Chapter 4

λ AND OVERLAPPING GENERATIONS

4.1 Introduction

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

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

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

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

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

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

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

4.2 Overlapping Generations via an Urn Model

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

4.2.1 Dependence structure with the Pacific salmon life history

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

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

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

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

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

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

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

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

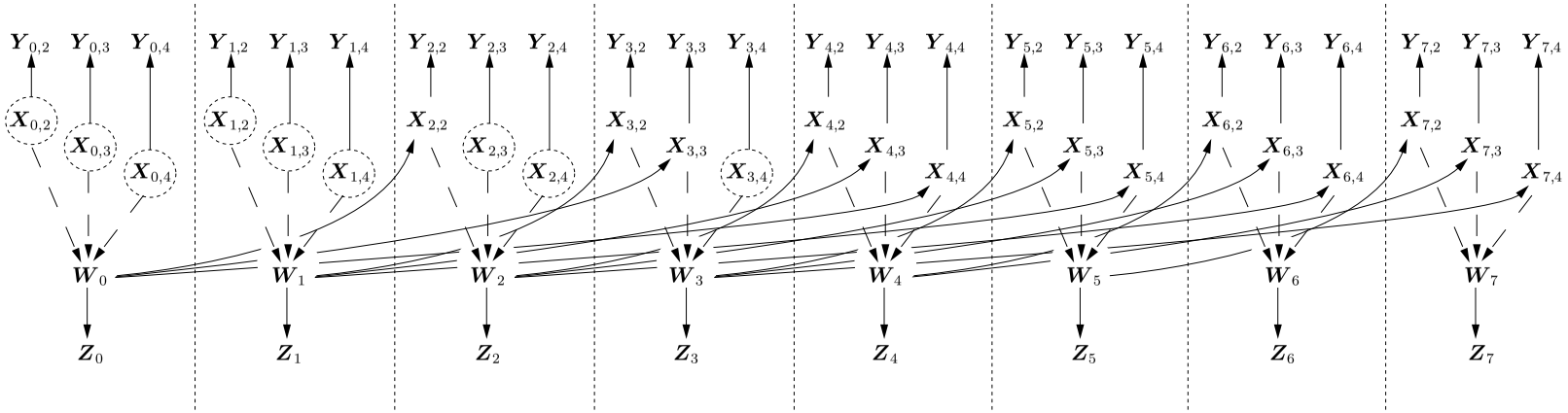


Figure 4.1: Acyclic directed graph describing the conditional dependence structure in the urn model with overlapping generations. Adults reproduce at ages 2, 3, and 4. The vertical, dotted lines separate time into different spawning years. The dotted circles represent latent variables in the set $\mathbf{X}_{\mathcal{P}}$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

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

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

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

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

4.2.2 Dependence structure under Sampling Scheme II and with null alleles

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

Nei and Tajima (1981) used the name "Sampling scheme II" for the case when the

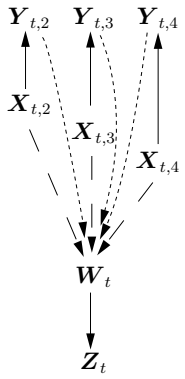


Figure 4.2: Acyclic directed graph describing the conditional dependence structure in the probability model for overlapping generations at year t, with three age classes of adults (2,3,4), under Sampling Scheme II—sampling adults destructively and before reproduction. The arrows connecting these variables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

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

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

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

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

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

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

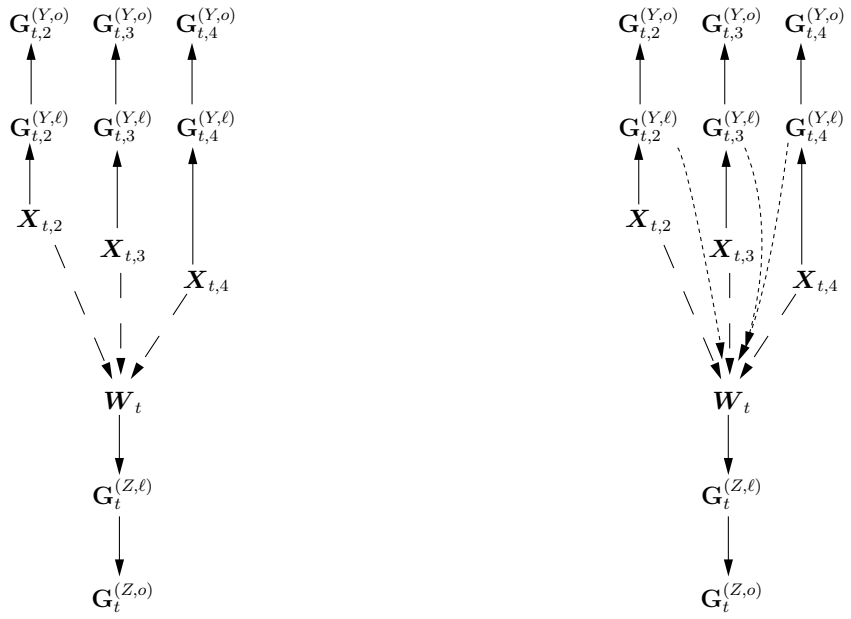


Figure 4.3: Acyclic directed graphs describing the conditional dependence structure in the probability model for overlapping generations at year t with three age classes (2,3,4) and with some alleles not codominantly expressed. The arrows connecting these graphs to other times are omitted in this figure.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

