Salmon, Genetics, and Monte Carlo

Eric C. Anderson

Interdisciplinary Program in Quantitative Ecology and Resource Management

University of Washington, Seattle, WA
Overview

§1 Genetic mixtures
  • Introductory genetics
  • Latent data formulation
  • Gibbs sampler for the posterior distribution

§2 A model for population admixture

§3 Simultaneous mixture/admixture analysis
  • Difficult full conditional calculations
  • A Baum et al. type of computation

§4 Results
Rapid River Fish Hatchery Example—Genetic Mixture

The Big Question:
Can we separate the unmarked hatchery fish from the wild fish using genetic markers?
And also estimate allele frequencies in the two different populations?

- Fish caught at the hatchery trap are a mixture of wild-spawning and hatchery fish.

- 95% of hatchery fish are adipose fin-clipped (i.e., “marked”).
Genetics Background

Each salmon cell has about 64 Pairs of Chromosomes. Very precise locations in the genome may be reliably found and analyzed. Such a location is called a LOCUS (plural = LOCI).

Genetic variants at a locus are known as alleles.

Each individual has two copies of genetic material at a locus which determine its single-locus genotype.

The probability that an individual carries a particular allele at a locus depends on how frequent that allele is in the population.

The alleles carried at one locus are typically (and often quite reasonably) assumed to be independent of the alleles carried at other loci.
The probability of the fish’s multilocus phenotype, \( y \), is a simple function of the population allele frequencies, \( \theta \).
A Finite Mixture Model

- The marked hatchery fish are treated as a learning sample, size \( n_ℓ \)
- The remaining fish are a sample of size \( n_m \) from a mixture
  - An unknown proportion \( π_0 \) in the mixture are wild fish
  - \( π_1 \) are unmarked hatchery fish \((π_0 + π_1 = 1)\)
- Allele frequencies at all the loci available are unknown and denoted by \( θ_0 \) and \( θ_1 \) for the two populations (wild and hatchery), respectively.
- The probability of the phenotype of the \( i^{th} \) fish conditional on its being from Population \( j \) is \( P(y_i|θ_j), j = 0, 1 \).
- The likelihood function is

\[
P(y|π, θ) = \left[ \prod_{i=1}^{n_m} \left( \sum_{j=0}^{1} π_j P(y_i|θ_j) \right) \right] \cdot \left\{ \prod_{i=1}^{n_ℓ} P(y_i|θ_1) \right\}
\]
The Latent Data Perspective

Something is *missing* here... Something that would make this whole, messy enterprise trivially easy...
**Posterior Probability of Population Origin**

We wish to distinguish wild fish from unmarked hatchery fish—This is typically impossible to do with certainty.

- We may compute the probability of a fish’s origin:

\[
P(\text{fish } i \text{ Wild} | y_i, \pi, \theta) = \frac{\pi_0 P(y_i | \pi_0, \theta_0)}{\pi_0 P(y_i | \pi_0, \theta_0) + \pi_1 P(y_i | \pi_1, \theta_1)},
\]

if \( \pi \) and \( \theta \) are known. . . But they aren’t!

- We take a Bayesian approach—inference based on the posterior distribution ([Diebolt and Robert 1994](#)):

- Posterior probability of population origin is

\[
P(\text{fish } i \text{ Wild} | y) = \int_{\pi, \theta} P(\text{fish } i \text{ Wild} | y_i, \pi, \theta) P(\pi, \theta | y) d\pi d\theta
\]

Which is impossible to compute directly. But note:
\[ = \mathbb{E}\left( P(\text{fish } i \text{ Wild}|y_i, \pi, \theta) \mid y \right) \]

Which suggests a Monte Carlo estimator

\[ = \frac{1}{M} \sum_{k=1}^{M} P(\text{fish } i \text{ Wild}|y_i, \pi^{(k)}, \theta^{(k)}) \]

with \((\pi^{(k)}, \theta^{(k)}) \sim P(\pi, \theta|y)\).

