5-24-2011

Using Dirichlet Process Priors For Bayesian Mixture Clustering

Xiao Huang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics

Dissertation Examination Committee:
Stanley Sawyer, Chair
Jeff Gill
Nan Lin
Michael Province
Edward Spitznagel
Victor Wickerhauser

USING DIRICHLET PROCESS PRIORS FOR BAYESIAN MIXTURE CLUSTERING

by

Xiao Huang

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2011

Saint Louis, Missouri

Abstract

We describe a non-parametric Bayesian model using genotype data to classify individuals among populations where the total number of populations is unknown. The model assumes that a population is characterized by a set of allele frequencies that follow multinomial distributions. The Dirichlet Process is applied as the prior distribution. The method estimates the number of populations together with the allele frequencies and the ancestry coefficients of each individual. Distance matrices and bootstrap support numbers based on MCMC runs are generated to create a phylogeny of the ancestral populations.

Acknowledgements

First I want to thank my advisor, Professor Stanley Sawyer, who has supported me in so many ways over the years. Thank you for sharing your vast knowledge of statistics and probability with me, as well as your truly impressive programming skill. Our many long conversations have been invaluable for my research and were some of the most enjoyable moments of my graduate career.

To my parents I owe the greatest debt of gratitude. I will never be able to adequately thank you for the infinite love and support you've given me during the course of my life. Thank you for always encouraging me to excel, for providing me with examples of what hard work and diligence can accomplish, and for believing in me and showing me how to live a meaningful life. I love both of you very much.

The exceptional WU faculty has been incredibly helpful, and I wish to thank all of you for teaching me mathematics. Specifically: Professor Nan Lin, thank you for introducing me to applied statistics in the summer of 2005 and organizing numerous talks for the stats group. Professor Jeff Gill, thank you for running the visiting professor program and giving me a chance to listen to some of the most prominent statisticians, and for offering statistical seminar courses. Professor Rachel Roberts, thank you for being my first year advisor and teaching me the geometry qualifying course. I also thank the remaining members of my dissertation committee for all their help and patience along the way.

Contents

List of Figures

1 Introduction

In a statistical study, the clustering of samples is often the question of interest. Grouping samples in a sensible manner is a common way to study a population. In such a study, the data set typically consists of traits of individuals measured for various attributes. For example, in population genetics, we are interested in inferring characteristics of different populations from those of a few representative individuals. With the original populations unknown, the most reasonable method is to group individuals with similar traits together. This is the context and starting point of our study.

The data will consist of the genotypes of N individuals at L loci. We are interested in the population structure available in such data. Pritchard *et al.* (2000) [23] proposed a model to detect cryptic populations. In this model, the number of populations K is pre-specified, and each population is assumed to be associated with a unique set of allele frequencies. In one version of the model, each individual is assigned to one ancestral population as a pure descendent. In a second version, each individual is assigned to multiple populations as a hybrid. The inference of the population number K is based on an estimation of the posterior probability: namely, the K with the highest posterior probability. The analysis is repeated for several different K values.

We will here consider models for a mixture of probability distributions for L types of counts among N individuals. For definiteness, assume we have data $X = \{m(i, a, j)\}$ corresponding to observations for N individuals ($1 \leq i \leq N$) for A observations ($1 \leq a \leq A$) in each of L categories ($1 \leq j \leq L$). Each observation for the j^{th} category is one of n_j types. In a genetic context, categories are genetic loci and observations are alleles at one locus. That is, the data $X = \{m(i, a, j)\}$ can be written

as

$$m(i,a,j) = b \qquad \text{for } \ 1 \le i \le N, \ \ 1 \le a \le A, \qquad (1.1)$$
$$1 \le j \le L, \ \ 1 \le b \le n_j$$

For simplicity, we assume that there are no missing data. If $A = 1$, the data has the simpler form

$$m(i,j) = b \ \text{ for } \ 1 \le i \le N, \ \ 1 \le j \le N, \ \ 1 \le b \le n_j \qquad (1.2)$$

The N individuals do not necessarily come from a homogeneous background. The simplest model assumes that each individual ($1 \le i \le N$) comes from exactly one of M cryptic source populations ($1 \le c \le M$), with probability $q(c)$ of coming from the c^{th} background population. The choices of background population are independent for different individuals. For the c^{th} background population, $p(c,j,b)$ is the proportion of the population that shares the b^{th} type of the j^{th} attribute, with independent choices of type for the L attributes. In particular

$$\sum_{b=1}^{n_j} p(c,j,b) = 1 \quad \text{for each pair } \ (c,j)$$

Thus the probability of observing $X = \{m(i,j)\}$ in (1.2) is

$$L_0(X \mid q,p) = \prod_{i=1}^{N} \Big(\sum_{c=1}^{M} q(c) \prod_{j=1}^{L} p(c,j,m(i,j)) \Big). \qquad (1.3)$$

Note that $q(c)$ and $p(c,j,b)$ are not individually identifiable if $L = 1$. The analog of

(1.3) for the A-multiple data (1.1) includes the product over a, given by

$$L_0(X \mid q, p) = \prod_{i=1}^{N} \Big(\sum_{c=1}^{M} q(c) \prod_{j=1}^{L} \prod_{a=1}^{A} p(c, j, m(i, a, j)) \Big). \tag{1.4}$$

The estimation algorithms for $q(c)$ and $p(c, j, b)$ (in our notation) in Sections 4 and 5 below are derived from Pritchard, Stephens, and Donnelly (2000)[23]. Some of the following material is also taken from unpublished lecture notes of S. Sawyer and from X. Ruibin *et al.*(2010)[30]. Section 8 and 10 are due to X. Huang. Section 9 s mostly from X. Ruibin *et al.* (2010)[30], of which X. Huang is a co-author.

2 Biological Background

Our motivation is the study of data from marker loci in population genetics. Marker loci are genetic loci in a species that are subject to mutation but not to Darwinian selection. These can be used to trace relatedness of individuals and to find the chromosomal locations of disease genes. In this case, the models above correspond to alleles or allelic values ($1 \leq b \leq n_j$) at L unlinked loci.

Here we consider the ploidy of the individual. More specifically, if $A = 1$ the data $X = \{m(i,j)\}$ in (1.2) correspond to marker data from N *haploid* individuals. If $A > 1$, the population is A - *ploid* rather than haploid, and the individuals have matching groups of A chromosomes. Most plants and animals (including humans) are *diploid*, corresponding to $A = 2$. Many animals also have sex chromosomes that may have different ploidy between sexes. For example, mammals typically have an X chromosome that is diploid in females but haploid in males. For simplicity, we assume that our marker loci are from *autosomal* chromosomes (i.e. non-sexual chromosomes). Viruses and bacteria are generally haploid ($A = 1$), bacteria often possessing a single circular chromosome. Although most higher plants and animals are diploid (except perhaps at sex chromosomes), there are a number of exceptions: several domestic plants have higher ploidy ($A = 4$ *tetraploid*, or $A = 6$ *hexaploid*) and a few non-domestic plants have higher ploidy as well (e.g., $A = 6$ for California redwoods).

The majority of higher plants and animals ($A = 2$) receive one allele at each locus from each of two parents. If $A = 6$, each offspring can receive one of two linked sets of three alleles from each parent, or else can receive two sets of three alleles, each randomly chosen from the genes at that locus in one parent. In particular, hexaploidy *does not* mean that each individual has six parents. Tetraploidy or hexaploidy usually results from a doubling or tripling of the number of chromosomes at some point in

the ancestry of the species. There are mechanisms for keeping the corresponding sets of two or three loci similar.

For $A > 1$, choices of allelic values at the same locus for the same individual are independent, corresponding to the Hardy-Weinberg law in population genetics. The likelihoods (1.3) and (1.4) say that each individual comes from one of M source populations, but that, given the source population for the i^{th} individual, all of its L alleles (or LA alleles if $A > 1$) are chosen independently from the same source population. Nothing is known about the M populations, or even how many populations there are. Information about the populations are to be inferred from inhomogeneities in genetic data from a sample at those loci.

As in (1.1) and (1.2), we assume that there are n_j distinct alleles out of a total number of NA alleles at the j^{th} locus. An example data set from Pritchard *et al.* (2000) [23] has $N = 200$ diploid ($A = 2$) individuals with data for $L = 5$ loci. Thus there are 400 possible allelic values at each locus. The number of *distinct* allelic values n_j ranges between 9 and 15 for $1 \leq j \leq 5$.

3 Bayesian Statistics and Hidden Variables

In general, suppose that we have independent data $X = (X_1, \cdots, X_n)$, and that each X_i has the same probability distribution that depends on a parameter θ. Both X_i and θ can be vector-valued. Specially, assume

$$P(X \mid \theta) = L_0(X \mid \theta). \tag{3.1}$$

Bayesian methods in general, and the Metropolis-Hastings (MH) and Markov chain Monte Carlo (MCMC) method in particular, depend on our ability to treat both θ and X as random variables on the same probability space. The first step is to specify an arbitrary marginal density $\pi_0(\theta)$ for θ, which is called the *prior density* or *prior distribution* for θ. The second step is to specify a joint probability density

$$L(X, \theta) = L_0(X \mid \theta)\pi_0(\theta) \tag{3.2}$$

for (X, θ) together. We assume that $L(X, \theta)$ and $\pi_0(\theta)$ are densities with respect to natural measures on x or θ respectively, for example combinations of Lebesgue measure (perhaps of high dimension) and/or counting measures. The natural measure on (x, θ) is assumed to be the product measure. Since $\int_x L(x \mid \theta)d\theta = 1$ by (3.1),

$$\int_\theta \int_x L(x, \theta)d\theta = \int_\theta \pi_0(\theta)d\theta = 1 \tag{3.3i}$$

$$\int_x L(x, \theta)dx = \int_x L_0(x \mid \theta)\pi_0(\theta)dx = \pi_0(\theta) \tag{3.3ii}$$

$$\int_\theta L(x, \theta)d\theta = \int_x L_0(x \mid \theta)\pi_0(\theta)d\theta = L(x) \tag{3.3iii}$$

$$L(X \mid \theta) = \frac{L(X, \theta)}{L(\theta)} = \frac{L_0(X, \theta)\pi_0(\theta)}{\pi_0(\theta)} = L_0(X \mid \theta) \tag{3.3iv}$$

Conditions (3.3i) and (3.3ii) say that $L(x, \theta)$ is a normalized joint probability density for x and θ with respect to which the marginal distribution of θ is $\pi_0(\theta)$, and conditions (3.3iii) and (3.3iv) say that the marginal distribution of X depends only on X, while the conditional density of X given θ is the same as $L_0(X \mid \theta)$ in (3.1).

Given a joint probability density of X and θ, and observed data X, it is natural to base inferences about an unknown value of θ on the conditional distribution of θ given X. This called the *posterior density* of θ given X, which is

$$\pi_1(\theta \mid X) = P(\theta \mid X) = L(\theta \mid X) = \frac{L(X, \theta)}{L(X)} = \frac{L_0(X \mid \theta)\pi_0(\theta)}{L(X)}$$

or

$$P(\theta \mid X) = C_X L_0(X \mid \theta)\pi_0(\theta) \tag{3.4}$$

where C_X depends only on the known data X. Basing inferences about θ on the posterior density is called *Bayesian statistics*, and (3.4) is Bayes's rule.

The goal of MCMC methods in Bayesian statistics is to marginalize a large joint distribution. We achieve the goal by finding an ergodic Markov chain that has the posterior density $P(\theta \mid X)$ in (3.4) as a stationary measure. If the chain is ergodic, we can use the mean or median of a long chain to estimate components of the parameter θ in (3.4).

The likelihood times prior in (3.1) is often sufficiently complicated that determining the distribution of the chain is prohibitively difficult. However, it may happen that $L_0(x \mid \theta)$ can be written in terms of a simpler expression $L(x, z \mid \theta)$ such that

$$L_0(x \mid \theta) = \int_z L_0(x, z \mid \theta) dz \tag{3.5}$$

where $L(x, z \mid \theta) \geq 0$. In this case

$$\int_x \int_z L_0(x, z \mid \theta) dz dx = \int_x L_0(x \mid \theta) dx = 1,$$

so the expression $L_0(x, z \mid \theta)$ in (3.5) can be viewed as a joint probability density for two random variables X and Z depending on the parameter θ. The newly-created random variable Z is called a *hidden variable* for $L_0(X, \theta)$. The process of introducing Z is called *data augmentation*[28]. Since Z may be multidimensional, more than one hidden variable can exist. In practice, X, Z, and theta are often highly multidimensional, but such that the corresponding Markov chains can be broken up into a series of lower-dimensional componentwise Markov steps.

