

Analyzing Genetic Mixtures using Reversible Jump Markov Chain Monte Carlo

Eric C. Anderson

*Interdisciplinary Program in Quantitative
Ecology and Resource Management*

University of Washington, Seattle, WA

Overview

§1 A motivating problem from salmon fisheries management

§2 A toy problem with the essential elements

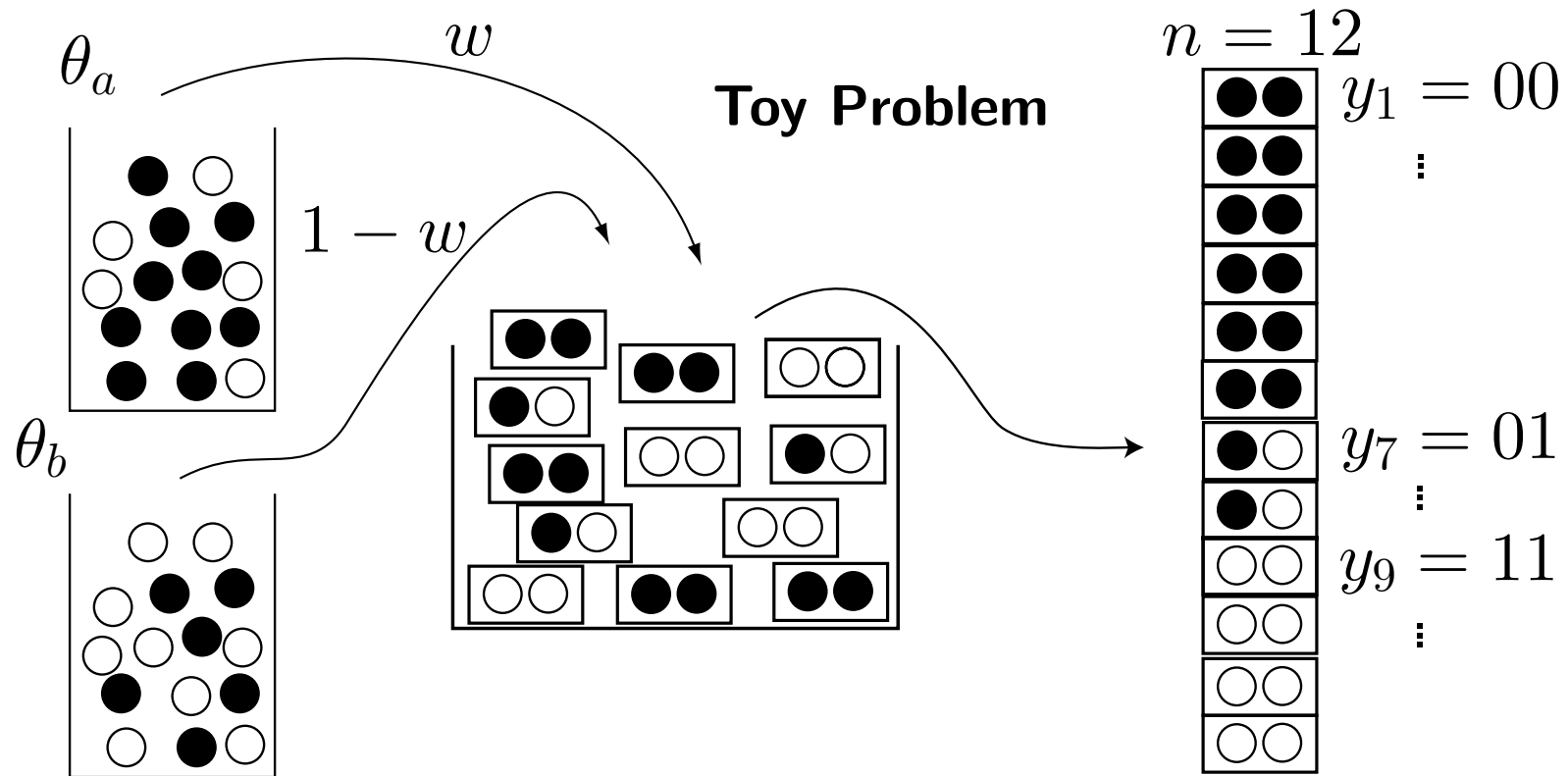
§3 Background on MCMC and RJMCMC

§4 Extension beyond the toy problem

- The problems
- Some solutions

§5 “Results”

