Biometrics 67, 381–390
June 2011

DOI: 10.1111/j.1541-0420.2010.01484.x

A Spatial Dirichlet Process Mixture Model for Clustering Population Genetics Data

Brian J. Reich* **and Howard D. Bondell**

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
**email:* reich@stat.ncsu.edu

Summary. Identifying homogeneous groups of individuals is an important problem in population genetics. Recently, several methods have been proposed that exploit spatial information to improve clustering algorithms. In this article, we develop a Bayesian clustering algorithm based on the Dirichlet process prior that uses both genetic and spatial information to classify individuals into homogeneous clusters for further study. We study the performance of our method using a simulation study and use our model to cluster wolverines in Western Montana using microsatellite data.

Key words: Bayesian nonparametrics; Dirichlet process prior; Landscape genetics; Microsatellite data; Model-based clustering.

1. Introduction

Recent advances in tools for molecular genetics, along with greater computational power, have led to many new applications with population genetics. For example, landscape genetics examines the interactions between environmental features and microevolutionary processes, such as gene flow, genetic drift, and selection. Two key steps in landscape genetics are to identify the location of genetic clusters and to then correlate these clusters with environmental features, such as mountains, rivers, roads, and human-assisted deforested areas. As a motivating example, consider a study on 88 wolverines (*Gulo gulo*), a highly mobile species, but extremely sensitive to human habitat disturbance (Banci, 1994). This sample was obtained in a region of Montana corresponding to an area of high human development and landscape disturbance (Banci, 1994; Cegelski, Waits, and Anderson, 2003). The individual wolverines were each genotyped at 10 microsatellite loci, and there is evidence for regional subpopulations among these wolverines (Cegelski et al., 2003; Guillot et al., 2005; Corander, Sirén, and Arjas, 2008; Guillot, 2008).

In this article, the focus is on the first step, identifying the genetic subpopulation clusters and their locations. These subpopulations are defined via allele frequencies, and the goal is to identify the spatial locations of these clusters. Clustering of genetic data has received a great deal of recent attention. This is, in part, due to the fact that many statistical procedures in population genetics are based on the assumption of a homogeneous population. For example, testing for selection (Nielsen, 2001), or association (Balding, 2006) assumes homogenous groups. Detection of these clusters can then be a preliminary step to perform these analyses by testing only within the homogenous subpopulations (Guillot, 2009). In addition, identification, and the subsequent investigation, of these clusters may yield information regarding underlying biological processes. The typical assumptions that can be found in Pritchard, Stephens, and Donnelly (2000) are that the individuals have ancestries in a fixed number of clusters. These clusters, or subpopulations, are assumed to be at Hardy–Weinberg equilibrium at each locus with linkage equilibrium between loci. For a further review of some population genetics methods and the associated computer code see Excoffier and Heckel (2006).

One difficulty with many clustering approaches is the determination of the number of clusters. In addition, for the particular goal at hand, even if the number of clusters were assumed to be known, typical approaches to clustering on the allele frequencies do not take the spatial locations into account, unless there is a model for the spatial component. The Bayesian paradigm gives a natural way to tackle this problem by placing a spatial model on the cluster membership probabilities. Francois, Ancelet, and Guillot (2006) use a Potts model as a spatial prior for cluster membership. Guillot et al. (2005) use Voronoi tessellations as the spatial model.

Assuming the clusters are in Hardy–Weinberg equilibrium and there is limited gene flow between clusters, the allele frequencies should vary across clusters for each locus. However, from a statistical perspective, isolating the loci with the largest variation across clusters and down-weighting the contribution of the other loci may improve classification by reducing the dimension of the model. For microarray data, both Wang and Zhu (2008) and Tadesse, Sha, and Vannucci (2005) use a finite mixture of multivariate normals to model the clusters, while also allowing for some of the dimensions to remain constant across mixture components, thus effectively eliminating the corresponding genes. The approach of Tadesse et al. (2005) is a Bayesian approach that yields a probability for each observation to fall into a particular cluster. Kim, Tadesse, and Vannucci (2006) use an infinite mixture of normals to avoid choosing the number of components, and fit the model via a Dirichlet process (DP) mixture.

As opposed to clustering using multivariate normality with microarray data, the genetic data in this case are categorical, as each locus gives rise to a multinomial distribution with probabilities given by the allele frequencies. Hence either the penalization or the Bayesian approach must be modified to a mixture of multinomial distributions for each locus. This approach, however, would still need to be modified to take into account the spatial aspect of the problem. A balance between genetic similarity and spatial proximity needs to drive the clustering.

In this article, we propose a semiparametric Bayesian model for landscape genetics. Following Kim et al. (2006), Pella and Masuda (2006), and Huelsenbeck and Andolfatto (2007), we model the cluster memberships using a DP prior. However, we also incorporate spatial information into the clustering algorithm. Unlike Francois et al. (2006) who discretize the spatial domain, we model the distribution of individuals within each cluster as a continuous process using a separate DP mixture model for the population density of each cluster. The majority of the parameters in our model have conjugate full conditionals that are common distributions, which leads to straightforward Markov chain Monte Carlo (MCMC) coding and tuning. Diagnostics show that for our model this leads to good MCMC convergence.

We apply our model to cluster Montana wolverines. However, the general approach may have other applications. For example, a common goal in spatial epidemiology is to identify clusters of cases to detect an outbreak of an infectious disease. Our model could be used to cluster cases based not only on spatial proximity, but also on other features of the cases such as age, gender, or school enrollment. Similarly, it is common to cluster crimes to identify groups of crimes committed by the same individual. Our model could be used to incorporate other features of the crime, such as the weapon used or the model of entry. In both cases, performing clustering using both spatial and nonspatial data could improve power for detecting homogeneous groups of events.

The remainder of the article is organized as follows. Section 2 details the spatial model. Section 3 discusses hyperpriors and Section 4 gives the computational algorithm. The performance of the model and the effect of hyperpriors is examined by a simulation study in Section 5. The method is applied to cluster Montana wolverines in Section 6. Section 7 concludes.

2. Statistical Model

Let z_{il} contain the genotype for individual i at locus l, $i = 1, \ldots, N$ and $l = 1, \ldots, L$. We assume a diploid organism and unlinked co-dominant neutral markers. The vector $\mathrm{z}_{il} = (z_{il1}, \ldots, z_{ilm_l})$ has length m_l, where m_l is the number of possible alleles at locus l and $z_{ila} \in \{0, 1, 2\}$ is the number of copies of allele a for individual i at locus l. In addition to genetic data, we also record the spatial location $\mathrm{s}_i = (s_{i1}, s_{i2})' \in \mathcal{R}^2$. We assume the observations form clusters, and denote the ith individual's cluster as $g_i \in \mathcal{N}$. Section 2.1 defines the model for z_{il} and s_i given g_i, and Section 2.2 defines the model for g_i and other features shared between groups.

2.1 *Model Within a Group*

Given the group labels g_i, all observations within a group are modeled as independent and identically distributed. The spatial locations and genetic data are modeled jointly. We assume Hardy–Weinberg equilibrium and model z_{il} as

$$\mathrm{z}_{il} \mid g_i = g \sim \mathrm{Multinomial}(2, \boldsymbol{\theta}_{gl}), \qquad (1)$$

where $\boldsymbol{\theta}_{gl}$ is the m_l-vector of allele probabilities at locus l for individuals in cluster g.