As in (3.2), the expression

$$L(x, z, \theta) = L_0(x, z \mid \theta)\pi_0(\theta) \tag{3.6}$$

defines a joint density of three random variables X, Z and θ. The relations (3.3) hold as before with x replaced by (x, z). Thus, by (2.6), the conditional distribution of Z given θ is

$$\begin{aligned} P(Z \mid \theta) &= \frac{L(Z, \theta)}{L(\theta)} = \int_x \frac{L(x, Z, \theta)}{L(\theta)} dx = \int_x \frac{L_0(x, Z, \theta)\pi_0(\theta)}{\pi_0(\theta)} dx \\ &= \int_x L_0(x, Z, \theta) dx \end{aligned} \tag{3.7}$$

The posterior density of (θ, Z) (with Z now viewed as additional parameters) is

$$\pi_1(\theta, Z \mid X) = P(\theta, Z \mid X) = \frac{P(\theta, Z, X)}{P(X)} = \frac{P(X \mid Z, \theta)}{P(X)} = \frac{P(X, \mid Z, \theta)P(Z \mid \theta)P(\theta)}{P(X)}$$

This leads to the useful identity

$$P(\theta, Z \mid X) = C_X P(X \mid Z, \theta) P(Z \mid \theta) P(Z \mid \theta) \pi_0(\theta) \tag{3.8}$$

where C_X does not depend on θ. The right-hand side of (3.8) (except for C_X) can usually be obtained from (3.5) and (3.7) very easily. The left-hand side of (3.8) is the analog of a posterior density for (θ, Z) together conditional on X, allowing MCMC methods to be used to estimate both θ and Z.

4 Mixtures of Multivariate Bernoulli Data

4.1 A Basic Haploid Model.

If $A = 1$, the haploid data (1.2) in Section 1 is

$$m(i,j) = b \quad \text{for} \quad 1 \le i \le N, \quad 1 \le j \le L, \quad 1 \le b \le n_j \tag{4.1}$$

If $X = \{m(i,j)\}$, the density

$$L_0(X \mid q,p) = \prod_{i=1}^{N} \Big(\sum_{c=1}^{M} q(c) \prod_{j=1}^{L} p(c,j,m(i,j)) \Big) \tag{4.2}$$

is fairly complex, but becomes simpler if we introduce hidden variables

$$Z(i) = c, \quad 1 \le c \le M \tag{4.3}$$

that denote the (unobserved) source population of the i^{th} individual. By the basic model assumptions, the $Z(i)$ are independent and

$$L_0(X, Z \mid q,p) = \prod_{i=1}^{N} q(Z(i)) \prod_{j=1}^{L} p(Z(i),j,m(i,j)). \tag{4.4}$$

In particular,

$$\int_Z L_0(X, Z \mid q,p) dZ = L_0(X \mid q,p)$$

where dZ is counting measure on part of an N-dimensional lattice. Thus the marginal distribution of X given (q,p) is (4.2). Similarly, the marginal distribution of Z is

$$L_0(Z \mid q,p) = \int_m L_0(m, Z \mid q,p) dm = \prod_{i=1}^{N} q(Z(i)) \tag{4.5}$$

with respect to counting measures on the values of $m(i,j)$, so that the $Z(i)$ are conditionally independent given q. Most importantly, (4.4) replaces the complicated product of sums in (4.2) by a simpler product in terms of Z. This represents an instance of data augmentation in the sense of Section 3.

Given any prior density $\pi_0(q,p)$ for (q,p), it follows as in (3.8) that the posterior density of (q,p,Z) given X is

$$\begin{aligned} P(q,p,Z \mid X) &= C_X P(X,Z \mid q,p)\pi_0(q,p) \\ &= C_X \Big(\prod_{i=1}^{N} q(Z(i)) \prod_{j=1}^{L} p(Z(i),j,m(i,j))\Big)\pi_0(p,q). \end{aligned} \tag{4.6}$$

This is the same as

$$C_X \Big(\prod_{c=1}^{M} q(c)^{N_3(c)}\Big)\Big(\prod_{c=1}^{M}\prod_{j=1}^{L}\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)}\Big)\pi_0(q,p), \tag{4.7}$$

where

$$\begin{aligned} N_2(c,j,b) &= \#\{i : Z(i) = c \text{ and } m(i,j) = b\}, \\ N_3(c) &= \#\{i : Z(i) = c\}. \end{aligned} \tag{4.8}$$

Inference about the parameters (q,p,Z) in (4.6) or (4.7) by MCMC depends on finding a Markov chain $W_n = (q,p,Z)$ for which (4.6) or (4.7) is a stationary measure. The most efficient Markov chains use *Gibbs sampling* of individual *parameters* [3, 11], for which we need to be able to identify the full conditional probability distributions from factors in (4.6) or (4.7). The likelihood (4.7) shows $q(c)$ and $p(c,j,b)$ for fixed (c,j) follow Dirichlet distribution, therefore we pick Dirichlet priors that are conjugate priors for q and p .

Dirichlet priors. We choose prior Dirichlet distributions for $q(c)$ and $p(c,j,b)$ that are relatively "uninformative." Specifically, we assume as priors

$$\begin{aligned} q(c) &\sim \mathcal{D}_M(\alpha,\alpha,\cdots,\alpha), \\ p(c,j,b) &\sim \mathcal{D}_{n_j}(\lambda,\lambda,\cdots,\lambda) \text{ for each } (c,j), \end{aligned} \tag{4.9}$$

where $\lambda > 0$ and $\alpha > 0$. According to Section 3, the posterior density (4.7) is now

$$\begin{aligned} L(q,p,Z \mid X) = C_X C_M(\alpha)\Big(\prod_{c=1}^{M} q(c)^{N_3(c)+\alpha-1}\Big) \\ \times \prod_{c=1}^{M}\prod_{j=1}^{L} C_{n_j}(\lambda)\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)+\lambda-1}, \end{aligned} \tag{4.10}$$

where $C_d(\lambda) = \Gamma(d\lambda)/\Gamma(\lambda)^d$. We now describe the step distribution of a Markov chain $W_n = (q,p,Z)_n$ that is ergodic and leaves the posterior density (4.10) invariant.

Initializing Variables in $W_n = (q,p,Z)_n$. The variables Z_i can be set, for example, by choosing $Z(i)$ independently with a uniform distribution for $1 \le Z(i) \le M$. Initial values of p and q can be set arbitrarily since their initial values will be immediately overwritten in the first step of the Markov chain by updates that depend only on $Z(i)$ and $m(i,j)$.

Each step of the Markov chain consists of a number of substeps. The first substep updates $p(c,j,b)$ for all $p(c,j,b)$, the second updates $q(c)$ for all c, and third updates $Z(i)$ for all i. If desired, updating of λ and/or α can be included as well.

Updating $p(c,j,b)$. From (4.10), the conditional distribution of $p(c,j,b)$ given the other variables and parameters is

$$L(p \mid Z,q,\lambda,\alpha,m) = C(X,Z,\lambda)\prod_{c=1}^{M}\prod_{j=1}^{L}\Big(\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)+\lambda-1}\Big)$$

for $N_2(c, j, b)$ in (4.8). This is a product of Dirichlet densities

$$L_{cj}(p \mid Z, q, \lambda) \sim \mathcal{D}_{n_j}(N_2(c, j) + \lambda)$$

where $N_2(c, j) = (N_2(c, j, 1), \cdots, N_2(c, j, n_j))$. Thus, for each pair (c, j) in sequence, the relation

$$p(c, j, b) \sim \mathcal{D}_{n_j}(N_2(c, j) + \lambda) \tag{4.11}$$

defines a Gibbs-sampler update for $p(c, j, b)$.

Updating $q(c)$. Also from (4.10), the conditional distribution of $q(c)$ given the other variables and parameter is

$$L(q \mid Z, p, \lambda, \alpha, m) = C(X, Z, \alpha) \prod_{c=1}^{N} q(c)^{N_3(c)+\alpha-1}$$

for N_3 in (4.8). Thus $q(c)$ has the Gibbs-sampler update

$$q(c) \sim \mathcal{D}_M(N_3 + \alpha) \tag{4.12}$$

for $N_3 = (N_3(1), \cdots, N_3(M))$.

Updating $Z(i)$. From the density $L(q, p, Z \mid X)$ in (4.6) and the prior $\pi_0(q, p)$ in (4.9), the conditional distribution of Z given the other variables and other parameters is

$$L(Z \mid q, p, \lambda, \alpha, X) = C(p, q, X) \prod_{i=1}^{N} q(Z(i)) \prod_{j=1}^{L} p(Z(i), j, m(i, j)).$$

It follows that the $Z(i)$ are conditionally independent given the other parameters and

variables, and

$$P(Z(i) = c \mid p, \lambda, X) = C_i q(c) \prod_{j=1}^{L} p(c, j, m(i, j)). \tag{4.13}$$

Thus we conclude that, for each i, $Z(i)$ has a multinomial Gibbs-sampler update for the discrete probabilities defined by the right-hand side of (4.13).

Updating $\boldsymbol{\lambda}$. If desired, the parameter λ can be updated as well. Metropolis updates for λ can be used based on the posterior density (4.10) with a uniform prior for λ [19]. The "update proposals" can be either a symmetric uniform or a standard normal distribution with reflect at $\lambda = 0$ [19, 29, 3] . In either case, they are scaled by a parameter $h > 0$ that is adjusted so that the acceptance probabilities fall within the range of 25-40%.

By (4.10), the conditional distribution of λ given the other parameters and variables is

$$L(\lambda \mid Z, p, q, \alpha, X) = C(p) \prod_{c=1}^{M} \prod_{j=1}^{L} \frac{\Gamma(n_j \lambda)}{\Gamma(\lambda)^{n_j}} \prod_{b=1}^{n_j} p(c, j, b)^{\lambda}. \tag{4.14}$$

The gamma function $\Gamma(x)$ takes on both very large and very small values on its domain, and there are numerically good algorithms for calculating $\log \Gamma(x)$. It is easier to calculate

$$\log L(\lambda \mid Z, p, q, \alpha, X) = M\Big(\sum_{j=1}^{L} \log \Gamma(n_j \lambda - n_L \log \Gamma(\lambda))\Big) + \lambda \sum_{c=1}^{M} \sum_{j=1}^{L} \sum_{b=1}^{n_j} \log p(c, j, b)$$

within a positive additive constant, where $n_L = \sum_{j=1}^{L} n_j$.

Updating $\boldsymbol{\alpha}$. The parameter α in the prior distribution for $q(c)$ in (4.9) can also be updated if necessary. By (4.10), the conditional distribution of α given the other

parameters and variables is

$$L(\alpha \mid Z, q, p, \lambda) = C_q \frac{\Gamma(M\alpha)}{\Gamma(\alpha)^M} \prod_{c=1}^{M} q(c)^{\alpha} \tag{4.15}$$

within a positive additive constant. Metropolis updates for α can be carried out in the same way as for λ.

4.2 Higher Ploidy ($A > 1$).

The basic data X is now

$$m(i, a, j) = b \qquad \text{for } \; 1 \le i \le N, \;\; 1 \le a \le A, \;\; 1 \le j \le L, \;\; 1 \le b \le n_j. \tag{4.16}$$

As before, we assume that individuals belong to one of M ancestral populations and mate only within those populations. The model likelihood is now

$$L_0(X \mid q, p) = \prod_{i=1}^{N} \Big(\sum_{c=1}^{M} q(c) \prod_{j=1}^{L} \prod_{a=1}^{A} p(c, j, m(i, a, j)) \Big)$$

by (1.4). For the hidden variables $Z(i)$ in (4.3) and the prior for (q,p) in (4.9), the posterior density is

$$\begin{aligned} P(q,p,Z \mid X) &= C_X P(X,Z \mid q,p)\pi_0(q,p) \\ &= C_X\Big(\prod_{i=1}^{N} q(Z(i))\Big)\Big(\prod_{i=1}^{N}\prod_{j=1}^{L}\prod_{a=1}^{A} p(Z(i),j,m(i,a,j))\Big)\pi_0(q,p) \\ &= C_X\Big(\prod_{c=1}^{M} q(c)^{N_3(c)}\Big)\Big(\prod_{c=1}^{M}\prod_{j=1}^{L}\prod_{b=1}^{n_j} p(c,j,b)\Big)\pi_0(q,p) \\ &= C_X C_M(\alpha)\Big(\prod_{c=1}^{M} q(c)^{N_3(c)+\alpha-1}\Big) \\ &\quad \times \prod_{c=1}^{M}\prod_{j=1}^{L} C_{n_j}(\lambda)\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)+\lambda-1}. \end{aligned} \tag{4.17}$$

This is formally identical to (4.10), where C_X depends on neither q nor p and $C_d(\lambda) = \Gamma(d\lambda)/\Gamma(\lambda)^d$, but now

$$\begin{aligned} N_2(c,j,b) &= \#\{(i,a) : Z(i) = c \ \text{ and } \ m(i,a,j) = b\} \\ N_3(c) &= \#\{i : Z(i) = c\} \end{aligned} \tag{4.18}$$

instead of (4.6). It follows that $p(c,j,b)$ and $q(c)$ have the same Dirichlet updates as in (4.11) and (4.12), and the only difference is between (4.8) and (4.18).

