Bayesian Analysis of Population-Genetic Mixture and Admixture

Eric C. Anderson
Interdisciplinary Program in Quantitative
Ecology and Resource Management
University of Washington, Seattle, WA 98195
email: eriq@cqs.washington.edu

November 5, 2015

Abstract

This file, “JASA_example.tex,” is just excerpted from a manuscript I wrote in the summer of 2000. The content will appear totally discontinuous, but the use of various sectioning commands and citation commands should be evident from it, and any other user should be able to use it as a template .tex file for using JASA_manu.sty. Biologists regularly encounter populations of organisms with disparate ancestries. Untangling the composition of such populations is a problem for conservation biologists and wildlife managers. In many cases the population under question is known to consist of individuals from two different subpopulations and their hybrids. This occurs, for example, in hybrid zones between two species or in regions recently colonized by exotics capable of reproducing with resident inhabitants. This paper develops techniques using multilocus genetic data for Bayesian clustering of individuals to purebred or genetically-mixed categories. The method relies on a novel application of the forward-backward recursions in a two-component, finite mixture model. Though developed in the context of the genetic admixture problem, these calculations are relevant more generally to Bayesian inference in finite mixtures; they may potentially improve mixing of the Gibbs sampler in such contexts. The technique is applied to genetic data on the Scottish wildcat, *Felis sylvestris*, a protected species whose distinctness from domestic housecats has been questioned. A high proportion ($\approx .60$) of the wild-living cats from which the sample was drawn are arguably purebred *F. sylvestris*.

Using the Bayes factor, we compare our new model, which allows for both purebred and admixed individuals, to a model in which all individuals are assumed genetically admixed to some degree. It is difficult to accurately compute the marginal likelihood directly in these models, so we compute the Bayes factor by reversible-jump MCMC. The approach follows from the original MCMC formulation of the problem, and should help to illustrate ways in which reversible-jump methods may be implemented for comparisons between a small set of closely-related models.

KEYWORDS: Forward-backward recursion, Gibbs sampler, reversible jump, MCMC, hybrid zone

1. INTRODUCTION

Pritchard, Stephens and Donnelly (2000a) propose a versatile model for genetic inheritance in admixed populations and use it in Bayesian analyses of population structure in several different species. Before I go any further, let's look at a "do-it-by-hand" citation/bibliography entry that was defined above: (Raftery 1992). A limitation of this model, however, is that it assumes every individual is admixed to some degree. In many situations, such as with populations spanning hybrid zones, there is reason to expect both purebred and admixed individuals. A probability model to accommodate such scenarios will include elements both of genetic mixture models and genetic admixture models. In this paper I extend the methods of Pritchard et al. (2000a) to handle explicitly purebred individuals. In section 2, I review mixture and admixture formulations for modeling population structure.

In Section 3, I develop a method for making joint, Gibbs updates of large blocks of variables in Pritchard et al.'s (2000a) model. The method uses the fact that the latent allocation variables of an i.i.d. finite mixture, with a Dirichlet prior on mixing proportions can be shown to follow a hidden Markov chain, after integrating out the mixing proportions. This computation facilitates MCMC simulation in a model, described later, that allows for both purebred and admixed individuals. Additionally, I describe in the Discussion how such a method could help the Gibbs sampler to escape from trapping states (Robert 1996) encountered in finite mixture problems.

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

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

2. A MODEL WITH ADMIXED INDIVIDUALS

With θ and y defined as in the previous section, the model of Pritchard et al. (2000a) is quickly described. Now, j indexes the J conceptual "gene pools" or "historical subpopulations" from which individuals may be descended. Allowing for admixed individuals requires a different model of genetic inheritance, which, in turn, requires different latent variables. The i^{th} individual in the sample gets a vector of probabilities $q_i = (q_{i1}, \ldots, q_{iJ})$, $\sum_{j=1}^{J} q_{ij} = 1$, which are the unobserved proportions of that individual's genome descended from each of the J gene pools. Also, let $w_i = (w_{i1}, \ldots, w_{i2L})$ be a vector of unobserved allocation variables which is parallel to the the vector of allelic types y_i. Hence, $w_{it} = j$ indicates that the $(t \text{ [mod 2]} + 1)^{\text{th}}$ allele at the $\lceil t/2 \rceil^{\text{th}}$ locus in the i^{th} individual is from the j^{th} gene pool. Given $w_{it} = j$ the type of allele is assumed to be drawn randomly according to θ_j. Under this model

$$p(y_i|\theta, w_i) = \prod_{t=1}^{2L} \theta\langle w_{it}; y_{it}\rangle \tag{1}$$

independently for each i. By assigning the prior $q_i \sim \text{Dir}(\alpha, \ldots, \alpha)$, $i = 1, \ldots, N$, and the hyperprior $\alpha \sim \text{Uniform}(0, A]$, Pritchard et al. (2000a)'s model is obtained. In effect this is a hierarchical model for N different finite mixtures—the genes carried by the i^{th} individual are a sample from a mixture with mixing proportions given by q_i, while the q_i themselves $(i = 1, \ldots, N)$ are drawn from a symmetrical $\text{Dir}(\alpha, \ldots, \alpha)$ distribution.

