Bayesian Analysis of Population-Genetic Mixture and Admixture

Submitted to the *Journal of Agricultural Biological and Environmental Statistics*

Please Do Not Cite

Eric C. Anderson

Interdisciplinary Program in Quantitative

Ecology and Resource Management

University of Washington, Seattle, WA 98195

email: eriq@stat.washington.edu

Mailing Address:

Department of Statistics

University of Washington

Box 354322

Seattle, WA 954322

January 2, 2001

Abstract

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

Populations studied by geneticists are seldom the ideal, randomly-mating and genetically-isolated collections of individuals for which much genetic theory has been developed. In particular, natural populations may possess internal structure which prohibits random mating, or they may be recently formed by migration and co-mingling of individuals from two or more originally separate populations. Such structure complicates many types of genetic studies. For instance, when using population-level data to map genetic diseases, population structure, if not accounted for, may lead to spurious associations between genetic markers and disease status (Ewens and Spielman 1995). Additionally, in the ecological study of plants and animals there is considerable interest in population structure, especially in regions of apparent overlap and interbreeding between different subpopulations—so-called "hybrid zones." For these, and other types of problems, it is desirable to be able to infer population structure from genetic data. To this end, models of population genetic *mixture* and *admixture* have been useful. I describe the application of such models to the inference of population structure, focusing primarily on applications to hybrid zones of two different groups of individuals. Such situations are now encountered frequently as a result of anthropogenic disturbance reducing barriers to gene flow between formerly separate subpopulations. Invasions of exotic species are one pervasive example.

As used here, a "genetic mixture model" attributes structure in a population to the presence of two or more subpopulations. Within these subpopulations individuals may mate at random, but no interbreeding occurs between subpopulations. Every individual in the mixture is considered to be a purebred descendant of only one subpopulation. Such models have been developed and used extensively in the field of fisheries management (Milner, Teel, Utter and Burley 1981; Pella and Milner 1987; Smouse, Waples and Tworek 1990; Millar 1991; Pella and Masuda in press).

On the other hand, admixture, throughout the genetics literature (Cavalli-Sforza and Bodmer 1971; Thompson 1973; Wijsman 1984; Long 1991) refers to interbreeding between members from different subpopulations. Accordingly, a "genetic admixture model" attempts to model a population's genetic structure by the presence of two or more previously separate subpopulations between which there has been some recent interbreeding. Such a population is said to be admixed. Additionally we will call an individual "admixed" if it possesses genes descended from more than one of the historically separate populations. Early investigations of admixed populations sought to

estimate the relative contributions of two founding populations to the admixed population. These studies assumed the admixed population had undergone enough generations of random mating to eliminate allelic associations between loci that accompanies genetic admixture. As a result, the individual allele frequencies observed in the two founding subpopulations and the admixed population were treated as sufficient statistics. It was not until recently that statistical methods were proposed whereby the information in *multilocus* data could be used to elucidate structure in a recently admixed population (Rannala and Mountain 1997; Paetkau, Calvert, Stirling and Strobeck 1995; Pritchard, Stephens and Donnelly 2000a).

Pritchard et al. (2000a) propose a versatile model for genetic inheritance in admixed populations and use it in Bayesian analyses of population structure in several different species. 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 sections 2 and 3, 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 in Section 4, 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. GENETIC MIXTURE MODELS

The formulation of a genetic mixture model follows that for a general finite mixture. Let N diploid individuals be sampled from a population and typed at L loci. The population is assumed to consist of J subpopulations indexed by $j = 1, \ldots, J$. The proportion of individuals in the mixed population from subpopulation j is the unknown parameter π_j with $\sum_{j=1}^{J} \pi_j = 1$. Assign to each individual a latent allocation variable z_i, $i = 1, \ldots, N$. We use $z_i = j$ to indicate that the i^{th} individual is from the j^{th} subpopulation.

Denote the multilocus phenotype of the i^{th} individual by y_i. I use "phenotype" as opposed to "genotype" because the phrase "multilocus genotype" typically implies knowledge of the gametic phase of the alleles present at different loci. No such knowledge is available here. The multilocus phenotype, y_i, consists of the allelic type of each of the two alleles carried by the i^{th} individual at L loci. Since we will later identify and label specific gene copies in an individual, we consider the two alleles carried at a locus to be ordered. This order may be arbitrary. For example, it can merely be the order in which the types of those two alleles are reported in the data on an individual. Thus, y_i can be regarded as a vector of length $2L$ with its first element giving the allelic type of the first allele at locus 1, its second element giving the type of the second allele at locus 1, its third element the type of the first allele at locus 2 and so forth. In general, y_{it}, $t = 1, \ldots, 2L$, is the type of allele number $(t \pmod 2 + 1)$ at locus number $\lceil t/2 \rceil$ in individual i, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

The allele frequencies in the j^{th} subpopulation are denoted by $\theta_j = (\theta_{j1}, \ldots, \theta_{jL})$, where $\theta_{j\ell}$, $\ell = 1, \ldots, L$, is a vector of length equal to K_ℓ—the number of types of alleles observed at locus ℓ across all the individuals sampled. The frequency in the j^{th} subpopulation of the k^{th} allelic type of the ℓ^{th} locus is $\theta_{j\ell k}$, $k = 1, \ldots, K_\ell$. We will adopt the notation $\theta\langle j; y_{it}\rangle$ to mean $\theta_{j\ell k}$ where $\ell = \lceil t/2 \rceil$ and k is the allelic type of the $(t \pmod 2 + 1)^{\text{th}}$ allele at the ℓ^{th} locus in the i^{th} individual.

Given that an individual is from subpopulation j, it is assumed to have a multilocus phenotype resulting from random mating and linkage equilibrium between the L loci *within* subpopulation j. Thus,

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

where $p(\cdot|\cdot)$ will be used throughout to denote conditional probability mass or density functions. The likelihood for $\pi = (\pi_1, \ldots, \pi_J)$ and $\theta = (\theta_1, \ldots, \theta_J)$, with y denoting $(y_1, \ldots, y_N)$, is

$$p(y|\pi, \theta) = \prod_{i=1}^{N} p(y_i|\pi, \theta) = \prod_{i=1}^{N} \sum_{j=1}^{J} \pi_j p(y_i|\theta_j, z_i = j). \tag{2}$$

Note that this formulation does not include the familiar binomial coefficient, 2, for heterozygotes because we have arbitrarily ordered the two alleles carried by an individual at each locus. This makes the likelihood in this model comparable to that in the admixture model of Pritchard et al. (2000a).

"Training" or "learning" samples may be available. They might take the form of specially tagged individuals which, though sampled along with the rest of the mixture, may be unambiguously assigned to a subpopulation. Such an individual, say i^*, known to come from subpopulation j^* is easily accommodated by setting $z_{i^*} = j^*$ and defining $p(y_{i^*}|\theta_j, z_{i^*} = j) \equiv 0$ for all $j \neq j^*$. However, if a learning sample from the j^{th} subpopulation is drawn separately (for example, if taken during a season when the subpopulations can be sampled separately) it contributes a term of the form $C \cdot \prod_{\ell=1}^{L} \prod_{k=1}^{K_\ell} \theta_{j\ell k}^{n_{j\ell k}}$ to the likelihood, where C is a product of multinomial coefficients and $n_{j\ell k}$ is the number of alleles of type k observed at locus ℓ in the learning sample taken separately from the j^{th} subpopulation. (In the Bayesian framework, these changes are equivalent to altering the prior for θ and π appropriately.)