We model locations of the individuals from group g using Bayesian nonparametric methods. Let

$$\mathrm{s}_i \mid g_i = g \sim F_g(\mathrm{s}), \qquad (2)$$

where $F_g(\mathrm{s})$ is the spatial distribution, that is, a density on $\mathcal{R}^2$. Rather than specify a parametric distribution for F_g, we model F_g as an unknown quantity to be estimated from the data. We use the potentially infinite mixture model

$$F_g(\mathrm{s}) \overset{d}{=} \sum_{j=1}^{M} p_{gj} \mathrm{G}(\mathrm{s} \mid \Theta_{gj}), \qquad (3)$$

where p_{gj} are the mixture probabilities with $\sum_{j=1}^{M} p_{gj} = 1$ for all g, and $\mathrm{G}(\mathrm{s} \mid \Theta_{gj})$ is a parametric distribution (e.g., bivariate normal) with parameters Θ_{gj} (e.g., the mean and covariance).

To complete the model, we must specify the model for the mixture probabilities p_{gj} and the mixing distribution G. The mixture probabilities are modeled using the stick-breaking construction of Sethuraman (1994). The stick-breaking model is an infinite mixture with $M = \infty$. The mixture probabilities "break the stick" into M pieces so the sum of the pieces is almost surely one, that is, $\sum_{j=1}^{M} p_{gj} = 1$. The first mixture probability is modeled as $p_{g1} = V_{g1}$, where $V_{g1} \sim \mathrm{Beta}(1, b_V)$. Subsequent mixture probabilities are $p_{gj} = V_{gj} \prod_{k=1}^{j-1}(1 - V_{gk})$, where $\prod_{k=1}^{j-1}(1 - V_k) = 1 - \sum_{k=1}^{j-1} p_k$ is the probability not accounted for by the first $j-1$ mixture components, and $V_{gj} \overset{iid}{\sim} \mathrm{Beta}(1, b_V)$ is the proportion of the remaining probability assigned to the jth component.

Assuming the stick-breaking weights, if the mixture distribution is $G(\mathrm{s} \mid \Theta) = \delta(\boldsymbol{\mu}_{gj})$, where $\delta(\boldsymbol{\mu}_{gj})$ is the Dirac distribution with point mass at $\boldsymbol{\mu}_{gj} \in \mathcal{R}^2$ and $\boldsymbol{\mu}_{gj} \overset{iid}{\sim} F_o$, then (3) becomes the DP prior (Ferguson, 1973) with centering distribution F_o. The DP prior is discrete, that is, it has mass only at a countable number of locations $\boldsymbol{\mu}_{gj}$. This can be undesirable in practice, therefore we use the DP mixture of normals model (Antoniak, 1974) for (3),

$$F_g(\mathrm{s}) \overset{d}{=} \sum_{j=1}^{M} p_{gj} \mathrm{N}(\mathrm{s} \mid \boldsymbol{\mu}_{gj}, \Sigma_g), \qquad (4)$$

where $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ is the bivariate normal density with mean $\boldsymbol{\mu}$ and covariance Σ. The mixture means are $\boldsymbol{\mu}_{gj} \overset{iid}{\sim} F_o$; we take F_o to be the uniform distribution on the spatial domain. The covariance matrices Σ_g can be modeled as draws from a conjugate inverse Wishart prior. This gives a nonstationary model as the covariance varies spatially. Simplifying to $\Sigma_g = \sigma_g^2 I_2$ gives an isotropic model. We also consider the stationary model $\sigma_g^2 \equiv \sigma^2$.

The full DP model takes M to be infinite. In practice, it may not be necessary to use an infinite mixture. Note that by construction the mixture probabilities are stochastically decreasing in j, for example, the prior mean of p_{gj} is $\{1/(b_V+1)\}\{b_V/(b_V+1)^{\{j-1\}}$. Therefore, little is lost by truncating the mixture, and we fix M to be a large number that approximates the full DP model, and take $V_M = 1$ to ensure that $\sum_{j=1}^{M} p_{gj} = 1$. Conveniently, the mass in the final term p_{gM} represents the truncation error, so to determine if the approximation is valid we inspect the posterior of p_{gM}. Monitoring truncation probability is discussed further in Section 4. Following this approach, the full model can be approximated to any degree of accuracy, although data sets with many clusters may require a very large M to provide an accurate approximation. Alternatively, to enforce strong spatial grouping within clusters, we could set $M = 1$ to model the spatial distribution as a single-component normal distribution.

An equivalent representation of the mixture model is

$$\begin{aligned} \mathrm{s}_i \mid g_i, h_i &\sim \mathrm{N}(\boldsymbol{\mu}_{g_i h_i}, \Sigma_{g_i}) \\ h_i \mid g_i &\sim \mathrm{Categorical}(p_{g_i 1}, \ldots, p_{g_i M}), \end{aligned} \tag{5}$$

where h_i is the label of the spatial mixture component assigned to observation i and $\mathrm{P}(h_i = h \mid g_i = g) = p_{gh}$. This representation is conducive to MCMC sampling (Section 4).

2.2 *Model Across Groups*

We assume there are as many as K (potentially $K = \infty$) clusters. Let

$$g_i \overset{iid}{\sim} \mathrm{Categorical}(q_1, \ldots, q_K), \tag{6}$$

where $\sum_{j=1}^{K} q_j = 1$. We again model the probabilities using the finite approximation to the stick-breaking model: K is fixed at a large number, $q_1 = U_1$, $q_j = U_j \prod_{k=1}^{j-1}(1 - U_k)$, $U_j \overset{iid}{\sim}$ Beta$(1, b_U)$, and $U_K = 1$. Although g_i can potentially take on K values under this model, the labels $\mathcal{G} = \{g_1, \ldots, g_N\}$ are partitioned into a smaller number of clusters. We define the number of clusters as the number of distinct labels that appear in $\mathcal{G}$ at least twice; labels assigned to only a single member are defined as outliers. We also examine the alternative definition of the number of clusters as simply the number of unique labels in $\mathcal{G}$. The parameter b_U controls the prior number of clusters. Small b_U gives U_j near one, which places most of the probability on the first few q_j, and thus gives a small number of clusters. In contrast, large b_U gives a U_j near zero and the probability is dispersed over many q_j, leading to a large number of clusters. The relationship between b_U and the prior number of clusters is discussed further in Section 3.

We pool information across groups to estimate the allele frequencies $\boldsymbol{\theta}_{gl}$. The allele frequencies are modeled as

$$\boldsymbol{\theta}_{gl} \sim \mathrm{Dirichlet}(\boldsymbol{\phi}_l), \tag{7}$$

where $\boldsymbol{\phi}_l = (\phi_{l1}, \ldots, \phi_{lm_l})$ and $p(\boldsymbol{\theta}_{gl}) \propto \prod_{j=1}^{m_l} \theta_{glj}^{\phi_{lj}-1}$. To aid in prior specification we denote $\boldsymbol{\phi}_l = \rho_l \boldsymbol{\alpha}_l$, where $\boldsymbol{\alpha}_l = (\alpha_{l1}, \ldots, \alpha_{lm_l})$ is a vector of probabilities with $\sum_{j=1}^{m_l} \alpha_{lj} = 1$ and $\rho_l > 0$. The Dirichlet prior is parameterized so that $E(\theta_{glj}) = \alpha_{lj}$ and $V(\theta_{glj}) = \alpha_{lj}(1 - \alpha_{lj})/(\rho_l + 1)$.

In this parameterization, $\rho_l^* = 1/\rho_l$ controls the variability in allele frequencies across clusters for locus l. Loci with $\rho_l^* = 0$ ($\rho_l = \infty$) have the same allele frequencies across clusters and do not help separate clusters. In many applications, the allele frequencies for a small subset of markers vary greatly across the clusters, and allele frequencies for many markers vary only slightly across clusters. To exploit this, we give the variances the two-component mixture prior (George and McCulloch, 1993, 1997)

$$\rho_l^* \sim (1 - \pi_l)\mathrm{Expo}(\lambda_{1l}) + \pi_l \mathrm{Unif}(0, \lambda_{2l}), \tag{8}$$

where $\pi_l \in [0, 1]$, Expo(λ_{1l}) is the exponential density with mean $\lambda_{1l} > 0$, and Unif$(0,\ \lambda_{2l})$ is the uniform density with upper bound $\lambda_{l2} > 0$. The first component of the mixture prior has mass near zero for loci that show little variation across clusters, while the second component gives a heavy tail to accommodate the loci that vary greatly across clusters.