In this model, Gibbs sampling proceeds using the full conditionals

$$\begin{aligned} q_i|\cdots &\sim \text{Dir}(\alpha_1 + \#\{w_i = 1\}, \ldots, \alpha_J + \#\{w_i = J\}), \quad i = 1, \ldots, N \\ \theta_{j\ell}|\cdots &\sim \text{Dir}(\lambda_{j\ell 1} + r_{j\ell 1}, \ldots, \lambda_{j\ell K_\ell} + r_{j\ell K_\ell}), \\ & \qquad j = 1, \ldots, J; \quad \ell = 1, \ldots, L \end{aligned}$$

$$p(w_{it}=j|\cdots) = \frac{q_{ij}\theta\langle j; y_{it}\rangle}{\sum_{k=1}^{J} q_{ij}\theta\langle k; y_{it}\rangle}\ ,\ i=1,\ldots,N;\ j=1,\ldots,J; \quad t=1,\ldots,2L$$

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

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

2.1 Block-updating w_i when $J=2$

In many situations involving invasions of exotic species, there is substantial prior knowledge that the number of major subpopulations or "gene pools" involved is two—the native population and the invading population. Additionally, many hybrid zones are known to be areas of hybridization (admixture) between two species or populations. Here I present novel computations, feasible when only two subpopulations or gene pools are involved, that eliminate the explicit need for the variable $q=(q_1,\ldots,q_N)$ in implementing a Gibbs sampler. Such a method slightly improves mixing of the chain, but is primarily useful as it makes possible Gibbs sampling in a simultaneous mixture and admixture analysis as will be described in Section Yippie!.

The computations themselves may be derived as follows. Let $J=2$, so that each allele in an individual may have originated from gene pool 1 or gene pool 2. Then, each q_{i1} will follow a Beta(α,α) distribution and $q_{i2}=1-q_{i1}$. Conditional on q_{i1}, each w_{it} will then be independently a Bernoulli trial with $p(w_{it}=1|q_{i1})=q_{i1}$. Marginalizing over q_{i1} (not conditioning on the data) it follows that $\#\{w_i=1\}$ follows a beta-binomial distribution with parameters (α,α). Of course, each allele in an individual is uniquely labelled so the elements of w_i may be interpreted as following a *labelled* beta-binomial distribution. Under such a distribution, the elements of

w_i are not independent, but they are exchangeable (deFinetti 1972), and hence their marginal distributions are invariant to permutations of their order (and thus the arbitrary order we have imposed upon them is acceptable).

This labelled beta-binomial sampling mechanism is easily visualized by a Pólya-Eggenberger urn scheme (Feller 1957; Johnson, Kotz and Kemp 1993). Imagine an urn initially filled with b_1 balls labelled "1" and b_2 balls labelled "2." Draw a ball randomly and record $w_{i1} = 1$ or 2 according to the ball's label. Then replace the ball to the urn, adding, at the same time, c more balls of the same type (1 or 2) as the ball just drawn. Repeat the process, assigning a value to w_{i2} and so forth until w_{i2L} has also been assigned a 1 or 2. If b_1, b_2, and c were chosen to satisfy $b_1/c = b_2/c = \alpha$, then the resulting vector w_i would be a realized value from the labelled beta-binomial distribution with parameters (α, α). (One should notice, also, that this extends to a non-symmetrical beta distribution, say $\text{Beta}(\alpha_1, \alpha_2)$, by choosing $b_1/c = \alpha_1$ and $b_2/c = \alpha_2$.)

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

$$\frac{b_1 + d_t c}{b_1 + b_2 + tc}. \tag{2}$$

And so the pairs (w_{it}, d_t), $t = 1, \ldots, 2L$, can be interpreted as forming a time-inhomogeneous Markov chain in time t with transition probabilities determined by (2) and the obvious fact that $d_{t+1} = d_t + 1\{w_{it+1} = 1\}$, where $1\{x = a\}$ takes the value one when $x = a$ and zero otherwise.

The foregoing has all been considered in the absence of data, y_i. However, given θ, the data provide some information about the true value of each w_{it} by the relation $p(y_{it}|w_{it}, \theta) = \theta\langle w_{it}; y_{it}\rangle$. Therefore, conditional on θ and y_{it}, the pairs (w_{it}, d_t) participate in a *hidden* Markov chain. Recognition of this fact allows application of a "filter-forward, simulate-backward" type of algorithm which may be derived following the computations of Baum, Petrie, Soules and Weiss (1970) in order to realize the elements of w_i from their joint full conditional distribution, $p(w_i|\alpha, \theta, y_i)$. Furthermore, using the Baum (1972) algorithm, it is possible to compute $p(y_i|\alpha, \theta)$, effectively performing a sum over all possible binary vectors of length $2L$ in an efficient manner. This is described below.