Treating this mixture model from the Bayesian perspective requires prior distributions for π and θ. The conjugate prior for π is the Dirichlet distribution, $\text{Dir}(\zeta_1, \ldots, \zeta_J)$. Prior information could be incorporated in the values of the ζ_j, or, if no prior information is available, the uniform distribution

$\zeta_j = 1$, $j = 1, \ldots, J$, is a reasonable choice. The conjugate prior for each $\theta_{j\ell}$ is $\text{Dir}(\lambda_{j\ell 1}, \ldots, \lambda_{j\ell K_\ell})$. In this paper, I use uniform Dirichlet priors, $\lambda_{j\ell k} = 1 \ \forall j, \ell, k$, which tend to de-emphasize the influence of rarely-occurring allelic types. This is a conservative assumption, and works well when the subpopulations are sufficiently genetically distinct. Note, however, that Pritchard et al. (2000a) discuss different Dirichlet priors and Pella and Masuda (in press) describe a useful Empirical Bayes approach to assigning allele frequency priors in the application of Bayesian methods to mixed-stock fishery analysis with closely-related subpopulations.

With the priors specified, the posterior distribution of π and θ, as well as other quantities of interest, may be investigated via Gibbs sampling as described by Diebolt and Robert (1994). The relevant full conditionals are

$$\begin{aligned}
\pi|\cdots &\sim \text{Dir}(\zeta_1 + \#\{z=1\}, \ldots, \zeta_J + \#\{z=J\}) \\
\theta_{j\ell}|\cdots &\sim \text{Dir}(\lambda_{j\ell 1} + m_{j\ell 1} + n_{j\ell 1}, \ldots, \lambda_{j\ell K_\ell} + m_{j\ell K_\ell} + n_{j\ell K_\ell}), \\
&\quad j = 1, \ldots, J; \quad \ell = 1, \ldots, L \\
p(z_i = j|\cdots) &= \frac{\pi_j p(y_i|\theta_j, z_i = j)}{\sum_{k=1}^{J} \pi_j p(y_i|\theta_j, z_i = k)}, \ i = 1, \ldots, N; \ j = 1, \ldots, J
\end{aligned}$$

where $\#\{z = j\}$ is the number of individuals currently allocated to subpopulation j, $m_{j\ell k}$ is the number of alleles of type k at locus ℓ in individuals currently allocated to subpopulation j, and the $n_{j\ell k}$ are, as before, the allele counts from the learning samples (if any) drawn separately from the mixture sample.

3. 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

$$p(y_i|\theta, w_i) = \prod_{t=1}^{2L} \theta\langle w_{it}; y_{it}\rangle \tag{3}$$

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

$$\begin{aligned}
q_i|\cdots &\sim \text{Dir}(\alpha_1 + \#\{w_i = 1\}, \ldots, \alpha_J + \#\{w_i = J\}), \quad i = 1, \ldots, N \\
\theta_{j\ell}|\cdots &\sim \text{Dir}(\lambda_{j\ell 1} + r_{j\ell 1}, \ldots, \lambda_{j\ell K_\ell} + r_{j\ell K_\ell}), \\
&\quad j = 1, \ldots, J; \quad \ell = 1, \ldots, L \\
p(w_{it} = j|\cdots) &= \frac{q_{ij}\theta\langle j; y_{it}\rangle}{\sum_{k=1}^{J} q_{ij}\theta\langle k; y_{it}\rangle}\ , \ i = 1, \ldots, N;\ j = 1, \ldots, J; \\
&\quad t = 1, \ldots, 2L
\end{aligned}$$

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.

3.1 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 4.

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 and Kotz 1977). 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 Beta(α_1, α_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

$$\frac{b_1 + d_t c}{b_1 + b_2 + tc}. \tag{4}$$

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 (4) 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. This Markov chain dependence structure was previously noted by Freedman (1965), who used it to obtain limiting distributions of quantities associated with urn models.

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,

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

Figure 1 about here.

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

In the forward step we compute and store values of $p(w_t, d_t|y_{\leq t})$ for $w_t = 1, 2$ and $d_t = 0, \ldots, t$, recursively for $t = 1, \ldots, 2L$, by the relations

$$p(w_{t+1}, d_{t+1}|y_{\leq t}) = \sum_{1 \leq w_t \leq 2} p(w_{t+1}, d_{t+1}|w_t, d_t^*) p(w_t, d_t^*|y_{\leq t}) \tag{5}$$

where $d_t^* = d_{t+1} - 1\{w_{t+1} = 1\}$, and

$$p(w_{t+1}, d_{t+1}|y_{\leq t+1}) = \frac{1}{\phi_{t+1}} p(w_{t+1}, d_{t+1}|y_{\leq t}) p(y_{t+1}|w_{t+1}, d_{t+1}) \tag{6}$$

where

$$\phi_{t+1} = p(y_{t+1}|y_{\leq t}) = \sum_{\substack{1\leq w_{t+1}\leq 2 \\ 0\leq d_{t+1}\leq t+1}} p(w_{t+1}, d_{t+1}|y_{\leq t})p(y_{t+1}|w_{t+1}, d_{t+1}). \tag{7}$$

At the end of the forward step, notice that $\prod_{t=1}^{2L} \phi_t = p(y_1, \ldots, y_{2L})$, which in the context of the Gibbs sampler (and if we were to reinstate the i subscript) is the desired quantity $p(y_i|\alpha, \theta)$ for the i^{th} individual. At the end of the forward step we have also obtained the distribution $p(w_{2L}, d_{2L}|y_{\leq 2L})$. After simulating values for w_{2L} and d_{2L} from that distribution, we are in a position to simulate values for w_t going backwards recursively for $t = 2L-1, 2L-2, \ldots, 1$, using the conditional distributions stored during the forward step and the values just realized for w_{t+1} and d_{t+1}. The backward step uses the following relations recursively to compute the conditional distribution from which to realize values of (w_t, d_t):

$$\begin{aligned} p(w_t, d_t|y_{\leq 2L}, w_{\geq t+1}, d_{\geq t+1}) &= p(w_t, d_t|y_{\leq t}, w_{t+1}, d_{t+1}) \\ &= \frac{1}{\psi_t} p(w_t, d_t|y_{\leq t})p(w_{t+1}, d_{t+1}|w_t, d_t), \end{aligned} \tag{8}$$

where ψ_t is a normalizing constant

$$\psi_t = \sum_{1\leq w_t\leq 2} p(w_t, d_t^*|y_{\leq t})p(w_{t+1}, d_{t+1}|w_t, d_t^*) \tag{9}$$

and where, again, $d_t^* = d_{t+1} - 1\{w_{t+1} = 1\}$. It is apparent that a realization of the variable $(w_1, \ldots, w_{2L})$ thus obtained is drawn from the distribution of $w_1, \ldots, w_{2L}$ conditional on $y_{\leq 2L}$. As such, in the context of the Gibbs sampler, and reinstating the i subscript, it is a realization from $p(w_i|\alpha, \theta, y_i)$ for the i^{th} individual, as desired.