§6 Further extensions and generalizations



●●	$p(y_i = 00) = w\theta_a^2 + (1-w)\theta_b^2$])
●○	$p(y_i = 01) = p(y_1 = 10) = 2w\theta_a(1-\theta_a) + 2(1-w)\theta_b(1-\theta_b)$		
○○	$p(y_1 = 11) = w(1-\theta_a)^2 + (1-w)(1-\theta_b)^2$		

Missing Data Specification

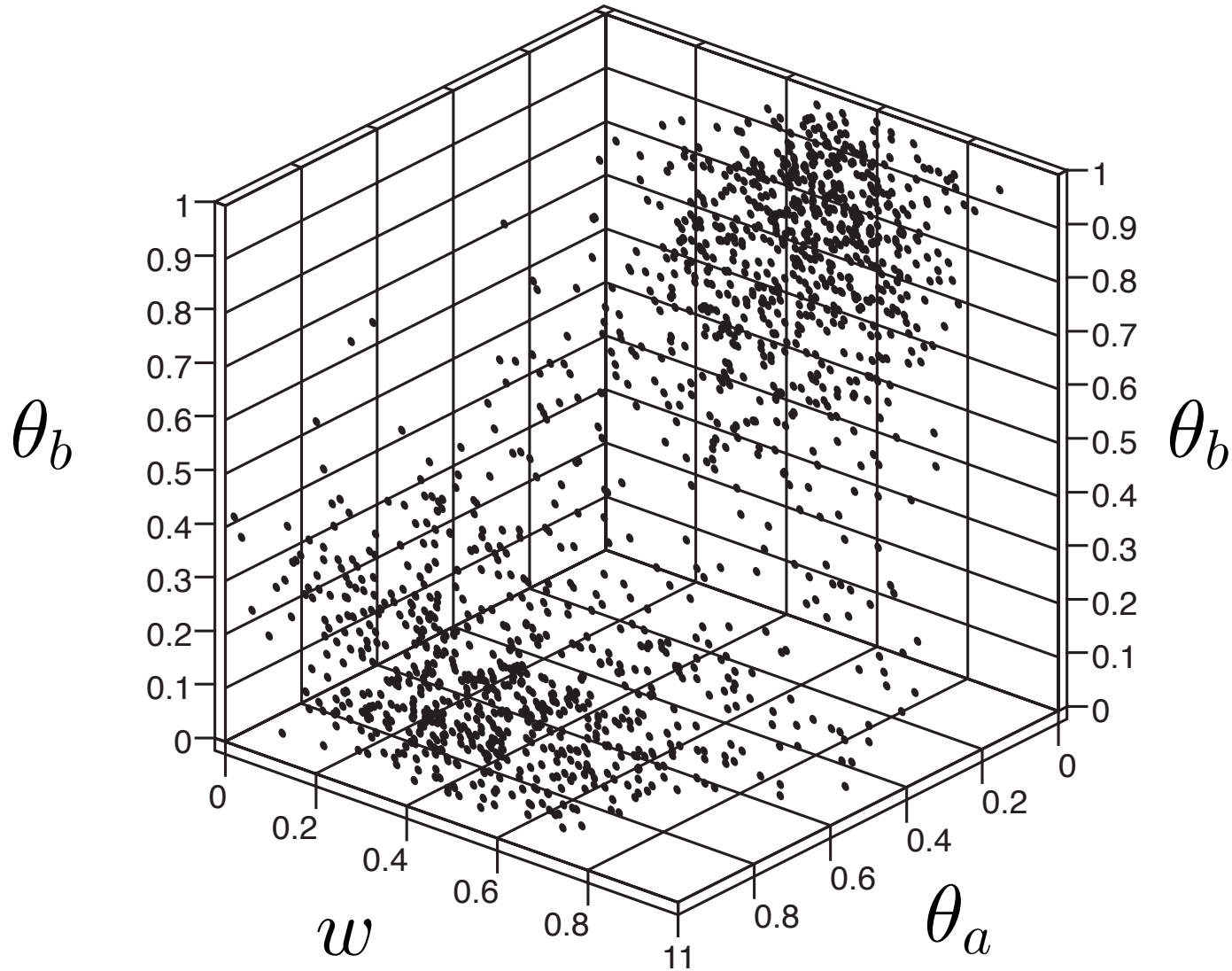
- Attach to each pair i an allocation variable z_i .
- Denotes the unknown bucket of origin $z_i = a$ or $z_i = b$
- Write $z = (z_1, \dots, z_{12})$ and $y = (y_1, \dots, y_{12})$