By the joint posterior density of q, p and Z in (4.17), the conditional distribution of Z given the other parameters and variables is

$$L(Z \mid q,p,\lambda,\alpha,X) = C(p,q,X)\prod_{i=1}^{N} q(Z(i))\prod_{j=1}^{L}\prod_{a=1}^{A} p(Z(i),j,m(i,a,j)).$$

It follows that the $Z(i)$ are conditionally independent given the other variables and

$$P(Z(i) = c \mid q, p, \lambda, \alpha, X) = C_i q(c) \prod_{j=1}^{L} \prod_{a=1}^{A} p(c, j, m(i, a, j)). \tag{4.19}$$

For each i, $Z(i)$ has a multinomial Bernoulli Gibbs-sampler update for the discrete probabilities defined by the right-hand side of (4.19). Updates for λ and/ or α are the same for $A = 1$.

An alternative mixture model that is more realistic for diploid populatons is described by the following.

5 Admixtures of Multivariate Bernoulli Data

We now assume that the individuals in a population of interest have interbred among M source or background populations, where the amount of interbreeding depends on the individual. Specifically, we assume that the i^{th} individual ($1 \le i \le N$) has ancestral mixture proportions

$$q(i,c) \;\; (1 \le c \le M) \quad \text{with} \quad \sum_{c=1}^{M} q(i,c) = 1 \;\; \text{for each} \;\; i. \tag{5.1}$$

As before, the basic data X with ploidy $A > 1$ is

$$\begin{aligned} m(i,a,j) = b \qquad & \text{for} \;\; 1 \le i \le N, \;\; 1 \le a \le A, \\ & 1 \le j \le L, \;\; 1 \le b \le n_j \end{aligned} \tag{5.2}$$

as in (1.1), where $m(i,a,j)$ is the allelic value of the a^{th} allele in the i^{th} individual at the j^{th} locus.

We assume that, for each allele in the i^{th} individual, the ancestral population of the allele is chosen with distribution $q(i,c)$, and the specific allelic value is chosen with probability $p(c,j,b)$. Thus the JA alleles for the i^{th} individual are chosen independently with probabilities

$$p_2(i,j,b) = \sum_{c=1}^{M} q(i,c) p(c,j,b)$$

for each j ($1 \le j \le L$) and a ($1 \le a \le A$). This differs from the basic ploidy model, in which the alleles for the i^{th} individual are not chosen independently, since they are conditioned to come from the same ancestral population. The model likelihood

is now

$$\begin{aligned} L_0(X \mid q,p) &= \prod_{i=1}^{N}\prod_{a=1}^{A}\prod_{j=1}^{L} p_2(i,j,m(i,a,j)) \\ &= \prod_{i=1}^{N}\prod_{a=1}^{A}\prod_{j=1}^{L}\Big(\sum_{c=1}^{M} q(i,c)p(c,j,m(i,a,j))\Big). \end{aligned} \tag{5.3}$$

We introduce the hidden variables

$$Z(i,a,j) = c, \quad 1 \le c \le M \tag{5.4}$$

which give the ancestral population of the a^{th} allele in the i^{th} individual at the j^{th} locus, which we assume are chosen independently with probabilities $q(i,c)$ for all NAL alleles. The augmented data model likelihood is then

$$L_0(X,Z \mid q,p) = \prod_{i=1}^{N}\prod_{a=1}^{A}\prod_{j=1}^{L} q(i,Z(i,a,j))p(Z(i,a,j),j,m(i,a,j)). \tag{5.5}$$

As in the previous section,

$$\int_Z L_0(X,z \mid q,p)\,dz = L_0(X \mid q,p)$$

for $L_0(X \mid q,p)$ in (5.3), where "dz" denotes counting measure on an $NAL-$ dimensional lattice. Note that (5.5) can also be written

$$L_0(X,Z \mid q,p) = \Big(\prod_{i=1}^{N}\prod_{c=1}^{M} q(i,c)^{N_3(i,c)}\Big)\Big(\prod_{c=1}^{M}\prod_{j=1}^{L}\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)}\Big),$$

where

$$\begin{aligned} N_2(c,j,b) &= \#\{(i,a) : Z(i,a,j) = c, \quad \text{and} \quad m(i,a,j) = b\} \\ N_3(i,c) &= \#\{(a,j) : Z(i,a,j) = c\}. \end{aligned} \tag{5.6}$$

We assume the prior distributions

$$\begin{aligned} q(i,c) &\sim \mathcal{D}_M(\alpha,\alpha,\cdots,\alpha) \quad \text{for each} \quad i, \\ P(c,j,b) &\sim \mathcal{D}_{n_j}(\lambda,\lambda,\cdots,\lambda) \quad \text{for each} \quad (c,j), \end{aligned} \tag{5.7}$$

where $\mathcal{D}_d$ is the Dirichlet distribution for $\alpha > 0$, $\lambda > 0$. The posterior density is then

$$\begin{aligned} L(q,p,Z \mid X) = C_X\Big(\prod_{i=1}^{N} C_M(\alpha) \prod_{c=1}^{M} q(i,c)^{N_3(i,c)+\alpha-1}\Big) \\ \times \prod_{c=1}^{M}\prod_{j=1}^{L} C_{n_j}(\lambda) \prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)+\lambda-1}, \end{aligned} \tag{5.8}$$

where $C_d(\lambda) = \Gamma(d\lambda)/\Gamma(\lambda)^d$.

As before, we estimate the model parameters using MCMC.

The Markov Chain: Initializing Variables. As in the previous section we initialized the Markov chain $W_n = (q,p,Z)_n$ by assigning arbitrary initial values to the hidden values $Z(i,a,j)$ in (5.4), for example, independently and uniformly distributed for $1 \le Z \le M$ for all triples (i,a,j). If it is desirable to vary λ and/or α, they can also be assigned arbitrary positive initial values. Note that the initial values of q and p are irrelevant since any initial values will be overwritten at the first step of the following Gibbs-sampler updates.

Each step of the Markov chain will involve the following substeps.

Updating $p(c,j,b)$. It follows from (5.8) that the conditional distribution of

$p(c, j, b)$ given the other parameters and variables is

$$L(p \mid Z, q, \lambda, \alpha) = C(X, Z, \lambda) \prod_{c=1}^{M} \prod_{j=1}^{L} \Big(\prod_{b=1}^{n_j} p(c, j, b)^{N_2(c,j,b)+\lambda-1} \Big) \tag{5.9}$$

for $N_2(c, j, b)$ in (5.6). This is the distribution of ML independent Dirichlet distributions

$$L_{cj}(p \mid Z, q, \lambda, \alpha) \sim \mathcal{D}_{n_j}(N_2(c, j) + \lambda),$$

where $N_2(c, j) = (N_2(c, j, 1), \cdots, N_2(c, j, n_j)$. For each (c, j) in sequence, $p(c, j, b)$ has the Dirichlet Gibbs-sampler update

$$p(c, j, b) \sim \mathcal{D}_{n_j}(N_2(c, j) + \lambda) \tag{5.10}$$

Updating $q(i, c)$. It follows from (5.8) that the conditional distribution of $q(i, c)$ given the other parameters and variables is

$$L_0(q \mid Z, p, \lambda, \alpha) = C(X, Z, \alpha) \prod_{i=1}^{N} \prod_{c=1}^{M} q(i, c)^{N_3(i,c)+\alpha-1} \tag{5.11}$$

for N_3 in (5.6). This leads to the Gibbs-sampler updates

$$q(i, c) \sim \mathcal{D}_M(N_3(i) + \alpha) \tag{5.12}$$

for $1 \leq i \leq N$ where $N_3(i) = (N_3(i, 1), \cdots, N_3(i, M))$.

Updating $Z(i, a, j)$. It follows from (5.5) that

$$P(Z \mid q, p, X) = \frac{P(Z, q, p, X)}{P(q, p, X)} = C_1 P(X, Z \mid q, p)$$
$$= C_1 \prod_{i=1}^{N} \prod_{a=1}^{A} \prod_{j=1}^{L} \Big(q(i, Z(i, a, j)) p(Z(i, a, j), j, m(i, a, j)) \Big), \tag{5.13}$$

where C_1 depends only on q, p and X. This implies that the $Z(i,a,j)$ are conditionally independent given q, p, X, and that

$$P(Z(i,a,j) = c \mid q, p, X) = C_2 q(i,c) p(c, j, m(i,a,j)). \tag{5.14}$$

As in the non-admixture case (4.19), each $Z(i,a,j)$ can be updated in sequence by a multinomial Bernoulli random variate. This is a discrete Gibbs-sampler update for the probabilities for $1 \le c \le M$ on the right-hand side of (5.14).

Updating $\boldsymbol{\alpha}$. If α is also to be estimated (for example, with a uniform prior on α) we consider the distribution of α conditional on the other parameters and variables. This is

$$L(\alpha \mid q, p, Z, \lambda, X) = C_0 C_M(\alpha)^N \prod_{i=1}^{N} \prod_{c=1}^{M} q(i,c)^{\alpha} \tag{5.15}$$

for $C_M(\alpha) = \Gamma(M\alpha)/\Gamma(\alpha)^M$. We use Metropolis updates for α, as opposed to the Gibbs-sampler updates of q, p and Z [19]. The proposal distribution is a symmetric uniform distribution of length h about α. The parameter h is chosen so that the proportion of time that the new value is accepted in the range 25-40%.

When comparing the likelihood of the newly proposed α to that of the old α, it is often easier to compute $\log L(\alpha)$ instead of $L(\alpha)$. Here

$$\begin{aligned} \log L(\alpha) &= N \log C_M(\alpha) + \alpha \sum_{i=1}^{N} \sum_{c=1}^{M} \log(q(i,c)) \qquad (5.16) \\ &= N(\log \Gamma(M\alpha) - M \log \Gamma(\alpha)) + \alpha Q(q), Q(q) = \sum_{i=1}^{N} \sum_{c=1}^{M} \log(q(i,c)) \end{aligned}$$

On one hand, $\log L(\alpha)$ is easier to compute. But on the other, the standard algorithm for generating Dirichlet random variables using gamma distributions can underflow to zero. Attempting to compute $\log(q(i,c))$ in (5.16) with $q(i,c) = 0$ may cause the

computer to crash. Usually, this only happens if α is small, for example $\alpha < 0.005$. The solution is to either replace $q(i,c) = 0$ by the smallest positive number that can be handled by the computer, or to put a lower bound on the prior of α.

Updating $\boldsymbol{\lambda}$. If λ is also to be estimated (for example, using a uniform prior) we consider the distribution of λ from the likelihood (5.8) conditional on q, p, Z and α. By (5.8), the conditional distribution is a constant times

$$L(\lambda) = \Big((\prod_{j=1}^{L} C_{n_j}(\lambda))^M \Big) \prod_{c=1}^{M} \prod_{j=1}^{L} \prod_{b=1}^{n_j} p(c,j,b)^{\lambda} \tag{5.17}$$

for $C_d(\lambda) = \Gamma(d\alpha)/\Gamma(\alpha)^d$. Thus

$$\log L(\lambda) = M \sum_{j=1}^{L} log(C_{n_j}(\lambda)) + \lambda \sum_{c=1}^{M} \sum_{j=1}^{L} \sum_{b=1}^{n_j} \log(p(c,j,b)),$$

where $\log C_d(\lambda) = \log \Gamma(d\alpha) - d \log \Gamma(\alpha)$. The same considerations about the possibility of small values of λ or $p(c,j,b)$ apply here as in the previous steps.

6 Dirichlet Process Priors

This section introduces a different approach to mixture models. Rather than assuming a fixed number of M of mixture components as in Section 4 and Section 5 and estimating mixture coefficients $q(c)$ and within-component parameters $p(c, j, b)$ for those fixed numbers of components, we initially associate multidimensional parameters $\theta_i = \{p(i, j, b)\}$ to individuals. The parameters θ_i will be updated in a way that causes them to form clusters of identical values. The mixture components will then be the unique values of θ_c, or equivalently individual atoms or tie groups in the sample distribution of the θ_i. In principle, this allows the data to determine the number of clusters as well as the cluster parameters.

Suppose Ω is a set and $\mathscr{A}$ is a σ-algebra of subsets of Ω. Assume G_0 is a probability measure on the space $(\Omega, \mathscr{A})$ and $\mu \in \mathbb{R}$ is a positive number. The DP (Dirichlet process) G with parameter μG_0 is a random probability on the space $(\Omega, \mathscr{A})$ such that for any measurable partition $(B_1, \cdots, B_k)$ the joint distribution of $(G(B_1), \cdots, G(B_k))$ is the Dirichlet distribution $\mathcal{D}(\mu G_0(B_1), \cdots, \mu G_0(B_k))$. G can be viewed as distribution on the set of distributions on $(\Omega, \mathscr{A})$ and it has the following important properties [7]:

(1) For any measurable set $A \in \mathscr{A}$, we have $E[G(A)] = G_0(A)$.