The amount of computation required for the backward step is linear in L. The forward step at time t requires a handful of elementary operations for each of the $2t$ states that the pair (w_t, d_t) may take. This makes the entire forward step $O(L^2)$ for the case of two subpopulations. Depending on how many loci are available this will typically be computationally reasonable. However, extending this method to $J > 2$ will be computationally difficult. With $J > 2$, d_t becomes a vector whose elements record the number of balls of type $1, \ldots, J$ which have been drawn up to and including time t. The number of possible states for the pair (w_t, d_t) is then $J(t + J - 2)!/[(t-1)!(J-1)!]$ which gets large rapidly with t and J.

4. A MODEL FOR SIMULTANEOUS POPULATION MIXTURE AND ADMIXTURE

Continuing in the case of two subpopulations ($J = 2$), a common goal in applications would be to identify purebred versus admixed individuals and to estimate the proportion of those types in the population. This corresponds to partitioning one's sample into purebred and admixed groups. The i^{th} individual's inclusion in one of the two groups can be denoted by a latent variable v_i taking the values

$$v_i = \begin{cases} \text{P}, & \text{if purebred} \\ \text{A}, & \text{if admixed following the Pritchard et al. (2000a) model} \end{cases}$$

Using the calculation of Section 3.1, this partition problem can be treated as a mixture problem using Gibbs sampling. The proportion of individuals of the two types in the population are given by the new parameter $\xi = (\xi_{\text{P}}, \xi_{\text{A}})$, with $\xi_{\text{P}} + \xi_{\text{A}} = 1$. The full conditional distribution for v_i is then, for example, for $v_i = \text{P}$

$$p(v_i = \text{P}|\cdots) = \frac{\xi_{\text{P}} p(y_i|\alpha, \theta, v_i = \text{A})}{\xi_{\text{P}} p(y_i|\pi, \theta, v_i = \text{P}) + \xi_{\text{A}} p(y_i|\alpha, \theta, v_i = \text{A})}. \tag{10}$$

Calculating each of the necessary phenotype probabilities, $p(y_i|\pi, \theta, v_i = \text{P})$ and $p(y_i|\alpha, \theta, v_i = \text{A})$, has been described in Equation 2 and Section 3.1.

The conjugate prior for ξ_P is a Beta$(\delta_{\text{P}}, \delta_{\text{A}})$ which gives the full conditional

$$\xi|\cdots \ \sim \ \text{Beta}(\delta_{\text{P}} + \#\{v = \text{P}\}, \delta_{\text{A}} + \#\{v = \text{A}\}). \tag{11}$$

I have used a uniform ($\delta_{\text{P}} = \delta_{\text{A}} = 1$) prior for ξ_{P}. This prior corresponds to each individual i having prior probability of 1/2 of being either purebred or admixed.

In this expanded model, which we will call model $M_{\text{P,A}}$ a sweep consists of

1. Gibbs update for π using only the individuals with $v_i = \text{P}$,
2. Gibbs update for θ where contributions to the full conditionals are determined by z_i for individuals with $v_i = \text{P}$ and by w_i for those with $v_i = \text{A}$,
3. Gibbs updates for each individual's z_i if $v_i = \text{P}$ and for w_i if $v_i = \text{A}$,
4. Gibbs update for ξ from Equation 11,
5. Gibbs update for each v_i using Equation 10,

6. Metropolis-Hastings update for α as described in Appendix A.

The output from the resulting Markov chain can provide Rao-Blackwellized (Liu, Wong and Kong 1994) estimates for the posterior probability that individuals in the sample are purebred or admixed as well as esimates of the posterior distributions for ξ, π, θ, α, and each q_i given $v_i = A$ (though the q_i's are not necessary for running the chain, they may still be realized from their full conditional distributions and they provide good summary statistics).

5. BAYESIAN MODEL COMPARISON

Once able to entertain the model $M_{\mathrm{P,A}}$, it is natural to ask whether that expanded model has gained us anything. One way to pose the question is to ask whether the data provide more support for $M_{\mathrm{P,A}}$ than for the model we will call M_{A} which requires all individuals to be admixed and governed by a single α. A rough estimate of the Bayes factor, $B = p(y|M_{\mathrm{P,A}})/p(y|M_{\mathrm{A}})$, might be obtained by observing the proportion of time the Markov chain defined under $M_{\mathrm{P,A}}$ spends in states with zero or almost zero individuals allocated to the purebred group (since retricting ξ_{P} to zero in $M_{\mathrm{P,A}}$ essentially gives M_{A}). However, this is unsatisfactory as there is no prior probability mass on the point $\xi_{\mathrm{P}} = 0$. Furthermore, the chain may visit states with low ξ_{P} so infrequently, that it is impossible to get a good estimate of B that way.

Gelfand et al. (1992) suggest approximating B by the "pseudo-Bayes factor" formed as the product over all observations of the ratio of cross-validitation predictive densities under the two models. The cross-validation predictive density for the i^{th} individual, may be approximated by the harmonic mean of the values $p(y_i|\alpha_s, \theta_s)$ under M_{A} and the values $p(y_i|\alpha_s, \pi_s, \theta_s, \xi_{\mathrm{P}_s})$, computed as the denominator in (10), under $M_{\mathrm{P,A}}$, where s subscripts the states visited by the chain over which the harmonic mean is taken. Raftery (1992) cautions that the pseudo-Bayes factor, being akin to a pseudo-likelihood, may be an inaccurate approximation to the Bayes factor and should not be used for model comparison if the latter is available. However as discussed by Pritchard et al. (2000a), it is difficult to reliably estimate the marginal likelihood $p(y|M_{\mathrm{A}})$, and the same is true for $p(y|M_{\mathrm{P,A}})$, making computation of the Bayes factor by that route challenging.

As an alternative, I have developed a reversible-jump MCMC scheme (Green 1995) for computing the Bayes factor $B = p(y|M_{\mathrm{P,A}})/p(y|M_{\mathrm{A}})$. While reversible jump methods have recently received widespread attention for sampling over numerous models in complex model spaces (Rue

and Hurn 1999; Dellaportas and Forster 1999; Giudici and Green 1999), it seems they have been used less often when a small number of closely-related models are being considered, as in the present case. The posterior distributions estimated from separate runs under $M_{\mathrm{P,A}}$ and M_{A} can guide us in devising reversible-jump proposals that are easy to implement and offer a good estimate of B. Details appear in Appendix B. This circumvents the potential problem of instability in a direct, sampling-based estimate of $p(y|M_{\mathrm{P,A}})$ or $p(y|M_{\mathrm{A}})$, and affords an opportunity to compare the pseudo-Bayes factor to the full Bayes factor in the comparison of two complex, hierarchical models.

6. DATA AND RESULTS

The data from Scottish wildcats were provided by Mark Beaumont (University of Reading, UK) and are fully described in Beaumont et al. (in press). The data set is freely available at `http://www.rubic.rdg.ac.uk/ mab/data.html`. Briefly, genetic samples were collected from wild-living cats throughout Scotland by a variety of methods including trapping and tissue collection from road kills and carcasses. Samples were also obtained from 13 museum specimens. In all, 230 wild-living cats were sampled and typed at eight microsatellite loci with numbers of alleles ranging from nine to 17 per locus. Additionally, 74 housecats were typed at those eight loci using blood samples from veterinary centers in the south of England. These 74 cats can be considered a learning sample for the domestic cat subpopulation.