In summary, the full hierarchical model described above can be written as

$$\begin{aligned} \text{Genetic data:}\quad \mathrm{z}_{il} \mid g_i &\sim \mathrm{Multinomial}(2, \boldsymbol{\theta}_{g_i l}) \\ \boldsymbol{\theta}_{gl} &\sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_l, \rho_l) \\ \rho_l^* &\sim (1 - \pi_l)\mathrm{Expo}(\lambda_{1l}) + \pi_l \mathrm{Unif}(0, \lambda_{2l}) \end{aligned} \tag{9}$$

$$\text{Spatial data:}\quad \mathrm{s}_i \mid g_i, h_i \sim \mathrm{N}(\boldsymbol{\mu}_{g_i h_i}, \Sigma_{g_i}) \tag{10}$$

$$h_i \mid g_i \sim \mathrm{Categorical}(p_{g_i 1}, \ldots, p_{g_i M})$$

$$\text{Cluster model:}\quad g_i \sim \mathrm{Categorical}(q_1, \ldots, q_K) \tag{11}$$

In Section 3, we specify hyperpriors for Σ_g, $\boldsymbol{\alpha}_l$, b_U, and b_V and the values of π_l, λ_{1l}, and λ_{2l}.

2.3 *Factors Influencing Clustering*

For the given model, clusters are formed based on both genetic and spatial similarity. To see this in a simple case, assume that there are only two individuals in the sample and that $M = K = \infty$. Integrating over the allele frequencies $\boldsymbol{\theta}$, the spatial knots $\boldsymbol{\mu}$ (assuming for algebraic convenience that F_0 is flat on the spatial domain and surrounding areas so that $\int N(\mathrm{s} \mid \boldsymbol{\mu}, \Sigma) d\boldsymbol{\mu} = 1$ for all s), and the stick-breaking parameters U and V, the posterior cluster probability is

$$P(g_1 = g_2) \propto D_0(\mathrm{s}_1, \mathrm{s}_2)^{-1} \prod_{l=1}^{L} D_l(z_{1l}, z_{2l})^{-1}, \tag{12}$$

where

$$D_0(\mathrm{s}_1, \mathrm{s}_2) = \left[b_V + \frac{|\,\Sigma\,|^{-1/2}}{2\sqrt{\pi}} \exp\{-0.25(\mathrm{s}_1 - \mathrm{s}_2)'\Sigma^{-1}(\mathrm{s}_1 - \mathrm{s}_2)\} \right]^{-1} \tag{13}$$

$$D_l(z_{1l}, z_{2l}) = \prod_{j=1}^{m_l} \frac{z_{1lj}! z_{2lj}!}{\Gamma(z_{1lj} + z_{2lj} + \alpha_{lj}/\rho_l^*)} \text{ for } l > 0. \tag{14}$$

$P(g_1 = g_2)$ is the product of terms representing spatial distance (D_0) and genetic distance at each locus (D_l), respectively. Clearly $D_0(\mathrm{s}_1, \mathrm{s}_2)$ increases with spatial distance $(\mathrm{s}_1 - \mathrm{s}_2)'\Sigma^{-1}(\mathrm{s}_1 - \mathrm{s}_2)$ and the covariance Σ determines the rate of increase. To illustrate how D_l measures genetic distance, assume there are two equally likely alleles at marker l ($m_l = 2$, $\alpha_{l1} = \alpha_{l2} = 1/2$) and the first subject has two copies of the first allele ($z_{1l} = (2,\ 0)$). The ratio of D_l assuming individuals' alleles match completely ($z_{2l} = (2,\ 0)$) and are completely mismatched ($z_{2l} = (0,\ 2)$) at marker l is

$$\frac{\Gamma\big(2 + 0.5/\rho_l^*\big)^2}{\Gamma\big(4 + 0.5/\rho_l^*\big)\Gamma\big(0.5/\rho_l^*\big)}. \tag{15}$$

This ratio is always less than one and is decreasing in ρ_l^*. ρ_l^* determines the sensitivity of the clustering probability to marker l; the ratio (15) is 0.93, 0.53, and 0.09, respectively, for ρ_l^* equal 0.01, 0.1, and 1.

3. Prior Specification

Commonly in Bayesian analysis results can be sensitive to the parameters that determine the model probabilities, for example the number of clusters (b_U), and less sensitive to prior for parameters that determine the fit of a given model (e.g., b_V). The prior b_U determines the prior distribution for the number of clusters. Although there is no closed-form expression for the distribution of the number clusters as a function of b_U, we examine this relationship by drawing samples of the labels $\mathcal{G}$ from the prior for various b_U. We assume $N = 88$ as in the wolverine data and truncate the mixture distribution at $K = 25$ terms, and we draw samples from $U_j \mid b_U$ and then $g_i \mid U_1, \ldots, U_K$. Figure 1a plots the distribution of the number of clusters as a function of b_U. The prior mean number of clusters is increasing in b_U. Different priors for b_U can induce a wide variety of priors for the number of clusters. Figure 1b integrates over b_U using $b_U \sim$ Gamma(1,1) and $b_U \sim$ Gamma(1,1/4) priors. These hyperpriors represent very different prior beliefs regarding the number of clusters. The prior number of clusters assuming $b_U \sim$ Gamma(1,1) has mode 1, median 3, and 95th quantile 9 and reflects the prior belief that the number of clusters is likely to be small. In contrast, the prior number of clusters assuming $b_U \sim$ Gamma(1,1/4) is fairly flat for 1 to 10 clusters.

Preliminary analysis suggests that an uninformative prior for the prior mean of the allele frequencies, $\boldsymbol{\alpha}_l$, gives a large posterior variance, especially for data with a small number of clusters. Therefore, to improve MCMC convergence we fix the prior mean of the allele frequencies at the sample mean, that is, $\alpha_{lj} = \bar{z}_{lj} = \sum_{i=1}^N z_{ilj}/(2N)$. The strength of the prior for the allele frequencies at locus l is determined by ρ_l^*; if $\rho_l^* \approx 0$ locus l the allele frequencies at locus l are strongly smoothed toward $\boldsymbol{\alpha}_l$. We assume $\pi_l = 0.5$ to give equal weight to each component in the mixture prior (8). The prior scale parameters λ_{1l} and λ_{2l} must be inversely proportional to the number of possible alleles at marker l, m_l, for the prior to be comparable across markers. To calibrate the prior according to an overall measure of variability in allele frequencies across clusters, we compute the expected value (with respect to $\boldsymbol{\theta}_{gl}$'s prior) of the usual chi-squared test of the hypothesis $\boldsymbol{\theta}_{gl} = \boldsymbol{\alpha}_l$,

$$\mathrm{E}(\chi^2) = \mathrm{E}\left\{\sum_{j=1}^{m_l} \frac{(\theta_{gjl} - \alpha_{jl})^2}{\alpha_{jl}}\right\} = (m_l - 1)\frac{\rho_l^*}{\rho_l^* + 1}. \tag{16}$$

Therefore, under the null hypothesis that allele frequencies are constant across cluster for all markers, we expect more variability for markers with many possible alleles. To account for this, we assume $\lambda_{1l} = (m_l - 1)/100$ and $\lambda_{2l} = 10(m_l - 1)$. This prior provides considerable mass near zero and has a heavy tail to prevent over-smoothing loci that vary considerable across clusters.