(2) If $X_1, \cdots, X_n$ is a sample of size n from a single realization of G, then the conditional distribution of G given $X_1, \cdots, X_n$ is a Dirichlet process with parameter $\mu G_0 + \sum_{i=1}^{n} \delta_{X_i}$.

(3) G is, with probability one, a discrete distribution on $(\Omega, \mathscr{A})$ with a finite number of atoms.

The DP can also be defined equivalently as a stick breaking process [27], or as a Chinese restaurant process, or through Polya urn schemes [2].

It is often helpful to model an unknown complicated distribution as a mixture of simpler ones. However, making inference on the number of mixing components is challenging since the definition of clusters is subjective and vague under most circumstances. In this scenario, we can avoid specifying the number of mixing components by using the Dirichlet process prior [4, 21]. For example, suppose the data set consists of n data points $x_1, \ldots, x_n$ and each has the distribution of the form $F(\theta)$ with the parameter θ mixed over the distribution G. Then the model has the form

$$\begin{aligned} x_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i &\sim G \\ G &\sim \mathrm{DP}(\alpha G_0) \end{aligned}$$

where θ_i and θ_j are not necessarily distinct for $i \neq j$. (For general reference for Dirichlet processes in probability and statistics, see Ferguson 1973 [7] , Ferguson 1974 [8], Antoniak 1974 [1]; see also Blackwell and MacQueen 1973 [2] ; Kingman 1975 [16]; Sawyer and Hartl 1985 [26]; Ibrahim *et al.* 2001 [15]; Gill 2008 [12])

7 A Mixture Model with Dirichlet Process Prior

Let i denote the individual, a the ploidy, j the locus and b the genotype. Assume that we have allelic value data for N individuals in the form of

$$m(i,a,j) = b \quad \text{for} \quad 1 \leq i \leq N, \quad 1 \leq a \leq A, \quad 1 \leq j \leq L, \quad 1 \leq b \leq n_j. \tag{7.1}$$

The ancestral population c is characterized by a set of allele frequencies

$$\theta_c = \{p(c,j,b) : 1 \leq j \leq L, \quad 1 \leq b \leq n_j\}.$$

Suppose the observed data for the i^{th} individual is

$$X_i = m(i,a,j) \quad \text{with} \quad 1 \leq a \leq A, \quad 1 \leq j \leq L.$$

In the non-admixture model, every individual is considered as an entirely pure descendent from one ancestral population. We let θ_i denote the allele frequencies for the i^{th} individual's ancestral population. Then

$$Pr(X_{i,a,j} \mid \theta_i) = p(i,j,m(i,a,j)).$$

Recall that we assume all loci are unlinked and linkage equilibrium is achieved within populations. Otherwise said,by the Hardy-Weinberg law the genotypes of an individual are sampled independently according to the allele frequencies of its ancestral population. Then the model has the form

$$Pr(X_i \mid \theta_i) = \prod_{a=1}^{A} \prod_{j=1}^{L} p(i,j,m(i,a,j)) = \prod_{j=1}^{L} \prod_{b=1}^{n_j} p(i,j,b)^{N_2(i,j,b)}, \tag{7.2}$$

where

$$N_2(i,j,b) = \#\{a : m(i,a,j) = b\}.$$

Thus

$$\sum_{b=1}^{n_j} N_2(i,j,b) = A \quad \text{and} \quad 0 \le N_2(i,j,b) \le A.$$

Our goal is to find a suitable prior distribution for the parameters θ_i for $1 \le i \le N$ and estimate the number of ancestral populations, i.e. the number of distinct values of the θ_i. Here we introduce some more notation to streamline the model. Let $X = \{X_1, \ldots, X_N\}$ and $\theta_{-i} = \{\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_N\}$. The full conditional likelihood of θ_i can be written as

$$\begin{aligned}
\pi(\theta_i \mid \theta_{-i}, X) &= \frac{f(\theta_i, X \mid \theta_{-i})}{f(X \mid \theta_{-i})} \\
&\propto f(X \mid \theta_1, \cdots, \theta_N) f(\theta_i \mid \theta_{-i}) \\
&= \prod_{k=1}^{N} f(X_k \mid \theta_k) f(\theta_i \mid \theta_{-i}) \\
&\propto f(X_i \mid \theta_i) f(\theta_i \mid \theta_{-i}),
\end{aligned}$$

where the first factor is written out explicitly as the product of corresponding frequencies (7.2) and the second factor is derived as below.

Suppose we observe N independent samples $\theta = \{\theta_1, \cdots, \theta_N\}$ from a single realization of Dirichlet distribution $G \sim \mathrm{DP}(\mu G_0)$. Let G^* denote the conditional distribution of G given θ_{-N}, the first $N-1$ samples. The new process G^* is still a Dirichlet process with a new base probability measure parameter $\frac{\mu G_0 + \sum_{i=1}^{N-1} \delta_{\theta_i}}{\mu + N - 1}$ by property (2) of DP, namely

$$G^* \sim \mathrm{DP}(\mu G_0 + \sum_{i=1}^{N-1} \delta_{\theta_i}) = \mathrm{DP}((\mu + N - 1)\frac{\mu G_0 + \sum_{i=1}^{N-1} \delta_{\theta_i}}{\mu + N - 1}).$$

By property (1) of DP, the second factor now can be written as

$$\begin{aligned} Pr(\theta_N \in d\theta \mid \theta_{-N}) &= E(G^*(d\theta)) \\ &= \frac{\mu G_0 + \sum_{i=1}^{N-1} \delta_{\theta_i}}{\mu + N - 1}(d\theta) \\ &= \frac{\mu G_0(d\theta) + \sum_{i=1}^{N-1} \delta_{\theta_i}(d\theta)}{\mu + N - 1} \end{aligned}$$

Since the θ_i are allele frequencies, we let G_0 be $\mathcal{D}(\lambda)$, the Dirichlet distribution with parameter λ and $g_0 = dG_0$, the corresponding p.d.f. of G_0. That is,

$$\begin{aligned} g_0(\theta_i) &= \prod_{j=1}^{L} D_{n_j}(\lambda)(\theta_{ij}) \\ &= \prod_{j=1}^{L} \frac{\Gamma(\lambda n_j)}{(\Gamma(\lambda))^{n_j}} \prod_{j=1}^{n_j} p(i,j,b)^{\lambda-1}. \end{aligned}$$

Thus the Gibbs sampler for θ_i is a mixture distribution of a Dirichlet distribution and Dirac functions, from which it is convenient to sample.

$$\begin{aligned} &\pi(\theta_i \mid \theta_{-i}, X) \\ &\propto \mu g_0(\theta_i) f(X_i \mid \theta_i) + \sum_{k=1, k\neq i}^{N} f(X_i \mid \theta_k)\delta_{\theta_k}(d\theta_i) \\ &= \mu C(\lambda, X_i) \prod_{j=1}^{L} D_{n_j}(N_2(i,j) + \lambda)(\theta_{ij}) + \sum_{k=1, k\neq i}^{N} f(X_i \mid \theta_k)\delta_{\theta_k}(d\theta_i), \end{aligned}$$

where $N_2(i,j) = \Big(N_2(i,j,1), \cdots, N_2(i,j,n_j)\Big)$

$$C(\lambda, X_i) = \prod_{j=1}^{L} \Big(\frac{\prod_{b=1}^{n_j} \lambda^{(N_{2\nu}(i,j,b))}}{(n_j\lambda)^{(A)}}\Big)$$

and

$$\lambda^{(n)} = \lambda(\lambda + 1) \cdots (\lambda + n - 1)$$

is the increasing factorial power.

8 M-Component Mixture Model with a Dirichlet Process Prior

In the non-admixed mixture model of Section 7, the ancestral population of the i^{th} individual is implicitly indicated by the corresponding allele frequencies θ_i. The θ_i are updated in sequence. In effect the individuals is assigned to the group whose members share the same allele frequencies. The maximum possible number of populations is the total number of individuals. On one hand, this allows for variety in the number of populations, but on the other, it can markedly slow the convergence rate of the MCMC procedure and, in practice, tends to lead to many small classes that are difficult to interpret. We will introduce a few intermediate steps into the model to control the number of populations and improve the mixing of the MCMC.

As in previous models, suppose X is the genotype data of N individuals at L loci with ploidy A and the ath allele of the i^{th} individual at j^{th} locus has value b. Then X takes the form $X = \{m(i,a,j) = b\}$ where $i = 1,\ldots,N$, $a = 1,\ldots,A$, $j = 1,\ldots,L$ and $b = 1,\ldots,n_j$. In addition, there are M unknown background populations. Each individual is assumed to be a pure descendant from one of them. Let $Z(i) = c$ for $1 \le c \le M$ denote the origin population for the i^{th} individual. The background population c is characterized by the allele frequencies $\theta_c = (p_1,\ldots,p_L)$ at L loci, where $p_j = (p(j,1),\ldots,p(j,n_j))$. The set $\theta = (\theta_1,\ldots,\theta_M)$ constitutes the allele frequencies of all populations. The likelihood of X given the hidden variable Z and the parameter θ is

$$P(X \mid Z,\theta) = \prod_{i=1}^{N}\prod_{a=1}^{A}\prod_{j=1}^{L} p\big(Z(i),j,m(i,a,j)\big) \tag{8.1}$$

$$= \prod_{c=1}^{M}\prod_{j=1}^{L}\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)}. \tag{8.2}$$

where

$$N_2(c, j, b) = \#\{(i, a) : Z(i) = c \text{ and } m(i, a, j) = b\}. \tag{8.3}$$

The posterior distribution of θ and Z can be estimated by MCMC method through Gibbs-sampler updates in a fashion similar to that used in the simple pure mixture models. Let $\theta_{-c} = (\theta_1, \ldots, \theta_{c-1}, \theta_{c+1}, \ldots, \theta_M)$ denote the allele frequencies omitting the c^{th} population. The full conditional distribution of θ_c is

$$\begin{aligned} P(\theta_c \mid X, Z, \theta_{-c}) &= \frac{P(\theta_c, \theta_{-c}, X, Z)}{P(X, Z, \theta_{-c})} \\ &\propto P(\theta, X, Z) \\ &\propto P(X \mid \theta, Z)P(\theta, Z) \\ &\propto P(X \mid \theta, Z)P(\theta)P(Z) \\ &\propto P(X \mid \theta, Z)P(\theta) \end{aligned} \tag{8.4}$$

since by (3.6) one can assume that the hidden variable Z have a prior that is independent of the prior of θ.

The first factor in (8.4)is

$$\begin{aligned} P(X \mid \theta, Z) &= \prod_{d=1}^{M} \prod_{j=1}^{L} \prod_{b=1}^{n_j} p(d, j, b)^{N_2(d,j,b)} \\ &\propto \prod_{j=1}^{L} \prod_{b=1}^{n_j} p(c, j, b)^{N_2(c,j,b)} \quad \text{conditional on} \theta_c \\ &\triangleq \phi_c(\theta_c) \end{aligned}$$

for $N_2(c, j, b)$ in (8.3). The second factor (8.4) has the same form as in the simple

pure mixture model

$$\begin{aligned} P(\theta) &\propto P(\theta_c \mid \theta_{-c}) \\ &\propto \frac{\mu g_0(\theta_c) + \sum_{d=1,d\neq c}^{M} \delta_{\theta_d}(\mathrm{d}\theta_c)}{\mu + M - 1}. \end{aligned}$$

The product of these factors then gives the Gibbs-sampler for θ_i

$$P(\theta_c \mid X, Z, \theta_{-c}) \propto \mu\phi_c(\theta_c)g_0(\theta_c) + \sum_{d=1,d\neq c}^{M} \phi_c(\theta_d)\delta_{\theta_d}(\mathrm{d}\theta_c), \tag{8.5}$$

where $\phi_c(\theta_d) = \prod_{j=1}^{L}\prod_{b=1}^{n_j} p(d,j,b)^{N_2(c,j,b)}$ and μg_0 is the scale measure of the Dirichlet process prior. To put the Gibbs-sampler in a more transparent form, we rewrite the first part as a Dirichlet distribution multiplied by constants that only depend on some

allele counts.