Take b_1, b_2, and c as defined above. Suppressing the i subscript for clarity, let $w_t \in \{1, 2\}$, $t = 1, \ldots, 2L$, and define $d_t = \sum_{\tau=1}^{t} 1\{w_\tau = 1\}$. We adopt the notation $w_{\leq t}$ $(w_{\geq t})$ to mean

$w_1, \ldots, w_t$ ($w_t, \ldots, w_{2L}$) for components of w, and use the same notation with y and d. The pairs (w_t, d_t) can be interpreted as following a Markov chain in t:

$$\begin{aligned} p(w_{t+1}, d_{t+1}|w_{\leq t}, d_{\leq t}) &= p(w_{t+1}, d_{t+1}|w_t, d_t) \\ &= \frac{b_1 + d_t c}{b_1 + b_2 + tc} 1\{d_{t+1} = d_t + 1\{w_{t+1} = 1\}\}. \end{aligned}$$

The "perturbed" or "degraded" observations of the chain are the allelic types $y_1, \ldots, y_{2L}$ which depend in hidden Markov fashion on w. For notational clarity, we assume implicit dependence on the allele frequencies θ,

$$p(y_t|w_{\leq 2L}, d_{\leq 2L}) = p(y_t|w_t) = \theta\langle w_t; y_t\rangle.$$

This dependence structure is shown in the undirected graph of Figure 1.

[Figure 1 about here.]

ACKNOWLEDGMENTS

This work developed out of conversations during a brief visit to University of Oxford, where the author was hosted by the lab of Peter Donnelly in the Department of Statistics. Partial support came from National Science Foundation Grant BIR–9807747. The author thanks Jonathan Pritchard, Elizabeth Thompson, and Matthew Stephens for helpful discussion and comments on earlier drafts. Special thanks to Mark Beaumont for providing the data on Scottish wildcats.

APPENDIX A. METROPOLIS UPDATES FOR α

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

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

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

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

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

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

When able to eliminate the q_i's (as in Section 2.1), then with only admixed individuals (model M_{A}) the acceptance probability may be written as

$$\min\left\{1, \frac{\prod_{i=1}^N p(y_i|\alpha^*, \theta)}{\prod_{i=1}^N p(y_i|\alpha, \theta)}\right\}. \tag{A.2}$$

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

$$\min\left\{1, \frac{\prod_{i=1}^N [\xi_{\mathrm{P}} p(y_i|\pi, \theta) + \xi_{\mathrm{A}} p(y_i|\alpha^*, \theta)]}{\prod_{i=1}^N [\xi_{\mathrm{P}} p(y_i|\pi, \theta) + \xi_{\mathrm{A}} p(y_i|\alpha, \theta)]}\right\}. \tag{A.3}$$

REFERENCES

Baum, L. E. (1972), "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," in *Inequalities–III: Proceedings of the Third Symposium on Inequalities Held at the University of California, Los Angeles, September 1–9, 1969*, ed. O. Shisha, New York: Academic Press, pp. 1–8.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), "A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains," *Annals of Mathematical Statistics*, 41, 164–171.

Beaumont, M., Gotelli, D., Barratt, E. M., Kitchener, A. C., Daniels, M. J., Pritchard, J. K., and Bruford, M. W. (in press), "Genetic diversity and introgression in the Scottish wildcat," , .

deFinetti, B. (1972), *Probability, Induction and Statistics. The Art of Guessing*, New York: John Wiley & Sons.

Feller, W. (1957), *An Introduction to Probability Theory and Its Applications, 2nd Edition*, New York: John Wiley & Sons.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), "Model determination using predictive distributions with implementation via sampling-based methods," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 147–167.

Johnson, N. L., Kotz, Z., and Kemp, A. W. (1993), *Univariate Discrete Distributions, 2nd Edition*, New York: Wiley & Sons.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a), "Inference of Population Structure Using Multilocus Genotype Data," *Genetics*, 155, 945–959.

Raftery, A. E. (1992), Discussion of "Model determination using predictive distributions with implementation via sampling-based methods," by A. E Gelfand, D. K. Dey, and H. Chang, in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 147–167. .

Robert, C. P. (1996), "Mixture of distributions: inference and estimation," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman and Hall, pp. 441–464.

List of Figures

1 An undirected graph showing the dependence between w, d and y in Section 2.1. This graph describes hidden Markov structure for the pairs (w_t, d_t). The dependence on θ is implicit and not shown. 11

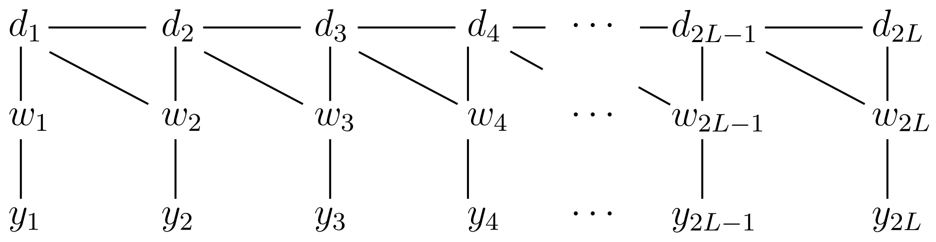

Figure 1: An undirected graph showing the dependence between w, d and y in Section 2.1. This graph describes hidden Markov structure for the pairs (w_t, d_t). The dependence on θ is implicit and not shown.