BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–9

StructHDP: Automatic inference of number of clusters and population structure from admixed genotype data

Suyash Shringarpure [1], Daegun Won [1] and Eric P. Xing [1*]

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Clustering of genotype data is an important way of understanding similarities and differences between populations. A summary of populations through clustering allows us to make inferences about the evolutionary history of the populations. Many methods have been proposed to perform clustering on multi-locus genotype data. However, most of these methods do not directly address the question of how many clusters the data should be divided into and leave that choice to the user.

Methods: We present StructHDP, which is a method for automatically inferring the number of clusters from genotype data in the presence of admixture. Our method is an extension of two existing methods, *Structure* and *Structurama*. Using a Hierarchical Dirichlet Process, we model the presence of admixture of an unknown number of ancestral populations in a given sample of genotype data. We use a Gibbs sampler to perform inference on the resulting model and infer the ancestry proportions and the number of clusters that best explain the data.

Results: To demonstrate our method, we simulated data from an island model using the neutral coalescent. Comparing the results of StructHDP with *Structurama* shows the utility of combining HDPs with the *Structure* model. We used StructHDP to analyze a data set of 155 Taita thrush, *Turdus helleri*, which has been previously analyzed using *Structure* and *Structurama*. StructHDP correctly picks the optimal number of populations to cluster the data. The clustering based on the inferred ancestry proportions also agrees with that inferred using *Structure* for the optimal number of populations. We also analyzed data from 1048 individuals from the Human Genome Diversity project from 53 world populations. We found that the clusters obtained correspond with major geographical divisions of the world, which is in agreement with previous analyses of the dataset.

Availability: StructHDP is written in C++. Please contact the authors for the source code.

Contact: suyash@cs.cmu.edu

1 INTRODUCTION

The clustering of individuals into populations is a frequently-undertaken task in studies of genetic data. Clustering allows the summarization of individuals into groups based on genetic similarity. Such a summary is easy to visualize and understand. It enables us to make simple inferences about the similarities and differences between different groups of individuals. We can also make inferences about evolutionary history of populations. An understanding of the population stratification present in a set of individuals is also required for avoiding false positives in genetic association studies.

Phenomena like migration and admixture make the clustering problem harder, since individuals need not belong to exactly one population cluster, but can have partial membership in multiple clusters. Different parts of an individual's genome can be inherited from individuals belonging to different populations. The problem of clustering individuals while allowing partial membership in multiple clusters was addressed using a Bayesian model by Pritchard et al. [13]. Extensions to the underlying model that account for other evolutionary processes such as mutation [17] and recombination [5] have also been proposed.

An important question that needs to be addressed when solving the problem of population stratification is deciding how many populations are required to best explain the variation observed in a given set of individuals. The Bayesian models mentioned earlier require the user to specify a number of clusters (also called ancestral populations) into which the individuals are divided. However, this might not always be possible or desirable. To address this problem, an extension of *Structure* was developed by Pella et al. [12]. Based on their method, Huelsenbeck et al. [9] developed *Structurama*. *Structurama* automatically infers the number of population clusters into which a given data set should be divided provided individuals only belong to a single population. A commonly-used solution to this problem is to use fixed-dimensionality models in combination with an information criterion [1, 16] to decide the optimal number of populations.

In this paper, we present StructHDP, a method for automatically inferring the number of population clusters present in a group of individuals, while accounting for admixture between ancestral populations. Using the Hierarchical Dirichlet Process framework for clustering developed by Teh et al. [20], we extend the *Structure* model so that the number of populations is inferred by the model and need not be specified by the user. This work is also an extension of the Dirichlet process model developed by Pella et al. [12] which has been implemented in *Structurama*.

We simulated data from an island model using the neutral coalescent to test the performance of our method at recovering the true number of ancestral populations. Comparing the results of StructHDP with *Structurama* shows the utility of combining HDPs with the *Structure* model. We used StructHDP to analyze

*to whom correspondence should be addressed

© Oxford University Press 2011.

a set of 155 Taita thrush individuals, *Turdus helleri*. This dataset has been previously analyzed using *Structure* and *Structurama*. We found that StructHDP correctly identifies the optimal number of populations to cluster the data. The clustering enforced by the inferred ancestry proportions for individuals also agrees with that inferred using *Structure* with the appropriate choice of the number of populations K. We also analyzed a set of 1048 individuals from the Human Genome Diversity Project (HGDP) using StructHDP. We found that the clusters inferred coincide with the major geographical divisions present in the data. We also observed that the distance between populations (based on their cluster memberships) is strongly positively correlated with F_{st} between populations, which suggests that the inferred cluster memberships capture the genetic variation present in the data well.