I analyzed the data under model $M_{\mathrm{P,A}}$ using runs of length 62,000 sweeps of ten different chains started from overdispersed starting points by initializing values of all parameters $(\alpha, \theta, \xi, \pi)$ with values simulated from their prior distributions. All ten chains converged very quickly to the same part of the parameter space. The first 2,000 sweeps were discarded as burn-in, as observing the estimated scale reduction potential factor (Gelman 1996) suggests this is more than adequate burn-in. I give the results in the next section. I performed an analagous run under model M_{A} and compare the differences between the results obtained under $M_{\mathrm{P,A}}$ and M_{A} in Section 6.2. For both runs I used an upper bound of $A = 3$ for the parameter α. Each run took about 11 hours on a laptop computer with a 266 Mhz G3 (Macintosh) processor.

It should be noted that the learning sample of housecats breaks the symmetry in the posterior with respect to permutations on the labels for the two components (*F. sylvestris* and housecats) in

the model. Thus, there is not a substantial label-switching (Stephens 2000) problem in this case.

6.1 Results for model $M_{\mathrm{P,A}}$

The posterior mean estimate of ξ_{P}, the proportion of purebred cats, is .65, with a 90% credible set spanning the range from .47 to .79. The MCMC estimate of the posterior density of ξ_{P} is given in Figure 2(a). The distribution is long-tailed to the left. These low values of ξ_{P} coincide with low values of the parameter α (Figure 2(d)). This correlation is expected; when α is low, then admixed individuals are expected to have admixture proportions near to zero or one, and hence the ability to distinguish between admixed and purebred individuals is diminished. The estimated posterior density for α itself is shown in Figure 2(c). It has a peak around 0.7, and tapers off with larger values, but it is still rather high at the upper bound imposed on it of 3. The choice of $A = 3$ is clearly a choice of prior to which the final result will be sensitive. A larger A would reduce the posterior probability for low values of ξ_{P}, reducing the skewness of the posterior for ξ_{P} and increasing its posterior mean estimate. This issue will be taken up again in the Discussion.

Figure 2 about here.

Figure 2(b) gives the estimated posterior density for the probability that a cat is *F. sylvestris* conditional on its being of purebred type. The posterior mean is .83 with a 90% credible interval from .73 to .94. This suggests that a large proportion ($> 60\%$) of the wild-living cats in Scotland are purebred *F. sylvestris*. On the other hand, there is evidence that between 21% and 53% of the wild-living cats are admixed individuals with ancestry from both *F. sylvestris* and domestic cats. Further, it cannot be ruled out that some wild-living cats are pure housecats that have gone feral.

Figure 3 about here.

Figure 3 summarizes the results for individual cats. On the horizontal axis, is the posterior probability of being purebred. On the vertical axis is the posterior probability of being *F. sylvestris* conditional on being purebred. A cluster in the upper right represents 102 of the cats in the sample, all with posterior probability of being pure greater than .80. Given that these cats are pure, they have posterior probability close to one of being *F. sylvestris*. Also evident is a small cluster of cats with $p(v_i = \mathrm{P}|y) > .65$ but which, if they are purebred cats, are almost certainly domestic cats. At the other end of the scale are several cats with very high probability of being admixed.

6.2 Comparison of results for models $M_{\mathrm{P,A}}$ and M_{A}

Analyazing the data under M_{A}, using the new approach in Section 3.1, I obtained the same results as Beaumont et al. (in press) did. The log of the Bayes factor, $\log B$, comparing the support of the

data for $M_{\mathrm{P,A}}$ versus M_{A}, given the Scottish cat data, is ≈ 20.3. Thus, $2 \log B > 40$, indicating overwhelming support for $M_{\mathrm{P,A}}$ (Raftery 1996). All details appear in Appendix B. The log of the pseudo-Bayes factor is 12.3 which, quite notably, differs by eight from the true $\log B$.

For parameters shared by M_{A} and $M_{\mathrm{P,A}}$, the estimates differ between models most for α. Under M_{A}, α is much smaller, so as to accommodate the purebred cats as admixed individuals with admixture proportions close to 0 or 1 (Figure 4). Additionally, a significantly smaller proportion of the alleles in the sample get allocated to the housecat population under $M_{\mathrm{P,A}}$ than under M_{A}. Histograms of the proportion of alleles allocated to the housecat subpopulation for $M_{\mathrm{P,A}}$ and M_{A} are shown in Figure 5.

Figure 4 about here.

Figure 5 about here.

7. DISCUSSION

In applications to conservation biology and ecology, populations of interest may be pure mixtures of two subpopulations, or they may contain admixed individuals from two originally separate subpopulations. Genetic data, in conjunction with statistical models of genetic mixture and admixture have been useful for clustering individuals and genes from different subpopulations. This paper presents a novel application of the "filter-forward-simulate-backward" algorithm akin to the computations presented in Baum et al. (1970) to the population-admixture model of Pritchard et al. (2000a). This computation makes it possible to expand that model to one that allows for individuals to be either purebred or admixed. Such an expanded model ($M_{\mathrm{P,A}}$) is vastly more supported by the Scottish cat data than one including admixture only. It is likely that $M_{\mathrm{P,A}}$ will fit other datasets much better as well, because samples from recently admixed populations will typically include some purebred individuals.

While the dramatic improvement of model fit is encouraging, it also raises some issues that bear further investigation. The first of these is that $M_{\mathrm{P,A}}$ may fit the data better not simply because it allows separate classes of purebred and admixed individuals. It may be that a great deal of improvement comes from including the parameter π which allows different contributions of pure cats from the two subpopulations. This contrasts to the formulation in M_{A} where, due to the symmetry of the Beta(α, α) prior for the q_i's, the marginal probabilities are equal that any gene copy is from the housecat or the *F. sylvestris* subpopulation. That is to say, under M_{A}, $p(w_{it} = 1|\alpha) = p(w_{it} = 2|\alpha) = .5$ for all i, t, and α. By contrast, under $M_{\mathrm{P,A}}$, for different values

of ξ, π, and α, the marginal probability that any gene copy is from the housecat population is not constrained to equal the marginal probability that it is from the *F. sylvestris* population. The symmetry imposed by M_{A} might explain some systematic biases for estimates of q_i that Pritchard, Stephens, Rosenberg and Donnelly (2000b) report for simulated data with unequal admixture proportions. Furthermore, the issue may have implications for model $M_{\mathrm{P,A}}$. If the population admixture proportions depart from .5, then $p(y_i|\alpha, \theta)$ might be inflated for individuals with large amounts of ancestry from the lesser-represented subpopulation, and deflated for individuals with more ancestry from the greater-represented subpopulation. For this reason, in the Scottish cat problem, one might expect that the posterior probability of being a purebred individual will be overestimated for cats that resemble *F. sylvestris* and underestimated for cats that appear to be housecats. This may also induce some bias in the posterior estimate of ξ_{P}. All this suggests that a fruitful extension to the model M_{A} of Pritchard et al. (2000a) would be to allow population-specific α's. For example, in the case of J subpopulations, $q_i \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_J)$.