- But, we must simulate each \((\pi^{(k)}, \theta^{(k)})\) from the posterior distribution.

\[ P(\pi, \theta|y) = \frac{1}{C} \rho(\pi) \rho(\theta) P(y|\pi, \theta) \]

- The prior \(\rho(\pi)\) is a Beta distribution and the prior for allele frequencies at each locus are Dirichlet distributions.
  - Often taken to be uniform

- \(C\) is typically unknown. But MCMC is an option.
Latent Data Revisited

Let the $i^{th}$ fish’s flag be $z_i = 0$ for wild origin or $z_i = 1$ for hatchery origin.

- Use Gibbs sampling to simulate values $(\pi^{(k)}, \theta^{(k)}, z^{(k)})$, $k = 1, \ldots, M$ then use the values of $\pi^{(k)}$ and $\theta^{(k)}$ that come out.

- Consecutively simulate new values from their full conditional distributions:
  - Update all the $z_i$'s ($i = 1, \ldots, n_m$).
  - Update $\pi$.
  - Update $\theta$.

- The full conditionals are simple distributions because of conjugacy and the use of latent variables.

$$P(z_i = 0|\cdots) = P(z_i = 0|y_i, \pi, \theta) = P(\text{fish } i \text{ Wild}|y_i, \pi, \theta)$$

$$(\pi_0|\cdots) = (\pi_0|z) \sim \text{Beta}(1 + \#\{z = 0\}, 1 + \#\{z = 1\})$$

- $\theta_0$ and $\theta_1$ have full conditional distributions which are Dirichlet distributions arising from standard multinomial sampling of allelic types.
What This Does/Doesn’t Offer You

Subject to the assumptions made in choosing the prior distributions for $\pi$ and $\theta$ this gets you:

- Estimates of the posterior probability of population of origin of each fish in the mixture.
- Estimates of the posterior distribution of allele frequencies of the wild and hatchery populations
- An estimate of the posterior distribution of $\pi$

However, the model is rather inflexible in assuming that every individual is purely descended from either the wild population or the hatchery population.
A Model For Admixed Populations
(Pritchard et al., in press)

\[ q_{0i} \sim \text{Beta}(\alpha, \alpha) \quad q_{0i} + q_{1i} = 1 \]

For the \(i^{th}\) fish:

- The goal is to estimate \(q_{0i}\) for the \(i^{th}\) fish \((i = 1, \ldots, n_m)\),
  the posterior probability of population of origin for each gene copy, and \(\theta\).
Latent Data, \( w \), for the Admixture Model

- The \( t^{th} \) gene copy in the \( i^{th} \) fish gets \( w_{it} = 0 \) or \( 1 \), \( (t = 1, \ldots, G) \)

Knowing \( w \) and \( q \) renders simple the calculation of many quantities;

Particularly the full conditionals.
Gibbs Sampling in the Admixture Model

- Perform a Gibbs sampling sweep as follows, updating from the full conditional distributions:
  - Update all the $q_{0i}$'s, $(i = 1, \ldots, n_m)$
  - Update $\theta$
  - Update all the $w_{it}$'s, $(i = 1, \ldots, n_m; t = 1, \ldots, G)$

- $(q_{0i} \mid \cdots) = (q_{0i} \mid w_i, \alpha) \sim \text{Beta}(\alpha + \#\{w_i = 0\}, \alpha + \#\{w_i = 1\})$

- Allele frequency full conditionals remain simple Dirichlet distributions

- Furthermore for the $i^{th}$ fish, knowing $q_{0i}$ makes it easy to compute the full conditional for each $w_{it}$:

$$P(w_{it} = 0 \mid \cdots) = P(w_{it} = 0 \mid y, \theta, q) = \frac{q_{0i}P(y_{it} \mid \theta_0)}{q_{0i}P(y_{it} \mid \theta_0) + q_{1i}P(y_{it} \mid \theta_1)}$$
What The Admixture Model Delivers

Subject to the assumptions made in choosing the prior distributions for $\theta$ and $q$ (recall that the prior on $q$ is determined by $\alpha$) this gets you:

- Estimates of the posterior probability of population of origin of each copy of each gene within each fish in the sample of unmarked fish.