2 RELATED WORK

Model-based clustering methods have become popular since the *Structure* model by Pritchard et al. [13]. The method uses a Bayesian model to capture admixture in populations, with model parameters inferred by Markov Chain Monte Carlo (MCMC) sampling. Extensions to the underlying model that account for other evolutionary processes such as mutation [17] and recombination [5] have also been proposed. These methods have the advantage that they use an explicit demographic model of population stratification. However, they are not efficient for large datasets.

Population clustering methods have also been developed in a complementary direction, using low-dimensional projections and eigenanalysis to cluster individuals [11]. These methods assume no underlying evolutionary model but have been shown to be good at modeling population structure. They are efficient and also allow the user to compute the statistical significance of the results obtained.

Recently, a set of methods such as *Admixture* [2] and *Frappe* [19] have been developed that use the likelihood model of *Structure* but use faster optimization methods for estimating model parameters. *Frappe* uses an expectation-maximization algorithm to determine individual ancestries. *Admixture* uses a block relaxation algorithm to speed up the optimization. Both these models are faster than *Structure* and allow the analysis of large datasets with hundreds of thousands of markers.

In all the models mentioned above, an important choice that must be made by the user is the number of populations (K) or clusters that the dataset must be divided into. To help the user make this choice, some methods provide an approximation to the posterior probability of the number of populations given the data. Alternatively, the user can use information criterion such as the Akaike Information Criterion (AIC) [1] or the Bayesian Information Criterion (BIC) [16] to choose the number of populations that best explain the data. The choice of number of populations is a trade-off between model expressiveness (how well the model captures the variation in the data) and model complexity (how many parameters the model needs to capture the variation).

Huelsenbeck et al. [9] developed an extension to *Structure*, called *Structurama*, that automatically infers the number of populations that are present in the data. Their method assumes that an individual only belongs to a single cluster and that the number of populations is also a random variable. They use a Dirichlet process prior to infer the number of populations automatically. Coalescent simulations by Huelsenbeck et al. [9] using island models show that inference of the number of populations is accurate when migration rates are low and differentiation between populations is high.

We propose to extend the *Structure* model beyond *Structurama* to allow for admixture between ancestral populations. An individual can have membership in multiple clusters and the number of populations will automatically be inferred by the model. In the following sections, we describe the model and discuss the results of analyzing some simulated and real data using StructHDP.

3 APPROACH

Bayesian models for clustering of genotype data use the framework of mixture models to model individuals. In a mixture model, an individual is assumed to be made up of a number of genetic markers. Each of those markers is assumed to originate from one of a finite number of ancestral populations. Ancestral populations are usually defined as a collection of allele frequencies at the markers under consideration. An individual can thus be considered to be a mixture of one or more ancestral populations, which explicitly accounts for admixture between populations. The proportions of different populations within an individual's genome are usually individual-specific and provide a compact summary of the individual's genome. This is the basis of the *Structure* model by Pritchard et al. [13].

We approach the problem of finding the optimal number of ancestral populations by extending the above model to a setting where there are potentially infinite ancestral population components in the mixture. Performing inference then allows us to examine the number of ancestral populations that have a non-zero contribution to the set of individuals under consideration. We use the Hierarchical Dirichlet Process framework [20] to model the mixture of infinite ancestral populations.

Consider the problem of clustering the markers within a single individual based on their population of origin. We can assume that the number of populations that contribute to the single individual's genome is unknown and is a random variable. The Dirichlet process (DP) [6] was proposed to solve a problem of this nature, where objects (genetic markers) belong to one of a potentially infinite number of mixture components (ancestral populations). In the case of multiple individuals, we can posit multiple DPs, one for each individual, that will address the problem of not knowing the optimal number of populations. We also require that the ancestral populations inferred for the DPs be the same across all the individuals. Mathematically, this is analogous to ensuring that mixture components are shared across DPs.