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

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

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

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

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

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

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

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

4.2.3 Specifying probability distributions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2.4 Probabilities with recessive alleles

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2.5 The prior distribution for allele counts

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

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

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

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

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

(1990b) empirically shows that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

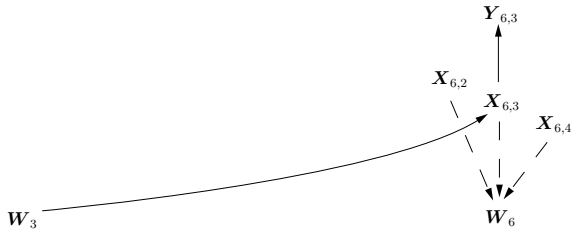
4.3.1 Neighborhood structures and joint probability ratios

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

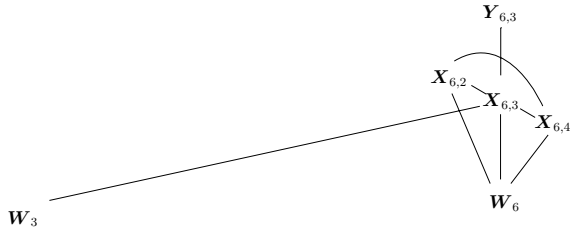
Let the data from the genetic samples be considered fixed at $\mathbf{Y}$ and $\mathbf{Z}$, and suppose the current values for $\mathbf{X}$ and $\mathbf{W}$ are denoted by $\mathbf{X}$ and $\mathbf{W}$, respectively. Let $\mathbf{X}'$ and $\mathbf{W}'$ differ from $\mathbf{X}$ and $\mathbf{W}$ only at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

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

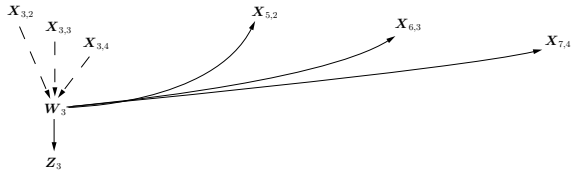
Calculating such ratios is done quickly by noting that they are functions only of a small collection of variables adjacent in the graph to the altered component. The variables adjacent to the altered component in the graph are members of its neighborhood, and the factors in the joint density including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



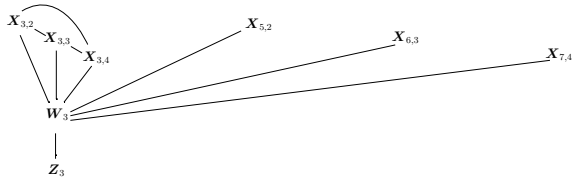
(a)



(b)



(c)



(d)

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.4 Special Cases

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

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

4.5 Simulated Data

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2. An update of λ.