The results also suggest that estimation in $M_{\mathrm{P,A}}$ may be sensitive to the upper bound, A, chosen for α. Had A been chosen greater than three, then values of $\alpha > 3$ would surely have been visited in the MCMC simulation, and the resulting estimate for ξ_{P} would have been somewhat larger, since α and ξ_{P} are positively correlated. This is observed in a separate run made with $A = 10$—the chain visits values of α between 3 and 10 quite frequently. In fact, the estimated posterior density for α decreases only slightly between 3 and 10. However, the effect on the other parameters is not overwhelming. For example, with $A = 10$, the posterior mean (90% credible interval) for ξ_{P} was .71 (.53, .82), as opposed to .65 (.47, .79) with $A = 3$. It is interesting that the choice of A has little effect in the poorer-fitting model M_{A}, because that model tries to fit purebred cats as admixed individuals. This keeps α low regardless of A. Under $M_{\mathrm{P,A}}$, however, once the purebred individuals are removed from the admixed class there is little information left for estimating α. So, paradoxically, to use the better-fitting model $M_{\mathrm{P,A}}$ requires imposing more prior information. In the case of A, however, biological knowledge can guide the choice.

I chose $A = 3$ because, with only two subpopulations, large values of α indicate that admixed individuals carry close to half of their ancestry from one subpopulation and half from the other. In a population like this, the most plausible explanation for such a pattern would be that the admixed individuals were all first-generation (F1) hybrids between individuals from the two subpopulations.

If this is the case, then, at each locus, an admixed individual will carry exactly one allele from one subpopulation, with the other allele coming from the other subpopulation. This condition can be used to compute the posterior probability that an individual in the sample is an F1 hybrid. I leave the description of such a procedure to a separate paper, however, I have found that none of the individuals in the sample had posterior probability greater than .5 of being an F1 hybrid. In fact, for all but seven of the individuals, the posterior probability of being an F1 hybrid was below .10. For this reason, it seemed implausible that α should be allowed to range past about three.

The Bayesian model comparison revealed that $M_{\mathrm{P,A}}$ is a much better model for the Scottish cat data. Computing Bayes factors in these models is often difficult because calculating the marginal likelihood can require a difficult computation of an unkown normalizing constant. Rather than directly computing the marginal likelihood, Appendix B gives an example of how approximations to posterior densities of several parameters in different models may be used to formulate reversible-jump moves between a small set of closely-related models. This gives us a good approximation to the Bayes factor. In turn, that allows us to compare the true Bayes factor to the pseudo-Bayes factor (a product of ratios of cross-validation predictive densities). The pseudo-Bayes factor has been advocated as a computationally manageable approximation to the Bayes factor. While cross-validation and predictive densities offer a fine level of detail for exploring which observations, in particular, are poorly fit by a model, their use in overall model comparison via the pseudo-Bayes factor should be done with reservation. In the current problem, the pseudo-Bayes factor provided a poor approximation of the Bayes factor.

Quite apart from genetic mixtures, the forward-backward computation here may be useful in more general mixture problems. Sometimes, the Gibbs sampler mixes poorly in the Bayesian analysis of mixtures. Robert (1996) describes this in terms of *trapping states* in finite normal mixtures: when only one or a few observations are allocated to a component, the parameters for that component fit the few observations so tightly that few if any of the other observations would likely get allocated to the component. Reparametrizing the normal mixture model, as done by Mengerson and Robert (1995), corrects the problem by keeping the component-specific parameters from fitting the observations in a near-empty component too tightly. However, this does not address trapping states that may occur simply because the mixing proportion for a component becomes small. If the mixing proportion of a component happens to reach a value near zero, then the

probability of allocating any observations to that component will also be small, and the component may remain empty through many iterations of the chain.

The block-updating scheme of Section 3.1 can provide a useful Gibbs move that could be executed to restore empty components, by the following rationale: in a J-component finite mixture with a $\mathrm{Dir}(\zeta_1, \ldots, \zeta_J)$ prior on the mixing proportions, the latent allocation variables, z_i, marginally follow a labelled compound multinomial-Dirichlet distribution. Consequently, conditional on current values of all the z_i's, the subset of those having any two values will follow a labelled beta-binomial distribution (Johnson, Kotz and Balakrishnan 1997). That is, the marginal distribution of $\{z_i : z_i = j_a \cup z_i = j_b, j_a \neq j_b, i = 1, \ldots, N\}$ follow a labelled beta-binomial distribution with parameters $(\zeta_{j_a}, \zeta_{j_b})$. Thus, the methods of Section 3.1 could be applied to redistribute elements amongst the two components j_a and j_b, having marginalized over the mixing proportions π_{j_a} and π_{j_b}. And so, observations may be reallocated to component j_a (or j_b), according to their full conditional distributions, even if π_{j_a} (or π_{j_b}) is close to zero.

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 $\mathrm{Normal}(\alpha, \sigma^2)$ 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

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

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 3.1), then with only admixed individuals (model M_{A}) the acceptance probability may be written as

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

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

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

APPENDIX B. REVERSIBLE JUMP MCMC FOR MODEL COMPARISON

We may compute the Bayes factor, B, by reversible jump MCMC (Green 1995). This method allows for the construction of a Markov chain that may jump between state spaces of varying dimension. In our case we construct a chain which takes values in two spaces indexed by $m = 1$ or 2. If $m = 1$ then the chain is currently in the space associated with model M_{A}, and it moves to new values within that space as described in Section 3. If $m = 2$, then the chain is currently in the state space associated with model $M_{\mathrm{P,A}}$, and it moves to new values within that space as described in Section 4. Since $M_{\mathrm{P,A}}$ includes the variables ξ and π (ξ has one degree of freedom, ξ_{P}, and, in the case of $J = 2$, π has one degree of freedom as well) which are absent in model M_{A}, there are two extra degrees of freedom when $m = 2$. For this reason, reversible-jump moves are required to move

from $m = 1$ to $m = 2$. The formulation of these moves is such that detailed balance is satisfied, ensuring that the proportion of time the chain spends with $m = 1$ converges to $p(M_{\rm A}|y)$ as the chain is run for infinite time, and so, for a run of the chain of length n, the quantity

$$\frac{\sum_{i=1}^{n} 1\{m_i = 1\}}{\sum_{i=1}^{n} 1\{m_i = 2\}} \tag{A.1}$$

estimates the posterior odds, which, upon division by the prior odds, gives B.

For a reversible jump move from $m = 2$ to $m = 1$ we leave θ unchanged and propose a new value for α, say α', that is a deterministic, many-to-one, function g of the current values of α, $\xi_{\rm P}$, and π. We are at liberty to choose any appropriate and suitable g. For the Scottish cat problem, by examining the posterior distribution of $\xi_{\rm P}$, π, and α under $M_{\rm P,A}$, and by surmising that high values of $\xi_{\rm P}$ in $M_{\rm P,A}$ should correspond to low values of α' in $M_{\rm A}$, I empirically chose

$$\alpha' = g(\alpha, \xi, \pi) = 0.0925 + 0.13638\alpha - 0.21 \sin^{-1}(\xi_{\rm P}^2). \tag{A.2}$$

In this case, $\sin^{-1}(\xi_{\rm P}^2)$ was chosen, since that transformed variable has a simpler (i.e. more linear) relationship with α than does $\xi_{\rm P}$, itself, in the MCMC output from $M_{\rm P,A}$ (see Figure 2(d) in the main text). Notice that π does not actually appear in the function g, since this simplifies the Jacobian, and it does not seem essential (i.e. there is not large correlation between α and π). Taking the 5000 pairs $(\alpha, \xi_{\rm P})$ that were plotted in Figure 2(d) and applying g to them gives values of α' summarized by their histogram in Figure 6(a). This histogram resembles the posterior distribution of α under $M_{\rm A}$ (broken line in Figure 4), as desired.