We pick uninformative priors for the parameters that control the spatial density. We transform the spatial coordinates to $[0,\ 1]^2$ and assume the isotropic spatial model with the same spread for each cluster, $\Sigma_g = \sigma^2 I_2$. The uninformative priors are $\sigma^{-2} \sim$ Gamma(0.1,0.1) and $b_V \sim$ Gamma(0.1,0.1). Section 6 conducts a sensitivity analysis to determine the effect of these priors and the assumption of a common spread parameters on the posterior number of clusters.

Summarizing, the priors and hyperparameters used and their motivation are:

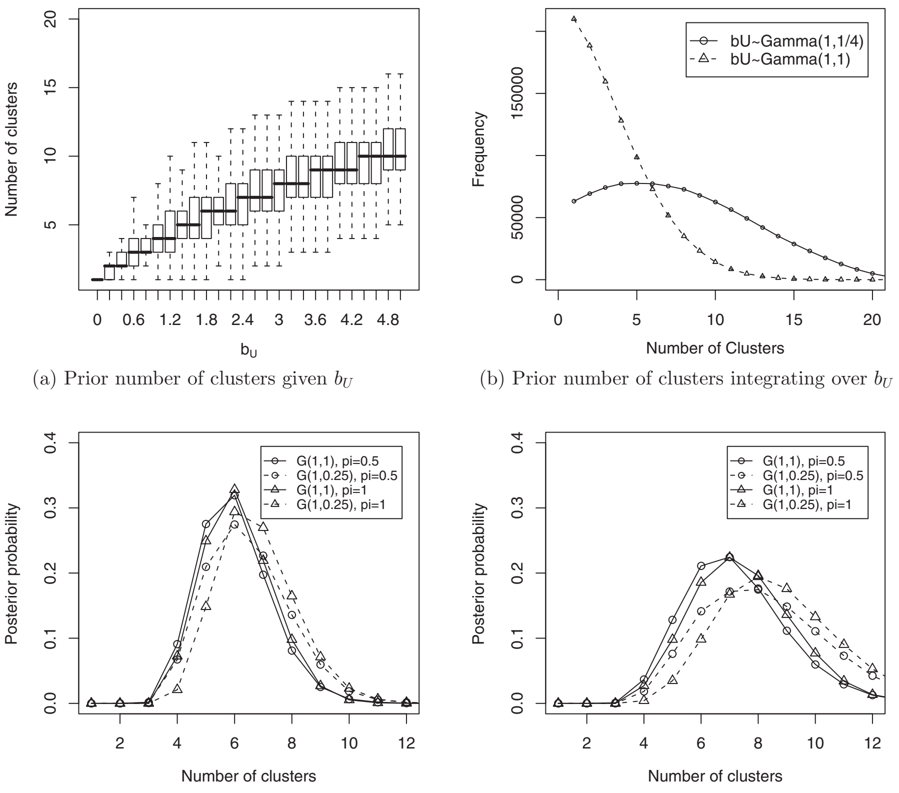
- $M = K$ are fixed at 25 to approximate the full DP model
- $b_U \sim$ Gamma(1,1) or Gamma(1,1/4) to give the prior number of clusters in Figure 1
- $\boldsymbol{\alpha}_l$ fixed at the sample frequencies to improve MCMC convergence
- $\rho_l^* \sim 0.5 * \mathrm{Expo}(\frac{m_l - 1}{100}) + 0.5 * \mathrm{Unif}(0,\ 10 * [m_l - 1])$ to represent the prior belief that some loci are vary greatly across clusters, while others are similar across clusters
- Uninformative priors $\sigma^{-2} \sim$ Gamma(0.1,0.1) and $b_V \sim$ Gamma(0.1,0.1)

4. Computational Details

Although it is straightforward to fit the infinite mixture model using the sampling method of Papaspiliopoulos and Roberts (2008), in practice we truncate the mixture distribution by fixing K, $M < \infty$. We assume $K = M = 25$ and set $U_K = V_{gM} = 1$. The truncation error can easily be accessed by inspecting the distributions of q_K and p_{gK}, the masses of the final components of the mixtures, which are readily available from the MCMC samples. For the analysis in Section 6 the posterior medians of the final mixture probabilities are less than 0.01 for all models. Truncating the infinite mixtures with a finite K allows the model to be fit using `WinBUGS`. `WinBUGS` can be freely downloaded from `http://www.mrc-bsu.cam.ac.uk/bugs/`. Although `WinBUGS` can be used, we perform MCMC sampling using `R` (`http://www.r-project.org/`). Gibbs sampling is used for all parameters except ρ_l^*, which is updated with Metropolis sampling using a Beta($100r_l$, $100[1 - r_l]$) candidate distribution, where r_l is the previous draw for ρ_l^*. A complete description of the algorithm is given in the Appendix. `R` code is available from the first author upon request.

For the simulation study in Section 5, we generate 20,000 samples and discard the first 5000 samples as burn-in; for the analysis of the wolverines data in Section 6, we generate 100,000 samples and discard the first 20,000 samples as burn-in. This sampling takes around 1 hour on an ordinary PC for the wolverines data with $N = 88$, $L = 10$, and $m_l \in \{3,\ 4,\ 5,\ 6\}$. Convergence is monitored using trace plots and autocorrelation plots for several representative parameters.

Monitoring convergence is challenging because of the label-switching problem, that is, that labels for the clusters are arbitrary and change from iteration to iteration. To monitor convergence, we use parameters that are not directly functions of the group labels. For example, for Section 6's wolverines data, we ran four independent chains (we present only the first in Section 6) with $b_U \sim$ Gamma(1,1) and π= 0.5. The mode



(a) Prior number of clusters given b_U

(b) Prior number of clusters integrating over b_U

(c) Posterior number of clusters

(d) Posterior number of clusters, alternate cluster definition

Figure 1. Panels (a) and (b) plot the prior number of clusters (defined as a group of at least two) conditioned on b_U and assuming $b_U \sim$ Gamma(1,0.25) or $b_U \sim$ Gamma(1,1). Panels (c) and (d) plot the posterior number of clusters for various models for the wolverines data, defining a cluster as a group of at least two (panel (c)) or at least one (panel (d)).

number of clusters was six for each chain, the posterior probability of six clusters ranged from 0.299 to 0.305, and the posterior mean number of clusters varied from 5.95 to 6.17. Similarly, for each pair of wolverines we computed the probability they were in the same cluster for each chain, and computed that standard deviation of these pairwise probabilities across chains. The median standard deviation was 0.004.

5. Simulation Study

In this section, we conduct a brief simulation study. In Section 5.1, we explore the benefits of accounting for less-informative markers and the effect of the prior for the number of clusters. In Section 5.2, we study the benefits of including spatial information to the DP model.

5.1 *Analysis of Data with Many Nondifferentiated Loci*

We assume there are $L = 20$ loci, each with $m_l = 2$ possible alleles, and $N = 100$ individuals drawn from as many as four clusters. The frequencies of the first allele for the four clusters are as follows. The clusters have frequencies 0.2, 0.8, 0.2, and 0.8, respectively, for the first locus and frequencies 0.2, 0.2, 0.8, and 0.8, respectively, for the second, third, and fourth loci. Therefore the first locus separates clusters 1 and 3 and the second through fourth loci separates the third and fourth clusters. The remaining 16 loci have allele frequency $0.9 * l/L$ for all four clusters. We generate data from four designs.

