rate of loss of rare alleles, in a salmon population all depend on the average generation length and the effective number of breeders per year, N_b. WAPLES (1990b) demonstrates that there is an approximately linear relationship between Wright's F-statistic and $1/N_b$ in salmon populations. He then shows how that relationship may be used to estimate the harmonic mean N_b from genetic samples of juveniles descended from temporally-spaced brood years.

The goals of this chapter are different. Rather than estimating an overall effective number of breeders for the population, the interest here is in estimating a λ-like quantity—a ratio of effective spawners to the census number of spawners—given data on the census sizes of fish of different age-groups and genetic data either from adults or juveniles or both. This goal is pursued within the context of a long time series of demographic and genetic data of the sort that should become increasingly available due to the falling costs of genotyping

and the ability to amplify DNA from archived fish materials (NIELSEN *et al.* 1999). For example, ARDREN (1999) describes extensive fish scale collections from two intensively-studied steelhead (*Oncorhynchus mykiss*) populations on the West Coast. These fish scales, taken from both spawning adults and outmigrating juveniles allow the age of each fish to be determined. Also, as MILLER and KAPUSCINSKI (1997) and ARDREN (1999) have shown, microsatellite loci may be reliably amplified from these fish scales. Furthermore, Canadian fisheries agencies together with other scientists have proposed launching a program of close genetic monitoring of a "reference" stream on the coast of Vancouver Island, in which spawners are carefully counted and samples from the population are genotyped on a regular basis (William Ardren, pers. comm.). The data-analysis framework described in this chapter would be very appropriate for such monitoring programs.

While we will conceptually think in terms of a λ for each age group of adults, we will rely heavily on the urn model for genetic inheritance, described in the last chapter, in order to derive a probability model and develop Markov chain Monte Carlo (MCMC) methods for computing the posterior probabilities of the parameters. Having such a model in which the census number of breeders is considered known, and is used in the probabilistic model for the population, but in which the corresponding effective size may be altered by changing a simple parameter which does not alter the census sizes, is crucial to formulating a reasonable probability model. In the following section I develop the probability model and several extensions to accommodate different sampling strategies and the occurrence of null alleles. In Section 4.3, I exploit the simple neighborhood structure in the model to develop single-site Metropolis-Hastings updates for the latent variables in the model. These updates form the basis of a Markov chain from which we may sample from the posterior distribution of the parameters of interest. I represent the dependence structures using the intuitively appealing "language" of graphical models. Since I use only the simplest results from the theory of graphical models, it should be self-explanatory to most. However, the reader interested in learning more about graphical models in statistics is referred to the comprehensive text by LAURITZEN (1996). Finally, in Section 4.5, I demonstrate the potential of the method in several small trials on genetic data simulated using census size estimates of chinook salmon from a Snake River tributary. The results suggest that the method works under such

conditions. However, future work assessing the robustness of the method to departures from the assumed model and characterizing the mixing properties of the sampler under different data scenarios is warranted.

Earlier work that I did on this topic involved an extension of the methods of Chapter 2 to a case with a Pacific-salmon-like life history. I did not pursue that approach any further, but I include a brief description of it in Appendix B. The urn model provides a superior approach.

4.2 Overlapping Generations via an Urn Model

The urn model for genetic inheritance described in the previous chapter provides a good mechanism for modeling genetic drift in populations with complex life histories, like those of Pacific salmon. This section describes how it may be applied in such a context. First we shall examine a model for the conditional dependence structure of the variables in such a population, without reference to specific probability distributions. We then "clothe that backbone" with the specific probability distributions chosen to represent the population-genetic sampling, as well as the taking of genetic samples from juveniles and adults.

4.2.1 Dependence structure with the Pacific salmon life history

We consider a population of dioecious, diploid, semelparous organisms, in which adults may mature and mate between the ages of a^- and a^+, inclusive, and from which it is straightforward to sample and count the reproductive adults separately from the rest of the population (as is the case with Pacific salmon). For example, a pink salmon population would have $a^- = a^+ = 2$, while for a species like chinook salmon in some rivers a^- might be 3 and a^+ might be 5 or 6. Assume that accurate estimates of the census sizes of adults of different age classes are available over a specific time period beginning at $t = 0$ and ending at $t = T$. The census of a-year-old adults breeding at time t is denoted $C_{t,a}$. We shall regard these estimates as known without error. Additionally, we shall assume that the number of juveniles each year has been estimated, or can be specified (to within a rough approximation, at least) based on the number of adults giving rise to them. We denote the estimated juvenile

population size at time t by J_t. We will assume that this population behaves in genetic terms as if it were an ideal population governed by a parameter $\boldsymbol{\lambda}$ which can be construed as a vector having several components—one for each age group $(\lambda_{a-}, \ldots, \lambda_{a+})$ and one for the sampling from adult into juvenile or gamete stages, say $\lambda^{(w)}$. This will become more clear when we actually start assigning probability distributions in this model.

Furthermore, assume that genetic samples are available from the adult and juvenile populations at $t = 0$, $t = T$, and at least some (and preferably many) time points in between. It must be possible to determine the age of adults, so that the genetic samples can be regarded as drawn from adults of known ages. Adult ages can be determined from scales or otoliths taken from individuals. Likewise, when sampling juveniles we shall assume that it is possible to sample reliably from a single age class of juveniles, so that they are known to have descended from a particular brood year of adults. This is possible, in practice, because juveniles of many species of salmon will migrate to the ocean at a single, early age; thus, the juveniles in a stream in a given season will all be of a known age class. For species, like steelhead, in which the freshwater residence time of juveniles may vary widely from individual to individual, juvenile age, like adult age, can be determined from scales or otoliths as well.

The genetic samples involve typing individuals at L loci assumed to be independently segregating. In such a case, it is easy to combine data from the multiple loci, so I will describe the methodology in detail for a single locus only, and then later describe how to combine data from multiple loci. From this single locus, let K alleles be observed in the genetic samples from adults and juveniles. $S_{t,a}$ denotes the sample size of adults of age a taken at time t, and $\boldsymbol{Y}_{t,a} = (Y_{t,a,1}, \ldots, Y_{t,a,K})$ is a vector of allele counts for the K different alleles observed in the sample of a-year-olds at time t. Likewise, we denote sample sizes from juveniles at time t by R_t, and the observed numbers of alleles from a sample of juveniles at time t by the K-vector, $\boldsymbol{Z}_t = (Z_{t,1}, \ldots, Z_{t,K})$.

The unobservable, or *latent* variables in this model are the allele counts in the adults of different ages at each of the times t, $\boldsymbol{X}_{t,a} = (X_{t,a,1}, \ldots, X_{t,a,K})$, and the allele counts amongst the juveniles at the different times t, $\boldsymbol{W}_t = (W_{t,1}, \ldots, W_{t,K})$. Note that the sum of the K components of $\boldsymbol{X}_{t,a}$ is $2C_{t,a}$, and the sum of the components of $\boldsymbol{W}_t$ is $2J_t$.

Before specifying probability distributions for the observed genetic samples and for the transitions between the latent variables, it is helpful to simply consider the conditional dependence structure between the variables, given the overlapping year-class nature of the population's reproduction. We will first investigate this dependence structure under the assumption that individuals sampled from amongst the adults are not then precluded from reproducing themselves. This corresponds to Sampling Scheme I of NEI and TAJIMA (1981) (without the restriction that the census size is equal to the effective size of the population). This sort of sampling would be realized if non-invasive genetic sampling (*e.g.*, fin clips) was used, or if adults were sampled destructively *after* spawning. I will consider Sampling Scheme II in Section 4.2.2.

Figure 4.1 shows an acyclic directed graph for a hypothetical population in which $T = 7$, $a^- = 2$, and $a^+ = 4$. In this graph, the arrows may be taken to represent a temporally-defined dependence. That is, $c \longrightarrow d$ may be read to mean "c is a variable that 'occurs' before d in time, and upon which the distribution of d depends."[1] The form of the graph thus follows exactly from what we know about reproduction in a population of Pacific salmon from which we sample both juveniles and adults. The shape of the graph also admits a simple factorization of the joint probability of all the variables involved. To express this succinctly, the following notation will be useful: let the set of relevant times and ages be denoted $\mathcal{T} = \{(t,a) : 0 \leq t \leq T, a^- \leq a \leq a^+\}$. The set of times and ages which are "initial points" are those for which we must posit a prior distrubution for adult allele counts over which we will integrate. This set is $\mathcal{P} = \{(t,a) \in \mathcal{T} : t - a < 0\}$, and we will use the shorthand $\mathbf{X}_{\mathcal{P}}$ to refer to the latent allele counts in adults of those ages and times. In the graph of Figure 4.1, the elements of $\mathbf{X}_{\mathcal{P}}$ are surrounded by dotted circles. We will refer to the set of pairs, (t,a) which are not in $\mathcal{P}$ as being in the set $\mathcal{P}^c = \{(t,a) \in \mathcal{T} : t - a \geq 0\}$. We shall denote by $\mathcal{S}_{\mathbf{Y}} = \{(t,a) \in \mathcal{T} : S_{t,a} > 0\}$ the set of times and ages for which we have drawn genetic samples from the adults. Similarly, the set of all times for which a genetic sample from the juveniles has been taken will be

[1]The variable c is said to be a "parent" of d, and variable d is called a "child" of variable c. This terminology will be used later in the context of moralizing directed graphs to find neighborhoods of variables.

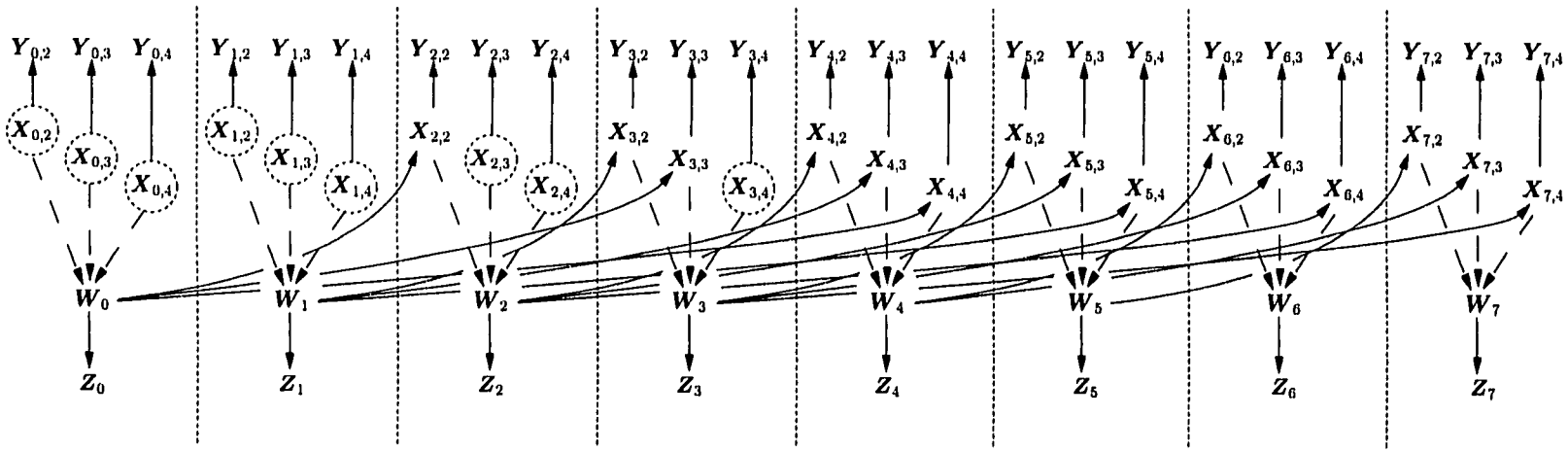


Figure 4.1: Acyclic directed graph describing the conditional dependence structure in the urn model with overlapping generations. Adults reproduce at ages 2, 3, and 4. The vertical, dotted lines separate time into different spawning years. The dotted circles represent latent variables in the set $\mathbf{X}_{\mathcal{P}}$ (see text).

denoted by $\mathcal{R}_{\mathbf{Z}} = \{t : 0 \leq t \leq T, R_t > 0\}$. And finally let the bold roman versions of each variable refer to sets of variables as follows: $\mathbf{Y} = \{\boldsymbol{Y}_{t,a} : (t,a) \in \mathcal{S}_{\mathbf{Y}}\}$, $\mathbf{Z} = \{\boldsymbol{Z}_t : t \in \mathcal{R}_{\mathbf{Z}}\}$, $\mathbf{X} = \{\boldsymbol{X}_{t,a} : (t,a) \in \mathcal{T}\}$, and $\mathbf{W} = \{\boldsymbol{W}_t : 0 \leq t \leq T\}$.

The joint probability of the observed and latent variables may then be written as

$$\begin{aligned} P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W}) &= P_{\boldsymbol{\lambda}}(\mathbf{X}_{\mathcal{P}}) \\ &\times \prod_{(t,a)\in\mathcal{S}_{\mathbf{Y}}} P(\boldsymbol{Y}_{t,a}|\boldsymbol{X}_{t,a}) \times \prod_{t\in\mathcal{R}_{\mathbf{Z}}} P(\boldsymbol{Z}_t|\boldsymbol{W}_t) \\ &\times \prod_{(t,a)\in\mathcal{P}^c} P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t,a}|\boldsymbol{W}_{t-a}) \times \prod_{0\leq t\leq T} P_{\boldsymbol{\lambda}}(\boldsymbol{W}_t|\boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+}) \end{aligned} \tag{4.1}$$

where $P(\cdot|\cdot)$ denotes a conditional probability distribution function not depending on $\boldsymbol{\lambda}$, $P_{\boldsymbol{\lambda}}(\cdot|\cdot)$ a conditional distribution depending on $\boldsymbol{\lambda}$ and $P_{\boldsymbol{\lambda}}(\mathbf{X}_{\mathcal{P}})$ is the prior probability of $\mathbf{X}_{\mathcal{P}}$, which also depends on $\boldsymbol{\lambda}$. This prior distribution, $P_{\boldsymbol{\lambda}}(\mathbf{X}_{\mathcal{P}})$, must necessarily be a joint distribution on the components of $\mathbf{X}_{\mathcal{P}}$, since we expect that those components will be dependent. I will treat this in more detail in Section 4.2.5, but for now we take the joint prior distribution as given. The two terms on the second line of (4.1) are the probabilities of the observed allele counts in all the samples of adults and juveniles, respectively. The two terms on the third line of the equation are 1) the probabilities due to population-genetic sampling of the latent allele counts in the adult groups given the juvenile cohorts to which they belonged, and 2) the probability of the latent allele counts amongst a juvenile cohort given all the adult age classes contributing to it.

4.2.2 Dependence structure under Sampling Scheme II and with null alleles

The dependence structure described in the previous section applies to many situations, but one may encounter other cases which require extensions to that basic dependence structure. Here I will deal with two such cases: 1) that when the genetic sampling is destructive and occurs before reproduction, so that individuals which are sampled do not have the opportunity to contribute offspring to the following years, and 2) the case of alleles that are not codominantly expressed.

NEI and TAJIMA (1981) used the name "Sampling scheme II" for the case when the

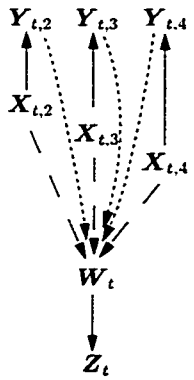


Figure 4.2: Acyclic directed graph describing the conditional dependence structure in the probability model for overlapping generations at year t, with three age classes of adults (2,3,4), under Sampling Scheme II—sampling adults destructively and before reproduction. The arrows connecting these variables to other times are omitted in this figure.

census size is larger than the effective size, and the genetic samples are destructively obtained before the organism is able to reproduce. WAPLES (1989) showed that the two sampling schemes could be handled within the same general F-statistic framework, with only a slight difference in the formulae for converting estimates of F to estimates of N_e. In our case, using a probability model derived from the urn model of the previous chapter, if the census size of the population is known, then the two different sampling plans can be treated using the different probability distributions that they give rise to.

The dependence structure of Figure 4.1 applies to Sampling Scheme I. For Scheme II, the dependence structure is different. Because the sampling is destructive, the gene copies sampled are not available to contribute gametes to the gamete pool. Hence, W_t will depend upon both $X_{t,a}$ *and* $Y_{t,a}$ for $a^- \leq a \leq a^+$. Figure 4.2 shows the dependence structure between the variables in a year t under Sampling Scheme II. The arrows between years are not shown, though they occur in the same places and directions as in Figure 4.1. Note the inclusion of the arrows (shown with finely dotted lines) from the samples to the gamete pool. This implies a modification of (4.1), changing the last factor to be

$$\prod_{0 \leq t \leq T} P_{\boldsymbol{\lambda}}(\boldsymbol{W}_t | \boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+}, \boldsymbol{Y}_{t,a^-}, \ldots, \boldsymbol{Y}_{t,a^+}). \tag{4.2}$$

It would not be difficult to modify the dependence structure further to account for the destructive genetic sampling of juveniles. However, I do not pursue that here, assuming instead that the gamete/juvenile pool from which samples are drawn is large enough that the effect of removing a sample of juveniles has little impact on the allele frequencies which will occur in the spawning populations of mature organisms.

Another complication which is frequently encountered is the occurrence of alleles that are not codominantly expressed. In this case, it is often possible to detect homozygotes of a particular allele, but the heterozygotes appear to be homozygotes of an alternate allele. This adds another layer of complexity to the model. The reason for this is that, when some alleles cannot be reliably detected in heterozygote form, it is not possible to actually observe allele counts $\boldsymbol{Y}_{t,a}$ in samples taken from the adults. Instead one observes only the counts of phenotypes (heterozygotes and apparent homozygotes) of different types, which I shall denote by $\mathbf{G}_{t,a}^{(Y,o)}$ for $(t,a) \in \mathcal{S}_{\mathbf{Y}}$. The superscript $^{(Y,o)}$ refers to the fact that these are the observed phenotypes in the sample from adults. Similarly, the samples from juveniles permit only the observation of phenotypes which will be denoted by $\mathbf{G}_{t}^{(Z,o)}$, $t \in \mathcal{R}_{\mathbf{Z}}$. Part of $\mathbf{G}_{t,a}^{(Y,o)}$ and $\mathbf{G}_{t}^{(Z,o)}$ should be thought of as symmetrical matrices with $(i,j)^{\text{th}}$ element equal to $(j,i)^{\text{th}}$ element and giving the number of observed phenotypes with a copy of allele i and a copy of allele j (i, j codominant). One additional category of phenotypes must be included in both $\mathbf{G}_{t,a}^{(Y,o)}$ and $\mathbf{G}_{t}^{(Z,o)}$. For this we use $G_{t,a,-}^{(Y,o)}$ and $G_{t,-}^{(Z,o)}$, to denote the number of individuals in the samples from adults and juveniles, respectively, in which no bands on a gel were detected. For example, if only allele i at a locus was undetectable, then for the sample from juveniles at time t, $G_{t,i,j}^{(Z,o)}$ would be zero, $G_{t,j,j}^{(Z,o)}$ would be the sum of the number of (j,j) genotypes and the number of heterozygotes of i and j, and $G_{t,-}^{(Z,o)}$ would be the number of (i,i) homozygotes.

Computing the probability of $\mathbf{G}_{t,a}^{(Y,o)}$ given $\boldsymbol{X}_{t,a}$ or $\mathbf{G}_{t}^{(Z,o)}$ given $\boldsymbol{W}_t$ would require a sum over all possible unobserved genotypes consistent with the observed phenotypes. To avoid having to do that sum directly, we will introduce more latent variables and effectively sum over them using MCMC. This also greatly simplifies the joint probability function in the case of Sampling Scheme II in the presence of null alleles. The new latent variables are $\mathbf{G}_{t,a}^{(Y,\ell)}$ and $\mathbf{G}_{t}^{(Z,\ell)}$, which are analogous to the symmetrical matrix portions of $\mathbf{G}_{t,a}^{(Y,o)}$ and

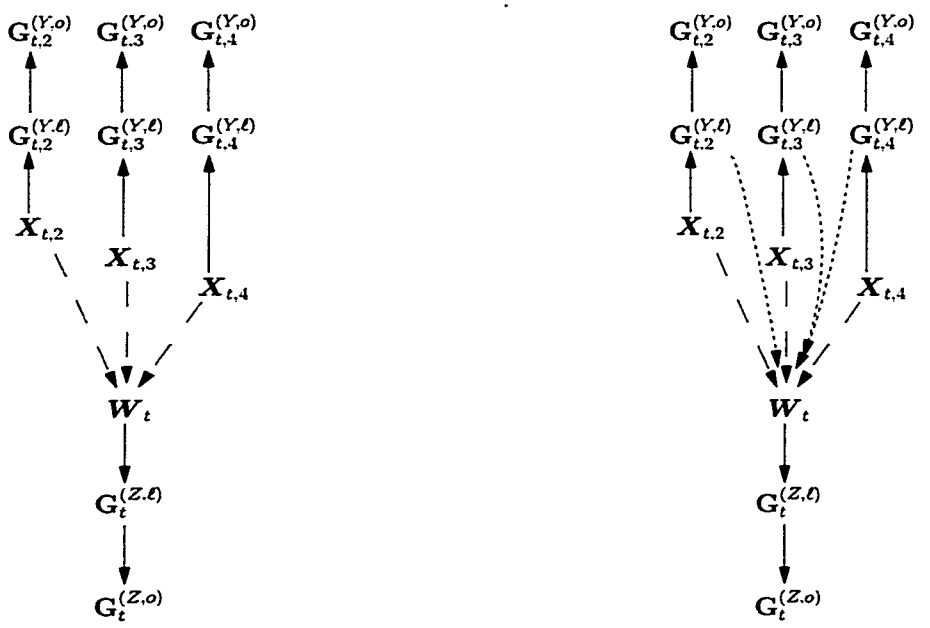


Figure 4.3: Acyclic directed graphs describing the conditional dependence structure in the probability model for overlapping generations at year t with three age classes (2,3,4) and with some alleles not codominantly expressed. The arrows connecting these graphs to other times are omitted in this figure. (a) Sampling Scheme I, sampled adults still contribute offspring to future generations (b) Sampling Scheme II, adults sampled destructively.

$\mathbf{G}_t^{(Z,o)}$, except that they count the number of different types of genotypes that would be observed if all the alleles were fully penetrant. Note that there is a many-to-one map from the space of $\mathbf{G}_{t,a}^{(Y,\ell)}$ to that of $\mathbf{G}_{t,a}^{(Y,o)}$, and similarly from the space of $\mathbf{G}_t^{(Z,\ell)}$ to $\mathbf{G}_t^{(Y,o)}$.

So long as the genetic transmission processes we consider are exchangeable, the dependence structure between these new variables within a year is given by the graph of Figure 4.3(a) for Sampling Scheme I and Figure 4.3(b) for Sampling Scheme II. The joint distribution of all the variables involved can then be written similarly to (4.1). Using the notation $\mathbf{G}^{(Y,o)} = \{\mathbf{G}_{t,a}^{(Y,o)} : (t,a) \in \mathcal{S}_\mathbf{Y}\}$ and $\mathbf{G}^{(Z,o)} = \{\mathbf{G}_t^{(Z,o)} : t \in \mathcal{R}_\mathbf{Z}\}$ along with

$\mathbf{G}^{(Y,\ell)} = \{\mathbf{G}^{(Y,\ell)}_{t,a} : (t,a) \in \mathcal{S}_\mathbf{Y}\}$ and $\mathbf{G}^{(Z,\ell)} = \{\mathbf{G}^{(Z,\ell)}_{t} : t \in \mathcal{R}_\mathbf{Z}\}$, we have

$$P_{\boldsymbol{\lambda}}(\mathbf{G}^{(Y,o)}, \mathbf{G}^{(Z,o)}, \mathbf{G}^{(Y,\ell)}, \mathbf{G}^{(Z,\ell)}, \mathbf{X}, \mathbf{W}) = \tag{4.3}$$

$$\begin{aligned} P_{\boldsymbol{\lambda}}(\mathbf{X}_\mathcal{P}) \quad & \times \prod_{(t,a)\in\mathcal{S}_\mathbf{Y}} P(\mathbf{G}^{(Y,o)}_{t,a} | \mathbf{G}^{(Y,\ell)}_{t,a}) \times \prod_{t\in\mathcal{R}_\mathbf{Z}} P(\mathbf{G}^{(Z,o)}_{t} | \mathbf{G}^{(Z,\ell)}_{t}) \\ & \times \prod_{(t,a)\in\mathcal{S}_\mathbf{Y}} P(\mathbf{G}^{(Y,\ell)}_{t,a} | \boldsymbol{X}_{t,a}) \times \prod_{t\in\mathcal{R}_\mathbf{Z}} P(\mathbf{G}^{(Z,\ell)}_{t} | \boldsymbol{W}_t) \\ & \times \prod_{(t,a)\in\mathcal{P}^c} P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t,a} | \boldsymbol{W}_{t-a}) \times \prod_{0\le t\le T} P_{\boldsymbol{\lambda}}(\boldsymbol{W}_t | \boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+}) \end{aligned}$$

for Sampling Scheme 1. For Sampling Scheme II, the final term in the product must be replaced by

$$\prod_{0\le t\le T} P_{\boldsymbol{\lambda}}(\boldsymbol{W}_t | \boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+}, \mathbf{G}^{(Y,\ell)}_{t,a^-}, \ldots, \mathbf{G}^{(Y,\ell)}_{t,a^+}).$$

Specification of the probability functions specific to sampling with recessive alleles is deferred until Section 4.2.4.

4.2.3 Specifying probability distributions

The graph of Figure 4.1 and the corresponding factorization of Equation 4.1 (as well as their extensions for the special cases described above) indicate that the probability model here may be fully defined by assigning distributions to $P_{\boldsymbol{\lambda}}(\mathbf{X}_\mathcal{P})$ and the different $P_{\boldsymbol{\lambda}}(\cdot|\cdot)$ and $P(\cdot|\cdot)$ distributions. Specifying these distributions requires several assumptions to be made about how reproduction occurs. In general, I shall model population-genetic sampling by Pólya urn models, and the drawing of genetic samples by sampling without replacement from the populations. In this context, sampling "without replacement" is *not* referring to whether or not sampled individuals are able to reproduce; it is referring to how the genetic samples are obtained. Certainly, destructive genetic sampling will occur without replacement, but even non-invasive sampling could occur without replacement since any previously-sampled fish will bear marks (for example the loss of a fin clipped for genetic sampling) that should prevent it from being sampled twice. For the samples taken from a large pool of juveniles, there will be little difference between the multivariate hypergeometric sampling implied

by sampling without replacement and the multinomial sampling implied by sampling with replacement.

The simplest transition to model is that described by $P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t,a}|\boldsymbol{W}_{t-a})$. This is the population-genetic sampling which occurs when juveniles from time $t-a$ are "selected" or "sampled" to survive to be reproducing adults of age a at time t. This depends on λ_a, and, following Equation 3.17 on Page 63, may be parametrized in terms of a stochastic replacement quantity $\varphi_{t,a}$ which depends on the juvenile census size, J_{t-a}, the adult census size, $C_{t,a}$, and λ_a:

$$\varphi_{t,a} = \frac{2J_{t-a}(1-\lambda_a)}{2\lambda_a C_{t,a} - 1}. \tag{4.4}$$

Thus, given $\boldsymbol{W}_{t-a}$, $\boldsymbol{X}_{t,a}$ follows the compound multinomial distribution (3.3). The probability mass function may be expressed, similarly to (3.3), as a normalizing constant times a product of K terms corresponding to the K different alleles:

$$\begin{aligned} P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t,a}|\boldsymbol{W}_{t-a}) &= P(\boldsymbol{X}_{t,a}|\boldsymbol{W}_{t-a}, \lambda_a, C_{t,a}, J_{t-a}) \\ &= P(\boldsymbol{X}_{t,a}|\boldsymbol{W}_{t-a}, \varphi_{t,a}, C_{t,a}) \\ &= \frac{(2C_{t,a})!\,\Gamma(\alpha_\bullet)}{\Gamma(2C_{t,a}+\alpha_\bullet)} \prod_{i=1}^{K} \left(\frac{\Gamma(X_{t,a,i}+\alpha_i)}{X_{t,a,i}!\,\Gamma(\alpha_i)} \right) \end{aligned} \tag{4.5}$$

where $\alpha_i = W_{t-a,i}/\varphi_{t,a}$ and $\alpha_\bullet = \sum_{i=1}^{K} \alpha_i$.

Modeling the stochastic process and distribution for $P_{\boldsymbol{\lambda}}(\boldsymbol{W}_t|\boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+})$ is more difficult, and requires that more assumptions be made about reproduction and survival in the population. The particular problem that arises is that the distribution of $\boldsymbol{W}_t$ depends not only on the vagaries of sampling alleles from within each age class of adults (*i.e.*, non-multinomial sampling of gene copies from amongst the $C_{t,a}$ a year-olds), but also on the fact that adults of different age classes may produce different mean numbers of juvenile offspring, either by producing more gametes or by producing individuals with higher survival to the juvenile stage. This second effect is akin to that discussed in RYMAN and LAIKRE (1991), in which the inbreeding effective size of a population is decreased due to the higher survivorship of a segment of the population included in a supportive breeding program. Since it is impossible to determine the age of the parent of any gene copy sampled amongst

juveniles, these two sources of variation in $\boldsymbol{W}_t$ are confounded and may not be separated. Rather than include these two confounded processes in a model which is not identifiable, I assume an ideal model for the production of juveniles from adults of different age classes, and then account for both of the above-mentioned processes by a single parameter in an urn model scheme.

This ideal model assumes that each adult at time t produces an age-specific number of gametes, and the survivors to the juvenile stage are sampled from those gametes by an urn scheme with stochastic replacement parameter ψ_t. More specifically, each diploid adult of age a contributes γ_a copies of each of its two gene copies to the gamete pool. Thus, the counts of the different alleles in the gamete pool at time t are given by the K-vector $\boldsymbol{B}_t = (B_{t,1}, \ldots, B_{t,K}) = \sum_{a=a^-}^{a^+} \gamma_a \boldsymbol{X}_{t,a}$. Then, the $2J_t$ gene copies in the juveniles are sampled from this gamete pool via a Pólya urn scheme in which the stochastic replacement quantity depends on the parameter $\lambda^{(w)}$—the conceptual ratio of "effective juveniles" to the census number of juveniles. Letting $B_{t,\bullet}$ denote the total number of gametes in the gamete pool at time t ($B_{t,\bullet} = \sum_{a=a^-}^{a^+} 2\gamma_a C_{t,a} = \sum_{i=1}^{K} B_{t,i}$), then, once again by Equation 3.17 on Page 63, we have the stochastic replacement

$$\psi_t = \frac{B_{t,\bullet}(1 - \lambda^{(w)})}{2\lambda^{(w)} J_t - 1}. \tag{4.6}$$

And so, the conditional probability $P_{\boldsymbol{\lambda}}(\boldsymbol{W}_t | \boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+})$ may now be expressed as

$$\begin{aligned} P_{\boldsymbol{\lambda}}(\boldsymbol{W}_t | \boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+}) &= P(\boldsymbol{W}_t | \boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+}, C_{t,a^-}, \ldots, C_{t,a^+}, \boldsymbol{\gamma}, J_t) \\ &= P(\boldsymbol{W}_t | \boldsymbol{B}_t, \psi_t, J_t) \\ &= \frac{(2J_t)!\,\Gamma(\alpha_\bullet)}{\Gamma(2J_t + \alpha_\bullet)} \prod_{i=1}^{K} \left(\frac{\Gamma(W_{t,i} + \alpha_i)}{W_{t,i}!\,\Gamma(\alpha_i)} \right) \end{aligned} \tag{4.7}$$

where $\alpha_i = B_{t,i}/\psi_t$ and $\alpha_\bullet = \sum_{i=1}^{K} \alpha_i$.

The quantities $\boldsymbol{\gamma} = (\gamma_{a^-}, \ldots, \gamma_{a^+})$ may be interpreted as fitness measures for different age classes expressing how successful they are at producing juveniles of sampling age. In practice, γ_a can be chosen to reflect the biology of the situation. For example, a reasonable choice for salmon would be one half the fecundity of age a females. It should be clear from

the above expression, that the absolute magnitudes of the γ_a's are actually irrelevant; the parametrization of ψ_t in terms of J_t and the relationship between α_i, $B_{t,i}$, and ψ_t ensure that the relative sizes of the γ_a's are all that matter. Nonetheless, it is computationally convenient to think of the γ_a's in terms of the number of gametes produced.

Another, and a possibly more elegant, interpretation of this population-genetic sampling scheme for juveniles is provided by the conditional branching process model of KARLIN and MCGREGOR (1965) with negative binomial distributions of offspring number (see Section 3.4). In this interpretation, the total number of juvenile gene copies is fixed to be $2J_t$, however the distribution of the number of copies of each gene within an age a adult appearing among the juveniles is exchangeably negative binomial with arbitrary (but equal for all genes) scale parameter β, and shape parameter γ_a/ψ_t. By such an interpretation it is perhaps even more clear that ψ_t, the stochastic replacement quantity for reproduction into juveniles at time t, represents both non-Wright-Fisher sampling within age classes, but also a departure from our best guess as biologists as to the fitnesses/fecundities of adults of different age classes. Since both the non-Wright-Fisher sampling within age classes, and the unkown differential survival between age classes reduce effective size of populations, and will therefore affect λ, it seems quite reasonable that both are accounted for in the parameter λ_t.

In the case of Sampling Scheme II, in which adults are destructively sampled before reproduction, defining the probability function $P_\lambda(\boldsymbol{W}_t|\boldsymbol{X}_{t,a^-},\ldots,\boldsymbol{X}_{t,a^+},\boldsymbol{Y}_{t,a^-},\ldots,\boldsymbol{Y}_{t,a^+})$ requires only a simple modification to the above mechanism for transmission of genes to juveniles. Since sampled adults do not contribute to future generations, we need merely define $B_{t,i}$ so as to reflect that. Namely, $B_{t,i} = \sum_{a=a^-}^{a^+} \gamma_a(X_{t,a,i} - Y_{t,a,i})$, and $B_{t,\bullet}$ must be modified accordingly: $(B_{t,\bullet} = \sum_{a=a^-}^{a^+} 2\gamma_a(C_{t,a} - S_{t,a}) = \sum_{i=1}^{K} B_{t,i})$. For Sampling Scheme II in the presence of null alleles, $Y_{t,a,i}$ in the immediately preceding sentence may be replaced by the quantity $Y_{t,a,i}^{(\ell)}$ described in the next section.

Finally, we only have to specify probability distributions for the genetic samples drawn from adults and juveniles, $P(\boldsymbol{Y}_{t,a}|\boldsymbol{X}_{t,a})$ and $P(\boldsymbol{Z}_t|\boldsymbol{W}_t)$. As stated at the beginning of this section, sampling without replacement is a good model for the acquisition of genetic samples. With multiple alleles, this leads to the multivariate hypergeometric distribution (see

JOHNSON *et al.* 1997, Chapter 39). This distribution may also be written as a normalizing constant multiplied by a product of K terms, one for each of the K alleles. So, for the genetic samples taken from adults we have

$$P(\boldsymbol{Y}_{t,a}|\boldsymbol{X}_{t,a}) = \frac{(2C_{t,a}-2S_{t,a})!(2S_{t,a})!}{(2C_{t,a})!}\prod_{i=1}^{K}\frac{X_{t,a,i}!}{(X_{t,a,i}-Y_{t,a,i})!Y_{t,a,i}!}. \tag{4.8}$$

For genetic samples taken from the juveniles, we can also use the multivariate hypergeometric distribution

$$P(\boldsymbol{Z}_t|\boldsymbol{W}_t) = \frac{(2J_t-2R_t)!(2R_t)!}{(2J_t)!}\prod_{i=1}^{K}\frac{W_{t,i}!}{(W_{t,i}-Z_{t,i})!W_{t,i}!}, \tag{4.9}$$

or, since the number of juveniles is typically large, modeling the process as sampling with replacement will yield essentially the same result, and so the multinomial probability distribution is appropriate:

$$P(\boldsymbol{Z}_t|\boldsymbol{W}_t) = (2R_t)!\prod_{i=1}^{K}\frac{[W_{t,i}/(2J_t)]^{Z_{t,i}}}{Z_{t,i}!}. \tag{4.10}$$

Notice that (4.10) also includes a simple product of terms over alleles.

4.2.4 Probabilities with recessive alleles

For recessive or null alleles at a locus with K alleles, I assume Hardy-Weinberg equilibrium and a simple penetrance model which may be summarized by the matrix $\mathbf{A}$ having elements $a_{i,j}$, $1 \le i,j \le K$. $a_{i,j} = 0$ implies that an allele of type i is detectable (*i.e.*, leaves a band on a gel) when it occurs in the same individual as an allele of type j. If i subscripts a null allele, than $a_{i,i} = 0$ and also $a_{i,j} = 0$ for all other j. This penetrance model can also account for other simple dominance relationships between alleles (*e.g.*, $a_{i,j} = 0$ but $a_{i,i} = 1$).

Given the latent genotypes of sampled juveniles, $\mathbf{G}_t^{(Z,\ell)}$, the observed phenotypes can be found by $G_{t,i,i}^{(Z,o)} = a_{i,i}G_{t,i,i}^{(Z,\ell)} + \sum_{j\neq i} a_{i,j}|a_{j,i}-1|G_{t,i,j}^{(Z,\ell)}$ and, for $j \neq i$, by $G_{t,i,j}^{(Z,o)} = a_{i,j}a_{j,i}G_{t,i,j}^{(Z,\ell)}$. The number of individuals showing no bands, $G_{t,-}^{(Z,o)}$, is found by subtraction, being half the number of gene copies not otherwise accounted for. Since there is a deterministic map from $\mathbf{G}_{t,a}^{(Y,\ell)}$ to $\mathbf{G}_{t,a}^{(Y,o)}$, $P(\mathbf{G}_{t,a}^{(Y,o)}|\mathbf{G}_{t,a}^{(Y,\ell)})$ will take the value one whenever $\mathbf{G}_{t,a}^{(Y,\ell)}$ is consistent with $\mathbf{G}_{t,a}^{(Y,o)}$ and zero otherwise. The map from $\mathbf{G}_{t,a}^{(Y,\ell)}$ to $\mathbf{G}_{t,a}^{(Y,o)}$ works similarly. Notice also

that the allele counts in the sample may be easily obtained from $\mathbf{G}_{t,a}^{(Y,\ell)}$ or $\mathbf{G}_{t}^{(Z,\ell)}$. We will denote these as $\boldsymbol{Y}_{t,a}^{(\ell)}$ and $\boldsymbol{Z}_{t}^{(\ell)}$, respectively.