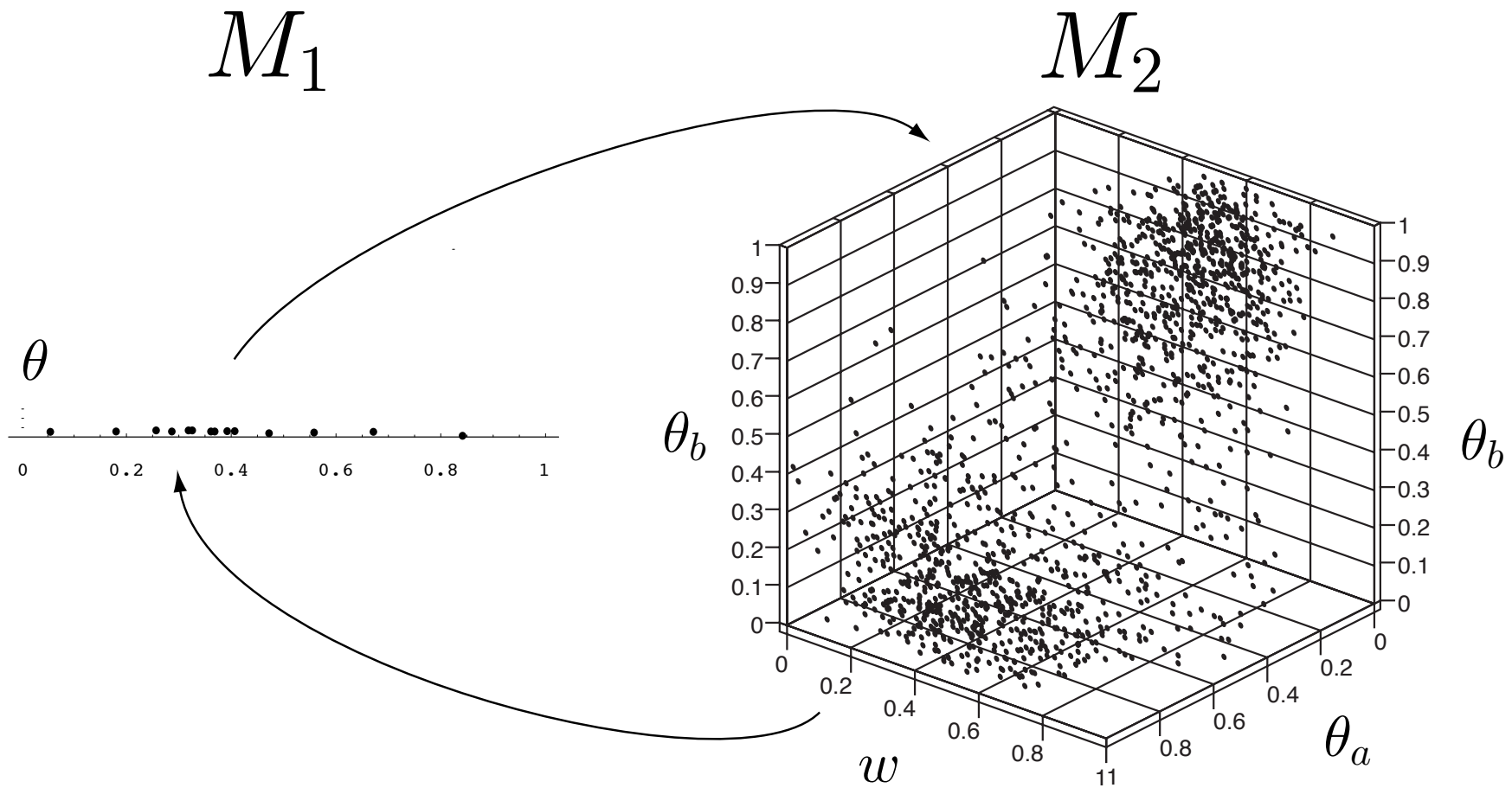
The joint density factorizes nicely,

$$p(w, \theta_a, \theta_b, z, y) = p(w)p(z|w)p(\theta_a, \theta_b)p(y|\theta_a, \theta_b, z).$$

- Facilitates Gibbs sampling:
 - Update z from $\propto p(z|w)p(y|\theta_a, \theta_b, z)$
 - Update w from $\propto p(w)p(z|w)$
 - Update θ_a, θ_b from $\propto p(\theta_a, \theta_b)p(y|\theta_a, \theta_b, z)$

**1500 realizations from a Markov chain with limit
distribution $p(w, \theta_a, \theta_b|y)$**





Conceptually we can think of a sort of Hastings' Ratio

$$\alpha = \min \text{ of } 1 \text{ or } \frac{q(\theta_a^*, \theta_b^*, w^*; \theta)p(\theta_a^*, \theta_b^*, w^*, y)}{q(\theta^*; \theta_a, \theta_b, w)p(\theta^*, y)}$$

but there are two main difficulties. . .

Reversible Jump MCMC

(Green 1995; Richardson and Green 1997)

- Provides a way to construct a Markov chain which has a limit distribution over regions of different dimensionality.
- Dimension-changing moves have two flavors:
 - One which jumps the chain to higher-dimensional space
 - The other which jumps the chain to lower-dimensional space
- Reversibility requirement
- Jumps to lower dimensional space via a many-to-one, deterministic map
- Jumps “up” draw some extra variables and uses them in a one-to-one “inverse” map to reach a point in the higher-dimensional space.

Split-Combine Proposals

In the Toy Problem

- $M_2 \rightarrow M_1$: deterministic “combine” proposal from $(w, \theta_a, \theta_b, z) \rightarrow \theta^*$
- May use the intuitively reasonable weighted average