Figure 6 about here.

To propose the reverse move from $m = 1$ with a current value α' to $m = 2$ with proposed values for the parameters α, $\xi_{\rm P}$ and π requires simulating new values for $\xi_{\rm P}$ and π from known densities and then using those values and the inverse of the function g to determine what value of α shall be proposed. The known densities were chosen to approximate the posterior density estimates of $\xi_{\rm P}$ and π under $M_{\rm P,A}$. Letting π_0 denote the proportion of purebred cats that are *F. sylvestris*, the densities used were $f_\xi(\xi_{\rm P}) \equiv \text{Beta}(8, 4)$ and $f_\pi(\pi_0) \equiv .8\text{Beta}(30, 9) + .2\text{Beta}(15, 2)$. These densities are shown in Figure 6(b). Comparison to Figures 2(a) and 2(b) shows that they resemble overdispersed versions of $p(\xi_{\rm P}|y)$ and $p(\pi_0|y)$. With values of $\xi_{\rm P}$ and π_0 drawn from these densities, α is determined by

$$\alpha = g^{-1}(\alpha', \xi_{\rm P}, \pi_0) = \frac{0.21 \sin^{-1}(\xi_{\rm P}^2) + \alpha' - .0925}{.13638}.$$

We may propose a reversible jump move at the end of each sweep. Thus, if $m = 1$, after a sweep updating all the variables associated with $M_{\rm A}$, we propose a jump up to $m = 2$. If $m = 2$, then after a sweep updating all the variables associated with $M_{\rm P,A}$ we propose a jump down to $m = 1$. Under such a scheme, the acceptance probability for a proposed move from $m = 1$ with $\alpha = \alpha'$ to $m = 2$ and parameter values $(\alpha, \xi_{\rm P}, \pi_0)$, is $\min\{1, \mathcal{A}\}$, with $\mathcal{A}$ reducing to

$$\begin{aligned} \mathcal{A} &= \frac{p(M_{\rm P,A})}{p(M_{\rm A})} \times \frac{p(\alpha|M_{\rm P,A})}{p(\alpha'|M_{\rm A})} \times \frac{p(\xi_{\rm P}|M_{\rm P,A})p(\pi_0|M_{\rm P,A})}{f_\xi(\xi_{\rm P})f_\pi(\pi_0)} \\ &\times \frac{\prod_{i=1}^N [\xi_{\rm P} p(y_i|\pi,\theta) + \xi_{\rm A} p(y_i|\alpha,\theta)]}{\prod_{i=1}^N p(y_i|\alpha',\theta)} \times \frac{1}{0.13638} \end{aligned} \tag{A.3}$$

where $p(\cdot|M)$ denotes prior densities for parameters under different models M. If proposing a move down from $m = 2$ with current values $(\alpha, \xi_{\rm P}, \pi_0)$ to $m = 1$ with $\alpha = \alpha'$, the acceptance probability is $\min\{1, \mathcal{A}^{-1}\}$. The factor of $(0.13638)^{-1}$ is the Jacobian from the transformation g.

Figure 7(a) shows a trace of $\log \mathcal{A}$ from a chain forced to stay in $m = 1$ (i.e. it makes proposals to $m = 2$ but is not allowed to accept them) using the Scottish cat data with learning samples, and assuming prior odds for the models $p(M_{\rm P,A})/p(M_{\rm A}) = 1$. Figure 7(b) shows a similar trace of $\log \mathcal{A}^{-1}$ for a chain restricted to $m = 2$.

Figure 7 about here.

It is apparent from these traces that, without imposing strong prior support for $M_{\rm A}$, it is unlikely that a chain in $m = 2$ would ever move to $m = 1$. Thus, I made three different runs with prior log-odds, $\log[p(M_{\rm P,A})/p(M_{\rm A})]$, equal to -19, -20, and -21. From each of these runs, I estimated the posterior log adds by taking the log of (A.1). The value of the posterior log-odds calculated as the average over ten chains started from overdispersed starting points as a function of sweep number is shown for the three different prior odds in Figure 8. Though the chains may not have been run long enough for an extremely precise estimate of the posterior log-odds, an order-of-magnitude estimate can clearly be made. Subtracting the prior log-odds from the estimated posterior log-odds gives, for each of the three different prior odds used, an estimate of ≈ 20.3 for the log of the Bayes factor. Hence, $2 \log B > 40$, indicating overwhelming support in the data for model $M_{\rm P,A}$ over $M_{\rm A}$.

Figure 8 about here.

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," , .

Cavalli-Sforza, L. L., and Bodmer, W. F. (1971), *The Genetics of Human Populations*, San Francisco: W. H. Freeman.

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

Dellaportas, P., and Forster, J. J. (1999), "Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models," *Biometrika*, 86, 615–613.

Diebolt, J., and Robert, C. P. (1994), "Estimation of finite mixture distributions through Bayesian sampling," *Journal of the Royal Statistical Socity, Series B*, 56, 363–375.

Ewens, W., and Spielman, R. (1995), "The transmission/disequilibrium test: history, subdivision, and admixture," *Am J Hum Genet*, 57, 455–464.

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

Freedman, D. A. (1965), "Bernard Friedman's urn," *Annals of Mathematical Statistics*, 36, 956–970.

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.

Gelman, A. (1996), "Inference and monitoring convergence," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman and Hall, pp. 131–143.

Giudici, P., and Green, P. J. (1999), "Decomposable graphical Gaussian model determination," *Biometrika*, 86, 785–801.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Johnson, N. L., and Kotz, Z. (1977), *Urn Models and Their Application*, New York: Wiley & Sons.

Johnson, N. L., Kotz, Z., and Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, New York: Wiley & Sons.

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, 81, 27–40.

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

Mengerson, K. L., and Robert, C. P. (1995), "Testing for mixtures via entropy distance and Gibbs sampling," in *Bayesian Statistics 5*, eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Lindley, D V Smith, Oxford: Oxford University Press, pp. 147–167.

Millar, R. B. (1991), "Selecting loci for genetic stock identification using maximum likelihood, and the connection with curvature methods," *Canadian Journal of Fisheries and Aquatic Sciences*, 48, 2173–2179.

Milner, G. B., Teel, D. J., Utter, F. M., and Burley, C. L. (1981), Columbia River stock identification study: validation of method,, Annu. rep. res., NOAA, Northwest and Alaska Fisheries Center, Seattle, Washington.

Paetkau, D., Calvert, W., Stirling, I., and Strobeck, C. (1995), "Microsatellite analysis of population structure in Canadian polar bears," *Molecular Ecology*, 4, 347–354.

Pella, J., and Masuda, M. (in press), "Bayesian methods for stock-mixture analysis from genetic characters," *Fisheries*, .

Pella, J., and Milner, G. B. (1987), "Use of genetic marks in stock composition analysis," in *Genetics and Fishery Management*, eds. N. Ryman, and F. Utter, Seattle: University of Washington Press, pp. 247–276.