Deriving the probability distribution $P(\mathbf{G}_{t,a}^{(Y,\ell)}|\boldsymbol{X}_{t,a})$ under the assumption of sampling without replacement requires some combinatoric calculations. Since the fates of gene copies in the genetic transmission and sampling models adopted here are exchangeable, the probability of every ordering of gene copies into the adults in the population, and therefore into the sampled adults from the population, is the same. Therefore $P(\mathbf{G}_{t,a}^{(Y,\ell)}|\boldsymbol{X}_{t,a})$ may be found by counting the ways of drawing particular combinations of pairs of genes, $\mathbf{G}_{t,a}^{(Y,\ell)}$, from the allele counts in the adults $\mathbf{X}_{t,a}$, and dividing by the total number of ways of drawing any $S_{t,a}$ pairs from the population. I show this below, suppressing the $_{t,a}$ subscript and $^{(Y,\ell)}$ superscript on elements of $\mathbf{G}_{t,a}^{(Y,\ell)}$, the $_{t,a}$ subscript on elements of $\boldsymbol{X}_{t,a}$ and on the population and sample sizes $C_{t,a}$ and $S_{t,a}$, and the $^{(\ell)}$ superscript and $_{t,a}$ subscript on elements of $\boldsymbol{Y}_{t,a}^{(\ell)}$.

First, the denominator of the probability $P(\mathbf{G}_{t,a}^{(Y,\ell)}|\boldsymbol{X}_{t,a})$ is the number of ways of drawing an unordered collection of S unordered pairs from a population of $2C$ gene copies, which is

$$\frac{1}{S!}\prod_{i=0}^{S-1}\binom{2C-2i}{2} = \frac{(2C)!}{2^S S!(2C-2S)!}. \tag{4.11}$$

The product of binomial coefficients arises from sequentially choosing unordered pairs without replacement, and the $1/(S!)$ accounts for the different orders in which those pairs may be drawn.

The numerator of $P(\mathbf{G}_{t,a}^{(Y,\ell)}|\boldsymbol{X}_{t,a})$ may be written as

$$\prod_{i=1}^{K}\left(\binom{X_i}{Y_i}\cdot\frac{Y_i!}{(2G_{i,i})!\prod_{j\neq i}G_{i,j}!}\cdot\frac{(2G_{i,i})!}{2^{G_{i,i}}G_{i,i}!}\cdot\prod_{j<i}G_{i,j}!\right) \tag{4.12}$$

and explained as follows: we have a product over alleles of four factors; the first factor is a binomial coefficient that counts the number of ways of choosing Y_i gene copies of type i from a population having X_i such gene copies. The second factor is a multinomial coefficient which counts the ways of partitioning those Y_i gene copies into the groups of genes participating in the different categories of genotypes. The third factor counts the number of ways $2G_{i,i}$ gene copies of allelic type i can be paired up into $G_{i,i}$ unordered

homozygous genotypes (this is a special case of (4.11)). And, finally, $G_{i,j}!$ is the number of ways of making $G_{i,j}$ heterozygote genotypes from $G_{i,j}$ copies of alleles of type i and $G_{i,j}$ copies of alleles of type j. The product of $G_{i,j}!$ is taken over $j < i$ since, in combination with the other product from $i = 1$ to K, this leads to the product over all heterozygote classes.

Equation 4.12 simplifies modestly so we may write our desired probability as

$$P(\mathbf{G}_{t,a}^{(Y,\ell)}|\boldsymbol{X}_{t,a}) = \frac{\prod_{i=1}^{K}\left(\frac{X_i!}{(X_i-Y_i)!}\cdot\frac{1}{2^{G_{i,i}}\prod_{j=1}^{K}G_{i,j}!}\cdot\prod_{j<i}G_{i,j}!\right)}{\frac{(2C)!}{2^S S!(2C-2S)!}}. \tag{4.13}$$

This is (4.12) divided by (4.11). The same is true for $P(\mathbf{G}_t^{(Z,\ell)}|\boldsymbol{W}_t)$ using J and R, and with Z's replacing Y's and W's replacing X's, under the assumption that sampling from juveniles is done without replacement. Under the assumption that sampling from juveniles is done with replacement, $P(\mathbf{G}_t^{(Z,\ell)}|\boldsymbol{W}_t)$ is a simple expression given by a multinomial distribution with cell probabilities being the genotype frequencies expected under Hardy-Weinberg equilibrium.

4.2.5 The prior distribution for allele counts

The prior distribution $P(\mathbf{X}_\mathcal{P})$ presents some interesting difficulties. Ideally, we would like to use some sort of stationary distribution of allele counts $\mathbf{X}_\mathcal{P}$ for the salmon population under study. However, this is difficult, first, because with fluctuating sizes, the population allele frequencies won't strictly have a stationary distribution, and second, because even if we knew the historical sizes of the population, it would not be straightforward to determine the distribution of $\mathbf{X}_\mathcal{P}$. Below, I present, in series, several different ways of handling the prior, $P_\lambda(\mathbf{X}_\mathcal{P})$, starting with the most naive. In practice, some combination of the methods described below will probably work best. The choice of which to use is a matter of balance between reflecting the reality of the situation and imposing too much (and possibly incorrect) structure on the latent variables, which will affect the inferences made.

The most naive approach would be to use independent priors for the components of $\mathbf{X}_\mathcal{P}$. Independent, uniform priors, for example, would assert very little *a priori* structure on the model. This is naive because some of the components of $\mathbf{X}_\mathcal{P}$ reflect fish that have matured

from the same pool of juveniles. Clearly, allele frequencies amongst adults matured from the same cohort of juveniles will be correlated. Fortunately, if a large sample has been taken from every time and age in $\mathcal{P}$, then the choice of prior may have little effect since those data (let us call them $\mathbf{Y}_{\mathcal{P}}$) will constrain $\mathbf{X}_{\mathcal{P}}$ considerably.

An improvement on the above can be made by adding to the model the allele counts in juvenile pools contributing to the adult populations of $\mathcal{P}$. For example, extending the graph in Figure 4.1 to include juvenile pools in "negative time," we could have the variable $\boldsymbol{W}_{-1}$, from which $\boldsymbol{X}_{1,2}$, $\boldsymbol{X}_{2,3}$, and $\boldsymbol{X}_{3,4}$ are drawn; $\boldsymbol{W}_{-2}$ parental to $\boldsymbol{X}_{1,3}$ and $\boldsymbol{X}_{2,4}$ in the graph; and $\boldsymbol{W}_{-3}$ parental to $\boldsymbol{X}_{1,4}$. Then, even with independent prior distribution on $\boldsymbol{W}_{-3}$, $\boldsymbol{W}_{-2}$, and $\boldsymbol{W}_{-1}$, the correlation between $\boldsymbol{X}_{1,2}$, $\boldsymbol{X}_{2,3}$, and $\boldsymbol{X}_{3,4}$ would be modeled, as well as that between the other elements of $\boldsymbol{X}_{\mathcal{P}}$. In general, this approach requires specifying a^+ new variables, $\boldsymbol{W}_{-a^+}, \ldots, \boldsymbol{W}_{-1}$, and giving them independent priors. While this is a great improvement over the first approach, it still does not account for the correlation that is bound to exist between the allele counts in the juvenile pools in "negative time." An *ad hoc* approach to doing so is described in the following method.

An approximate relationship between the variables $(\boldsymbol{W}_{-a^+}, \ldots, \boldsymbol{W}_{-1})$ can be derived using the work of WAPLES (1990b) which explores the expected F-statistics between the allele frequencies corresponding to $\boldsymbol{W}_{-a^+}$ and the remaining components, as a function of the effective number of breeders N_b and the proportion of fish maturing at different ages. Consider a salmon population progressing through time with effective numbers of spawners N_b, possibly changing each year, and with $\boldsymbol{f} = (f_{a^-}, \ldots, f_{a^+})$ being a vector of proportions giving the probability that a fish matures at a particular age. Through computer simulation of such a population, WAPLES (1990b) found a linear relationship between the expected value of F computed from allele frequencies in the gamete pools separated by t years and the inverse of twice the harmonic mean effective number of breeders $(\tilde{N}_b)$ in the t years between the gamete pools considered. He also showed that the slope of this linear relationship depends on t and the proportions $\boldsymbol{f}$. We will denote this slope by the function $\Delta_t(\boldsymbol{f})$. TAJIMA (1992) gives a convenient recursive algorithm for computing $\Delta_t(\boldsymbol{f})$. In our case, denoting the allele *frequency* in a juvenile or gamete pool at time i by p_i, WAPLES

(1990b) empirically shows that

$$\frac{\mathbb{E}[(p_{-a^+} - p_i)^2]}{p_{-a^+}(1 - p_{-a^+})} \approx \frac{\Delta_{i+a^+}(\boldsymbol{f})}{2\bar{N}_b} \tag{4.14}$$

for $i = -a^+ + 1, \ldots, -2, -1$. If we assume that the juvenile/gamete pools in these years are all of the same size, $J^{(-)}$ diploids (the superscript $^{(-)}$ refers to these being in "negative time"), and the expected value of each of $\boldsymbol{W}_{-a^+ +1}, \ldots, \boldsymbol{W}_{-1}$ is $\boldsymbol{W}_{-a^+}$, then, by (4.14), we have for allele j at time i

$$\begin{aligned} \mathrm{Var}(W_{i,j} | W_{a^+,j}) &= \mathbb{E}[(2J^{(-)})^2 (p_{-a^+} - p_i)^2] \\ &\approx (2J^{(-)}) p_{-a^+}(1 - p_{-a^+}) \left(\frac{(2J^{(-)})\Delta_{i+a^+}(\boldsymbol{f})}{2\bar{N}_b} \right). \end{aligned} \tag{4.15}$$

This is the variance of a binomial random variable with $2J^{(-)}$ trials and success probability p_{a^+}, multiplied by the term in the large parentheses. That, in turn, is the form of the variance of a beta binomial random variable. From the discussion in Section 3.3 (Page 55) of the relationship between the variance of beta-binomial and binomial random variables, it may be seen that a distribution satisfying the variance relationship in (4.15) is the beta-binomial distribution with $2J^{(-)}$ trials and parameters α_j and $\alpha_\bullet - \alpha_j$ such that $a_j/\alpha_\bullet = p_{-a^+}$ and

$$\frac{2J^{(-)} + \alpha_\bullet}{1 + \alpha_\bullet} = \frac{(2J^{(-)})\Delta_{i+a^+}(\boldsymbol{f})}{2\bar{N}_b}. \tag{4.16}$$

This suggests that the following would be a reasonable way to construct a prior for the vector $(\boldsymbol{W}_{-a^+}, \ldots, \boldsymbol{W}_{-1})$:

1. Assume reasonable values for the proportions of individuals maturing at different ages, $\boldsymbol{f} = (f_{a^-}, \ldots, f_{a^+})$.

2. Let $\boldsymbol{W}_{-a^+}$ follow a discrete uniform prior (since $J^{(-)}$ does not change in the MCMC simulations, this term in the distribution will conveniently never change, either).

3. Given $\boldsymbol{W}_{-a^+}$, assume that $\boldsymbol{W}_i$ $(i = -a^+ + 1, \ldots, -1)$ are drawn from the gene copies in the juvenile pool at time $-a^+$ via independent Pólya urn schemes with stochastic

replacement quantities $\psi_i^{(-)} = 2J^{(-)}/\alpha_{\bullet_i}$, where $\alpha_{\bullet_i}$ is calculated according to (4.16):

$$\alpha_{\bullet_i} = \frac{2\tilde{N}_b - \Delta_{i+a^+}(\boldsymbol{f})}{\Delta_{i+a^+}(\boldsymbol{f}) - \tilde{N}_b/J^{(-)}}. \tag{4.17}$$

It can be shown that conditional on $\boldsymbol{W}_{-a^+}$ such a distribution will have the variance of (4.15). Although it will not properly reflect the covariance between the gamete pools, it should be a very reasonable approximation. In practice, of course, the harmonic mean effective number of breeders will be unknown, but one should be able to make a reasonable estimate at the harmonic mean census number of spawners of all age groups in the a^+ years before data started being recorded for the population. Denoting that quantity as $\tilde{C}^{(-)}$, a simple way of estimating $\tilde{N}_b$ given $\tilde{C}^{(-)}$ and $\boldsymbol{\lambda}$ is $\tilde{N}_b = \lambda^{(-)}\tilde{C}^{(-)}$ where $\lambda^{(-)} = \sum_{a=a^-}^{a^+} f_a\lambda_a$. This is the way in which the prior distribution depends on $\boldsymbol{\lambda}$.

Finally, if census sizes (or estimates thereof) of the different aged fish in the population are known in the years before the genetic data started being collected, that information can similarly be used to help define a prior distribution for $\mathbf{X}_{\mathcal{P}}$. Doing so is simple—one merely defines time 0 to be the time at which the first census size data are available. Then, everything from the previous several paragraphs still applies for constructing a prior on initial gamete pools, but one also has several years of census data over which $\boldsymbol{X}_{t,a}$'s and $\boldsymbol{W}_t$'s may be sampled in an MCMC sampler, helping to more accurately reflect the joint distribution of $\boldsymbol{X}_{t,a}$'s when genetic samples are finally taken.

In concluding this section, I point out that while the *ad hoc* approach described above is reasonable and practical, it is not deeply satisfying. The derivation of an elegant prior $P_{\boldsymbol{\lambda}}(\mathbf{X}_{\mathcal{P}})$ remains an interesting, open problem.

4.3 A Bayesian Formulation and MCMC Simulation from $P(\boldsymbol{\lambda}|\mathbf{X},\mathbf{W})$

A Bayesian formulation of this problem is obtained by assigning a prior distribution $P(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}$. This leads to the posterior distribution

$$P(\boldsymbol{\lambda}|\mathbf{Y},\mathbf{Z}) = \frac{P(\boldsymbol{\lambda})\sum_{\mathbf{X},\mathbf{W}} P_{\boldsymbol{\lambda}}(\mathbf{Y},\mathbf{Z},\mathbf{X},\mathbf{W})}{\int_{\boldsymbol{\lambda}} P(\boldsymbol{\lambda})\sum_{\mathbf{X},\mathbf{W}} P_{\boldsymbol{\lambda}}(\mathbf{Y},\mathbf{Z},\mathbf{X},\mathbf{W})d\boldsymbol{\lambda}} \tag{4.18}$$

where the integral in the denominator is over all values of $\boldsymbol{\lambda}$ and the sum is over all possible values of $\mathbf{X}$ and $\mathbf{W}$. This sum and integral are intractable. However, it is possible to simulate

values of $\boldsymbol{\lambda}$ from this posterior distribution using the Metropolis-Hastings algorithm, and thus, the posterior distribution may be evaluated by Markov chain Monte Carlo. This is presented in overview in the following two paragraphs, and in detail in the remainder of the section.

Given current values of $\mathbf{X}$, $\mathbf{W}$, and $\boldsymbol{\lambda}$, a Metropolis-Hastings update step for $\mathbf{X}$ involves simulating a new value $\mathbf{X}'$ from a proposal distribution $q_{\mathbf{X}}(\mathbf{X}'|\mathbf{X}, \cdots)$ that depends on $\mathbf{X}$, and possibly on the current values of other variables in the model (denoted by "$\cdots$"). A uniform random variable on the unit interval U is then drawn. If $U < H_{\mathbf{X}}$ then the proposal is accepted and the value of $\mathbf{X}$ is changed to $\mathbf{X}'$. If $U > H_{\mathbf{X}}$, then the value of $\mathbf{X}$ remains unchanged. If

$$H_{\mathbf{X}} = \frac{q_{\mathbf{X}}(\mathbf{X}|\mathbf{X}', \cdots) P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}', \mathbf{W})}{q_{\mathbf{X}}(\mathbf{X}'|\mathbf{X}, \cdots) P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W})}, \tag{4.19}$$

then, if $q_{\mathbf{X}}$ is such that successively applying the updates using U and $H_{\mathbf{X}}$ above leads to an irreducible Markov chain of $\mathbf{X}$ values, that Markov chain will have limit distribution $P_{\boldsymbol{\lambda}}(\mathbf{X}|\mathbf{Y}, \mathbf{Z}, \mathbf{W})$. Similarly, updates to $\mathbf{W}$ can be made by proposing new values $\mathbf{W}'$ from $q_{\mathbf{W}}(\mathbf{W}'|\mathbf{W}, \cdots)$, drawing U and accepting the proposal if U is less than

$$H_{\mathbf{W}} = \frac{q_{\mathbf{W}}(\mathbf{W}|\mathbf{W}', \cdots) P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W}')}{q_{\mathbf{W}}(\mathbf{W}'|\mathbf{W}, \cdots) P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W})}. \tag{4.20}$$

In the same way, updates to $\boldsymbol{\lambda}$ are made with a proposal distribution $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}'|\boldsymbol{\lambda}, \cdots)$ and accepted according to the Hastings ratio

$$H_{\boldsymbol{\lambda}} = \frac{q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}|\boldsymbol{\lambda}', \cdots) P(\boldsymbol{\lambda}') P_{\boldsymbol{\lambda}'}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W})}{q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}'|\boldsymbol{\lambda}, \cdots) P(\boldsymbol{\lambda}) P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W})}. \tag{4.21}$$

Applying these updates in series (update $\mathbf{X}$, update $\mathbf{W}$, update $\boldsymbol{\lambda}$, update $\mathbf{X}$, update $\mathbf{W}$, and so on...) leads to a Markov chain with limit distribution $P(\boldsymbol{\lambda}, \mathbf{X}, \mathbf{W}|\mathbf{Y}, \mathbf{Z})$. Sampling n values of $\boldsymbol{\lambda}$ visited by this chain gives a sequence $\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(n)}$ which may be used to estimate $P(\boldsymbol{\lambda}|\mathbf{Y}, \mathbf{Z})$ by Monte Carlo. The following three sections provide greater detail on the calculations involved. Section 4.3.1 shows how to exploit the conditional dependence structure of the graph in Figure 4.1 to simplify the calculation of Hastings ratios for $\mathbf{X}'$ and $\mathbf{W}'$. Then Section 4.3.2 gives a prescription for the proposal distributions $q_{\mathbf{X}}$ and $q_{\mathbf{W}}$. Finally, in Section 4.3.3, proposal distributions for $\boldsymbol{\lambda}$ are considered, and a Rao-Blackwellized estimator for $P(\boldsymbol{\lambda}|\mathbf{Y}, \mathbf{Z})$ is given.

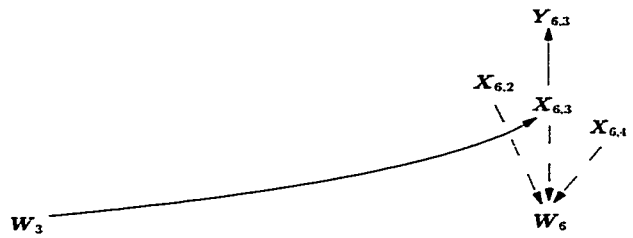
4.3.1 Neighborhood structures and joint probability ratios

Making updates to $\mathbf{X}$ and $\mathbf{W}$ requires repeated calculation of the Hastings ratios (4.19) and (4.20). This task is made easy by proposing changes only to small parts (two components, for example) of either $\mathbf{X}$ or $\mathbf{W}$ at any one time. The neighborhood structure inherent in the graph of Figure 4.1 and the fact that the probabilities described above can all be written in terms of a product over the K alleles make this particularly attractive, as described below.

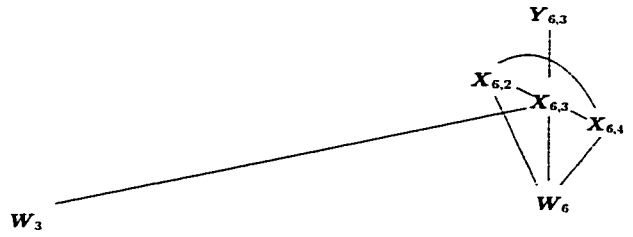
Let the data from the genetic samples be considered fixed at $\mathbf{Y}$ and $\mathbf{Z}$, and suppose the current values for $\mathbf{X}$ and $\mathbf{W}$ are denoted by $\mathbf{X}$ and $\mathbf{W}$, respectively. Let $\mathbf{X}'$ and $\mathbf{W}'$ differ from $\mathbf{X}$ and $\mathbf{W}$ only at an arbitrary, single component subscripted by $(t', a') \in \mathcal{T}$ for $\mathbf{X}'$ and by $t' \in \{0, \ldots, T\}$ for $\mathbf{W}'$. In doing MCMC we will make frequent use of the ratios

$$\frac{P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}', \mathbf{W})}{P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W})} \quad \text{and} \quad \frac{P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W}')}{P_{\boldsymbol{\lambda}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W})}. \tag{4.22}$$

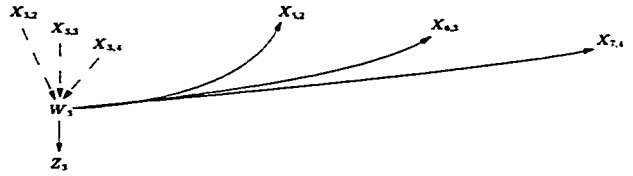
Calculating such ratios is done quickly by noting that they are functions only of a small collection of variables adjacent in the graph to the altered component. The variables adjacent to the altered component in the graph are members of its neighborhood, and the factors in the joint density including those neighbors are the only ones that are changed by the alteration in that component. Hence, the other factors cancel out in the ratio. The neighborhoods can be graphically found and represented via the moralized, undirected graph associated with the directed graph (LAURITZEN 1996). The moralized subgraph around $\boldsymbol{X}_{t',a'}$ (or $\boldsymbol{W}_{t'}$) is formed by starting with the subgraph containing all variables which are either connected to $\boldsymbol{X}_{t',a'}$ (or $\boldsymbol{W}_{t'}$) by arrows in either direction or which are parents of any children of $\boldsymbol{X}_{t',a'}$ (or $\boldsymbol{W}_{t'}$), and then converting all the arrows between those variables to undirected edges and moralizing the subgraph. Moralizing is done by including edges between any unconnected parents in the directed graph. Directed and moralized versions of the subgraphs around $\boldsymbol{X}_{6,3}$ and $\boldsymbol{W}_3$ from Figure 5.1 are shown in Figure 4.4. The corresponding distribution associated with each undirected graph in the figure may be factorized by their cliques (maximally connected subgraphs). Therefore the ratios in (4.22) may be written as ratios of terms corresponding to the cliques. Using the notation $\mathbf{X}_{\{t',\setminus a'\}}$ to refer



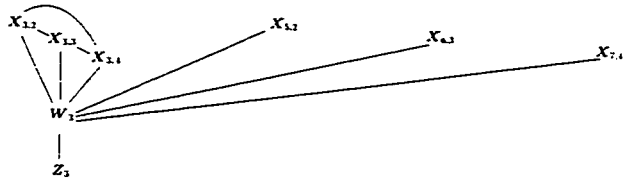
(a)



(b)



(c)



(d)

Figure 4.4: Neighborhoods for the allele count amongst the juveniles and adults. (a) and (b) are respectively the directed and the moralized, undirected subgraphs for the relevant neighborhood in $\mathbf{X}'$ with $t' = 6$ and $a' = 3$. (c) and (d) are the same for the neighborhood around $\boldsymbol{W}_3$ (*i.e.*, $t' = 3$).

to the set $\{\boldsymbol{X}_{t',a^-},\ldots,\boldsymbol{X}_{t',a^+}\}$, *excluding* $\boldsymbol{X}_{t',a'}$, we have

$$\frac{P_{\boldsymbol{\lambda}}(\mathbf{Y},\mathbf{Z},\mathbf{X}',\mathbf{W})}{P_{\boldsymbol{\lambda}}(\mathbf{Y},\mathbf{Z},\mathbf{X},\mathbf{W})} = \frac{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{X}'_{t',a'},\mathbf{X}_{\{t',\backslash a'\}})}{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{X}_{t',a^-},\ldots,\boldsymbol{X}_{t',a^+})} \times \frac{P(\boldsymbol{Y}_{t',a'}|\boldsymbol{X}'_{t',a'})}{P(\boldsymbol{Y}_{t',a'}|\boldsymbol{X}_{t',a'})} \times \frac{P_{\boldsymbol{\lambda}}(\boldsymbol{X}'_{t',a'}|\boldsymbol{W}_{t'-a'})}{P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t',a'}|\boldsymbol{W}_{t'-a'})} \tag{4.23}$$

for the ratio involving an altered version of $\mathbf{X}$. Note that if $(t',a') \in \mathcal{P}$ then a term corresponding to the prior $P(\mathbf{X}_{\mathcal{P}})$ would also appear in the ratio. For the ratio involving the altered version of $\mathbf{W}$ we have

$$\frac{P_{\boldsymbol{\lambda}}(\mathbf{Y},\mathbf{Z},\mathbf{X},\mathbf{W}')}{P_{\boldsymbol{\lambda}}(\mathbf{Y},\mathbf{Z},\mathbf{X},\mathbf{W})} = \frac{P_{\boldsymbol{\lambda}}(\boldsymbol{W}'_{t'}|\boldsymbol{X}_{t',a^-},\ldots,\boldsymbol{X}_{t',a^+})}{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{X}_{t',a^-},\ldots,\boldsymbol{X}_{t',a^+})} \times \frac{P(\boldsymbol{Z}_{t'}|\boldsymbol{W}'_{t'})}{P(\boldsymbol{Z}_{t'}|\boldsymbol{W}_{t'})} \times \prod_{(t,a)\in\mathcal{W}_{t'}} \frac{P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t,a}|\boldsymbol{W}'_{t'})}{P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t',a'}|\boldsymbol{W}_{t'})} \tag{4.24}$$

where $\mathcal{W}_{t'}$ represents the times and ages of adults descended from the juvenile pool at time t'. That is, $\mathcal{W}_{t'} = \{(t,a) \in \mathcal{T} : t - a = t'\}$.

Let us now make the further restriction that $\mathbf{X}'$ differs from $\mathbf{X}$ only in two components of $\mathbf{X}'_{t',a'}$. That is to say $X'_{t',a',i}$ and $X'_{t',a',j}$ can take any non-negative values so long as $X'_{t',a',i} + X'_{t',a',j} = X_{t',a',i} + X_{t',a',j}$. We shall make a similar restriction on $\mathbf{W}'$. In such a case, the probability ratios in (4.23) and (4.24) simplify further still, with the normalizing constants and the terms for unaltered allele counts cancelling out. Hence we have

$$\frac{P(\boldsymbol{Y}_{t',a'}|\boldsymbol{X}'_{t',a'})}{P(\boldsymbol{Y}_{t',a'}|\boldsymbol{X}_{t',a'})} = \frac{X'_{t',a',i}!(X_{t',a',i} - Y_{t',a',i})!}{X_{t',a',i}!(X'_{t',a',i} - Y_{t',a',i})!} \cdot \frac{X'_{t',a',j}!(X_{t',a',j} - Y_{t',a',j})!}{X_{t',a',j}!(X'_{t',a',j} - Y_{t',a',j})!} \tag{4.25}$$

when $(t',a') \in \mathcal{S}_{\mathbf{Y}}$ and 1 otherwise. A similar expression applies to $\frac{P(\boldsymbol{Z}_t|\boldsymbol{W}'_t)}{P(\boldsymbol{Z}_t|\boldsymbol{W}_t)}$ for sampling without replacement from juveniles. For sampling with replacement from juveniles we have

$$\frac{P(\boldsymbol{Z}_{t'}|\boldsymbol{W}'_{t'})}{P(\boldsymbol{Z}_{t'}|\boldsymbol{W}_{t'})} = \left(\frac{W'_{t',i}}{W_{t',i}}\right)^{Z_{t',i}} \left(\frac{W'_{t',j}}{W_{t',j}}\right)^{Z_{t',j}} \tag{4.26}$$

for $t' \in \mathcal{R}_{\mathbf{Z}}$, and 1 otherwise. For the terms having to do with population genetic sampling into the adult stage, we have

$$\frac{P_{\boldsymbol{\lambda}}(\boldsymbol{X}'_{t',a'}|\boldsymbol{W}_{t'-a'})}{P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t',a'}|\boldsymbol{W}_{t'-a'})} = \frac{\Gamma(X'_{t',a',i} + W_{t'-a',i}/\varphi_{t',a'})X_{t',a',i}!}{\Gamma(X_{t',a',i} + W_{t'-a',i}/\varphi_{t',a'})X'_{t',a',i}!} \cdot \frac{\Gamma(X'_{t',a',j} + W_{t'-a',j}/\varphi_{t',a'})X_{t',a',j}!}{\Gamma(X_{t',a',j} + W_{t'-a',i}/\varphi_{t',a'})X'_{t',a',j}!} \tag{4.27}$$

and

$$\frac{P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t'+a,a}|\boldsymbol{W}'_{t'})}{P_{\boldsymbol{\lambda}}(\boldsymbol{X}_{t'+a,a}|\boldsymbol{W}_{t'})} = \frac{\Gamma(X_{t'+a,a,i}+W'_{t',i}/\varphi_{t'+a,a})\Gamma(W_{t',i}/\varphi_{t'+a,a})}{\Gamma(X_{t'+a,a,i}+W_{t',i}/\varphi_{t'+a,a})\Gamma(W'_{t',i}/\varphi_{t'+a,a})} \times \frac{\Gamma(X_{t'+a,a,j}+W'_{t',j}/\varphi_{t'+a,a})\Gamma(W_{t',j}/\varphi_{t'+a,a})}{\Gamma(X_{t'+a,a,j}+W_{t',j}/\varphi_{t'+a,a})\Gamma(W'_{t',j}/\varphi_{t'+a,a})}. \tag{4.28}$$

For the terms having to do with population sampling into the juvenile stage we compute the two relevant ratios using the quantity $\boldsymbol{B}_t$ (defined on Page 86 in Section 4.2.3) and its altered version $\boldsymbol{B}'_t$ when necessary. Thus we have

$$\begin{aligned}\frac{P_{\boldsymbol{\lambda}}(\boldsymbol{W}'_{t'}|\boldsymbol{X}_{t',a^-},\ldots,\boldsymbol{X}_{t',a^+})}{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{X}_{t',a^-},\ldots,\boldsymbol{X}_{t',a^+})} &= \frac{P_{\boldsymbol{\lambda}}(\boldsymbol{W}'_{t'}|\boldsymbol{B}_{t'})}{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{B}_{t'})} \\ &= \frac{\Gamma(W'_{t',i}+B_{t',i}/\psi_{t'})W_{t',i}!}{\Gamma(W_{t',i}+B_{t',i}/\psi_{t'})W'_{t',i}!}\cdot\frac{\Gamma(W'_{t',j}+B_{t',j}/\psi_{t'})W_{t',j}!}{\Gamma(W_{t',j}+B_{t',j}/\psi_{t'})W'_{t',j}!}\end{aligned} \tag{4.29}$$

and

$$\begin{aligned}\frac{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{X}'_{t',a'},\mathbf{X}_{\{t',\setminus a'\}})}{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{X}_{t',a^-},\ldots,\boldsymbol{X}_{t',a^+})} &= \frac{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{B}'_{t'})}{P_{\boldsymbol{\lambda}}(\boldsymbol{W}_{t'}|\boldsymbol{B}_{t'})} \\ &= \frac{\Gamma(W_{t',i}+B'_{t',i}/\psi_{t'})\Gamma(B_{t',i}/\psi_{t'})}{\Gamma(W_{t',i}+B_{t',i}/\psi_{t'})\Gamma(B'_{t',i}/\psi_{t'})} \\ &\quad\times\frac{\Gamma(W_{t',j}+B'_{t',j}/\psi_{t'})\Gamma(B_{t',j}/\psi_{t'})}{\Gamma(W_{t',j}+B_{t',j}/\psi_{t'})\Gamma(B'_{t',j}/\psi_{t'})}.\end{aligned} \tag{4.30}$$

Both of the above extend immediately to the case of Sampling Scheme II with $\boldsymbol{B}_t$ defined appropriately (*i.e.*, with the $\boldsymbol{Y}_{t,a}$'s subtracted out as on Page 87 in Section 4.2.3).

The derivation of the ratios of joint probabilities when non-penetrant alleles are present proceeds in similar fashion to the above treatment, but is omitted for brevity.

4.3.2 Proposal distributions for $\mathbf{X}'$ *and* $\mathbf{W}'$

In the preceding, we saw that it is advantageous to consider changes to pairs of alleles at a single time and age for $\mathbf{X}$ and a single time for $\mathbf{W}$. Consequently the proposal distribution $q_{\mathbf{X}}$ can be a function just of those components, and can be written $q_{\mathbf{X}}(X'_{t,a,i},X'_{t,a,j}|X_{t,a,i},X_{t,a,j},\ldots)$. Since $X'_{t,a,i}+X'_{t,a,j}$ must equal $X_{t,a,i}+X_{t,a,j}$, the proposal distribution is simply a distribution on $X'_{t,a,i}$, imposing, for uniqueness of reverse moves in this sampler, the condition $i<j$.

The proposal distribution $q_{\mathbf{X}}(X'_{t,a,i}|X_{t,a,i}, X_{t,a,j}, \cdots)$ should reflect a compromise between statistical efficiency and computational efficiency. From a statistical perspective, it is most efficient to simulate $X'_{t,a,i}, X'_{t,a,j}$ from their full conditional distribution. However, not much efficiency is gained this way, and calculating the full conditional distribution incurs a heavy computational cost. Instead, I define $q_{\mathbf{X}}$ to be a uniform distribution with width determined by the current values $X_{t,a,i}$ and $X_{t,a,j}$. That is, $X'_{t,a,i}$ is drawn from a uniform distribution on the integers (excluding the current value, $X_{t,a,i}$) between X_{lo} and X_{hi}, inclusive, where the values of X_{lo} and X_{hi} are chosen as a linear function of the approximate standard deviation of $X_{t,a,i}$ conditional only upon its parents in the graph. Namely $X_{\text{lo}} = X_{t,a,i} - w$ and $X_{\text{hi}} = X_{t,a,i} + w$ where w is the greatest integer less than or equal to

$$2\beta\left(\frac{L+\alpha}{1+\alpha}X_{t,a,i}(1 - X_{t,a,i}/L)\right)^{1/2} \tag{4.31}$$

where $L = X_{t,a,i} + X_{t,a,j}$, $\alpha = L/\varphi_{t,a}$, and β is a scaling factor that may be adjusted to achieve a desired acceptance proportion. It can be tuned automatically during run time if desired. The width of $q_{\mathbf{W}}$ may be tuned similarly.

It is also desirable to include some checking in the computer code to ensure that $q_{\mathbf{X}}$ does not give positive probability to any values of $X'_{t,a,i}$ which would be incompatible with the descendants of $X_{t,a,i}$ and $X_{t,a,j}$ in the graph.

4.3.3 Proposal distributions for $\boldsymbol{\lambda}$

To make updates to $\boldsymbol{\lambda}$, we consider changes to just one of its components at a time, λ_a for the discussion here. A naive, computationally simple proposal distribution for λ_a is less desirable than a full conditional update for λ_a, because the latter allows for a Rao-Blackwellized (see Section 1.5.3 on Page 19) Monte Carlo estimator of $P(\boldsymbol{\lambda}_a|\mathbf{Y}, \mathbf{Z})$. This does require that the parameter space for $\boldsymbol{\lambda}$ be discretized. This has little effect on the final inferences one can make if the discretization is fine enough. For example one could choose to consider n values for λ_a say, $\lambda_{a,0}, \lambda_{a,1} \ldots, \lambda_{a,n}$ where $\lambda_{a,i} = .02 * i$. For most situtations, this will be a fine enough discretization. Writing Λ_a for the set $\{\lambda_{a,0}, \ldots, \lambda_{a,n}\}$, and $\boldsymbol{\lambda}'$ for

$\boldsymbol{\lambda}$ with its a^{th} component set to λ'_a, I use for $q_{\boldsymbol{\lambda}}(\lambda'_a)$ the full conditional distribution

$$q_{\boldsymbol{\lambda}}(\lambda'_a|\cdots) = P(\lambda_a|\mathbf{Y},\mathbf{Z},\mathbf{X},\mathbf{W}) = P(\lambda'_a|\mathbf{X},\mathbf{W}) = \frac{P(\lambda')P_{\boldsymbol{\lambda}'}(\mathbf{X},\mathbf{W})}{\sum_{\lambda'_a\in\Lambda_a} P(\lambda')P_{\boldsymbol{\lambda}'}(\mathbf{X},\mathbf{W})} \tag{4.32}$$

For each update to λ_a, (4.32) must be computed for all $\lambda'_a \in \Lambda_a$. This is computationally expensive, but that is more than offset by the fact that at the i^{th} update, one is realizing the values $P(\lambda'_a|\mathbf{X}^{(i)},\mathbf{W}^{(i)})$ for each $\lambda'_a \in \Lambda_a$ with $(\mathbf{X}^{(i)},\mathbf{W}^{(i)})$ being simulated from their posterior distribution given $\mathbf{Y}$ and $\mathbf{Z}$. Therefore the successive values $P(\lambda'_a|\mathbf{X}^{(i)},\mathbf{W}^{(i)})$ may be averaged over the course of a run of the Markov chain to yield an efficient, Rao-Blackwellized estimate of $P(\lambda_a|\mathbf{Y},\mathbf{Z})$. Furthermore, since $q_{\boldsymbol{\lambda}}(\lambda_a)$ is the full conditional distribution for λ_a, the above scheme defines a Gibbs sampling proposal for λ_a, and the Hastings ratio $H_{\boldsymbol{\lambda}}$ reduces to unity, always.