$$
\begin{aligned}
& \phi_c(\theta_c) g_0(\theta_c) \qquad (8.6)\\
= & \prod_{j=1}^{L} \prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)} \quad \prod_{j=1}^{L} \frac{\Gamma(n_j\lambda)}{(\Gamma(\lambda))^{n_j}} \prod_{b=1}^{n_j} p(c,j,b)^{\lambda-1} \\
= & \prod_{j=1}^{L} \frac{\Gamma(n_j\lambda)}{(\Gamma(\lambda))^{n_j}} \quad \prod_{j=1}^{L} \prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)+\lambda-1} \\
= & \prod_{j=1}^{L} \frac{\Gamma(n_j\lambda)}{(\Gamma(\lambda))^{n_j}} \quad \prod_{j=1}^{L} \frac{\prod_{b=1}^{n_j} \Gamma(N_2(c,j,b)+\lambda)}{\Gamma(N_{2s}(c,j)+n_j\lambda)} \times \\
& \prod_{j=1}^{L} \frac{\Gamma(N_{2s}(c,j)+n_j\lambda)}{\prod_{b=1}^{n_j} \Gamma(N_2(c,j,b)+\lambda)} \prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)+\lambda-1} \\
= & \prod_{j=1}^{L} \frac{\Gamma(n_j\lambda)}{(\Gamma(\lambda))^{n_j}} \frac{\prod_{b=1}^{n_j} \Gamma(N_2(c,j,b)+\lambda)}{\Gamma(N_{2s}(c,j)+n_j\lambda)} \times \\
& \mathcal{D}(N_2(c,j,1)+\lambda-1,\ldots,N_2(c,j,n_j)+\lambda-1)(p(c,j,1),\ldots,p(c,j,n_j)) \\
= & \prod_{j=1}^{L} \frac{\big(\lambda(\lambda+1)\ldots(\lambda+N_2(c,j,1)-1)\big)\ldots\big(\lambda(\lambda+1)\ldots(\lambda+N_2(c,j,n_j)-1)\big)}{n_j\lambda(n_j\lambda+1)\ldots(n_j\lambda+N_{2s}(c,j)-1)} \times \\
& \mathcal{D}(N_2(c,j,1)+\lambda-1,\ldots,N_2(c,j,n_j)+\lambda-1)(p(c,j,1),\ldots,p(c,j,n_j))
\end{aligned}
$$

where

$$
\begin{aligned}
N_{2s}(c,j) = \sum_{b=1}^{n_j} N_2(c,j,b) &= \sum_{b=1}^{n_j} \#\{(i,a) : Z(i)=c \text{ and } m(i,a,j)=b\} \\
&= A \times \#\{i : Z(i)=c\}. \qquad (8.7)
\end{aligned}
$$

Finally, the posterior distribution of the hidden variable Z is

$$\begin{aligned} P(Z \mid \theta, X) &= \frac{P(Z, \theta, X)}{P(\theta, X)} \\ &= P(X \mid \theta, Z)\frac{P(\theta, Z)}{P(\theta, X)} \\ &= P(X \mid \theta, Z)\frac{P(\theta)P(Z)}{P(\theta, X)} \\ &\propto P(X \mid \theta, Z) \end{aligned}$$

Since by (3.6) we can assume Z has the non-informative prior $P(Z(i) = c) = 1/M$, independent of θ. Here the hidden variable Z is updated as a multinomial variable according to

$$P(Z(i) = c \mid \theta, X) \propto \prod_{j=1}^{L}\prod_{a=1}^{A} p\big(c, j, m(i, a, j)\big) \tag{8.8}$$

To summarize, we first initialize $\theta_c = (p_1, \ldots, p_L), 1 \le c \le M$, by sampling from Dirichlet distributions. That is, $p_j \sim D(\alpha_{j1}, \ldots, \alpha_{jn_j})$ where the parameter α_{jb} is any positive constant or can be updated also through a Metropolis-Hastings update. Z is initialized from a uniform distribution, and then θ and Z are updated iteratively as described above in (8.5) and (8.8).

8.1 Three-step update

We can add one more step to the model to assist the update of the hidden variable Z. Assume that the i^{th} individual has probability $q(i, c)$ of being from the c^{th} population:

$$P(Z(i) = c \mid \theta, q) = q(i, c), \quad 1 \le c \le M.$$

Naturally the sum of these probabilities over all the populations is 1: $\sum_{c=1}^{M} q(i, c) = 1$. Let us use the short-hand notations $q_i = (q(i, 1), \ldots, q(i, M))$, $i = 1, \ldots, N$ and

$q = (q_1, \ldots, q_N)$. The likelihood of the data X given Z, q and θ can be written as

$$\begin{aligned} P(X \mid \theta, q, Z) &= P(X \mid \theta, Z) \\ &= \prod_{i=1}^{N} \prod_{a=1}^{A} \prod_{j=1}^{L} p\big(Z(i), j, m(i,a,j)\big) \\ &= \prod_{c=1}^{M} \prod_{j=1}^{L} \prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)}, \end{aligned}$$

where $N_2(c,j,b)$ is defined as in (8.3). The variable θ_c has the same Gibbs-sampler (8.5) as in the two-step update model

$$\begin{aligned} P(\theta_c \mid X, Z, \theta_{-c}, q) &= \frac{P(\theta_c, X, Z, \theta_{-c}, q)}{P(X, Z, \theta_{-c}, q)} \\ &= \frac{P(\theta, X, Z, q)}{P(X, Z, \theta_{-c}, q)} \\ &\propto P(X \mid \theta, Z, q) P(\theta, Z, q) \\ &\propto P(X \mid \theta, Z, q) P(Z \mid \theta, q) P(\theta \mid q) \pi_0(q) \\ &= P(X \mid \theta, Z, q) P(Z \mid q) P(\theta \mid q) \pi_0(q) \\ &\propto P(X \mid \theta, Z, q) P(\theta \mid q) \\ &\propto P(X \mid \theta, Z, q) P(\theta). \end{aligned}$$

The posterior probability of Z has the form

$$\begin{aligned}
P(Z(i) = c \mid \theta, q, X) &= \frac{P(Z(i) = c, \theta, q, X)}{P(\theta, q, X)} \\
&= C_1(\theta, q, X) P(Z(i) = c, \theta, q, X) \\
&= C_1(\theta, q, X) P(X \mid Z(i) = c, \theta, q) P(Z(i) = c, \theta, q) \\
&= C_1(\theta, q, X) P(X \mid Z(i) = c, \theta, q) P(Z(i) = c \mid \theta, q) P(\theta, q) \\
&= C_2(\theta, q, X) P(X \mid Z(i) = c, \theta, q) P(Z(i) = c \mid \theta, q) \\
&= C_2(\theta, q, X) \big(\prod_{j=1}^{L} \prod_{a=1}^{A} p(c, j, m(i, a, j)) \big) q(i, c).
\end{aligned}$$

Since $C_2(\theta, q, X)$ is independent of i, the full conditional distribution for Z is

$$P(Z(i) = c \mid \theta, q, X) = \frac{q(i, c) \big(\prod_{j=1}^{L} \prod_{a=1}^{A} p(c, j, m(i, a, j)) \big)}{\sum_{d=1}^{M} q(i, d) \big(\prod_{j=1}^{L} \prod_{a=1}^{A} p(d, j, m(i, a, j)) \big)}.$$

The full conditional distribution for q is given by

$$\begin{aligned}
P(q \mid \theta, Z, X) &\propto P(q, \theta, Z, X) \\
&= P(X \mid q, \theta, Z) P(q, \theta, Z) \\
&= P(X \mid \theta, Z) P(q, \theta, Z) \\
&\propto P(Z \mid q, \theta) P(q, \theta) \\
&\propto P(Z \mid q, \theta) P(q) \\
&\propto P(Z \mid q) \pi_0(q) \pi_0(\theta) \\
&\propto \prod_{i=1}^{N} q(i, Z(i)) \prod_{i=1}^{N} \pi_0(q_i).
\end{aligned}$$

We assume q_i has the Dirichlet prior distribution $D(\alpha, \ldots, \alpha)$. The Gibbs-sampler for q_i is $q_i \sim D(\alpha, \ldots, \alpha + 1, \ldots, \alpha)$ with the c^{th} parameter equal to $\alpha + 1$ if $Z(i) = c$ i.e. the i^{th} individual is assigned to the c^{th} population.

In practice, a larger value of M (the initial number of components) yields more clusters for the same data. Those clusters could be combined, using methods like hierarchical clustering to form larger groups using the pairwise distance between the smaller clusters. For example, Medvedovic *et al.* 2002 [18] suggested using a complete linkage clustering. Clusters were generated by separating the profiles in groups with the maximum possible complete linkage distance. That is, for any pair of clusters formed, there was at least one profile in the first cluster that had a zero posterior pairwise probability and at least one profile in the second cluster [18].

Alternatively, we have found that setting M at a reasonable upper bound for the expected number of clusters often leads to a stable estimate for the number of clusters. We will discuss this further in Chapter 10 when we describe our simulations.

9 An Admixture Model with a Dirichlet Process Prior

We now give an account of the Gibbs-sampler for q and θ, and again indicate the updating procedures as in the previous section. The data we have consists the allelic values of N individuals

$$\begin{aligned} m(i,a,j) &= b \text{ for } 1 \le i \le N, \quad 1 \le a \le A, \\ & \qquad 1 \le j \le L, \quad 1 \le b \le n_j. \end{aligned}$$

Assume that the sample data come from a possible maximum of M cryptic populations or mixture components. For example, we could set $M = N$. However, a choice of M which is closer to the correct value helps with the convergence of MCMC procedure and provides better clustering. In the admixed model, each allele of the individual is chosen independently from the c^{th} population $1 \le c \le M$ with probability $q(i,c)$, as opposed to all alleles being chosen from the same c^{th} population, with probability $q(i,c)$. As in previous sections, each of the M populations has parameters $\theta_c = \{p(c,j,b)\}$, where $p(c,j,b)$ is the probability that the b^{th} allele is observed at the j^{th} locus.

As before, the allele frequencies $\theta_1, \ldots, \theta_M$ are M independent samples from the same realization of a Dirichlet process G with the base measure G_0 which follows a Dirichlet distribution. We also assume as part of the prior that

$$q(i,c) \sim \mathcal{D}_M(\alpha, \cdots, \alpha) \text{ for each } i.$$

Here we introduce hidden variables

$$Z(i,a,j) = c \ \ (1 \le c \le M)$$

for the ancestral population associated with the allele (i,a,j). The joint likelihood of the data X and the hidden variables Z conditional on $\theta = (\theta_1, \ldots, \theta_M)$ and q is

$$\begin{aligned} P(X, Z \mid \theta, q) &= \prod_{i=1}^{N} \prod_{a=1}^{A} \prod_{j=1}^{L} q(i, Z(i,a,j)) p(Z(i,a,j), j, m(i,a,j)) \\ &= \prod_{i=1}^{N} \prod_{c=1}^{M} q(i,c)^{N_3(i,c)} \prod_{c=1}^{M} \prod_{j=1}^{L} \prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)}, \end{aligned}$$

where

$$N_2(c,j,b) = \#\{(i,a) : Z(i,a,j) = c \text{ and } m(i,a,j) = b\} \tag{9.1}$$

$$N_3 = \#\{(a,j) : Z(i,a,j) = c\}, \tag{9.2}$$

and the joint posterior probability is

$$f(\theta, q, Z \mid X) = C_X P(X, Z \mid \theta, q) \pi_0(\theta) \pi_0(q). \tag{9.3}$$

The parameters $q(i,c)$, θ_c and the hidden variables $Z(i,a,j)$ can be estimated by a Markov chain with Gibbs-sampler updates. In principle the parameters α, μ and λ could be also updated by a Metropolis-Hastings step. We initialized the hidden variable $Z(i,a,j)$ as uniform in $1, 2, \ldots, M$ independently for all triples (i,a,j). The variables $\theta = (\theta_1, \ldots, \theta_M)$ can be initialized by, for instance, independent uniforms normalized so that $\sum_{b=1}^{n_j} \theta(c,j,b) = 1$ for all (c,j). The initializations for $q(i,c)$ are

not important since the initial values will be immediately overwritten by a Gibbs-sampler update.

Gibbs sampler for q. It follows from (9.3) that the full conditional distribution of $q(i,c)$ is

$$f(q \mid \theta, Z, X) = C(X, Z, \alpha) \prod_{i=1}^{N} \prod_{c=1}^{M} q(i,c)^{N_3(i,c)+\alpha-1}$$

for $N_3(i,c)$ in (9.2). This leads to the Gibbs-sampler for $q(i,c)$

$$q(i,c) \sim \mathcal{D}_M(N_3(i) + \alpha)$$

for $1 \le i \le N$, where $N_3(i) = (N_3(i,1), \ldots, N_3(i,M))$ for $N_3(i,c)$ in (9.2).