- Estimates of the posterior distribution of allele frequencies of the wild and hatchery populations

- An estimate of the posterior distribution of $q_{0i}$ for the $i^{th}$ fish

This is great for populations which have been interbreeding for a long time. However, this is not entirely appropriate if many individuals in the mixture are of pure descent (i.e., non-hybrids).
Clearly we desire a probability model in which some fish may be either “pure” or ”admixed”

Conceptually this is easy. Declare:

- Each fish is of pure origin with unknown probability $\xi_P$ or admixed with unknown probability $\xi_A$. \hspace{1cm} ($\xi_P + \xi_A = 1$)

The probability of the $i^{th}$ fish’s genotype is a weighted average of its probability under the mixture and admixture models.

- Simulation from the posterior distribution is again possible with Gibbs sampling

- Requires one final piece of latent data, $v_i$, for the $i^{th}$ fish.
Gibbs Sampling in the Merged Model

**Pure Descent**

\[ v_i = P \]

Additional latent datum is \( z_i \)

\[ P(v_i = P|\cdots) = \frac{\xi_P P(y_i|v_i = P, \cdots)}{\xi_P P(y_i|v_i = P, \cdots) + \xi_A P(y_i|v_i = A, \cdots)} \]

**Admixed Descent**

\[ v_i = A \]

Additional latent data are \( w_{it} \)'s and \( q_{0i} \)

Additional latent datum is \( z_i \)
**Updating $v_i$ Correctly**

Imagine a model without the latent data. Then:

$$ P(v_i = A|y_i, \pi, \theta, \alpha, \xi) = \frac{\xi_A P(y_i|\theta, \alpha, v_i = A)}{\xi_A P(y_i|\theta, \alpha, v_i = A) + \xi_P P(y_i|\theta, \pi, v_i = P)} $$

With latent data, we may update $v_i$ and the “additional latent data” (ALD) jointly from the joint full conditional by the following factorization:

$$ P(v_i, ALD|\cdots) = P(v_i|y_i, \pi, \theta, \alpha, \xi) P(ALD|y_i, \pi, \theta, \alpha, \xi, v_i) $$

The main intricacy lies in computing $P(y_i|\pi, \theta, \alpha, v_i)$ which is a marginal distribution achieved by integrating and/or summing out the additional latent data (either $q_{0i}$ and the $w_{it}$'s or $z_i$).
Integrating Out the Latent Data

- Easy for $v_i = P$

\[
P(y_i|\pi, \theta, v_i = P) = \sum_{z \in \{0,1\}} P(y_i, z|\pi, \theta, v_i = P)
= \sum_{j=0}^{1} P(y_i|\theta, v_i = P, z_i = j)P(z_i = j|\pi)
\]

- More difficult for $v_i = A$

\[
P(y_i|\theta, \alpha, v_i = A) = \int_0^1 \sum_w P(y_i|\theta, w_i)P(w_i|q_0)P(q_0|\alpha)dq_0
\]

But there is a neat way to compute this using some facts about the beta-binomial distribution, Pólya-Eggenberger urn schemes, and hidden Markov chains.
An Urn Scheme Interpretation

- First imagine that the genotypes are not known, and consider the distribution of the $w_{it}$’s given only $\alpha$.

- Recall that $w_{it}|q_{0i} \sim \text{Bernoulli}(q_{0i})$ and that $q_{0i} \sim \text{Beta}(\alpha, \alpha)$ so

$$P(w_i|\alpha) = \int_0^1 q_{0i}^{\alpha-1}(1 - q_{0i})^{\alpha-1} \prod_{t=1}^{G} q_{0i}^{w_{it}} (1 - q_{0i})^{1-w_{it}} dq_{0i}$$

- For illustration, take $\alpha = 1$

$$\frac{1}{2} \times \frac{1}{3} \times \frac{2}{4} \times \frac{2}{5} \times \frac{3}{6} \times \frac{3}{7} \times \frac{4}{8} \times \frac{5}{9}$$

