Monte Carlo Methods for Inference of Population-Genetic Parameters

A Summary of Proposed Ph.D. Dissertation Research

Eric C. Anderson

Interdisciplinary Program in Quantitative Ecology and Resource Management

University of Washington
Seattle, WA 98195

April 7,1999

Preface

My dissertation research explores the use of Markov chain Monte Carlo techniques in inference from allelic and genotype count data from population genetic models. I am mainly concerned with applying these techniques to natural populations, especially salmon populations. I have made forays into two main areas: a likelihood-based temporal method for inference on the ratio of effective size to census size, and, more recently, a Bayesian approach to estimating the contribution of known and unknown stocks to a mixed fishery. As I have been moving forward with the effective size project for some time, and because the general exam coincides with several other deadlines for manuscripts and smaller proposals, Elizabeth Thompson (committee chair) and I determined that I would benefit most by preparing a document of the sort that you will find in the following pages. This packet starts with a short CV and a list of course work that I have completed, but then, rather than a lengthy proposal for future work, I have included a collection of recently completed and ongoing work in the form of contributed papers and manuscripts, an earlier proposal that inspired some of this work, and then one chapter which is a proposal for work that is still in its infancy. This is certainly not to suggest that I am close to completing my work in this area. In fact, I have been pleased to find challenging terrain around almost every corner in these projects. Numerous extensions, specific to my applied goals, remain to be made, and occasionally problems are encountered for which these projects may prove a convenient context for delving deeper into issues of wider statistical interest.

Chapter 2 is a short contributed paper to an upcoming session of the International Statistics Institute which describes our efforts at implementing MCMC for the effective size problem. Chapter 3 is a portion of a manuscript that I am preparing as part of a collaborative project with Ellen Williamson at Berkeley. This may be the most novel piece of work I have done to date; it draws upon methods for inference from hidden Markov chains and some classical population genetics to create a block-updating sampler which is a sort of analogue to the M-sampler in pedigree analysis (Thompson and Heath 1998). This will be the topic which I discuss in the most detail during my general exam. Chapter 4 is a proposal for work that I am just starting on a Bayesian approach to mixed-stock fishery problems, using computational methodologies developed within the last five years. Chapter 5 contains the complete Proposed Research section for an NSF proposal that I assisted Elizabeth in writing. It provides an extensive literature review and a comprehensive description of the effective size project. The final chapter (6) describes some preliminary work I did exploring exact computation of likelihoods for the effective size project and comparing the performance of different estimators. With some modifications I will probably include it as an early chapter in the dissertation.

Finally, Chapter 1 is a brief summary of my research and my personal interests in the projects. I include there references to sections of this document where you may find more detailed information on specific topics.

Contents

Eric Christopher Anderson; Brief CV

Interdisciplinary Program in Quantitative
Ecology and Resource Management
Box 351720
University of Washington
Seattle, WA 98103

Born January 25, 1970
Email: `eriq@cqs.washington.edu`
Phone: (206) 685-8969

EDUCATION

B.A. (Human Biology)	1993	Stanford University
M.S.(Fisheries)	1998	University of Washington
Ph.D. (QERM)	(in progress)	University of Washington

EMPLOYMENT

1999–	Research Assistant, NSF Grant #BIR-9807747, "Computational Methods for Inference of Population Parameters," (PI: E.A. Thompson).
1994–98	Research Assistant, Center for Streamside Studies, School of Fisheries, and Quantitative Ecology and Resource Management.
1997	Graduate Assistant, Curriculum development, Department of Statistics, University of Washington.

ACADEMIC HONORS AND AWARDS

1996–	Recipient, National Science Foundation Mathematical Biology Training Grant, University of Washington.
1994–96	Recipient of Fellowship from H. Mason Keeler Endowment for Excellence. School of Fisheries, University of Washington.
1993	Phi Beta Kappa, Stanford University.

PUBLICATIONS

Naiman, R.J., Anderson, E.C. (1997) Streams and rivers: their physical and biological variability. In *The Rain Forests of Home: Profile of a North American Bioregion*, ed. P.K. Schoonmaker, B. von Hagen, E.C. Wolf, pp. 131–148. Washington, D.C.: Island Press.

Anderson, E.C., Thompson, E.A. (submitted) MCMC likelihoods for population genetics. (contributed paper at the 52nd session of the International Statistics Institute).

M.S. THESIS

Anderson, E.C. (1998) Inferring the ancestral origin of Sockeye Salmon, *Oncorhynchus nerka*, in the Lake Washington basin: A statistical method in theory and application. (Advisor: Dr. Thomas Sibley)

REVIEWER FOR

Journal of Fish Biology (UK)
Genetics

Coursework Applied to Ph.D. Program

Graded 500-level and approved 400-level courses

Course #	Course Name	Quarter	Units	Grade
STAT 513	Statistical Inference	WIN 97	4.0	3.9
QERM 550	Ecological Modelling	AUT 97	4.0	3.8
STAT 516	Stochastic Modelling of Scientific Data	AUT 97	4.0	3.8
STAT 517	Stochastic Modelling of Scientific Data	WIN 98	4.0	3.8
ZOOL 470	Techniques for Mathematical Biology	WIN 98	4.0	I[†]
ZOOL 471	Models in Biology	SPR 98	4.0	4.0
QERM 514	Analysis of Ecological Data	SPR 98	3.0	3.8

[†] I completed this course long ago. Garry has been quite remiss about reporting my grade.

Ungraded Courses, Ongoing Seminars, and Courses Below the 400 Level

Course #	Course Name	Quarters Attended	Units/Qtr
ZOOL 525	Seminar in Mathematical Biology	All Qtrs since AUT 96	2.0
BIOST 580	Statistical Genetics Seminar	A96, W97, Sp97, W98, A98, W99	1.0
GENET 590	Population Genetic Seminar	A96, A97, A98	1.0
QERM 597	Seminar in Quantitative Ecology	A97, Sp98	2.0
STAT 592	Meioses, Pedigrees, Populations	WIN 99	3.0
CSE 142	Computer Programming I	SUM 97	4.0

Currently Enrolled, Ungraded 500-level courses

Course #	Course Name	Quarter	Units
STAT 518	Stochastic Modelling of Scientific Data	SPR 99	3.0
STAT 593C	Grahical Markov Models	SPR 99	3.0

To the best of my knowledge I have fulfilled all of the QERM course requirements.

Examinations Passed at the Ph.D. Level

QERM Statistical Theory Exam June 1997

QERM Applied Statistics and Modelling Exam June 1998

Examinations Pending

Ph.D. General Exam to be April 21, 1999

Ph.D. Final Exam , projected date: Spring Quarter 2001

Chapter 1

Summary of Research

Markov chain Monte Carlo (MCMC) methods have been profitably applied to inference problems in genetics since the early 1990's. First they were applied to likelihood inference from pedigree data. More recently they have permitted likelihood inference from genetics models based on Kingman's coalescent (Kuhner *et al.* 1995, 1997). Between these two ends of the genetics modelling spectrum are a number of classical population-genetic models to which MCMC methods, though they may prove useful, have not yet been applied. This dissertation proposes to develop, investigate, and implement MCMC methods for both likelihood-based and Bayesian inference of population genetic parameters associated with population models related to the Wright-Fisher model.

The data for these sorts of models are sample counts of different alleles and/or genotypes. The likelihood inference problem considered is that of estimating, from temporally spaced allele frequency samples, the effective size of a population of constant size, or of estimating the ratio λ of the effective number of breeding adults to the census number of breeding adults, when the census number of breeding adults for the generations between genetic samples is known. MCMC techniques are useful in this context for approximating the enormous sums over latent variables which appear in the likelihood function.

The Bayesian inference problem considered is one of determining the posterior distribution of mixture contributions of salmon stocks to a mixed-stock fishery. This approach could yield posteriors marginalized over the unknown number of contributing populations not represented in the baseline data, and over the various lumping/splitting possibilities of baseline populations. MCMC techniques allow sampling from unnormalized posterior distributions, and, in this case, reversible jump MCMC (Green 1995) allows sampling from parameter spaces of varying dimension (*i.e.*, different numbers of populations contributing to the mixture).

1.1 Likelihood estimation of λ

A comprehensive overview of this project can be found in the proposal of Chapter 5. The rationale for pursuing a likelihood approach is detailed in Sections 5.1, 5.2, 5.4, and 6.1. My interest in the problem grew from reading (Miller and Kapuscinski 1997), a paper describing the use of archived fish scales for obtaining allele frequency estimates from a small population of northern pike over long time periods. The authors used these data to estimate the genetically effective size of that fish population. Most striking to me in this study, (and apparently an almost ubiquitous feature in other studies estimating effective size) is that as soon as the authors had estimated the population's effective size, they computed the ratio of effective size to observed size of the population. They did not however, have a way of estimating that ratio directly. Such a direct estimation scheme would be useful, I believed, for incorporating uncertainty in the estimate of population sizes into the precision estimate for the ratio, and also for dealing with populations having overlapping generations and other complicating life-history features by estimating a quantity we call λ (see Page 46 of Section 6.1). My interest in the problem grew when I learned of a similar project at the University of Minnesota investigating two steelhead (*Oncorhynchus mykiss*) populations in the Northwest.

The method of maximum likelihood is an obvious choice for estimating this sort of ratio, but one that had not been previously used. With large amounts of data maximum likelihood estimators typically outperform other types of estimators. Just as important, it is conceptually straightforward to apply the method of maximum likelihood to populations with sophisticated life-history patterns—so long as you can describe in probabilistic terms the stochastic process which generates the data, you can write down a likelihood for the parameters. (An example of a likelihood function for a simple population structure may be found in Equation 5.7). The difficulty, however, comes with trying to compute the likelihood function. In the present case, computing the likelihood requires performing an enormous summation over latent variables. In most cases exact summation is completely, computationally infeasible. This summation, however, may be cast as an expectation, and hence approximated by Monte Carlo. Implementing Monte Carlo methods which actually work has been the major focus of this project.

We have been pursuing techniques of Monte Carlo likelihood (Thompson and Guo 1991; Geyer and Thompson 1992) which rely on realization of latent variables from Markov chains constructed by Metropolis-Hastings methods (Metropolis *et al.* 1953; Hastings 1970). Chapter 2 describes a component-wise Hastings sampler approach. More recently, following a similar evolution of MCMC samplers in other fields of statistical genetics (Thompson and Heath 1998) I have developed a block-updating sampler which uses as its proposal distribution the distribution $P^*_N(\mathbf{X})$ described in Chapter 3.

Many extensions have yet to be made, especially as regards incorporating more complex life histories and accounting for uncertainty in population census size estimates.

1.2 Bayesian mixed fishery estimation

This project is compactly described in Chapter 4 and I refer the reader there immediately. I only state here that the mixed fishery problem was one of the first inference problems I encountered (while having little statistics background at the time) which made me extremely excited about parameter estimation and mathematical statistics. It is a problem that is still very near to my heart, and it is (seriously) with immeasurable joy that I work toward making a contribution in the field.

1.3 Sideshows and future directions

While working on these projects, particularly the Monte Carlo likelihood project, I have encountered many of the same difficulties that Elizabeth and others have encountered in Monte Carlo computations on pedigrees, albeit in a slightly different guise. It has been extremely rewarding for me to see the similarities between these two different fields of statistical genetics. Nonetheless, I must concede that my Monte Carlo problem is in many ways much simpler. Most notably, the space of latent variables I deal with (unobserved allele frequencies) is considerably less complex than that of the space of genotype assignments to pedigrees or the space of possible coalescent trees uniting samples of DNA sequences.

In some ways, however, this simplicity is advantageous in making it possible to explore and apply new computational technologies that do not lend themselves well to more complex latent variable spaces. As one example, I have successfully applied Coupling From The Past (Propp and Wilson 1996) to obtain exact samples from the distribution of latent variables conditional on the data. Though, in its current form, this exact sampling is not particularly useful to my project, methods for exact sampling from Markov chains may someday prove themselves useful, and fluency in these emerging techniques may become valuable.

Additionally, since we are interested in estimating a complete Monte Carlo likelihood curve for λ, we are forced to realize values from a number of different Markov chains, each indexed by a different parameter. Reweighting those realizations appropriately is difficult. Geyer (1994) presents a method which is limited by its demands on computer storage. The λ-inference problem may provide a useful setting to explore alternative reweighting schemes and compare their merits amongst themselves and to a Bayesian approach.

Finally, I remark briefly on a tantalizing future direction. It seems that some sort of combination of the Monte Carlo likelihood framework for estimating λ and the Bayesian mixed fishery problem could lead to an advantageous technique for estimating proportions of admixed populations (populations in which mixture occurred in the past and one or more generations of reproduction have taken place since then) as in Long (1991) and Thompson (1973). The key here would be to recognize as the sufficient statistic the multilocus phenotype of individuals in the samples, rather than merely the allele frequencies. Especially with linked markers, this technique would derive information from the linkage disequilibrium resulting from the original mixing of contributing populations. Though this may not be part of my dissertation it seems to merit further consideration.

Chapter 2

ISI paper as submitted

On the following two pages is a brief contributed paper as it was submitted electronically on April 9, 1999, for the 52nd Session of the International Statistics Institute in Finland. Elizabeth Thompson will be presenting the paper at the meeting in August.

MCMC Likelihoods for Population Genetics

Eric C. Anderson and Elizabeth A. Thompson
Department of Statistics, University of Washington
Box 354322
Seattle, WA 98195 U.S.A.
eriq@stat.washington.edu, thompson@stat.washington.edu

1. The Problem and the Likelihood

It is important to consider the genetic health of small or threatened populations. A key quantity in this regard is the effective number of breeding adults, N_a, each generation. Population geneticists define N_a by comparison to an ideal population in which the genes of the next generation are sampled with replacement from the current one. N_a is usually smaller than the census number of breeding adults, C. Often C can be measured, and then it is valuable to know $\lambda = N_a/C$. It is difficult to measure λ by demographic methods, particularly for some species of fish and amphibians, which produce thousands of offspring, very few of which survive to adulthood. Instead, λ may be estimated from temporal changes of allele frequencies sampled from the population. We present a Monte Carlo approach for approximating the likelihood curve for λ.

Assume a discrete-generation, semelparous population with C_t haploid individuals reproducing at t, giving rise to C_{t+1} individuals at $t+1$. We take genetic samples of size $S_1, \ldots, S_T$ (assume $S_1 > 0,\ S_T > 0$) individuals, and find counts of the k different allelic types at a locus, $\mathbf{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)$ where $\boldsymbol{Y}_t = (Y_{t1}, \ldots, Y_{tk})$. The population at t is modelled as $\lfloor \lambda C_t \rfloor$ ideally-reproducing adults, where $\lfloor x \rfloor$ is the largest integer $\leq x$. Underlying the data are latent allele counts $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$ with $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{tk})$. $\{\boldsymbol{X}_t,\ t \geq 1\}$ is a first-order Markov chain with transition probabilities $P_\lambda(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t)$ being multinomial with cell probabilities $\boldsymbol{X}_t/\lfloor \lambda C_t \rfloor$ and number of trials $\lfloor \lambda C_{t+1} \rfloor$. The genetic data at time t are samples from the gamete pool produced by the $\lfloor \lambda C_t \rfloor$ adults. Hence, for $S_t > 0$, $P_\lambda(\boldsymbol{Y}_t|\mathbf{X}) = P_\lambda(\boldsymbol{Y}_t|\boldsymbol{X}_t)$ is multinomial with parameters $\boldsymbol{X}_t/\lfloor \lambda C_t \rfloor$ and S_t. For $S_t = 0$, $P_\lambda(\boldsymbol{Y}_t|\boldsymbol{X}_t) \equiv 1$. Summing out the nuisance parameters $\boldsymbol{X}_1$ over an uninformative prior $\pi(\boldsymbol{X}_1)$ gives the likelihood

$$L(\lambda) = P_\lambda(\mathbf{Y}) = \sum_{\mathbf{X}} P_\lambda(\mathbf{Y}, \mathbf{X}) = \sum_{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T} \pi(\boldsymbol{X}_1) P_\lambda(\boldsymbol{Y}_1|\boldsymbol{X}_1) \prod_{t=2}^{T} P_\lambda(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) P_\lambda(\boldsymbol{Y}_t|\boldsymbol{X}_t).$$

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.

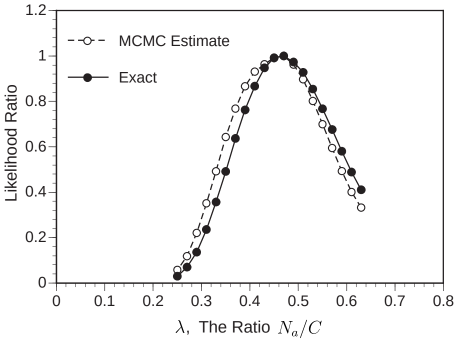
2. Monte Carlo Likelihood

To obtain an efficient Monte Carlo estimate of $L(\lambda)$, we consider the likelihood ratios $L(\lambda)/L(\lambda_0)$, (Thompson and Guo 1991, Geyer and Thompson 1992)

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

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

Estimating a curve for $L(\lambda)$, the range of λ's of interest may be large. In such cases it does not suffice to realize $\mathbf{X}^{(i)}$'s under a single λ_0. We 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.



The figure at left shows the estimated likelihood curve (open circles) from a simulated dataset for which the exact likelihood curve (filled circles) can be computed. The data were simulated for 20 loci with $k = 2$, $\lambda = .4$, C_t varying between 90 and 130, $S_t = 200$, and $\boldsymbol{X}_1$'s drawn from a uniform distribution. The estimated curve is the product of likelihood ratios estimated for each locus with $\Lambda = \{.25, .27, \ldots, .61, .63\}$, $m = 10,000$ and $u = 1,000$.

Figure 1. Estimated and Exact Likelihoods

We are currently investigating more effective MCMC updates, incorporating uncertainty in census size estimates, and extending the approach to more complex models, including age-structured populations with overlapping generations.

This research was supported in part by NSF grants BIR-9256537 and BIR-9807747.

REFERENCES

Geyer, C.J., 1994. Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Tech. Rep. 568r, School of Statistics, University of Minnesota.

Geyer, C.J. and E.A. Thompson, 1992. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). J. Roy. Statist. Soc. Ser. B 54:657–699.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Thompson, E.A. and S.-W. Guo, 1991. Monte Carlo evaluation of likelihood ratios. IMA J. Math. Appl. Med. & Biol. 8:149–169.

RÉSUMÉ

Le ratio de l'effectif par rapport au nombre total d'adultes reproducteurs est un paramètre important dans la structure d'une population. Nous présentons une approche de vraisemblance de Monte Carlo permettant d'estimer ce paramètre à partir de données génétiques, au niveau de plusieurs loci multialléliques.

Chapter 3

Excerpts from "Sequential Forward-Backward Realization of Latent Allele Counts for Efficient Importance Sampling"

These are excerpts from a manuscript in preparation with Ellen Williamson (a postdoc in Monty Slatkin's lab in Berkeley) and Elizabeth Thompson. The manuscript describes a method for realizing latent allele counts by a Baum *et al.* (1970) algorithm. This is the same sort of technique used in implementing the M-sampler for MCMC in pedigrees (Thompson and Heath 1998), though in this case there is a twist; we make several approximations so that we ultimately do not sample latent variables $\mathbf{X}$ exactly from their distribution conditional on the observed data $\mathbf{Y}$. Nonetheless, this method does sample $\mathbf{X}$'s from a distribution that is close to being proportional to $P_N(\mathbf{Y}, \mathbf{X})$.

In this project with Ellen, we cast the problem as one of estimating the effective size N of a population of constant size by straightforward Monte Carlo from an importance sampling distribution which is not a Markov chain. However, the extension to estimating λ in a population of fluctuating, but known, census size is straightforward (as should be the extension to overlapping year-classes). Finally, it should be readily evident that the $P_N^*(\mathbf{X})$ constructed below is practically tailor-made to be a proposal distribution for making multi-component Hastings updates in an MCMC approach to the problem.