The Hierarchical Dirichlet process (HDP) is a framework for clustering of observations when the observations are present in groups. Each group can be modeled using a finite mixture model or a Dirichlet process. The mixture models or DPs across groups are linked by sharing mixture components. It is useful to think of each group as having its own Dirichlet processes, with the processes linked to each other by the parameters of the HDP. StructHDP is based on the Hierarchical Dirichlet process described by Teh et al. [20]. In the following section, we provide a description and mathematical representation of the HDP model.

4 METHODS

A commonly used analogy for representing HDPs is the Chinese Restaurant Franchise (CRF). This is an extension of the representation of the Dirichlet process (DP) as a Chinese restaurant with customers. The DP representation and its application to *Structurama* are described in more detail by Huelsenbeck et al. [9]. A CRF comprises of a number of Chinese restaurants which share a common (possibly infinite) menu of dishes. In a CRF, each restaurant corresponds to a group of observations, and the customers are observations. The dishes served in the restaurant are the mixture components, and sharing of mixture components across groups corresponds to sharing of dishes across restaurants. In the CRF metaphor, a new customer (observation) arrives at the restaurant corresponding to its group. The customer chooses a previous occupied table in the restaurant with a probability proportional to the number of customers already at the table, or, with a constant probability, chooses a new table. Every table serves a dish from the possible set of dishes, and every customer at the table is assigned that particular dish, i.e, the observation is assigned the particular mixture component that is associated with the table. All observations that are assigned to a particular table are considered to originate from the same mixture component, clustering the observations within the group. The same mixture component might also be shared across multiple tables within a group. The method of choosing a table for a new customer is similar to a "rich gets richer" model which is regulated by the probability of starting a new table. This is the property of the HDP that is responsible for its clustering behavior.

This analogy can be easily extended to the case of genetic data, with every individual considered to be a separate group corresponding to a restaurant. The loci within an individual are the customers in the restaurant, and the ancestral populations are the mixture components or the dishes in the CRF. A minor subtlety that arises in this case is that the set of possible alleles at each locus might be different, which needs to be accounted for in the inference process. This can be accomplished easily with some minor bookkeeping without changing the inference process significantly.

Consider a dataset having N individuals genotyped at M loci. The observed allele for individual j at locus i is denoted by x_{ji}. For ease of representation, we will ignore the diploid nature of genotype data. In implementation, we shall allow our method to handle data of any fixed ploidy. The HDP can then be used to generate the allele x_{ji} for the j^{th} individual at the i^{th} locus as follows:

$$\begin{aligned} G_0|\gamma, H_i &\sim DP(\gamma, H_i) \\ G_j|\alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ \theta_{ji}|G_j &\sim G_j \\ x_{ji}|\theta_{ji} &\sim \mathrm{F}(\theta_{ji}) \end{aligned}$$

Here, H_i is the base distribution over alleles at locus i, commonly a Dirichlet distribution. γ and α_0 are parameters of the HDP that control how fast new populations are added to the model. G_0 is an intermediate probability distribution over alleles at locus i and G_j is a distribution specific to individual j. The individual-specific distributions G_j are connected to one another through G_0 and α_0, ensuring the sharing of ancestral populations across individuals. G_0 and G_j are both generated by Dirichlet processes (DP) that use γ and α_0 as parameters. The θs denotes the mixture components. x_{ji} is a sample from a distribution $F(\theta_{ji})$, a multinomial distribution over alleles in our case.

For modeling purposes, it is helpful to modify the representation of the HDP so that the generative process looks as follows.

$$\beta|\gamma \sim \mathrm{GEM}(\gamma) \quad (1)$$

$$\pi_j|\alpha_0, \beta \sim \mathrm{DP}(\alpha_0, \beta) \quad (2)$$

$$\phi_{ik}|H_i \sim H_i \quad (3)$$

$$z_{ji}|\pi_j \sim \pi_j \quad (4)$$

$$x_{ji}|z_{ji}, (\phi_{ik})_{k=1}^{\infty} \sim \mathrm{F}(\phi_{z_{ji}}) \quad (5)$$