Gibbs sampler for θ. Following previous models, we update the parameter $\theta_c = \{p(c,j,b)\}$ in sequence. The full conditional distribution of θ_c given the other parameters (including those in $\theta_{-c} = \{\theta_d : d \neq c\}$) is

$$\begin{aligned} f(\theta_c \mid X, q, Z, \theta_{-c}) &= \frac{f(\theta_c, \theta_{-c}, q, X, Z)}{f(\theta_{-c}, q, X, Z)} = C_1 f(\theta, q, X, Z) \\ &= C_2 f(X, Z \mid \theta, q) f(\theta, q) \end{aligned} \tag{9.4}$$

where we have combined factors that depend only on q, X, Z, and θ_{-c} into constants C_1 and C_2. Recall that the θ_c are not independent for different c. The last factor above is

$$P(\theta_c \mid \theta_{-c}, q) = \big(\mu g_0(\theta_c) + \sum_{d=1, d\neq c}^{M} \delta_{\theta_d}(d\theta_c)\big)/(\mu + M - 1).$$

By (9.1), the other main factor in (9.4) is

$$\begin{aligned} P(X, Z \mid \theta, q) &= \Big(\prod_{i=1}^{N}\prod_{c=1}^{M} q(i,c)^{N_3(i,c)}\Big)\Big(\prod_{c=1}^{M}\prod_{j=1}^{L}\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)}\Big) \\ &= C_3(\theta_{-c}, q)\phi_c(\theta_c) \end{aligned}$$

where

$$\phi_c(\theta_d) = \prod_{j=1}^{L}\prod_{b=1}^{n_j} p(d,j,b)^{N_2(c,j,b)}. \tag{9.5}$$

Thus by (9.4) we have that

$$P(\theta_c \mid q, X, Z, \theta_{-c}) = C_4\big(\mu\phi_c(\theta_c)g_0(\theta_c) + \sum_{d=1, d\neq c}^{M} \phi_c(\theta_d)\delta_{\theta_d}(d\theta_c)\big). \tag{9.6}$$

Since $\theta_c = \{p(c,j,b)\}$,

$$\begin{aligned} \phi_c(\theta_c)g_0(\theta_c) &= \prod_{j=1}^{L}\frac{\Gamma(n_j\lambda)}{\Gamma(\lambda)^{n_j}}\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)+\lambda-1} \\ &= \Big(\prod_{j=1}^{L}\frac{\Gamma(n_j\lambda)}{\Gamma(\lambda)^{n_j}}\Big)\Big(\prod_{j=1}^{L}\frac{\prod_{b=1}^{n_j}\Gamma(N_2(c,j,b)+\lambda)}{\Gamma(N_{2s}(c,j)+n_j\lambda)}\Big) \\ &\quad \times \prod_{j=1}^{L}\frac{\Gamma(N_{2s}(c,j)+n_j\lambda)}{\prod_{b=1}^{n_j}\Gamma(N_2(c,j,b)+\lambda)}\prod_{b=1}^{n_j} p(c,j,b)^{N_2(c,j,b)+\lambda-1} \\ &= C(c,\lambda,X,Z)\prod_{j=1}^{L}\mathcal{D}_{n_j}(N_{2v}(c,j)+\lambda)(p(c,j)) \end{aligned}$$

where $\mathcal{D}_n(\lambda_1, \cdots, \lambda_n)(x_1, \cdots, x_n)$ is the Dirichlet density and

$$\begin{aligned} C(c,\lambda) = C(c,\lambda,X,Z) &= \prod_{j=1}^{L}\Big(\frac{\prod_{b=1}^{n_j}\lambda^{(N_2(c,j,b))}}{(n_j\lambda)^{(N_{2s}(c,j))}}\Big) \\ N_{2v}(c,j) &= (N_2(c,j,1), \cdots, N_2(c,j,n_j)) \\ N_{2s}(c,j) &= \sum_{b=1}^{n_j} N_2(c,j,b) = \#\{(i,a) : Z(i,a,j) = c\} \end{aligned}$$

where $\lambda^{(n)} = \lambda(\lambda+1)\cdots(\lambda+n-1)$ denotes the factorial power. Thus by (9.6)

$$P(\theta_c \mid q, X, Z, \theta_{-c}) = C_4 \cdot \big(\mu C(c,\lambda) \prod_{j=1}^{L} \mathcal{D}_{n_j}(N_{2v}(c,j)+\lambda)(\theta_c) + \sum_{d=1, d\neq c}^{M} \phi_c(\theta_d)\delta_{\theta_d}(d\theta_c)\big).$$

Gibbs-sampler updates can be carried out in sequence for θ_c by sampling from a mixture, one of which is a product of Dirichlet densities, in a similar manner as in previous section.

Updating $Z(i,a,j)$. It follows from (9.1) that the conditional density of Z given all of the other parameters and variables is

$$\begin{aligned} P(Z \mid \theta, q, X) &= \frac{P(Z,\theta,q,X)}{P(\theta,q,X)} = C_1(\theta,q,X)P(X,Z \mid \theta,q) \\ &= C_1 \prod_{i=1}^{N}\prod_{a=1}^{A}\prod_{j=1}^{L}\big(q(i,Z(i,a,j))p(Z(i,a,j),j,m(i,a,j))\big) \end{aligned}$$

Thus the $Z(i,a,j)$ are conditionally independent given (θ,q,X), and

$$P(Z(i,a,j)=c \mid \theta,q,X) = C_2 q(i,c)p(c,j,m(i,a,j)) \tag{9.7}$$

This implies that $Z(i,a,j)$ has a Gibbs-sampler multinomial Bernoulli update with discrete probabilities given by the right-hand side of (9.7).

Updating α. If desired, α can also be updated. We assume as a prior that α is uniformly distributed in the range $0 < A_1 \leq \alpha \leq A_2$. The posterior densities of the parameters (9.3) can be extended to

$$P(\theta,q,Z,\alpha \mid X) = C_X P(X,Z \mid \theta,q,\alpha)\pi_0(\theta)\pi_0(q)I_{[A_1,A_2]}(\alpha). \tag{9.8}$$

It follows from (9.1) and (9) that the distribution of α conditional on the other

parameters and the variance can be expressed

$$L(\alpha \mid q, p, Z, \lambda, X) = C_0 \pi_0(q) = C_1 C_M(\alpha)^N \prod_{i=1}^{N} \prod_{c=1}^{M} q(i,c)^{\alpha} I_{[A_1, A_2]}(\alpha) \tag{9.9}$$

within the bounds on α, where $C_M(\alpha) = \Gamma(M\alpha)/\Gamma(\alpha)^M$. We use Metropolis updates for α. Note that $\log L(\alpha)$ will be is easier to work with than $L(\alpha)$ for computing acceptance probabilities. Here

$$\log L(\alpha) = N \log C_M(\alpha) + \alpha \sum_{i=1}^{N} \sum_{c=1}^{M} \log(q(i,c)) = N(\log \Gamma(M\alpha) - M \log \Gamma(\alpha)) + \alpha Q(q) \tag{9.10}$$

where $Q(q) = \sum_{i=1}^{N} \sum_{c=1}^{M} \log(q(i,c))$.

10 Applications to Data

The simulation study is done on data sets generated from a probability model either with or without admixture using standard coalescent techniques, and then the simulation is performed again on a real data set. Although strictly speaking data sets generated from coalescents are not equivalent to those generated from the the non-admixture or admixture model (in the sense that they are from different probability models), nevertheless coalescent techniques are widely used for generating data sets in simulation studies. Data set Testdata1 from Pritchard's paper [23], for example, is used to test their non-admixture probability model.

The Number of Components. In clustering analysis, determining an appropriate value for the number of components K, is a formidable problem. Pritchard *et al* addressed this problem in [23] as an important issue. They examined the performance of the harmonic mean estimator method, but this is computationally infeasible when estimating K because of the high dimension of the data[23]. However, once K is fixed, the rest of the parameter sets of interest can be estimated using MCMC method. The Bayesian paradigm has helped the development of several new model-based parametric methods, the most representative of which can be implemented in the computer program STRUCTURE, whose the algorithm is explained in [23]. The algorithm is implemented as follows. First, all the possible values of K should be specified, together with a prior distribution, usually a non-informative one, such that each population is equally likely to occur, with probability $1/K$. Second, the Gibbs samplers are run independently for each value of K. The posterior distribution of K is calculated using a normal distribution approximation of the *Bayesian deviance*. The K value with the highest posterior probability is then chosen, and the estimation of other parameters from that particular MCMC run is reported.

A couple of strategies have been proposed to determine K automatically in the non-admixture scenario. Stephens *et al.* (2000) introduced a method using birth-death process to model the number of populations. Although computationally feasible, it is not as efficient as Pritchard's *ad hoc* approach. Huelsenbeck and Andolfatto 2007 [14] describe an algorithm in Neal(2000)[21] which involves sampling the number of populations K and the allocations Z jointly under the Dirichlet process prior. At one iteration of the Gibbs sampling, each individual has certain probability to join one of the existing populations and certain probability to form a new population.

Despite its great success in application, this method of STRUCTURE has some drawbacks. First, an independent run is required for each value of K, second, the estimation of the posterior probability of K is *ad hoc* [23], and a common feature of both STRUCTURE and our methods is that the estimation of the key parameter of q (the ancestral population coefficient) depends on the "poor mixing" [23] of the MCMC sampler. Otherwise the elements of q should give identical proportions of the origin populations due to the non-identifiability inherent in the problem using MCMC clustering method.

Our method requires no pre-specified value of K, although our simulation study shows that the number of components yielded by the models with Dirichlet process prior highly depends on the choice of the initial number of initial components (source populations). A large number of components suggests that the model should look into deeper branches of the coalescent structure, and therefore will give a finer partition of the sample.One can apply other methods to combine smaller subpopulations into larger ones, depending on which level one wished to study the populations.

Summarizing output from MCMC. Huelsenbeck and Andolfatto (2007) [14] suggested to find the mean partition with the minimum square distance to all partitions sampled from MCMC. The algorithm was first applied to find the mean partition

as an application of Kuhn's "Hungarian method" in combinatorial optimization [17] by Gusfield [13]. In practice, there usually exists a unique partition with minimum squared distance. Therefore the average of multiple partitions is well defined. Practically, the method is implemented using an greedy algorithm[14]: that is any proposed change that reduces the squared distance of the mean partition is immediately accepted. The convergency of the algorithm depends on whether we observe no updates for "sufficiently long" [14]. Our simulation study shows that time to convergence varies dramatically with partition length and the sample size. Generally speaking, it takes more iterations to possibly converge if the partition is long. For example, it took $\sim$ 133,000 iterations ($\sim$ 144 hours on a computer with Intel i7 processor) to find the mean of 5,000 partitions of length $\sim$ 2,100 and $\sim$ 7,5000 iterations ($\sim$ 350 seconds) of 5 partitions of the same length.

Results for the Taita thrush. We now present results from applying our method to genotype data from an endangered bird species, the Taita thrush, *Turdus helleri*. Individual birds were sampled at four locations in southeast Kenya [Chawia (17 individuals), Ngangao (54), Mbololo (80) and Yale (4)], and each individual was genotyped at seven microsatellite loci [10]. The neighbor joining tree method is used to display the Thrush data as in Figure 10 from Pritchard (2000) [23] which is reproduced as Figure 1 here.

The use of distance-based clustering methods to visualize this sort of genotype data is quite common. The disadvantage is that the method is not based on a probability model, and thus it is difficult to make inference on the population structure.

A crude, but still useful, way of summarizing the thrush data is the following. (Figure 5 below will give a cleaner picture.) Define a distance metric $d_q(i, j)$ by summing the differences of ancestry population proportions Q for all iterations. Otherwise

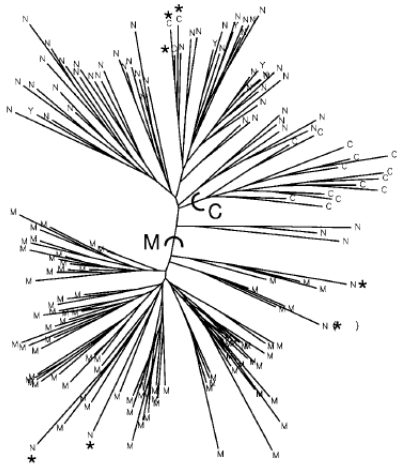


Figure 1: Neighbor-joining tree of individuals in the thrush data set. Each tip represents a single individual. C, M, N and Y indicate the populations of origin. Individuals who appear to be misclassified are marked $*$. The tree was constructed using the program *Neighbor* included in Phylip by Felsenstein [6]. The pairwise distance matrix was computed as follows: for each pair of individuals, we added $1/L$ for each locus at which they had no alleles in common, $1/2L$ for each locus at which they had one allele in common(*e.g.*, AA:AB or AB:AC)and 0 for each locus at which they had two alleles in common (*e.g.*, AA:AA or AB:AB), where L is the number of loci compared.

said,

$$d_q(i,j) = \sum_{t=1}^{T} \sum_{c=1}^{M} \frac{|q(t,i,c) - q(t,j,c)|}{T}$$

where T is the number of sampled iterations. Figure 2 shows a neighbor-joining tree, constructed based on $d_q(i,c)$ with the program *VTSD* by S. Sawyer [20] implementing the neighbor joining method by Saitou and Nei (1987) [25]. The tree in Figure 2 visibly separates three distinct populations. The seven individuals shown as possible outliers in Figure 1 are logically placed in the plot. The long branch lengths before the tips in Figure 2 are due to MCMC noise in $q(t,i,c)$.