1. All individuals belong to Cluster 1
2. Individuals are drawn from Cluster 1 or 2 with equal probability
3. Individuals are drawn from Cluster 1 with probability 1/3 or Cluster 2 otherwise
4. Individuals are drawn from Cluster 1, 2, 3, or 4 with equal probability

Given the group memberships g_i, the genetic data are drawn from multinomial distributions with allele frequencies specified above, and spatial locations are generated as $\text{s}_i \sim \text{N}(\bar{\text{s}}_{g_i}, 0.25^2 I_2)$, where $\bar{\text{s}}_{g_i}$ is the cluster center. The cluster centers are $\bar{\text{s}}_1 = (-0.5,-0.5)$, $\bar{\text{s}}_2 = (0.5,\text{-}0.5)$, $\bar{\text{s}}_3 = (-0.5, 0.5)$, and $\bar{\text{s}}_4 = (0.5,0.5)$. Design 1 is the null design with a single cluster, Designs 2 and 3 are nonnull with two clusters, and Design 4 has four clusters. The informative loci are sparse, with only Locus 1 being informative for Designs 2 and 3 and only Loci 1 though 4 being informative for Design 4.

For each design, we simulate $S = 25$ data sets. For each data set we fit three special cases of Section 2's spatial clustering model by varying the prior number of cluster (b_U) and the prior for the allele frequencies (π):

1. $b_U \sim$ Gamma(1,1), $\pi = 1$
2. $b_U \sim$ Gamma(1,0.25), $\pi = 0.5$
3. $b_U \sim$ Gamma(1,1), $\pi = 0.5$.

Table 1

Analysis of Section 5.1*'s simulated data. Mean* (*SD*) *over the simulated data sets of the posterior mode number of clusters* (*"Mode # clust"*), *the posterior probability of the correct number of clusters* (*"P correct clust"*), *and the true* (*"TPC"*) *and false* (*"FPC"*) *pairwise cluster probabilities. The true and false pairwise cluster probabilities, TPC and FPC, are defined in* (17).

Design	Model	Mode # clust	P correct clust	TPC	FPC
1	1	1.08 (0.40)	0.79 (0.21)	0.97 (0.06)	–
	2	1.08 (0.40)	0.72 (0.18)	0.94 (0.06)	–
	3	1.00 (0.00)	0.87 (0.10)	0.98 (0.02)	–
2	1	3.16 (1.65)	0.27 (0.21)	0.75 (0.14)	0.05 (0.02)
	2	2.08 (0.40)	0.52 (0.16)	0.87 (0.05)	0.09 (0.13)
	3	1.96 (0.35)	0.67 (0.17)	0.91 (0.03)	0.12 (0.21)
3	1	3.08 (1.15)	0.30 (0.25)	0.77 (0.14)	0.05 (0.03)
	2	2.20 (0.41)	0.53 (0.18)	0.88 (0.05)	0.06 (0.04)
	3	1.96 (0.45)	0.61 (0.22)	0.92 (0.05)	0.18 (0.27)
4	1	6.00 (1.29)	0.11 (0.10)	0.73 (0.09)	0.03 (0.01)
	2	4.16 (1.28)	0.27 (0.17)	0.84 (0.05)	0.10 (0.12)
	3	4.00 (1.04)	0.42 (0.23)	0.87 (0.06)	0.07 (0.10)

Table 2

Analysis of Section 5.2*'s simulated data. Mean* (*SD*) *over the simulated data sets of the posterior mode number of clusters* (*"Mode # clust"*), *the posterior probability of the correct number of clusters* (*"P correct clust"*), *and the true* (*"TPC"*) *and false* (*"FPC"*) *pairwise cluster probabilities. The true and false pairwise cluster probabilities, TPC and FPC, are defined in* (17).

Spatial	F_{ST}	Mode # clust	P correct clust	TPC	FPC
Yes	0.01	5.40 (0.55)	0.35 (0.07)	0.59 (0.01)	0.10 (0.01)
	0.02	6.40 (0.89)	0.10 (0.09)	0.67 (0.02)	0.06 (0.01)
	0.03	5.60 (0.55)	0.28 (0.16)	0.81 (0.01)	0.04 (0.00)
	0.04	6.20 (1.10)	0.40 (0.31)	0.88 (0.02)	0.03 (0.00)
	0.05	5.00 (0.00)	0.75 (0.17)	0.94 (0.01)	0.01 (0.00)
	0.06	5.00 (0.00)	0.82 (0.12)	0.96 (0.01)	0.01 (0.00)
	0.07	5.00 (0.00)	0.95 (0.03)	0.98 (0.00)	0.00 (0.00)
	0.08	5.00 (0.00)	0.97 (0.02)	0.99 (0.00)	0.00 (0.00)
	0.09	5.00 (0.00)	0.99 (0.01)	0.99 (0.00)	0.00 (0.00)
No	0.01	12.60 (2.51)	0.02 (0.02)	0.18 (0.02)	0.17 (0.02)
	0.02	15.60 (3.91)	0.02 (0.02)	0.17 (0.02)	0.16 (0.03)
	0.03	17.60 (4.34)	0.00 (0.00)	0.24 (0.06)	0.09 (0.01)
	0.04	9.20 (2.05)	0.03 (0.07)	0.50 (0.05)	0.09 (0.00)
	0.05	6.00 (1.22)	0.29 (0.25)	0.74 (0.02)	0.06 (0.00)
	0.06	5.80 (0.84)	0.39 (0.31)	0.85 (0.02)	0.04 (0.00)
	0.07	5.00 (0.00)	0.81 (0.13)	0.93 (0.02)	0.02 (0.00)
	0.08	5.00 (0.00)	0.91 (0.07)	0.96 (0.01)	0.01 (0.00)
	0.09	5.00 (0.00)	0.91 (0.12)	0.97 (0.00)	0.01 (0.00)

For each model and each data set, we compute the posterior mode number of clusters ("Mode # clust"), the posterior probability of the correct number of clusters ("P correct clust"), and the true ("TPC") and false ("FPC") pairwise cluster probabilities, defined as

$$TPC = \frac{\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N} I(g_i^* = g_j^*) P(g_i = g_j \mid \text{data})}{\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N} I(g_i^* = g_j^*)}$$

$$FPC = \frac{\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N} I(g_i^* \neq g_j^*) P(g_i = g_j \mid \text{data})}{\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N} I(g_i^* \neq g_j^*)}, \tag{17}$$

where g_i^* is the true cluster for individual i. TPC (FPC) is the average pairwise cluster probability over all pairs of individuals that are (are not) from the same group. Table 1 reports the mean (SD) over the $S = 25$ simulated data sets of these summary measures.

For the first design, all three models successfully identify that the data are generated from a single cluster. For Designs 2 through 4 with multiple clusters, the model that does not account for nondifferentiated loci (Model 1) overestimates the number of clusters. This may be due to spurious clusters formed based on loci that are truly not different across clusters but are included in the clustering model. This leads to poor performance in the pairwise cluster probabilities. The models that account for nondifferentiated loci are able to identify important loci. For example, for Design 4 only the first four loci are differentiated, and for Model 4 the average (standard error) of ρ_l^* over the 25 data sets for the first five loci are 0.45 (0.02), 0.53 (0.02), 0.52 (0.02), 0.54 (0.02), and 0.00 (0.00), respectively (loci 6–20 are similar to loci 5). For Models 2 and 3, we used two very different priors for the number of clusters (Figure 1), and the prior for the number of clusters has only a small effect on the posterior number of clusters. As expected based on Figure 1, the Gamma(1,1) prior for b_U (Model 3) favors slightly fewer clusters than the Gamma(1,0.25) prior (Model 2).