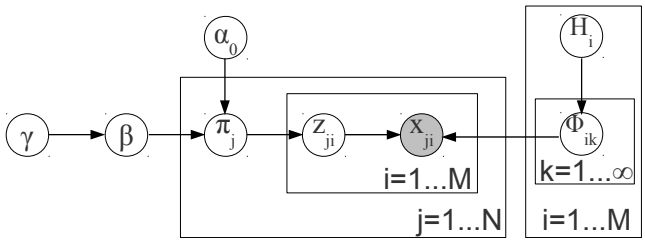

Fig. 1. Graphical model representation of the generative process of StructHDP. Nodes represent random variables and edges indicate dependencies between random variables. The shaded circle indicates the observed alleles. The dataset has N individuals, each genotyped at M loci. For ease of representation, we do not show the ploidy of the individual in the graphical model.

where we say that $\beta = (\beta_k)_{k=1}^{\infty} \sim \mathrm{GEM}(\gamma)$ if it satisfies the following construction:

$$\beta'_k|\gamma \sim \mathrm{Beta}(1, \gamma) \quad (6)$$

$$\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \quad (7)$$

This construction ensures that $\sum_{k=1}^{\infty} \beta_k = 1$. The β thus represents the fractional contributions of the potentially infinite populations to the given set of individuals.

In the HDP representation above, ϕ_{ik} represents the allele frequencies of the k^{th} population at the i^{th} locus. π_j is a vector that denotes the ancestry proportions (contributions from all populations) for individual j, and its components sum to 1. The indicator variable z_{ji} denotes which population the observed allele x_{ji} at locus i originates from. We will use this notation for representing the HDP model for our problem due to its similarity with the *Structure* generative process. This representation also shows how the model can account for diploid individuals by changing the step of sampling z_{ji} and x_{ji} to the following:

$$\begin{aligned} z_{ji,1}|\pi_j &\sim \pi_j \\ z_{ji,2}|\pi_j &\sim \pi_j \\ x_{ji,1}|z_{ji,1}, (\phi_{ik})_{k=1}^{\infty} &\sim \mathrm{F}(\phi_{z_{ji,1}}) \\ x_{ji,2}|z_{ji,2}, (\phi_{ik})_{k=1}^{\infty} &\sim \mathrm{F}(\phi_{z_{ji,2}}) \end{aligned}$$

where $x_{ji,1}$ and $x_{ji,2}$ now represent the two alleles at locus i in individual j and $z_{ji,1}$ and $z_{ji,2}$ are their respective population indicator variables. This allows the model to account for mixed ancestries at a single locus as well. For ease of representation, we will drop the subscript indicating the ploidy in the analysis.

Figure 1 shows the graphical model representation of the StructHDP generative process. In this graphical model representation, the nodes represent random variables which have been described earlier. The edges denote dependencies between the random variables due to the sampling steps in the generative process. The shaded nodes represent the random variables we observe, viz, the alleles observed at each locus.

To allow for more flexibility with the parameter settings, we impose priors on α_0, γ and the base distributions H_i. We assume that α_0 and γ have Gamma priors with parameters (α_a, α_b) and (γ_a, γ_b) respectively and that H_i has a symmetric Dirichlet distribution with parameter λ. The graphical model with all priors shown can be seen in Figure 11 in the Appendix.

$$\alpha_0 \sim \mathrm{Gamma}(\alpha_a, \alpha_b) \quad (8)$$

$$\gamma \sim \mathrm{Gamma}(\gamma_a, \gamma_b) \quad (9)$$

$$H_i \sim \mathrm{Dir}(\lambda) \quad (10)$$

4.1 Inference

For performing inference on the model, we use Gibbs sampling, a MCMC sampling method described for the HDP by Teh et al. [20]. For inference in the CRF representation of the HDP, we create some bookkeeping variables $\mathbf{m}$ that keep count of the number of tables at the restaurant and franchise levels. More details about the distributions used for sampling the variables are presented in the Appendix.

4.2 Inference steps

Using all the variable updates, the inference process can be described as:

1. Set the values for the prior parameters $\alpha_a, \alpha_b, \gamma_a, \gamma_b, \lambda$.
2. Start with random values for all other variables.
3. Sample $\mathbf{z}$ variables given all other variables.
4. Sample $\mathbf{m}$ variables given all other variables, using updated value of $\mathbf{z}$.
5. Sample β given all other variables, using updated values of $\mathbf{z}$ and $\mathbf{m}$.
6. Sample α_0 using updated values of $\mathbf{z}$, $\mathbf{m}$ and β.
7. Sample γ using updated values of all other variables.
8. Repeat 3-7 until convergence.