$$\theta^* = w\theta_a + (1 - w)\theta_b$$

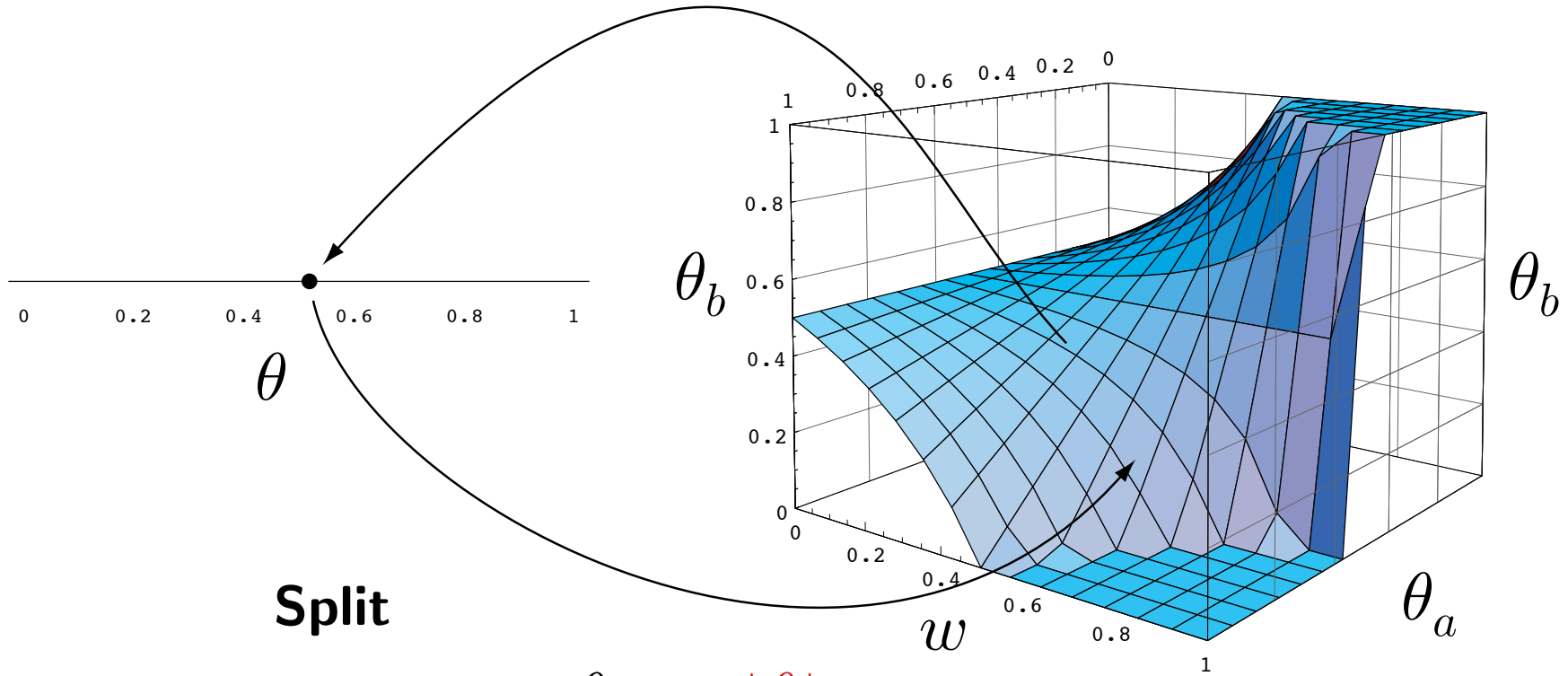
- $M_1 \rightarrow M_2$ uses a non-deterministic “split” proposal from $\theta \rightarrow (w^*, \theta_a^*, \theta_b^*, z^*)$ carried out as follows:
 - Draw u_w from $f_{u_w}(\cdot)$; Draw u_θ from $f_{u_\theta}(\cdot)$ then use the 1 : 1 map $g : (u_w, u_\theta, \theta) \rightarrow (w^*, \theta_a^*, \theta_b^*)$ defined by

$$w^* = u_w \quad \theta_a^* = u_\theta \quad \theta_b^* = \frac{\theta - w^*\theta_a^*}{1 - w^*}$$

- Propose z^* from $\propto p(z^*|w^*)p(y|\theta_a^*, \theta_b^*, z^*)$

$$\theta^* = w\theta_a + (1 - w)\theta_b$$

Combine



Split

$$\theta_b^* = \frac{\theta - w^* \theta_a^*}{1 - w^*}$$

The Acceptance Probability

- Think of it in terms of proposing a *split* move.
- Then the acceptance probability is $\min\{1, A\}$ where

$$A = \frac{p(\text{choosing to combine given the state in } \uparrow \text{ Dim space})}{p(\text{choosing to split given the state in } \downarrow \text{ Dim space})}$$
$$\times \frac{1}{f_{u_w}(u_w) f_{u_\theta}(u_\theta)} \times \frac{p(w^*, \theta_a^*, \theta_b^*, z^*, y, M_2)}{p(\theta, y, M_1)} \times \frac{1}{p(z^* | w^*, \theta_a^*, \theta_b^*)}$$
$$\times \left| \frac{\partial g(u_w, u_\theta, \theta)}{\partial (u_w, u_\theta, \theta)} \right|$$

- For a combine move, the acceptance probability is $\min\{1, A^{-1}\}$.