5.2 *Analysis of the Five-island Data*

Latch et al. (2006) simulated several data sets, which are available at `www2.imm.dtu.dk/ gigu/Bioinformatics-HMRF/`. These data sets have five clusters, each with 200 members. For each individual, there are 10 loci, each with 10 possible alleles. The data are generated with differentiation across clusters controlled by $F_{ST} \in \{0.01, 0.02, \ldots, 0.09\}$. For each value of F_{ST} there are $S = 5$ simulated data sets. For each data set, we fit the full model with $\pi = 0.5$ and $b_U \sim$ Gamma(1,2), which corresponds to a (1,10) 90% prior interval for the number of clusters. We also fit the nonspatial DP model, similar to Huelsenbeck and Andolfatto (2007), by eliminating the model for the spatial location in (10). The results are given in Table 2.

For all data sets, the posterior mode number of clusters is at least five, so the spatial model is able to detect some clustering for all values of F_{ST}, although the true pairwise cluster probability is less than 0.8 for $F_{ST} = 0.01$ and $F_{ST} = 0.02$. The number of clusters is consistently identified as 5 for F_{ST} higher than 0.04. Due to the large sample size, for small F_{ST} there are often some extra clusters with only a few members. This could be remedied with a more conservative definition of a cluster. For example, rather than considering a group with at least two members as a cluster, we could require a certain percentage of the population. Alternatively, we could consider a more general model for the cluster labels. In the stick-breaking representation of the DP we have $U_j \sim$ Beta(1, b_U). A more general model is $U_j \sim$ Beta(a_j, b_j) which could be tuned to have less mass in the tail of the prior for the number of clusters. For example, $a_j = 1 - a_U$ and $b_j = b_U + ja_U$ gives the Pitman–Yor process (Pitman and Yor, 1997).

The spatial DP model is far more effective than the nonspatial DP model. The nonspatial DP model often identifies more than 10 clusters and has low true pairwise cluster probabilities. Guillot (2009) analyzes these data using the program `Tess` (Chen et al., 2007) and finds that, similar to the nonspatial DP, `Tess` often reports 10 clusters and identifies spurious clusters when no differentiation is present. Although Durand, Chen, and Francois (2009) show that even in cases where the number of clusters is over-estimated, the 5 main clusters are often clear for F_{ST} as low as 0.03 (Durand et al., 2009, Figure 1c), but there are also additional clusters with smaller membership probabilities. When the parameters are tuned to match a known pattern, `Tess` does identify the correct number of clusters with $F_{ST} = 0.05$ or higher (Guillot, 2009). However, in practice, such tuning is usually not possible.

6. Analysis of the Wolverine Data

In this section, we apply our spatial clustering algorithm to wolverine data, originally analyzed by Cegelski et al. (2003). There are $N = 88$ wolverines and $L = 10$ loci. For each locus, there are between $m_l = 3$ and $m_l = 6$ possible alleles. Previous analyses have found between 3 (Cegelski et al., 2003; Corander et al., 2008) and 6 (Guillot et al., 2005; Guillot, 2008) clusters.

We fit four models by varying the prior for the number of clusters, $b_U \sim$ Gamma(1,1) or Gamma(1,0.25), and the prior for the allele frequencies, $\pi = 0.5$ or $\pi = 1$. Figure 1c plots the posterior of the number of clusters for each model. The posterior number of clusters is robust to the prior for b_U; comparing models with $\pi = 0.5$ (circles), the $b_U \sim$ Gamma(1,1) prior (solid line) leads to slightly more mass on small numbers of clusters than the $b_U \sim$ Gamma(1,0.25) prior (dashed line). However, these differences are small compared to the prior difference (Figure 1b) and both priors lead to a posterior mode of six clusters, as in Guillot et al. (2005) and Guillot (2008).

Figure 2 shows the posterior of the ρ_l^*, which control the variation of allele frequencies across clusters, for the two models with $b_u \sim$ Gamma(1,1). Markers 1, 3, 4, and 5 emerge as highly informative. The model with $\pi = 0.5$ pushes the mass of the remaining loci to zero and effectively removes these loci from the clustering model. Also plotted in Figure 2 is the posterior of ρ_l^* for the model that ignores the number of possible alleles at each marker in the prior for ρ_l^* and simply takes $\lambda_1 = 0.001$ for all l. The effect is largest for the first marker, which has only three possible alleles compared to the other markers that have either five or six.

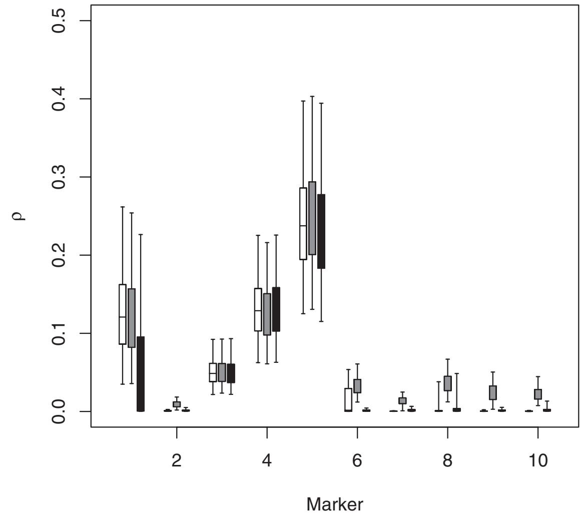


Figure 2. Posterior distribution of ρ_l^* for various models for the wolverine data. All models assume $b_U \sim$ Gamma(1,1); the left (white) boxplots correspond to the model with variable selection ($\pi = 0.5$), the middle (gray) boxplots correspond to the model without variable selection ($\pi = 1$), and the right (black) right boxplots have variable selection because we do not adjust for the number of alleles and take $\lambda = 0.001$ for all markers. The number of possible alleles (m_l) is three for marker 1; six for markers 2, 6, and 7; and five for all the other markers.