In empirical tests, this method of updating $\boldsymbol{\lambda}$ takes more computational time, but yields far superior estimates of $P(\lambda_a|\mathbf{Y},\mathbf{Z})$ than a naive (*i.e.*, uniform) proposal distribution for each λ_a in fewer updates of the chain.

4.4 Special Cases

There are some situations in which it might be advantageous or imperative to consider a model which is simpler and has fewer parameters than the one just described. One obvious simplification would be to restrict the λ_a's of each age group to be equal. This would be appropriate if data were only available on juveniles, since in that case the different λ_a's would be unidentifiable. It might also be desirable if data are relatively sparse, and/or if one has prior reason to believe that λ_a's would not differ greatly over different age classes.

Similarly, it is possible to restrict $\mathbf{W}$ to match perfectly the allele frequencies implied by $\mathbf{X}$. This corresponds to the assumption that juvenile populations are very large and all of the population-genetic sampling that is *not* random with respect to an individual's family of origin occurs in the mortality between juvenile and adult stages. While this restriction would not allow the independent estimate of a λ for juveniles, it would be prudent in the case when data are available only on adults or only on juveniles. Any non-random (with respect to family) sampling that occurred before the juvenile stage would then be estimated as part of the population-genetic sampling from juvenile to adult.

4.5 Simulated Data

It is always the case that the process of formulating and describing this type of MCMC method is far simpler and takes less time than the cycle of implementation, testing, and debugging that is required to produce software to actually carry out the Markov chain simulations described. I have not yet been able to test all the parts of the current version of the software implementing the method described in the chapter. However, I am satisfied that some of its modules are functioning properly, and the results are sufficiently promising to present the method's performance on simulated data. For this purpose I have used the census data in Table 4.1, from the Inmaha Creek chinook salmon population. These data appear in BEAMSDERFER *et al.* (1998), and were kindly provided to me in electronic format by Robin Waples at the National Marine Fisheries Service. As in Chapter 2, the purpose of this demonstration is not to assess the bias and variance of the Bayesian estimator for λ derived in this chapter. Doing so would require a prohibitive amount of computing in order to average the results over a large number of simulated datasets. Rather, this section demonstrates that the MCMC method itself is able to provide a good approximation to the posterior probability for λ given a single dataset.

Inmaha Creek is a tributary of the Snake River in which the chinook salmon population has declined dramatically in the last four decades. Fish return at ages 3, 4, and 5. The 3-year-olds are almost all small males called "jacks." Census size estimates, broken down by age class, are available for the years 1954 to 1999. While jacks certainly contribute some to future generations, it is unlikely that the contribution, on a per-fish basis, is nearly as great as that of four and five year-olds. Further, since there are so few of them, and because their occasional zero census size estimates cause conflicts with the current version of my software, they were excluded from the dataset.

Genetic data were simulated for a single locus given these census sizes by initializing a juvenile pool in year 1949 with 5 alleles having counts in the proportions $(.4, .2, .2, .1, .1)$. Allele counts in the juvenile pool in years 1950 to 1953 were then considered to be $\boldsymbol{W}_{-4}$ to $\boldsymbol{W}_{-1}$ and were drawn according to the urn scheme describing the prior distribution (Section 4.2.5) with $\tilde{C}$ being 900. In other words, $\boldsymbol{W}_{-4}$ to $\boldsymbol{W}_{-1}$ were simulated by independent,

Table 4.1: Estimates of the number of chinook spawners returning to Inmaha Creek (a tributary on the Snake River drainage) in years 1954 to 1999. Age 3 fish are young males known as "jacks." (Data source: BEAMSDERFER *et al.* (1998))

Brood Year	Age 3	Age 4	Age 5	Brood Year	Age 3	Age 4	Age 5
1954	146	507	1079	1977	0	460	230
1955	232	1473	1638	1978	0	87	1914
1956	62	985	619	1979	13	113	124
1957	242	1438	1875	1980	10	87	95
1958	31	508	655	1981	24	214	236
1959	18	231	299	1982	32	279	307
1960	40	655	845	1983	23	206	226
1961	149	341	575	1984	17	219	321
1962	60	678	458	1985	0	363	278
1963	113	207	321	1986	43	214	235
1964	58	684	464	1987	0	139	262
1965	49	385	474	1988	13	92	411
1966	136	385	555	1989	18	85	49
1967	30	666	326	1990	0	70	14
1968	12	450	687	1991	12	36	34
1969	57	843	556	1992	3	58	16
1970	0	350	480	1993	4	81	282
1971	176	1039	529	1994	0	17	34
1972	20	364	1235	1995	3	26	28
1973	0	602	1905	1996	5	130	13
1974	0	590	711	1997	0	95	58
1975	0	139	579	1998	0	39	50
1976	0	306	284	1999	0	0	15

random draws from an urn containing alleles in the initial frequencies, (.4, .2, .2, .1, .1). The remaining latent variables $\mathbf{X}$ and $\mathbf{W}$ were simulated throughout the graph via the urn scheme described in this chapter, with $\lambda_4 = \lambda_5 = 0.4$, and using the age-specific fitnesses of $\gamma_4 = 450$ and $\gamma_5 = 650$. These values were obtained by using rough fecundity/length, length/age, and juvenile survivorship relationships for chinook salmon (both stream- and ocean-type combined) given in HEALEY (1991). The latent data were simulated under the assumption that no genetic drift occurs between the adult and the juvenile stage. Genetic data were not simulated from 1954 to 1963. However, genetic drift was simulated in the population during that interval. This allowed the allele frequencies between different years to settle closer to their joint stationary distribution before starting the simulated sample collection. Other simulations (Robin Waples, National Marine Fisheries Service, unpublished result) show that 20 years is sufficient to allow the alleles frequencies to "warm-up." For the purposes of the present demonstration, ten years should be sufficient.

From 1963 to 1988 I simulated datasets with samples of varying sizes drawn every year from the same simulated set of latent variables. The three different sample sizes considered were:

1. $S_4 = S_5 = 10$ and $R = 30$

2. $S_4 = S_5 = 25$ and $R = 60$

3. $S_4 = S_5 = 60$ and $R = 125$

In years when the sample size would have been larger than one half the census size of the population of a particular age (4 or 5), the sample size for that age group was decreased to be one half of the census size of the population. Data were not simulated and used for the last eleven years (1989–1999) of the census data because the small population sizes in those years meant that even with very small samples from the adult populations, a good estimate of $\boldsymbol{\lambda}$ was possible, and I wanted a more challenging scenario for demonstrating the method.

The simulated data were analyzed under the assumption that $\lambda_4 = \lambda_5$ (which shall hereafter be referred to as λ) and that no drift occurs in the transmission of genes to the

juvenile stages; hence $\boldsymbol{W}_t$ is completely determined by $(\boldsymbol{X}_{t,a^-}, \ldots, \boldsymbol{X}_{t,a^+})$. λ values in the set $\{.02, .04, \ldots, .98\}$ were considered. To reduce burn-in time, $\mathbf{X}$ was initialized to the value that was realized in the original simulation. I have subsequently verified that the burn-in time required for other reasonable starting configurations (like all allele frequencies of $\mathbf{X}$ initialized to the average frequency of the alleles observed in the samples) is short. A sweep of the algorithm consisted of:

1. E updates in series, first with a random pair $(X_{t,a,i}, X_{t,a,j})$ between the years 1963 and 1988 and then with a random pair $(W_{t,i}, W_{t,j})$ from the juvenile pools used to construct the prior distribution in the years 1958 to 1962.

2. An update of λ.

E was chosen so that each component of $\mathbf{X}$ was updated twice on average during a sweep. For the different scenarios I simulated, I performed 70,000 sweeps of the algorithm. Inspection of the estimated posterior for λ suggests that the estimate changed imperceptibly over the last 50,000 sweeps of the algorithm. 70,000 sweeps required 2 hours on a laptop computer with a 266 Mhz G3 (Macintosh) processor.

At each of the different sampling intensities I analyzed the data under the assumption that all the samples were available (Figure 4.5), and also under the assumption that only the adult samples were available (Figure 4.6(a)). I also did one simulation in which samples from the adults were not available, but samples of size $R = 125$ from juveniles at all years were available (Figure 4.6(b)). For comparison, I have plotted each of these posterior distributions next to the posterior distribution that one would obtain if $\mathbf{X}$ and $\mathbf{W}$ were known without error.

The results, as shown in Figures 4.5 and 4.6, suggest that the MCMC sampler is functioning appropriately and computing the posterior distribution for λ. The curvature generally decreases with sample size, reflecting the loss of information, as it should.[2] Furthermore,

[2]The posterior distribution for sample sizes $S = 25$, $R = 60$, being more peaked than the posterior distribution for $S = 60$, $R = 125$, is an exception to the trend. This results from the fact that for the particular set of data simulated for $S = 25$, $R = 60$, the estimated λ happens to be smaller than for the data simulated with $S = 60$, $R = 125$. The credible set will be smaller for a lower estimated value of λ

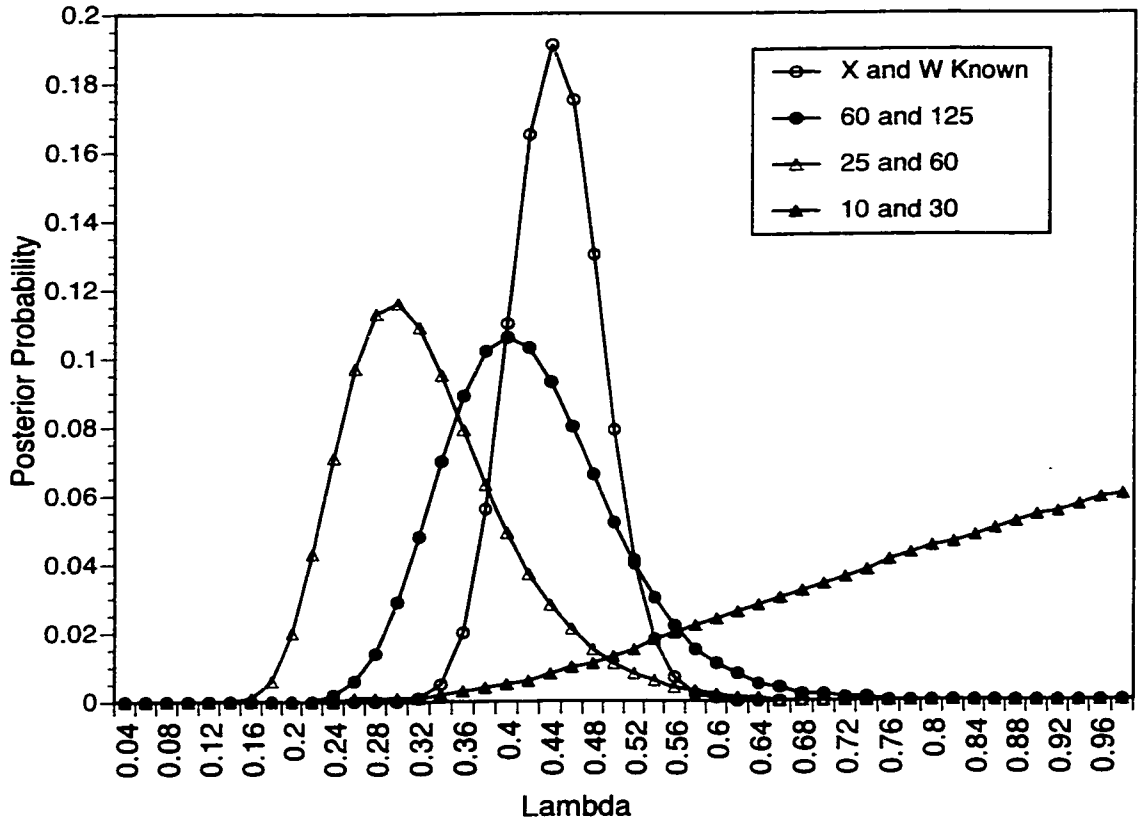
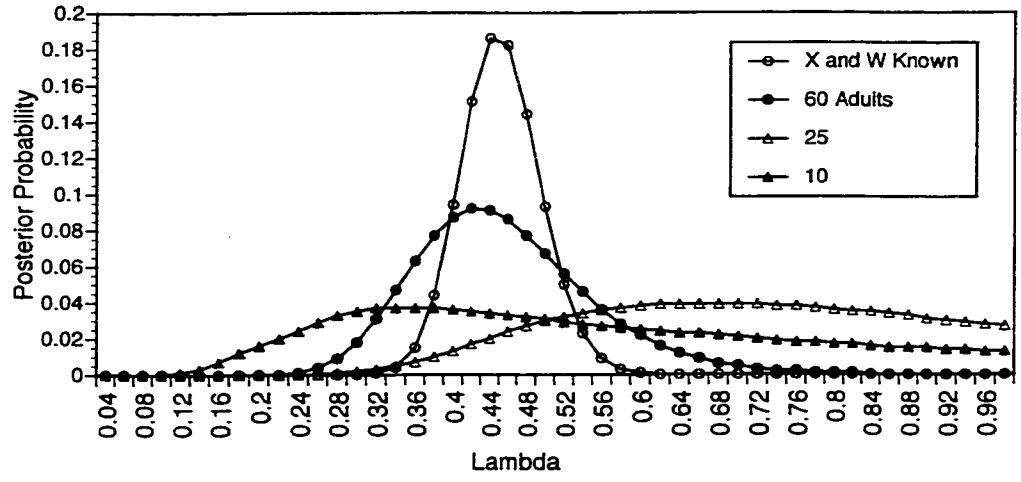
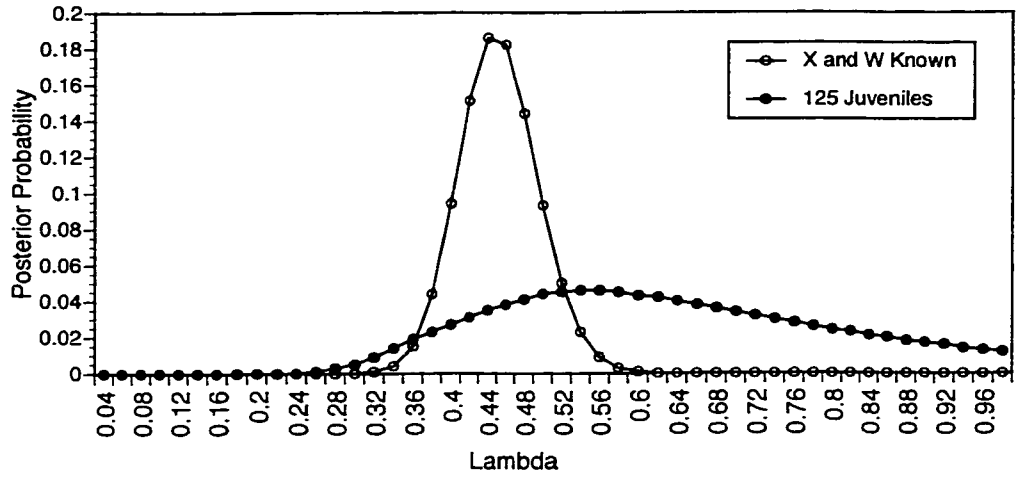


Figure 4.5: Plot of the posterior probabilities for $\lambda = \lambda_4 = \lambda_5$ from genetic data simulated on the Inmaha Creek chinook population (Table 4.1), when data were available on both adults and juveniles. The first line in the graph corresponds to the estimate with the latent variables known without error. The other three lines correspond to the different sample sizes of adults and juveniles. The true $\lambda_4 = \lambda_5 = 0.4$.



(a) Samples taken only from adults



(b) Samples taken only from juveniles

Figure 4.6: Graphs as described in Figure 4.5, but under different sampling scenarios. (a) Juvenile sample sizes all zero, and adult sample sizes as shown. (b) Adult sample sizes not all zero, but samples of 125 taken each year from the juveniles.

in all cases except one, the 90 percent credible interval for λ overlaps the true value of 0.4. While the narrowest posterior distributions occur with samples from both adults and juveniles, there still seems to be a substantial signal in the data, even when samples are taken either only from adults or only from juveniles.

4.6 Discussion

These results are encouraging. They demonstrate that the MCMC sampler devised here is able to compute the posterior probability distribution for λ, suggesting that the method presented in this chapter permits use of data over multiple years from a salmon population with known census sizes to estimate the ratio λ with good precision. It should be kept in mind that these simulations exploit the data from only a single locus with five alleles. Narrower credible sets would be obtained with data on multiple loci. The posterior distribution for $\boldsymbol{\lambda}$ given data on multiple, independently segregating loci is proportional to the product of the posterior probabilities for $\boldsymbol{\lambda}$ from each of the loci treated separately, as described here. Therefore, the extension to multiple loci is simple.

The method developed herein would be particularly appropriate for estimating $\boldsymbol{\lambda}$ in hatchery populations of salmon where the census sizes of spawning adults are well known. As in the previous chapter, the method thus far developed in the current chapter assumes that $\boldsymbol{\lambda}$ remains constant over time. Future work is required to assess how robust this estimate is to departures from the underlying model. However, like the method of Chapter 3, it would also be possible here to propose new models in which $\boldsymbol{\lambda}$ varied over time, and to compare those models within a Bayesian framework using reversible jump MCMC (GREEN 1995). Such a method would be well-suited to using genetic data to detect the impact of supportive breeding programs (RYMAN and LAIKRE 1991; HANSEN *et al.* 2000) on $\boldsymbol{\lambda}$ in salmon populations.

---

because, when λ is smaller, then the amount of genetic drift expected will be larger, relative to the amount of error due to random sampling of genes. In other simulations (not shown) in which the maximum *a posteriori* estimate of λ for the simulated data with $S = 25$, $R = 60$ was closer to that for the simulated data with $S = 60$, $R = 125$, the credible set is wider for the dataset with smaller samples.

Chapter 5

BAYESIAN INFERENCE IN MIXED AND ADMIXED POPULATIONS

5.1 Introduction

Populations studied by geneticists are seldom the ideal, randomly-mating and genetically-isolated collections of individuals for which much genetic theory has been developed. In particular, natural populations may possess internal structure which prevents random mating, or they may be recently formed by migration and co-mingling of individuals from two or more originally separate populations. Such structure complicates many types of genetic studies. For instance, when using population-level data to map genetic diseases, population structure, if not accounted for, may lead to spurious associations between genetic markers and disease status (EWENS and SPIELMAN 1995). Additionally, in the ecological study of plants and animals there is considerable interest in population structure, especially in regions of apparent overlap and interbreeding between different subpopulations—so-called "hybrid zones." For these, and other types of problems, it is desirable to be able to infer population structure from genetic data. To this end, models of population genetic *mixture* and *admixture* have been useful. I describe the application of such models to the inference of population structure, focusing primarily on applications to hybrid zones of two different groups of individuals. Such situations are now encountered frequently as a result of anthropogenic disturbance reducing barriers to gene flow between formerly separate subpopulations. Invasions of exotic species are one pervasive example.

As used here, a "genetic mixture model" attributes structure in a population to the presence of two or more subpopulations. Within these subpopulations individuals may mate at random, but no interbreeding occurs between subpopulations. Every individual in the mixture is considered to be a purebred descendant of only one subpopulation. Such models

have been developed and used extensively in the field of fisheries management (MILNER *et al.* 1981; PELLA and MILNER 1987; SMOUSE *et al.* 1990; MILLAR 1991; PELLA and MASUDA 2001).

On the other hand, admixture, throughout the genetics literature (CAVALLI-SFORZA and BODMER 1971; THOMPSON 1973; WIJSMAN 1984; LONG 1991) refers to interbreeding between members from different subpopulations. Accordingly, a "genetic admixture model" attempts to model a population's genetic structure by the presence of two or more previously separate subpopulations between which there has been some recent interbreeding. Such a population is said to be admixed. Additionally we will call an individual "admixed" if it possesses genes descended from more than one of the historically separate populations. Early investigations of admixed populations sought to estimate the relative contributions of two founding populations to the admixed population. These studies assumed the admixed population had undergone enough generations of random mating to eliminate the allelic associations between loci that accompanies genetic admixture. In such cases, the individual allele frequencies observed in the two founding subpopulations and the admixed population are the sufficient statistics. It was not until recently that statistical methods were proposed whereby the Hardy-Weinberg and linkage disequilibrium information captured in *multilocus* data could be used to elucidate structure in a recently admixed population (RANNALA and MOUNTAIN 1997; PAETKAU *et al.* 1995; PRITCHARD *et al.* 2000).

PRITCHARD *et al.* (2000) propose a versatile model for genetic inheritance in admixed populations and use it in Bayesian analyses of population structure in several different species. A limitation of this model, however, is that it assumes every individual is admixed to some degree. In many situations, such as with populations spanning hybrid zones, there is reason to expect both purebred and admixed individuals. A probability model to accommodate such scenarios will include elements both of genetic mixture models and genetic admixture models. In this chapter, I extend the methods of PRITCHARD *et al.* (2000) to handle explicitly purebred individuals.

In sections 5.2 and 5.3, I review mixture and admixture formulations for modeling population structure. In Section 5.3.1, I develop a method for making joint, Gibbs updates of large blocks of variables in the PRITCHARD *et al.* (2000) model. The method uses the

fact that the latent allocation variables of an i.i.d. finite mixture, with a Dirichlet prior on mixing proportions can be shown to follow a hidden Markov chain, after integrating out the mixing proportions. This computation facilitates MCMC simulation in a model, described in Section 5.4, that allows for both purebred and admixed individuals. Additionally, I describe in the Discussion how such a method could help the Gibbs sampler to escape from trapping states (ROBERT 1996) encountered in other finite mixture problems.

I apply these techniques to data on the Scottish wildcat *Felis sylvestris*. In Scotland, *F. sylvestris* evolved for thousands of years with little or no genetic exchange with cats in continental Europe. Within the last 2,000 years these Scottish cats have suffered population declines due to human influences and have been exposed to possible interbreeding with domestic cats. It can be difficult to distinguish *F. sylvestris* from domestic cats on the basis of morphological characters alone and conservation biologists are concerned that the wild-living cats in Scotland may now represent an admixture of *F. sylvestris* and domestic cats. These data were previously analyzed by BEAUMONT *et al.* (2001) using the method of PRITCHARD *et al.* (2000). However, this analysis does not address the issue of particular interest—that of estimating the proportion of purebred *F. sylvestris* individuals in the population. Nor does that analysis allow estimation of posterior probabilities that particular individuals in the sample are purebred cats. These questions about the Scottish wildcat population are similar to those for many species of conservation interest to which the present methods apply.

Finally, using reversible-jump MCMC, it is possible to compute the Bayes factor for comparing the new, expanded model to that of PRITCHARD *et al.* (2000) given the Scottish cat data. While the reversible-jump sampler allows estimation of the true Bayes factor, it is also possible to compute the "pseudo-Bayes factor" (GELFAND *et al.* 1992), and assess how accurately that estimates the Bayes factor.

5.2 Genetic Mixture Models

The formulation of a genetic mixture model follows that for a general finite mixture. Let N diploid individuals be sampled from a population and typed at L loci. The population

is assumed to consist of J subpopulations indexed by $j = 1, \ldots, J$. The proportion of individuals in the mixed population from subpopulation j is the unknown parameter π_j with $\sum_{j=1}^{J} \pi_j = 1$. Assign to each individual a latent allocation variable Z_i, $i = 1, \ldots, N$. We use $Z_i = j$ to indicate that the i^{th} individual is from the j^{th} subpopulation.

Denote the multilocus phenotype of the i^{th} individual by $\boldsymbol{Y}_i$. I use "phenotype" as opposed to "genotype" because the phrase "multilocus genotype" sometimes implies knowledge of the gametic phase of the alleles present at different loci. No such knowledge is available here. The multilocus phenotype, $\boldsymbol{Y}_i$, consists of the allelic type of each of the two alleles carried by the i^{th} individual at L loci. Since we will later identify and label specific gene copies in an individual, we consider the two alleles carried at a locus to be ordered. This order may be arbitrary. For example, it can merely be the order in which the types of those two alleles are reported in the data on an individual. Thus, $\boldsymbol{Y}_i$ can be regarded as a vector of length $2L$ with its first element giving the allelic type of the first allele at locus 1, its second element giving the type of the second allele at locus 1, its third element the type of the first allele at locus 2 and so forth. In general, $Y_{i,t}$, $t = 1, \ldots, 2L$, is the type of allele number (t [mod 2] + 1) at locus number $\lceil t/2 \rceil$ in individual i, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x and t [mod 2] is the remainder after dividing t by 2.

The allele frequencies in the j^{th} subpopulation are denoted by $\boldsymbol{\Theta}_j = (\boldsymbol{\theta}_{j,1}, \ldots, \boldsymbol{\theta}_{j,L})$, where $\boldsymbol{\theta}_{j,\ell}$, $\ell = 1, \ldots, L$, is a vector of length equal to K_ℓ—the number of distinct types of alleles observed at locus ℓ across all the individuals sampled. The frequency in the j^{th} subpopulation of the k^{th} allelic type of the ℓ^{th} locus is $\theta_{j,\ell,k}$, $k = 1, \ldots, K_\ell$. We will adopt the notation $\theta\langle j; Y_{i,t}\rangle$ to mean $\theta_{j,\ell,k}$ where $\ell = \lceil t/2 \rceil$ and k is the allelic type of the $(t \text{ [mod 2]} + 1)^{\text{th}}$ allele at the ℓ^{th} locus in the i^{th} individual.

Given that an individual is from subpopulation j, it is assumed to have a multilocus phenotype resulting from random mating and linkage equilibrium between the L loci *within* subpopulation j. Thus,

$$P(\boldsymbol{Y}_i | \boldsymbol{\Theta}_j, Z_i = j) = \prod_{t=1}^{2L} \theta\langle j; Y_{i,t}\rangle \tag{5.1}$$

where $P(\cdot|\cdot)$ will be used throughout to denote conditional probability mass or density

functions. The likelihood for $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ and $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_J)$, with $\mathbf{Y}$ denoting $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$, is

$$P(\mathbf{Y}|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \prod_{i=1}^{N} P(\boldsymbol{Y}_i|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \prod_{i=1}^{N} \sum_{j=1}^{J} \pi_j P(\boldsymbol{Y}_i|\boldsymbol{\Theta}_j, Z_i = j). \tag{5.2}$$

Note that the product in (5.1) does not include the familiar binomial coefficient, 2, for heterozygotes because we have arbitrarily ordered the two alleles carried by an individual at each locus. This makes the likelihood in this model comparable to that in the admixture model of PRITCHARD *et al.* (2000).

"Training" or "learning" samples may be available. They might take the form of specially tagged individuals which, though sampled along with the rest of the mixture, may be unambiguously assigned to a subpopulation. Such an individual, say i^*, known to come from subpopulation j^*, is easily accommodated by setting $z_{i^*} = j^*$ and defining $P(\boldsymbol{Y}_{i^*}|\boldsymbol{\Theta}_j, Z_{i^*} = j) \equiv 0$ for all $j \neq j^*$. However, if a learning sample from the j^{th} subpopulation is drawn separately (for example, if taken during a season when the subpopulations can be sampled separately) it contributes a term of the form $C \cdot \prod_{\ell=1}^{L} \prod_{k=1}^{K_\ell} \theta_{j,\ell,k}^{n_{j,\ell,k}}$ to the likelihood, where C is a product of multinomial coefficients and $n_{j,\ell,k}$ is the number of alleles of type k observed at locus ℓ in the learning sample taken separately from the j^{th} subpopulation. (In the Bayesian framework, these changes are equivalent to altering the prior for $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$ appropriately.)

Treating this mixture model from the Bayesian perspective requires prior distributions for $\boldsymbol{\pi}$ and $\boldsymbol{\Theta}$. The conjugate prior for π is the Dirichlet distribution, $\text{Dir}(\zeta_1, \ldots, \zeta_J)$. Prior information could be incorporated in the values of the ζ_j, or, if no prior information is available, the uniform distribution $\zeta_j = 1$, $j = 1, \ldots, J$, is a reasonable choice when J is not large. The conjugate prior for each $\boldsymbol{\Theta}_{j,\ell}$ is $\text{Dir}(\lambda_{j,\ell,1}, \ldots, \lambda_{j,\ell,K_\ell})$. In this chapter, I use uniform Dirichlet priors, $\lambda_{j,\ell,k} = 1\ \forall j, \ell, k$, which tend to de-emphasize the influence of rarely-occurring allelic types. This is a conservative assumption, and works well when the subpopulations are sufficiently genetically distinct. Note, however, that PRITCHARD *et al.* (2000) discuss different Dirichlet priors and PELLA and MASUDA (2001), who independently developed the same Bayesian scheme for genetic mixture analysis, describe other approaches to assigning allele frequency priors in mixed-stock fishery problems with closely-related

subpopulations.

With the priors specified, the posterior distribution of $\boldsymbol{\pi}$ and $\boldsymbol{\Theta}$, as well as other quantities of interest, may be investigated via Gibbs sampling as described by DIEBOLT and ROBERT (1994). The relevant full conditionals are

$$\begin{aligned}
\boldsymbol{\pi}|\cdots &\sim \text{Dir}(\zeta_1 + \#\{\boldsymbol{Z}=1\}, \ldots, \zeta_J + \#\{\boldsymbol{Z}=J\}) \\
\boldsymbol{\theta}_{j,\ell}|\cdots &\sim \text{Dir}(\lambda_{j,\ell,1} + m_{j,\ell,1} + n_{j,\ell,1}, \ldots, \lambda_{j,\ell,K_\ell} + m_{j,\ell,K_\ell} + n_{j,\ell,K_\ell}), \\
&\qquad j = 1, \ldots, J; \quad \ell = 1, \ldots, L \\
P(Z_i = j|\cdots) &= \frac{\pi_j P(\boldsymbol{Y}_i|\boldsymbol{\Theta}_j, Z_i = j)}{\sum_{k=1}^{J} \pi_j P(\boldsymbol{Y}_i|\boldsymbol{\Theta}_j, Z_i = k)}, \; i = 1, \ldots, N; \; j = 1, \ldots, J
\end{aligned}$$

where $\#\{\boldsymbol{Z}=j\}$ is the number of individuals currently allocated to subpopulation j, $m_{j,\ell,k}$ is the number of alleles of type k at locus ℓ in individuals currently allocated to subpopulation j, and the $n_{j,\ell,k}$ are, as before, the allele counts from the learning samples (if any) drawn separately from the mixture sample.

5.3 A Model with Admixed Individuals

With $\boldsymbol{\Theta}$ and $\mathbf{Y}$ defined as in the previous section, the model of PRITCHARD *et al.* (2000) is quickly described. Now, j indexes the J conceptual "gene pools" or "historical subpopulations" from which individuals may be descended. Allowing for admixed individuals requires a different model of genetic inheritance, which, in turn, requires different latent variables. For the i^{th} individual in the sample, a vector of probabilities $\boldsymbol{Q}_i = (Q_{i,1}, \ldots, Q_{i,J})$, $\sum_{j=1}^{J} Q_{i,j} = 1$, denotes the unobserved proportions of that individual's genome descended from each of the J gene pools. Also, let $\boldsymbol{W}_i = (W_{i,1}, \ldots, W_{i,2L})$ be a vector of unobserved allocation variables which is parallel to the the vector of allelic types $\boldsymbol{Y}_i$. Hence, $W_{i,t} = j$ indicates that the $(t \text{ [mod 2]} + 1)^{\text{th}}$ allele at the $\lceil t/2 \rceil^{\text{th}}$ locus in the i^{th} individual is from the j^{th} gene pool. Given $W_{i,t} = j$ the type of allele is assumed to be drawn randomly according to θ_j. Under this model

$$P(\boldsymbol{Y}_i|\boldsymbol{\Theta}, \boldsymbol{W}_i) = \prod_{t=1}^{2L} \theta\langle W_{i,t}; Y_{i,t}\rangle \tag{5.3}$$

independently for each i. By assigning the prior $\boldsymbol{Q}_i \sim \text{Dir}(\alpha, \ldots, \alpha)$, $i = 1, \ldots, N$, and the hyperprior $\alpha \sim \text{Uniform}(0, A]$, PRITCHARD *et al.* (2000)'s model is obtained. In effect this is a hierarchical model for N different finite mixtures—the genes carried by the i^{th} individual are a sample from a mixture with mixing proportions given by $\boldsymbol{Q}_i$, while the $\boldsymbol{Q}_i$ themselves ($i = 1, \ldots, N$) are drawn from a symmetrical $\text{Dir}(\alpha, \ldots, \alpha)$ distribution. This model allows for individuals to carry alleles in Hardy-Weinberg and linkage disequilibrium *with respect to the allele frequencies found in the single, admixed population.* The method of PRITCHARD *et al.* (2000) indirectly uses the information in this disequilibrium (which may occur even between unlinked loci) to assign genes to different subpopulations or "gene pools."

Gibbs sampling proceeds using the full conditionals

$$\begin{aligned}
\boldsymbol{Q}_i|\cdots &\sim \text{Dir}(\alpha_1 + \#\{\boldsymbol{W}_i = 1\}, \ldots, \alpha_J + \#\{\boldsymbol{W}_i = J\}), \quad i = 1, \ldots, N \\
\boldsymbol{\theta}_{j,\ell}|\cdots &\sim \text{Dir}(\lambda_{j,\ell,1} + r_{j,\ell,1}, \ldots, \lambda_{j,\ell,K_\ell} + r_{j,\ell,K_\ell}), \\
&\quad j = 1, \ldots, J; \quad \ell = 1, \ldots, L \\
P(W_{i,t} = j|\cdots) &= \frac{Q_{i,j}\theta\langle j; Y_{i,t}\rangle}{\sum_{k=1}^{J} Q_{i,j}\theta\langle k; Y_{i,t}\rangle}, \; i = 1, \ldots, N; \; j = 1, \ldots, J; \\
&\quad t = 1, \ldots, 2L
\end{aligned}$$

where $\#\{\boldsymbol{W}_i = j\}$ is the number of alleles in the i^{th} individual currently allocated to gene pool j and $r_{j,\ell,k}$ denotes the number of alleles of type k at locus ℓ currently allocated to gene pool j. PRITCHARD *et al.* (2000) update α by a Metropolis-Hastings method as described in Section 5.5. The posterior distribution of α thus estimated provides some insight into the degree to which admixture has occurred across individuals.

Learning samples would be available if there were substantial prior knowledge about the gene pools contributing to the admixture and if known, purebred descendants from them were separately sampled. By assuming any effects of genetic drift to be negligible, such samples could be treated as learning samples in the mixture model. The full conditional for $\boldsymbol{\theta}_{j,\ell}$ would then be modified to include the $n_{j,\ell,k}$ as before.

5.3.1 *Block-updating $\boldsymbol{W}_i$ when $J = 2$*

In many situations involving invasions of exotic species, there is substantial prior knowledge that the number of major subpopulations or "gene pools" involved is two—the native population and the invading population. Additionally, many hybrid zones are known to be areas of hybridization (admixture) between two species or populations. Here I present novel computations, feasible when only two subpopulations or gene pools are involved, that eliminate the explicit need for the variable $\mathbf{Q} = (\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_N)$ in implementing a Gibbs sampler. Such a method slightly improves mixing of the chain, but is primarily useful as it makes possible Gibbs sampling in a simultaneous mixture and admixture analysis described in Section 5.4.

The computations themselves may be derived as follows. Let $J = 2$, so that each allele in an individual may have originated from gene pool 1 or gene pool 2. Then, each $Q_{i,1}$ will follow a Beta(α, α) distribution and $Q_{i,2} = 1 - Q_{i,1}$. Conditional on $Q_{i,1}$, each $W_{i,t}$ will then be independently a Bernoulli trial with $P(W_{i,t} = 1|Q_{i,1}) = Q_{i,1}$. Marginalizing over $Q_{i,1}$ (not conditioning on the data) it follows that $\#\{\boldsymbol{W}_i = 1\}$ follows a beta-binomial distribution with parameters (α, α). Of course, each allele in an individual is uniquely labelled so the elements of $\boldsymbol{W}_i$ may be interpreted as following a *labelled* beta-binomial distribution. Under such a distribution, the elements of $\boldsymbol{W}_i$ are not independent, but they are exchangeable (DEFINETTI 1972), and hence their marginal distributions are invariant to permutations of their order (and thus the arbitrary order we have imposed upon them is acceptable).

This labelled beta-binomial sampling mechanism can be interpreted as arising from a Pólya-Eggenberger urn scheme (FELLER 1957; JOHNSON and KOTZ 1977). Imagine an urn initially filled with b_1 balls labelled "1" and b_2 balls labelled "2." Draw a ball randomly and record $W_{i,1} = 1$ or 2 according to the ball's label. Then replace the ball to the urn, adding, at the same time, c more balls of the same type (1 or 2) as the ball just drawn. Repeat the process, assigning a value to $W_{i,2}$ and so forth until $W_{i,2L}$ has also been assigned a 1 or 2. If b_1, b_2, and c were chosen to satisfy $b_1/c = b_2/c = \alpha$, then the resulting vector $\boldsymbol{W}_i$ would be a realized value from the labelled beta-binomial distribution with parameters

(α, α). (One should notice, also, that this extends to a non-symmetrical beta distribution, say Beta(α_1, α_2), by choosing $b_1/c = \alpha_1$ and $b_2/c = \alpha_2$.)

By such a scheme it is apparent that if D_t balls of type 1 have been drawn in the first t drawings from the urn, then the probability that the next ball drawn is a 1 is given by

$$\frac{b_1 + D_t c}{b_1 + b_2 + tc}. \tag{5.4}$$

And so the pairs $(W_{i,t}, D_t)$, $t = 1, \ldots, 2L$, can be interpreted as forming a Markov chain in time t with time-inhomogeneous transition probabilities determined by (5.4) and the obvious fact that $D_{t+1} = D_t + \mathcal{I}\{W_{i,t+1} = 1\}$, where $\mathcal{I}\{X = a\}$ takes the value one when $X = a$ and zero otherwise. This Markov chain dependence structure was previously noted by FREEDMAN (1965), who used it to obtain limiting distributions of quantities associated with urn models.