The individual thrush are better distinguished by

$$\bar{q}(i,c) = \frac{1}{T} \sum_{t=1}^{T} q(t,i,c)$$

where T is the total number of iterations (See Figure 3). Now the initial number of components is $M = 7$ for this run. The number of clusters converges at $K = 3$, and the plot shows a separation of clusters reasonably consistent with the output from the neighbor-joining tree on genotype data and STRUCTURE.

The distance matrix of each pair of individuals can also be calculated using $\bar{q}(i,c)$. Specifically, let

$$d_{\bar{q}}(i,j) = \frac{1}{M} \sum_{c=1}^{M} |\bar{q}(i,c) - \bar{q}(j,c)|.$$

The corresponding phylogeny using $d_{\bar{q}}$ is shown in Figure 4.

We also apply the neighbor joining tree method to build the phylogeny of the starting components (Figure 5). Figure 5 also has bootstrap support numbers on several links. These are the number of estimated phylogenies for 1000 bootstrap replications of the thrush data (that is, the list of thrush is resampled) that contain that link. It strongly suggests that the number of population K should be 3, which

Figure 2: The neighbor-joining tree of individuals in the thrush data. X, A and B indicate the populations of origin: X for Chawia, A for Mbololo, and B for Ngangao and Yale, instead of using "C", "M", "N" and "Y" for a clearer display. The long branches are due to the MCMC noise. The pairwise distance matrix is computed as follows: for each pair of individuals, add the abusolute difference in their ancestry proportions over populations and MCMC iterations. The initial number of components is set to $M = 7$ for this run. The tree was constructed using the program $VTSD$ by S. Sawyer [20] implementing the neighbor joining method by Saitou and Nei [25].

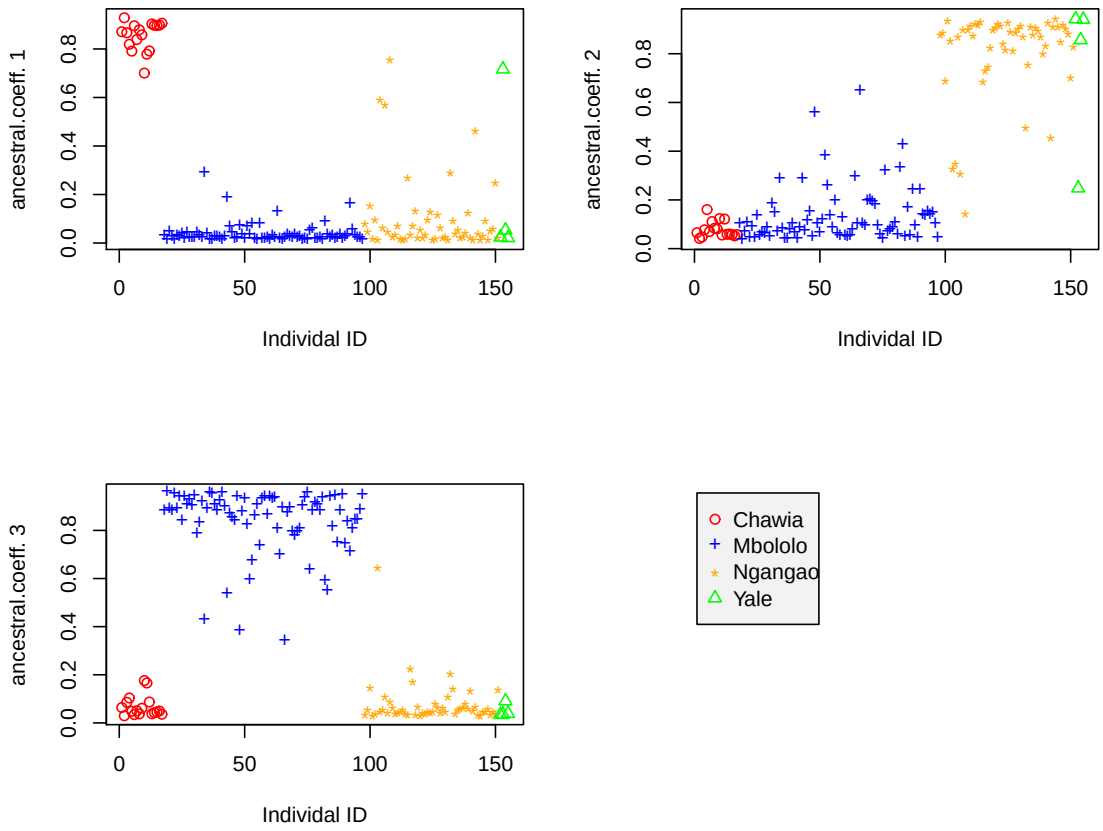


Figure 3: Summary of the clustering results for Taita thrush data set using Dirichlet prior admixture model. For each individual, the mean value of $q(i, c)$ (the proportion of ancestry) is computed over a single run of the MCMC. The points are labeled according to the sampling location. The location information is not used to estimate ancestry.

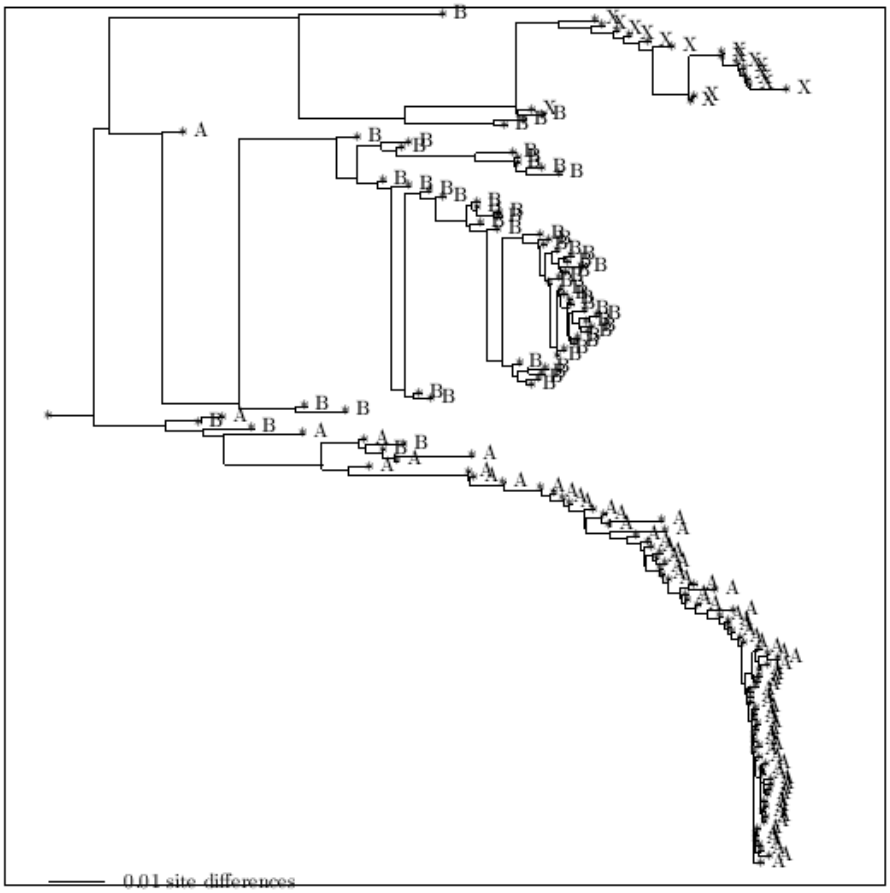


Figure 4: The neighbor-joining tree of individuals in the thrush data. X, A and B indicate the populations of origin: X for Chawia, A for Mbololo, and B for Ngangao and Yale, instead of using "C", "M", "N" and "Y" for a clearer display. The pairwise distance matrix is computed as follows: for each pair of individuals, add the difference in their ancestry proportions over populations and MCMC iterations. The initial number of components is set to $M = 8$ for this run. The tree was constructed using the program $VTSD$ by S. Sawyer [20] implementing the neighbor joining method by Saitou and Nei [25].

is more difficult to see directly from Figure 1.

The no-admixture model with DP prior. Testdata1 from the simulation study in Pritchard *et al.* (2000) [23], is available on STRUCTURE homepage [22]. It consists of microsatellite-length data for 2 populations with 100 individuals each at 5 loci.

```
Distinct allelic values for data read from  testdata1.txt
  Locus 01: (n= 9)  -5 -4 -3 -2 -1  0  1  2  3
  Locus 02: (n=12)  -4 -3 -2 -1  0  1  2  3  4  5  6  7
  Locus 03: (n=11)  -3 -2 -1  0  1  2  3  4  5  6  7
  Locus 04: (n=14)   2  3  4  5  6  7  8  9 10 13 14 15 16 17
  Locus 05: (n=15)  -5 -4 -3 -2 -1  0  1  2  3  4  5  6  7  8  9
```

We use the Dirichlet process prior model with no admixture (that is, no inter-breeding occurs among populations) and first set the starting number of components equal to the number of individuals (200). After 10k burn-ins and 500k iterations, the number of components K oscillates around 10. Medvedovic and Sivaganesan (2002) [18] and Rodriguez and Papaspiliopoulos (2009) [24] suggest calculating the proportion $p(i,j)$ of pairwise joint classification defined in (10.1) of (θ_i, θ_j) for all $1 \leq i, j \leq N$ and applying a hierarchical clustering analysis method on $p(i,j)$. We call the matrix with entries $p(i,j)$ the *pairwise posterior probability*, and the matrix with entries $1-p(i,j)$ the *MCMC distance matrix*.

$$p(i,j) \;=\; \frac{\#\text{ of iterations with } \theta_i \text{ and } \theta_j \text{ in the same cluster}}{\#\text{ of iterations}}. \qquad (10.1)$$

Usually the hierarchical clustering method requires a criterion to aggregate smaller clusters into bigger ones, or a partition of larger clusters into smaller ones [5]. This data set contains a great deal of information about the population origin, as illustrated in [23]. As a result, the clustering result is robust to the choice of different criterion

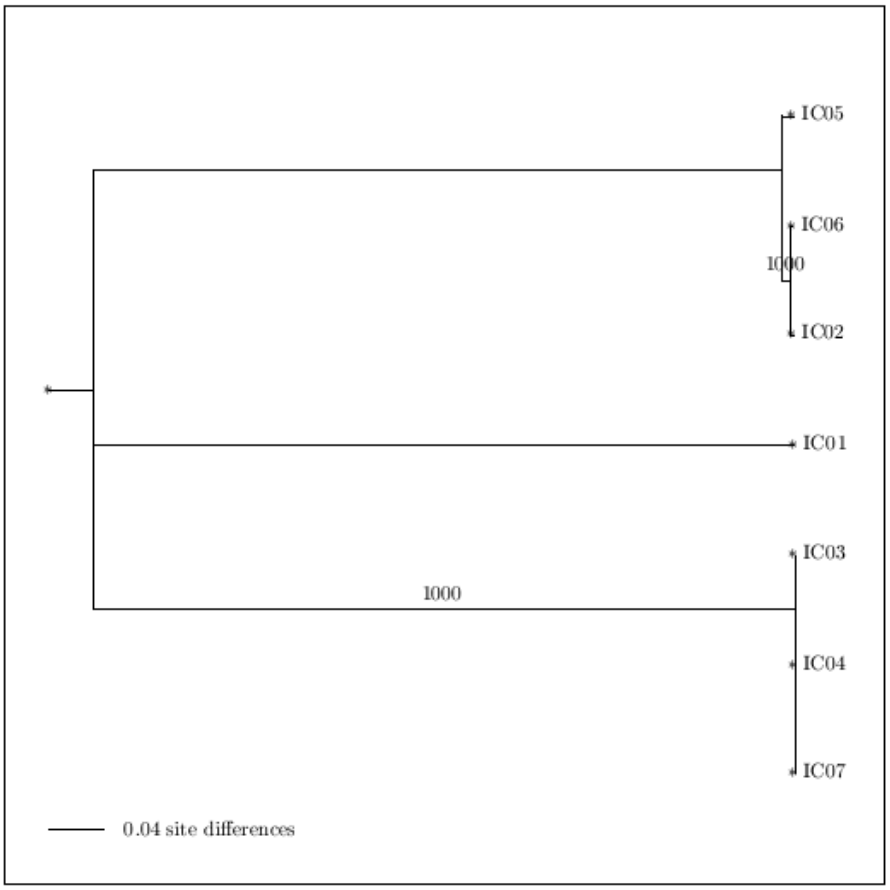


Figure 5: The neighbor joining tree of ancestral populations with the initial number of component $M = 7$. The labels IC01-IC07 refer to the initial components. The pairwise distance matrix is computed as follows: for each pair of starting components, add $1/T$ at each iteration for which it is not in the cluster, and add 0 if it is. The tree was constructed using the program $VTSD$ by S. Sawyer [20] which implementing the neighbor joining method by Saitou and Nei [25] together with bootstrap support number based on 1000 bootstrap replications to show the strength with which the data supports the existence of the branch.