3.1 Background on the Problem

Here we describe the problem as one of inference from a hidden Markov chain, and bring to bear a number of techniques previously developed for such inference. In particular we describe Baum *et al.* (1970)-type algorithms for an importance-sampling-based Monte Carlo method of approximating the likelihood.

3.2 Inference from a Hidden Markov Chain

A researcher collects genetic samples at $r+1$ different generations, ($t_0, \ldots, t_r$, with $0 = t_0 < t_r = T$) from a discretely-reproducing population. For notational simplicity the following will pertain to a single observed locus, but the extension to multiple, independently-segregating loci is made by

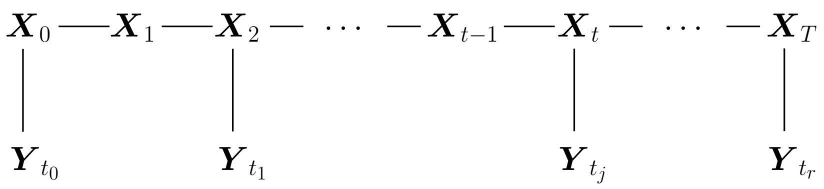


Figure 3.1: An undirected graph showing the dependencies between components of $\mathbf{X}$ and $\mathbf{Y}$. The $\boldsymbol{Y}_t$'s are observations of a hidden Markov chain.

simply multiplying the likelihoods for each locus together. Amongst all of the samples, there are K different allelic types observed, indexed by $k = 1, \ldots, K$. The observed frequencies of the different allelic types will differ through time. The maximum likelihood approach to estimating N, assumes that the population can be modelled as a Wright-Fisher population of size N and then finds the value of N that is most likely given the data.

Let $\boldsymbol{Y}_t = (Y_{t_0,1}, \ldots, Y_{t_r,K})$ be the counts of the K different allelic types in the sample at time t. Denote the sample size at time t by S_t. The unobserved population allele counts at time t are $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,K})$. The $\boldsymbol{X}_t$ form a Markov chain in time, with transitions defined by multinomial probabilities depending on N,

$$P_N(\boldsymbol{X}_t|\boldsymbol{X}_0, \ldots, \boldsymbol{X}_{t-1}) = P_N(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) = N! \prod_{k=1}^{K} \frac{[X_{t-1,k}/N]^{X_{t,k}}}{X_{t,k}!}. \tag{3.1}$$

Observations at time t_j are conditionally independent of everything else, given X_{t_j}, and also follow the multinomial distribution depending on the parameter N and the known sample sizes S_{t_j}:

$$P_N(\boldsymbol{Y}_{t_j}|\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T) = P_N(\boldsymbol{Y}_{t_j}|\boldsymbol{X}_{t_j}) = S_{t_j}! \prod_{k=1}^{K} \frac{[X_{t_j,k}/N]^{Y_{t_j,k}}}{Y_{t_j,k}!}. \tag{3.2}$$

This system is a hidden Markov chain. (Figure 3.1 depicts in graph format). The likelihood for N is the probability of the data $\mathbf{Y} = (\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_T)$ given the parameter N. The nuisance parameters $\boldsymbol{X}_0$ may be integrated out by assuming a prior $\pi(\boldsymbol{X}_0)$ on them. The probability of $\mathbf{Y}$ is the sum of the joint probability of the data, and the latent variables $\mathbf{X} = (\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T)$ over the space of all latent variables

$$L(N) = P_N(\mathbf{Y}) = \sum_{\mathbf{X}} P_N(\mathbf{Y}, \mathbf{X}) = \sum_{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T} \left(\pi(\boldsymbol{X}_0) \prod_{t=1}^{T} P_N(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) \right) \left(\prod_{j=0}^{r} P_N(\boldsymbol{Y}_{t_j}|\boldsymbol{X}_{t_j}) \right). \tag{3.3}$$

For the case of $K = 2$ and N small the likelihood in (3.3) may be computed exactly. Williamson and Slatkin (in press) effected the summation in (3.3) in terms of multiplication of transition probability matrices. We note that the hidden Markov form of the system allows a more efficient computation of the likelihood by way of the algorithm of Baum (1972). However, as K and N increase, the number of terms in the sum increases dramatically, and exact evaluation by either method is infeasible. An alternative is to approximate (3.3) by Monte Carlo.

3.3 Monte Carlo Evaluation

We want to estimate $P_N(\mathbf{Y})$ for a number of different values of N. We can express this probability as an expectation with respect to the distribution of $\mathbf{X}$.

$$P_N(\mathbf{Y}) = \sum_{\mathbf{X}} P_N(\mathbf{Y}, \mathbf{X}) = \sum_{\mathbf{X}} P_N(\mathbf{Y}|\mathbf{X}) P_N(\mathbf{X}) = E_N\Big(P_N(\mathbf{Y}|\mathbf{X})\Big). \tag{3.4}$$

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

$$P_N(\mathbf{Y}) \approx \frac{1}{m} \sum_{i=1}^{m} P_N(\mathbf{Y}|\mathbf{X}^{(i)}) \tag{3.5}$$

for large m, with $\mathbf{X}^{(i)}$ being the i^{th} realization from the marginal distribution of $\mathbf{X}$. The $\mathbf{X}^{(i)}$ may be realized by drawing $\boldsymbol{X}_0$ from the prior distribution, then realizing $\boldsymbol{X}_1$ given $\boldsymbol{X}_0$ by (3.1), then $\boldsymbol{X}_2$ given $\boldsymbol{X}_2$ and so forth to $\boldsymbol{X}_T$. Doing so, however, will almost always yield $\mathbf{X}$'s that bear no resamblance to the data, or are simply incompatible with it. For example, suppose that on a given realization $\boldsymbol{X}_{t,k} = 0$ but $\boldsymbol{Y}_{t_j,k} > 0$ for some $t_j > t$. Then, $P_N(\mathbf{Y}|\mathbf{X}) = 0$ and that realization contributes nothing to our Monte Carlo sum. Williamson and Slatkin (unpublished data) avoided the problem of zero contributions to the Monte Carlo sum by first realizing $\boldsymbol{X}_0$ from its distribution conditional on $\boldsymbol{Y}_)$ and then ensuring that alleles could not go extinct in the simulated $\mathbf{X}$ if they appeared in future samples, and then reweighting the summand in (3.5) appropriately. Though this eliminates the zero contributions, it does not address the problem that $P_N(\mathbf{Y}, \mathbf{X})$ is still very small for all but a tiny fraction of the $\mathbf{X}$'s realized—a situation that leads to huge Monte Carlo variance; indeed, Williamson and Slatkin (unpublished) simulated $\mathbf{X}$'s for 30 days on one dataset, and still did not observe adequate convergence of their Monte Carlo estimate.

Here, we pursue a more refined importance sampling (Hammersley and Handscomb 1964) scheme. We may express $P_N(\mathbf{Y})$ as an expectation with respect to some other distribution of $\mathbf{X}$, say $P_N^*(\mathbf{X})$. Then the term after the second equals sign in (3.4) may be rewritten so that:

$$P_N(\mathbf{Y}) = \sum_{\mathbf{X}} \frac{P_N(\mathbf{Y}|\mathbf{X}) P_N(\mathbf{X})}{P_N^*(\mathbf{X})} P_N^*(\mathbf{X}) \tag{3.6}$$

which is equal to the expectation

$$P_N(\mathbf{Y}) = E_N^* \left(\frac{P_N(\mathbf{Y}, \mathbf{X})}{P_N^*(\mathbf{X})} \right) \tag{3.7}$$

where E_N^* indicates that the expectation is over $\mathbf{X}$'s weighted by the distribution $P_N^*(\mathbf{X})$. The expectation (3.7) may be approximated by Monte Carlo. Hence,

$$P_N(\mathbf{Y}) \approx \tilde{P}_N(\mathbf{Y}) = \frac{1}{m} \sum_{i=1}^{m} \frac{P_N(\mathbf{Y}, \mathbf{X}^{(i)})}{P_N^*(\mathbf{X}^{(i)})} \tag{3.8}$$

for suitably large m where $\mathbf{X}^{(i)}$ is the i^{th} realization of $\mathbf{X}$ drawn from $P_N^*(\mathbf{X})$. The Monte Carlo variance of the estimator $\tilde{P}_N(\mathbf{Y})$ will be minimized by choosing a $P_N^*(\mathbf{X})$ which is exactly proportional to $P_N(\mathbf{Y}, \mathbf{X})$. Of course that distribution would be $P_N(\mathbf{X}|\mathbf{Y})$. However, if we could compute $P_N(\mathbf{X}^{(i)}|\mathbf{Y})$ for any $\mathbf{X}^{(i)}$, then we could compute $P_N(\mathbf{Y}) = P_N(\mathbf{Y}, \mathbf{X}^{(i)})/P_N(\mathbf{X}^{(i)}|\mathbf{Y})$ and would not need Monte Carlo at all! Since this is obviously not the case, we take up the task in the following section of constructing a distribution $P_N^*(\mathbf{X})$ that is close to proportional to $P_N(\mathbf{X}|\mathbf{Y})$, is easily sampled from, and for which $P_N^*(\mathbf{X}^{(i)})$ can be quickly computed.

3.4 Sampling from $P^*_N(\mathbf{X})$ by a Forwards-Backwards Method

We could obtain a realization exactly from $P_N(\mathbf{X}|\mathbf{Y})$ by employing an exact version of the Baum *et al.* (1970) algorithm. Unfortunately doing so would require more computation than computing $P_N(\mathbf{Y})$ outright. However, by assuming that the $\mathbf{X}$ and the $\mathbf{Y}$ follow an approximate multivariate normal distribution with special treatment for the parts of that distribution that fall beyond the bounds of 0 or N, we have developed a computationally cheap method for generating $\mathbf{X}$'s from a distribution that is very close to $P_N(\mathbf{X}|\mathbf{Y})$. We describe below the computational procedures which gives us realizations from $P^*_N(\mathbf{X})$ and allow us to compute $P^*_N(\mathbf{X}^{(i)})$ while doing so.

The method we use is a "forward-backward" method assuming a normal approximation to genetic drift and sampling in each generation. We realize each of the K alleles sequentially. To describe this, we introduce several more pieces of notation. First, we will consider the possibility that the size of the population changes from generation to generation so we have $\boldsymbol{N} = (N_0, \ldots, N_T)$. When we are simulating from P^*_N, we will set the population sizes constant at N, *i.e.*, $\boldsymbol{N} = (N, \ldots, N)$, for realizing the first allele, but then will update $\boldsymbol{N}$ based on our realization for the first allele. Also, 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 $\boldsymbol{Y}_{(k)} = (Y_{t_0,k}, \ldots, Y_{t_r,k})$, and we let $\boldsymbol{S} = (S_{t_0}, \ldots, S_{t_r})$. Finally, we introduce two new terms which will be used to ensure that we don't realize any $\mathbf{X}$'s for which $P_N(\mathbf{Y}, \mathbf{X}) = 0$. Let $\delta_{t,k} = 1$ denote that the realized value of $X_{t,k}$ must be greater than zero while $\delta_{t,k} = 0$ implies that $X_{t,j}$ may be greater than or equal to zero, and let $\kappa_{t,k}$ be the number of alleles with subscripts $\ell : k < \ell \leq K$ for which $Y_{t_j,\ell} > 0$ for at least one $t_j \geq t$.

The method for obtaining a realization for $\boldsymbol{X}_{(k)}$—the latent counts of the k^{th} allele through all the generations—given an $\boldsymbol{N}$, an $\boldsymbol{S}$, and the data $\mathbf{Y}$ is described below. With a multiallelic locus, one works through the alleles sequentially, first realizing $\boldsymbol{X}_{(1)}$ then setting $\boldsymbol{N} \leftarrow \boldsymbol{N} - \boldsymbol{X}_{(1)}$ and $\boldsymbol{S} \leftarrow \boldsymbol{S} - \boldsymbol{Y}_{(1)}$, and then realizing $\boldsymbol{X}_{(2)}$, and so forth. To pursue this method, it is easiest to deal not with $\boldsymbol{X}_{(k)}$ and $\boldsymbol{Y}_{(k)}$ directly, but rather with the corresponding angularly-transformed allele frequencies. Thus our sample information about allele 1 at time t when a sample exists becomes the random variable $\phi_{t,k} = \sin^{-1}(Y_{t,k}/S_t)^{1/2}$, and the latent variable corresponding to the first allele at time t is $\theta_{t,k} = \sin^{-1}(X_{t,k}/N_t)^{1/2}$.

Following Cavalli-Sforza and Edwards (1967), if $\theta_{t-1,k}$ is normally distributed with mean μ_{t-1} and variance σ^2_{t-1} then, after a generation of genetic drift, $\theta_{t,k}$ will be approximately normal with mean μ_{t-1} and variance $\sigma^2_t = \sigma^2_{t-1} + 1/(4N_t)$. If there happens to be data $\phi_{t,k}$ at time t, then $\phi_{t,k}$ has an approximate normal distribution with mean $\theta_{t,k}$ and variance $1/(4S_t)$, so, given that $\theta_{t,k} \sim \mathcal{N}(\mu_t, \sigma^2_t)$, the conditional distribution of $\theta_{t,k}$ given $\phi_{t,k}$ will also be normally distributed (like the Bayesian posterior distribution given a normal prior and normal data). These relations form the basis of the forward step in the Baum *et al.* (1970) algorithm which works as follows:

3.4.1 The Forward Step:

For $t = 0$ we assume that the uniform prior on $\boldsymbol{X}_0$ is equivalent to a diffuse prior on $\theta_{0,k}$, so $\theta_{0,k}|\phi_{0,k} \sim \mathcal{N}(\mu_0, \sigma^2_0)$ with $\mu_0 = \phi_{0,k}$ and $\sigma^2_0 = 1/(4S_0)$. With that as a starting point, we work forward in time assigning values at time t:

$$\mu_t \longleftarrow \mu_{t-1} \tag{3.9}$$

$$\sigma^2_t \longleftarrow \sigma^2_{t-1} + 1/(4N_t) \tag{3.10}$$

to account for drift. If there are genetic data at time t, then the values of μ_t and σ_t^2 are updated to reflect that before moving on to $t+1$, *i.e.*,

$$\mu_t \longleftarrow \frac{\mu_t/(4S_t) + \sigma_t^2 \phi_{t,k}}{1/(4S_t) + \sigma_t^2} \tag{3.11}$$

$$\sigma_t^2 \longleftarrow \frac{\sigma_t^2/(4S_t)}{1/(4S_t) + \sigma_t^2}. \tag{3.12}$$

Carrying this process on to $t = T$ gives values for the approximate (*i.e.*, assuming the normal approximation to genetic drift and genetic sampling) mean and variance of $\theta_{T,k}$ given $\phi_{t_0,k}, \dots, \phi_{t_r,k}$. In fact, for each t, it gives us the parameters of the assumed normal distribution for $\theta_{t,k}$ conditional on $\phi_{t_j,k}$ for all $t_j \leq t$. We are thus in a position to realize the $\theta_{t,k}$'s in the "backward" step and transform those $\theta_{t,k}$'s back into $X_{t,k}$'s.

3.4.2 The Backward Step:

In undertaking the backward step, we first realize $\theta_{T,k}$ from a $\mathcal{N}(\mu_T, \sigma_T^2)$ distribution. Then we convert that to a realized value of $X_{T,k}$ Here we encounter a difficulty; a special rule is required to convert any realized value of $\theta_{t,k}$ to the realized value for $X_{t,k}$. The simple inverse transformation $X_{t,k} = N_t \sin^2 \theta_{t,k}$ will not work, first because $X_{t,k}$ must be an integer, and second because $\theta_{t,k}$ may take values which would then yield $X_{t,k}$'s less than zero or greater than N_t. Thus, though we realize each $\theta_{t,k}$ from a $\mathcal{N}(\mu_t, \sigma_t^2)$ distribution, we then convert that to a realization of $X_{t,k}$ by the many-to-one mapping $\mathcal{M}$ that sends real numbers to integers between $\delta_{t,k}$ and $N_t - \kappa_{t,k}$ inclusive. $\mathcal{M}$ has the effect of folding and translating parts of the distribution of $\theta_{t,k}$ so that it is bounded between 0 and $\pi/2$, and then discretizing $\theta_{t,k}$ into appropriate values of $X_{t,k}$. A complete description of $\mathcal{M}$ is found in Section 3.5.1.

After realizing $X_{T,k}$ you work backward, updating μ_{T-1} and σ_{T-1}^2 based on the realized value of $\theta_{T,k}$. This process is repeated until $X_{),k}$ has been realized. The probability of realizing each $X_{t,k}$ can come computed as detailed in Section 3.5.2. So long as the sequence of alleles is followed for realizing $\boldsymbol{X}_{(k)}$'s each time, the product of $\mathcal{R}$'s gives the desired probability of realizing $\mathbf{X}$, $P_N^*(\mathbf{X})$. (That last section is a bit hasty.)