The foregoing has all been considered in the absence of data, $\boldsymbol{Y}_i$. However, given $\boldsymbol{\Theta}$, the data provide some information about the true value of each $W_{i,t}$ by the relation $P(Y_{i,t}|W_{i,t}, \boldsymbol{\Theta}) = \theta\langle W_{i,t}; Y_{i,t}\rangle$. Therefore, conditional on $\boldsymbol{\Theta}$ and $\boldsymbol{Y}_i$, the pairs $(W_{i,t}, D_t)$ participate in a *hidden* Markov chain. Recognition of this fact allows application of a "filter-forward, simulate-backward" type of algorithm which may be derived following the computations of BAUM *et al.* (1970) in order to realize the elements of $\boldsymbol{W}_i$ from their joint full conditional distribution, $P(\boldsymbol{W}_i|\alpha, \boldsymbol{\Theta}, \boldsymbol{Y}_i)$. Furthermore, using the BAUM (1972) algorithm, it is possible to compute $P(\boldsymbol{Y}_i|\alpha, \boldsymbol{\Theta})$, effectively performing a sum over all possible binary vectors of length $2L$ in an efficient manner. This is described below.

Take b_1, b_2, and c as defined above. Suppressing the i subscript for clarity, let $W_t \in \{1, 2\}$, $t = 1, \ldots, 2L$, and define $D_t = \sum_{\tau=1}^{t} \mathcal{I}\{W_\tau = 1\}$. We adopt the notation $W_{\leq t}$ $(W_{\geq t})$ to mean $W_1, \ldots, W_t$ $(W_t, \ldots, W_{2L})$ for components of W, and use the same notation with Y and D. The pairs (W_t, D_t) can be interpreted as following a Markov chain in t:

$$\begin{aligned} P(W_{t+1}, D_{t+1}|W_{\leq t}, D_{\leq t}) &= P(W_{t+1}, D_{t+1}|W_t, D_t) \\ &= \frac{b_1 + d_t c}{b_1 + b_2 + tc} \mathcal{I}\{D_{t+1} = D_t + \mathcal{I}\{W_{t+1} = 1\}\}. \end{aligned}$$

The "perturbed" or "degraded" observations of the chain are the allelic types $Y_1, \ldots, Y_{2L}$ which depend in hidden Markov fashion on W. For notational clarity, we assume implicit

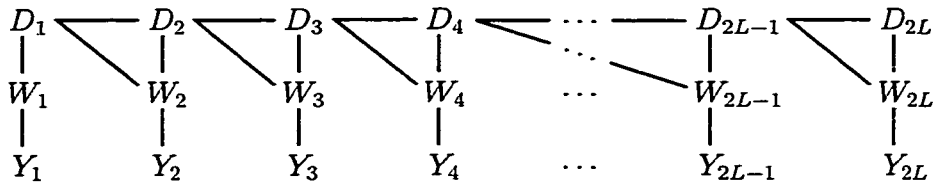


Figure 5.1: An undirected graph showing the dependence between $\boldsymbol{W}_i$, $\boldsymbol{D} = (D_1, \ldots, D_{2L})$ and $\boldsymbol{Y}_i$ in Section 5.3.1. This graph describes hidden Markov structure for the pairs $(W_{i,t}, D_t)$. Note that in the figure the i subscript has been omitted on the elements of $\boldsymbol{W}_i$ and $\boldsymbol{Y}_i$. The dependence on θ is implicit and not shown.

dependence on the allele frequencies $\boldsymbol{\Theta}$,

$$P(Y_t|W_{\leq 2L}, D_{\leq 2L}) = P(Y_t|W_t) = \theta\langle W_t; Y_t\rangle.$$

This dependence structure is shown in the undirected graph of Figure 5.1.

In the forward step we compute and store values of $P(W_t, D_t|Y_{\leq t})$ for $W_t = 1, 2$ and $D_t = 0, \ldots, t$, recursively for $t = 1, \ldots, 2L$, by the relations

$$P(W_{t+1}, D_{t+1}|Y_{\leq t}) = \sum_{1\leq W_t\leq 2} P(W_{t+1}, D_{t+1}|W_t, D_t^*)P(W_t, D_t^*|Y_{\leq t}) \tag{5.5}$$

where $D_t^* = D_{t+1} - \mathcal{I}\{W_{t+1} = 1\}$, and

$$P(W_{t+1}, D_{t+1}|Y_{\leq t+1}) = \frac{1}{\phi_{t+1}} P(W_{t+1}, D_{t+1}|Y_{\leq t})P(Y_{t+1}|W_{t+1}, D_{t+1}) \tag{5.6}$$

where

$$\phi_{t+1} = P(Y_{t+1}|Y_{\leq t}) = \sum_{\substack{1\leq W_{t+1}\leq 2 \\ 0\leq D_{t+1}\leq t+1}} P(W_{t+1}, D_{t+1}|Y_{\leq t})P(Y_{t+1}|W_{t+1}, D_{t+1}). \tag{5.7}$$

At the end of the forward step, notice that $\prod_{t=1}^{2L} \phi_t = P(Y_1, \ldots, Y_{2L})$, which in the context of the Gibbs sampler (and if we were to reinstate the i subscript) is the desired quantity $P(\boldsymbol{Y}_i|\alpha, \boldsymbol{\Theta})$ for the i^{th} individual. At the end of the forward step we have also obtained the distribution $P(W_{2L}, D_{2L}|Y_{\leq 2L})$. After simulating values for W_{2L} and D_{2L} from that distribution, we are in a position to simulate values for W_t going backwards recursively

for $t = 2L-1, 2L-2, \ldots, 1$, using the conditional distributions stored during the forward step and the values just realized for W_{t+1} and D_{t+1}. The backward step uses the following relations recursively to compute the conditional distribution from which to realize values of (W_t, D_t):

$$\begin{aligned} P(W_t, D_t | Y_{\leq 2L}, W_{\geq t+1}, D_{\geq t+1}) &= P(W_t, D_t | Y_{\leq t}, W_{t+1}, D_{t+1}) \\ &= \frac{1}{\psi_t} P(W_t, D_t | Y_{\leq t}) P(W_{t+1}, D_{t+1} | W_t, D_t), \end{aligned} \quad (5.8)$$

where ψ_t is a normalizing constant

$$\psi_t = \sum_{1 \leq W_t \leq 2} P(W_t, D_t^* | Y_{\leq t}) P(W_{t+1}, D_{t+1} | W_t, D_t^*) \quad (5.9)$$

and where, again, $D_t^* = D_{t+1} - \mathcal{I}\{W_{t+1} = 1\}$. It is apparent that a realization of the variable $(W_1, \ldots, W_{2L})$ thus obtained is drawn from the distribution of $W_1, \ldots, W_{2L}$ conditional on $Y_{\leq 2L}$. As such, in the context of the Gibbs sampler, and reinstating the i subscript, it is a realization from $P(W_i | \alpha, \boldsymbol{\Theta}, \boldsymbol{Y}_i)$ for the i^{th} individual, as desired.

The amount of computation required for the backward step is linear in L. The forward step at time t requires a handful of elementary operations for each of the $2t$ states that the pair (W_t, D_t) may take. This makes the entire forward step $O(L^2)$ for the case of two subpopulations. Depending on how many loci are available this will typically be computationally reasonable. However, extending this method to $J > 2$ will be computationally difficult. With $J > 2$, D_t becomes a vector whose elements record the number of balls of type $1, \ldots, J$ which have been drawn up to and including time t. The number of possible states for the pair (W_t, D_t) is then $J(t+J-2)!/[(t-1)!(J-1)!]$ which gets large rapidly with t and J.

5.4 *A Model for Simultaneous Population Mixture and Admixture*

Continuing in the case of two subpopulations ($J = 2$), a common goal in applications is to identify purebred versus admixed individuals and to estimate the proportion of those types in the population. This corresponds to partitioning one's sample into purebred and

admixed groups. The i^{th} individual's inclusion in one of the two groups can be denoted by a latent variable V_i taking the values

$$V_i = \begin{cases} \text{P}, & \text{if purebred} \\ \text{A}, & \text{if admixed following the PRITCHARD } \textit{et al.} \text{ (2000) model} \end{cases}$$

Using the calculation of Section 5.3.1, this partition problem can be treated as a mixture problem using Gibbs sampling. The proportion of individuals of the two types in the population are given by the new parameter ξ_{P} and $\xi_{\text{A}} = 1 - \xi_{\text{P}}$. The full conditional distribution for V_i is then, for example, for $V_i = \text{P}$

$$P(V_i = \text{P}|\cdots) = \frac{\xi_{\text{P}} P(\boldsymbol{Y}_i|\alpha, \boldsymbol{\Theta}, V_i = \text{A})}{\xi_{\text{P}} P(\boldsymbol{Y}_i|\boldsymbol{\pi}, \theta, V_i = \text{P}) + \xi_{\text{A}} P(\boldsymbol{Y}_i|\alpha, \theta, V_i = \text{A})}. \tag{5.10}$$

Calculating the necessary phenotype probabilities, $P(\boldsymbol{Y}_i|\boldsymbol{\pi}, \boldsymbol{\Theta}, V_i = \text{P})$ and $P(\boldsymbol{Y}_i|\alpha, \boldsymbol{\Theta}, V_i = \text{A})$, has been described in Equation 5.2 and Section 5.3.1.

The conjugate prior for ξ_{P} is Beta$(\delta_{\text{P}}, \delta_{\text{A}})$ which gives the full conditional

$$\xi_{\text{P}}|\cdots \sim \text{Beta}(\delta_{\text{P}} + \#\{\boldsymbol{V} = \text{P}\}, \delta_{\text{A}} + \#\{\boldsymbol{V} = \text{A}\}). \tag{5.11}$$

I have used a uniform ($\delta_{\text{P}} = \delta_{\text{A}} = 1$) prior for ξ_{P}.

In this expanded model, which we will call model $M_{\text{P,A}}$ a sweep consists of

1. Gibbs update for $\boldsymbol{\pi}$ using only the individuals with $v_i = \text{P}$,

2. Gibbs update for $\boldsymbol{\Theta}$ where contributions to the full conditionals are determined by Z_i for individuals with $V_i = \text{P}$ and by $\boldsymbol{W}_i$ for those with $V_i = \text{A}$,

3. Gibbs updates for each individual's Z_i if $v_i = \text{P}$ and for w_i if $v_i = \text{A}$,

4. Gibbs update for ξ from Equation 5.11,

5. Gibbs update for each V_i using Equation 5.10,

6. Metropolis-Hastings update for α as described in Section 5.5.

The output from the resulting Markov chain can provide Rao-Blackwellized (LIU *et al.* 1994) estimates for the posterior probability that individuals in the sample are purebred or admixed as well as estimates of the posterior distributions for ξ, $\boldsymbol{\pi}$, $\boldsymbol{\Theta}$, α, and each Q_i given $V_i = A$ (though the Q_i's are not necessary for running the chain, they may still be realized from their full conditional distributions and they provide good summary statistics).

5.5 *Metropolis Updates for* α

The method of Metropolis sampling is used to update values of α. A new value for α, denoted α^*, is drawn from a proposal distribution. Since α is constrained to the interval $(0, A]$, I use a folded normal distribution, centered at α. Hence a variable a is drawn from a Normal(α, σ^2) distribution. If $0 < a \leq A$ then $\alpha^* = a$. Otherwise if $-A \leq a < 0$ then $\alpha^* = -a$ and if $A < a \leq 2A$ then $\alpha^* = 2A - a$. In all other cases ($a < -A$ or $a > 2A$) the proposal is rejected without further consideration. The proposal density is then still symmetrical

$$h(\alpha^*|\alpha) = \mathcal{N}(\alpha^*; \alpha, \sigma^2) + \mathcal{N}(-\alpha^*; \alpha, \sigma^2) + \mathcal{N}(2A - \alpha^*; \alpha, \sigma^2) = h(\alpha|\alpha^*)$$

with $\mathcal{N}(\alpha^*; \alpha, \sigma^2)$ denoting the normal density function of α^* having mean α and variance σ^2. The standard deviation, σ, of the proposal distribution requires some tuning. Under model M_{A}, $\sigma \approx .12$ seems to work well, while when individuals may be purebred or admixed (model $M_{\mathrm{P,A}}$) then $\sigma \approx .5$ encourages better mixing for the Scottish cat data.

The proposed value α^* is accepted as the new value with probability given by the minimum of 1 or the Hastings ratio. For PRITCHARD *et al.* (2000)'s model, using, the q_i's, the acceptance probability is

$$\min\left\{1, \frac{\prod_{i=1}^{N} \mathcal{D}(Q_i; \alpha^*, J)}{\prod_{i=1}^{N} \mathcal{D}(Q_i; \alpha, J)}\right\}$$

where $\mathcal{D}(Q; \alpha, J)$ denotes the density of a Dirichlet random vector Q of J components with all J parameters equal to α.

When able to eliminate the Q_i's (as in Section 5.3.1), then with only admixed individuals (model M_{A}) the acceptance probability may be written as

$$\min\left\{1, \frac{\prod_{i=1}^{N} P(\boldsymbol{Y}_i|\alpha^*, \boldsymbol{\Theta})}{\prod_{i=1}^{N} P(\boldsymbol{Y}_i|\alpha, \boldsymbol{\Theta})}\right\}.$$

In the model $M_{\mathrm{P,A}}$ which includes both purebred and admixed individuals, the acceptance probability is

$$\min\left\{1, \frac{\prod_{i=1}^{N}[\xi_{\mathrm{P}}P(\boldsymbol{Y}_i|\boldsymbol{\pi},\boldsymbol{\Theta},V_i=\mathrm{P})+\xi_{\mathrm{A}}P(\boldsymbol{Y}_i|\alpha^*,\boldsymbol{\Theta},V_i=\mathrm{A})]}{\prod_{i=1}^{N}[\xi_{\mathrm{P}}P(\boldsymbol{Y}_i|\boldsymbol{\pi},\boldsymbol{\Theta},V_i=\mathrm{P})+\xi_{\mathrm{A}}P(\boldsymbol{Y}_i|\alpha,\boldsymbol{\Theta},V_i=\mathrm{A})]}\right\}.$$

5.6 Data and Results

The data from Scottish wildcats were provided by Mark Beaumont (University of Reading, UK) and are fully described in BEAUMONT *et al.* (2001). The dataset is freely available at `http://www.rubic.rdg.ac.uk/mab/data.html`. Briefly, genetic samples were collected from wild-living cats throughout Scotland by a variety of methods including trapping and tissue collection from road kills and carcasses. Samples were also obtained from 13 museum specimens. In all, 230 wild-living cats were sampled and typed at eight microsatellite loci with numbers of alleles ranging from nine to 17 per locus. Additionally, 74 housecats were typed at those eight loci using blood samples from veterinary centers in the south of England. These 74 cats can be considered a learning sample for the domestic cat subpopulation.

I analyzed the data under model $M_{\mathrm{P,A}}$ using runs of length 62,000 sweeps of ten different chains started from overdispersed starting points by initializing values of all parameters $(\alpha, \boldsymbol{\Theta}, \xi_{\mathrm{P}}, \boldsymbol{\pi})$ with values simulated from their prior distributions. All ten chains converged very quickly to the same part of the parameter space. The first 2,000 sweeps were discarded as burn-in, as observing the estimated scale reduction potential factor (GELMAN 1996) suggests this is more than adequate burn-in. I give the results in the next section. I performed an analogous run under model M_{A} and compare the differences between the results obtained under $M_{\mathrm{P,A}}$ and M_{A} in Section 5.6.2. For each run I used an upper bound of $A = 3$ for the parameter α. Each run took 11 hours on a laptop computer with a 266 Mhz G3 (Macintosh) processor.

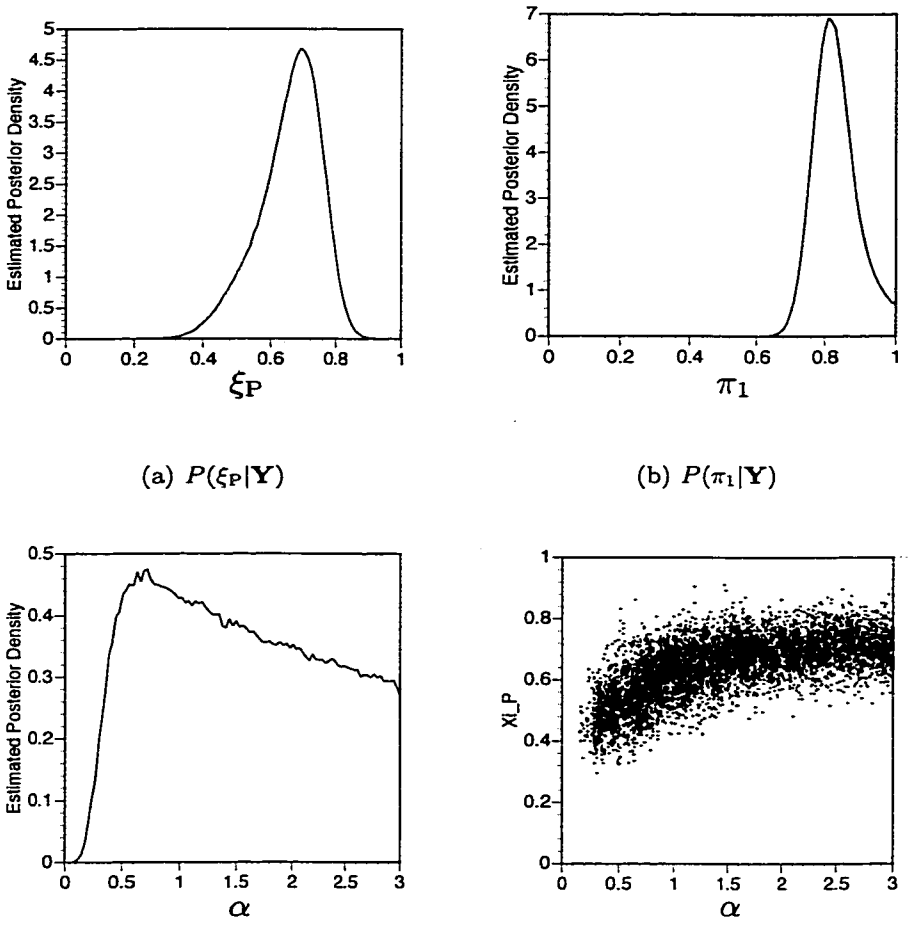
It should be noted that the learning sample of housecats breaks the symmetry in the posterior with respect to permutations on the labels for the two components (*F. sylvestris* and housecats) in the model. Thus, there is not a substantial label-switching (STEPHENS 2000) problem in this case.

5.6.1 *Results for model* $M_{\mathrm{P,A}}$

The posterior mean estimate of ξ_{P}, the proportion of purebred cats, is .65, with a 90% credible set spanning the range from .47 to .79. The category of purebred cats includes both pure *F. sylvestris* and pure housecats found in the admixed sample (but *not* the housecats in the learning sample). The MCMC estimate of the posterior density of ξ_{P} is given in Figure 5.2(a). The distribution is long-tailed to the left. These low values of ξ_{P} coincide with low values of the parameter α (Figure 5.2(d)). This correlation is expected; when α is low, then admixed individuals are expected to have admixture proportions near to zero or one, and hence the ability to distinguish between admixed and purebred individuals is diminished. The estimated posterior density for α itself is shown in Figure 5.2(c). It has a peak around 0.7, and tapers off with larger values, but it is still rather high at the upper bound imposed on it of 3. The choice of $A = 3$ is clearly a choice of prior to which the final result will be sensitive. A larger A would reduce the posterior probability for low values of ξ_{P}, reducing the skewness of the posterior for ξ_{P} and increasing its posterior mean estimate. This issue will be taken up again in the Discussion.

Figure 5.2(b) gives the estimated posterior density for the probability that a cat is *F. sylvestris* conditional on its being of purebred type. The posterior mean is .83 with a 90% credible interval from .73 to .94. This suggests that a large proportion ($> 53\%$) of the wild-living cats in Scotland are purebred *F. sylvestris*. On the other hand, there is evidence that between 21% and 53% of the wild-living cats are admixed individuals with ancestry from both *F. sylvestris* and domestic cats. Further, it cannot be ruled out that some wild-living cats are pure housecats that have gone feral.

Figure 5.3 summarizes the results for individual cats. On the horizontal axis is the posterior probability of being purebred. On the vertical axis is the posterior probability of being *F. sylvestris* conditional on being purebred. A cluster in the upper right represents 102 of the cats in the sample, all with posterior probability of being pure greater than .80. Given that these cats are pure, they have posterior probability close to one of being *F. sylvestris*. Also evident is a small cluster of cats with $P(V_i = \mathrm{P}|\mathbf{Y}) > .65$ but which, if they are purebred cats, are almost certainly domestic cats. At the other end of the scale



(a) $P(\xi_P|\mathbf{Y})$ (b) $P(\pi_1|\mathbf{Y})$

(c) $P(\alpha|\mathbf{Y})$ (d) α vs. ξ_P

Figure 5.2: Graphical summaries of the aggregate-level parameters for the Scottish cat dataset. (a–c) are unsmoothed posterior density estimates taken by scaling histograms with bin widths of 0.01 in (a) and (b) and 0.03 in (c): (a) proportion of purebred cats, ξ_P in the population from which the cats were sampled, (b) proportion of cats, π_1, from the *F. sylvestris* subpopulation, conditional on being purebred, (c) the parameter α. (d) a scatterplot of 5,000 pairs (α, ξ_P) sampled from the Markov chain. The two parameters are clearly correlated. Lower values of α correspond to lower values of ξ_P, as expected. The correlation is most apparent for values of α less than 1.5. For $\alpha > 1.5$, the value of α has little effect on the value of ξ_P.

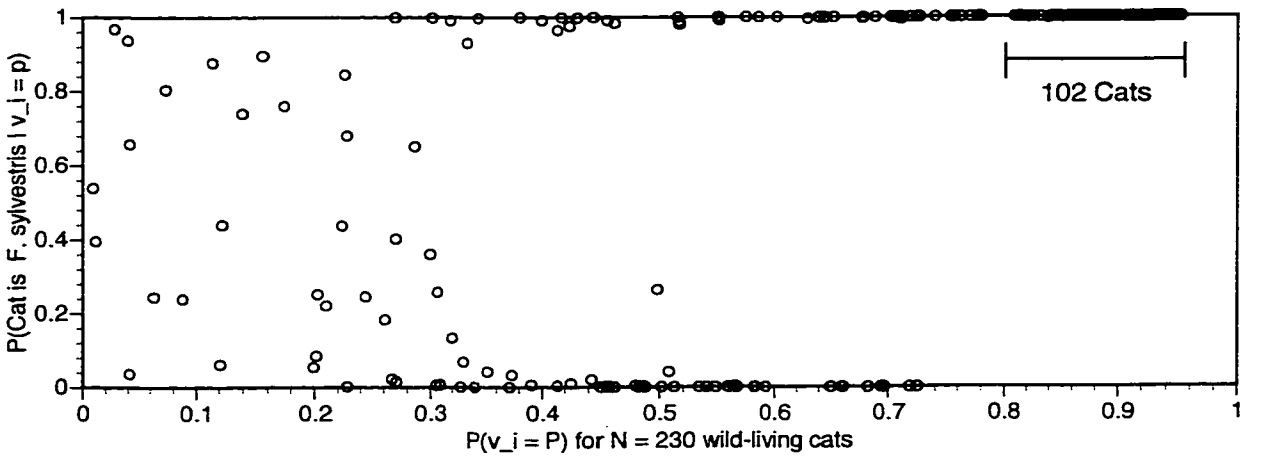


Figure 5.3: A plot of posterior mean estimates for $P(V_i = \mathrm{P})$ on the horizontal axis against $P(Z_i =$ "*F. sylvestris*"$|V_i = \mathrm{P})$ on the vertical axis. Each open circle represents one of the 230 wild-living cats in the sample. The cluster in the upper right includes 120 individuals all with posterior probability of being purebred *F. sylvestris* greater than .80. At far left are some eight individuals with high posterior probability of being admixed. For cats with intermediate estimates of $P(V_i = \mathrm{P}|\mathbf{Y})$, the credible sets tend to be quite wide (not shown).

are several cats with very high probability of being admixed.

5.6.2 Comparison of results for models $M_{\mathrm{P,A}}$ and M_{A}

For parameters shared by M_{A} and $M_{\mathrm{P,A}}$, the estimates differ between models most for α. Under M_{A}, α is much smaller, so as to accommodate the purebred cats as admixed individuals with admixture proportions close to 0 or 1 (Figure 5.4). Additionally, a significantly smaller proportion of the alleles in the sample get allocated to the housecat population under $M_{\mathrm{P,A}}$ than under M_{A}. Histograms of the proportion of alleles allocated to the housecat subpopulation for $M_{\mathrm{P,A}}$ and M_{A} are shown in Figure 5.5.

5.7 Bayesian Model Comparison

Once able to entertain the model $M_{\mathrm{P,A}}$, it is natural to ask whether that expanded model has gained us anything. One way to pose the question is to ask whether the data provide more support for $M_{\mathrm{P,A}}$ than for the model we will call M_{A} which requires all individuals to

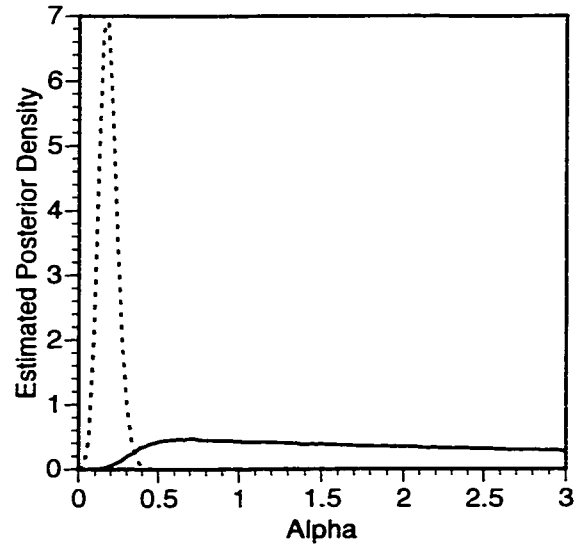


Figure 5.4: Posterior densities for α from the Scottish cat data. Broken line shows $P(\alpha|\mathbf{Y})$ under model M_{A}. Solid line shows $P(\alpha|\mathbf{Y})$ under model $M_{\mathrm{P,A}}$. It appears that under M_{A}, most of the information constraining values of α in the posterior comes from individuals of pure, or mostly pure origin.

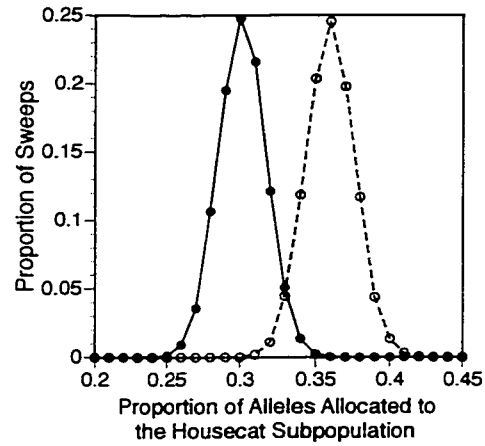


Figure 5.5: Comparison of the proportion of all the alleles in the sample allocated to the housecat subpopulation. Solid line and circles are a histogram from $M_{\mathrm{P,A}}$; broken line and open circles are a histogram from M_{A}. There is little overlap between the two. A higher proportion of alleles is allocated to the housecat population under M_{A}.

be admixed and governed by a single α. This may be assessed via the Bayes Factor (KASS and RAFTERY 1995), $B = P(\mathbf{Y}|M_{\mathrm{P,A}})/P(\mathbf{Y}|M_{\mathrm{A}})$. A rough estimate of B might be obtained by observing the proportion of time the Markov chain defined under $M_{\mathrm{P,A}}$ spends in states with zero or almost zero individuals allocated to the purebred group (since retricting ξ_{P} to zero in $M_{\mathrm{P,A}}$ essentially gives M_{A}). However, this is unsatisfactory as there is no prior probability mass on the point $\xi_{\mathrm{P}} = 0$. Furthermore, the chain may visit states with low ξ_{P} so infrequently, that it is impossible to get a good estimate of B that way.

GELFAND *et al.* (1992) suggest approximating B by the "pseudo-Bayes factor" formed as the product over all observations of the ratio of cross-validitation predictive densities under the two models. The cross-validation predictive density for the i^{th} individual, may be approximated by the harmonic mean of the values $P(\boldsymbol{Y}_i|\alpha^{(s)}, \boldsymbol{\Theta}^{(s)})$ under M_{A} and the values $P(\boldsymbol{Y}_i|\alpha^{(s)}, \boldsymbol{\pi}^{(s)}, \boldsymbol{\Theta}^{(s)}, \xi_{\mathrm{P}}^{(s)})$, computed as the denominator in (5.10), under $M_{\mathrm{P,A}}$, where the superscript $^{(s)}$ denotes the states visited by the chain over which the harmonic mean is taken. RAFTERY (1992) cautions that the pseudo-Bayes factor, being akin to a pseudo-likelihood, may be an inaccurate approximation to the Bayes factor and should not be used for model comparison if the latter is available. However as discussed by PRITCHARD *et al.* (2000), it is difficult to estimate reliably the marginal likelihood $P(\boldsymbol{Y}|M_{\mathrm{A}})$, and the same is true for $P(\boldsymbol{Y}|M_{\mathrm{P,A}})$, making computation of the Bayes factor by that route challenging.

As an alternative, I have developed a reversible-jump MCMC scheme (GREEN 1995) for computing the Bayes factor. While reversible jump methods have recently received widespread attention for sampling over numerous models in complex model spaces (RUE and HURN 1999; DELLAPORTAS and FORSTER 1999; GIUDICI and GREEN 1999), it seems they have been used less often when a small number of closely-related models are being considered, as in the present case. The posterior distributions estimated from separate runs under $M_{\mathrm{P,A}}$ and M_{A} can guide us in devising reversible-jump proposals that are easy to implement and offer a good estimate of B. Details appear in Section 5.7.1. This circumvents the potential problem of instability in a direct, sampling-based estimate of $P(\mathbf{Y}|M_{\mathrm{P,A}})$ or $P(\mathbf{Y}|M_{\mathrm{A}})$, and affords an opportunity to compare the pseudo-Bayes factor to the full Bayes factor in the comparison of two complex, hierarchical models.

5.7.1 *Reversible Jump MCMC for Model Comparison*

Reversible jump MCMC (GREEN 1995) allows for the construction of a Markov chain that may jump between state spaces of varying dimension. In our case we construct a chain which takes values in two spaces indexed by $m = 1$ or 2. If $m = 1$ then the chain is currently in the space associated with model M_{A}, and it moves to new values within that space as described in Section 5.3. If $m = 2$, then the chain is currently in the state space associated with model $M_{\mathrm{P,A}}$, and it moves to new values within that space as described in Section 5.4. Since $M_{\mathrm{P,A}}$ includes the variables ξ_{P} and $\boldsymbol{\pi}$ (ξ_{P} has only one degree of freedom and, in the case of $J = 2$, $\boldsymbol{\pi}$ has one degree of freedom as well—π_1, which I will just denote by π_1 in the following) which are absent in model M_{A}, there are two extra degrees of freedom when $m = 2$. For this reason, reversible-jump moves are required to move from $m = 1$ to $m = 2$. The formulation of these moves is such that detailed balance is satisfied, ensuring that the proportion of time the chain spends with $m = 1$ converges to $P(M_{\mathrm{A}}|\mathbf{Y})$ as the chain is run for infinite time, and so, for a run of the chain of length n, the quantity

$$\frac{\sum_{i=1}^{n} \mathcal{I}\{m_i = 1\}}{\sum_{i=1}^{n} \mathcal{I}\{m_i = 2\}} \tag{5.12}$$

estimates the posterior odds, which, upon division by the prior odds, gives B.

For a reversible jump move from $m = 2$ to $m = 1$ we leave $\boldsymbol{\Theta}$ unchanged and propose a new value for α, say α', that is a deterministic, many-to-one, function g of the current values of α, ξ_{P}, and π. We are at liberty to choose any appropriate and suitable g. For the Scottish cat problem, by examining the posterior distribution of ξ_{P}, π_1, and α under $M_{\mathrm{P,A}}$, and by surmising that high values of ξ_{P} in $M_{\mathrm{P,A}}$ should correspond to low values of α' in M_{A}, I empirically chose

$$\alpha' = g(\alpha, \xi, \pi_1) = 0.0925 + 0.13638\alpha - 0.21 \sin^{-1}(\xi_{\mathrm{P}}^2). \tag{5.13}$$

In this case, $\sin^{-1}(\xi_{\mathrm{P}}^2)$ was chosen, since that transformed variable has a more linear relationship with α than does ξ_{P}, itself, in the MCMC output from $M_{\mathrm{P,A}}$ (see Figure 5.2(d)). Notice that π_1 does not actually appear in the function g, since this simplifies the Jacobian, and it does not seem essential (i.e. there is not large correlation between α and π_1). Taking the 5000 pairs $(\alpha, \xi_{\mathrm{P}})$ that were plotted in Figure 5.2(d) and applying g to them gives

values of α' summarized by their histogram in Figure 5.6(a). This histogram resembles the posterior distribution of α under M_{A} (broken line in Figure 5.4), as desired.

To propose the reverse move from $m = 1$ with a current value α' to $m = 2$ with proposed values for the parameters α, ξ_{P} and π_1 requires simulating new values for ξ_{P} and π_1 from known densities and then using those values and the inverse of the function g to determine what value of α shall be proposed. The known densities were chosen to approximate the posterior density estimates of ξ_{P} and π_1 under $M_{\mathrm{P,A}}$. Letting π_1 denote the proportion of purebred cats that are *F. sylvestris*, the densities used were $f_\xi(\xi_{\mathrm{P}}) \equiv \mathrm{Beta}(8,4)$ and $f_\pi(\pi_1) \equiv .8\mathrm{Beta}(30,9) + .2\mathrm{Beta}(15,2)$. These densities are shown in Figure 5.6(b). Comparison to Figures 5.2(a) and 5.2(b) shows that they resemble overdispersed versions of $P(\xi_{\mathrm{P}}|\mathbf{Y})$ and $P(\pi_1|\mathbf{Y})$. With values of ξ_{P} and π_1 drawn from these densities, α is determined by

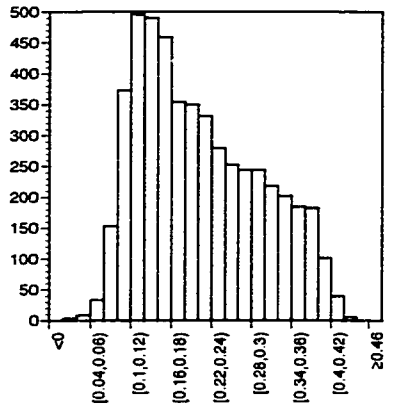
$$\alpha = g^{-1}(\alpha', \xi_{\mathrm{P}}, \pi_1) = \frac{0.21 \sin^{-1}(\xi_{\mathrm{P}}^2) + \alpha' - .0925}{.13638}.$$

We may propose a reversible jump move at the end of each sweep. Thus, if $m = 1$, after a sweep updating all the variables associated with M_{A}, we propose a jump up to $m = 2$. If $m = 2$, then after a sweep updating all the variables associated with $M_{\mathrm{P,A}}$ we propose a jump down to $m = 1$. Under such a scheme, the acceptance probability for a proposed move from $m = 1$ with $\alpha = \alpha'$ to $m = 2$ and parameter values $(\alpha, \xi_{\mathrm{P}}, \pi_1)$, is $\min\{1, \mathcal{A}\}$, with $\mathcal{A}$ reducing to

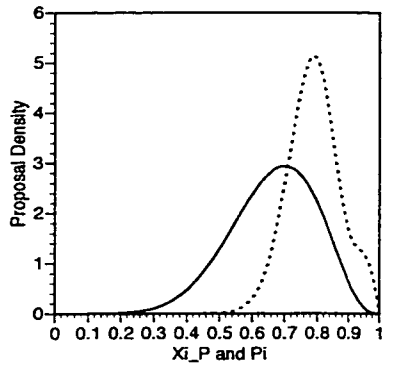
$$\begin{aligned} \mathcal{A} &= \frac{P(M_{\mathrm{P,A}})}{P(M_{\mathrm{A}})} \times \frac{P(\alpha|M_{\mathrm{P,A}})}{P(\alpha'|M_{\mathrm{A}})} \times \frac{P(\xi_{\mathrm{P}}|M_{\mathrm{P,A}})P(\pi_1|M_{\mathrm{P,A}})}{f_\xi(\xi_{\mathrm{P}})f_\pi(\pi_1)} \\ &\times \frac{\prod_{i=1}^{N}[\xi_{\mathrm{P}}P(\boldsymbol{Y}_i|\pi_1, \boldsymbol{\Theta}) + \xi_{\mathrm{A}}P(y_i|\alpha, \theta)]}{\prod_{i=1}^{N} P(\boldsymbol{Y}_i|\alpha', \boldsymbol{\Theta})} \times \frac{1}{0.13638} \end{aligned} \tag{5.14}$$

where $P(\cdot|M)$ denotes prior densities for parameters under different models M. If proposing a move down from $m = 2$ with current values $(\alpha, \xi_{\mathrm{P}}, \pi_1)$ to $m = 1$ with $\alpha = \alpha'$, the acceptance probability is $\min\{1, \mathcal{A}^{-1}\}$. The factor of $(0.13638)^{-1}$ is the Jacobian from the transformation g.