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

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b), "Association mapping in structured populations," *American Journal of Human 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. .

Raftery, A. E. (1996), "Hypothesis testing and model selection," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman and Hall, pp. 164–187.

Rannala, B., and Mountain, J. L. (1997), "Detecting immigration by using multilocus genotypes," *Proc. Natl. Acad. Sci. USA*, 94, 9197–9102.

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.

Rue, H., and Hurn, M. A. (1999), "Bayesian object identification," *Biometrika*, 86, 649–660.

Smouse, P. E., Waples, R. S., and Tworek, J. A. (1990), "A genetic mixture analysis for use with incomplete source population data," *Canadian Journal of Fisheries and Aquatic Science*, 47, 620–634.

Stephens, M. (2000), "Dealing with label-switching in mixture models," *Journal of the Royal Statistical Society, Series B*, 62, 795–809.

Thompson, E. A. (1973), "The Icelandic admixture problem," *Annals of Human Genetics*, 37, 69–80.

Wijsman, E. M. (1984), "Techniques for estimating genetic admixture and applications to the problem of the origin of the Icelanders and the Ashkenazi Jews," *Human Genetics*, 67, 441–448.

List of Figures

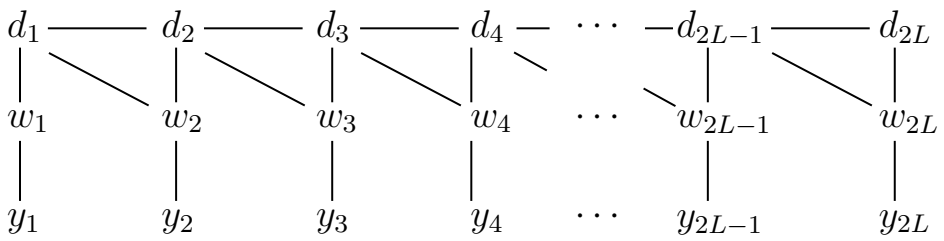
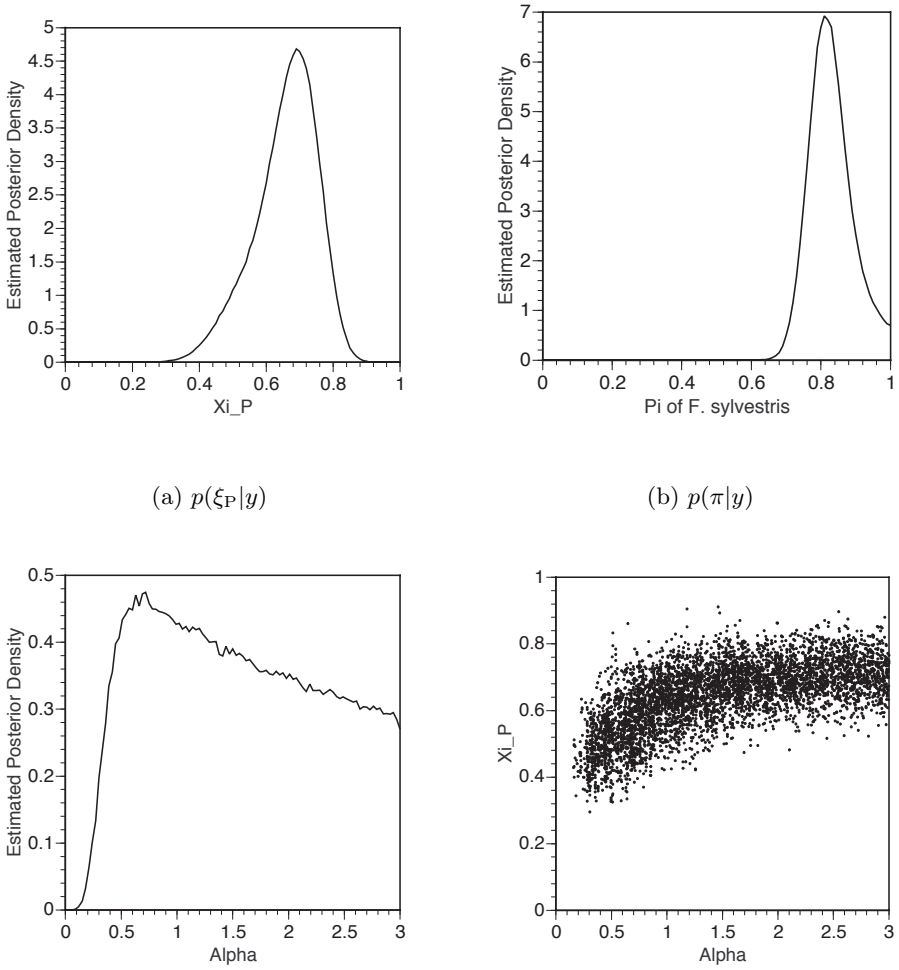

d_1 d_2 d_3 d_4 $\cdots$ d_{2L-1} d_{2L}
w_1 w_2 w_3 w_4 $\cdots$ w_{2L-1} w_{2L}
y_1 y_2 y_3 y_4 $\cdots$ y_{2L-1} y_{2L}

Figure 1



(a) $p(\xi_{\mathrm{P}}|y)$

(b) $p(\pi|y)$

(c) $p(\alpha|y)$

(d) α vs. ξ_{P}

Figure 2

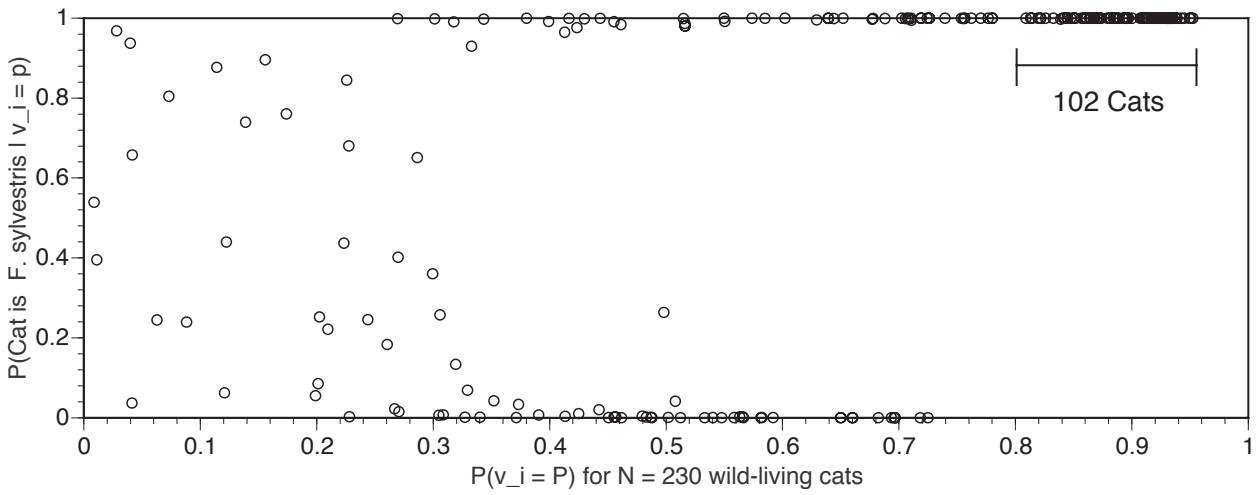

P(Cat is F. sylvestris | v_i = p)
P(v_i = P) for N = 230 wild-living cats
102 Cats
0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9