3.5 Details of $\mathcal{M}$ and $\mathcal{R}$

3.5.1 The Map $\mathcal{M}$

Let $\mathcal{M}(\theta; \delta, \kappa, N) : \Re^1 \to (\delta, \dots, N - \kappa)$ be the many-to-one map that takes a realization of $\theta \in (-\infty, \infty)$ to the corresponding realization of the integer X such that $\delta \leq X \leq N - \kappa$. (Should note somewhere that $\delta \in \{0, 1\}$ and $\kappa \in \{0, 1, \dots, N - \delta\}$.)) $\mathcal{M}$ may be described by the following pseudocode. We first define the quantities $L = \sin^{-1}(.5/N)^{1/2}$ and

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

Then, given a realized value of θ we map it to an X as follows:

```
**if** $(L \leq \theta < H)$ **then** $X \longleftarrow \lfloor N \sin^2 \theta + .5 \rfloor$
**else if** $(\theta < L)$
```

```
        and if $(\delta = 0)$ then $X \longleftarrow 0$
        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 $X \longleftarrow \lfloor N \sin^2 \theta^{(L)} + .5 \rfloor$
            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 $X \longleftarrow \lfloor N_t \sin^2 \theta^{(i^*)} + 1/2 \rfloor$ 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 $X \longleftarrow N$
        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 $X \longleftarrow \lfloor N \sin^2 \theta^{(H)} + .5 \rfloor$
            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 $X \longleftarrow \lfloor N_t \sin^2 \theta^{(j^*)} + 1/2 \rfloor$ where $j^*$ is the least $j$ such that
                $L \leq \theta^{(j)} < H$. (The sequence $\theta^{(j)}$ represents successive translation rightward).
```

3.5.2 The probability $\mathcal{R}_{\mu,\sigma^2}(x; \delta, \kappa, N)$ of realizing $X = x$

If θ is realized from a $\mathcal{N}(\mu, \sigma^2)$ distribution, then the probability that the corresponding realization of $X = x$ can be expressed using the notation from the above section. First, $\mathcal{R}_{\mu,\sigma^2}(x; \delta, \kappa, N) = 0$ for $x < \delta$ or $x > N - \kappa$, and $\mathcal{R}_{\mu,\sigma^2}(\delta; \delta, \kappa, N) = 1$ when $\delta = N - \kappa$ (VERIFY THIS, ERIC!). Otherwise, for $x = 0$ and $x = N$ we have

$$\begin{aligned} \mathcal{R}_{\mu,\sigma^2}(0; 0, \kappa, N) &= P(-\infty < \theta < L) \\ \mathcal{R}_{\mu,\sigma^2}(N; \delta, 0, N) &= P(H \leq \theta < \infty). \end{aligned} \tag{3.13}$$

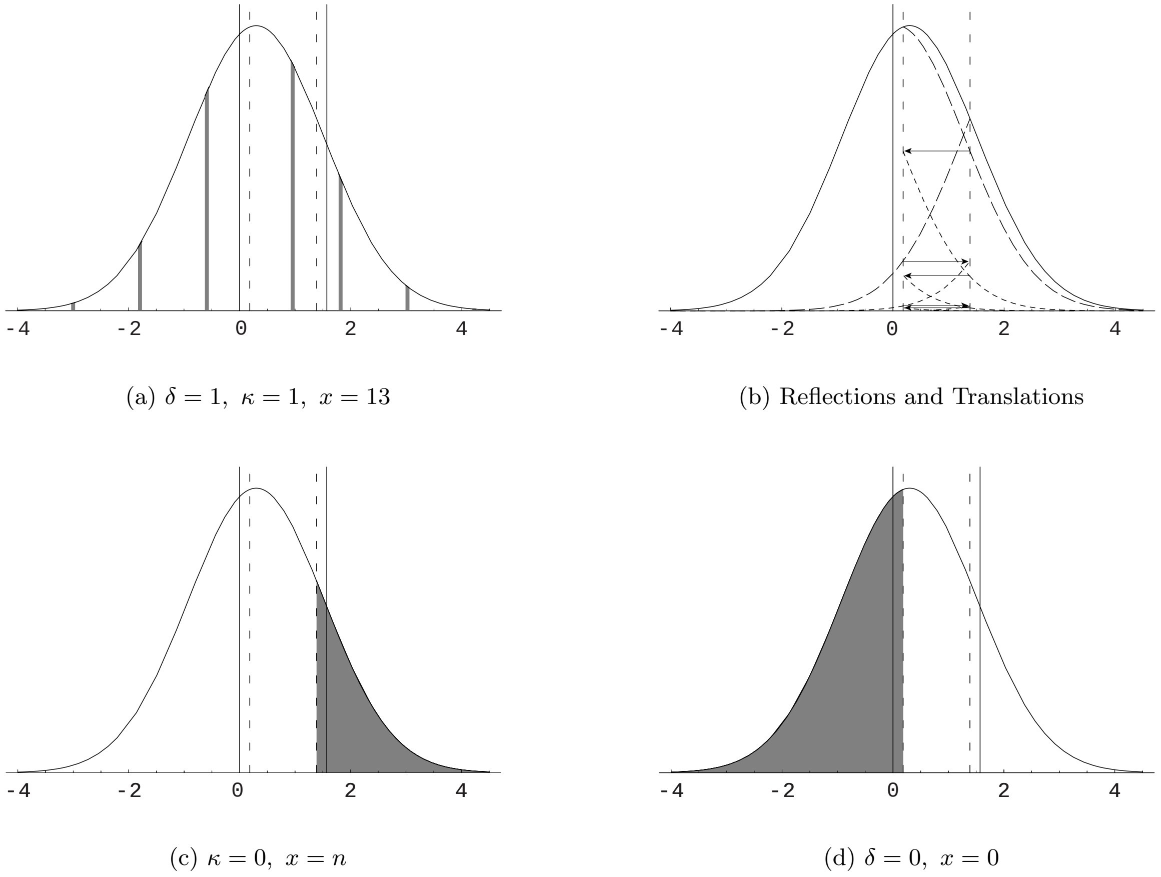
For values of x between 0 and $N - \kappa$, defining $a = \sin^{-1}[(X - 1/2)/N]^{1/2}$ and $b = \sin^{-1}[(X + 1/2)/N]^{1/2}$,

$$\begin{aligned} \mathcal{R}_{\mu,\sigma^2}(x; \delta, \kappa, N) &= P(a \leq \theta < b) \\ &\quad + I\{\delta = 1\} P(a \leq \theta^{(L)} < b) + I\{\kappa > 0\} P(a \leq \theta^{(H)} < b) \\ &\quad + I\{\delta = 1\} \sum_{i=1}^{\infty} P(a \leq \theta^{(i)} < b) + I\{\kappa > 0\} \sum_{j=1}^{\infty} P(a \leq \theta^{(j)} < b) \end{aligned} \tag{3.14}$$

where $I\{\cdot\}$ is an indicator function and $P(a \leq \theta < b)$ is the probability that a normal random variable with mean μ and variance σ^2 is between a and b, namely $\int_a^b (2\pi\sigma^2)^{-1/2} \exp\{[-(\theta - \mu)^2]/(2\sigma^2)\} d\theta$. 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).

3.6 The Performance of our $P_N^*(\mathbf{Y})$

I have assessed how well this scheme works in two ways. First, for a diallelic locus, it is possible to compute, by a different sort of Baum algorithm, the marginal probabilities of each $\boldsymbol{X}_t$ given the above scheme for realizing them. Comparing this to the exact marginal probabilities of each $\boldsymbol{X}_t$ given $\mathbf{Y}$ shows that the approximation is quite good. I have also visually inspected the realizations from our $P_N^*(\mathbf{X})$ with some real-time graphics displays I have coded up. It does produce realizations that are close to the desired distribution. Predictably it works best when all the alleles at a locus have been found at intermediate frequencies in the samples. In loci with many (say ≥ 6) alleles, some of which are at low frequency ($< .05$,, say), $P^*_N(\mathbf{X})$ doesn't perform quite as well. However, it does generate enough reasonable realizations that it should be an excellent proposal distribution in a Metropolis-Hastings framework.



(a) $\delta = 1,\ \kappa = 1,\ x = 13$

(b) Reflections and Translations

(c) $\kappa = 0,\ x = n$

(d) $\delta = 0,\ x = 0$

Figure 3.2: Figures representing the action of the map $\mathcal{M}$. The normal curve is the density for θ. (a) When $\delta = 1$ and $\kappa = 1$ recall that x is constrained so that $x \in \{1 \ldots, N-1\}$. In this case, $N = 20$; the shaded regions correspond to the probability that $\mathcal{M}(\theta, 1, 1, 25) = 13$. (b) A diagram of reflections and translations implicit in $\mathcal{M}(\theta, 1, 1, N)$. 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. (c,d) Shaded areas show values of θ for which $\mathcal{M}$ returns an x of N or 0 respectively.

Chapter 4

A Proposal for Fully-Bayesian Mixed Stock Fishery Analysis Using Reversible Jump MCMC

This is a general outline of a project I am pursuing in Dr. Julian Besag's Stochastic Modelling course (STAT 518). I initiated this project only recently, but have been looking forward to it since Peter Green's talk in Biostatistics last year. I have not worked out the exact details on the pairs of reversible jump moves. Nor have I yet given enough thought to the many advantageous ways of post-processing the output, but this will be a primary focus during the spring quarter.

4.1 General Introduction and Overview

Managing the salmon populations of the West coast in the face of the many competing interests which impact their well-being is an outstandingly difficult problem. Crucial to appropriate management is the recognition that salmon populations are (typically) reproductively-isolated, distinct populations, and their management should reflect that fact (Ricker 1972). Commercial harvest of salmon is a major source of mortality in some populations, but one which may be regulated through policy. However, an obstacle to regulating fishing from a population-based management perspective is that salmon are typically caught in the ocean (or the mainstems of rivers) where fish from many different populations are intermingled. It is therefore difficult to monitor the impact of certain commercial fisheries on specific salmon populations. Fortunately, genetic data from different salmon populations of interest, in conjunction with statistical techniques, allow salmon managers to determine the relative contributions of different salmon populations to the mixture of fish being caught in a fishery.

Beginning in the early 1980's, electrophoretically-detectable allozyme polymorphisms and maximum-likelihood, mixed-stock fishery analysis (Milner *et al.* 1981; Fournier *et al.* 1984; Millar 1987; Smouse *et al.* 1990) have been used to assist managers in estimating the proportion of catch coming from different source populations from which "baseline" genetic data have previously been collected. Smouse *et al.* (1990) extended such analyses to include the possibility that one or more populations for which no baseline data were available, were contributing to the mixture. Though Millar (1987) suggested the infinitesimal jackknife as a way of assessing uncertainty in estimates from genetic stock identification, bootstrap methods are typically used today (Pella and Milner 1987).

I propose to deal with the contribution of unsampled stocks to the mixture by methods for fully-Bayesian mixture analysis with unknown number of components as detailed by Richardson and Green (1996) using Reversible Jump Markov Chain Monte Carlo (RJMCMC). Such an approach should provide a more useful, Bayesian criterion for model selection (where different models correspond to different numbers of populations contributing to the mixture), and might alleviate the need for an extensive "testing cascade" as pursued by Smouse *et al.* (1990). In addition, the output from the RJMCMC should provide direct insight into issues traditionally of importance in genetic stock identification: the question of whether different sources with similar baseline genotype frequencies should be split or lumped for the purposes of the analysis, the question of which loci are the most informative, and classifying individual fish in the mixture sample as to their population of origin. The MCMC approach should provide a natural framework for summing over latent variables at isoloci (proteins which are coded for at duplicate loci in the genome). The Bayesian nature of the approach will allow incorporation of prior knowledge (for example, if one population is many times larger than another, you might expect, *a priori* that it will make a greater contribution to the mixture), and should provide a flexible tool for decision-making fisheries managers.

4.2 Salmon and Genetics Background

Salmon have a remarkable talent for returning to the stream from whence they were born.[1] We call an assemblage of fish that all reproduce in the same river a salmon "population." This means different salmon populations have different evolutionary histories. In fact, many populations show evidence of adaptation to the physical characteristics of their home rivers. Each of these populations is, in effect, a unique entity with its own heritage; the populations vary with respect to their size and their resilience to fishing pressure, weather cycles, *etc.* Fishing pressure on salmon must be regulated so as to not drive to extinction populations of salmon which are at risk of being lost. However, fishing typically occurs in the ocean where it is impossible to tell if a netted salmon is from one population or another. It is, however, possible and practical to estimate the proportion that each of a number of populations contribute to the catch of commercial fishers in the ocean. This is done by exploiting the genetic differences between the populations.

Because each salmon population has its own evolutionary history, each will have different frequencies of genetic types. Typically the genetic markers used are neutral, electrophoretically-detectable, allozyme polymorphisms. These polymorphisms are coded for at different sites (loci) on the chromosomes of the fish. At a particular locus, the value of the two different alleles (one maternal in origin, the other, paternal) will determine the fish's genotype at that locus. The full complement of single-locus genotypes present in a fish, over all the loci assayed, and the relative dominance of alleles over one another (as well as the presence of isoloci) determines the fish's multilocus phenotype (MP).[2] Each salmon population will possess distinct allele frequencies which, in turn, leads each population to have different proportions of MP's. It is these differences that allow the contribution of different populations to the mixture to be estimated.

Crucial to making this scheme useful is the fact that allele frequencies from some salmon populations which might contribute to the mixture have been estimated from samples of fish spawning

[1] A salmon's homing to its natal stream is not perfect—if it were, none of the rivers in the Northwest would have salmon in them after the last Ice Age—nonetheless, for wild populations of some species, homing is quite good.

[2] I use the phrase "multilocus phenotype" instead of multilocus genotype because some loci have alleles that may not be scored individually, *i.e.*, the heterozygote of two alleles, may appear as a homozygote of one of them. Additionally, "multilocus genotype" in the human genetics context implies knowledge of the phase (maternal or paternal origin) of each of the alleles, and this is not the case here.

in their natal streams (or from samples of their outmigrating offspring). This sampling contributes several terms to the likelihood (a refinement introduced by Smouse *et al.* (1990)) allowing for uncertainty in the baseline estimates. The problem we wish to solve is that of estimating the proportion of fish from different source populations, given the estimates of allele frequencies in those source populations and a sample of fish taken from the mixture. If there are populations in the mixture which were not sampled previously (*i.e.*, for which no baseline data exist) then this becomes a mixture problem with unkown number of components. In that context, one wishes also to estimate the number of unknown populations contributing to the mixture, their contributions to the mixture, and the allele frequencies in those populations.

4.3 The Likelihood

We begin by reviewing the likelihood solution to the problem, and then extend that into a Bayesian framework. Let there be k populations contributing to the mixture, and let us have baseline data from k^* of these.[3] These baseline data come from samples of size n_i fish from each of the baseline populations. They are the set of $y_{i\ell}$—the multilocus phenotype of the ℓ^{th} fish from the i^{th} source population—with $i = 1, \dots, k^*$ and $\ell = 1, \dots, n_i$. From the mixture, a sample of $n_{(m)}$ fish are drawn. The data from the mixture are the $y_{(m)\ell}$, $\ell = 1, \dots, n_{(m)}$. All of the data are denoted by $y = \{y_{i\ell} : i = 1, \dots, k^*;\ \ell = 1, \dots, n_i\} \cup \{y_{(m)\ell} : \ell = 1, \dots, n_{(m)}\}$. The allele frequencies at each locus in each of the k source populations are parameters in this context. We denote these parameters for the i^{th} source population as x_i and let $x = (x_1, \dots, x_k)$. The probability of finding a fish of a particular multilocus phenotype depends on the allele frequencies, the assumptions of Hardy-Weinberg equilibrium at individual loci and independent segregation between different loci, and then dominance and isolocus considerations. Accounting for these factors is relatively straightforward, but for notational convenience, now, we just express the probability of finding multilocus phenotype j in the i^{th} source population, given its allele frequencies as $g(j, x_i)$. If the ℓ^{th} fish from the i^{th} source or from the mixture has multilocus phenotype j, we say that $y_{i\ell} = j$ or $y_{(m)\ell} = j$, respectively. Finally the parameter vector $w = (w_1, \dots, w_k)$ is the vector of mixture proportions that populations 1 through k contribute to the mixed stock fishery.

The likelihood for x and w given the data is:

$$L(w, x|y) = p(y|w, x) = C \cdot \left[\prod_{\ell=1}^{n_{(m)}} \left(\sum_{i=1}^{k} w_i g(y_{(m)\ell}, x_i) \right) \right] \cdot \prod_{i=1}^{k^*} \prod_{\ell=1}^{n_i} g(y_{i\ell}, x_i) \tag{4.1}$$

where C is the product of several multinomial coefficients. The contribution of the samples from each of the baselines when all alleles are codominant and there are no isoloci factorizes into a multinomial probability for counts of different allelic types (rather than MP's). Thus the number of factors may be reduced considerably when computing (4.1).

This likelihood may be maximized by an EM-algorithm as detailed in Smouse *et al.* (1990). To do so, one must create a complete-data specification by assigning to each fish in the mixture sample an allocation (random) variable which indicates which source that fish is from. Such a latent allocation variable will also be very useful in the Bayesian approach. For the ℓ^{th} fish from

[3] Note that, in reality, some of the baseline populations may not contribute at all to the mixture, so k^* might be greater than k. Primarily because I haven't come up with a decent notation for expressing such a situtuation, I assume here that $k^* < k$ and that the stocks sampled in the baseline are a subset of the stocks contributing to the mixture. This assumption will be relaxed as the project proceeds.

the mixture, we denote by z_ℓ the latent allocation variable which takes values between 1 and k, indicating the population of origin of the fish. We denote all the allocation variables by $z = (z_1, \dots, z_{n_{(m)}})$.

Smouse *et al.* (1990) treated the problem of unsampled stocks contributing to the mixture ($k^* < k$), and the desirability of lumping versus splitting of baseline populations by considering a series of non-nested models in which stocks were either split or lumped, and excluded or included, and in which either one or more unsampled stocks were in the mixture. The various models were compared on the basis of the increase in maximum log-likelihood, relative to increase in the number of free parameters, but the authors note that the models are not nested, so the log-likelihood differences are not necessarily asymptotically chi-square distributed. They suggest using the increases in log-likelihood as rough guidelines only.

4.4 A Bayesian Formulation

With priors for w and x and the likelihood in (4.1), a Bayesian formulation of the problem when k is constant would give the joint distribution

$$p(w, x, y) = p(w)p(x)p(y|w, x). \tag{4.2}$$

Furthermore we can consider the number of components, k, in the mixture to be a random variable with prior $p(k)$, and we may consider the allocation variables z in our model. Then we may write the joint density as

$$p(k, w, x, z, y) = p(k)p(w|k)p(x|k)p(z|w, k)p(y|x, z). \tag{4.3}$$

This factorization obtains because w and x have dimensionality depending on k, the prior for z (when the multilocus phenotypes are not known) depends on w (with k included there since the support of z depends on k), and finally, the data y are conditionally independent of w given x and z.

For greater flexibility one may wish to include some hyperparameters with appropriate hyperpriors. I will investigate this as I proceed. For now, we specify priors $p(k)$, $p(w|k)$, and $p(x|k)$.

The priors $p(w)$ and $p(x)$ shall be Dirichlet distributions. The Dirichlet prior is suitable for x on the grounds of population genetics theory. The allele frequencies at a single locus will presumably not be too far from existing in "drift-mutation" equilibrium such that they should follow a Dirichlet distribution (Wright 1937). Thus the components of x_i from a single locus would follow a Dirichlet distribution with number of variables equal to one less than the number of distinct allelic types found at that locus in all the baselines and the mixture.[4] Empirical evidence of allele frequencies in salmon populations supports this assumption; a histogram of allele frequencies for at least 50 loci in 177 chinook populations follows a classic, upward-facing, beta-distribution "U" shape (Waples 1990a). Such data from many salmon populations may be useful for determining the parameters of Dirichlet priors. Alternatively, it may be preferable to adapt "uniform" Dirichlet priors for each locus.

The weights w of the different populations are proportions and may also follow a Dirichlet prior. Often, a uniform prior may be appropriate, however, surely if one population is of size 40,000 individuals, and another population is of size 1,000 individuals, then there is some prior information

[4] It may be desirable to allow for allelic types that were not sampled at all, giving a Dirichlet prior with one extra component having very little weight. This will be explored.

about the sorts of relative contributions those two populations might make to a mixture. One way of deriving a prior for these weights given estimates of population sizes N_i, $i = 1, \ldots, k$, is to define the prior as

$$w \sim \text{Dirichlet}(QN_1/N, \ldots, QN_k/N) \Longrightarrow p(w) = C_D(w_1)^{\frac{QN_1}{N}} \cdots (w_k)^{\frac{QN_k}{N}} \tag{4.4}$$

where $\sum w_i = 1$, $N = \sum_{i=1}^k N_i$, and C_D is the normalizing constant for this Dirichlet distribution, and $Q \geq 0$ is a factor which expresses the strength we choose to impart to the prior on w. In effect, this prior arises by imagining that without population size information there would be a uniform prior on w, but that we draw an ideal sample of size Q well-labelled fish from the mixture and find that the fish in the sample are in the proportions expected from their population sizes. The posterior distribution for w from this thought experiment then becomes the prior distribution for w for the mixed fishery analysis. Larger Q implies that we give more weight to the population size data. Choice of Q could depend upon many things.

The prior on k could follow a Poisson distribution, or some other distribution deemed appropriate on the basis of knowledge of the number and sizes of populations for which no baseline data are available, but which may nonetheless contribute to the mixture.

4.5 MCMC Techniques

There will be three standard MCMC moves, and two pairs of reversible-jump type moves. The three standard moves are 1) updating w, 2) updating the allocation vector z, and 3) updating x, which may be done in three different Gibbs steps, the computations for which resemble those in the EM algorithm for finding the maximum likelihood estimate. The full conditional distribution for w is Dirichlet by conjugacy, since knowing z gives multinomial data depending on w (one merely classifies and counts all fish in the mixture according to z). Hence, if $p(w|k)$ is Dirichlet$(\delta_1, \ldots, \delta_k)$ then the full conditional for w is Dirichlet$(\delta_1 + \#\{z_\ell = 1\}, \ldots, \delta_k + \#\{z_\ell = k\})$, where $\#\{z_\ell = i\}$ is the number of fish in the mixture sample having allocation variable equal to i. We sample Dirichlet random variables by realizing independent gamma random variables with appropriate parameters and then scaling them so they sum to one.