Figure 5.7(a) shows a trace of $\log \mathcal{A}$ from a chain forced to stay in $m = 1$ (i.e. it makes proposals to $m = 2$ but is not allowed to accept them) using the Scottish cat data

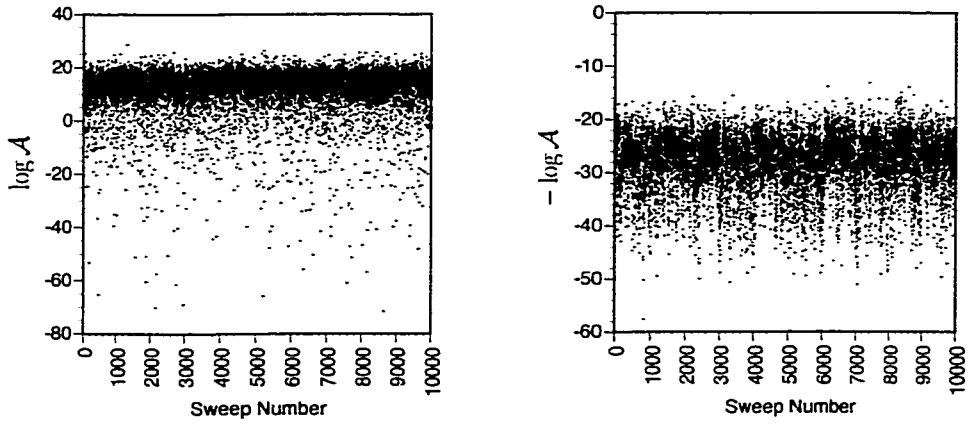


(a) α' histogram



(b) $f_\xi(\xi_P)$ & $f_\pi(\pi_1)$

Figure 5.6: Elements of the reversible-jump proposals. (a) Histogram of values of α' computed as $g(\xi_P, \alpha, \pi_1)$ for 5,000 points visited by a Markov chain run under $M_{P,A}$. The function g was chosen so that this histogram would be similar to the posterior density for α under M_A, shown in Figure 5.4. (b) The proposal densities: solid line is $f_\xi(\xi_P)$; broken line is $f_\pi(\pi_1)$. These were chosen to represent overdispersed versions of $P(\xi_P|\mathbf{Y})$ and $P(\pi_1|\mathbf{Y})$ under $M_{P,A}$, shown in Figures 5.2(a) and 5.2(b).



(a) Trace of $\log \mathcal{A}$ under M_{A}

(b) Trace of $-\log \mathcal{A}$ under $M_{\mathrm{P,A}}$

Figure 5.7: With prior odds of 1, (a) a trace of values of $\log \mathcal{A}$ plotted as unconnected points for 10,000 sweeps of a chain with m fixed at 1 (model M_{A}). The majority of points lie above 5, indicating that most proposals to move to $M_{\mathrm{P,A}}$ from M_{A} by the proposed reversible-jump move would be accepted. Note that some values of $\log \mathcal{A}$ are greater than 20. So, even with prior log-odds of $\log[P(M_{\mathrm{P,A}})/P(M_{\mathrm{A}})] \approx -20$, proposals from $m = 1$ to $m = 2$ will occasionally be accepted. (b) a trace of $-\log \mathcal{A}$ for 10,000 sweeps of a chain restricted to $m = 2$. Many values are less than -20. However, again, with $\log[P(M_{\mathrm{P,A}})/P(M_{\mathrm{A}})] \approx -20$ proposals from $m = 2$ to $m = 1$ will be occasionally accepted.

with learning samples, and assuming prior odds for the models $P(M_{\mathrm{P,A}})/P(M_{\mathrm{A}}) = 1$. Figure 5.7(b) shows a similar trace of $\log \mathcal{A}^{-1}$ for a chain restricted to $m = 2$. It is apparent from these traces that, without imposing strong prior support for M_{A}, it is unlikely that a chain in $m = 2$ would ever move to $m = 1$. Thus, I made three different runs with prior log-odds, $\log[P(M_{\mathrm{P,A}})/P(M_{\mathrm{A}})]$, equal to -19, -20, and -21. From each of these runs, I estimated the posterior log adds by taking the log of (5.12). The value of the posterior log-odds calculated as the average over ten chains started from overdispersed starting points as a function of sweep number is shown for the three different prior odds in Figure 5.8. Though the chains may not have been run long enough for an extremely precise estimate of the posterior log-odds, an order-of-magnitude estimate can clearly be made. Subtracting the prior log-odds from the estimated posterior log-odds gives, for each of the three different

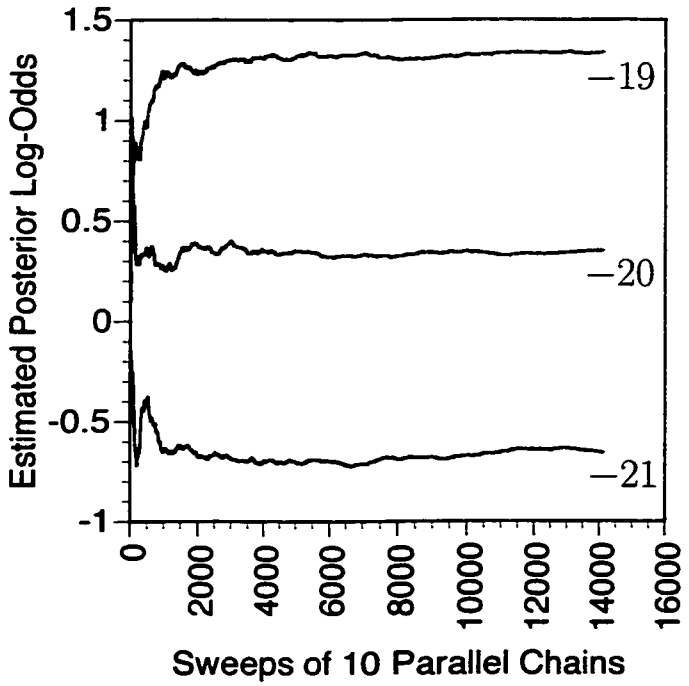


Figure 5.8: Estimated posterior log-odds, $\log[P(M_{\mathrm{P,A}}|\mathbf{Y})/P(M_{\mathrm{A}}|\mathbf{Y})]$ for three different prior log-odds. Each line shows the estimate as a function of number of sweeps. Ten separate chains started from overdispersed points were used for each estimate. The estimates shown are the log of the expression in (5.12) for $n = 10$ times the number of sweeps. The numbers at the right of each line are the prior log-odds assumed. $M_{\mathrm{P,A}}$ is so highly favored, that in order for the reversible-jump sampler to mix, huge prior weight must be given to M_{A}. The log of the Bayes factor, $\log B$, may be calculated by subtracting the prior log-odds from the estimated posterior log-odds. For all three values of the prior odds this gives about 20.3.

prior odds used, an estimate of ≈ 20.3 for the log of the Bayes factor. Hence, $2 \log B > 40$, indicating overwhelming support in the data for model $M_{\mathrm{P,A}}$ over M_{A}. The log of the pseudo-Bayes factor is 12.3 which, quite notably, differs by eight from the true $\log B$.

5.8 Discussion

In applications to conservation biology and ecology, populations of interest may be pure mixtures of two subpopulations, or they may contain admixed individuals from two originally separate subpopulations. Genetic data, in conjunction with statistical models of

genetic mixture and admixture have been useful for clustering individuals and genes from different subpopulations. This chapter presents a novel application of the "filter-forward-simulate-backward" algorithm akin to the computations presented in BAUM *et al.* (1970) to the population-admixture model of PRITCHARD *et al.* (2000). This computation makes it possible to expand that model to one that allows for individuals to be either purebred or admixed. Such an expanded model ($M_{\mathrm{P,A}}$) is vastly more supported by the Scottish cat data than one including admixture only. It is likely that $M_{\mathrm{P,A}}$ will fit other datasets much better as well, because samples from recently admixed populations will typically include some purebred individuals.

While the dramatic improvement of model fit is encouraging, it also raises some issues that bear further investigation. The first of these is that $M_{\mathrm{P,A}}$ may fit the data better not simply because it allows separate classes of purebred and admixed individuals. It may be that a great deal of improvement comes from including the parameter π which allows different contributions of pure cats from the two subpopulations. This contrasts to the formulation in M_{A} where, due to the symmetry of the Beta(α, α) prior for the q_i's, the marginal probabilities are equal that any gene copy is from the housecat or the *F. sylvestris* subpopulation. That is to say, under M_{A}, $P(W_{i,t} = 1|\alpha) = P(W_{i,t} = 2|\alpha) = .5$ for all i, t, and α. By contrast, under $M_{\mathrm{P,A}}$, for different values of ξ, π, and α, the marginal probability that any gene copy is from the housecat population is not constrained to equal the marginal probability that it is from the *F. sylvestris* population. The symmetry imposed by M_{A} might explain some systematic biases for estimates of q_i that PRITCHARD, STEPHENS, ROSENBERG and DONNELLY (2000) report for simulated data with unequal admixture proportions. Furthermore, the issue may have implications for model $M_{\mathrm{P,A}}$. If the population admixture proportions depart from .5, then $P(\boldsymbol{Y}_i|\alpha, \boldsymbol{\Theta})$ might be inflated for individuals with large amounts of ancestry from the lesser-represented subpopulation, and deflated for individuals with more ancestry from the greater-represented subpopulation. For this reason, in the Scottish cat problem, one might expect that the posterior probability of being a purebred individual will be overestimated for cats that resemble *F. sylvestris* and underestimated for cats that appear to be housecats. This may also induce some bias in the posterior estimate of ξ_{P}. All this suggests that a fruitful extension to the model M_{A}

of PRITCHARD *et al.* (2000) would be to allow population-specific α's. For example, in the case of J subpopulations, $q_i \sim \text{Dir}(\alpha_1, \ldots, \alpha_J)$.

The results also suggest that estimation in $M_{\text{P,A}}$ may be sensitive to the upper bound, A, chosen for α. Had A been chosen greater than three, then values of $\alpha > 3$ would surely have been visited in the MCMC simulation, and the resulting estimate for ξ_{P} would have been somewhat larger, since α and ξ_{P} are positively correlated. This is observed in a separate run made with $A = 10$—the chain visits values of α between 3 and 10 quite frequently. In fact, the estimated posterior density for α decreases only slightly between 3 and 10. However, the effect on the other parameters is not overwhelming. For example, with $A = 10$, the posterior mean (90% credible interval) for ξ_{P} was .71 (.53, .82), as opposed to .65 (.47, .79) with $A = 3$. It is interesting that the choice of A has little effect in the poorer-fitting model M_{A}, because that model tries to fit purebred cats as admixed individuals. This keeps α low regardless of A. Under $M_{\text{P,A}}$, however, once the purebred individuals are removed from the admixed class there is little information left for estimating α. So, paradoxically, to use the better-fitting model $M_{\text{P,A}}$ requires imposing more prior information. In the case of A, however, biological knowledge can guide the choice.

I chose $A = 3$ because, with only two subpopulations, large values of α indicate that admixed individuals carry close to half of their ancestry from one subpopulation and half from the other. In a population like this, the most plausible explanation for such a pattern would be that the admixed individuals were all first-generation (F_1) hybrids between individuals from the two subpopulations. If this is the case, then, at each locus, an admixed individual will carry exactly one allele from one subpopulation, with the other allele coming from the other subpopulation. This condition can be used to compute the posterior probability that an individual in the sample is an F_1 hybrid. The details of this are covered in the following chapter. I have found that none of the individuals in the sample had posterior probability greater than .5 of being an F1 hybrid. In fact, for all but seven of the individuals, the posterior probability of being an F1 hybrid was below .10. For this reason, it seemed implausible that α should be allowed to range past about three.

The Bayesian model comparison revealed that $M_{\text{P,A}}$ is a much better model for the Scottish cat data. Computing Bayes factors in these models is often difficult because calcu-

lating the marginal likelihood can require a difficult computation of an unkown normalizing constant. Rather than directly computing the marginal likelihood, Section 5.7.1 gives an example of how approximations to posterior densities of several parameters in different models may be used to formulate reversible-jump moves between a small set of closely-related models. This gives us a good approximation to the Bayes factor. In turn, that allows us to compare the true Bayes factor to the pseudo-Bayes factor (a product of ratios of cross-validation predictive densities). The pseudo-Bayes factor has been advocated as a computationally manageable approximation to the Bayes factor. While cross-validation and predictive densities offer a fine level of detail for exploring which observations, in particular, are poorly fit by a model, comparisons between models via the pseudo-Bayes factor should be made with reservation. In the current problem, the pseudo-Bayes factor provided a poor approximation of the Bayes factor.

Quite apart from genetic mixtures, the forward-backward computation here may be useful in more general mixture problems. Sometimes, the Gibbs sampler mixes poorly in the Bayesian analysis of mixtures. ROBERT (1996) describes this in terms of *trapping states* in finite normal mixtures: when only one or a few observations are allocated to a component, the parameters for that component fit the few observations so tightly that few if any of the other observations would likely get allocated to the component. Reparametrizing the normal mixture model, as done by MENGERSON and ROBERT (1995), corrects the problem by keeping the component-specific parameters from fitting the observations in a near-empty component too tightly. However, this does not address trapping states that may occur simply because the mixing proportion for a component becomes small. If the mixing proportion of a component happens to reach a value near zero, then the probability of allocating any observations to that component will also be small, and the component may remain empty through many iterations of the chain.

The block-updating scheme of Section 5.3.1 can provide a useful Gibbs move that could be executed to restore empty components, by the following rationale: in a J-component finite mixture with a $\mathrm{Dir}(\zeta_1, \ldots, \zeta_J)$ prior on the mixing proportions, the latent allocation variables, Z_i, marginally follow a labelled compound multinomial-Dirichlet distribution. Consequently, conditional on current values of all the Z_i's, the subset of those having any

two values will follow a labelled beta-binomial distribution (JOHNSON *et al.* 1997). That is, the marginal distribution of $\{Z_i : Z_i = j_a \text{ or } Z_i = j_b, j_a \neq j_b, i = 1, \ldots, N\}$ follow a labelled beta-binomial distribution with parameters $(\zeta_{j_a}, \zeta_{j_b})$. Thus, the methods of Section 5.3.1 could be applied to redistribute elements amongst the two components j_a and j_b, having marginalized over the mixing proportions π_{j_a} and π_{j_b}. And so, observations may be reallocated to component j_a (or j_b), according to their full conditional distributions, even if π_{j_a} (or π_{j_b}) is close to zero.

Chapter 6

EXPLICIT MODELING OF THE HYBRIDIZATION PROCESS

The previous chapter described an extension to the model proposed by PRITCHARD *et al.* (2000) for individuals from structured populations. The PRITCHARD *et al.* model is a mathematically convenient model which seems to apply well to many situations of population structure. However, when more information is available about the nature of the admixing process, then a more detailed analysis of the situation is possible by using a more detailed model of the process. This chapter describes such a detailed model for a commonly encountered sitatuation in conservation biology—the case of two species which are known to have been hybridizing for a limited number of generations or in which the hybrids have fitness which is reduced to the extent that all of the admixed individuals in regions of overlap between the species are the result of recent hybridization. The model applied is one in which individuals belong to one of many different hybrid categories (*e.g.*, F_1, F_2, and various backcrosses), and inference will be done in a similar way to that in the previous chapter. Using Markov chain Monte Carlo (MCMC) we compute for each individual the posterior probability of inclusion in a particular hybrid category, and simultaneously compute the posterior distribution of the allele frequencies in the populations under study of the two different species.

I present these methods in the context of hybridization between sympatric populations of a species A and a species B, and develop the probability model that arises from explicitly modeling the hybridization process and using data on multiple, unlinked markers. This probability is the product of probabilities for single-locus genotypes conditional on the hybrid category. There is a simple calculus for these conditional probabilities which is described below. With the model thus specified, I describe how MCMC proceeds in similar fashion to the previous chapter.

The main goals of this chapter are to derive the method, apply it to real and simu-

lated data, and then describe several modifications and extensions suggested by the results of these data analyses. I apply the method to real genetic data (a sample of juvenile salmonids containing rainbow trout (*Oncorhynchus mykss*), cutthroat trout (*O. clarki*), and their hybrids) and then to simulated data with many relatively uninformative markers and again with fewer very informative markers. The results of these analyses, in Section 6.5, suggest that distinguishing between different hybrid categories beyond the F_1 stage requires a substantial amount of data, and clear genetic differentiation between the species. They also underscore the importance of having an established sampling scheme when trying to detect hybrids. These issues suggest further extensions to the method which are beyond the current scope of the thesis.

6.1 Population and Probability Model

We consider a group of individuals in the wild which consists of sympatric populations of two species A and B, and hybrids of the two species which have occurred from n potential generations of interbreeding. We take n to be known or assumed. We also assume that we have a sample of M individuals drawn from this group for genetic analysis. For now, we shall assume that individuals are sampled randomly and independently of whether they are of species A, or species B, or are a hybrid of the two species. This sort of sampling would arise, if, for example, juveniles were sampled and were very difficult to distinguish on the basis of morphology between the two species or hybrids thereof. In Section 6.6, I briefly discuss a modeling/sampling approach for relaxing that assumption. We have genotype information from L unlinked loci on the individuals sampled. Let the ℓ^{th} locus possess K_ℓ alleles detected in the sample. We denote the allele frequencies in species A and B, n generations ago, by $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$, respectively. Each of these $\boldsymbol{\Theta}$'s is a collection of vectors, with each vector giving the allele frequencies at a particular locus. For example, for species A, $\boldsymbol{\Theta}_A = (\boldsymbol{\theta}_{A,1}, \ldots, \boldsymbol{\theta}_{A,L})$, where $\boldsymbol{\theta}_{A,\ell} = (\theta_{A,\ell,1}, \ldots, \theta_{A,\ell,K_\ell})$ are allele frequencies at the ℓ^{th} locus. The alleles found in individuals from species A and species B are assumed to be drawn randomly from the allele frequencies $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$ respectively, n generations ago. Likewise, individuals n generations before sampling are assumed to be in Hardy-Weinberg

and linkage equilibrium with reference to their contemporaneous conspecifics.

Each individual in the sample is genotyped at L loci. The gene copies carried at any locus are considered to be ordered, though that order may be arbitrary; for example it may merely be the order in which the genetic data on that locus in that individual happened to be recorded. This ordering arises because we will shortly introduce latent data that applies to each particular gene copy. The allelic types of the two gene copies at locus ℓ in individual i are denoted by $\boldsymbol{Y}_{i,\ell} = (Y_{i,\ell,1}, Y_{i,\ell,2})$, with each of $Y_{i,\ell,1}$ and $Y_{i,\ell,2}$ taking an integer value between 1 and K_ℓ, inclusive, corresponding to the possible allelic types at the ℓ^{th} locus. The L single-locus genotypes in the i^{th} individual are denoted by $\mathbf{Y}_i = (\boldsymbol{Y}_{i,1}, \ldots, \boldsymbol{Y}_{i,L})$, and all of the genetic data, over all M individuals in the sample is $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_M)$. We do not know from which species each of an individual's gene copies descended, but we denote that unknown information by the latent variable $\boldsymbol{W}_{i,\ell} = (W_{i,\ell,1}, W_{i,\ell,2})$. $W_{i,\ell,1}$ takes the value 0 if the first gene copy at the ℓ^{th} locus of the i^{th} individual originated from the species A population, and it takes the value 1 if that gene copy originated from species B. $W_{i,\ell,2}$ takes values analogously, depending on the origin of the second gene copy. We use $\mathbf{W}_i = (\boldsymbol{W}_{i,1}, \ldots, \boldsymbol{W}_{i,L})$ to denote all the latent gene origin indicators in the i^{th} individual, and $\mathbf{W}$ denotes the latent gene origin indicators in all the individuals.

To develop the probability for the observed data, $\mathbf{Y}$, we must explicitly state what is meant by a "hybrid category." For unlinked markers such a definition is given in the following subsection in terms of "genotype frequency classes."

6.1.1 Hybrid categories

When hybridization between two species has been potentially occurring for n generations, the possible hybrid categories into which an individual may fall can be enumerated and described by considering the possible arrangements of different species amongst the founders in an n-generational pedigree. The individual of interest is taken to be the member at the bottom of the pedigree, and is assumed to be non-inbred over the last n generations; hence we assume there are no loops in its n-generational pedigree. Figure 6.1 illustrates this for the case of $n = 2$. Depending on the type of genetic data available (*e.g.*, linked versus

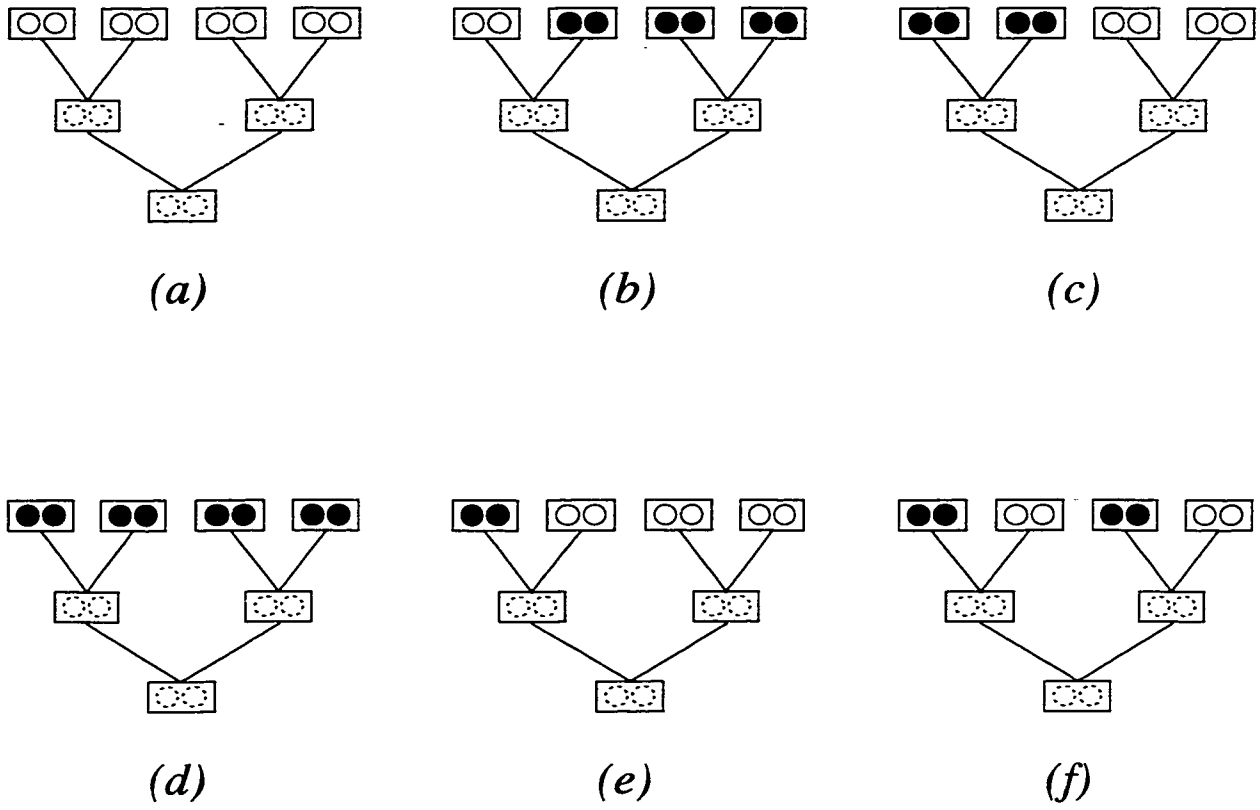


Figure 6.1: Six arrangements of founders on pedigrees of $n = 2$ two generations each. Each box represents a locus. The circles within each box represent the two gene copies possessed by the diploid organism at the locus. The founders are the individuals in the top row of each pedigree. Black gene copies amongst founders are those originating from the Species A population, and the white gene copies are from Species B. The individual at the bottom of each pedigree belongs to a different hybrid category, determined by the arrangement of species amongst the founders. (a) through (f) represent six distinct *hybrid classes*. (a) through (f) also represent six distinct *genotype frequency classes*. There are, however, only five distinct *gene frequency classes*; the individuals at the bottoms of pedigrees (c) and (f) are both in the same gene frequency class.

unlinked loci), it may not be possible to resolve all the possible arrangements of founders on the pedigree. I describe three ways of classifying these different arrangements (and hence the individuals at the bottom of the pedigree) the last two of which correspond to the level of resolution that is available in unlinked versus linked marker data.

At the simplest level, we can classify individuals into different *gene frequency classes* which correspond to the expected proportion of gene copies originating from one species or another within that individual. This is determined by the number of founders from each species on the pedigree. Since there are 2^n founders for a pedigree with n generations, there are $2^n + 1$ different gene frequency classes. Each class is determined by the number, a, of founders originating from the species A population ($a = 0, 1, \ldots, 2^n$). In Figure 6.1, both (c) and (f) belong to the gene frequency class with $a = 2$. The individuals at the bottoms of the other pedigrees belong to the remaining four distinct gene frequency classes.

Thinking in terms of gene frequency classes provides some perspective on the latent variable Q_i described in the previous chapter. PRITCHARD *et al.* (2000) introduced Q_i as the proportion of genome of the i^{th} individual originating from Population 0. Q_i reflects membership of the i^{th} individual in a gene frequency class as $n \rightarrow \infty$. As $n \rightarrow \infty$ the number of possible gene frequency classes becomes infinite—hence the continuous nature of Q_i in the previous chapter. In the present chapter, we will use Q to denote the proportion of an individual's genome derived from the Species A population; however, here, Q will be discrete in nature, rather than continuous. We refer to Q as the "genetic heritage proportion."

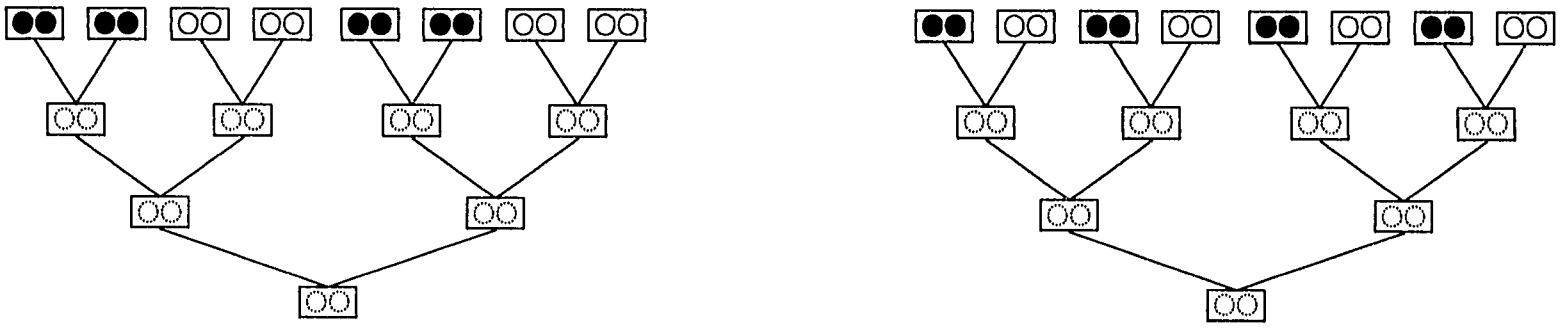
In recently hybridized populations, there is more information available than that used by considering only gene frequency classes. We may instead consider the different *genotype frequency classes* into which an individual may fall. The members of a genotype frequency class all share the same expected proportion of the three possible single-locus genotypes with respect to origin of the gene copies. These three genotypes are: 1) both gene copies originating from the Species A population; 2) one gene copy from Species A, the other from Species B; and 3) both gene copies from Species B . Enumerating these genotype frequency classes, and computing the expected proportions of the genotypes follows from Mendel's laws. Since each individual receives one gene copy randomly selected from the two in its mother, and another randomly selected from the two in its father, the expected proportions

of the genotypes in an individual are determined by the gene frequency classes to which its parents belong. We let $G_g = (G_{g,1}, G_{g,2}, G_{g,3})$ denote the expected proportions of the three different genotypes in an individual in genotype frequency class g; these proportions are

$$\begin{aligned} G_{g,1} &= Q_m Q_f \\ G_{g,2} &= 2Q_m(1-Q_f) \\ G_{g,3} &= (1-Q_m)(1-Q_f) \end{aligned} \tag{6.1}$$

where Q_m and Q_f are the genetic heritage proportions of the individual's mother and father, respectively. Straightforward algebra verifies that two individuals i and j will belong to the same genotype frequency class if and only if the parents of j belong to the same gene frequency classes as the parents of i. Consequently, the number of distinct genotype frequency classes after n generations of possible interbreeding between the species is the number of unordered pairs of different gene frequency classes after $n-1$ generations: $(2^{n-1}+1)(2^{n-1}+2)/2$. For $n \geq 2$ there are always more genotype frequency classes than there are gene frequency classes. With data on multiple unlinked loci, it is possible to distinguish between individuals in different genotype frequency classes. This is one of our inference goals, and in the Bayesian context will be pursued by computing the posterior distribution of a latent variable Z_i which takes the value g if the i^{th} individual belongs to genotype frequency class g.

Another classification of hybrids may be made into what I refer to as different *hybrid classes*. Members of the same hybrid class share the same number of founders from each species *and* the same arrangement of those founders, up to changes of branching order at any node on the binary tree of the pedigree. For $n > 2$ there are always more hybrid classes than there are genotype frequency classes. As an example, with $n = 3$, F_2 and F_3 hybrids are in different hybrid classes as shown by the different arrangements on the three generational pedigrees of Figure 6.2. Nonetheless, they are in the same genotype frequency class. With only unlinked markers, it is not possible to distinguish individuals that are in different hybrid classes, but in the same genotype frequency class. For the formulation with unlinked markers we will deal exclusively with the genotype frequency classes. A probability



(a) F_2 hybrid class

(b) F_3 hybrid class

Figure 6.2: Pedigrees of the F_2 and F_3 hybrid classes with $n = 3$. (a) is the pedigree of an F_2 individual following three generations of potential interbreeding, and (b) is the pedigree of an F_3 hybrid. Note that the gene frequency classes of the parents of the individual at the bottom of each pedigree are the same. Thus, the bottom individuals on each pedigree belong to the same genotype frequency class, even though they belong to different hybrid classes.

model with more complex dependence structure would be required for dealing with linked markers.

6.1.2 Probability of the data

Similar to the previous chapter we introduce notation here to avoid awkward subscripting: let $\theta_A\langle i;\ell;1\rangle$ denote the frequency in the species A population of the allele possessed by the i^{th} individual at the first gene copy of its ℓ^{th} locus. This is a shorthand for the doubly-subscripted $\theta_{A,\ell,Y_{i,\ell,1}}$. Similarly, for the second gene copy we will write $\theta_A\langle i;\ell;2\rangle$ and for the frequencies in the population from species B we have $\theta_B\langle i;\ell;1\rangle$ and $\theta_B\langle i;\ell;2\rangle$.

Given the population allele frequencies, the gene origin indicators, and the genotype frequency class to which an individual belongs, it is straightforward to compute the probability of that individual's single-locus genotype at the ℓ^{th} locus. For our purposes later, it is more useful to have an expression for the joint probability of the genotype and the gene origin indicators. In the i^{th} individual in the g^{th} genotype frequency class, this is

$$P(\boldsymbol{Y}_{i,\ell}, \boldsymbol{W}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell}) = \begin{cases} \theta_A\langle i;\ell;1\rangle\theta_A\langle i;\ell;2\rangle G_{g,1}, & \text{if } W_{i,\ell,1} = W_{i,\ell,2} = 0 \\ \theta_A\langle i;\ell;1\rangle\theta_B\langle i;\ell;2\rangle G_{g,2}/2, & \text{if } W_{i,\ell,1} = 0,\ W_{i,\ell,2} = 1 \\ \theta_B\langle i;\ell;1\rangle\theta_A\langle i;\ell;2\rangle G_{g,2}/2, & \text{if } W_{i,\ell,1} = 1,\ W_{i,\ell,2} = 0 \\ \theta_B\langle i;\ell;1\rangle\theta_B\langle i;\ell;2\rangle G_{g,3}, & \text{if } W_{i,\ell,1} = W_{i,\ell,2} = 1. \end{cases} \tag{6.2}$$

The product of the two allele frequencies in the above expressions follows from the assumption that each gene copy in the founders (n generations ago) of the i^{th} individual is sampled randomly from the alleles present in its population of origin. Then, $G_{g,1}$ is the probability that an individual in genotype frequency class g has both gene copies originating from species A, $G_{g,2}/2$ is the probability that the first (second) gene copy originates from species A and the second (first) originates from species B, and $G_{g,3}$ is the probability that both gene copies originated from speces B.

For a given genotype frequency class, the marginal probability of the i^{th} individual's

genotype at locus ℓ is computed by summing (6.2) over the latent gene origin indicators:

$$P(\boldsymbol{Y}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell}) = \sum_{\substack{0\le W_{i,\ell,1}\le 1 \\ 0\le W_{i,\ell,2}\le 1}} P(\boldsymbol{Y}_{i,\ell}, \boldsymbol{W}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell}). \tag{6.3}$$

And finally, under the assumption of unlinked markers in Hardy-Weinberg and linkage equilibrium among conspecifics n generations ago, the probability of the i^{th} individual's multilocus genotype is just the product over the L single-locus genotype probabilities:

$$P(\mathbf{Y}_i|Z_i = g, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B) = \prod_{\ell=1}^{L} P(\boldsymbol{Y}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell}). \tag{6.4}$$

This gives us an expression for the probability of the data on a single individual. We now must derive the probability for the data on all M individuals in the sample.

Given n generations of potential interbreeding between the species, there are $\mathcal{G}_n = (2^{n-1}+1)(2^{n-1}+2)/2$ genotype frequency classes that members of our genetic sample may fall into. We model the individuals in the sample as being randomly and independently drawn from a mixture of individuals, each belonging to one of the $\mathcal{G}_n$ genotype frequency classes with probability π_g, $g = 1, \ldots, \mathcal{G}_n$, $\sum_{g=1}^{\mathcal{G}_n} \pi_g = 1$. Using $\boldsymbol{\pi}$ to denote the vector of mixing proportions, $(\pi_1, \ldots, \pi_{\mathcal{G}_n})$, we may now write the probability of all the observed data, $\mathbf{Y}$, conditional on n, $\boldsymbol{\Theta}_A$, $\boldsymbol{\Theta}_B$ and $\boldsymbol{\pi}$ as the product over the members of the sample of the probability of each of their multilocus genotypes:

$$P(\mathbf{Y}|\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi}) = \prod_{i=1}^{M} \Big(\sum_{g=1}^{\mathcal{G}_n} \pi_g P(\mathbf{Y}_i|Z_i = g, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B) \Big). \tag{6.5}$$

6.2 A Bayesian Specification

Equation 6.5 is the likelihood for $\boldsymbol{\Theta}_A$, $\boldsymbol{\Theta}_B$, and $\boldsymbol{\pi}$. To pursue Bayesian inference in this problem requires prior distributions $P(\boldsymbol{\Theta}_A)$, $P(\boldsymbol{\Theta}_B)$, and $P(\boldsymbol{\pi})$, so that the posterior distribution may be computed. We wish to make inferences not only about $\boldsymbol{\Theta}_A$, $\boldsymbol{\Theta}_B$, and $\boldsymbol{\pi}$, but also the latent variables $\mathbf{W}$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_M)$, so we are concerned with the posterior distribution $P(\boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi}, \boldsymbol{Z}, \mathbf{W}|\mathbf{Y})$ and the marginalizations thereof. That posterior distribution is proportional to the joint probability of all those variables, $P(\mathbf{Y}, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi}, \boldsymbol{Z}, \mathbf{W})$. With

the latent variables present, this joint density factorizes neatly:

$$P(\mathbf{Y}, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B, \boldsymbol{\pi}, \boldsymbol{Z}, \mathbf{W}) = P(\boldsymbol{\Theta}_A)P(\boldsymbol{\Theta}_B)P(\boldsymbol{\pi}) \times \prod_{i=1}^{M} P(\mathbf{Y}_i|\mathbf{W}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)P(\boldsymbol{W}_i|Z_i)P(Z_i|\boldsymbol{\pi}), \tag{6.6}$$

which, as we will see in Section 6.3, allows straightforward MCMC sampling.

It is computationally convenient and biologically reasonable to take the specific form of the prior distributions $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$ to be Dirichlet distributions, independent over the L unlinked loci. That is, $P(\boldsymbol{\theta}_{A,\ell})$ is Dirichlet$(\lambda_{A,\ell,1}, \ldots, \lambda_{A,\ell,K_\ell})$. Since the Dirichlet distribution is the conjugate prior for the multinomial distribution, this choice facilitates simulation from the full conditional distributions for $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$. The Dirichlet distribution is also the multivariate generalization of the beta distribution which arises theoretically as the equilibrium distribution for gene frequencies in the presence of genetic drift and linear pressure from migration or mutation (WRIGHT 1938; WRIGHT 1952). Specification of the parameters $\boldsymbol{\lambda}_{A,\ell} = (\lambda_{A,\ell,1}, \ldots, \lambda_{A,\ell,K_\ell})$ and $\boldsymbol{\lambda}_{B,\ell} = (\lambda_{B,\ell,1}, \ldots, \lambda_{B,\ell,K_\ell})$ provides a way to incorporate prior information about the allele frequencies among the two species at the ℓ^{th} locus. For example, if at locus ℓ, previous studies have indicated that species B has a very low frequency of allele j while species A has a high frequency, then $\lambda_{B,\ell,j}$ should be chosen small, relative to the other components of $\boldsymbol{\lambda}_{B,\ell}$, while $\lambda_{A,\ell,j}$ should be chosen large. If one had very strong prior evidence that different species were fixed for different alleles, then this could also be incorporated in the prior in a similar manner. If, on the other hand, very little prior knowledge is available about allele frequencies in the two species, then a sensible choice of prior may be the Jeffreys prior (see GELMAN *et al.* (1996)). For species A this would be $\lambda_{A,\ell,j} = 1/K_\ell$ for $j = 1, \ldots, K_\ell$. Another "not-so-informative" prior density is the uniform Dirichlet distribution with $\lambda_{A,\ell,j} = 1$, $j = 1, \ldots, K_\ell$. This prior de-emphasizes the importance given to those alleles that only appear in several copies in the whole sample.

The conjugate prior for $\boldsymbol{\pi}$ is also a Dirichlet distribution, so we shall let $P(\boldsymbol{\pi}) \equiv$ Dirichlet$(\zeta_1, \ldots, \zeta_{\mathcal{G}_n})$. Here too, prior knowledge of the biology of the situation could be incorporated into the prior. For example, if it was well known that backcrosses between F_1 hybrids and species A had low fitness, then that could be reflected in the prior for $\boldsymbol{\pi}$.