The Gibbs sampling update distributions can be derived following the methodology in Teh et al. [20]. Due to space limitations, the details of the Gibbs sampling update distributions and their derivations are deferred to the Appendix.

4.3 Other inference details

Like all MCMC methods, the sampler is run for a large number of iterations, with some initial iterations discarded as burn-in. Samples from the posterior can then be used to estimate the ancestry proportions π_j for each individual. The posterior distribution for the individual ancestry proportions π_j can be shown to be a Dirichlet distribution.

$$\pi_j \sim \text{Dir}\left(\cdots, \alpha_0\beta_k + \sum_{i=1}^{M} \mathcal{I}[z_{ji} = k], \cdots\right) \tag{11}$$

where $\mathcal{I}[.]$ denotes an indicator function. If the number of populations remains constant across iterations in the sampling (as is often seen to happen in our experiments after a large number of iterations), this estimate can be averaged over multiple samples to get a more accurate estimate of the individual ancestry proportions.

As with the Gibbs sampler used in *Structure*, our method could have problems with the identifiability of clusters, if label switching for the clusters were a frequent occurrence. In practice, we find that label switching is infrequent, and is avoided by the use of the restricted growth function (RGF) notation of Stanton et al. [18] in summarizing MCMC results.

5 RESULTS

5.1 Coalescent simulation data

We performed coalescent simulations based on an island model similar to Huelsenbeck et al. [9]. We used the program *ms* [8] to simulate samples under a neutral coalescent model. As an initial evaluation of the performance of StructHDP in recovering the correct number of population clusters, we simulated data from a symmetric equilibrium island model with 4 demes of equal size, with the mutation rate $\theta = 4N_e\mu = 0.5$ and migration rate $M = 4N_em = \{1, 2, 4\}$. In each case, 100 diploid individuals were sampled with an equal number being sampled from each deme. 50 replicates were created for each parameter setting.

We analyzed the data using StructHDP, *Structurama* and *Admixture*. For StructHDP, the priors for both concentration parameters were set to (0.5,0.5) and the parameter for the Dirichlet distribution of H was set to 0.5. The StructHDP Gibbs sampler was run for 25,000 iterations, with the first 12,500 iterations discarded as burn-in. To thin the Markov chain, samples were taken every 25 iterations. We computed the expected value of the number of populations, K, using the sampled values of K from the Gibbs sampler. The expected value of K, $E(K|X)$ can then be compared against the true value of the number of demes, $K_T = 4$, across multiple replicates, to get an error measure that is given by $E(E(K|X) - K_T)$ [9].

Method ↓ / Migration rate →	M=1	M = 2	M =4
StructHDP	0.10	**0.01**	**0.15**
Structurama	**0.0**	-0.21	-1.31
Admixture+AIC	-1.8	-1.73	-1.65
Admixture+BIC	-2.6	-2.78	-2.62
Admixture+CV	2.5	2.63	2.71

Table 1. Comparison of simulation results for StructHDP, *Structurama* and *Admixture*. 50 replicates, consisting of 100 diploid individuals each, were sampled from a 4-deme symmetric island model, with $\theta = 0.5$ and $M = \{1, 2, 4\}$. The error in recovering the number of demes is shown, as computed by the error measure $E(E(K|X) - K_T)$.

For *Structurama*, the experiments for each parameter setting were performed with different priors on the expected number of populations in [9]. For comparison purposes, we chose the best result, i.e, the prior setting that gave the least error. Model selection with *Admixture* can be done in three different ways by choosing either the AIC, BIC or the cross-validation error as the measure of model fit. We present results for all three measures.

Table 1 shows the results of the simulation. We can see that the error in recovering K is much smaller for StructHDP than for *Structurama* and for *Admixture*, except when the migration rate is small. The underlying assumption of the Dirichlet process model of *Structurama* is that there is no admixture and individuals only belong to a single ancestral population. As a result, in a simulation setting with less admixture due to migration, the number of recovered populations for *Structurama* is almost perfect. As the amount of admixture increases, the error in the number of recovered populations increases. On the other hand, StructHDP explicitly accounts for admixture in the model. Therefore it recovers the true number of demes in the island model with low error for all parameter values. In terms of F_{st}, we can say that as the F_{st} between the demes decreases (as migration increases), the accuracy of *Structurama* drops while that of StructHDP is nearly unaffected.