Next we update z from its full conditional distribution. The full conditional for each z_ℓ can be found with Bayes' rule,

$$P(z_\ell = i | \cdots) = \frac{w_i g(y_{(m)\ell}, x_i)}{\sum_{i=1}^k w_i g(y_{(m)\ell}, x_i)} \ , \ i = 1, \ldots, k \tag{4.5}$$

and then easily sampled from by drawing a Uniform$(0,1)$ random variable. Finally, the allele frequency parameters need to be updated. The full conditionals for these parameters will be Dirichlet distributions for each locus within each population. This obtains because, conditioning on the z_ℓ, it is possible to assign fish from the mixture into the appropriate baseline samples, and then count the number of allelic types in all of those "combined samples."

Either the presence of null alleles or isoloci in the data could complicate the update steps for z and x. However, it seems plausible that the convenient conjugacy could be maintained in each situation by introducing some new variables. For example, in the null allele case, one could introduce a variable indicating the allelic type of the "masked" allele (either a null allele or another copy of the dominant allele) at the locus. These new variables could then be updated from their full conditionals (which would probably be multinomial in form) in separate steps during each sweep.

For the moves that are specifically of a reversible-jump nature, *i.e.*, the split-combine or the birth-death updates to the sources of the mixture, I will use the methodology developed by (GREEN 1995). I haven't figured out the specifics of how these moves will be done, and which variables will be allowed to undergo them. Since some of the sources have baseline data, and hence are permanently labelled, while sources for which no baseline data are available may appear and disappear in the sampling, and be "different" source populations each time, I must still do a little thinking about how to best approach this.

4.6 Using the Output

Most importantly, one gets samples from the posterior for w, marginalized over the unkwown number of stocks contributing to the mixture. Of course, one can also look at posterior distributions conditional on the values of other variables. Exploring these possibilities will be an important part of the project. Dr. Robin Waples from the National Marine Fisheries Service has kindly provided a dataset from the Columbia River in 1980 and 1981 as a good test case. It is the same data that he and others treated in SMOUSE *et al.* (1990).

Chapter 5

Proposed Research Statement Excerpted From NSF Grant #BIR-9807747, "Computational Methods for Inference of Population Parameters," (PI: E.A. Thompson)

This is the "proposed research" section from the grant proposal that Elizabeth Thompson and I wrote in December 1997. The proposal was submitted to the Computational Biology program at the National Science Foundation, and a three-year grant was awarded in September 1998. This grant is my primary funding source for the remainder of my student years at the University of Washington.

5.1 The genetic monitoring of populations

In conservation, much of the methodology of genetic management has focused on very small captive or semi-captive species or populations, the members of which are individually identified (LACY *et al.* 1995). However, in the monitoring of natural populations, and in the reestablishment of wild populations of endangered species, it may be neither feasible nor appropriate to identify each individual and the detailed pedigree structure of the population (BALLOU *et al.* 1995). Instead, there may often be information on population size and age structure, and genetic samples may be obtainable. In order to conserve the population or to provide successful population management, for example by providing migration corridors or by moving subpopulation groups, it is necessary first to be able to estimate well the relevant population parameters (FOOSE *et al.* 1995).

The effective population size N_e (WRIGHT 1931) is the most fundamental parameter of a population from the perspective of its genetic characteristics, affecting both the rate of inbreeding and of loss of genetic variability in the population. The variance effective size determines the increase of gene frequency variance in future generations, while the inbreeding effective size affects the increase in gene identity by descent. In some cases these two effective sizes take different values (CROW and MORTON 1955), but in others they are comparable (CHARLESWORTH 1980). Conservation geneticists must establish a population's N_e as a measure of its genetic risk; inbreeding due to small effective size greatly increases the probability of population extinction in typically outbreeding species (Frankham 1995b), and small effective population size over prolonged periods may lead to "mutational meltdown," resulting in the eventual loss of the population due to the load of deleterious recessive mutations (Lynch *et al.* 1995).

Effective size, N_e, is determined by N, the number of potentially reproducing adults in a population, and by a number of intergenerational factors such as sex-ratio, reproductive success, and family size distribution. Conservation biologists may often readily estimate N, for example by aerial surveys of herd species (such as bison, Wolfe and Kimball 1989) or weir counts of returning salmon (Smith *et al.* 1997). Thus, a key quantity in assessing genetic risk is λ, the ratio of the effective number of reproducing individuals to the census number of adults potentially contributing to a cohort of offspring. Knowing λ and the age-class sizes of a population, a biologist could make informed predictions of the expected loss of genetic variability in future years. An important feature of this perspective is the separation of multigenerational factors such as fluctuating population size from the factors affecting the distribution of family sizes in a single season of reproduction. In a population with non-overlapping generations and no age-structure, λ is the ratio N_e/N, but this definition does not prevail in populations with more complex structure.

Frankham (1995a) reviews 192 published N_e/N ratios from 102 species. The ratio varies widely among species and even among different populations of the same species. Much of this variation is attributable to the different life histories of the organisms studied, but it is also a reflection of the inability of the N_e/N ratio to capture much information in the context of overlapping generations and fluctuating population size. Except in organisms with the simplest life histories, the available methods for estimating effective size do not allow estimation of λ, the most relevant parameter in a genetic monitoring context. For example, of the demographic methods, the formulae of Crow and Denniston (1988) for estimating N_e from the sex ratio and variance in family size apply only to populations with discrete generations, while the method of Nunney and Elam (1994) for organisms with overlapping generations, being based on the formulae of Hill (1979), essentially assumes that the population is of constant size and age structure—the method yields an N_e that does not clearly relate to λ.

Demographic methods require knowledge of family sizes which may be difficult to measure, if not nearly impossible as in the case of organisms with high fecundity and high juvenile mortality. In such cases, estimating N_e or λ from genetic data is a possibility. One such method, which uses linkage disequilibrium data (Hill 1981; Bartley *et al.* 1992) requires a number of assumptions which limit its applicability. Existing methods that infer N_e from temporal changes in allele frequencies (Nei and Tajima 1981; Pollak 1983; Waples 1989; Waples 1990b) cannot separate multigenerational effects (fluctuating population size or age structure) from the intergenerational factors, and thus cannot estimate λ. We propose a statistical and computational approach to estimate λ (or N_e, if desired) from genetic data on population samples.

While λ should be estimated to monitor genetic risk, other population parameters are also important in conservation. To determine origins of subpopulations, admixture must be assessed. Estimation of admixture between strains and stocks has attracted much recent interest from population biologists in a wide variety of species from the hybridization of tree species (Bacilieri *et al.* 1996) or mammal breeds (Machugh *et al.* 1997) to the mixing of cultivated and wild strains of grasses (Faville *et al.* 1995) or wild and hatchery strains of salmonid stocks—the Washington Department of Fish and Wildlife is currently genetically monitoring steelhead (*Oncorhynchus mykiss*) populations for just that purpose (Steve Phelps, WDF&W, pers. comm.). Inference of migration structure is another key factor in monitoring genetic risks to a subdivided population. We will extend our computational framework for estimating λ to encompass problems of estimation of admixture and migration.

5.2 Availability of population and genetic data

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. Band-sharing methods for multilocus DNA fingerprints and RAPD's have been developed (Lynch 1988), and used to infer relationships among individuals. However, the power to infer individual relationships is slight, and, for natural populations, genetic monitoring without inferring a pedigree is an attractive option. The advent of Mendelian-inherited microsatellite markers (Tautz 1989; Wright and Bentzen 1994) has made informative genetic data increasingly available and inexpensive for such monitoring. 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 highly 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. Many 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 are being analyzed in the laboratory of Anne Kapucsinski (William Ardren, University of Minnesota, pers. comm.). In both of the above studies, population census data are available over the time periods in question, and genetic data are available in many more years than those considered by the studies.

Data, both genetic and demographic, in closely monitored or studied populations are increasingly available, but are also more complex in structure: genetic samples may be drawn from juveniles and/or reproducing adults at many points in time; fuller data on fluctuating population sizes might be available as well as detailed information on age composition of populations. Multiple multi-allelic loci, some potentially linked, and often having low-frequency alleles, are becoming widely used. Efficient and simultaneous use of all these sources of information demands explicit stochastic modeling of population-specific genetic processes. Inference on such models is analytically complex and computationally intensive.

5.3 The Monte Carlo likelihood approach

In many areas of scientific modeling, where a highly structured complex stochastic system underlies observable data, it may be impractical or even infeasible to compute a likelihood exactly. With the advent of ever faster computers, Monte Carlo likelihood (Geyer and Thompson 1992) is an increasingly valuable approach. In many of these structured systems there are latent variables, which define the dependence structure of the data. Specifically, suppose we have data random variables $\mathbf{Y}$ and a stochastic model indexed by parameters θ. Then, for any chosen latent variables $\mathbf{X}$ the likelihood may be written

$$L(\theta) \;=\; P_\theta(\mathbf{Y}) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{Y},\mathbf{X}) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X})\; P_\theta(\mathbf{X}). \tag{5.1}$$

The latent variable $\mathbf{X}$ is chosen to facilitate rapid computation of the component probabilities $P_\theta(\mathbf{Y}|\mathbf{X})$ and $P_\theta(\mathbf{X})$, either or both of which may depend on the model parameters θ. The computational difficulty then lies in the summation over the set of all possible $\mathbf{X}$ values. It is this, or an equivalent, summation which is to be effected by Monte Carlo.

Although methods for Monte Carlo evaluation of sums and integrals date back to Hammersley and Handscomb (1964), direct Monte Carlo evaluation of (5.1), for example by simulation from $P_\theta(\mathbf{X})$ is likely to be ineffective, since realizations will bear no relationship to the specific data $\mathbf{Y}$. More effective Monte Carlo estimates are obtained when latent variables $\mathbf{X}$ are realized from the conditional distribution $P_\theta(\mathbf{X}|\mathbf{Y})$. Then (5.1) may be rewritten in the form

$$\frac{L(\theta)}{L(\theta_0)} \;=\; \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} \;=\; \sum_{\mathbf{X}} \frac{P_\theta(\mathbf{Y},\mathbf{X})}{P_{\theta_0}(\mathbf{Y},\mathbf{X})} P_{\theta_0}(\mathbf{X} \mid \mathbf{Y}) \;=\; \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y},\mathbf{X})}{P_{\theta_0}(\mathbf{Y},\mathbf{X})} \middle| \mathbf{Y} \right) \tag{5.2}$$

(Thompson and Guo 1991). Thus the likelihood ratio is expressed as an expectation with respect to the distribution of $\mathbf{X}$ conditional upon $\mathbf{Y}$ under the model θ_0, and can be estimated by averaging values of the integrand $P_\theta(\mathbf{Y},\mathbf{X})/P_{\theta_0}(\mathbf{Y},\mathbf{X})$ over $\mathbf{X}$ values realized from this conditional distribution. This form has two major advantages. First, the values $\mathbf{X}$ realized from $P_{\theta_0}(\mathbf{X}|\mathbf{Y})$ will be in proportion to $P_{\theta_0}(\mathbf{Y},\mathbf{X})$, and so will be those giving high contributions to the likelihood, provided θ does not differ too far from θ_0. Second, realization at a single θ_0 provides a Monte Carlo estimate of the entire likelihood surface $L(\theta)/L(\theta_0)$, at least in the neighborhood of the model θ_0.

The disadvantage of the form (5.2) is that realizations from $P_{\theta_0}(\mathbf{X}|\mathbf{Y})$ are required. This conditional probability is $P_{\theta_0}(\mathbf{X}|\mathbf{Y}) = P_{\theta_0}(\mathbf{Y},\ \mathbf{X})/P_{\theta_0}(\mathbf{Y})$ which is proportional to $P_{\theta_0}(\mathbf{Y},\mathbf{X})$ as a function of $\mathbf{X}$. Now $P_{\theta_0}(\mathbf{Y})$ is not readily computable; if it were, Monte Carlo likelihood would be unnecessary. However, the latent variables $\mathbf{X}$ are chosen so that $P_{\theta_0}(\mathbf{Y},\mathbf{X})$ is easily evaluated, and Markov chain Monte Carlo (MCMC) is a method developed precisely for the purpose of sampling from a distribution known only up to a normalizing factor (Hastings 1970). Early forms of MCMC date to the Metropolis algorithm (Metropolis *et al.* 1953) and the Gibbs sampler (Geman and Geman 1984) and in recent years many forms of Metropolis-Hastings samplers have been developed and implemented (Gilks *et al.* 1996).

All MCMC methods are based on defining an irreducible Markov chain over the space of $\mathbf{X}$ values, the equilibrium distribution of the chain being the distribution from which realizations are desired. Thus expectations with respect to the distribution may be estimated by time-averages over realizations of the chain. Possible transitions of the chain are drawn from a *proposal distribution.* For each proposed change to $\mathbf{X}$, an *acceptance probability* is computed. If the change is accepted with this acceptance probability, $\mathbf{X}$ otherwise remaining unchanged, the equilibrium distribution of the chain will be as desired. The Gibbs sampler is a special case of a Metropolis-Hastings sampler in which the proposal distribution resamples components of $\mathbf{X}$ from their conditional distribution given $\mathbf{Y}$ and the remaining components: in this case the acceptance probability is always one. Components may be updated singly, but it is often more efficient to update several components of $\mathbf{X}$ jointly, where this is computationally feasible (Smith and Roberts 1993). Recently, Metropolis-Hastings samplers have been generalized to include reversible-jump MCMC samplers (Green 1995). This enables sampling of more complex structured spaces, where the dimension of the space of $\mathbf{X}$ values may vary within a single run of the chain.

MCMC methods have been used extensively in genetic analysis, particularly in the analysis of data on extended or complex pedigree structures. In segregation and linkage analysis of trait and genetic marker data observed on members of a known pedigree, the latent variables may be genotypes (Guo and Thompson 1992) or meiosis indicators (Thompson 1994) or both (Lange and Matthysse 1989). While too large a space of latent variables is undesirable, sometimes the addition of more variables makes it possible to develop better mixing MCMC methods. Single-site updating methods, in which a single component of $\mathbf{X}$ is proposed to be altered, are generally very slowly mixing. Long runs are then required to ensure both adequate convergence to the equilibrium distribution and precise Monte Carlo estimates. (If successive realizations are very highly correlated, large Monte Carlo sample sizes are needed to reduce the Monte Carlo standard error.) Methods to improve the mixing of samplers have been developed; 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). Recently, to improve mixing and ensure irreducibility, there has been a focus on developing methods in which multiple components of $\mathbf{X}$ 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 1997), and a whole-locus Gibbs sampler (Heath 1997). New MCMC methods to analyze more complex model spaces are being used in genetic analysis: reversible-jump MCMC (Green 1995) has been used in 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 (Heath 1997).

MCMC methods have 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 have 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 $\mathbf{X}$ 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) have proposed 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 have also been used 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). Again the realized latent variable is the ancestral coalescent, in these cases specified by the sequence of ancestral recombination, mutation, or coalescent events.

While MCMC methods have been used on pedigrees and on coalescents, there seems previously to have been no proposal to use them on the intervening population time-scale, where data consist of allele frequencies in samples of individuals from specified populations. The samples may be from the same population at different time points, or from different populations. A population may be subdivided, with some migration or admixture process among the subpopulations. A population may also be age-structured. In the data samples, there may or may not be information on the ages or subcomponent origins of the sampled individuals. The primary parameter of interest may often be effective population size N_e or the parameter λ (section C.2.1), although migration and admixture structure may also be of interest. This structure is ideally suited to Monte Carlo likelihood with a latent-variable framework (equation (5.2)); the latent variables are the allele frequencies in components of the population. MCMC is ideally suited to the sampling of these latent variables, since the population processes underlying changes in allele frequency can be simply specified in terms of the model parameters, and the probability of data samples conditional upon underlying allele frequencies are likewise easily specified.

Thus our objective is to develop methods with which to realize allele proportions or counts, conditional upon data from samples, in order to provide Monte Carlo likelihoods for genetically relevant population parameters. To be specific, let us consider the simplest possible case of a discrete generation population, with samples taken at two successive generations and typed for a single diallelic locus. For simplicity we consider binomial sampling (with replacement) from the two populations, although hypergeometric (without replacement) sampling could also be implemented. Suppose that the frequency (proportion) of one of the two alleles is x_1 in the first generation, and x_2 in the second. Suppose the samples are of size n_1 and n_2, and the frequencies observed in the samples are y_1 and y_2. Suppose that the (diploid) variance effective population size over this one generation of drift is N_e, and that N_e is small enough that standard diffusion approximations would be inappropriate. The adult census size N is assumed known, and in this case $\lambda = N_e/N$. The likelihood for $\theta = (x_1, \lambda)$ is

$$L(\theta) \;=\; P_\theta(Y_1 = y_1,\ Y_2 = y_2) \;=\; \sum_{x_2} P_\theta(Y_1 = y_1,\ Y_2 = y_2 \mid X_2 = x_2) P_\theta(X_2 = x_2) \qquad (5.3)$$

which involves summation over all possible values of x_2. However, the conditional probabilities are easily computed:

$$(Y_1|x_1) \;\sim\; n_1^{-1} B(n_1, x_1) \quad (Y_2|X_2 = x_2) \;\sim\; n_2^{-1} B(n_2, x_2), \quad (X_2|\lambda, x_1) \;\sim\; (2N_e)^{-1} B(2N_e, x_1), \qquad (5.4)$$

where $B(n, p)$ indicates a binomial probability distribution with index n and parameter p. Thus, Monte Carlo evaluation of likelihood ratios using equation (5.2) is straightforward. In this basic formulation we have treated the initial allele frequency x_1 as a parameter to be estimated jointly with λ. This is undesirable, particularly when there are data on many alleles at many loci, each with an initial frequency to be estimated. Ways to avoid this are discussed in section C.2.6.

Of course, in practice, the situation will be far more complex than this simple example. The data $\mathbf{Y}$ may be of marker phenotypes rather than genotypes, so sample allele proportions may not be observable. The markers will have multiple alleles, there will be data at multiple genetic marker loci, and some of these loci may be genetically linked. We may have data at multiple time points, and may wish to consider the evolution of the population over multiple generations. The generations may be overlapping, and the population age-structured. The population may have several components, or be an admixture of several ancestral stocks. Despite these complications, the same principle applies. We consider specific approaches to addressing more complex models later in the proposal, but first review current methods for the estimation of population parameters from the genetic characteristics of population samples.

5.4 Current methods for inference from allele frequencies

Estimating N_e by temporal methods. Current methods for estimating N_e from temporal changes in allele frequencies are either direct applications or extensions of the F-statistic based approaches of Nei and Tajima (1981) and Pollak (1983). F is the standardized allele frequency variance after t generations of genetic drift: $F = (x_t - x)^2/x(1-x)$, where x is the initial population allele frequency and x_t is the frequency at time t. Such methods are not well-suited to the data now available. They do not deal well with samples from multiple time points; Pollak's method does so only by assuming N_e is constant. Additionally they are based, implicitly or explicitly, on diffusion approximations for genetic drift—approximations which are unreliable for alleles at low frequencies in small populations. Anderson (1998) compares diffusion approximations with exact t-generation transition probabilities of allele counts, demonstrating a significant discrepancy when the probability of allele fixation is non-negligible. Felsenstein (1985) documents similar problems with such diffusion approximations. Waples (1989) shows that F-based methods exhibit a bias in the estimation of F when there are alleles in low frequencies, resulting in a bias in the estimation of N_e. Eliminating this bias currently requires grouping rare alleles, thus losing information.

Waples (1990b) and Jorde and Ryman (1995) extend the above methods to organisms with more complex life histories. By computer simulations Waples (1990b) derives an empirical relationship between F and the effective number of spawners per year, N_b, in a semelparous salmon population with three ages of maturation. He estimates N_b from two samples separated in time by one or more years. He also proposes a method of averaging F values over more time intervals to accommodate data from several time points, but reports added imprecision when using samples from many different time points. His basic model assumes that N_b is constant over time, and he points out via computer simulations that the method is less reliable when population size fluctuates. When data for the number of returning adults, N, are available, and N varies over time, a problematic feature of this approach is that it is not straightforward to estimate $\lambda = N_b/N$. The method estimates an average number of effective spawners over the time interval in question, but does not relate this quantity clearly to the observed population sizes. With this method, the effects on estimated effective population size of intergenerational (breeding) factors cannot be separated from the effects of changing population size.