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

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

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

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

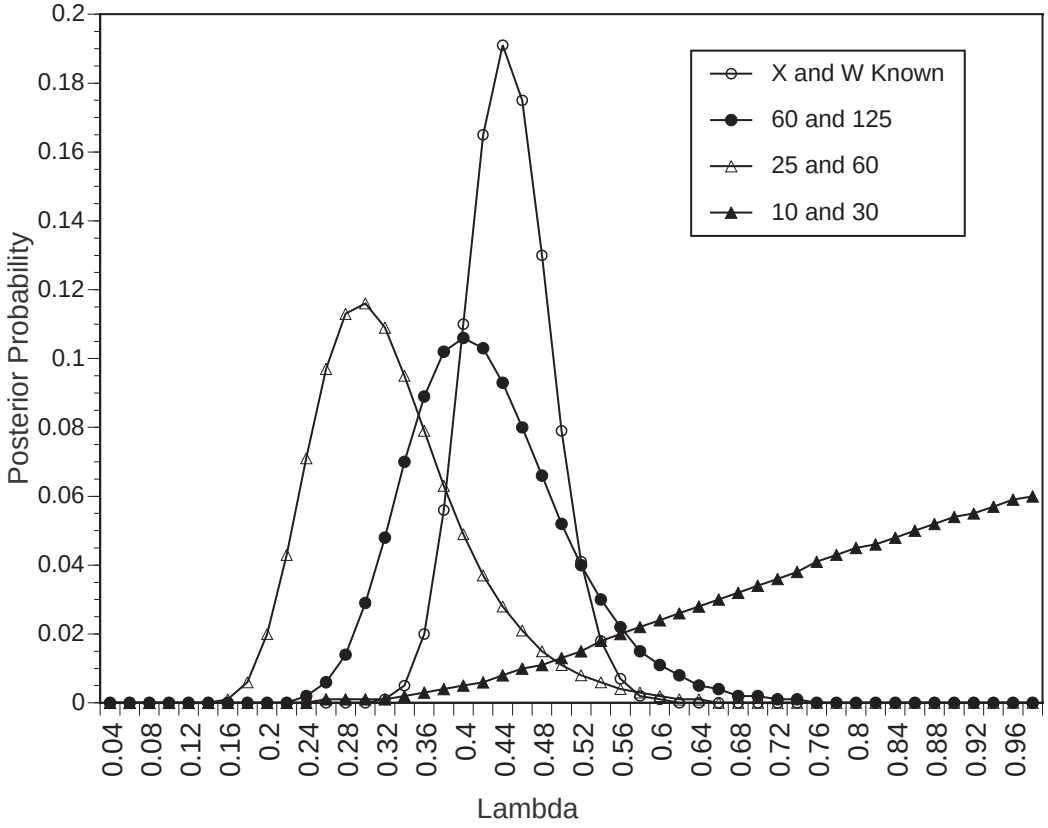
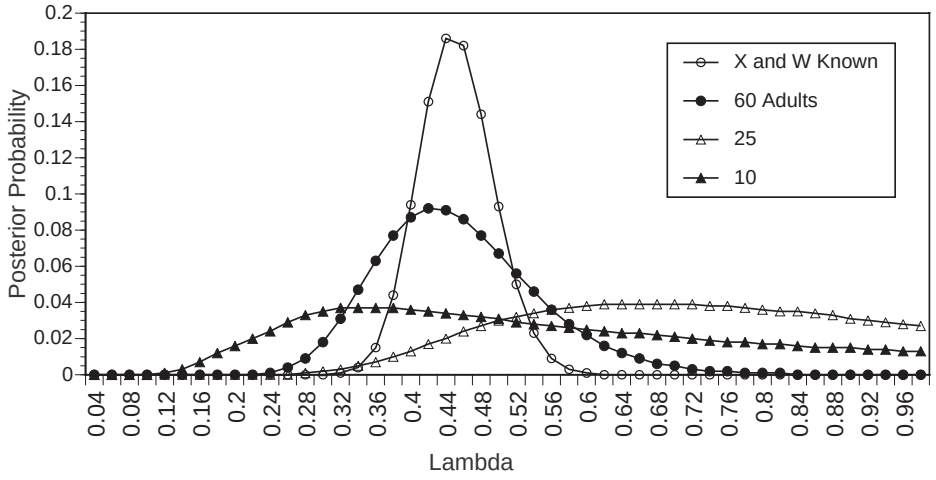
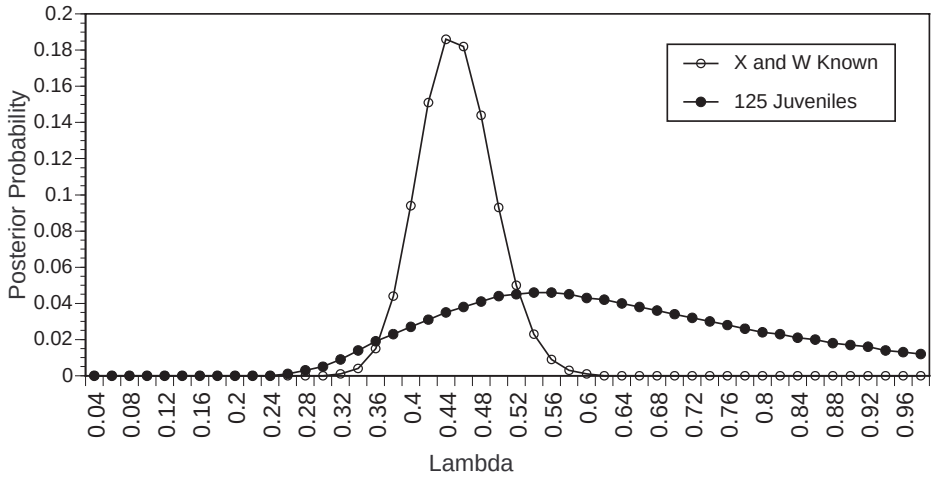


Figure 4.5: Plot of the posterior probabilities for $\lambda = \lambda_4 = \lambda_5$ from genetic data simulated on the Inmaha Creek chinook population (Table 4.1), when data were available on both adults and juveniles. The first line in the graph corresponds to the estimate with the latent variables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



(a) Samples taken only from adults



(b) Samples taken only from juveniles

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

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

4.6 Discussion

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

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

---

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