Posterior Probability for M_1 and M_2 in our toy problem

- Assume
 - Equal priors on model: $p(M_1) = p(M_2) = .5$
 - Uniform prior on θ under M_1
 - Independent uniform priors on w, θ_a, θ_b under M_2
- Take $f_{u_w}(\cdot)$ and $f_{u_\theta}(\cdot)$ to be uniform.
- The update cycle is then:
 - Update θ if in M_1 —OR—
 - Update z , then w , then θ_a and θ_b if in M_2
 - THEN: Propose a split if in M_1 and a combine if in M_2 .
- From the 12 pairs in our simple problem we get:

$$p(M_1|y) = .156 \quad p(M_2|y) = .844$$

which accords with the result from numerically integrating it.

Inference for Individual Components

- Likelihood is invariant to permutations of the component labels
- Traditionally managed by imposing a constraint
 - R & G: ordering by w , μ , σ^2 , or some combination thereof
 - Used adjacency criterion for proposing split/combine moves
- In genetic mixtures, there is no obviously useful ordering in the parameter space
 - Imposing one would probably slow mixing
- M. Stephens' methods seem much more attractive.

Extension to Multiple Loci and Alleles —The Jacobian—

$$\begin{array}{l}
 w_a \\
 w_b \\
 \vec{\theta}_{a1} \\
 \vdots \\
 \vec{\theta}_{ah} \\
 \vec{\theta}_{b1} \\
 \vdots \\
 \vec{\theta}_{bh}
 \end{array}
 \begin{pmatrix}
 w_c & u_w & \vec{\theta}_{c1} & \cdots & \vec{\theta}_{ch} & \vec{u}_{\theta_1} & \cdots & \vec{u}_{\theta_h} \\
 u_w & w_c & 0 & 0 & 0 & 0 & 0 & 0 \\
 1 - u_w & -w_a & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
 \vdots & \vdots & \vdots & \vdots & \vdots & 0 & 1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
 \frac{\partial \vec{\theta}_{b1}}{\partial w_c} & \frac{\partial \vec{\theta}_{b1}}{\partial u_w} & \text{Dg}\left(\frac{1}{1-u_w}\right) & 0 & 0 & \text{Dg}\left(\frac{u_w}{1-u_w}\right) & 0 & 0 \\
 \vdots & \vdots & 0 & \text{Dg}\left(\frac{1}{1-u_w}\right) & 0 & 0 & \text{Dg}\left(\frac{u_w}{1-u_w}\right) & 0 \\
 \frac{\partial \vec{\theta}_{bh}}{\partial w_c} & \frac{\partial \vec{\theta}_{bh}}{\partial u_w} & 0 & 0 & \text{Dg}\left(\frac{1}{1-u_w}\right) & 0 & 0 & \text{Dg}\left(\frac{u_w}{1-u_w}\right)
 \end{pmatrix}$$

For the “intuitive” lumping transformation, the Jacobian always takes the simple form

$$w_c \prod_{\ell=1}^h \left(\frac{1}{1-u_w} \right)^{K_\ell - 1}$$

where h is the number of loci and K_ℓ is the number of alleles at the ℓ^{th} locus

Extension to Multiple Loci and Alleles

Irreversible Split Proposals

- With w determined by u_w from f_{u_w} and θ_a^* from f_{θ_a} , it is possible that

$$\theta_b^* = \frac{\theta - w^* \theta_a^*}{1 - w^*}$$

will be less than zero or greater than one.