Jorde and Ryman (1995) present a method for estimating effective size in populations with overlapping generations, and they use it estimate the effective size of four iteroparous, brown trout (*Salmo trutta*) populations (Jorde and Ryman 1996). Their method is also F-statistic based. They assume that the population in question is of constant size with constant survivorship and mortality so that they can consider the allele frequency changes that would be observed once the population's gene frequency variance has settled into an asymptotic pattern of proportional increase each year (Hill 1979; Charlesworth 1980). These assumptions are unlikely to hold when population size, and hence also age structure. varies from year to year. For example, the steelhead population of the Keogh River, has varied about 200 reproducing adults to over 2,800 in ten years of record (Ward and Slaney 1988).

Inference of admixture. In a typical admixture problem, snapshot data of present-day gene frequencies from parental populations and the admixed one are used to infer the proportion that each parental population contributed at the time of admixture some known time in the past. Accordingly, the inference must take account of error due both to drift and to sampling. Thompson (1973) presented the first method to include both drift and sampling error. Long (1991) developed an iteratively reweighted least squares method incorporating drift and sampling for use with codominant and recessive marker alleles. Both of these methods adopt a diffusion approximation for genetic drift and have been used primarily in the context of human populations, since only for these, until recently, have there been adequate population samples. With small populations of endangered species, the effect of random genetic drift may far outweigh sampling variation. Particularly with highly polymorphic markers, inference then requires use of exact genetic drift transition probabilities such as may be effected by Monte Carlo methods.

The basic framework for inference of migration structure is similar to that of admixture; indeed, migration is continual admixture. Again, the available data have led to methods being developed in the context of human population studies, with earlier studies using distance methods based on diffusion approximations to infer, for example, patterns of migration among Italian villages (Cavalli-Sforza 1969). More recent studies have used likelihoods derived from migration matrix models and other multiallelic admixture models to estimate patterns of migration and admixture among Faroese Islands (Thompson 1984) or Amerindian villages (Long and Smouse 1983). With the advent of more polymorphic markers, there has been consideration of the migration information contained in patterns of presence or absence of rare alleles (Thompson *et al.* 1992). In these cases diffusion approximations are inadequate, and other methods are needed. Slatkin (1995) has developed a more general framework for inference of migration structure from data on polymorphic microsatellite loci. However, there is currently no method for computing likelihoods for, and hence testing, specific migration hypotheses.

5.5 Population sampling and relationship to life-history

Some components of our methods must necessarily reflect details of the life history of the species under study. Salmonid populations are of great conservation interest (Waples 1995; Nehlsen *et al.* 1991), and they present a range of life-history scenarios which may be addressed within a relatively coherent framework. Pink salmon (*Oncorhynchus gorbuscha*) have the simplest life-history. They are semelparous and mature exclusively at two years. Thus the odd-year and even-year populations follow a rigid, discrete non-overlapping generation model. Indeed, the two-year-old maturation age is so rigid that, throughout the species' natural range, odd- and even- year populations can be distinguished genetically (Zhivotovskii *et al.* 1989). Modeling the life history of this species is straightforward, but must accommodate population fluctuations that may be extreme (Heard 1991).

Sockeye salmon (*O. nerka*) offer another life history—this species is semelparous, but adults returning to spawn may be three, four, or five years old, providing the additional complication of an age-structured adult population, and contributions from three years of juveniles to each returning adult year class. Chinook salmon (*O. tshawytscha*) are similarly semelparous, with a returning population composed of several year classes. The maturation ages are different, however; fish mature at two to eight years (Healey 1991). More complex again are steelhead, (*O. mykiss*). These fish may remain in fresh water for several years before going to the ocean, and they may reside in the ocean and return to spawn several times. Since they are iteroparous, they have true, overlapping generations, and individuals could be sampled at several stages in their life span (Withler 1966).

Our initial modeling efforts will be on a semelparous salmonid, with a variable spawning age, such as *O. nerka* or *O. tshawytscha*. We shall assume we have an annual estimate of of the number of returning adults, N, which will enable us to separate the effects of fluctuating population size from other aspects of the breeding structure of the population. We aim to estimate the ratio $\lambda = N_b/N$ where N_b is the effective number of reproducing individuals in a given year (Waples and Teel 1990; Waples 1990b). To accommodate the variable age at reproduction we shall model the allele frequencies of the separate year classes contributing to a given adult pool of potential breeders.

We will assume genetic data are available from population samples. The potential to sample non-destructively either adults or juveniles raises several interesting modeling and computational issues. Nei and Tajima (1981) distinguish between two different sampling schemes: one in which adults are sampled non-destructively, and the other where samples are from the huge pool juveniles produced by these adults. In the latter case, sample size may be larger than effective population size. If returning adults are sampled, we have a direct estimate of the genetic characteristics of the breeding population. Estimates of the proportions in each age class may also be available. If only juveniles are sampled, we can consider variation only between offspring gamete pools. Inferences will be dependent on our ability to make life-history assumptions, for example about the age-composition of the parent population. Where both adults and their offspring are sampled, there is the potential to separate components of drift. In principle, at least, it would be possible to test for the differential contributions of adults in the different age classes. Additionally the contribution to drift from reproductive and freshwater life stage factors could be separated from the effects of differential ocean survival. By casting the inference problem in an explicit probability framework and computing likelihoods by MCMC, we will have the flexibility to treat any sampling scheme or a mixture of several.

We will assume the availability of samples taken at multiple time points, although not necessarily every year. In estimating admixture proportions we will pursue inference from both single-time-point data as well as data from several points in time. Rannala and Hartigan (1996) provide a pseudo-likelihood approach to the problem of estimating, from single-time-point data, the average number of migrants per generation among subpopulations: An MCMC framework could provide a likelihood estimation procedure. However, our primary focus will be on data taken at multiple time points, and on estimating migration as a process of ongoing admixture.

After treating the semelparous salmon life histories, we will consider the iteroparous life history of *O. mykiss* to begin extending our methods to organisms with truly overlapping generations. This will also allow us to estimate λ from the excellent datasets being compiled for two populations of this species (William Ardren pers. comm.), and to infer and estimate admixture from hatchery populations of *O. mykiss* to adjacent wild populations. Our modeling efforts for *O. mykiss* will be an important step in making these methods general to a broad class of iteroparous species of conservation interest.

5.6 MCMC approaches to multiple parameters and missing data

The aim of our proposed MCMC algorithms is to realize underlying population allele counts or frequencies, in order to provide likelihoods for genetically relevant population parameters, such as N_e or λ, and to estimate admixture proportions, and migration structure. In this section, we provide some preliminary details of our proposed methods; implementing and then refining these methods, to improve the statistical and Monte Carlo efficiencies of procedures, is the major component of the proposed research.

In section C.2.3 we gave the equations (5.3) and (5.4) for sampling the latent frequencies at a single generation for a single allele. Considering first a discrete-generation population, extension to samples at multiple time-points is straightforward. Each sample Y is a binomial proportion, with index the sample size n, and parameter the frequency of the allele X in the population from which the sample is taken (equation (5.4)). Each generation t, conditionally upon the previous one, has a population frequency X_t which is a binomial proportion, with index $2N_e(t)$ and parameter X_{t-1}, where $N_e(t)$ is the diploid effective population size over that generation (equation (5.4)). Conditional on the population frequencies $\mathbf{X}$, the data samples Y are independent, each depending only on the X-value in the population from which it is taken. The distribution of any given X_t, conditional on the other X-values and on the samples $\mathbf{Y}$, depends only on X_{t-1}, X_{t+1}, and, if a sample is taken at generation t, on the sample proportion Y_t. For example, equation (5.3) may be rewritten

$$\begin{aligned} L(x_1, \lambda) &= P_{(x_1,\lambda)}(Y_1 = y_1,\ Y_2 = y_2) \\ &= \sum_{x_2} P_\lambda(Y_1 = y_1,\ Y_2 = y_2 \mid X_2 = x_2, X_1 = x_1) P_\lambda(X_2 = x_2 \mid X_1 = x_1) \\ &= \sum_{x_2} P_\lambda(Y_1 = y_1 | X_1 = x_1) P_\lambda(\ Y_2 = y_2 \mid X_2 = x_2) P_\lambda(X_2 = x_2 \mid X_1 = x_1) \qquad (5.5) \end{aligned}$$

This conditional independence structure of hidden-Markov form makes implementation of a Metropolis-Hastings or Gibbs sampler for $\mathbf{X}$ straightforward. Note there need not be data from every generation; whatever data there are can be fully utilized without additional complexity.

Clearly, to have sufficient statistical power for useful inferences, data from a large number of polymorphic loci are required. While, for unlinked loci with negligible linkage disequilibrium among them, latent population allele frequencies can be realized independently, loci with multiple alleles cause potential difficulties. In place of binomial samples, we have multinomial samples. For binomial samples the probabilities can be evaluated rapidly even for populations or samples of several hundreds or even thousands, by pre-computing and storing the binomial coefficients. For multinomial samples, the number of such coefficients makes this impractical. One solution is to sample the allele counts at a locus sequentially over the alleles; each of the first $k-1$ allele counts at a k-allele locus is then again a binomial realization. Computational efficiency is improved by sampling first the larger allele classes.

In equation (5.3) in C.2.3, we treated the initial population allele count as a parameter to be estimated jointly with the parameter λ. While this is acceptable when data on only a few loci and alleles are to be estimated, it is undesirable to estimate an initial population frequency for every allele. The number of parameters increases with the number of alleles and loci, and as more and more data are included in the likelihood, the maximum likelihood estimator of a parameter such as λ is inconsistent (Neyman and Scott 1948). The situation is akin to that of estimating a rooted evolutionary tree (Thompson 1975), or to ML estimation of variance components in quantitative genetics. In these cases, it is usual to, in effect, integrate out the high-dimensional nuisance parameter, by considering an unrooted evolutionary tree (Felsenstein 1973), or by adopting REML methods in quantitative genetics (Henderson 1986). We propose to do likewise, considering an integrated likelihood, placing a diffuse prior distribution (Box and Tiao 1973) on initial population allele frequencies at each locus. For example, Painter (1997) puts Dirichlet priors on allele frequencies at polymorphic microsatellite loci. Population data from other studies can also be used to provide prior distributions. Waples (1990a) gives the empirical distribution of population allele frequencies over a large number of populations and a large number of allozyme loci. This empirical distribution can be very closely fitted by a Dirichlet distribution with appropriately chosen parameter values. In the simplest case, for example, equation (5.5) becomes

$$\begin{aligned} L(\lambda) &= P_\lambda(Y_1 = y_1,\ Y_2 = y_2) \\ &= \sum_{x_1,x_2} P_\lambda(Y_1 = y_1 | X_1 = x_1) P_\lambda(Y_2 = y_2 | X_2 = x_2) P_\lambda(X_2 = x_2 | X_1 = x_1) \pi(X_1 = x_1) (5.6) \end{aligned}$$

where $\pi()$ is the assumed prior on the initial frequency x_1. MCMC methods, in which latent variables are sampled, makes staightforward the consideration of an integrated likelihood, where nuisance parameters are given some prior distribution. Indeed, once the relevant priors are incorporated, there is no need to distinguish between a latent variable such as x_2 and a parameter such as x_1. For these reasons, such integrated likelihoods are increasingly used for inference in multiparameter structured stochastic systems, including those in genetic analysis (HEATH 1997).

For a semelparous salmonid, with variable age at maturity, the generations are no longer discrete. We may suppose that the age composition of a returning adult population is known, or can be estimated from the samples of returning adults. In this case, the latent variables $\mathbf{X}$ should specify population allele frequencies at each locus, and in each returning year class. With this partitioning of the population, we have more latent variables to sample, but the conditional independence structure is maintained. The allele frequency $X_{(t,a)}$ in a given age class, age a at year t, conditional upon all other components of $\mathbf{X}$ and on the data $\mathbf{Y}$, will depend only on the data samples $\mathbf{Y}$ at year t, on the allele frequency in the parents of that class (the spawners at year $t-a$), and the allele frequencies in the offspring age class a_j returning at year $t+a_j$. The adults sampled, in any given year, may or may not be age-classified. That is, we may have samples $Y_{(t,a)}$, each a Binomial sample proportion with index the sample size $n_{(t,a)}$ and parameter $X_{(t,a)}$, or we may simply have a sample allele count Y_t with a separate estimate of the age composition of the population, say a fraction q_a of age a. In this latter case, Y_t is a Binomial proportion with index the sample size n_t, and parameter $\sum_a q_a X_{(t,a)}$. In either case, $X_{(t,a)}$ may be sampled from its conditional distribution; a Gibbs sampler can be implemented. The variables q_a may be treated also as latent variables of the MCMC, and sampled accordingly.

A major advantage of the MCMC framework is that partial information can be incorporated easily. For example there may be annual census counts and age composition data, but only intermittent observation of genetic marker data. Or, in some years, there may be age-specific samples, but in other years only a general estimate of age composition of the spawners. Some loci may not be typed in all years for which there are genetic data. The Monte Carlo framework enables us to sample missing observations from the appropriate conditional distributions, so that all data observations are fully utilized in making estimates and inferences. In the above development, we have presented the sampling as a single-site updating MCMC scheme, in which each latent component $X_{(t,a)}$ for each of $k-1$ alleles at each k-allele locus is updated in turn. However, in some cases it is possible to update several components of $\mathbf{X}$ jointly. We propose to investigate the feasibility and practicality of joint updating schemes, to improve computational efficiency.

For ease of notation, we revert to the discrete-generation case to discuss estimation of population parameters, although the same general formulae apply to the more general age-class case. We also discuss estimation of the single parameter λ. We assume λ remains constant over the years, although the number of breeding adults may fluctuate widely. Note that this assumption is primarily for notational convenience: with sufficient genetic data, at multiple time points, it would be possible to test for variation in λ due to changing ecological or demographic conditions. Also, the same general formulation would apply if we were instead to estimate N_e directly, or assume some other parametrized relationship between N_e and N.

The integrated likelihood (5.6) is of the same general latent-variable form given in section C.2.3 (equation (5.1)). More generally, with data at multiple time-points $t_1, \ldots, t_r$, with $1 = t_1 < t_2 < \ldots < t_r = T$:

$$L(\lambda) \quad = \quad P_\lambda(Y_{t_1}, \ldots, Y_{t_r}) \quad = \quad \sum_{\mathbf{X}} P_\lambda(\mathbf{Y}, \mathbf{X})$$

$$= \sum_{x_1,\ldots,x_T} \left[\left(\prod_{j=1}^{r} P_\lambda(Y_{t_j}|X_{t_j}) \right) \left(\prod_{t=2}^{T} P_\lambda(X_t|X_{t-1}) \right) \pi(X_1) \right] \tag{5.7}$$

Note that the joint probability $P_\lambda(\mathbf{Y},\mathbf{X})$ is easily computed; the difficulty in exact likelihood evaluation is in the summation of $\mathbf{X}$-values. In obtaining Monte Carlo likelihood ratios, we shall follow the formulation of Thompson and Guo (1991) given in equation (5.2). At a given parameter value λ_0, we shall realize Monte Carlo samples of $\mathbf{X}$ from the conditional distribution $P_{\lambda_0}(\mathbf{X} \mid \mathbf{Y})$. Hence, we obtain Monte Carlo estimates of the likelihood ratio $L(\lambda)/L(\lambda_0)$ by averaging the "complete-data" likelihood ratio $P_\lambda(\mathbf{Y},\mathbf{X})/P_{\lambda_0}(\mathbf{Y},\mathbf{X})$ over the realized $\mathbf{X}$-values (equation (5.2)). From the Monte-Carlo estimate of the likelihood surface, an estimate of λ may be obtained, possibly after interating the process to obtain a simulation value λ_0 which is close to the maximum likelihood estimate (Geyer and Thompson 1992; Kuhner *et al.* 1995). In addition to investigating the performance and properties of this Monte-Carlo likelihood approach, for more complex models we will also consider a fully Bayesian approach, incorporating prior distributions on all parameters, and sampling also over values of λ. For simple models with a single parameter, we prefer the Monte Carlo likelihood approach, but for more complex models a fully Bayesian MCMC may provide clearer inferences (Heath 1997). While each approach has been used in several areas of genetic analysis, there has been no direct comparison. For the case of estimation of λ under a drift model, we will compare the two approaches.

Hypotheses of admixture and migration tend to be specific to the populations under study, but several classes of such problems can be formulated. One is where the admixture occurs in an initial founding event, and samples are taken at some later time point, both from the populations dscending from the unmixed contributors to the founding gene pool, and from the mixed population. If populations are small, or a substantial number of generations have elapsed, all three populations will have undergone genetic drift. One example is the Icelandic population, founded approximately 40 generations ago by a mix of Celtic and Norse populations (Thompson 1973). Another is of wild salmonid populations of the Lake Washington drainage, founded about 14 generations ago by stocks from several other lakes of the Pacific Northwest (Anderson 1998). In both these examples, genetic drift is a significant factor. Each was analyzed using diffusion approximations, but for highly polymorphic marker loci, where not all alleles are present in the samples from each population, this is not ideal. The structure of the likelihood is similar to that above, although now there are two initial populations (and the mixture), and data are taken only at a single final timepoint T. We assume, as is the case in these two examples, that there are good estimates of adult census size over the relevant period of history, and assume that the three populations share the same value of λ. Let $X_t^{(1)}$, $X_t^{(2)}$, and $X_t^{(3)}$ denote the allele frequencies at generation t, in the two contributing populations and in the admixed one, respectively. Of course, data for multiple alleles and loci will be used. If the mixing at generation $t = 1$ is in proportions μ and $(1 - \mu)$, then $X_1^{(3)} = \mu X_1^{(1)} + (1 - \mu) X_1^{(2)}$ and, analogously to equation (5.7), given data $\mathbf{Y} = (Y_T^{(j)}; j = 1, 2, 3)$ at time T on all three populations, we have

$$\begin{aligned} L(\lambda,\mu) &= P_\lambda(Y_T^{(1)}, Y_T^{(2)}, Y_T^{(3)}) = \sum_{\mathbf{X}} P_{(\lambda,\mu)}(\mathbf{Y},\mathbf{X}) \\ &= \sum_{\mathbf{X}} \left[\left(\prod_{j=1}^{3} P_\lambda(Y_T^{(j)}|X_T^{(j)}) \right) \left(\prod_{t=2}^{T} \prod_{j=1}^{3} P_\lambda(X_t^{(j)}|X_{t-1}^{(j)}) \right) \pi(X_1^{(1)})\pi(X_1^{(2)}) \right] \end{aligned} \tag{5.8}$$

where now the sum is over all the allele frequencies in all three populations over the T generations. Note that μ enters only through the initial value $X_1^{(3)} = \mu X_1^{(1)} + (1-\mu) X_1^{(2)}$ in the mixed population.

Since we have a likelihood with the same conditional independence structure as before, the latent allele frequencies $\mathbf{X}$ may again be sampled conditionally on the data $\mathbf{Y}$, at any given parameter values (λ_0, μ_0), to give a Monte Carlo estimate of the relative joint likelihood for λ and μ, using equation (5.2):

$$\frac{L(\lambda,\mu)}{L(\lambda_0,\mu_0)} \;=\; \sum_{\mathbf{X}} \frac{P_{(\lambda,\mu)}(\mathbf{Y},\mathbf{X})}{P_{(\lambda_0,\mu_0)}(\mathbf{Y},\mathbf{X})} P_{(\lambda_0,\mu_0)}(\mathbf{X} \mid \mathbf{Y}) \;=\; \mathrm{E}_{(\lambda_0,\mu_0)}\left(\frac{P_{(\lambda,\mu)}(\mathbf{Y},\mathbf{X})}{P_{(\lambda_0,\mu_0)}(\mathbf{Y},\mathbf{X})} \middle| \mathbf{Y} \right) \tag{5.9}$$

We will investigate the properties of this likelihood for joint estimation of λ and μ. Alternatively, there may be separate data on which to base population-specific values of λ, providing more statistical precision for the estimation of μ. We will investigate the loss of information about μ incurred by lack of knowledge of λ.