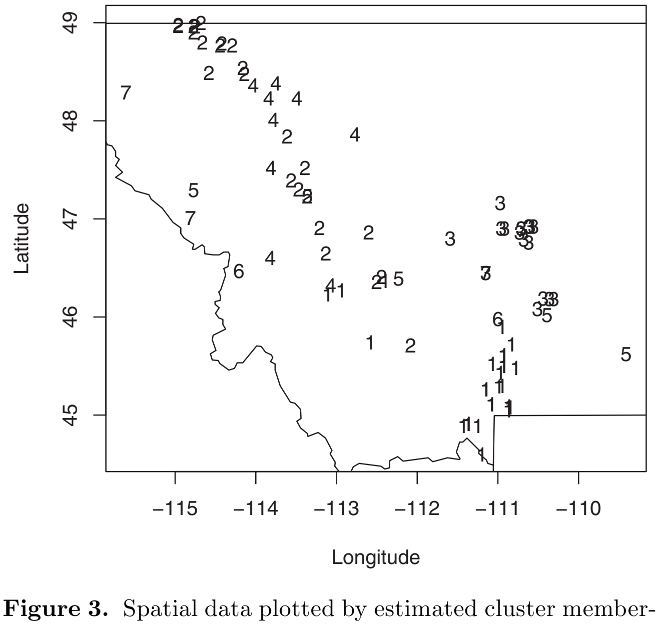
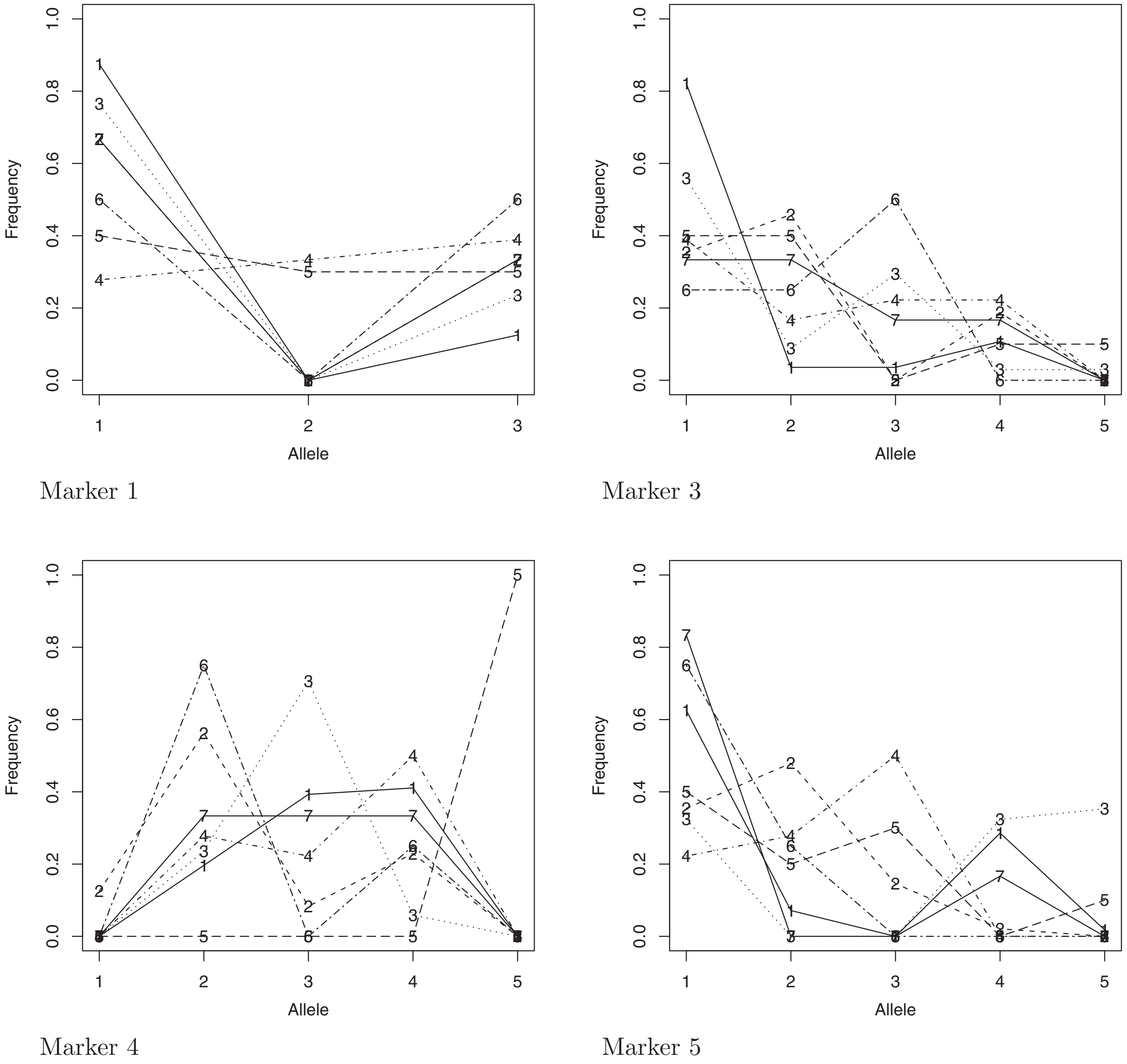


Figure 3. Spatial data plotted by estimated cluster membership.

Figures 3 and 4 summarize the spatial and genetic information for the clusters from the model with $b_U \sim$ Gamma(1,1) and $\pi = 0.5$ assuming there are six clusters. Summarizing the cluster membership in a Bayesian clustering model is notoriously difficult due to the label-switching problem; see Jasra, Holmes, and Stephens (2005), Guillot (2008), and Dawson and Belkhir (2009). Define $\mathcal{G}^s = \{g_1^s, \ldots, g_N^s\}$ as the labels for iteration s of the MCMC algorithm. We use only samples for which $\mathcal{G}^s$ has six clusters. Following Dahl (2006) and Dahl and Newton (2007), we estimate group membership by computing all pairwise posterior probabilities $d_{ij} = \sum_{s=1}^{M} I(g_i^s = g_j^s)/M$ for each pair of individuals, and then estimate the cluster membership as

$$\hat{\mathcal{G}} = \arg\min_{s} \sum_{i=1}^{N} \sum_{j=1}^{N} \{I\big(g_i^s = g_j^s\big) - d_{ij}\}^2. \tag{18}$$



Marker 1

Marker 3

Marker 4

Marker 5

Figure 4. Genetic data plotted by estimated cluster. In all plots, the number of the points indicates the estimated cluster membership. The vertical axis gives the sample frequency of each allele for each group.

Dahl (2006) shows this minimizes the posterior expected loss of Binder (1978) with equal costs of clustering mistakes.

The locations of the members of each cluster in Figure 3 show a strong spatial pattern. However, there is some overlap between groups; for example, between Clusters 2 and 4 in the north/central region. There are also clear genetic differences between the clusters in Figure 4. For example, only members of Cluster 2 have copies of the first allele for Marker 4 and no member of Cluster 3 has the second or third alleles for Marker 5.

Unlike the other clusters, members of Cluster 5 are spatially dispersed. This cluster is defined largely by Marker 4; all members of this cluster have two copies of the fifth allele for Marker 4, and no other wolverines have this allele. In addition to the six clusters, there are also three outliers, labeled Cluster 7 in Figures 3 and 4. Although these individual are in a spatial area predominated by Cluster 2, they are genetically dissimilar to this cluster; for example, unlike any member of Cluster 2 one outlier has a copy of the third allele for Marker 3.

To test for prior sensitivity, we also refit the model with $b_U \sim \text{Gamma}(1,1)$ and $\pi = 0.5$ using different priors for σ^2 and b_V. The analysis above assumed $\sigma^{-2} \sim \text{Gamma}(0.1,0.1)$ and $b_V \sim \text{Gamma}(0.1,0.1)$. We refit, altering one prior each analysis, with $\sigma^{-2} \sim \text{Gamma}(0.01,0.01)$ or $\sigma^{-2} \sim \text{Gamma}(1,1)$ and $b_V \sim \text{Gamma}(0.01,0.01)$ or $b_V \sim \text{Gamma}(1,1)$. These priors had little effect on the posterior mean number of clusters that ranged from 5.1 with $\sigma^{-2} \sim$ Gamma(0.1,0.1) and $b_V \sim$ Gamma(1,1) to 5.8 with $\sigma^{-2} \sim$ Gamma(0.01,0.01) and $b_V \sim$ Gamma(0.1,0.1). Finally, we fit the model with a different spread parameter for each group, $\sigma_g^{-2} \overset{iid}{\sim}$ Gamma(0.1,0.1), and the mean number of clusters was 5.8.

We also study the sensitivity to the definition of a cluster. In the above analysis, we define a cluster as a group with at least two members. We also consider defining a cluster as any group with at least one member. Figure 1d plots the posterior number of clusters under this alternative definition. As expected, the number of clusters is higher under the second definition. For example, if $b_U \sim$ Gamma(1,1) and $\pi = 0.5$, the posterior mean number of clusters is 6.01 and 7.23, for the first and second definitions, respectively. This comparison may be more appropriate after calibrating the prior for the new cluster definition. Assuming $b_U \sim$ Gamma(1,1) the prior median (95% quantile) number of clusters is 3 (9) and 4 (11) under the first and second cluster definitions, respectively. If we assume $b_U \sim$ Gamma(1,1.5), under the second definition the prior median (95% quantile) number of clusters is 3 (9) and the posterior mean number of clusters is 6.91. So the definition of a cluster has a moderate effect on the results.

7. Conclusions

In this article, we develop a Bayesian spatial model based on the DP prior to cluster individuals using both spatial and genetic information. By jointly modeling genotypes and geographic location, the majority of the parameters in our model have conjugate full conditionals that are common distributions, which leads to straightforward MCMC coding and tuning.