Figure 3

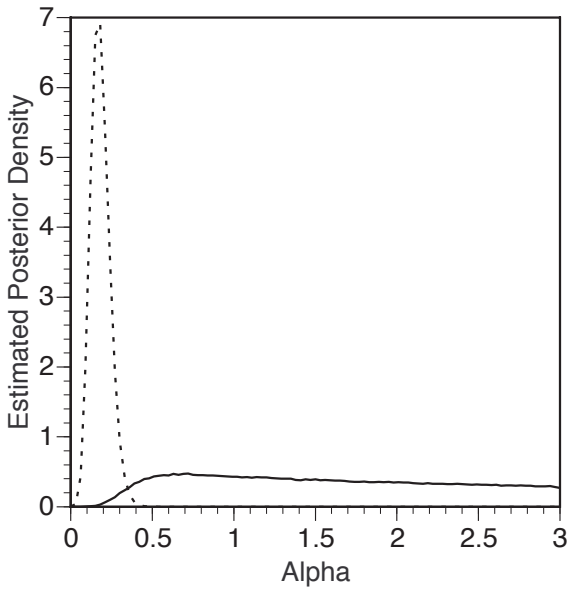


Figure 4

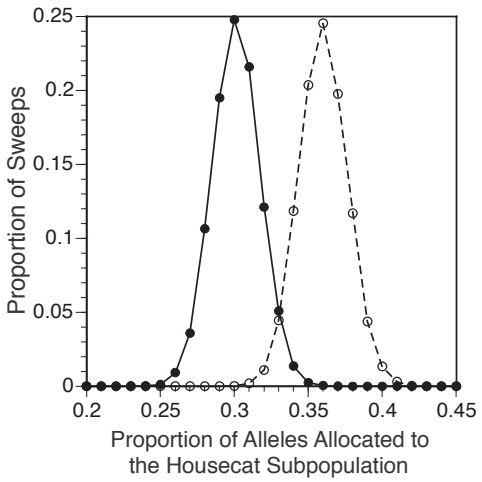
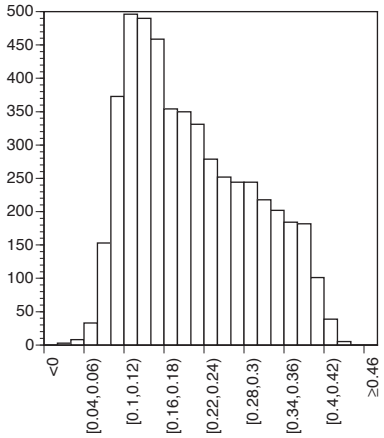
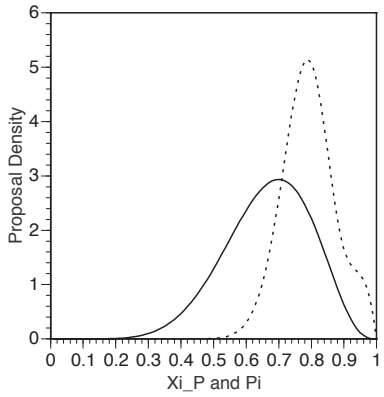

0.25
0.2
0.15
0.1
0.05
0
Proportion of Sweeps
0.2
0.25
0.3
0.35
0.4
0.45
Proportion of Alleles Allocated to
the Housecat Subpopulation

Figure 5

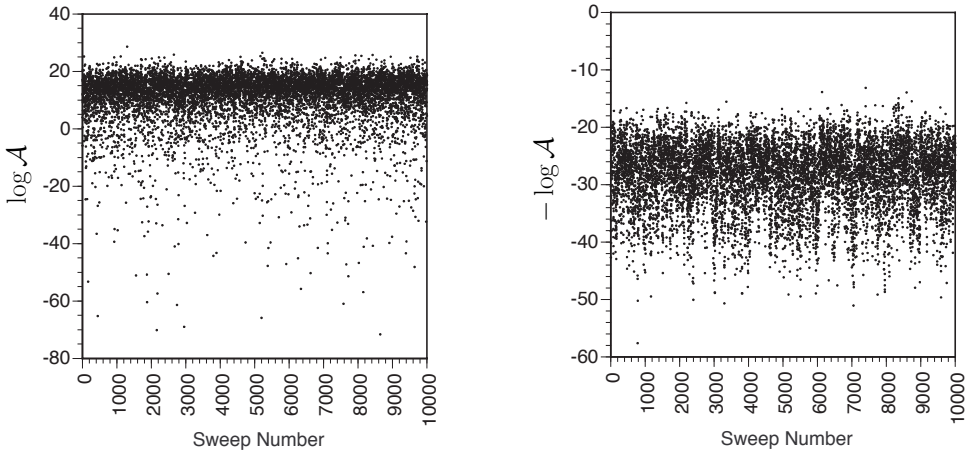


(a) α' histogram



(b) $f_\xi(\xi_{\mathrm{P}})$ & $f_\pi(\pi_0)$

Figure 6



(a) Trace of $\log \mathcal{A}$ under M_{A}

(b) Trace of $-\log \mathcal{A}$ under $M_{\mathrm{P,A}}$

Figure 7

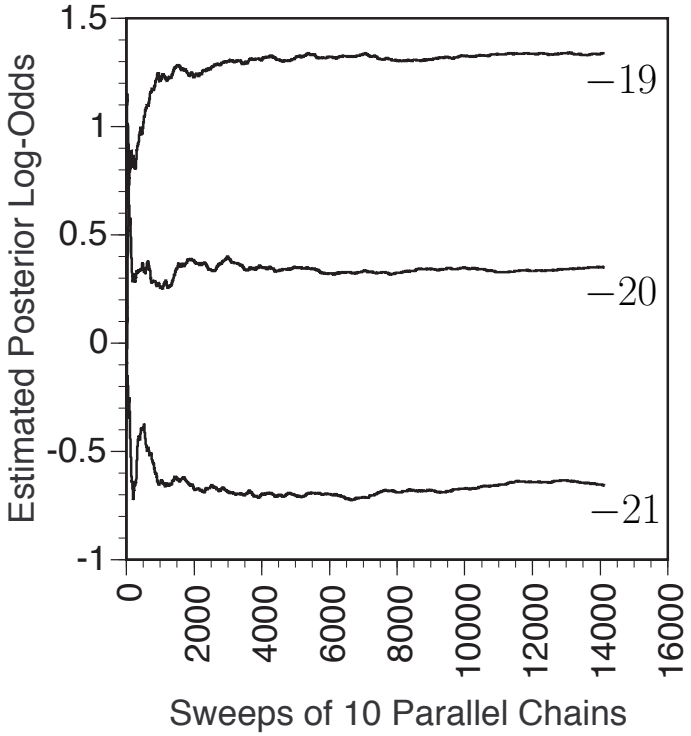

1.5
1
0.5
0
-0.5
-1
Estimated Posterior Log-Odds
−19
−20
−21
0
2000
4000
6000
8000
10000
12000
14000
16000
Sweeps of 10 Parallel Chains

Figure 8