Generally, there is little information in "snapshot data" taken at a single time point, although with data on a very large number of alleles at a very large number of genetically neutral loci it should be possible to estimate μ and λ quite precisely. We will also extend the above formulation to the case where data are available from the three populations at multiple time points. We will also develop tests of population origins, based on these admixture likelihoods. In this case, the hypotheses are of variable dimension, depending on whether a population did, or did not, originate as a mixture of other populations. To encompass these hypotheses within a single MCMC framework, reversible-jump MCMC can be used (Green 1995; Heath 1997).

So far we have considered only genetic drift, admixture, and migration, as the factors influencing population allele frequencies. Although mutation rates are non-negligible for some microsatellite loci, mutation will have little impact on population allele frequencies over the sort of time period of tens of generations relevant to small natural populations. If desired, mutation can be incorporated into the MCMC sampling process for latent allele frequencies. A more significant factor may be selection. Although it may be necessary to assume neutral markers to be able to make clear inferences about population structure (λ) or origins (μ), it should be possible to detect whether significant selection is present. Whereas effects of population history are common to all loci, selection affects loci differentially, and this, in principle, enables tests for selection to be developed (Lewontin and Krakauer 1973). Within our framework, selection could be incorporated into the distribution of allele frequencies given those at the preceeding generation, and hence into the MCMC sampling of population allele frequencies. To avoid a large number of selection parameters, which would again lead to inconsistent estimators (Neyman and Scott 1948), a prior on selection coefficients could be assumed. Time permitting, we will investigate the feasibility of this approach.

We have here presented our objective in terms of estimates to be derived from Monte Carlo likelihoods. However, assessment of the precision of estimation is no less important. In any MCMC analysis there are two sources of variance, or standard error. The first is the statistical standard error, due to variance of data random variables under a model. This may be assessed by curvature of the estimated likelihood surface, or by a parametric bootstrap procedure. Although the latter is computationally intensive, it is always feasible, since the Monte Carlo procedure provides a method for simulating under any specified parameter values. The second source of variance is the Monte Carlo standard error, which must also be assessed. Geyer (1996) has proposed several methods for estimation of the Monte Carlo standard error in MCMC procedures, and we propose to adopt these methods.

MCMC is a very powerful and flexible approach for addressing inferences from data taken over the course of history. Provided the population process is well-specified, it can be simulated. Moreover, the process of genetic inheritance defines a conditional independence structure: individuals receive genes from their parents and segregate them to their offspring. Provided a process can be simulated, it is always theoretically possible to implement an MCMC algorithm to realize latent variables conditionally upon observed data. The inheritance pattern of genes within a population makes the MCMC realization of latent allele frequencies conditionally upon observed sample allele frequencies not only theoretically possible, but also practically feasible, enabling us to develop practical computational methods for population genetics inference.

Chapter 6

Direct Estimation of the Ratio of Effective Number of Breeders to the Census Number from Temporal Changes in Allele Frequencies: A Collection of Working Notes on a Maximum Likelihood Approach and a Simulation Study

These working notes outline the rationale for a maximum likelihood approach to estimating λ, the ratio of effective breeding adults to the census number of breeding adults in a population. It provides background on F-statistic approaches to the same problem. Then it describes how to implement a Baum (1972)-type algorithm to efficiently compute the likelihood for a diallelic locus. It then describes extensive simulations comparing the moment-based estimators to the maximum likelihood estimator for the case of data on diallelic loci. This reads a bit like a manuscript that never got sent anywhere—which it is!

6.1 Introduction

Starting with Krimbas and Tsakas (1971), researchers have used temporal changes in observed allele frequencies to estimate the effective size N_e of a population. Until recently, all such efforts have used moment estimators derived from Wright's standardized allele frequency variance, $F = (p_t - p_0)^2/[p_0(1-p_0)]$, where p_0 is the initial population allele frequency and p_t is the population allele frequency at time t. These F-based methods have performed adequately with data at only a few temporally-spaced samples, however current biotechnology allows for obtaining long time series of allele-frequency data either by non-invasive sampling or by the amplification of DNA from archived tissue samples collected previously from populations. Among other examples, microsatellite 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. Scale samples, historically collected from fish populations to determine the age of individuals, are a valuable source of archived tissue for modern genetic studies. Miller and Kapuscinski (1997) isolated DNA from northern pike (*Esox lucius*) scales from a small lake in Wisconsin and used those samples to estimate N_e for the population and compare that to estimates of the number of adult fish in the lake. Currently a similar project is underway with steelhead (*Oncorhynchus mykiss*) on the West Coast in two populations with excellent spawner abundance data for over 40 years (William Ardren, Univ. of Minn., pers. comm.).

There is a growing amount of allele frequency data from these sorts of closely-monitored populations, however, the traditional F-based methods of estimating N_e are not particularly well-suited for these rich datasets. Specifically, there are no formally-derived methods, which work well, for many samples in time, and there is no facility for incorporating data on the census size of the population at points in time between the samples. We propose a maximum likelihood procedure which is particularly advantageous when data are available at multiple points (generations) in time from samples of varying size, and which also provides a natural way to incorporate population census or abundance estimates. A recent maximum likelihood estimator (Williamson and Slatkin in press) has demonstrated the utility of the maximum likelihood approach for estimating N_e. Our method differs in that it explicitly uses estimates of population abundance in each breeding season to make a direct estimate of the ratio of the effective number of reproducing adults (N_r) to the observed number of reproducing adults (C).

This ratio, which we call $\lambda = N_r/C$ is a quantity which separates the effects of intergenerational population-genetic sampling from those of fluctuating population size on the effective size of a population. It is useful to conservation biologists faced with the genetic management of populations whose annual census size is easily estimated, but whose effective size is more difficult to infer by demographic methods. Knowing λ and the annual population abundance, a biologist could make informed predictions of the expected loss of genetic variability in future years. In this regard, especially in the context of overlapping generations or year classes, λ is distinct from the ratio N_e/N—the ratio of the overall effective size of the population to some average of the census size over some period. As Frankham (1995a) notes in a review of N_e/N ratios from 102 species, the ratio N_e/N fails to capture much information in the context of overlapping generations and fluctuating population size.

Throughout this treatment we will assume that λ remains constant during the time of genetic sampling. It is easy to suspect that this assumption would not hold; for example, λ could be expected to vary under different environmental or ecological conditions, or under different numbers of breeding adults, N_r. Fortunately, estimating λ by maximum likelihood takes the first step toward testing hypotheses about λ using likelihood ratio tests. The model presented here may then be seen as a null hypothesis (*i.e.*, "λ is constant") which could be rejected by comparing it to models in which λ is allowed to vary. Future work will explore this.

Except in organisms with the simplest life histories, existing methods for estimating effective size do not allow estimation of λ. Our goal in this paper is to present a maximum likelihood method for estimating λ in such simple cases where the exact likelihood may be computated easily enough to allow sufficient replicates in a simulation study to compare our method with the F-based methods. Thus we restrict our attention, here, to multiple samples in time of diallelic marker loci from a very small population with discrete, non-overlapping generations. In a future paper we will discuss Monte Carlo likelihood methods (Thompson and Guo 1991; Geyer and Thompson 1992) to extend this approach to larger populations, multiallelic markers, and organisms with overlapping year classes. Below we introduce the scenario of the population and data we are dealing with. We review the F-based methods that can be used for this sort of estimation, then we develop the maximum likelihood approach. We compare the performance of different estimators by computer simulation.

6.2 The Population and Sampling Scenario

We assume a population of diploid organisms having discrete, non-overlapping generations. Each generation t there are C_t adults which produce offspring that become C_{t+1} adults in the next generation (Figure 6.1). The population does not reproduce as a Wright-Fisher population, but the effective number of reproducers in any generation t is given by $\lfloor \lambda C_t \rfloor$ where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. We observe census sizes C_t each generation ($t = 0, 1, \ldots, T$) and we observe allele frequencies y_{tj} by sampling with replacement from the adults at some (not necessarily all) times t at different loci indexed by $j = 1, \ldots, J$. The loci all have two codominant alleles, and the sample size for locus j in generation t is S_{tj}. We assume that migration, selection, and mutation are unimportant.

This type of sampling scheme, called sampling plan II by Waples (1989), applies to organisms with high fecundity which can be randomly sampled as gametes or juveniles. It is probabilistically identical to the sampling scheme shown in Figure 6.2, which may be interpreted as follows: at time $t = 0$ each of the C_0 adults produces an equal and infinite number of gametes according to its genotype. From this infinite pool of gametes (or juveniles, if we wish to call them that) we draw our genetic sample which gives an estimate of the population allele frequency at $t = 0$ (note that this is the same as sampling the C_0 adults with replacement). Then the intergenerational sampling occurs to form the C_1 adults in the next generation. The details of this sampling are unknown, but it is not simple binomial sampling; if it were, then N_{r_1} the effective number of reproducers at $t = 1$ would be C_1 and we would have no further work to do. Instead, the intergenerational sampling from $t = 0$ to $t = 1$ results in an allele frequency distribution at $t = 1$ commensurate with a single generation of genetic drift in a Wright-Fisher population of size N_{r_1}. (So, it is Wright-Fisher reproduction of a population of size N_{r_1}.

Our objective is to use the observed census sizes and the observed allele frequencies to estimate λ, the ratio of N_r/C, by maximum likelihood. We start with the assumption that C_t, the number of reproducing adults, is known without error for each t, but we will relax this assumption in a later paper, when we will also consider different sampling plans.

6.3 F-statistic Approaches

F-statistic methods for estimating N_e are based on classical theory of the increase in allele frequency variance due to genetic drift. We review the derivation of F-based estimators for N_e in the case of two temporally spaced samples. Then we consider methods for combining the information from multiple (> 2) temporally spaced samples. We review the method due to Pollak (1983), and another method suggested in an overlapping-generations context by Waples (1990b). These methods typically estimate the harmonic mean effective sizes of the population for the generations between samples. Thus, in populations of simple life-history structure it is possible to obtain an estimate of λ given the census sizes and the estimate of N_e. We compare the above two methods to our maximum likelihood estimator by computer simulation later in the paper.

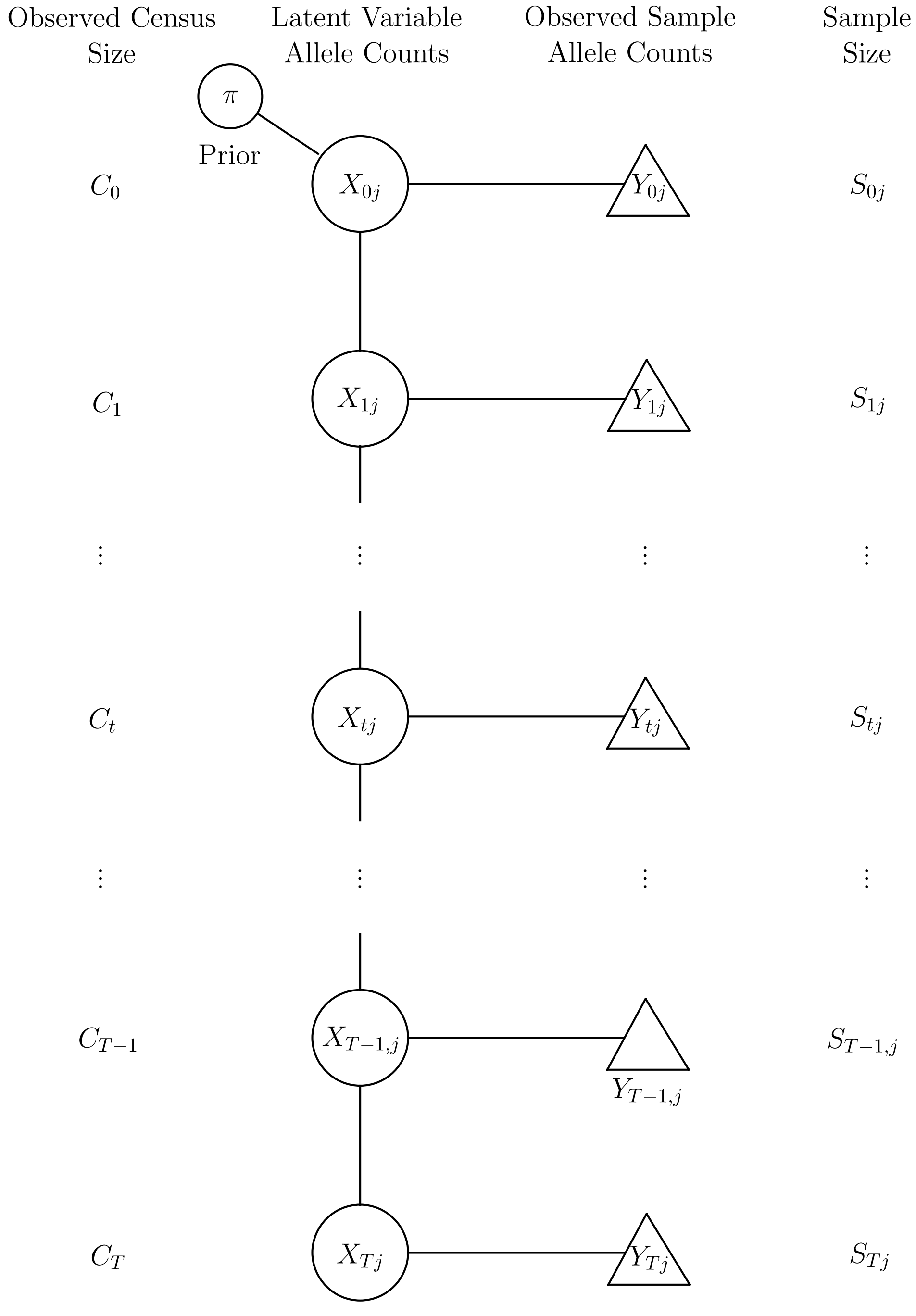


Figure 6.1: A graph showing the relationships between random variables at the j^{th} locus and associated quantities in the probability model. The circles denote populations, the triangles denote samples from those populations. Time proceeds from top to bottom in the figure. Note that there may be no samples in some years; S_{tj} in such years is zero.

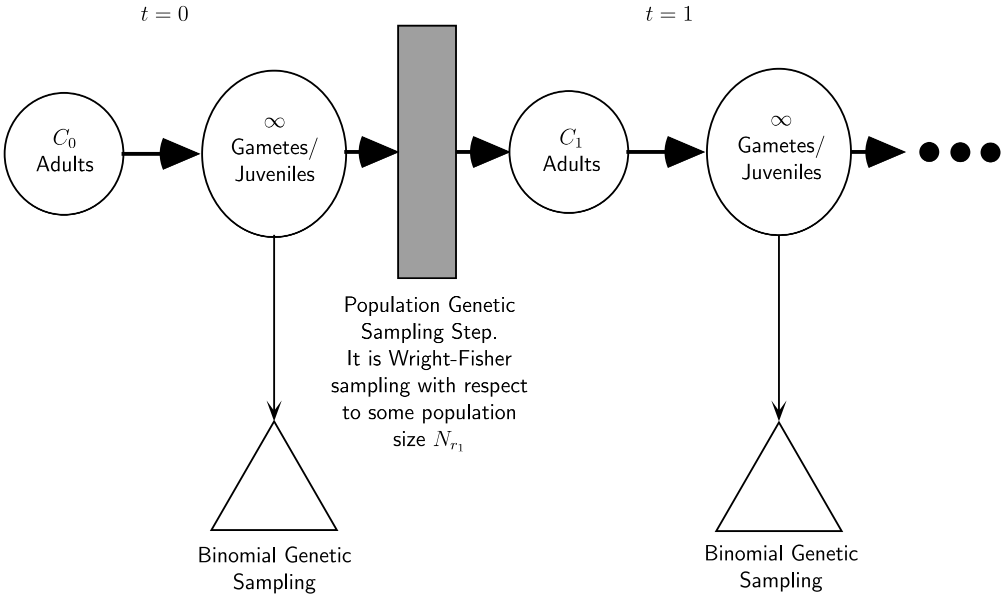

$t = 0$
$t = 1$
C_0 Adults
∞ Gametes/ Juveniles
C_1 Adults
∞ Gametes/ Juveniles
Population Genetic Sampling Step. It is Wright-Fisher sampling with respect to some population size N_{r_1}
Binomial Genetic Sampling
Binomial Genetic Sampling

Figure 6.2: Schematic showing how the sampling scheme described in the text might arise in organisms with high fecundity. The population genetic sampling associated with intergenerational inheritance is assumed to occur primarily between the juvenile and the adult stage.

6.3.1 Population Genetics Background

In a Wright-Fisher population of size N_e diploid individuals the frequency of an allele changes by genetic drift such that after each generation, the expectation of the frequency is the initial frequency, but the variance increases (Wright 1931). Thus at time t, the random variable p_t—the allele frequency—has for its first two central moments:

$$\begin{aligned} E(p_t) &= p_0 \\ \mathrm{Var}(p_t) &= p_0(1-p_0)\left[1-\left(1-\frac{1}{2N_e}\right)^t\right]. \end{aligned}$$

Expanding this, we have

$$\left(1-\frac{1}{2N_e}\right)^t = 1-\frac{t}{2N_e}+O\left(\frac{1}{N_e^2}\right),$$

so that when $t/2N_e$ is small,

$$\frac{\mathrm{Var}(p_t)}{p_0(1-p_0)} = \frac{E(p_t-p_0)^2}{p_0(1-p_0)} \approx \frac{t}{2N_e}. \tag{6.1}$$

The expectation of the squared deviation of p_t from its starting value p_0, when adjusted to compensate for the effects of its starting frequency [*i.e.*, when divided by $p_0(1-p_0)$], is approximately a linear function of t with slope equal to $1/2N_e$.

This forms the basis for Wright's F-statistic:

$$F = \frac{(p_t-p_0)^2}{p_0(1-p_0)}.$$

Since p_0 is a constant and p_t is the random variable the expectation of F is $\approx t/2N_e$ by (6.1). Thus, if we could measure K population allele frequencies without error, the estimator

$$\tilde{F} = \frac{1}{K}\sum_{i=1}^{K} \frac{(p_{ti} - p_{0i})^2}{p_{0i}(1 - p_{0i})}$$

would be unbiased for $t/2N_e$.

In practice, however, one cannot observe p_0 and p_t without error, but may only observe sample estimates y_0 and y_t. Several authors have suggested estimators for $E(F)$ in this case. For a single locus with K alleles Krimbas and Tsakas (1971) proposed

$$\tilde{F}_a = \frac{1}{K}\sum_{i=1}^{K} \frac{(y_{0i} - y_{ti})^2}{y_{0i}(1 - y_{ti})}.$$

They just use y_{0i} and y_{ti} in place of p_{0i} for p_{ti}. One notable problem with this is that if an allele appears in the sample at time t, but not at time 0, or vice-versa, then the estimate is infinite.

Others have proposed estimators, which differ only in how the $p_0(1 - p_0)$ term in the denominator is estimated. Nei and Tajima (1981) proposed

$$\tilde{F}_c = \frac{1}{K}\sum_{i=1}^{K} \frac{(y_{0i} - y_{ti})^2}{(y_{0i} + y_{ti})/2 - y_{0i}y_{ti}}. \tag{6.2}$$

and Pollak (1983) suggested the estimator

$$\tilde{F}_k = \frac{1}{K-1}\sum_{i=1}^{K} \frac{(y_{0i} - y_{ti})^2}{(y_{0i} + y_{ti})/2}. \tag{6.3}$$

both of which avoid the problem of the estimate going to infinity when an allele is detected in only one year. In extensive computer simulations Waples (1989) found that F_c and F_k yield similar results in most situations. Though the expectation of $\tilde{F}_a$ $\tilde{F}_c$ or $\tilde{F}_k$ is not the expectation of parametric F, because the expectation of a ratio of random variables is not the ratio of the expectations and the expectation of the square of a random variable is not the square of the expectation, the estimators are, in most instances, not badly biased, unless some of the alleles are at very low frequency (Waples 1989).