Additionally if hybridization and backcrossing was known to be fairly rare, then this prior knowledge could be reflected by having smaller ζ_g's for those genotype frequency classes that required more episodes of interbreeding within the last n generations. In the absence of prior information on hybridization rates, the Jeffreys prior, $\zeta_g = 1/\mathcal{G}_n$, $g = 1, \ldots, \mathcal{G}_n$, is once again a suitable choice.

6.3 MCMC Simulation from the Posterior Distribution

It is not possible to compute directly the posterior distributions for the variables that we are interested in. However, simulating from the joint posterior distribution of all the variables by MCMC can be done via Gibbs sampling in a manner similar to that for normal finite mixture models (DIEBOLT and ROBERT 1994). After a sufficient period of burn-in, a sample of variables drawn from this joint posterior distribution allows Monte Carlo estimation of the posterior distribution of any subset of variables of interest, either marginally, or conditional on the values taken by another subset of variables. Given initial starting values for all the variables in the model, Gibbs sampling proceeds by successively simulating new values for particular variables in the model from their full conditional distributions (GEMAN and GEMAN 1984). I shall denote full conditional distributions by $P(\cdot|\cdots)$.

We shall refer to a standard iteration of our MCMC algorithm as a "sweep." A sweep consists of a series of steps in which each of the variables in the probability model (except for the data, $\mathbf{Y}$, which are fixed) is updated once. Here, the steps in a single sweep are

1. For $\ell = 1, \ldots, L$, simulate new values for $\boldsymbol{\Theta}_{A,\ell}$ and $\boldsymbol{\Theta}_{B,\ell}$ from their full conditional distributions, $P(\boldsymbol{\Theta}_A|\cdots)$ and $P(\boldsymbol{\Theta}_B|\cdots)$, respectively,

2. Simulate a new value of $\boldsymbol{\pi}$ from $P(\boldsymbol{\pi}|\cdots)$,

3. For $i = 1, \ldots, M$ and $\ell = 1, \ldots, L$, simulate a new value of $\boldsymbol{W}_{i,\ell}$ from $P(\boldsymbol{W}_{i,\ell}|\cdots)$,

4. For $i = 1, \ldots, M$, simulate a new value of Z_i from $P(Z_i|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$.

By sampling the current states of all the variables after each sweep, one aquires a dependent sample suitable for Monte Carlo estimation of most quantities of interest. In

particular, a Rao-Blackwellized Monte Carlo estimate of the posterior probability that individual i is of the g^{th} genotype frequency class is obtained by averaging the values of $P(Z_i = g|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$ computed during each sweep.

The full conditional distributions are easily derived. By conjugacy,

$$P(\boldsymbol{\theta}_{A,\ell}|\cdots) \equiv \text{Dirichlet}(\lambda_{A,\ell,1} + r_{A,\ell,1}, \ldots, \lambda_{A,\ell,K_\ell} + r_{A,\ell,K_\ell}) \tag{6.7}$$

where $r_{A,\ell,j}$ is the number of gene copies of allelic type j at the ℓ^{th} locus currently allocated to species A. (*i.e.*, gene copies of allelic type j for which the corresponding $W_{\cdot,\ell,\cdot} = 0$). An analogous expression exists for $P(\boldsymbol{\theta}_{B,\ell}|\cdots)$. Again by conjugacy

$$P(\boldsymbol{\pi}|\cdots) \equiv \text{Dirichlet}(\zeta_1 + s_1, \ldots, \zeta_{\mathcal{G}_n} + s_{\mathcal{G}_n}) \tag{6.8}$$

where s_g is the number of individuals in the sample currently allocated to genotype frequency class g.

The full conditional distribution for the pair of gene origin indicators $\boldsymbol{W}_{i,\ell}$ in the i^{th} fish currently included in the g^{th} genotype frequency class is obtained by Bayes Law:

$$P(\boldsymbol{W}_{i,\ell}|\cdots) = \frac{P(\boldsymbol{Y}_{i,\ell}, \boldsymbol{W}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell})}{P(\boldsymbol{Y}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell})} \tag{6.9}$$

where the numerator and denominator are given in (6.2) and (6.3), respectively. The full conditional distribution for Z_i would be $P(Z_i|\boldsymbol{W}_{i,\ell}, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell})$, however, it is more efficient to simulate new values of Z_i from $P(Z_i|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$, since this represents a marginalization over the four possible values of $\boldsymbol{W}_{i,\ell}$. $P(Z_i|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B)$ is easy to compute, because the quantities needed to calculate it by Bayes Law have already been computed in Step 3 of the sweep. By Bayes Law

$$P(Z_i = j|\mathbf{Y}_i, \boldsymbol{\Theta}_A, \boldsymbol{\Theta}_B) = \frac{\pi_g P(\boldsymbol{Y}_{i,\ell}|Z_i = j, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell})}{\sum_{g=1}^{\mathcal{G}_n} \pi_g P(\boldsymbol{Y}_{i,\ell}|Z_i = g, \boldsymbol{\theta}_{A,\ell}, \boldsymbol{\theta}_{B,\ell})}. \tag{6.10}$$

It is a good idea to run multiple chains from different starting values to diagnose mixing problems. In this case, it is easy to assign overdispersed starting values by simulating values of $\boldsymbol{\Theta}_A$, $\boldsymbol{\Theta}_B$, and $\boldsymbol{\pi}$ from their prior distributions rather than their full conditional distributions in Steps 1 and 2 of the first sweep. I have used GELMAN (1996)'s estimated scale reduction potential factor to assess how quickly chains in this problem converge to the

target distribution. In the case of data from cutthroat trout and steelhead trout described in the following section this occurs very rapidly. Burn in then requires little time.

6.4 *Steelhead × Cutthroat Trout Hybrids in Whiskey Creek*

Hybrids of steelhead trout, *O. mykiss*, and coastal cutthroat trout, *O. clarki.* in various streams on the West Coast have been detected using genetic data in several studies (CAMPTON and UTTER 1985; NEILLANDS 1990). In a large genetic survey conducted by the National Marine Fisheries Service and the Washington Department of Fish and Wildlife (as part of the preparation of a status report to determine extinction risk for coastal cutthroat trout) widespread evidence for hybridization between the two species was found. JOHNSON *et al.* (1999) summarize this survey, as well as the available literature from the field and the laboratory on hybridization between *O. clarki* and *O. mykiss.* They report that no severe developmental abnormalities occur in hybrids of the two species; hybrid offspring are clearly viable. However, hybrids may possess morphological and behavioral traits that reduce their fitness in natural environments. This accords well with the observation that hybrid individuals are typically detected among juvenile trout, but adult hybrids are seldom observed, and with the observation that although hybrization may occur each year (in cases where it has been monitored over time it has been found to be ongoing) the two species still remain distinct. Nonetheless, at some locales, some fish sampled and analyzed possess genotypes suggesting they belong to a hybrid class involving more than just one generation of hybridization. Here, I apply the methods of this chapter to the survey data from Whiskey Creek in Western Washington to see if it is possible to distinguish between F_1 and later hybrids.

As described in JOHNSON *et al.* (1999), 74 juvenile trout, believed to be cutthroat, were sampled from Whiskey Creek. The sample was not a random sample of the juvenile fish inhabiting Whiskey Creek because the biologists were expressly trying to sample cutthroat trout only. It is quite evident, however, that they were unsuccessful at capturing a sample only of cutthroat trout. In the following, the sample is treated as if it were a random sample, and, given the difficulty of distinguishing the two species as juveniles, this may

not be very inaccurate. It should be kept in mind, however, that careful estimates of the frequency of hybrid individuals would require data from a carefully planned study conducted for the purpose of identifying hybrids (rather than, as with the Whiskey Creek data, a study intended to estimate the allele frequencies in cutthroat trout alone). I briefly discuss the statistical modeling of different sampling scenarios in the discussion.

The 74 fish were genotyped at 50 enzyme loci. The occurrence at some loci of several allelic types typically found only at low frequency in cutthroat populations but at high frequency in steelhead populations led JOHNSON *et al.* (1999) to separate the 74 fish into a group of 48 putative cutthroat, 21 putative steelhead, and 5 putative hybrids. These classifications were based on two rules: 1) fish homozygous for the steelhead common allele at *ADA-2**, *mAH-2**, and *CKA-2** were classified as steelhead; and 2) fish possessing a steelhead common allele in four or more of the eight loci, *sAAT-4**, *ADA-2**, *mAH-2**, *mAH-3**, *CKA-2**, *IDDH-1**, *sIDHP-2** and *PEPA**, were classified as hybrids. David Teel of the National Marine Fisheries Service kindly provided me with the data on those 74 fish, typed at 50 loci from Whiskey Creek. Somehow, this dataset did not include the four supposedly informative loci *mAH-2**, *CKA-2**, *sAAT-4**, or *PEPA**. Even without those loci, however, the dataset is still suitable for demonstrating some important points about the methods developed within this chapter. In the remainder of this section I describe four different analyses, two with real data and two with simulated data, that I carried out. The results of these analyses appear in Section 6.5.

6.4.1 Four analyses for demonstration

I performed four analyses:

1. I first analyzed the data from Whiskey Creek using the same method as that used for the Scottish cat data in Chapter 5. The probability model underlying this analysis allows fish to belong either to a pure cutthroat or pure steelhead category, or to a generic "admixed" category. The analysis also provides a posterior distribution for the putatively admixed individuals' genetic heritage proportions. Of the 50 loci available, 30 were polymorphic in the sample from Whiskey Creek and were used for

Table 6.1: Genotype frequency classes assumed for the analyses. 1–5 and 7 are the six genotype frequency classes that arise given $n = 2$ generations of potential interbreeding. $G_{g,1}$, $G_{g,2}$, and $G_{g,3}$ are as described in the text. The final column gives names that I use to refer to these genotype frequency classes. Classes 6 and 8 require $n = 3$ generations of potentical interbreeding, as they would be formed by Cutt Bx × Cutt Bx and St Bx × St Bx pairings, respectively.

g	Q	$G_{g,1}$ (A,A)	$G_{g,2}$ (A,B) or (B,A)	$G_{g,2}$ (B,B)	Name
1	1.00	1.0000	0.0000	0.0000	Pure Cutt
2	0.00	0.0000	0.0000	1.0000	Pure St
3	0.50	0.0000	1.0000	0.0000	F_1
4	0.50	0.2500	0.5000	0.2500	F_2
5	0.75	0.5000	0.5000	0.0000	Cutt Bx
6	0.75	0.5625	0.3750	0.0625	Cutt $(Bx)^2$
7	0.25	0.0000	0.5000	0.5000	St Bx
8	0.25	0.0625	0.3750	0.5625	St $(Bx)^2$

the analysis. For the MCMC I used 5 different chains, each with 2000 sweeps for burn in and then 40,000 sweeps during which samples were collected every sweep from each chain. This required 12 hours on a laptop computer with a 266 MHz G3 (Macintosh) processor.

2. I then re-analyzed the data from Whiskey Creek using the methods developed in the present chapter. For this analysis, I assumed there were 8 genotype frequency classes to which individuals might belong—the six classes arising from $n = 2$ generations of potential interbreeding, as well as two more corresponding to offspring generated by breeding between two backcrossed individuals. Table 6.1 lists the expected proportions of the different single locus genotypes in these eight classes, and also gives the names that I use to refer to them. I use these names only as representative "type" names—for example, it should be kept in mind that the "F_2" genotype frequency class contains other indistinguishable hybrid classes as well, such as F_3. Though prior information

is available on allele frequencies in other steelhead and cutthroat populations, I chose to do this analysis using the Jeffreys prior ($\lambda_{\ell,j} = 1/K_\ell$, $j = 1, \ldots, K_\ell$) for allele frequencies. I also used the Jeffreys prior ($\zeta_g = 1/8$, $g = 1, \ldots, 8$) for the mixing proportions (π) of the different genotype frequency classes. 100,000 sweeps of five chains started from overdispersed starting values were run. This required 5 hours on the same laptop computer with the 266 Mhz G3 (Macintosh) processor.

3. From the posterior mean estimates in Analysis 2 of the allele frequencies in the cutthroat and the steelhead populations of Whiskey Creek, I simulated a new sample of 391 fish. The purpose of this was to see whether inferences about genotype frequency classes could be made any more sharply with many more individuals in the sample, but still with loci that were roughly as informative as those in the Whiskey Creek dataset. It is also, of course, instructive to analyze simulated data in which the "truth" is known. In this case I included in the sample 200 Pure Cutt, 150 Pure St, 20 F_1, 10 F_2, 5 St Bx, and 2 each of Cutt Bx, Cutt $(Bx)^2$, and St $(Bx)^2$ individuals. I once again used a Jeffreys prior for allele frequency, but used the uniform Dirichlet prior ($\zeta_g = 1$, $g = 1, \ldots, 8$) for the mixing proportions π. 35,000 sweeps of five chains with different starting values required 8 hours.

4. The last dataset analyzed is another simulated dataset with 391 individuals and the same numbers of individuals from each of the 8 genotype frequency classes as in Analysis 3. However, genetic data were simulated for 12 diallelic loci with very large frequency differences between the two species—at each locus, species A has an allele at frequency .995, which is at frequency .005 in species B. This was designed to mimic the situation in which a researcher has a moderate number of nearly diagnostic loci. The analysis was once again carried out using the Jeffreys prior for allele frequencies and a uniform Dirichlet prior on the mixing proportions. 45,000 sweeps of five chains, each with overdispersed starting values, required 5.1 hours.

6.5 Results

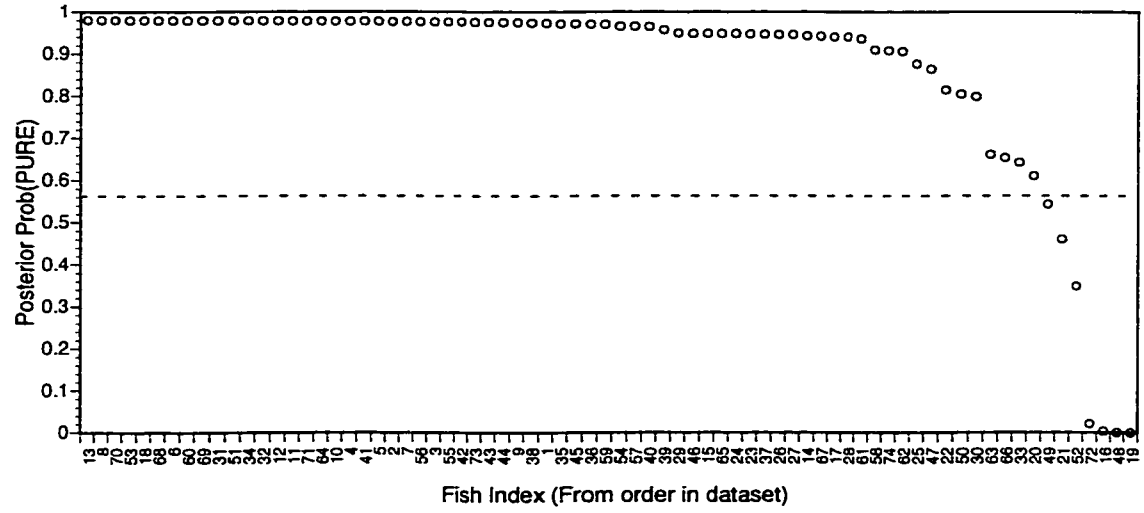
Analysis 1

The results of the simultaneous mixture/admixture analysis of Chapter 5 on the Whiskey Creek dataset motivated, to some extent, the development of the techniques in this chapter. Figure 6.3(a) shows the posterior probability that each of the 74 fish in the sample is Pure Cutt or Pure St. Seven of those fish have posterior probabilities of being purebred less than .55. In Figure 6.3 are posterior distributions of the genetic heritage proportions Q_i of these seven fish, suggesting that perhaps they belong to different hybrid classes. Fish 48, 19, and 16, have posterior means for Q_i that are near .5. This would be the genetic heritage proportion of either an F_1 or F_2 hybrid; the analysis of Chapter 5 is not able to distinguish those two different hybrid classes. Fish 49, has a Q_i near .25, which would correspond to the St Bx or St $(Bx)^2$ categories. Fish 21 and 52, on the other hand, appear to have roughly 3/4 of their ancestry from the cutthroat species suggesting that they may belong to the Cutt Bx or Cutt $(Bx)^2$ categories. Once again, it is not possible to distinguish between those possibilities without explicitly modeling the hybridization process as done in the next analysis.

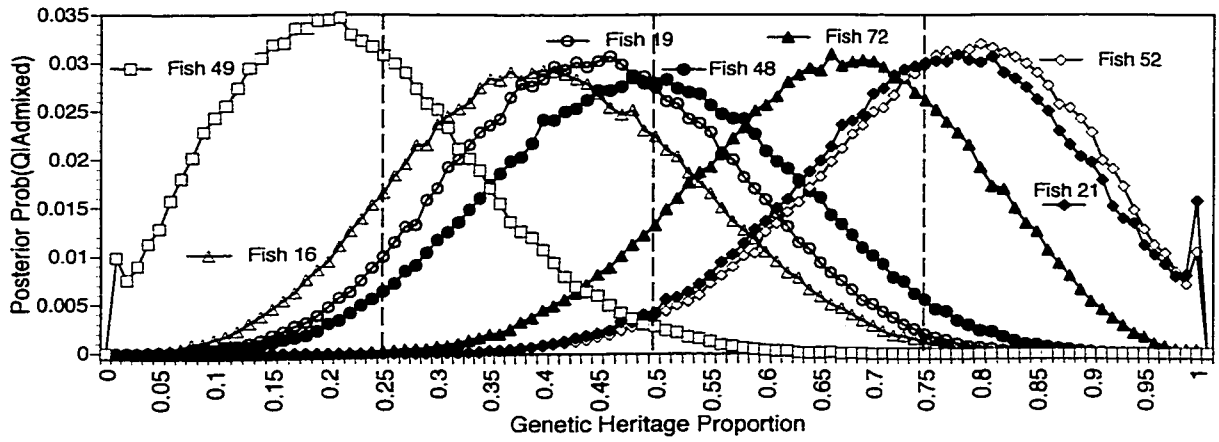
Analysis 2

In this analysis, each fish is assigned a posterior probability of belonging to each one of the 8 different genotype frequency classes. Two of those classes are purebred categories (Pure Cutt and Pure St). All the individuals with a posterior probability of being Pure Cutt greater than .5 have a negligible posterior probability of being Pure St, and vice versa. For these individuals, Figure 6.4(a) shows their posterior probabilities of being Pure Cutts (open circles) or Pure St (closed cirles), respectively. Only three fish have posterior probability of being purebred less than .55. These three are 48, 19, and 16. Fish 72, 49, 21 , and 52 are no longer among them—they all have posterior probabilities greater than .7 of being purebred.

Figure 6.4 shows the posterior probabilities that fish 48, 19, and 16 belong to each of the eight different genotype frequency classes. Fish 19 has posterior probability of being an F_1 hybrid that is more than three times that of being in the F_2 category. Nonetheless,

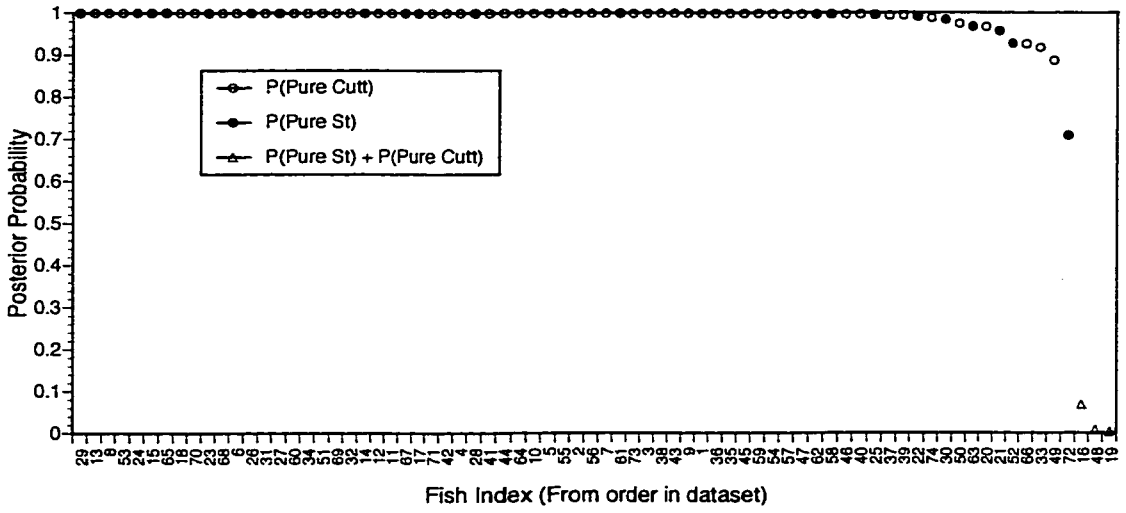


(a) P(PURE) by Chapter 5 Methods

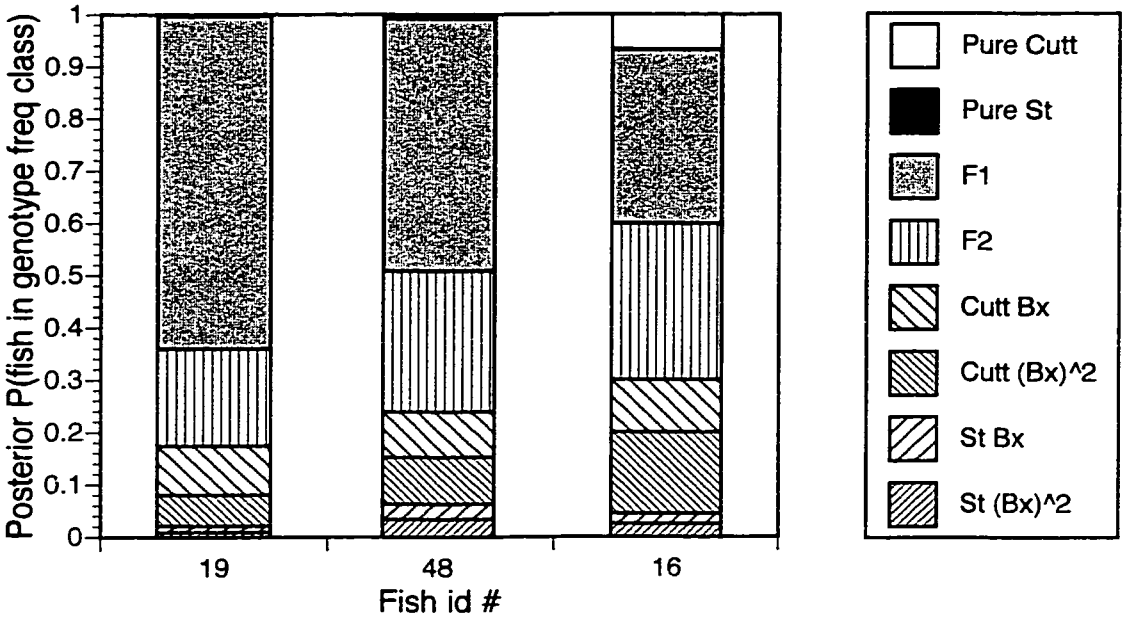


(b) Histograms showing $P(Q_i|\mathbf{Y})$ for 7 fish

Figure 6.3: Results from analysis 1. (a) shows the posterior probability that an individual is purebred (Pure Cutt or Pure St) for all 74 fish ordered by that posterior probability. The labels of the fish (as given in the dataset) are on the horizontal axis. Four fish have high posterior probability of being admixed—19, 48, 16, and 72, while three more have probability of being admixed greater than .45—52, 21, and 49. (b) Q_i for the seven fish that appear most admixed: 48,19, and 16 have genetic heritage proportions around that expected for F_1 or F_2 hybrids while 49, 21, and 52 have Q_i's closer to those expected for backcrossed hybrids.



(a) P(PURE) with 8 genotype frequency classes



(b) Posterior probabilities of genotype frequency class for 3 fish

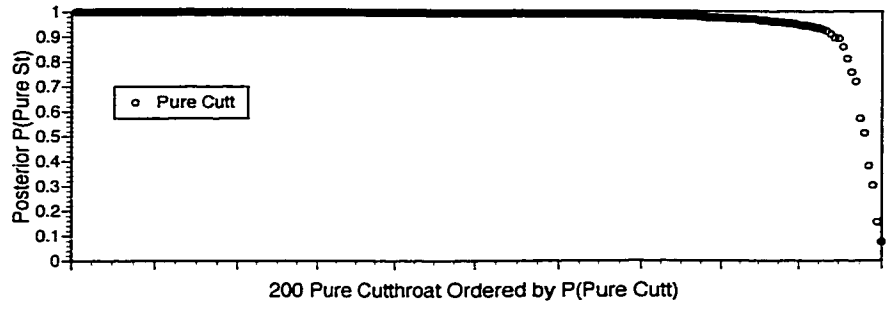
Figure 6.4: Results from Analysis 2. Description in text.

the posterior probability that it belongs to a genotype frequency class other than F_1 is .36; there is not sufficient evidence to assign it to a single genotype frequency class with any sort of confidence. This is even more true for fish 48 and 16.

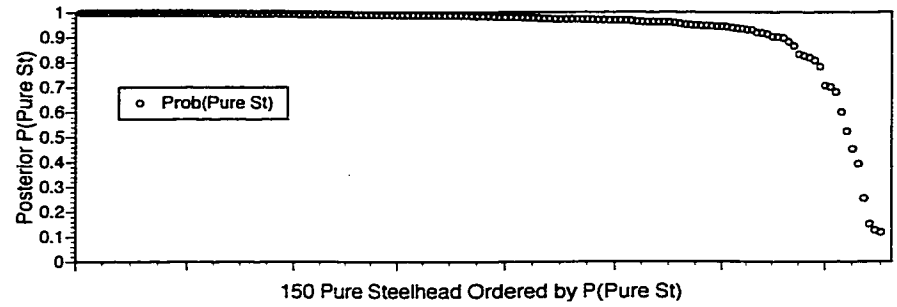
Analysis 3

The posterior probability that each of the 200 simulated cutthroat trout belong to the Pure Cutt class is shown in Figure 6.5(a). Most of these fish have high posterior probability of being Pure Cutts. The same is true for the 150 simulated steelhead—all but a few of them have posterior probability greater than .9 of being in the Pure St category. In other words, with data similar to those from Whiskey Creek, and with many purebred individuals of each species sampled, it is unlikely that any purebred individual will receive high posterior probability of being in a non-purebred genotype frequency category. Figure 6.5(c) shows the probability of being in either of the Pure Cutt or Pure St categories for the 41 simulated hybrid fish. The different symbols denote the true genotype frequency class of each fish. The F_1 and F_2 individuals are most readily indentified as being hybrids of some sort. The various backcrossed categories, however, contain some members with high posterior probabilities of being Pure Cutt or Pure St—they are not easily detected as hybrids.

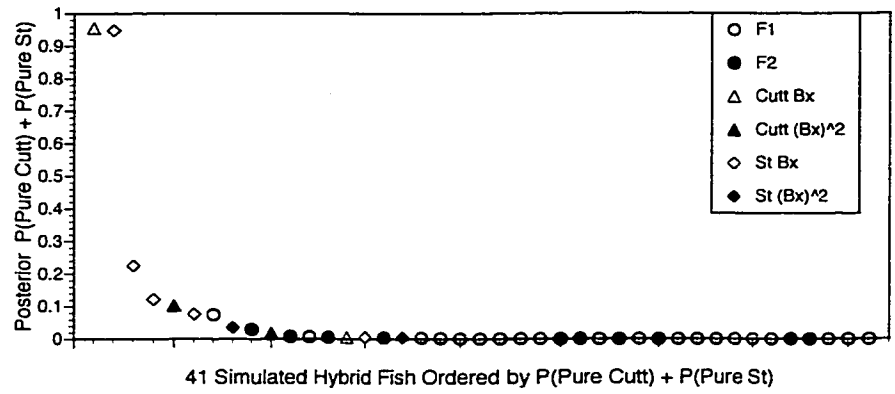
It appears difficult to make clear distinctions between the different hybrid genotype frequency classes with the genetic data used. Figure 6.6(a) shows the posterior probabilities of inclusion in the 8 genotype frequency classes for the 20 F_1 hybrids in descending order of the posterior probability that they are F_1's. While many of them have high posterior probability of being F_1, there is also one with posterior probability greater than .5 of being in the Pure St category. For the non-F_1 hybrids, the situation is even less promising. Of the 10 F_2 individuals (the first 10 individuals in Figure 6.6(b)), not one of them has posterior probability greater than .5 of being in the F_2 category. Finally, for the backcrossed genotype frequency classes, as the final 11 columns in Figure 6.6(b) reveal, there is little relationship between a fish's true genotype frequency class and the estimated posterior probability of being in that class.



(a) P(Pure Cutt) for 200 simulated cutthroat trout

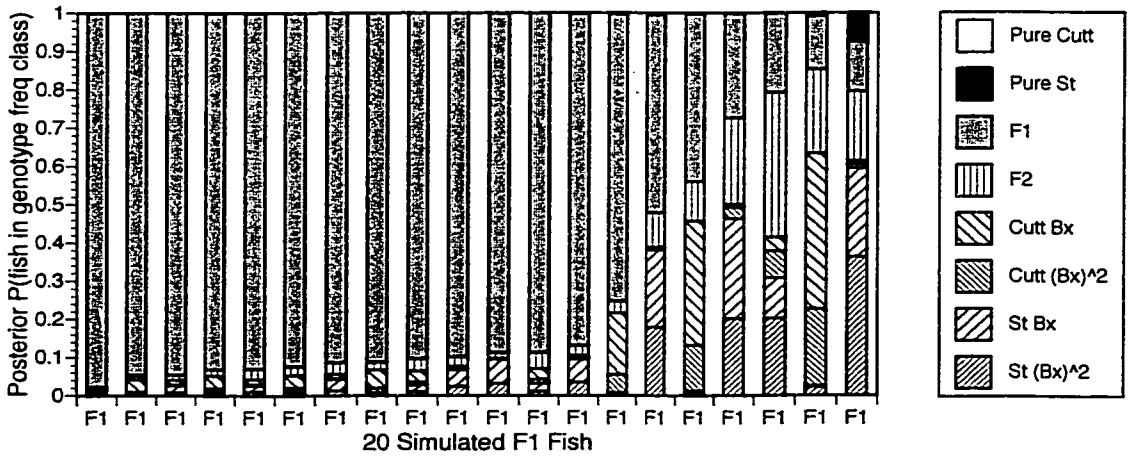


(b) P(Pure St) for 150 simulated steelhead trout

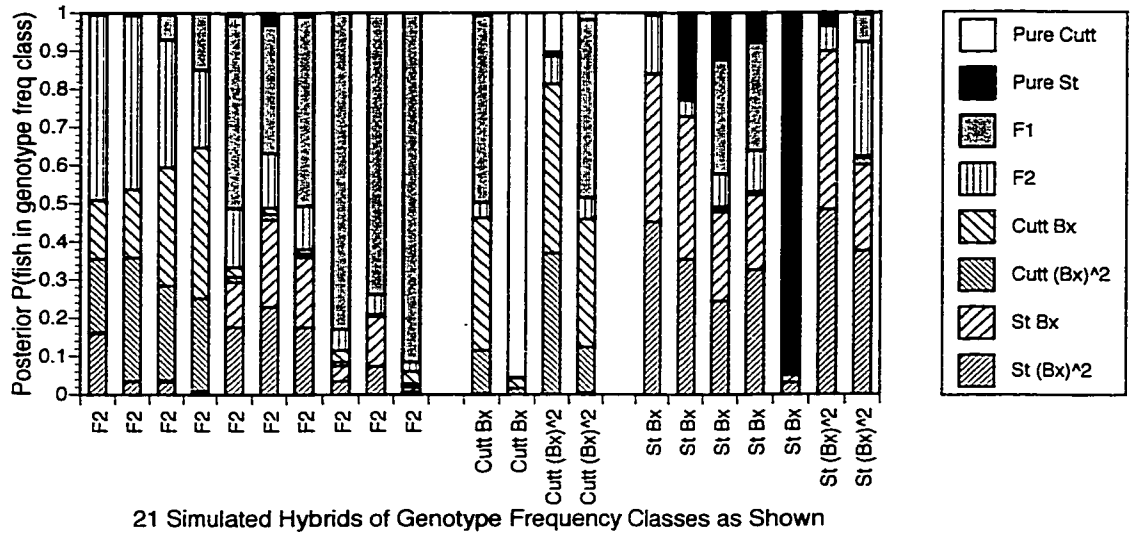


(c) Probability of pure descent of 41 simulated, hybrid juvenile trout

Figure 6.5: Results from Analysis 3. Description in text.



(a) 20 simulated F_1's



(b) 21 simulated hybrids as indicated

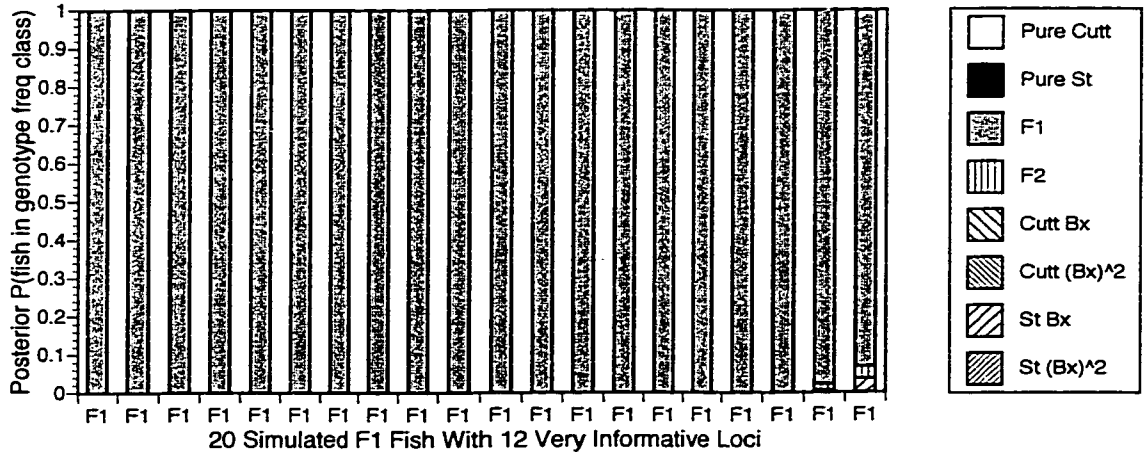
Figure 6.6: More results from Analysis 3. Description in text.

Analysis 4

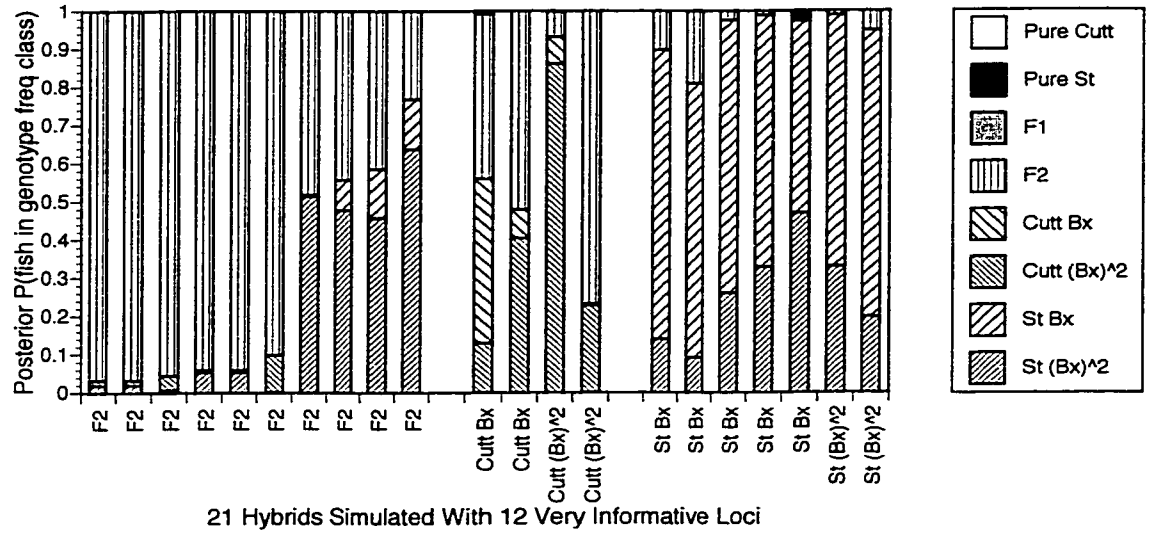
This analysis reveals that, even if all the loci available have great allele frequency differences between the species, many loci are still required to distinguish individuals from genotype frequency classes beyond the two pure categories and F_1's. All the simulated Pure Cutt and Pure St fish had very high posterior probabilities of being from their respectively correct categories. The same is true of the F_1's as shown in Figure 6.7(a). Each of them has posterior probability near 1 of being from the F_1 category. However, for the F_2's and beyond, the story is mixed. Four of the F_2's have posterior probability greater than .9 of being in the F_2 category. However, four of them also have posterior probabilities lower than .5. Within the backcrossed categories, distinctions are even less reliable. It is not difficult to see why this might be the case with as few as 12 loci, even if all the loci were perfectly diagnostic (*i.e.*, there were only private "cutthroat" and "steelhead" alleles at each). For example, an individual in the St $(Bx)^2$ category, with probability $(1 - .0625)^{12} = .46$, will have no single locus genotypes homozygous for cutthroat origin, and will thus look very much like a St Bx individual. For all of these 21 second-generation (or beyond) hybrids the posterior probability of being in a genotype frequency class "beyond" F_1 is quite high ($> .99$ for all but one). In that sense, they may be reliably categorized as "F_2 or backcrossed" by the method.