- For illustration, take $\alpha = 1$

$$\begin{array}{cccc}
0 & 1 & 1 & 1 \\
\text{0} & \text{1} & \text{0} & \text{1} \\
1 & 2 & 3 & 4 & 5 \\
2 & 3 & 4 & 5 & 6 & 7 & 8 & 9
\end{array}$$
A Markov Chain Interpretation

• Sampling $w_{it}$’s from this urn scheme defines a time inhomogeneous Markov chain on the pairs $(w_{it}, d_t)$

  \[(w_{i1}, d_1) \rightarrow (w_{i2}, d_2) \rightarrow (w_{i3}, d_3) \rightarrow \cdots \rightarrow (w_{i7}, d_7) \rightarrow (w_{i8}, d_8)\]

• Ultimately, though, we must include the data $y_i$.

• This gives us a hidden Markov chain

  \[
  \begin{align*}
    (w_{i1}, d_1) & \rightarrow (w_{i2}, d_2) \rightarrow (w_{i3}, d_3) \rightarrow \cdots \rightarrow (w_{i7}, d_7) \rightarrow (w_{i8}, d_8) \\
    \downarrow & \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\
    y_{i1} & \quad y_{i2} \quad y_{i3} \quad \cdots \quad y_{i7} \quad y_{i8}
  \end{align*}
  \]

• The Baum et al. family of algorithms provides an efficient way to sum out the $w_{it}$’s, giving us $P(y_i|\theta, \alpha, v_i = A)$, as desired.
Results

- Dataset kindly provided by Paul Moran (Nat’l Marine Fisheries Service):
  - 49 Marked hatchery fish (learning sample)
  - 49 Unmarked fish (sample from the mixture)
  - 20 loci (with between 2 and 16 alleles)

- Found no strong evidence for the existence of two separate populations in the mixture. Either:
  - Putative wild population has allele frequencies similar to hatchery’s
  - Putative wild population no longer exists

- Fortunately the methods/software are applicable to situations in many different species
Scottish Wildcat Data and Analysis

• Data kindly provided to me by Mark Beaumont (Reading, UK)
  – 239 “Wild-living” cats
  – 65 Domestic cats collected at veterinary centers
  – 9 Moderately polymorphic microsatellite loci

• Analysis
  – Assumed we had no learning sample
  – Uniform priors on $\xi$, $\pi$, and $q$ (i.e., $\alpha = 1$)
  – 10,000 Sweeps of burn-in
  – 100,000 Sweeps for inference (took about 2 hours)
Mixing and Inference for $\xi$

![Graph showing realized value of Xi_P over sweep number and posterior probability distributions for Xi_A and Xi_P.](image-url)
Individual Posterior Probabilities of Type of Descent

\[ P(v_i = P|y) \]
\[ P(v_i = P|y) \text{ vs. } P(z_i = 0|y, v_i = P) \]
$P(v_i = P|y)$ vs. $P(z_i = 0|y, v_i = P)$

(Zoomed In and Jiggled)
Posterior Probability of First-Generation Hybrids

$P(F_1 \text{ Hybrid} \mid y)$

Average $q_0 \mid \text{Admixed}, y$
MCMC and Gibbs Sampling

- MCMC is just like standard Monte Carlo except that the simulated values are states visited by a Markov chain.

- Requires knowing the target distribution “up to a normalizing constant”

- Gibbs Sampling: constructing multi-dimensional Markov chains
  - Successively update individual components/variables of a probability model conditional on the current state of all remaining variables in the model.
  - Full conditional distributions: $P(X|\cdots)$
  - Works best when the full conditionals are quickly computed.
Summary

- Models of genetic mixture and genetic admixture
- Model for “simultaneous” mixture and admixture
- Bayesian inference via MCMC (Gibbs sampling)
- Enjoyable full conditional calculations for the “merged model”
- Salmon data: a fine demonstration that there must be genetic differences between the populations in question if one wishes to separate individuals.
- Scottish wildcat data: admirable performance of the method.
  - Extension to more specific models of admixture.