- Such split proposals must be rejected without further consideration
- Irreversible proposals are more probable with many alleles at low frequency

Finding Sensible Proposal Distributions, f_{u_w} and f_{θ_a}

- Dirichlet family is most reasonable, yet our task becomes finding suitable parameters for those distributions.
- In choosing those parameters “anything goes” so long as the selection process is deterministic
 - Balance between computational effort and quality of proposal
 - Ugly problem with multivariate data
- One thing I’ve tried: use a fixed number (say ν) of steps of an EM algorithm to get oneself in the vicinity of posterior modes for a proposed split of a particular component
 - (EM: Slow later convergence but “rapid” initial convergence)
- Imagine a component labelled c which we have chosen to try to split it into two components a and b , as before. Think of diallelic loci for now.

- Take all the individuals currently allocated to c and imagine that they are a single sample that is drawn from a mixture of two components a and b
- We could implement an EM algorithm to find \hat{w} , $\hat{\theta}_a$, and $\hat{\theta}_b$.

- Recall that θ_b is determined. There is a constraint,

$$\theta_b = \frac{\theta_c - w\theta_a}{1 - w},$$

which one can impose in the M -step.

- Solving the exact likelihood equations with that constraint is messy, but an approximation is available. . .

- If you had a sample from population b from which you determined $\hat{\theta}_b$ you could have just as well have used it to find $(\hat{\theta}_a)_b = [\theta_c - (1 - w)\hat{\theta}_b]/w$
- With a different sample from a you could find $\hat{\theta}_a$ and then propose a weighted average of $\hat{\theta}_a$ and $(\hat{\theta}_a)_b$ as an estimate of θ_a using both samples.

- For example, at the j th iteration of the EM algorithm

1. Find $\theta_a^{(j)}$, $\theta_b^{(j)}$, $w^{(j)}$, by a standard M -step.
2. Set

$$(\theta_a^{(j)})_b \leftarrow \frac{\theta_c - (1 - w^{(j)})\theta_b^{(j)}}{w^{(j)}}$$

3. Revise $\theta_a^{(j)}$ as an inverse-variance weighted average:

$$\theta_a^{(j)} \leftarrow \frac{\theta_a^{(j)} V_b + \theta_b^{(j)} V_a}{V_a + V_b}$$

where

$$V_a = \frac{\theta_a^{(j)}(1 - \theta_a^{(j)})}{w^{(j)}} \quad \text{and} \quad V_b = \frac{[1 - w^{(j)}]^2 \theta_b^{(j)}(1 - \theta_b^{(j)})}{(w^{(j)})^2}$$

4. Set $\theta_b^{(j)} \leftarrow \frac{\theta_c - w^{(j)}\theta_a^{(j)}}{1 - w^{(j)}}$
5. Proceed to $j + 1^{\text{st}}$ iteration.

- At the end of ν such cycles, determine parameters for f_{u_w} which will be $\text{Beta}(\beta_1, \beta_2)$ where

$$\beta_1 = \alpha + \phi_w \sum_{\{z_i: z_i=c\}} P(z_i = a | w^{(\nu)}, \theta_a^{(\nu)}, \theta_b^{(\nu)})$$

$$\beta_2 = \alpha + \phi_w \sum_{\{z_i: z_i=c\}} P(z_i = b | w^{(\nu)}, \theta_a^{(\nu)}, \theta_b^{(\nu)})$$

Where α is related to the prior for w and ϕ_w weights how much you want to emphasize the results of the ν cycles of EM. (as $\phi \rightarrow 0$ you end up drawing u_w from what would be reasonable given the prior on w .)

- Parameters for f_{u_θ} determined similarly. Let d_{ij} be the number of alleles of type j at a locus in individual i , then for $\text{Dirichlet}(\beta_1, \dots, \beta_{K-1})$

$$\beta_j = \alpha + \phi_\theta \sum_{\{z_i: z_i=c\}} d_{ij} P(z_i = a | w^{(\nu)}, \theta_a^{(\nu)}, \theta_b^{(\nu)})$$