6.6 Discussion

I have described a method for Bayesian inference in populations of recently hybridized species. The four short exercises in data analysis using the method bring three main issues to attention. The first is that it requires many markers, and substantial genetic differentiation between the species at those markers to be able to reliably distinguish hybrids in classes beyond the F_1 class. For many problems, it may be preferable to include only four categories analogous to Pure Cutt, Pure St, F_1, and "F_2 and beyond." Alternatively, one could plan to collect data on many loci. Doing so, however, increases the chance that the assumption of unlinked loci will be violated. It would be possible to model the case of linked loci if one knew the recombination fractions between loci. The genotype frequency class at one marker would



(a) 20 simulated F_1's



(b) 21 simulated hybrids as indicated

Figure 6.7: Analysis 4 results. Description in text.

then depend on the genotype frequency classes at the other markers on the chromosome. Just like the identity-by-descent process along homologous chromosomes, this "genotype frequency class" process would be non-Markovian. It is from this non-Markovian process that one would gain the information (though probably very little) to distinguish different hybrid classes within the same genotype frequency class using linked markers. In a strict analysis, linked markers and the possibility of linkage disequilibrium among conspecifics at those markers would also necessitate the use of haplotype frequency parameters rather than the allele frequency parameters $\boldsymbol{\Theta}_A$ and $\boldsymbol{\Theta}_B$. While the necessary computations for this sort of model could all be done in terms of the underying vector of segregation patterns at each locus, the use of some sort of approximation, for example a Markov approximation to the "genotype frequency class" process along the chromosome as taken by McKeigue (1998) in a related problem, is also a possibility.

Another issue is the importance of the sampling model used. In the case of the model as described in this chapter, individuals are assumed drawn at random from a population which is a mixture in the proportions $\boldsymbol{\pi}$ of individuals from the different genotype frequency classes. This is violated in the case of the cutthroat trout data, because when the biologists collected the specimens, they were explicitly trying to obtain pure cutthroat, and hence throwing back those individuals that looked like steelhead or hybrids. Clearly, it must not be easy to distinguish cutthroat juveniles from steelhead or hybrid juveniles on the basis of morphological characters. In order to estimate accurately the proportion of hybrids in a locale, or even to estimate accurately the posterior probability that an individual is a hybrid, it is imperative to design the study with those goals in mind. Having an explicit model, like the one described in this chapter, that includes the sampling of the organisms is an asset, since the model may be tailored to particular sampling schemes. For example, it would be possible to model stratified sampling in which sampled organisms were first put into "possibly hybrid" and "probably purebred" categories on the basis of their morphological traits, and then a random subset of individuals from each of those categories was genetically typed. Extending the present model to such a scenario requires only that two different mixing proportions be used for the respective samples, *i.e.*, a vector $\boldsymbol{\pi}_h$ for the "possibly hybrid" sample and a $\boldsymbol{\pi}_p$ for the "probably pure" sample. The specification

of the latent data and allele frequency parameters would remain the same, and the MCMC would proceed with very little modification. Such a sampling scheme might be very useful for species in which hybridization is rare, and leads to some morphological distinction of the hybrid individuals that could be used to concentrate sampling upon them.

Finally, the analysis of the real data exposes a difficulty in the Bayesian analysis of finite mixtures. In this case, there are only three individuals with high posterior probability of being hybrids, yet there are 6 different hybrid genotype frequency classes. Hence, much of the time during the running of the Markov chain for MCMC, some of the components corresponding to those genotype frequency classes will be empty, and they will never have many individuals allocated to them. This makes the results more sensitive to the prior chosen for $\boldsymbol{\pi}$ than it would be if all the genotype frequency classes were well represented in a large sample, and provides another argument for reducing the number of components used in the model to two pure categories, an F_1 category, and then the "F_2 and beyond" category. Fortunately, however, in the Gibbs sampling, individuals still mix easily between the empty or nearly-empty genotype frequency classes. This contrasts with the Bayesian analysis of normal mixtures in which trapping states are often encountered with nearly-empty components (DIEBOLT and ROBERT 1994). The distinction here occurs because the component-specific parameters are the fixed quantities $\boldsymbol{G}_g$, which are not affected by the number of fish allocated to each component. This feature would also make it straightforward to implement a reversible jump MCMC (GREEN 1995) sampler to allow the number of genotype frequency classes to be modeled as an unknown random variable whose posterior distribution was to be determined.

Chapter 7

CONCLUSIONS

Each of the preceding five chapters presents a Monte Carlo method for computing the likelihood function or the posterior probability in an inference problem in population genetics. The Monte Carlo method is a computational tool for approximating expectations, and, therefore, many of the advances presented in this dissertation are concerned with methods for computing particular quantities. However, the novel developments herein are not to be found solely in the computational methods—each chapter includes some degree of stochastic modeling necessary to actually define the quantities (likelihoods, posterior probabilities, *etc.*) that are to be computed. While some chapters have a computational focus, others deal more directly with the underlying modeling issues. However, in each case, the computational and modeling elements are complementary to one another in a synergistic fashion—with the new computational methods it is possible to do inference using stochastic models that are more faithful representations of the actual population genetic processes than the models previously used. Likewise, judicious modeling choices may simplify the computational challenges that must be confronted. Here I summarize the work in each of Chapters 2 to 6 with particular attention to the distinction between the stochastic modeling and the computational component of each.

Chapter 2 focuses heavily on a computational method—importance sampling—to compute the likelihood for the effective size N_e of a population with discrete generations. The stochastic modeling involved is standard, with the genetic samples taken from the population being easily recognized as observations of a hidden Markov chain. This recognition, however, makes possible the importance sampling method which employs forward-backward methods for hidden Markov chains (BAUM *et al.* 1970) and a variate transformation previously used in theoretical genetics (CAVALLI-SFORZA and EDWARDS 1967).

Chapter 3, on the other hand, deals primarily with issues of the stochastic modeling of

genetic inheritance in populations. It details some of the inaccuracies of the Wright-Fisher model as regards fixation probabilities when the census size of the population is known to be larger than the variance effective size of the population. It then describes a different model for genetic inheritance. This model is based on a Pólya urn scheme, and is a special case of the conditional branching process models investigated by KARLIN and MCGREGOR (1965). This urn scheme provides a better model of genetic inheritance in populations of known census size, and is used to define a stochastic model that depends on the parameter λ, the ratio of effective to census breeders in a population. At the end of Chapter 3, it is shown that implementing Markov chain Monte Carlo (MCMC) using the urn model is simpler than trying to implement MCMC while adhering strictly to the Wright-Fisher model. This is thus a case in which the more accurate stochastic model also carries certain computational advantages.

The main modeling development in Chapter 4 is a stochastic model for genetic transmission that faithfully represents the life history features of salmon populations with semelparous individuals maturing at different ages. This model uses the urn model to represent genetic inheritance between different components of the salmon population (*e.g.*, juveniles and adults of different ages). The connections between these different population components are summarized by a directed graph between them. The first half of the chapter provides a clear example of how graphical modeling can assist in the development, portrayal, and understanding of probabilistic dependence in complex biological systems. The second half of the chapter shows how, with the neighborhood structure easily inferred from the graph, an MCMC algorithm may be implemented to estimate λ in salmon populations.

Chapters 5 and 6 both deal with genetic admixture. Chapter 5 is more concerned with computational issues than modeling. There are a few very slight modeling concerns in extending the model of PRITCHARD *et al.* (2000) to include a purebred category, however, the primary advances in the chapter are 1) the application of a forward-backward algorithm to compute the probability of an individual's multilocus genotype, and 2) the development of a reversible jump MCMC scheme that allows computation of the Bayes factor for comparing the model with purebred categories to the model without those categories. Chapter 6 employs standard MCMC techniques for the Bayesian analysis of mixtures, but it does so

within a different modeling framework than that pursued in Chapter 5. While the model adopted in Chapter 5 derives from an heuristic model for structured populations, proposed by PRITCHARD *et al.* (2000) to encompass a wide range of population scenarios, the efforts of Chapter 6 are directed toward explicitly modeling the hybridization process between different species. This allows a more natural and interpretable analysis of genetic variation in sympatric populations of occasionally hybridizing species.

In conclusion, I have presented five applications of Monte Carlo methods to inference from population genetics data. These methods allow inferences to be made based on the likelihood function in the frequentist setting, or on the posterior probability function in the Bayesian setting. These Monte Carlo methods allow the computation of likelihoods or posterior probabilities from complex models, and this, in turn, permits a greater degree of biological reality in the stochastic models used.

There is considerable room for future work in the field of inference from population genetics data. In some of the chapters, I have indicated extensions that could be made. There are also certainly some estimation problems in conservation genetics that would require either custom-tailoring of the models used here or completely new models altogether. The treatment of inference problems in this dissertation provides several examples of the use of Monte Carlo, likelihood and Bayesian analysis, and graphical modeling, and I hope that these examples will illuminate and stimulate the future application of these useful tools to problems in conservation genetics.

BIBLIOGRAPHY

ALLENDORF, F. W. and R. S. WAPLES, 1996 Conservation and genetics of salmonid fishes. In J. C. Avise and J. L. Hamrick (Eds.), *Conservation Genetics: Case Histories from Nature*. New York: Chapman and Hall.

ANDERSON, E. C. and E. A. THOMPSON, 1999 MCMC likelihoods for population genetics. In *Proceedings of the 52nd Session of the International Statistical Institute, 3*, pp. 347–348.

ANDERSON, E. C., E. G. WILLIAMSON, and E. A. THOMPSON, 2000 Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. Genetics **156:** 2109–2118.

ARDREN, W. R., 1999 Effective number of breeders, rates of loss of genetic variability, and population productivity in two steelhead trout populations: implications for conservation and restoration. Ph.D. Thesis, University of Minnesota.

AVISE, J. C., S. M. HAIG, O. A. RYDER, M. LYNCH, and C. J. GEYER, 1995 Descriptive genetic studies: applications in population management and conservation biology. In J. D. Ballou, M. Gilpin, and T. J. Foose (Eds.), *Population Management for Survival and Recovery*, pp. 183–244. New York: Columbia University Press.

BARTLEY, D., M. BAGLEY, G. GALL, and B. BENTLEY, 1992 Use of linkage disequilibrium data to estimate effective size of hatchery and natural populations. Conservation Biology **6:** 365–375.

BAUM, L. E., 1972 An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In O. Shisha (Ed.), *Inequalities–III: Proceedings of the Third Symposium on Inequalities Held at the University of California, Los Angeles, September 1–9, 1969*, pp. 1–8. New York: Academic Press.

BAUM, L. E., T. PETRIE, G. SOULES, and N. WEISS, 1970 A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains. Annals of Mathematical Statistics **41:** 164–171.

BEACHAM, T. D., 1996 The use of minisatellite DNA variation for stock identification of chum salmon, *Oncorhynchus keta*. Fisheries Bulletin **94:** 611–627.

BEACHAM, T. D., K. D. LE, M. R. RAAP, K. HYATT, W. LUEDKE, and R. E. WITHLER, 2000 Microsatellite DNA variation and estimation of stock composition of sockeye salmon, *Oncorhynchus nerka*, in Barkley Sound, British Columbia. Fishery Bulletin **98:** 14–24.

BEAMSDERFER, R. C. P., H. A. SCHALLER, M. P. SIMMERMAN, C. E. PETROSKY, O. P. LANGNESS, and OTHERS, 1998 Spawner-recruit data for spring and summer chinook salmon populations in Idaho, Oregon, and Washington. In D. R. Marmorek and C. N. Peters (Eds.), *Plan for Analyzing and Testing Hypotheses (PATH): Retrospective Analyses of Spring/Summer Chinook Reviewed in FY 1997.* Vancouver, Canada: ESSA Technologies.

BEAUMONT, M., 1999 Detecting population expansion and decline using microsatellites. Genetics **153:** 2013–2029.

BEAUMONT, M., E. M. BARRATT, D. GOTELLI, A. C. KITCHENER, M. J. DANIELS, J. K. PRITCHARD, and M. W. BRUFORD, 2001 Genetic diversity and introgression in the Scottish wildcat. Molecular Ecology **10:** 319–336.

BEGON, M., C. B. KRIMBAS, and M. LOUKAS, 1980 The genetics of *Drosophila subobscura* populations XV. Effective size of a natural population estimated by three independent methods. Heredity **45:** 335–350.

CAMPTON, D. E. and F. M. UTTER, 1985 Natural hybridization between steelhead trout (*Salmo gairdneri*) and coastal cutthroat trout (*Salmo clarki clarki*) in two Puget Sound streams. Canadian Journal of Fisheries and Aquatic Sciences **42:** 110–119.

CANNINGS, C., 1974 The latent roots of certain Markov chains arising in genetics: a

new approach 1. Haploid models. Advances in Applied Probability **6:** 260–290.

CAVALLI-SFORZA, L. L. and W. F. BODMER, 1971 *The Genetics of Human Populations.* San Francisco: W. H. Freeman.

CAVALLI-SFORZA, L. L. and A. W. F. EDWARDS, 1967 Phylogenetic analysis: models and estimation procedures. Evolution **21:** 550–570.

CLIFFORD, S. L., P. MCGINNITY, and A. FERGUSON, 1998 Genetic changes in Atlantic salmon (*Salmo salar*) populations of northwest Irish rivers resulting from escapes of adult farm salmon. Canadian Journal of Fisheries and Aquatic Sciences **55:** 358–363.

DEFINETTI, B., 1972 *Probability, Induction and Statistics. The Art of Guessing.* New York: John Wiley & Sons.

DELLAPORTAS, P. and J. J. FORSTER, 1999 Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika **86:** 615–613.

DEVLIN, R. H., B. K. NCNEIL, T. D. D. GROVES, and E. M. DONALDSON, 1991 Isolation of a Y-chromosomal DNA probe capable of determining genetic sex in chinook salmon (*Oncorhynchus tshawytscha*). Canadian Journal of Fisheries and Aquatic Sciences **48:** 1606–1612.

DIEBOLT, J. and C. P. ROBERT, 1994 Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Socity, Series B **56:** 363–375.

ELO, K., J. ERKINARO, J. A. VUORINEN, and E. M. NIEMELAE, 1995 Hybridization between Atlantic salmon (*Salmo salar*) and brown trout (*S. trutta*) in the Teno and Naeaetaemoe River systems, northernmost Europe. Nordic Journal of Freshwater Research **70:** 56–61.

ELO, K., S. IVANOFF, J. A. VUORINEN, and J. PIIRONEN, 1997 Inheritance of RAPD markers and detection of interspecific hybridization with brown trout and Atlantic salmon. Aquaculture **152:** 55–65.

EWENS, W. and R. SPIELMAN, 1995 The transmission/disequilibrium test: history, subdivision, and admixture. American Journal of Human Genetics **57:** 455–464.

EWENS, W. J., 1979 *Mathematical Population Genetics.* New York: Springer-Verlag.

FELLER, W., 1957 *An Introduction to Probability Theory and Its Applications, 2nd Edition.* New York: John Wiley & Sons.

FRANKHAM, R., 1995 Inbreeding and extinction: a threshold effect. Conservation Biology **9:** 792–799.

FREEDMAN, D. A., 1965 Bernard Friedman's urn. Annals of Mathematical Statistics **36:** 956–970.

GELFAND, A. E., D. K. DEY, and H. CHANG, 1992 Model determination using predictive distributions with implementation via sampling-based methods. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 147–167. Oxford: Oxford University Press.

GELFAND, A. E. and A. F. M. SMITH, 1990 Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association **85:** 398–409.

GELMAN, A., 1996 Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 131–143. New York: Chapman and Hall.

GELMAN, A., J. B. CARLIN, H. S. STERN, and D. B. RUBIN, 1996 *Bayesian Data Analysis.* New York: Chapman and Hall.

GEMAN, S. and D. GEMAN, 1984 Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence **6:** 721–741.

GEYER, C. J., 1994 Estimating normalizing constants and reweighting mixtures in Markov Chain Monte Carlo. Technical Report 568r, School of Statistics, University of Minnesota.

GEYER, C. J., O. A. RYDER, L. G. CHEMNICK, and E. A. THOMPSON, 1993 Analysis of Relatedness in the California Condors from DNA fingerprints. Molecular Biology and Evolution **10:** 571–589.

GEYER, C. J. and E. A. THOMPSON, 1992 Constrained Monte Carlo maximum likelihood for dependent data (with dicussion). Journal of the Royal Statistical Society, Series B **54:** 657–699.

GEYER, C. J. and E. A. THOMPSON, 1995 Annealing Markov chain Monte Carlo with applications to ancestral inference. Journal of the American Statistical Association **90:** 909–920.

GIANOLA, D., 2001 Inferences about breeding values. In D. J. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*, pp. 645–672. New York: John Wiley & Sons.

GIUDICI, P. and P. J. GREEN, 1999 Decomposable graphical Gaussian model determination. Biometrika **86:** 785–801.

GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

GRIFFITHS, R. C. and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. Journal of Computational Biology **3:** 479–502.

GRIFFITHS, R. C. and S. TAVARÉ, 1994 Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society (London), Series B **344:** 403–410.

GROOT, C. and L. MARGOLIS (Eds.), 1991 *Pacific Salmon Life Histories.* Vancouver: UBC Press.

GROSS, R. and J. NILSSON, 1995 Application of heteroduplex analysis for detecting variation within the growth hormone 2 gene in *Salmo trutta* L. (brown trout). Heredity **74:** 286–295.

GUO, S.-W. and E. A. THOMPSON, 1992 A Monte Carlo method for combined segregation and linkage analysis. American Journal of Human Genetics **51:** 1111–1126.

GYLLENSTEIN, U. and A. C. WILSON, 1987 Mitochondrial DNA of salmonids: inter- and

intraspecific variability detected with restriction enzymes. In N. Ryman and F. Utter (Eds.), *Genetics and Fishery Management*, pp. 301–317. Seattle: University of Washington Press.

HAMMERSLEY, J. M. and D. C. HANDSCOMB, 1964 *Monte Carlo Methods*. London: Methuen & Co Ltd.

HANSEN, M. M., E. E. NIELSEN, D. E. RUZZANTE, C. BOUZA, and K. L. D. MENSBERG, 2000 Genetic monitoring of supportive breeding in brown trout (*Salmo trutta* L.), using microsatellite DNA markers. Canadian Journal of Fisheries and Aquatic Sciences **57:** 2130–2139.

HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

HEALEY, M. C., 1991 Life history of chinook salmon (*Oncorhynchus tshawytscha*). In C. Groot and L. Margolis (Eds.), *Pacific Salmon Life Histories*, pp. 311–394. Vancouver: UBC Press.

HEATH, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. American Journal of Human Genetics **61:** 748–760.

HEDRICK, P. W., D. HEDGECOCK, and S. HAMELBERG, 1995 Effective population size in winter-run chinook salmon. Conservation Biology **9:** 615–624.

HILL, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genetical Research (Cambridge) **38:** 209–216.

HOESCHELE, I., 2001 Mapping of quantitative trait loci in outbred pedigrees. In D. J. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*, pp. 599–644. New York: John Wiley & Sons.

HUELSENBECK, J. P. and J. P. BOLLBACK, 2001 Application of the likelihood function in phylogenetic analysis. In D. J. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*, pp. 415–443. New York: John Wiley & Sons.

HUSBAND, B. C. and S. C. H. BARRETT, 1995 Estimating effective population size: a reply to Nunney. Evolution **49:** 392–394.

JANSS, L. L. G., R. THOMPSON, and J. A. M. ARENDONK, 1995 Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theoretical and Applied Genetics **91:** 1137–1147.

JANSSON, H. and T. OEST, 1997 Hybridization between Atlantic salmon (*Salmo salar*) and brown trout (*S. trutta*) in a restored section of the River Dalaelven, Sweden. Canadian Journal of Fisheries and Aquatic Sciences **54:** 2033–2039.

JOHNSON, N. L. and Z. KOTZ, 1977 *Urn Models and Their Application.* New York: Wiley & Sons.

JOHNSON, N. L., Z. KOTZ, and N. BALAKRISHNAN, 1997 *Discrete Multivariate Distributions.* New York: Wiley & Sons.

JOHNSON, O. W., M. H. RUCKELSHAUS, W. S. GRANT, F. W. WAKNITZ, A. M. GARRETT, G. J. BRYANT, K. NEELY, and J. J. HARD, 1999 Status review of coastal cutthroat trout from Washington, Oregon, and California. NOAA Technical Memorandum NMFS-NWFSC-37, National Marine Fisheries Service.

JORDE, P. E. and N. RYMAN, 1995 Temporal allele frequency change and estimation of effective size in populations with overlapping generations. Genetics **139:** 1077–1090.

JORDE, P. E. and N. RYMAN, 1996 Demographic genetics of brown trout (*Salmo trutta*) and estimation of effective population size from temporal change of allele frequencies. Genetics **143:** 1369–1381.

KARLIN, S. and J. MCGREGOR, 1965 Direct product branching processes and related induced Markoff chains. I. Calculations of rates of approach to homozygosity. In L. LeCam and J. Neyman (Eds.), *Bernoulli (1793), Bayes (1773), Laplace (1813): Anniversary Volume*, pp. 111–145. New York: Springer.

KASS, R. E. and A. E. RAFTERY, 1995 Bayes factors and model uncertainty. Journal of the American Statistical Association **90:** 773–795.

KINCAID, H. L., 1995 An evaluation of inbreeding and effective population size in salmonid broodstocks in federal and state hatcheries. Uses and Effects of Cultured Fishes in Aquatic Ecosystems **15:** 193–204.

KINGMAN, J. F. C., 1982 On the genealogy of large populations. Journal of Applied Probability **19A:** 27–43.

KITADA, S., T. HAYASHI, and H. KISHINO, 2000 Empirical Bayes procedure for estimating genetic distance between populations and effective population size. Genetics **156:** 2063–2079.

KONG, A., 1991 Analysis of pedigree data using methods combining peeling and Gibbs sampling. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 379–385. Interface Foundation of North America.

KRIMBAS, C. B. and S. TSAKAS, 1971 The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control—selection or drift? Evolution **25:** 454–460.

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN, 1997 Applications of Metropolis-Hastings genealogy sampling. In P. Donnelly and S. Tavaré (Eds.), *Progress in Population Genetics and Human Evolution: IMA Volumes in Mathematics and its Applications, volume 87*, pp. 183–192. Berlin: Springer Verlag.

KUHNER, M. K., J. YAMATYO, and F. J, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

LANGE, K. and S. MATTHYSSE, 1989 Simulation of pedigree genotypes by random walks. American Journal of Human Genetics **45:** 959–970.

LAURITZEN, S. L., 1996 *Graphical Models*. Oxford: Clarendon Press.

LEARY, R. F. and F. W. ALLENDORF, 1997 Genetic confirmation of sympatric bull trout and Dolly Varden in western Washington. Transactions of the American Fisheries Society **126:** 715–720.

LIU, J. S., W. H. WONG, and A. KONG, 1994 Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika **81:** 27–40.

LONG, J. C., 1991 The genetic structure of admixed populations. Genetics **127:** 417–428.

MCKEIGUE, P., 1998 Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations by conditioning on parental admixture. American Journal of Human Genetics **63:** 241–251.

MENGERSON, K. L. and C. P. ROBERT, 1995 Testing for mixtures via entropy distance and Gibbs sampling. In J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Lindley, D V Smith (Eds.), *Bayesian Statistics 5*, pp. 147–167. Oxford: Oxford University Press.

METROPOLIS, N., 1987 The beginning of the Monte Carlo method. In *Stanislaw Ulam 1909–1984. Los Alamos Science, Special Issue Number 15*, pp. 125–130. Los Alamos, NM: Los Alamos National Laboratory.

MILLAR, R. B., 1987 Maximum likelihood estimation of mixed stock fishery composition. Canadian Journal of Fisheries and Aquatic Sciences **44:** 583–590.

MILLAR, R. B., 1991 Selecting loci for genetic stock identification using maximum likelihood, and the connection with curvature methods. Canadian Journal of Fisheries and Aquatic Sciences **48:** 2173–2179.

MILLER, L. and A. R. KAPUSCINSKI, 1997 Historical analysis of genetic variation reveals low effective population size in a northern pike (*Esox lucius*) population. Genetics **147:** 1249–1258.

MILNER, G. B., D. J. TEEL, F. M. UTTER, and C. L. BURLEY, 1981 Columbia River stock identification study: validation of method. Annual report of research, NOAA, Northwest and Alaska Fisheries Center, Seattle, Washington.

MORAN, P. A. P. and G. A. WATTERSON, 1958 The genetic effects of family structures in natural populations. Australian Journal of Biological Science **12:** 1–15.

MORIN, P. A., J. WALLIS, J. J. MOORE, R. CHAKRABORTY, and D. S. WOODRUFF, 1993 Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees. Primates **34:** 347–356.

NEHLSEN, W., J. E. WILLIAMS, and J. A. LICHATOVICH, 1991 Pacific salmon at the crossroads: stocks at risk from California, Oregon, Idaho, and Washington. Fisheries **16:** 4–21.

NEI, M. and F. TAJIMA, 1981 Genetic drift and estimation of effective population size. Genetics **98:** 625–640.

NEILLANDS, W. G., 1990 Natural hybridization between coastal cutthroat trout (*Oncorhynchus clarki*) and steelhead trout (*Oncorhynchus mykiss*) within Redwood Creek, Caliifornia. Master's thesis, Humboldt State University, Arcata, CA.

NEWTON, M. A., B. MAU, and B. LARGET, 1997 MCMC for Bayesian analysis of evolutionary trees from aligned molecular sequences. In F. Seillier-Moseiwitch, T. P. Speed, and M. Watterman (Eds.), *Statistics and Molecular Biology.* in press: Monograph Series of the Institute of Mathematical Statistics.

NIELSEN, E. E., M. M. HANSEN, and V. LOESCHCKE, 1999 Analysis of DNA from old scale samples: technical aspects, applications and perspectives for conservation. Hereditas **130:** 265–276.

NIELSEN, R., 1997 A likelihood approach to populations samples of microsatellite alleles. Genetics **146:** 711–716.

NUNNEY, L., 1995 Estimating the ratio of effective population size to adult numbers using genetic and ecological data. Evolution **49:** 389–392.

OLSEN, J. B., P. BENTZEN, M. A. BANKS, J. B. SHAKLEE, and S. YOUNG, 2000 Microsatellites reveal population identity of individual pink salmon to allow supportive breeding of a population at risk of extinction. Transactions of the American Fisheries Society **129:** 232–242.

OLSEN, J. B., J. K. WENBURG, and P. BENTZEN, 1996 Semiautomated multilocus genotyping of Pacific salmon (*Oncorhnychus* spp.) using microsatellites. Molecular Marine Biology and Biotechnology **5:** 259–272.

PAETKAU, D., W. CALVERT, I. STIRLING, and C. STROBECK, 1995 Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology **4:** 347–354.

PAINTER, I., 1997 Sibship reconstruction without parental information. Journal of Agricultural, Biological, and Environmental Statistics **2:** 212–229.

PELLA, J. and M. MASUDA, 2001 Bayesian methods for analysis of stock mixtures from genetic characters. Fisheries Bulletin Seattle **99:** 151–167.

PELLA, J. and G. B. MILNER, 1987 Use of genetic marks in stock composition analysis. In N. Ryman and F. Utter (Eds.), *Genetics and Fishery Management*, pp. 247–276. Seattle: University of Washington Press.

PENDAS, A. M., P. MORAN, J. L. MARTINEZ, and E. GARCIA-VAZQUEZ, 1995 Applications of 5S rDNA in Atlantic salmon, brown trout, and in Atlantic salmon x brown trout hybrid identification. Molecular Ecology **4:** 275–276.

PEREZ, J., E. GARCIA-VAZQUEZ, and P. MORAN, 1999 Physical distribution of SINE elements in the chromosomes of Atlantic salmon and rainbow trout. Heredity **83:** 575–579.

POLLAK, E., 1983 A new method for estimating the effective population size from allele frequency changes. Genetics **104:** 531–548.

PRITCHARD, J. K., M. STEPHENS, and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG, and P. DONNELLY, 2000 Association mapping in structured populations. American Journal of Human Genetics **67:** 170–181.

PRODOHL, P. A., J. B. TAGGART, and A. FERGUSON, 1995 A panel of minisatellite (VNTR) DNA locus specific probes for potential application to problems in salmonid aquaculture. Aquaculture **137:** 87–97.

RAFTERY, A. E., 1992 Discussion of: "Model determination using predictive distributions with implementation via sampling-based methods," by A. E. Gelfand, D. K. Dey, and H. Chang. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 147–167. Oxford: Oxford University Press.

RANNALA, B. and J. L. MOUNTAIN, 1997 Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences, USA **94:** 9197–9102.

RAO, B. R., S. M. MAZUMDAR, J. H. WALLER, and C. C. LI, 1973 Correlation between the numbers of two types of children in a family. Biometrics **29:** 271–279.

RIPLEY, B. D., 1987 *Stochastic Simulation.* New York: Wiley & Sons.

ROBERT, C. P., 1996 Mixture of distributions: inference and estimation. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 441–464. New York: Chapman and Hall.

RUBINSTEIN, B. Y., 1981 *Simulation and the Monte Carlo Method.* New York: Wiley & Sons.

RUE, H. and M. A. HURN, 1999 Bayesian object identification. Biometrika **86:** 649–660.

RYMAN, N. and L. LAIKRE, 1991 Effects of supportive breeding on the genetically effective population size. Conservation Biology **5:** 325–329.

RYMAN, N. and F. M. UTTER (Eds.), 1987 *Population Genetics and Fishery Management.* Seattle, WA: University of Washington Press.

SMOUSE, P. E., R. S. WAPLES, and J. A. TWOREK, 1990 A genetic mixture analysis for use with incomplete source population data. Canadian Journal of Fisheries and Aquatic Science **47:** 620–634.

SOULÉ, M. E. (Ed.), 1986 *Conservation Biology: The Science of Scarcity and Diversity.* Sunderland, MA: Sinauer and Associates.

STEPHENS, M., 2000 Dealing with label-switching in mixture models. Journal of the Royal Statistical Society, Series B **62:** 795–809.

STEPHENS, M., 2001 Inference under the coalescent. In D. J. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*, pp. 213–238. New York: John Wiley & Sons.

TABERLET, P., J.-J. CAMARRA, S. GRIFFIN, E. UHRES, O. HANNOTE, L. P. WAITS, and C. DUBOIS-PAGANON, 1997 Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Molecular Ecology **6:** 869–876.

TAJIMA, F., 1992 Statistical method for estimating the effective population size in Pacific salmon. Journal of Heredity **83:** 309–311.

TAUTZ, D., 1989 Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research **17:** 6463–6471.

TAYLOR, E. B., Z. REDENBACK, A. B. COSTELLO, S. M. POLLARD, and C. J. PACAS, 2001 Nested analysis of genetic diversity in northwestern North American char, Dolly varden (*Salvelinus malma*) and bull trout (*Salvelinus confluentus*). Canadian Journal of Fisheries and Aquatic Sciences **58:** 406–420.

TESSIER, N., L. BERNATCHEZ, and J. M. WRIGHT, 1997 Population structure and impact of supportive breeding inferred from mitochondrial and microsatellite DNA analyses in land- locked Atlantic salmon *Salmo salar* L. Molecular Ecology **6:** 735–750.

THOMPSON, E. A., 1973 The Icelandic admixture problem. Annals of Human Genetics **37:** 69–80.

THOMPSON, E. A., 1994 Monte Carlo likelihood in genetic mapping. Statistical Science **9:** 355–366.

THOMPSON, E. A. and S.-W. GUO, 1991 Monte Carlo evaluation of likelihood ratios. IMA Journal of Mathematics Applied in Medicine and Biology **8:** 149–169.

THOMPSON, E. A. and S. C. HEATH, 1999 Estimation of conditional multilocus gene identity among relatives. In F. Seillier-Moseiwitch (Ed.), *Statistics in Molecular Biology and Genetics: selected proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology.* Hayward, CA: Institute of Mathematical Statistics Lecture Notes—Monograph Series, Volume 33.

UTTER, F. M., 1999 Ecological genetics: introductory note. Ecology of Freshwater Fish **8:** 111–113.

UTTER, F. M., P. AEBERSOLD, and G. WINANS, 1987 Interpreting genetic variation detected by electrophoresis. In N. Ryman and F. M. Utter (Eds.), *Population Genetics and Fishery Management*, pp. 21–46. Seattle, WA: University of Washington Press.

VUCETICH, J. A., T. A. WAITE, and L. NUNNEY, 1997 Fluctuating population size and the ratio of effective to census population size. Evolution **51:** 2017–2021.

WAPLES, R. S., 1989 A generalized approach for esimating effective population size from temporal changes in allele frequency. Genetics **121:** 379–391.

WAPLES, R. S., 1990a Conservation genetics of Pacific salmon. II. Effective population size and the rate of loss of genetic variability. Journal of Heredity **81:** 267–276.

WAPLES, R. S., 1990b Conservation genetics of Pacific salmon: III. Estimating effective population size. Journal of Heredity **81:** 277–289.

WAPLES, R. S., 1995 Evolutionarily significant units and the conservation of biological diversity under the Endangered Species Act. In J. L. Nielsen (Ed.), *Evolution and the Aquatic Ecosystem: Defining Unique Units in Population Conservation*, pp. 8–27. Bethesda, MD: American Fisheries Society Symposium 17.

WAPLES, R. S., O. W. JOHNSON, P. B. AEBERSOLD, C. K. SHIFLETT, D. M. VANDOORNIK, D. J. TEEL, and A. E. COOK, 1993 A genetic monitoring and evaluation program for supplemented populations of salmon and steelhead in the Snake River

basin. Annual report of research, Coastal Zone and Estuarine Studies Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA.

WAPLES, R. S. and D. J. TEEL, 1990 Conservation genetics of Pacific salmon I. Temporal changes in allele frequency. Conservation Biology **4:** 144–156.

WIJSMAN, E. M., 1984 Techniques for estimating genetic admixture and applications to the problem of the origin of the Icelanders and the Ashkenazi Jews. Human Genetics **67:** 441–448.

WILLIAMSON, E. G. and M. SLATKIN, 1999 Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics **152:** 755–761.

WOOD, C. C., D. T. RUTHERFORD, and S. MCKINNELL, 1989 Identification of sockeye salmon (*Oncorhynchus nerka*) stocks in mixed-stock fisheries in British Columbia and southeast Alaska using biological markers. Canadian Journal of Fisheries and Aquatic Sciences **46:** 2108–2120.

WRIGHT, J. M. and P. BENTZEN, 1994 Microsatellites: genetic markers for the future. Reviews in Fish Biology and Fisheries **4:** 384–388.

WRIGHT, S., 1938 The distribution of gene frequencies under irreversible mutation. Proceedings of the National Academy of Sciences, USA **24:** 253–259.

WRIGHT, S., 1952 The theoretical variance within and among subdivisions of a population that is in a steady state. Genetics **37:** 313–321.

Appendix A

MONTE CARLO VARIANCE OF A PRODUCT

It seems intuitively reasonable that if one is estimating a product of independent probabilities by Monte Carlo (for example a likelihood that factorizes over loci), then an efficient Monte Carlo estimator will be obtained by estimating each probability separately and then multiplying the estimators together at the end. This will be more efficient than taking the Monte Carlo average of the product of the probabilities. Though this is well-known, I could not find a reference to it in the literature. Therefore, I prove it here, with the caveat that I would not be surprised if a more succinct proof were available.

THEOREM I: MONTE CARLO VARIANCE OF A PRODUCT.

Let $X_{j1}, \ldots, X_{jm}$, $j = 1, \ldots, J$ be $J \geq 2$ sequences of independent random variables, each of length m. For each j, $X_{j1}, \ldots, X_{jm}$ are identically and independently distributed with $\mathbb{E}X_{ji} = \mu_j$. Suppose that we are interested in estimating $\prod_{j=1}^{J} \mu_j$. Then the two estimators

$$\bar{\mu} = \prod_{j=1}^{J} \frac{1}{m} \sum_{i=1}^{m} X_{ji} \qquad \text{and} \qquad \tilde{\mu} = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{J} X_{ji}$$

are both unbiased for $\prod_{j=1}^{J} \mu_j$, but $\mathrm{Var}(\bar{\mu}) \leq \mathrm{Var}(\tilde{\mu})$, with strict inequality holding when the X_{ji} are non-degenerate random variables for at least two $j \in \{1, \ldots, J\}$.