6.3.2 Estimating N_e from the Estimates of $E(F)$

Having estimated $E(F)$, one must convert that estimate (say, $\tilde{F}_c$, though the following also applies to $\tilde{F}_a$ and $\tilde{F}_k$) into an estimate of the effective size. This is typically done by first assuming that the expectation of $\tilde{F}_c$ is approximately

$$E(\tilde{F}_c) \approx \frac{E(y_0 - y_t)^2}{p_0(1 - p_0)} = \frac{\mathrm{Var}(y_0 - y_t)}{p_0(1 - p_0)}.$$

Then, the variance of $(y_0 - y_t)$ may be obtained by considering conditional variances and expectations. Dropping terms of $O(1/N_e^2)$ yields

$$E(\tilde{F}) \approx \frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t}{2N_e},$$

and solving for N_e gives the estimator

$$\tilde{N}_e = \frac{t}{2[\tilde{F}_w - 1/(2S_0) - 1/(2S_t)]} \tag{6.4}$$

(S_0 and S_1 are the sample sizes in number of diploid individuals) for the sampling scheme of Figure 6.2 (Waples 1989).

6.3.3 Estimators for Multiple Samples in Time

Pollak (1983), assuming a population of constant effective size N_e develops an estimator for $r+1$ sampling points at generations $t_0, \ldots, t_r$, with sample sizes $S_0, \ldots, S_r$. Though he developed this method for multiallelic loci, we restrict our attention to the diallelic case where the observed allele frequencies are $y_0, \ldots, y_r$, and for notational convenience we denote the frequencies of the alternate allele at each time point $w_0, \ldots, w_r$ (thus $w = 1 - y$). Pollak (1983) proposes the F-statistic F_{K_r} for this scenario:

$$F_{K_r} = 2\sum_{k=1}^{r}\left(\frac{(y_k - y_{k-1})^2}{y_k + y_{k-1}} + \frac{(w_k - w_{k-1})^2}{w_k + w_{k-1}}\right),$$

and derives, via a diffusion approximation, its expectation for sampling plan II as

$$E_P(F_{K_r}) \approx \frac{t_r - (2r-1)}{N_e} + \sum_{k=1}^{r}\left(\frac{1}{S_k} + \frac{1}{S_{k-1}}\right).$$

The corresponding estimator of effective size is thus[1]

$$\tilde{N}_{K_r} = \frac{t_r - (2r-1)}{F_{k_r} - \sum_{k=1}^{r}\left(\frac{1}{S_k} + \frac{1}{S_{k-1}}\right)}. \tag{6.5}$$

An undesirable feature of this estimator is that the numerator is always negative when the number of samples exceeds $(t_r + 3)/2$. (Note that t_r here is the same as T in our population scenario.) Therefore the numerator will always be negative when a sample is available from every generation of a population for any interval longer than one generation.

Waples (1990b) suggests a different method. When there are $r+1$ different samples available from a population with overlapping generations (such as Pacific salmon) he notes that there are $r(r+1)/2$ different time intervals for which $\tilde{F}$ could be computed, and he advocates using the mean $\tilde{F}$ computed from each of those $r(r+1)/2$ comparisons as the estimator for $E(F)$. In estimating N_e from that, however, the value of t in Equation 6.4 is not just length of time between the two most distant samples. In the case dealt with by Waples (1990b), t in (6.4) is replaced by a "b" to account for the overlapping generations, and when using $r+1$ samples the appropriate b is the mean b over all $r(r+1)/2$ comparisons. This is easily translated into the present scenario because, in the discrete generation case, b is merely t; thus the appropriate t in (6.4) is the mean t over all the $r(r+1)/2$ comparisons, weighted by the number of comparisons. This approach gives the estimator:

$$F_{N_r} = \frac{2}{r(r+1)}\sum_{i=0}^{r}\sum_{j>i}\frac{1}{2}\left(\frac{(y_i - y_j)^2}{(y_j + y_j)/2 - y_i y_j} + \frac{(w_i - w_j)^2}{(w_i + w_j)/2 - w_i w_j}\right) \tag{6.6}$$

[1] In retrospect, the other estimator that Pollak (1983) gives would be less biased in the simulations that I did. However, it does seem to me that (6.3.3) *is* the estimator Pollak (1983) gives for the sampling plan under which I simulate data, later in this chapter.

if the formula for $\tilde{F}_c$ (6.2) is used, and

$$F_{P_r} = \frac{2}{r(r+1)} \sum_{i=0}^{r} \sum_{j>i} 2 \left(\frac{(y_i - y_j)^2}{y_j + y_j} + \frac{(w_i - w_j)^2}{w_i + w_j} \right) \tag{6.7}$$

when the formula for F_k (6.3) is used. An estimate for N_e is then found by inserting either F_{P_r} or F_{N_r} into (6.4) and using for t in that expression the average

$$\tilde{t} = \frac{2}{r(r+1)} \sum_{i=0}^{r} \sum_{j>i} (t_j - t_i). \tag{6.8}$$

6.3.4 Estimating λ from $\tilde{N}_e$

Since the quantity estimated by the above estimators is the harmonic mean effective size (Waples 1990b), λ in this simple situation may be easily estimated. If each generation t there were $N_{r_t} = \lfloor \lambda C_t \rfloor$ effective adults, then the F-based estimate, $\tilde{N}_e$ would yield an estimate of

$$\frac{T}{\displaystyle\sum_{t=1}^{T} \frac{1}{\lfloor \lambda C_t \rfloor}}.$$

Thus, assuming that λC_t is an integer (or otherwise disregarding the small error from rounding) the estimate of λ by one of the above F-based methods is

$$\tilde{\lambda} = \frac{\tilde{N}_e \sum_{t=1}^{T} \frac{1}{C_t}}{T}, \tag{6.9}$$

which is just the estimate of the effective size divided by the harmonic mean of the observed census sizes.

6.4 The Maximum Likelihood Approach

6.4.1 The Probability Model

Referring back to Figure 6.1, let $\mathbf{Y}$ denote the vector of all Y_{tj}, *i.e.*, the genetic data. Since genetic drift in this population is a Markov chain (albeit time-inhomogeneous if the population size is inconstant), the probability of $\mathbf{Y}$ given the parameters λ and $\mathbf{X}_0$, the unobserved population allele counts at $t = 0$, can be computed as the sum over latent variables $\mathbf{X}$, the unobserved allele counts in the population. Thus we have

$$\begin{aligned} L(\lambda, \mathbf{X}_0) &= P_{\lambda,\mathbf{X}_0}(\mathbf{Y}) = \sum_{\mathbf{X}} P_{\lambda,\mathbf{X}_0}(\mathbf{Y}, \mathbf{X}) & (6.10) \\ &= \sum_{x_1,\ldots,x_T} \left(\prod_{t=0}^{T} \prod_{j=1}^{J} P_\lambda(Y_{tj}|X_{tj}) P_\lambda(X_{t+1,j}|X_{tj}) \right), & (6.11) \end{aligned}$$

where $P_\lambda(Y_{tj}|X_{tj})$ is understood to be unity any time that no genetic sample is taken, *i.e.*, when $S_{tj} = 0$. It is easy to compute the joint probability, $P_\lambda(\mathbf{X}, \mathbf{Y})$, if the $\mathbf{X}$ are known; then $P_\lambda(\mathbf{X}, \mathbf{Y})$ is merely the probability of a fully-specified path through a Markov Chain (with no parts of it "hidden" any longer). The probabilities of the individual transitions of this chain are easily computed. In generation t the effective number of diploid breeding adults is $N_{r_t} = \lfloor \lambda C_t \rfloor$, so at the j^{th} locus the population allele count in the next generation follows a binomial distribution depending on the current generation:

$$X_{t+1,j} \sim \text{Binomial}(2N_{r_{t+1}}, x_{tj}/(2N_{r_t})). \tag{6.12}$$

Likewise, the allele count in a sample drawn at time t at the j^{th} locus is binomially distributed,

$$Y_{tj} \sim \text{Binomial}(2S_{tj}, x_{tj}/(2N_{r_t})). \tag{6.13}$$

Though the individual transition probabilities are readily computed, the summation over $\mathbf{X}$-values is difficult (if not impossible) to perform because there are so many terms in it—the number of terms increases with C, λ, and T, as well as with additional alleles. To apply this approach to real scenarios will require Monte Carlo evaluation of the sum over those $\mathbf{X}$ values. Below, however, we explore the method in cases where exact computation is feasible.

In the above development of the likelihood, $\mathbf{X}_0$, the initial population allele counts are treated as parameters to be estimated jointly with λ. It is undesirable, however, to estimate an initial frequency for every allele, and, in fact with the number of nuisance parameters increasing with each new allele employed in a sample, the estimate of λ is inconsistent (Neyman and Scott 1948). To avoid these problems, we instead seek to maximize an integrated likelihood (Kalbfleisch and Sprott JRSS 1976 or so) formed by integrating out the nuisance parameters over a prior distribution, $\pi(\mathbf{X}_0)$, for the initial allele frequencies. $\pi(\mathbf{X}_0)$ may be an uninformative prior, or one that comes from previous observations on the types of locus systems being used (*e.g.*, allozyme markers vs. microsatellites). The integrated likelihood has the form:

$$\begin{aligned} L(\lambda) &= P_\lambda(\mathbf{Y}) = \sum_{\mathbf{X}} P_\lambda(\mathbf{Y}, \mathbf{X}) & (6.14) \\ &= \sum_{x_0,\dots,x_T} \left(\pi(X_{0j}) \prod_{t=0}^{T} \prod_{j=1}^{J} P_\lambda(Y_{tj}|X_{tj}) P_\lambda(X_{t+1,j}|X_{tj}) \right). & (6.15) \end{aligned}$$

6.4.2 Exact Computation of the Likelihood by Baum (1972) Algorithm

In the diallelic case, one may compute the likelihood in (6.15) exactly by a "peeling" method. (Without a peeling type of algorithm, the number of terms in the sum increases roughly exponentially with T—exactly so if C_t is constant.) The simple Markov structure of the process lets us start from time T and then work our way backward in time storing the probability of all the future events conditional on a current state, and using those in the sum over allele frequencies in the preceding time step. First, rewrite the right hand side of (6.15) without the locus subscripts for notational clarity:

$$\sum_{x_0,\dots,x_T} \left(\pi(X_0) \left(\prod_{t=0}^{T-1} P(Y_t|X_t) P(X_{t+1}|X_t) \right) P(Y_T|X_T) \right). \tag{6.16}$$

Since any X_T terms in the product appear with no other X's except X_{T-1} we may factor those out and consider two separate sums:

$$\sum_{x_0,\dots,x_{T-1}} \left(\pi(X_0) \prod_{t=0}^{T-2} P(Y_t|X_t) P(X_{t+1}|X_t) \right) \sum_{x_T} P(X_T|X_{T-1}) P(Y_T|X_T). \tag{6.17}$$

The second sum in that expression is the sum over all intervening values between X_{T-1} and Y_T. It is thus the probability of Y_T given X_{T-1}. We can compute and store this $P(Y_T|X_{T-1})$ for each of the values of X_{T-1} by summing over the possible values of X_T, and we can store the $P(Y_T|X_{T-1})$ in memory. We have now "pruned back" our sum to look like

$$\sum_{x_0,\ldots,x_{T-1}} \left(\pi(X_0) \prod_{t=0}^{T-2} P(Y_t|X_t)P(X_{t+1}|X_t) \right) P(Y_T|X_{T-1}). \tag{6.18}$$

Now we may treat all the terms involving X_{T-1}'s separately from the rest, proceeding as in (6.17) only one "layer" further back:

$$\sum_{x_0,\ldots,x_{T-2}} \left(\pi(X_0) \prod_{t=0}^{T-3} P(Y_t|X_t)P(X_{t+1}|X_t) \right) \sum_{x_{T-1}} P(X_{T-1}|X_{T-2})P(Y_{T-1}|X_{T-1})P(Y_T|X_{T-1}). \tag{6.19}$$

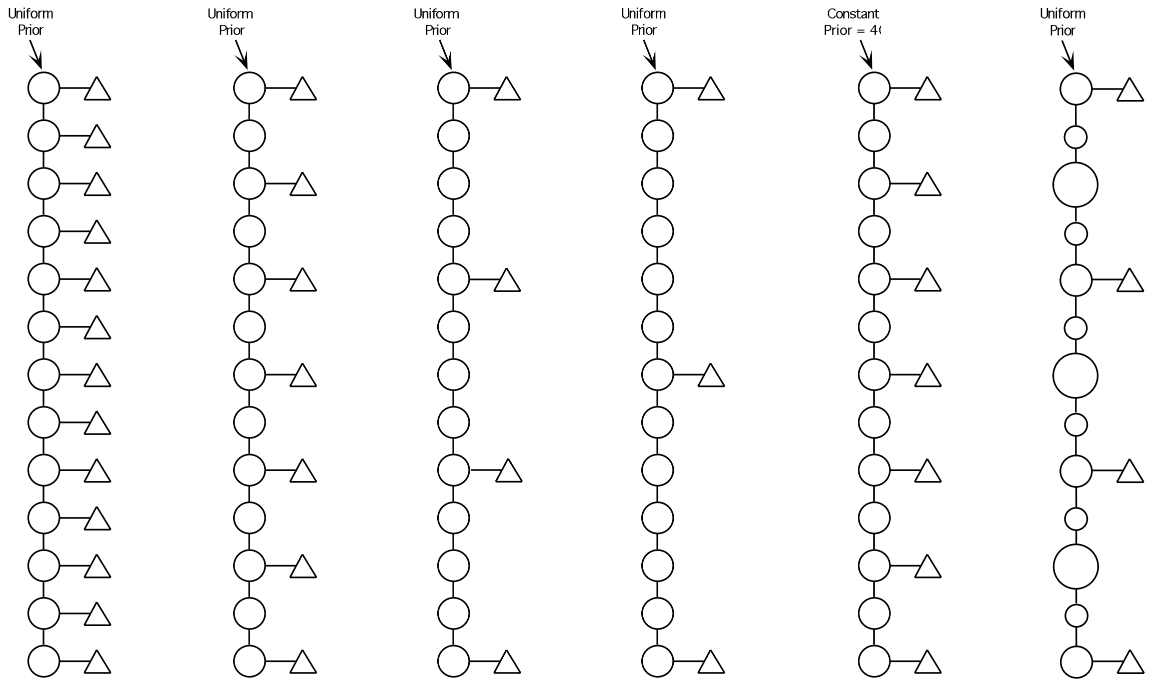
The second sum in the above expression is $P(Y_T, Y_{T-1}|X_{T-2})$ which we can compute for each of the values of X_{T-2} by summing over the previously stored values of $P(Y_T|X_{T-1})$. Once we have computed all the $P(Y_T, Y_{T-1}|X_{T-2})$'s we may discard the $P(Y_T|X_{T-1})$'s, freeing up computer memory. We proceed thus until finally summing the values of $P(Y_0, \ldots Y_T|X_0)$ weighted by the prior on X_0.

By performing the sum this way, a problem that seemed to increase exponentially in T is now only linear in T. The difficulty arises in the storage requirements. Though this is not a great problem for diallelic loci, for loci with multiple (say k) alleles the memory requirements become severe. In general the memory required is twice the number of possible allele frequency states—a number that grows very quickly with both N_r and k, being the number of ordered k-tuples whose sum is $2N_r$ (for diploids). For example, if k is 6 and N_r is 150 diploid individuals, there are more than 21 billion possible allele frequency states. To compute the sum as above for the exact likelihood would require more than 170 gigabytes of RAM. (Also, when $\lfloor \lambda C_t \rfloor$ is large, this method consumes a large amount of CPU time.)

6.5 The Simulations

I have performed simulations to investigate the variance of a maximum likelihood estimator for λ to that of two different F-statistic approaches. Scenarios for simulation are described in Figure 6.3 and some summary statistics for the estimators presented in Table 6.1. Histograms of the MLE and the F-based estimator using Equation 6.7, follow in several figures.

The basic conclusion of the simulations is that though the F-statistic estimator has smaller variance, it becomes more biased as more samples are available in time, so that its mean squared error is greater than that of the MLE in data-rich situations. It should be noted that one might be able to develop a different F-statistic estimator for multiple samples that performs better than the "ACE" estimator and the "PE" estimator used here. Nonetheless, the MLE compares quite favorably with other methods available today.



(a) 13 samples

(b) 7 samples

(c) 4 samples

(d) 3 samples

(e) 7 samples, constant prior

(f) 7 samples, C cycling

Figure 6.3: Graphs showing six different scenarios for performing simulations. Each one involves 12 generations between the first and the last sample drawn. Sample sizes for each locus were 50 haploid individuals. In scenarios a–e, the observed census size C_t each generation was 100 haploid individuals, genetic drift was simulated at $\lambda = .3$, and 8 diallelic loci were drawn in each sample. In scenarios a–d the initial allele count in the population was drawn from a discrete uniform distribution on $[1, \lfloor 100\lambda \rfloor - 1]$, and the prior π assumed for computing the likelihood (6.15) was the same (discrete uniform). In scenario e, the initial allele count for the simulations was constant (a frequency of .4, thus 40 out of 100) while the prior in (6.15) was still the discrete uniform. In scenario f, the observed census size cycles $100 \rightarrow 50 \rightarrow 200 \rightarrow 50 \rightarrow 100 \rightarrow \ldots$ and samples are drawn from the years with 100 individuals observed. The "true" λ for this simulation was .4.

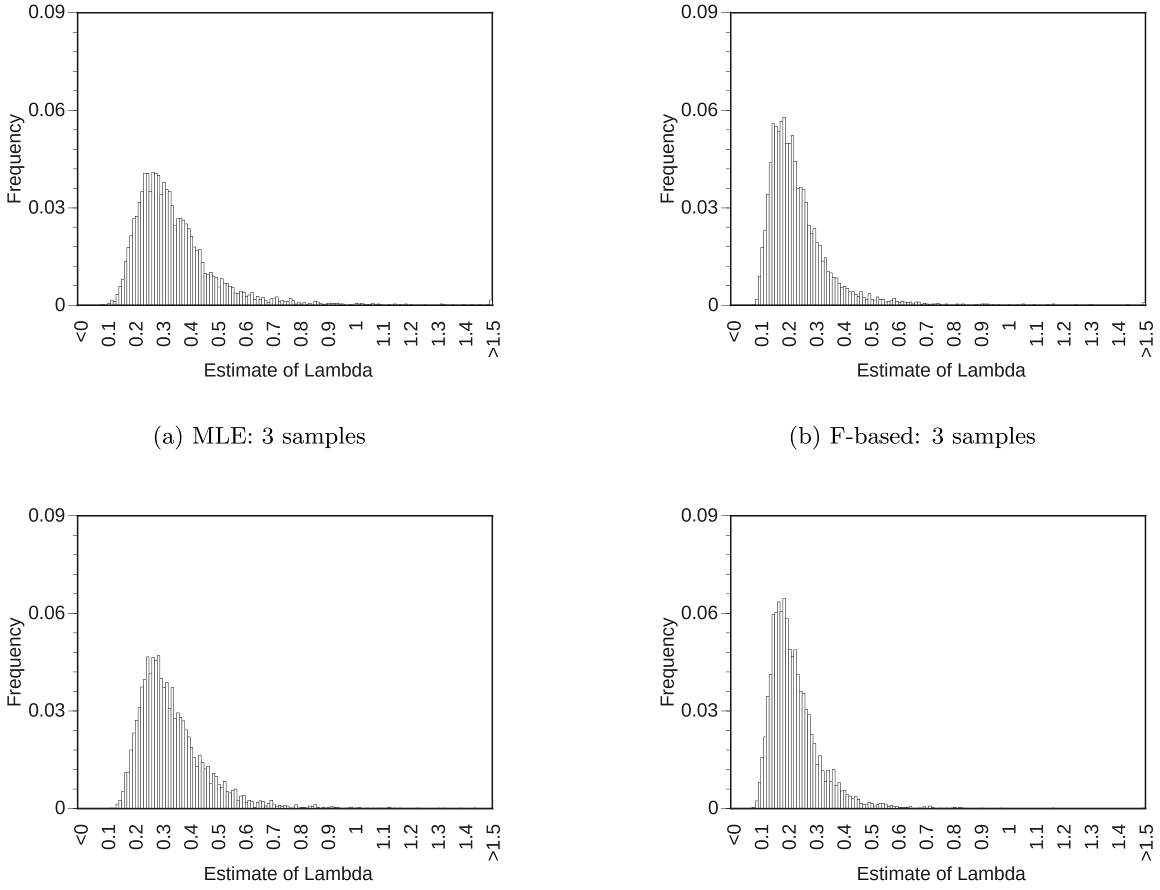
Table 6.1: Summary statistics for the three estimators of λ: "MLE"=maximum likelihood estimate; "ACE"="all comparisons estimator"= F-based method using F from Equation 6.7; "PE"=Pollak's estimator (6.5) with the obvious adjustment made for a haploid population. All simulations were of 5,000 replicate estimates of λ, except for the one with cycling population size which included 1,250 replicates.