(that is, the clustering results should not be sensitive to the choice of method).

The clustering method is applied on the MCMC distance matrix for two reasons. The first is that, with a large number of initial components, say $M = 20$ for 200 individuals, the clustering result from a particular iteration of MCMC is unstable, in the sense that very different clustering could result within the next few iterations. The second is that it tends to produce small clusters. For example, on Testdata1 [23] setting $M = 200$, if the MCMC is stopped after 10k burn-in and 500k post burn-in iterations, the individuals form 10 clusters with a misclassifying rate of 13.5% (27 out of 200). A hierarchical clustering method based on the pairwise posterior probability strongly suggests two clusters with a 4% (8/200) misclassifying rate.

As mentioned earlier in this section, we also did a simulation study based on data sets generated from the non-admixed probability model. Three sets of allele frequencies for three background populations at 5 loci are generated from a Dirichlet distribution with the parameter λ set to 1. This data set is denoted by Testdata2A. The genotypes of 100 individuals are sampled according to each set of allele frequencies and the analysis is performed. We then repeat this for another data set with $\lambda = 0.1$, denoted Testdata2. As expected, Figure 6 shows that a data set providing strong information on origin population gives cleaner separation of clusters when using the no-admixture model with DP prior.

The Admixture Model with DP Prior. We now apply the model with admixture to the data sets Testdata1, Testdata2A and Testdata2B analyzed by the un-admixed model above. We expect to see that, like other programs analyzing population structure, the model with admixture should uncover the no-admixture population structure but the opposite is often not true [9]. Setting $M > 2$ for Testdata1 [23], the model yields more than two clusters and often recognizes the hybrid of two ancestral population as the third population as shown in Figure 7. However, with $M = 2$, the model

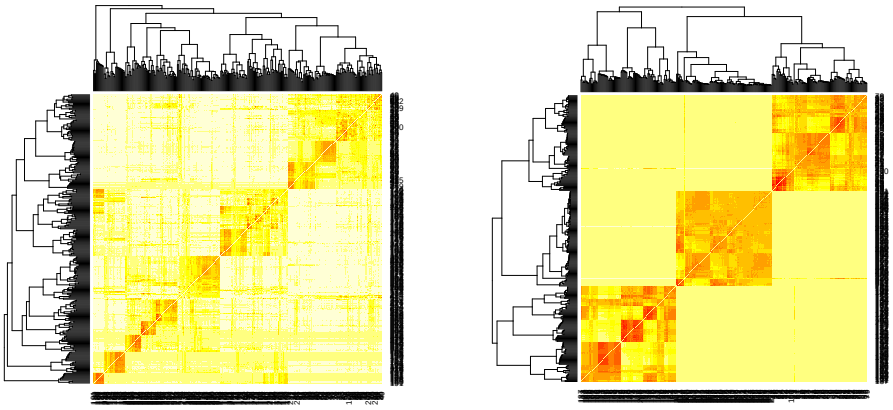


Figure 6: Summary of analyzing two data sets each with 300 individuals from three populations (100 individual for each population). Each data set is generated as following: the allele frequencies for each population is a vector of length 10 generated from $D_{10}(\lambda, \ldots, \lambda)$. Left: Testdata2A, $\lambda = 1$. Right : Testdata2B, $\lambda = 0.1$. The heatmap with the probabilities of pairwise joint classification. Pixiel(i,j) represents the posterior probability, defined in 10.1, of the individual i and j being clustered together. The dendrogram shows hierarchical clustering.

can easily separate the two background populations as shown in Figure 8.

We repeat the analysis to the data sets representing individuals from no admixture populations. We apply a similar analysis to Testdata2A and Testdata2B, setting the initial number of ancestral populations to 30.

Figure 7: Summary of analyzing Testdata1 using the admixture model with DPprior. Testdata1 is the microsatellite data consists of 2 cryptic populations with 100 individuals each at five loci. The initial number of components is set at $M = 20$. Left: the heat-map plot of the MCMC distance matrix. Right: The plot of ancestry coefficients of each individual for each population.

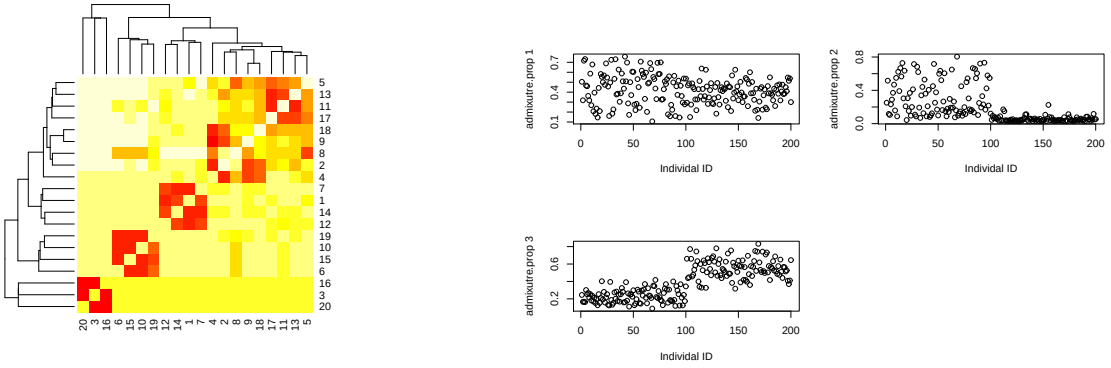


Figure 8: Summary of analyzing Testdata1 using the admixture model with DP prior. Testdata1 is the microsatellite data consists of 2 cryptic populations with 100 individuals each at five loci. The initial number of components is set at $M = 2$. Left: the heat-map plot of the MCMC distance matrix. Right: The plot of ancestry coefficients of each individual for each population.

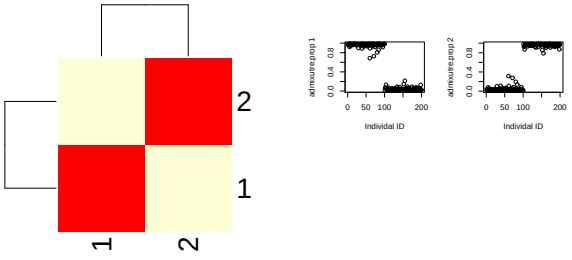


Figure 9: Summary of analyzing Testdata2A using the admixture model with DP prior. Testdata2A is the generated from no admixture model with 3 cryptic populations with 100 individuals each at five loci with the parameter $\lambda = 1$. The initial number of components is set at $M = 30$. Left: the heat-map plot of the MCMC distance matrix with dendrogram display. Right: The plot of ancestry coefficients of each individual for each population.

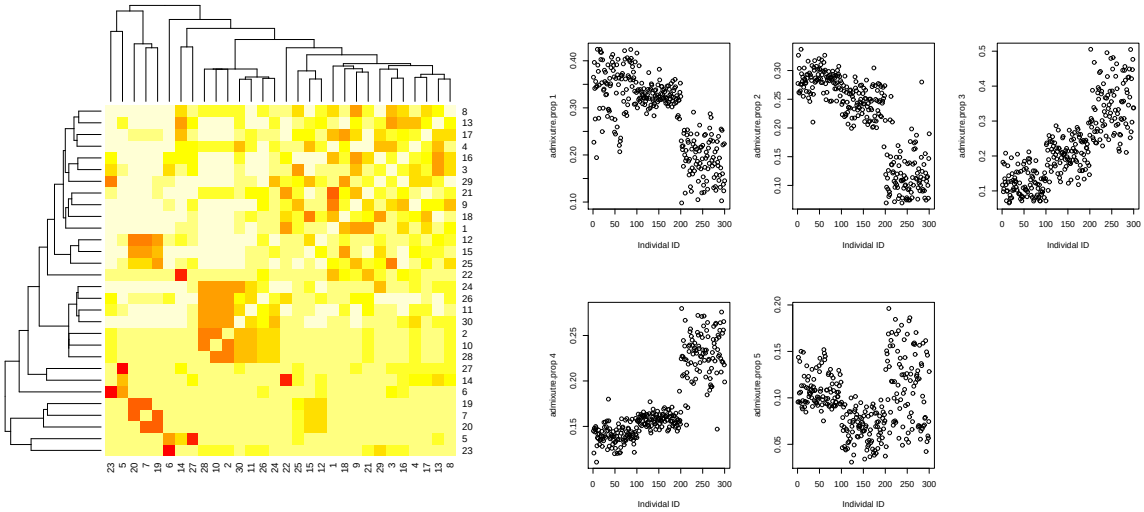
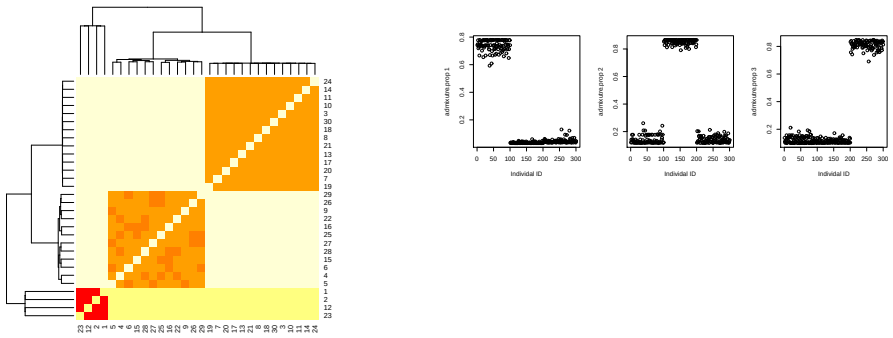


Figure 10: Summary of analyzing Testdata 2B using the admixture model with DPprior. Testdata 2B is the generated from no admixture model with 3 cryptic populations with 100 individuals each at five loci with the parameter $\lambda = 0.1$. The initial number of components is set at $M = 30$. Left: the heat-map plot of the MCMC distance matrix with dendrogram display. Right: The plot of ancestry coefficients of each individual for each population.



References

[1] C Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

[2] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[3] Bradley P. Carlin and Thomas A. Lewis. *Bayes and Empirical Bayes Methods for Data Analysis, 2nd edn.* Chapman & Hall/CRC, 2000.

[4] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[5] B. S. Everitt. *Clustering Analysis*. Edward Arnold, London, 1993.

[6] J. Felsenstein. Phylip (phylogeny inference package) version 3.5c. Technical report, University of Washington, Department of Genetics.

[7] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

[8] T. S. Ferguson. Prior distribution on spaces of probability measures. *The Annals of Statistics*, 2:615–629, 1974.

[9] Oliver François and Eric Durand. Spacially explicit bayeian clustering models in population genetics. *Molecular Ecology Resources*, 27:1257–1268.

[10] P. Galbusera, L. Lens, E. Waiyaki, T. Schenck, and E. Mattysen. Effective population size and gene flow in the globally, critically endangered taita thrush, *Turdus helleri*. *Conserv. Genet.*, 1:45–55, 2000.

[11] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[12] Jeff Gill. *Bayesian methods: a social and behavioral sciences approach*. Chapman-Hill/CRC, 2008.

[13] Dan Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82:159–164.

[14] John P. Huelsenbeck and Peter Andolfatto. Inference of population structure under a dirichlet process model. *Genetics Society of America*, 175:1787–1802, 2007.

[15] M.-H. Chen Ibrahim, J. G and D. Sinha. *Bayesian survival analysis*. Springer Series in Statistics, Springer-Verlag, 2001.

[16] J Kingman. Random discrete distributions. *Proc. Royal Statist. Soc.*, B37:1–22, 1975.

[17] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[18] Mario Medvedovic and Siva Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.

[19] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[20] R. A. Naidu, S. A. Sawyer, and C. M. Deom. Molecular diversity of rna-2 genome segments in pecluviruses causing peanut clump disease in west africa and india. *Archives of Virology*, 148:83–98, 2003.

[21] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[22] J.K. Pritchard. The sturcture homepage. `http://pritch.bsd.uchicago.edu/structure.html`, August 2010.

[23] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[24] G. Rodriguez and O. Papaspiliopoulos. Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, 96:149–162, 2009.

[25] Naruya Saitou and Nei Masatoshi. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, 1987.

[26] S. A. Sawyer and D. Hartl. A sampling therory for local selection. *Journal of Genetics*, 64:21–29, 1985.

[27] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[28] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[29] L Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1728, with discussion 1728–1762., 2006.

[30] Ruibin Xi, X. Huang, N. Lin, and S. Sawyer. Inferring population structure using dirichlet process., 2010.