PROOF: Unbiasedness is straightforward to show:

$$\mathbb{E}(\bar{\mu}) = \mathbb{E}\left(\prod_{j=1}^{J} \frac{1}{m} \sum_{i=1}^{m} X_{ji} \right) = \prod_{j=1}^{J} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}X_{ji} = \prod_{j=1}^{J} \mu_j \tag{A.1}$$

and

$$\mathbb{E}(\tilde{\mu}) = \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{J} X_{ji} \right) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{J} \mathbb{E}X_{ji} = \prod_{j=1}^{J} \mu_j. \tag{A.2}$$

To show that $\mathrm{Var}(\bar{\mu}) \leq \mathrm{Var}(\tilde{\mu})$, it will suffice to show that $\mathbb{E}\bar{\mu}^2 \leq \mathbb{E}\tilde{\mu}^2$, because the squares of the expected values of each estimator will be equal and $\mathrm{Var}(X) = \mathbb{E}X^2 - [\mathbb{E}X]^2$. We

start by simplifying the expression for $\mathbb{E}\bar{\mu}^2$.

$$\begin{aligned}
\mathbb{E}\bar{\mu}^2 &= \mathbb{E}\Big(\prod_{j=1}^{J}\frac{1}{m}\sum_{i=1}^{m}X_{ji}\Big)^2 \\
&= \mathbb{E}\Big(\prod_{j=1}^{J}\frac{1}{m^2}\Big(\sum_{i=1}^{m}X_{ji}\Big)^2\Big) \\
&= \prod_{j=1}^{J}\frac{1}{m^2}\mathbb{E}\Big(\sum_{i=1}^{m}X_{ji}^2+2\sum_{i=1}^{m}\sum_{k<i}X_{ji}X_{jk}\Big) \\
&= \prod_{j=1}^{J}\frac{1}{m^2}\Big(\sum_{i=1}^{m}\mathbb{E}X_{ji}^2+2\sum_{i=1}^{m}\sum_{k<i}\mu_j^2\Big) \qquad \text{(by independence)} \\
&= \prod_{j=1}^{J}\frac{1}{m^2}\Big(m\mathbb{E}X_{ji}^2+m(m-1)\mu_j^2\Big) \\
&= \prod_{j=1}^{J}\frac{1}{m}\Big(\mathbb{E}X_{ji}^2+(m-1)[\mathbb{E}X_{ji}^2-\mathrm{Var}(X_{ji})]\Big) \\
&= \prod_{j=1}^{J}\Big(\mathbb{E}X_{ji}^2-\frac{(m-1)}{m}\mathrm{Var}(X_{ji})\Big).
\end{aligned}$$

We know that $\mathrm{Var}(X_{ji}) \leq \mathbb{E}X_{ji}^2$, so we may write $\mathrm{Var}(X_{ji}) = \delta_j\mathbb{E}X_{ji}^2$, where $0 \leq \delta_j \leq 1$. Doing so, we may now rewrite $\mathbb{E}\bar{\mu}^2$.

$$\begin{aligned}
\mathbb{E}\bar{\mu}^2 &= \prod_{j=1}^{J}\Big(\mathbb{E}X_{ji}^2-\frac{(m-1)\delta_j}{m}\mathbb{E}X_{ji}^2\Big) \\
&= \prod_{j=1}^{J}\Big([(1-\delta_j)+\delta_j/m]\mathbb{E}X_{ji}^2\Big) \\
&= \prod_{j=1}^{J}\Big[(1-\delta_j)+\delta_j/m\Big]\prod_{j=1}^{J}\mathbb{E}X_{ji}^2. \qquad \text{(A.3)}
\end{aligned}$$

In similar fashion, we simplify the expression for $\mathbb{E}\tilde{\mu}^2$:

$$\mathbb{E}\tilde{\mu}^2 = \mathbb{E}\Big(\frac{1}{m}\sum_{i=1}^{m}\prod_{j=1}^{J}X_{ji}\Big)^2$$

$$
\begin{aligned}
&= \frac{1}{m^2}\mathbb{E}\Big[\sum_{i=1}^{m}\Big(\prod_{j=1}^{J}X_{ji}\Big)^2 + 2\sum_{i=1}^{m}\sum_{k<i}\Big(\prod_{j=1}^{J}X_{ji}X_{jk}\Big)\Big] \\
&= \frac{1}{m^2}\mathbb{E}\Big[\sum_{i=1}^{m}\prod_{j=1}^{J}X_{ji}^2 + 2\sum_{i=1}^{m}\sum_{k<i}\Big(\prod_{j=1}^{J}\mu_j^2\Big)\Big] \\
&= \frac{1}{m^2}\Big(\sum_{i=1}^{m}\prod_{j=1}^{J}\mathbb{E}X_{ji}^2 + m(m-1)\prod_{j=1}^{J}\mu_j^2\Big) \\
&= \frac{1}{m}\Big(\prod_{j=1}^{J}\mathbb{E}X_{ji}^2 + (m-1)\prod_{j=1}^{J}[\mathbb{E}X_{ji}^2 - \mathrm{Var}(X_{ji})]\Big) \\
&= \frac{1}{m}\Big(\prod_{j=1}^{J}\mathbb{E}X_{ji}^2 + (m-1)\prod_{j=1}^{J}[\mathbb{E}X_{ji}^2 - \delta_j\mathbb{E}X_{ji}^2]\Big) \\
&= \Big[\frac{1}{m} + \frac{(m-1)}{m}\prod_{j=1}^{J}(1-\delta_j)\Big]\prod_{j=1}^{J}\mathbb{E}X_{ji}^2. \qquad \text{(A.4)}
\end{aligned}
$$

And so, inspecting (A.3) and (A.4) it is clear that $\mathrm{Var}(\bar{\mu}) \leq \mathrm{Var}(\tilde{\mu})$ if and only if

$$\prod_{j=1}^{J}\Big[(1-\delta_j) + \delta_j/m\Big] \leq \frac{1}{m} + \frac{(m-1)}{m}\prod_{j=1}^{J}(1-\delta_j), \qquad \text{(A.5)}$$

an inequality which may be verified as follows: note that $\prod_{j=1}^{J}[(1-\delta_j)+\delta_j] = 1$, but the product may be written as a sum

$$\prod_{j=1}^{J}[(1-\delta_j)+\delta_j] = \prod_{j=1}^{J}(1-\delta_j) + \phi = 1$$

where ϕ is a sum of $2^J - 1$ terms, each of the form

$$\prod_{j\in\mathcal{A}}\delta_j \prod_{j\in\mathcal{A}^c}(1-\delta_j)$$

where $\mathcal{A}$ is a non-empty subset of $\{1,\ldots,J\}$, and $\mathcal{A}^c$ is its complement. Observe then,

$$
\begin{aligned}
\prod_{j=1}^{J}(1-\delta_j) + \phi &= 1 \\
\frac{1}{m}\prod_{j=1}^{J}(1-\delta_j) + \frac{\phi}{m} &= \frac{1}{m}
\end{aligned}
$$

$$\frac{\phi}{m} = \frac{1}{m} - \frac{1}{m}\prod_{j=1}^{J}(1-\delta_j)$$

$$\prod_{j=1}^{J}(1-\delta_j) + \frac{\phi}{m} = \frac{1}{m} + \frac{m-1}{m}\prod_{j=1}^{J}(1-\delta_j). \tag{A.6}$$

The right side of (A.6) is the same as the right side of (A.5), so to prove the inequality in (A.5), we need merely demonstrate that

$$\prod_{j=1}^{J}\Big[(1-\delta_j) + \delta_j/m\Big] \leq \prod_{j=1}^{J}(1-\delta_j) + \frac{\phi}{m}.$$

This may be done by noting that the left side expands into

$$\prod_{j=1}^{J}(1-\delta_j) + \varphi$$

where φ is a sum of $2^J - 1$ terms, each of which has the form

$$\prod_{j\in\mathcal{A}}\frac{\delta_j}{m^z}\prod_{j\in\mathcal{A}^c}(1-\delta_j)$$

where $\mathcal{A}$ and $\mathcal{A}^c$ are as above and z is the number of elements in the set A. (Since $\mathcal{A}$ is non-empty, $1 \leq z \leq J$.) For $m > 1$, each such term is clearly less than or equal to the corresponding term in the sum for ϕ/m so $\varphi \leq \phi/m$. Transparently, the equality holds only when all but one of the δ_j are zero. □

Appendix B

OVERLAPPING GENERATIONS VIA IMPORTANCE SAMPLING WITH THE MULTIVARIATE NORMAL DISTRIBUTION

I investigated the extension of the importance sampling methods of Chapter 2 to a population with a Pacific salmon-like life history. While it was possible to do so, the method allowed little flexibility in modeling and required more assumptions and approximations than the method developed in Chapter 4. Nonetheless, the importance sampling techniques developed are instructional, so I include a brief description of them here, as applied to diallelic loci.

In the following, the mathematical notation departs from that adopted in the remainder of the dissertation. In order to express vectors as bold Roman lowercase characters, and matrices as bold, uppercase Roman characters, it is necessary here to denote random vectors by bold, lowercase Roman letters. This should not create confusion.

B.1 Introduction to the Problem and Notation

A population is perpetuating itself forward in time. Individuals reproduce only once in their lives, either at age 1 or at age 2. It is assumed that the census number of individuals reproducing at time t of the different ages, $C_{t,1}$ and $C_{t,2}$, respectively, is known. It is assumed that $C_{t,a}$ individuals in a census represent $N_{t,a} = \lfloor \lambda_a C_{t,a} \rfloor$ effective individuals, $a = 1, 2$. We denote the effective proportions of individuals each year as either $\alpha_{t,1}$ or $\alpha_{t,2}$ where

$$\alpha_{t,1} = \frac{N_{t,1}}{N_{t,1} + N_{t,2}} \quad \text{and} \quad \alpha_{t,2} = \frac{N_{t,2}}{N_{t,1} + N_{t,2}}.$$

We assume that we have genetic data at a single locus with alleles B and b. We observe the variable Y_t—the number of copies of B found in a sample of size S_t taken from the gametes produced by the adults reproducing at time t. It is assumed that each age class

contributes to the gamete pool in proportion to their α.

The frequency of the B gene among the members of different age classes depends on the frequency of the B gene in the gamete pool from which their genes were drawn and the N for that age class via sampling with replacement as in the Wright-Fisher model. Hence, if we let $X_{t,a}$ be the number of copies of the B gene amongst a-year-olds reproducing at time t, we have:

$$X_{t,a} \sim \text{Binomial}\left(N_{t,a} \ , \ (\alpha_{t-a,1})\frac{X_{t-a,1}}{N_{t-a,1}} + (\alpha_{t-a,2})\frac{X_{t-a,2}}{N_{t-a,2}}\right).$$

Our genetic data samples, being assumed drawn with replacement from the gamete pool of a given year follow a similar probability distribution

$$Y_t \sim \text{Binomial}\left(S_t \ , \ (\alpha_{t,1})\frac{X_{t,1}}{N_{t,1}} + (\alpha_{t,2})\frac{X_{t,2}}{N_{t,2}}\right).$$

This is all presented in terms of haploid populations. The extension to diploid populations is straightforward, and won't be discussed.

B.2 The Likelihood

Our objective is to use the census sizes (assumed known without error, at this point) and the genetic sample data to compute a likelihood surface for different values of λ_1 and λ_2. We do this by Monte Carlo, using an importance sampling function which we construct by applying a multivariate normal approximation to this process and then discretizing it. The likelihood function may be derived using the dependence structure of the latent and observed variables.

The dependence structure is that shown in Figure B.1. The joint probability of all the $X_{t,a}$'s (denoted by $\mathbf{X}$) and all the Y_t's (denoted by $\mathbf{Y}$) for different values of λ_1 and λ_2 (denoted collectively by $\boldsymbol{\lambda}$) is

$$\begin{aligned} P_{\boldsymbol{\lambda}}(\mathbf{X},\mathbf{Y}) \ = \ & P_{\lambda}(X_{0,1},X_{0,2},X_{1,2})P_{\boldsymbol{\lambda}}(X_{1,1}|X_{0,1},X_{0,2})P_{\boldsymbol{\lambda}}(Y_0|X_{0,1},X_{0,2})P_{\boldsymbol{\lambda}}(Y_1|X_{1,1},X_{1,2}) \\ & \times \prod_{t=2}^{T}\left[P_{\boldsymbol{\lambda}}(Y_t|X_{t,1},X_{t,2})\prod_{a=1}^{2}P_{\boldsymbol{\lambda}}(X_{t,a}|X_{t-a,1},X_{t-a,2})\right] \end{aligned} \tag{B.1}$$

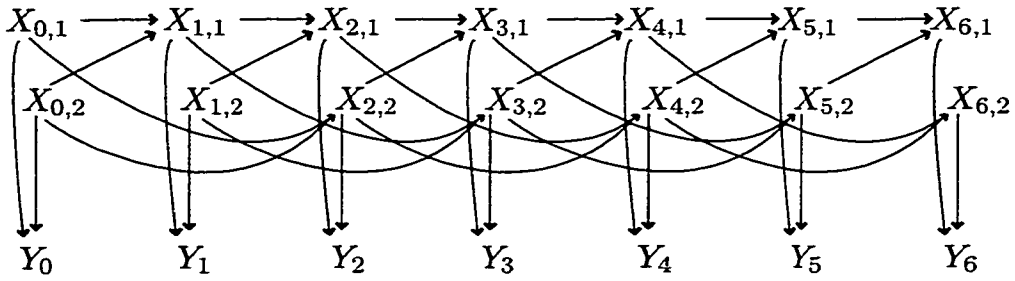


Figure B.1: A directed graph showing the dependence structure of the $X_{t,a}$'s and the Y_t's.

where $P_{\boldsymbol{\lambda}}(X_{0,1}, X_{0,2}, X_{1,2})$ is a prior for $X_{0,1}, X_{0,2}, X_{1,2}$—three variables that we will integrate out (so we are actually considering an integrated likelihood).

The likelihood for $\boldsymbol{\lambda}$ is just the sum over all $\mathbf{X}$ of that quantity and we will approximate it by Monte Carlo:

$$P_{\boldsymbol{\lambda}}(\mathbf{Y}) = \sum_{\mathbf{X}} P_{\boldsymbol{\lambda}}(\mathbf{X}, \mathbf{Y}) \approx \frac{1}{m} \sum_{i=1}^{m} \frac{P_{\boldsymbol{\lambda}}(\mathbf{X}^{(i)}, \mathbf{Y})}{Q_{\boldsymbol{\lambda}}(\mathbf{X}^{(i)})} \tag{B.2}$$

where $Q_{\boldsymbol{\lambda}}(\mathbf{X})$ is a discrete probability distribution for $\mathbf{X}$ and each $\mathbf{X}^{(i)}$ is independently simulated from the distribution $Q_{\boldsymbol{\lambda}}(\mathbf{X})$. This is going to work best when $Q_{\boldsymbol{\lambda}}(\mathbf{X})$ is as close as possible to $P_{\boldsymbol{\lambda}}(\mathbf{X}|\mathbf{Y})$. The next section describes a method for simulating from a $Q_{\boldsymbol{\lambda}}(\mathbf{X})$ which is similar to $P_{\boldsymbol{\lambda}}(\mathbf{X}|\mathbf{Y})$.

B.3 Constructing $Q_{\boldsymbol{\lambda}}(\mathbf{X})$ so it is close to $P_{\boldsymbol{\lambda}}(\mathbf{X}|\mathbf{Y})$

We proceed much as in Chapter 2, using the normal approximation to the binomial. We first apply the variance-stabilizing, arc-sine square-root transformation, and think in terms of continuous variables:

$$\begin{aligned} \gamma_{t,a} &= \sin^{-1}\left(\frac{X_{t,a}}{N_{t,a}}\right)^{1/2} \quad t = 1, \ldots, T \;\; ; \;\; a = 1, 2 \\ \phi_t &= \sin^{-1}\left(\frac{Y_t}{S_t}\right)^{1/2} \quad t = 1, \ldots, T. \end{aligned} \tag{B.3}$$

Recall that if $X \sim \text{Binomial}(n, p)$ then the corresponding transformed variable has an approximate normal distribution with mean $\sin^{-1}\sqrt{p}$ and variance $1/(4n)$. Thus we have

for $2 \leq t \leq T$ or $t = 1$ and $a = 1$, that $\gamma_{a,t}$ is approximately normal with mean

$$\mu = \sin^{-1}\left((\alpha_{t-a,1})\frac{X_{t-a,1}}{N_{t-a,1}} + (\alpha_{t-a,2})\frac{X_{t-a,2}}{N_{t-a,2}}\right)^{\frac{1}{2}} \tag{B.4}$$

and variance $1/(4N_{t,a})$.

We can approximate the mean given in (B.4) by a linear combination:

$$\begin{aligned} \mu &\approx (\alpha_{t-a,1})\sin^{-1}\left(\frac{X_{t-a,1}}{N_{t-a,1}}\right)^{1/2} + (\alpha_{t-a,2})\sin^{-1}\left(\frac{X_{t-a,2}}{N_{t-a,2}}\right)^{1/2} \\ &\approx (\alpha_{t-a,1})\gamma_{t-a,1} + (\alpha_{t-a,2})\gamma_{t-a,2} \end{aligned} \tag{B.5}$$

It is extremely useful to be able to express the mean of each $\gamma_{a,t}$ as this linear combination of the preceding $\gamma_{a,t}$'s. The same approximation is useful for determining the mean of the distributions for each ϕ_t. This approximation is quite good for values of γ that are in the middle of the range 0 to $\pi/2$, but becomes less accurate with γ's near the ends of that range. (This is one of the approximations that is not altogether satisfactory).

The fact that the mean can be written as a linear combination of the preceding γ's means that for every $\gamma_{a,t}$, $2 \leq t \leq T$ or $t = 1$ and $a = 1$, and for every ϕ_t, $0 \leq t \leq T$, we may write:

$$\gamma_{t,a} = (\alpha_{t-a,1})\gamma_{t-a,1} + (\alpha_{t-a,2})\gamma_{t-a,2} + \epsilon_{t,a} \tag{B.6}$$

$$\phi_t = (\alpha_{t,1})\gamma_{t,1} + (\alpha_{t,2})\gamma_{t,2} + \delta_t, \tag{B.7}$$

approximately, where $\epsilon_{t,a}$ and δ_t are independent, normally-distributed random variables with mean zero and variance $1/(4N_{t,a})$ or $1/(4S_t)$, respectively; thus, we discover that we can write every $\gamma_{t,a}$ or ϕ_t as a linear combination of the ϵ's and δ's and the three initial γ's: $\gamma_{0,1}$, $\gamma_{0.2}$, and $\gamma_{1,2}$.

Continuing, we find for those three initial γ's a prior distribution that is commensurate with the prior on the initial allele counts— $P_{\boldsymbol{\lambda}}(X_{0,1}, X_{0,2}, X_{1,2})$. We construct this as a multivariate normal distribution, $\mathrm{MVN}_3(\boldsymbol{\mu}_{\mathrm{init}}, \boldsymbol{\Sigma}_{\mathrm{init}})$, with three components. This is a column vector with three components, $\gamma_{0,1}, \gamma_{0,2}$, and $\gamma_{1,3}$. By appending to this vector, additional components, one for each independent $\epsilon_{t,1}$ $(0 < t \leq T)$ and $\epsilon_{t,2}$ $(1 < t \leq T)$

and δ_t $(0 \leq t \leq T)$, we may construct a new multivariate random vector with $3(T+1)$ components, which we shall call $\mathbf{w}$. We have

$$\mathbf{w} \sim \mathrm{MVN}_{3(T+1)}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where the first three components of $\boldsymbol{\mu}$ are the three elements of $\boldsymbol{\mu}_{\text{init}}$, with the rest being zero (because the ϵ's and δ's are error terms with mean zero). And with the upper 3×3 square of $\boldsymbol{\Sigma}$ holding the elements of $\boldsymbol{\Sigma}_{\text{init}}$ with the remaining entries in $\boldsymbol{\Sigma}$ being zero (because all the ϵ's and δ's are independent), except for the diagonal entries which include the appropriate variance terms $1/(4N_{t,a})$ for $\epsilon_{t,a}$ and $1/(4S_t)$ for δ_t. The notation here departs from the rest of the thesis in that $\mathbf{w}$, though lowercase, is a random variable—this is to make the notation more familiar in the matrix algebra that is coming up.

Recall that we can express every γ and every ϕ as a linear combination of variables which precede them in the directed graph of Figure B.1, and hence they may all be expressed as linear combinations of the elements of $\mathbf{w}$. Let us denote all the γ's by $\boldsymbol{\gamma} = (\gamma_{t,a})_{t=0,a=1}^{T,2}$ and all the ϕ's by $\boldsymbol{\phi} = (\phi_t)_{t=0}^{T}$. Then appending those together into a single vector, say $(\boldsymbol{\gamma}, \boldsymbol{\phi})$, we have a vector of all the variables we are interested in (the angularly transformed versions of the $X_{t,a}$'s and the Y_t's). Since these may all be expressed as a linear combination of $\mathbf{w}$ we may write $(\boldsymbol{\gamma}, \boldsymbol{\phi})$ as the product of a matrix $\mathbf{A}$ and the vector $\mathbf{w}$:

$$(\boldsymbol{\gamma}, \boldsymbol{\phi}) = \mathbf{A}\mathbf{w} \tag{B.8}$$

where $\mathbf{A}$ is a matrix with $3(T+1)$ rows and $3(T+1)$ columns. It is not difficult to compute the entries of $\mathbf{A}$. Each row of $\mathbf{A}$ is a linear combination of preceding rows, so the matrix $\mathbf{A}$ may be constructed recursively. For example, write $(\boldsymbol{\gamma}, \boldsymbol{\phi})$ as the column vector

$$(\gamma_{0,1}, \gamma_{0,2}, \gamma_{1,1}, \gamma_{1,2}, \gamma_{2,1}, \gamma_{2,2}, \ldots, \gamma_{T,1}, \gamma_{T,2}, \phi_0, \phi_1, \ldots, \phi_T)'$$

and $\mathbf{w}$ as the column vector

$$(\gamma_{0,1}, \gamma_{0,2}, \epsilon_{1,1}, \gamma_{1,2}, \epsilon_{2,1}, \epsilon_{2,2}, \ldots, \epsilon_{T,1}, \epsilon_{T,2}, \delta_0, \delta_1, \ldots, \delta_T)'.$$

Then, to satisfy (B.8), the first row of $\mathbf{A}$, which we denote as the row vector $\mathbf{A}_{[1]}$, must clearly be $(1, 0, 0, \ldots, 0)$. Likewise, $\mathbf{A}_{[2]} = (0, 1, 0, 0, \ldots, 0)$ and the fourth row, $\mathbf{A}_{[4]} =$

$(0,0,0,1,0,0,\ldots,0)$. By (B.8), the dot product of the third row, $\mathbf{A}_{[3]}$ with $\mathbf{w}$ must $\gamma_{1,1}$; hence by (B.6), $\mathbf{A}_{[3]} = (\alpha_{0,1}, \alpha_{0,2}, 1, 0, 0, \ldots, 0)$. The fifth row of $\mathbf{A}$ dotted with $\mathbf{w}$ must be $\gamma_{2,1}$. Again, by (B.6) we have

$$\gamma_{2,1} = \alpha_{1,1}\gamma_{1,1} + \alpha_{1,2}\gamma_{1,2} + \epsilon_{2,1}.$$

However, recalling that $\gamma_{1,1} = \mathbf{A}_{[2]} \cdot \mathbf{w}$ and $\gamma_{1,2} = \mathbf{A}_{[4]} \cdot \mathbf{w}$, we may write

$$\begin{aligned} \gamma_{2,1} &= \alpha_{1,1}(\mathbf{A}_{[2]} \cdot \mathbf{w}) + \alpha_{1,2}(\mathbf{A}_{[4]} \cdot \mathbf{w}) + \epsilon_{2,1} \\ &= (\alpha_{1,1}\mathbf{A}_{[2]} + \alpha_{1,2}\mathbf{A}_{[4]}) \cdot \mathbf{w} + \epsilon_{2,1} \\ &= (\alpha_{1,1}\mathbf{A}_{[2]} + \alpha_{1,2}\mathbf{A}_{[4]} + (0,0,0,0,1,0,0,\ldots,0)) \cdot \mathbf{w}. \end{aligned} \tag{B.9}$$

In other words, $\mathbf{A}_{[5]} = (\alpha_{1,1}\mathbf{A}_{[2]} + \alpha_{1,2}\mathbf{A}_{[4]} + (0,0,0,0,1,0,0,\ldots,0))$ is a linear combination of $\mathbf{A}_{[2]}$, $\mathbf{A}_{[4]}$, and a vector containing a single non-zero element which picks out the error term $\epsilon_{2,1}$ from amongst the components in $\mathbf{w}$. Proceeding forward, computing the entries of $\mathbf{A}_{[6]}$, $\mathbf{A}_{[7]}$, and so forth, is done similarly.

By the standard properties of multivariate normals, it follows that the vector $(\boldsymbol{\gamma}, \boldsymbol{\phi})$ has a $\mathrm{MVN}_{3(T+1)}(\boldsymbol{\mu}_{\text{joint}}, \boldsymbol{\Sigma}_{\text{joint}})$ distribution, where

$$\boldsymbol{\mu}_{\text{joint}} = \mathbf{A}\boldsymbol{\mu} \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\text{joint}} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$$

and $\mathbf{A}'$ denotes the matrix transpose of $\mathbf{A}$. This gives us a normal approximation to the joint distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, the angularly transformed allele frequencies in the population age groups and the genetic samples, respectively. However, to construct $Q_{\boldsymbol{\lambda}}(\mathbf{X})$ which is close to $P_{\boldsymbol{\lambda}}(\mathbf{X}|\mathbf{Y})$, we desire an approximation to the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$. Such an approximation may be derived using the *conditional* distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{\phi}$, which we will now deduce from the joint distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ by standard methods for multivariate normal variables.

Notice that $\boldsymbol{\Sigma}_{\text{joint}}$ is composed of a part for $\boldsymbol{\gamma}$, a part for $\boldsymbol{\phi}$ and a part for the covariances between $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$. Also $\boldsymbol{\mu}_{\text{joint}}$ consists of a part for $\boldsymbol{\gamma}$ and a part for $\boldsymbol{\phi}$. They may be partitioned as

$$\boldsymbol{\mu}_{\text{joint}} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{\gamma}} \\ \boldsymbol{\mu}_{\boldsymbol{\phi}} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\text{joint}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} & \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\phi}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\phi}\boldsymbol{\gamma}} & \boldsymbol{\Sigma}_{\boldsymbol{\phi}\boldsymbol{\phi}} \end{pmatrix}.$$

Therefore, in order to find the conditional distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{\phi}$ we may use a standard result about the distribution of a subset of components of a multivariate normal random vector, conditional upon the values of the remaining components:

$$\boldsymbol{\gamma}|\boldsymbol{\phi} \sim \mathrm{MVN}_{2(T+1)}(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$$

where

$$\boldsymbol{\mu}_Q = \boldsymbol{\mu}_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\phi}}\boldsymbol{\Sigma}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}_{\boldsymbol{\phi}})$$

and

$$\boldsymbol{\Sigma}_Q = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\phi}}\boldsymbol{\Sigma}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\phi}}'.$$

With $\boldsymbol{\mu}_Q$ and $\boldsymbol{\Sigma}_Q$ as calculated above, we are in a position to simulate values of $\boldsymbol{\gamma}$ and then transform them back to $\mathbf{X}$'s, essentially discretizing the distribution of $\boldsymbol{\gamma}$. Some care must be taken to ensure that alleles are not lost from the population when they appear in later samples (for example, we must never simulate $X_{t,1}$ and $X_{t,2}$ to be zero if the B allele appears in the sample from the gametes produced at time t), and further, some care should be taken to ensure that it is not very difficult to compute the probability $Q(\mathbf{X}^{(i)})$ of each resulting $\mathbf{X}^{(i)}$ simulated from this scheme. First, we describe a general method for simulating $\boldsymbol{\gamma}$ from a $\mathrm{MVN}_{2(T+1)}(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$ distribution.

Since $\boldsymbol{\Sigma}$ is symmetric matrix and positive definite, it is possible to compute its *Cholesky factor*, $\mathbf{L}$. $\mathbf{L}$ is a lower triangular matrix having the property that

$$\mathbf{L}\mathbf{L}' = \boldsymbol{\Sigma}_Q.$$

It is well known that if $\mathbf{z} = (z_1, \ldots, z_{2(T+1)})'$ is a column vector of independent, univariate normal random variables with variance one and mean zero, then the quantity

$$\boldsymbol{\mu}_Q + \mathbf{L}\mathbf{z}$$

will have the $\mathrm{MVN}_{2(T+1)}(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$ distribution. This is one way to simulate realizations of $\boldsymbol{\gamma}$. We note here, however, that even after the Cholesky factor has been computed, the computation to simulate $\boldsymbol{\gamma}$ this way increases quadratically in $2(T+1)$. I believe it should actually be possible to exploit the second-order Markov structure here do something that is linear in $2(T+1)$. But I have not pursued this.

Having the Cholesky factor, we can simulate γ's forward in time from their distribution condition on ϕ. For simplicity, I am going to assume that no alleles are going extinct, and all frequencies are intermediate, so we don't have to worry about writing code for careful bookkeeping. From above, we have that we can simulate the first element of γ, that is $\gamma_{0,1}$, by simulating a standard unit normal, $z_1^{(i)}$ and transforming it—$\gamma_{0,1}^{(i)} = \mu_{Q_1} + z_1^{(i)} L_{1,1}$. We transform that back to an $X_{0,1}$ by the discretization

$$X_{0,1}^{(i)} = \lfloor N_{0,1} \sin(\gamma_{0,1}^{(i)})^2 + 1/2 \rfloor.$$

The probability of simulating this $X_{0,1}$ is the area under the curve of a $N(\mu_{Q,1}, L_{1,1}^2)$ random variable between

$$\sin^{-1}\left(\frac{X_{0,1}^{(i)}}{N_{0,1}} - \frac{1}{2}\right)^{\frac{1}{2}} \qquad \text{and} \qquad \sin^{-1}\left(\frac{X_{0,1}^{(i)}}{N_{0,1}} + \frac{1}{2}\right)^{\frac{1}{2}}.$$

Proceeding we could then simulate $\gamma_{0,2}^{(i)} = \mu_{Q,2} + z_1^{(i)} L_{2,1} + z_2^{(i)} L_{2,2}$, and similarly transform that into an $X_{0,2}^{(i)}$. Now, however, the probability of simulating that $X_{0,2}^{(i)}$, given $X_{0,1}^{(i)}$ still depends on the original value of $z_1^{(i)}$ (and hence on the actual continuous value $\gamma_{0,1}^{(i)}$). This problem also occurs in the non-overlapping generations case. By this approach, in order to be correct in computing $Q_{\boldsymbol{\lambda}}(\mathbf{X}^{(i)})$ (and, hence, most efficient in the Monte Carlo estimation of $P_{\boldsymbol{\lambda}}(\mathbf{Y})$) one should do the multivariate integral of the joint density of all the $\gamma_{t,a}$'s given $\mathbf{X}^{(i)}$. It can be shown that even if the multivariate integral is not performed, then the Monte Carlo estimator for $P_{\boldsymbol{\lambda}}(\mathbf{Y})$ still has the correct expectation. Nonetheless it is more efficient and cleaner to simulate γ's and transform them to X's as follows:

1. Realize $\gamma_{0,1}^{(i)} = \mu_{Q,1} + z_1^{(i)} L_{1,1}$.

2. Convert that to an $X_{0,1}^{(i)} \leftarrow \lfloor N_{0,1} \sin(\gamma_{0,1}^{(i)})^2 + 1/2 \rfloor$

3. Define $z_1^* \leftarrow \left(\sin^{-1}\left(\frac{X_{0,1}^{(i)}}{N_{0,1}}\right)^{\frac{1}{2}} - \mu_{Q1}\right) / L_{1,1}$

4. Simulate $\gamma_{0,2}^{(i)} = \mu_{Q,2} + z_1^* L_{2,1} + z_2^{(i)} L_{2,2}$,

and so forth. The key difference here being the use of z_1^* rather than the $z_1^{(i)}$ in determining the distributions for the future $\gamma_{t,a}$'s.

By this procedure, it is possible to simulate independent realizations $\mathbf{X}^{(i)}$, $i = 1, \ldots, m$, and to compute the probability $Q_{\boldsymbol{\lambda}}(\mathbf{X}^{(i)})$ of each of these as the simple product of areas under normal densities (rather than by doing a multidimensional integration). The details are omitted, but follow quite simply from the formulation given above. These realizations may then be used in (B.2) to approximate $P_{\boldsymbol{\lambda}}(\mathbf{Y})$.

B.4 Simulated Data

To assess the potential of this method, I applied it to simulated data. I simulated the data as follows: I entered a census size, $C = 200$ haploids, which was assumed to be the total census size each year. Then I entered fractions of the census population that were age 1 or 2. These were r_1 and $r_2 = 1 - r_1$, which I set to be each .5. I then simulated each $C_{t,1}$ from a Binomial(C, r_1) distribution and set $C_{t,2} = C - C_{t,1}$. I then set "true" values of $\lambda_1 = 1.0$ and $\lambda_2 = 1.2$ at which to simulate the data and converted the census sizes into effective size by $N_{t,a} = \lfloor \lambda_a C_{t,a} \rfloor$. Then, for J diallelic loci, I drew $X_{0,1}$, $X_{0,2}$, and $X_{1,2}$ from respective Binomial$(N_{t,a}, .5)$ distributions, then I simulated gamete frequencies and gene copy numbers forward in time, and for each time step drew a sample of 500 haploid gametes.

I used an initial mean vector $\boldsymbol{\mu}_{\text{init}}$ and initial variance vector $\boldsymbol{\Sigma}_{\text{init}}$ that approximated the distribution used to simulate $X_{0,1}$, $X_{0,2}$, and $X_{1,2}$. In actual practice one would want to use some sort of diffuse prior that accounted for the correlation between those three variables, but I did not derive such a diffuse prior for those allele frequencies. Instead, $\boldsymbol{\mu}_{\text{init}}$ was set to a three-vector of $\sin^{-1}\sqrt{.5}$'s, and $\boldsymbol{\Sigma}_{\text{init}}$ included values that were commensurate with the prior probability described above. This is not a "non-informative" prior, at all, but I used it for this trial.

For a mesh of values of λ_1 and λ_2 between the values of 0.25 and 2.25 in steps of .25, I approximated $P_{\boldsymbol{\lambda}}(\mathbf{Y})$ by (B.2) using $m = 10,000$ for each different pair (λ_1, λ_2). The resulting likelihood surface is shown in the contour plot of Figure B.2, which has a peak

that is not too far from $\lambda_1 = 1.0$ and $\lambda_2 = 1.2$.

This appendix has described the implementation of an importance sampling method for $\boldsymbol{\lambda}$ in the context of overlapping generations under the simplest possible data scenario—diallelic loci at intermediate frequencies. Despite its application to a simple scenario, the formulation of the problem in this manner is quite challenging, and it would be even more challenging to extend it to more complex situations. Therefore I did not pursue this method any further. The techniques presented in Chapter 4 are clearly far more versatile than the technique investigated here.

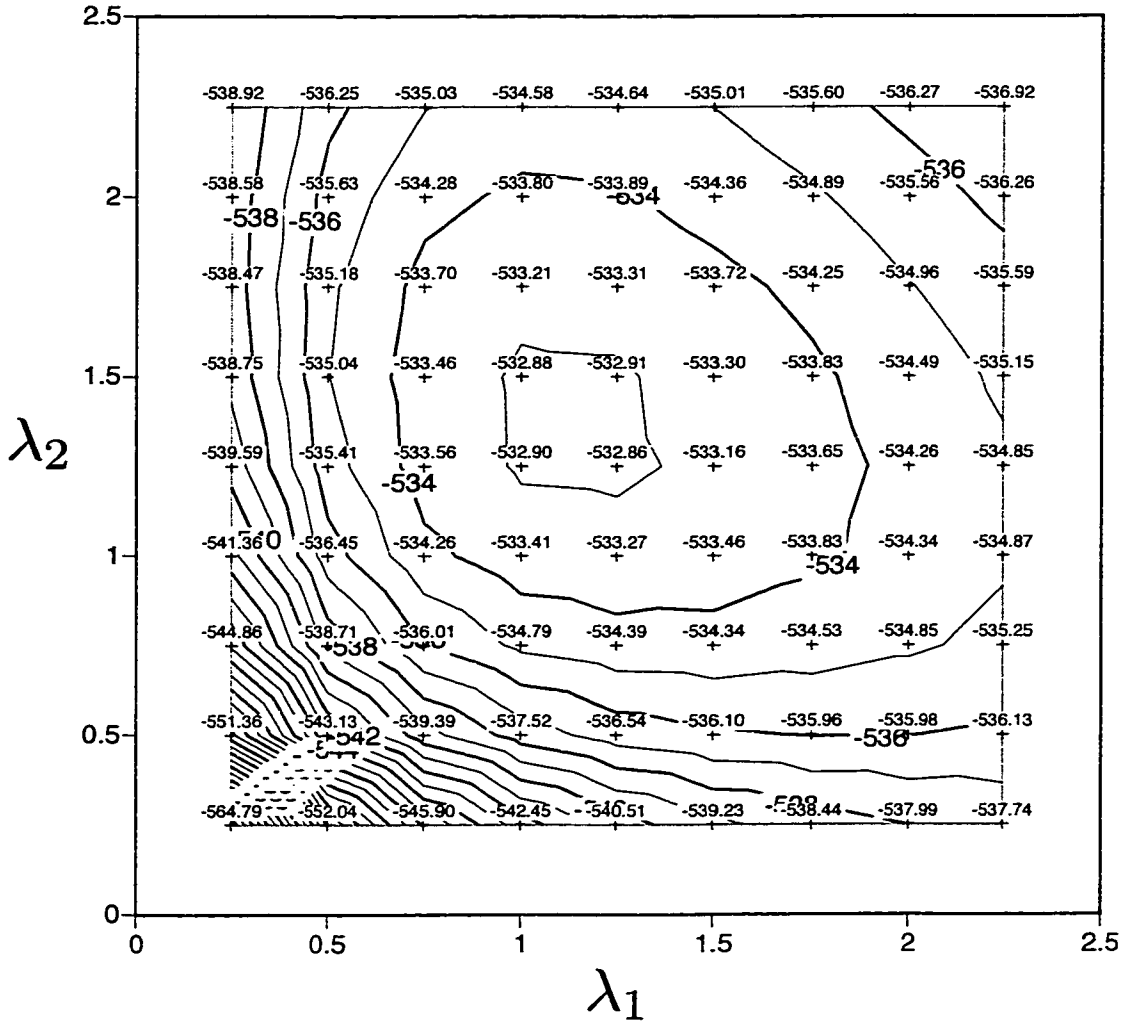


Figure B.2: Contour plot of the likelihood surface for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ from simulated data. Importance sampling was used to estimate the likelihood at an array of points (λ_1, λ_2) from (0.25,0.25) to (2.25,2.25) in increments of 0.25. These points are shown, with the estimated likelihood value appearing above them. The lines in the figure are interpolated contour lines of the likelihood surface. The true values, under which the data were simulated, are $\lambda_1 = 1.0$ and $\lambda_2 = 1.2$.

VITA

Eric Anderson was born during the winter of 1970 in Anchorage, Alaska to Richard Anderson and Dorothy Ingebretsen. He spent most of his childhood years in Ojai, California. After earning a high school diploma in 1988 from The Thacher School in Ojai, he studied at Stanford University in California for his first year at college. During his sophomore year he studied at Prescott College in Arizona. In 1991, he returned to Stanford and earned a B.A. in Human Biology in 1993. In 1994, he entered the School of Fisheries at the University of Washington to study River Ecology. After his first year of graduate school, his interests became focused on statistical issues in the analysis of genetic data from wild populations. He earned his M.S. degree from U.W. in Fisheries in 1998. In addition to his academic interests, Eric enjoys many active, outdoor pursuits, including skiing, hiking, rock-climbing, kayaking, trail-running, cycling, and tight-rope walking. His permanent email address is `eric.anderson@stanfordalumni.org`.