Our model uses only latitude and longitude to summarize the individual's spatial location. It would be straightforward to include additional geographic information into the clustering model. For example, distance to a river or land-use classification could be added as additional responses. The variable selection approach could be used to identify a subset of geographic predictors that aid in clustering.

Acknowledgements

The authors thank the editor, associate editor, and two referees for helpful comments and suggestions that greatly improved the manuscript. We also thank Drs Jung-Ying Tzeng and Jeffrey Thorne of NCSU for discussion of population genetics and the National Institutes of Health (1 R01 MH084022-01A1) and National Science Foundation (DMS-0705968 and CMG-0934595) for supporting this work.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2,** 1152–1174.

Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Review Genetics* **7,** 781–791.

Banci, V. (1994). Wolverine. In *The Scientific Basis for Conserving Forest Carnivores, American Marten, Fisher, Lynx, and Wolverine in the Western United States,* L. F. Ruggiero, K. B. Aubry, S. W. Buskirk, L. J. Lyon, and W. J. Zielinski (eds), General Technical Report RM-254: 1–184, 99-12. Fort Collins, CO: United States Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65,** 31–38.

Cegelski, C., Waits, L. P., and Anderson, N. J. (2003). Assessing population structure and gene flow in Montana wolverines (*Gulo gulo*) using assignment-based approaches. *Molecular Ecology* **12,** 2907–2918.

Chen, C., Durand, E., Forbes, F., and François, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. *Molecular Ecology Notes* **7,** 747–756.

Corander, J., Sirén, J., and Arjas, E. (2008). Bayesian spatial modeling of genetic population structure. *Computational Statistics* **23,** 111–129.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics*, K. A. Do, P. Muller, and M. Vannucci (eds), 201–218. Cambridge, U.K.: Cambridge University Press.

Dahl, D. B. and Newton, M. A. (2007). Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association* **102,** 517–526.

Dawson, K. J. and Belkhir, K. (2009). An agglomerative hierarchical approach to visualisation in Bayesian clustering problems. *Heredity* **103,** 32–45.

Durand, E., Chen, C., and Francois, O. (2009). Comment on "On the inference of spatial structure from population genetics data." *Bioinformatics* **25,** 1802–1804.

Excoffier, L. and Heckel, G. (2006). Computer programs for population genetics data analysis: A survival guide. *Nature Review Genetics* **7,** 745–758.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1,** 209–230.

Francois, O., Ancelet, S., and Guillot, G. (2006). Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174,** 805–816.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88,** 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7,** 339–373.

Guillot, G. (2008). Inference of structure in subdivided populations at low levels of genetic differentiation. The correlated allele frequencies model revisited. *Bioinformatics* **24,** 2222–2228.

Guillot, G. (2009). On the inference of spatial structure from population genetics data. *Bioinformatics* **25,** 1796–1801.

Guillot, G., Estoup, A., Mortier, F., and Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170,** 1261–1280.

Huelsenbeck, J. P. and Andolfatto, P. (2007). Inference of population structure under a Dirichlet process model. *Genetics* **175,** 1787–1802.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* **20,** 50–67.

Kim, S., Tadesse, M. G., Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93,** 877–893.

Latch, E. K., Dharmarajan, G., Glaubitz, J. C., Rhodes, O. E., Jr. (2006). Relative performance of Bayesian clustering softwares for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics* **7,** 295–302.

Nielsen, R. (2001). Statistical tests of neutrality at the age of genomics. *Heredity* **86,** 641–647.

Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika* **95,** 169–186.

Pella, J. and Masuda, M. (2006). The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fishery and Aquatic Sciences* **63,** 576–596.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25,** 855–900.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4,** 639–650.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100,** 602–617.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64,** 440–448.

Received July 2009. *Revised June* 2010.
Accepted June 2010.

Appendix

We initialize the MCMC algorithm by drawing a sample from the prior. Sampling then proceeds using a combination of Gibbs and Metropolis updates in the following steps:

1. The labels $(g_i,\, h_i)$ are drawn as a block for each individual with probabilities

$$P(g_i = g, h_i = h \,|\, \text{rest}) \propto p_{gh} q_g \exp\left\{-\frac{(\text{s}_i - \boldsymbol{\mu}_{gh})'(\text{s}_i - \boldsymbol{\mu}_{gh})}{2\sigma^2}\right\} \prod_{l=1}^{L} \prod_{j=1}^{m_l} (\theta_{glj})^{z_{ilj}}$$

2. $\boldsymbol{\theta}_{jl} \,|\, \text{rest} \sim \text{Dirichlet}(\alpha_l/\rho_l^* + \sum_{i=1}^{n} I(g_i = j) z_{il})$
3. $U_j \,|\, \text{rest} \sim \text{Beta}(1 + \sum_{i=1}^{n} I(g_i = j), b_U + \sum_{i=1}^{n} I(g_i > j))$
4. $V_{gj} \,|\, \text{rest} \sim \text{Beta}(1 + \sum_{i=1}^{n} I(g_i = g, h_i = j), b_V + \sum_{i=1}^{n} \times I(g_i = g, h_i > j))$
5. $\sigma^2 \,|\, \text{rest} \sim \text{InvGamma}(a_1 + n, b_1 + \sum_{i=1}^{n} (\text{s}_i - \mu_{g_i h_i})' \times (\text{s}_i - \mu_{g_i h_i})/2)$
6. $\mu_{jgh} \,|\, \text{rest} \sim \text{N}_{[-1,1]}(\frac{\sum_{i=1}^{n} I(g_i = g, h_i = h) s_{ij}}{\sum_{i=1}^{n} I(g_i = g, h_i = h)}, \frac{\sigma}{\sqrt{\sum_{i=1}^{n} I(g_i = g, h_i = h)}})$
7. $b_U \,|\, \text{rest} \sim \text{Gamma}(K - 1 + a_2, b_2 - \sum_{j=1}^{K-1} \log(1 - U_j))$
8. $b_V \,|\, \text{rest} \sim \text{Gamma}(K(K-1) + a_3, b_3 - \sum_{k=1}^{K} \sum_{j=1}^{K-1} \times \log(1 - V_{kj}))$
9. ρ_l^* is updated using Metropolis sampling using a $\rho_l^* \sim \text{Beta}(100 r_l,\, 100(1 - r_l))$ candidate distribution, where r_l is the previous draw for ρ_l^*. The acceptance probability is

$$\frac{\text{Dirch}(\theta_l \,|\, \boldsymbol{\alpha}_l/\rho_l^*)\{(1 - \pi_l)\exp(-\rho_l^*/\lambda_{1l})/\lambda_{1l} + \pi_l I(0 < \rho_l < \lambda_{2l})\}\text{Beta}(r_l \,|\, 100\rho_l^*, 100[1 - \rho_l^*])}{\text{Dirch}(\theta_l \,|\, \boldsymbol{\alpha}_l/r_l)\{(1 - \pi_l)\exp(-r_l/\lambda_{1l})/\lambda_{1l} + \pi_l I(0 < r_l < \lambda_{2l})\}\text{Beta}(\rho_l^* \,|\, 100 r_l, 100[1 - r_l])},$$

where $I(\cdot)$ is the indicator function, $\boldsymbol{\mu}_{gh} = (\mu_{1gh}, \mu_{2gh})'$, $\text{N}_A(\mu,\, \sigma)$ is the truncated normal distribution with center μ, scale σ, and domain A, and the hyperparameters are $\sigma^{-2} \sim \text{Gamma}(a_1,\, b_1)$, $b_U \sim \text{Gamma}(a_2,\, b_2)$, $b_V \sim \text{Gamma}(a_3,\, b_3)$.