	# of	Sample Means			Sample Variances			Est. Bias			Est. MSE		
Census	Samples	MLE	ACE	PE	MLE	ACE	PE	MLE	ACE	PE	MLE	ACE	PE
100	3	0.353	0.246	0.180	0.0273	0.0160	0.0170	0.053	−0.054	−0.120	0.0301	0.0189	0.0313
100	4	0.345	0.233	0.125	0.0168	0.0091	1.1579	0.045	−0.067	−0.175	0.0189	0.0135	1.1882
100	7	0.330	0.219	0.021	0.0098	0.0058	0.0381	0.030	−0.081	−0.279	0.0107	0.0123	0.1159
100	13	0.321	0.213	−0.416	0.0065	0.0043	70.3162	0.021	−0.087	−0.716	0.0069	0.0118	70.8140
100[a]	7	0.337	0.197	0.015	0.0118	0.0042	< 0.0001	0.037	−0.103	−0.285	0.0132	0.0149	0.0815
50–200[b]	4	0.455	0.309	0.198	0.0232	0.0164	0.0559	0.055	−0.091	−0.203	0.0262	0.0247	0.0969

Notes:

[a] For this simulation, initial population allele frequencies x_0 were not drawn from a uniform distribution. Instead, $x_0 = .4 = 40/100$ was used to start each simulation.

[b] The census size each generation was variable and followed four-generation cycle of 100 → 50 → 200 → 50 → 100 →. . . , for generations 0 to 12. The harmonic mean of the census size in such a population is 72.73.



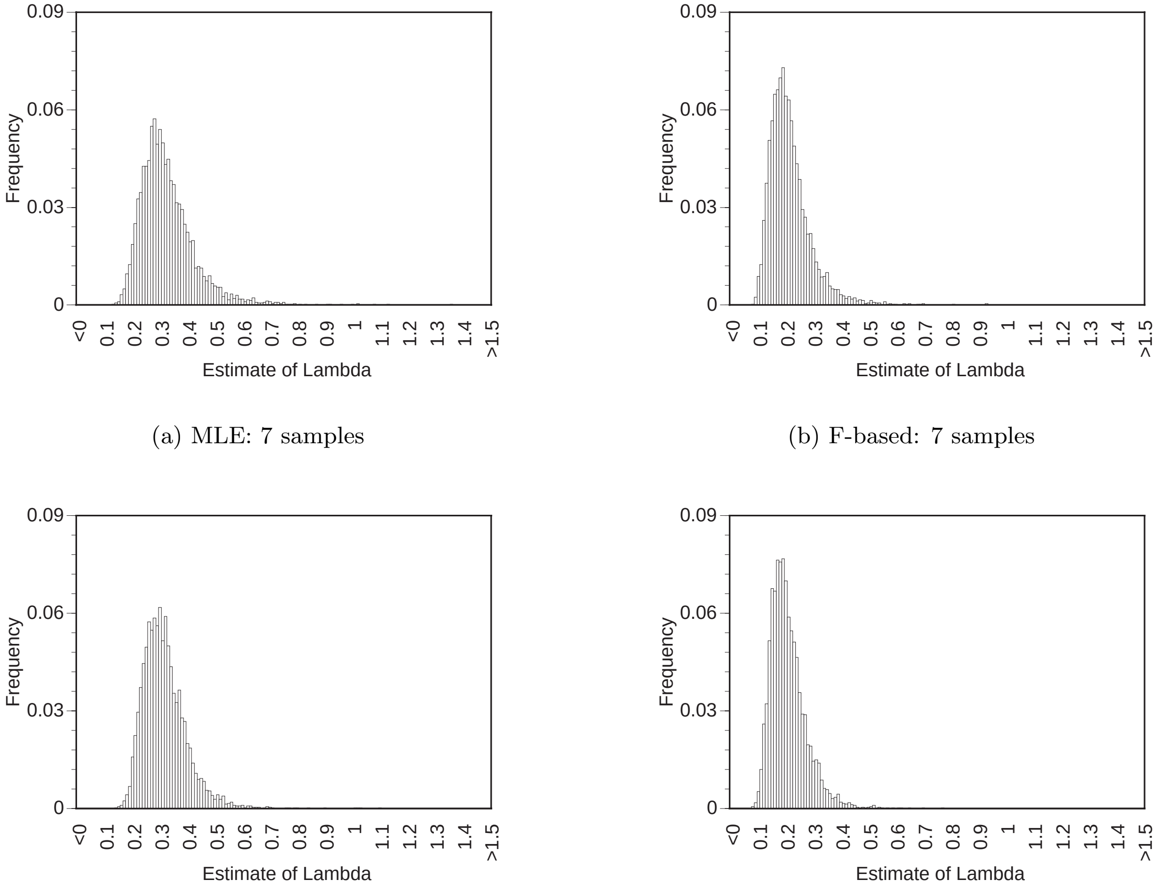
(a) MLE: 3 samples

(b) F-based: 3 samples

(c) MLE: 4 samples

(d) F-based: 4 samples

Figure 6.4: Histograms of the estimators. MLE is the maximum likelihood method. “F-based” is the ACE estimator. See caption in Table 6.1



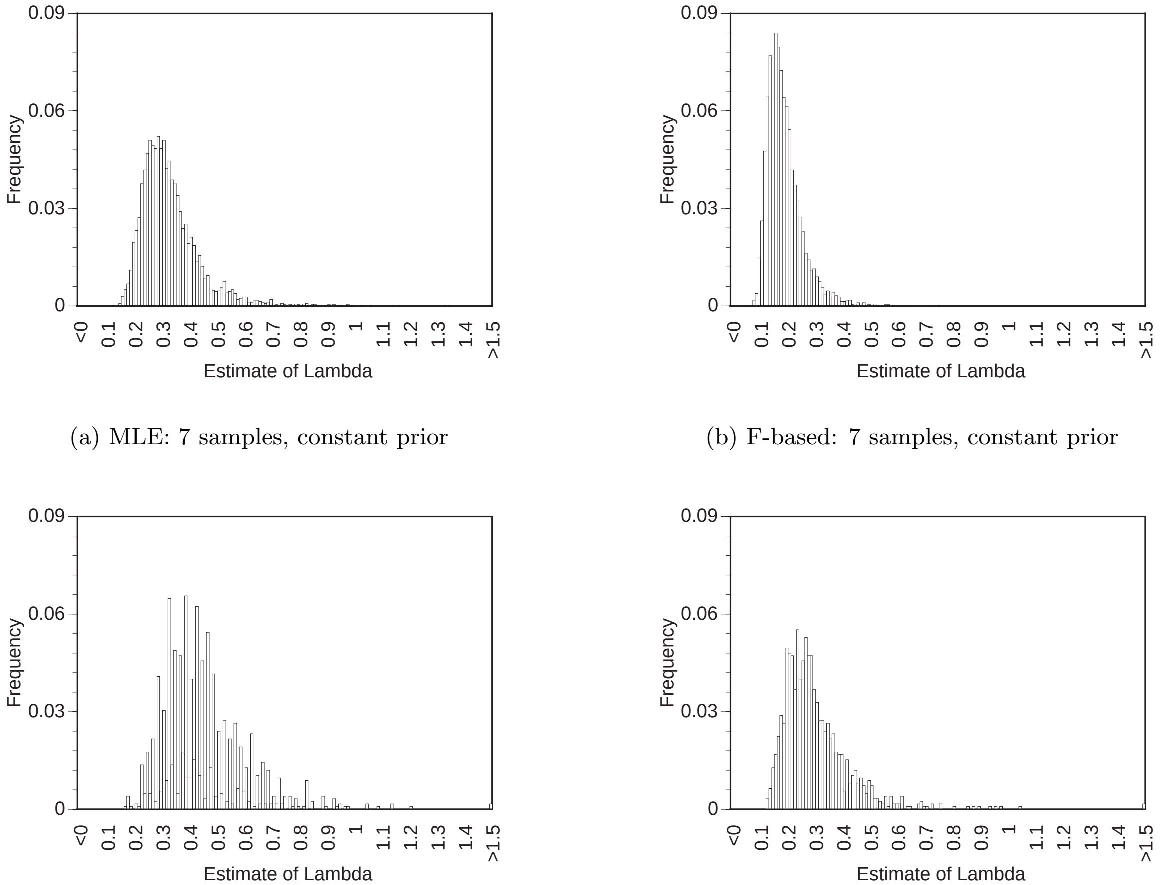
(a) MLE: 7 samples

(b) F-based: 7 samples

(c) MLE: 13 samples

(d) F-based: 13 samples

Figure 6.5: Histograms of the estimators. MLE is the maximum likelihood method. “F-based” is the ACE estimator. See caption in Table 6.1



(a) MLE: 7 samples, constant prior

(b) F-based: 7 samples, constant prior

(c) MLE: 4 samples, cycling C_t, true $\lambda = .4$

(d) F-based: 4 samples, cycling C_t, true $\lambda = .4$

Figure 6.6: Histograms of the estimators. MLE is the maximum likelihood method. “F-based” is the ACE estimator. See caption in Table 6.1

Bibliography

Anderson, E. C., 1998 Inferring the ancestral origin of Sockeye Salmon, *Oncorhynchus nerka*, in the Lake Washington basin: A statistical method in theory and application. Master's thesis, University of Washington.

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.

Bacilieri, R. B., A. Ducousso, R. J. Petit, and A. Kremer, 1996 Mating system and asymmetric hybridization in a mixed stand of European oaks. Evolution **50:** 900–908.

Ballou, J. D., M. Gilpin, and T. J. Foose (Eds.), 1995 *Population Management for Survival and Recovery.* New York: Columbia University Press.

Bartley, D., M. Bagley, G. Gall, and B. Bentley, 1992 Use of linkage disequilibrium data to estimate the effective size of hatchery and natural fish 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.

Box, G. E. P. and G. C. Tiao, 1973 *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

Cavalli-Sforza, L. L., 1969 Genetic drift in an Italian Population. Scientific American **221:** 30–37.

Cavalli-Sforza, L. L. and A. W. F. Edwards, 1967 Phylogenetic analysis: models and estimation procedures. Evolution **21:** 550–570.

Charlesworth, B., 1980 *Evolution in age-structured populations.* Cambridge, UK: Cambridge University Press.

Crow, J. F. and C. Denniston, 1988 Inbreeding and variance effective numbers. Evolution **42:** 482–495.

Crow, J. F. and N. E. Morton, 1955 Measurement of gene frequency drift in small populations. Evolution **9:** 202–214.

Faville, M. J., K. D. Adam, and M. E. Wedderburn, 1995 Allozyme variation within and between three populations of browntop (*Agrostis capillaris*). New Zealand Journal Of Agricultural Research **38:** 65–70.

Felsenstein, J., 1973 Maximum likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics **25:** 471–492.

Felsenstein, J., 1985 Phylogenies from gene frequencies: a statistical problem. Systematic Zoology **34:** 300–311.

Foose, T. J., L. de Boer, U. S. Seal, and R. Lande, 1995 Conservation management strategies based on viable populations. In J. D. Ballou, M. Gilpin, and T. J. Foose (Eds.), *Population Management for Survival and Recovery*, pp. 273–294. New York: Columbia University Press.

Fournier, D. A., T. D. Beacham, B. E. Riddell, and C. A. Busack, 1984 Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. Canadian Journal of Fisheries and Aquatic Sciences **41:** 400–408.

Frankham, R., 1995a Effective population size/adult population size ratios in wildlife: a review. Genetical Research Cambridge **66:** 95–107.

Frankham, R., 1995b Inbreeding and extinction: a threshold effect. Conservation Biology **9:** 792–799.

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 Monte Carlo. Technical Report 568r, School of Statistics, University of Minnesota.

Geyer, C. J., 1996 Estimation and optimization of functions. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 241–258. New York: Chapman and Hall.

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). J. Roy. Statist. Soc. Ser. 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.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Eds.), 1996 *Markov Chain Monte Carlo in Practice*. New York: Chapman and Hall.

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. Phil. Trans. Roy. Soc. (London) Ser. B **344:** 403–410.

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.

Hammersley, J. M. and D. C. Handscomb, 1964 *Monte Carlo Methods*. London: Methuen & Co Ltd.

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.

Heard, W. R., 1991 Life history of pink salmon. In C. Groot and L. Margolis (Eds.), *Pacific Salmon Life Histories*, pp. 119–230. 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.

Henderson, C. R., 1986 Recent developments in variance and covariance estimation. Journal of Animal Science **63:** 208–216.

Hill, W. G., 1979 A note on effective population size with overlapping generations. Genetics **92:** 317–322.

Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genetical Research (Cambridge) **38:** 209–216.

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.

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.

Kingman, J. F. C., 1982 On the genealogy of large populations. Journal of Applied Probability **19A:** 27–43.

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.

Lacy, R. C., J. D. Ballou, F. Princeé, A. Starfield, and E. A. Thompson, 1995 Pedigree analysis for population management. In J. D. Ballou, M. Gilpin, and T. J. Foose (Eds.), *Population Management for Survival and Recovery*, pp. 57–75. New York: Columbia University Press.

Lange, K. and S. Matthysse, 1989 Simulation of pedigree genotypes by random walks. American Journal of Human Genetics **45:** 959–970.

Lewontin, R. C. and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175–195.

Long, J. C., 1991 The genetic structure of admixed populations. Genetics **127:** 417–428.

Long, J. C. and P. E. Smouse, 1983 Intertribal gene flow between the Ye'cuana and Yanomama: genetic analysis of an admixed village. American Journal of Physical Anthropology **61:** 411–412.

Lynch, M., 1988 Estimation of relatedness by DNA fingerprinting. Molecular Biology and Evolution **5:** 584–599.

Lynch, M., J. Conery, and R. Buerger, 1995 Mutation accumulation and the extinction of small populations. American Naturalist **146:** 489–518.

Machugh, D. E., M. D. Shriver, R. T. Loftus, P. Cunningham, and D. G. Bradley, 1997 Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). Genetics **146:** 1071–1086.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953 Equations of state calculations by fast computing machines. Journal of Chemical Physics **21:** 1087–1092.

Millar, R. B., 1987 Maximum likelihood estimation of mixed stock fishery composition. Canadian Journal of Fisheries and Aquatic Sciences **44:** 583–590.

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. Annu. rep. res., NOAA, Northwest and Alaska Fisheries Center, Seattle, Washington.

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.

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.

Neyman, J. and E. L. Scott, 1948 Consistent estimates based on partially consistent observations. Econometrica **16:** 1–32.

Nielsen, R., 1997 A likelihood approach to populations samples of microsatellite alleles. Genetics **146:** 711–716.

Nunney, L. and D. R. Elam, 1994 Estimating the effective size of conserved populations. Conservation Biology **8:** 175–184.

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.

Painter, I., 1997 Sibship reconstruction without parental information. Journal of Agricultural, Biological, and Environmental Statistics **2:** 212–229.

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.

Pollak, E., 1983 A new method for estimating the effective population size from allele frequency changes. Genetics **104:** 531–548.

Propp, J. G. and D. B. Wilson, 1996 Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures and Algorithms **9:** 223–252.

Rannala, B. and J. A. Hartigan, 1996 Estimating gene flow in island populaions. Genetical Research (Cambridge) **67:** 147–158.

Richardson, S. and P. J. Green, 1996 On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society, Series B **59:** 731–792.

Ricker, W. E., 1972 Hereditary and environmental factors affecting certain salmonid populations. In P. A. Larkin and R. C. Simon (Eds.), *The Stock Concept in Pacific Salmon. H.R. MacMillan lectures in fisheries*, pp. 1–8. Vancouver, B.C.: University of British Columbia.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.

Smith, A. F. M. and G. O. Roberts, 1993 Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. Roy. Statist. Soc. Ser. B **55:** 3–24.

Smith, I. P., A. D. F. Johnstone, and G. W. Smith, 1997 Upstream migration of adult Atlantic salmon past a fish counter weir in the Aberdeenshire Dee, Scotland. Journal of Fish Biology **51:** 266–274.

Smouse, P. E., R. S. Waples, and J. A. Tworek, 1990 A genetic mixture analysis for use with incomplete source population data. Can. J. Fish. Aquat. Sci. **47:** 620–634.

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.

Tautz, D., 1989 Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research **17:** 6463–6471.

Thompson, E. A., 1973 The Icelandic mixture problem. Annals of Human Genetics, London **37:** 69–80.

Thompson, E. A., 1975 *Human Evolutionary Trees.* Cambridge, UK: Cambridge University Press.

Thompson, E. A., 1984 Inferring migration history from genetic data: application to the Faroe Islands. In A. J. Boyce (Ed.), *Migration and Mobility: SSHB Symposium Volume 23*, pp. 123–142. London, UK: Taylor and Francis.

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 J. Math. Appl. Med. & Biol. **8:** 149–169.

Thompson, E. A. and S. C. Heath, 1997 Estimation of conditional multilocus gene identity among relatives. 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.

Thompson, E. A. and S. C. Heath, 1998 Estimation of conditional multilocus gene identity among relatives. IMS Lecture Note Series (in press).

Thompson, E. A., J. V. Neel, P. E. Smouse, and R. Barrantes, 1992 Microevolution of the Chibcha-speaking peoples of lower Central America: rare genes in an Amerindian Complex. Amer. J. Hum. Genet. **51:** 609–626.

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. and D. J. Teel, 1990 Conservation genetics of Pacific salmon I. Temporal changes in allele frequency. Conservation Biology **4:** 144–156.

Ward, B. R. and P. A. Slaney, 1988 Life history and smolt-to-adult survival of Keogh River steelhead trout (*Salmo gairdneri*) and the relationship to smolt size. Canadian Journal of Fisheries and Aquatic Sciences **45:** 1110–1122.

Withler, I. L., 1966 Variability in life history characteristics of steelhead trout (*Salmo gairdneri*). Journal of the Fisheries Research Board of Canada **23:** 365–392.

Wolfe, M. L. and J. F. Kimball, 1989 Comparison of bison population estimates with a total count. Journal of Wildlife Management **53:** 593–596.

Wright, J. M. and P. Bentzen, 1994 Microsatellites: genetic markers for the future. Reviews in Fish Biology and Fisheries **4:** 384–388.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Wright, S., 1937 The distribution of gene frequencies in populations. Proceedings of the National Academy of Sciences **23:** 307–320.

Zhivotovskii, L. A., M. K. Gulbokovskii, R. M. Viktorovskii, K. I. Afanas'ev, V. V. Efremov, L. N. Ermolenko, and B. Kalabushkin, 1989 Genetic differentiation in pink salmon. Genetika **25:** 1261–1274.