Admixture performs worse than both StructHDP and *Structurama* in recovering the true number of populations. This may be due to the small number of markers that are used in the simulation study.

5.2 Real data analysis

5.2.1 ***Taita thrush data:*** We used our method to analyze a data set of N = 155 Taita thrush, *Turdus helleri* [7]. Each individual was genotyped at M = 7 microsatellite loci. Individuals were sampled at four locations in southeast Kenya [Chawia (17 individuals), Ngangao (54), Mbololo (80), and Yale (4)]. The thrush data were previously analyzed in [13, 9] so we use it to verify the correctness of StructHDP.

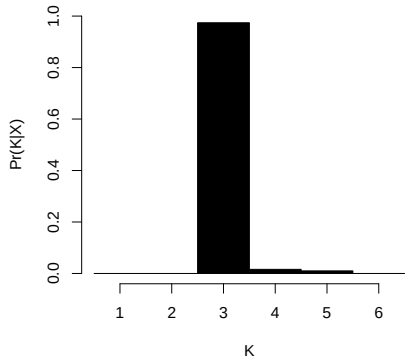

Fig. 2. Posterior distribution for number of populations, $Pr(K|X)$ for the thrush data.

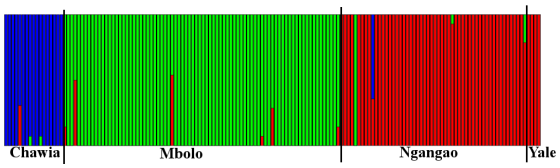

Fig. 3. A single sample of the ancestry proportions for the thrush data. The black lines separate the individuals according to their geographic labels. The analysis did not use any geographical information.

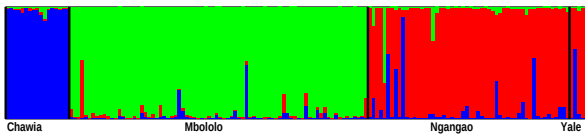

Fig. 4. The ancestry proportions for the thrush data from a single *Structure* run for K=3.

We ran StructHDP for 25,000 iterations, with the first 12,500 iterations as burn-in. Samples were taken every 25 iterations to thin the Markov chain. The priors for both concentration parameters were set to (0.5,0.5) and the parameter for the Dirichlet distribution of H was set to 0.5.

We find that our method converges to K=3 populations in a few thousand iterations. The posterior distribution for K is shown in Figure 2. From the posterior, we can see that $K = 3$ is the most likely value for K. Figure 3 shows a single sample for the ancestry proportions of the thrush data. The clusters agree with geographical labels well except for a few individuals. We also see that the 4 Yale individuals fall into the same cluster as the Ngangao individuals. All of these findings agree with those of Pritchard et al. [13] when *Structure* is initialized with $K = 3$ clusters. Figure 4 shows the results of *Structure* analysis of the thrush data with $K = 3$. In their analysis, Pritchard et al. also found that $K = 3$ explains the data best. Their conclusion was based on an *ad hoc* approximation to $Pr(K|X)$, the posterior likelihood of K given the data X while StructHDP automatically infers this from the data.

5.2.2 ***Human Genome Diversity Project:*** The Human Genome Diversity Project dataset we analyze consists of 1048 individuals from 53 world populations genotyped at 783 microsatellite loci.

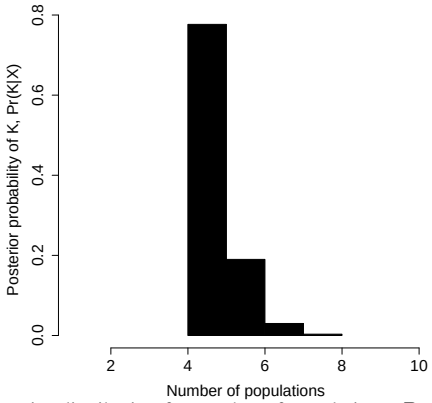

Fig. 5. Posterior distribution for number of populations, $Pr(K|X)$ for the HGDP data.

Along with genotype information, the individuals are also labeled with the geographical divisions to which they belong. Using *Structure*, Rosenberg et al. [15] have previously analyzed the genotype data and found that the population clusters correspond to major geographical divisions of the world. We used StructHDP to reanalyze this data (without making use of the geographical information). The sampler was run for 20,000 iterations with the first 10,000 iterations discarded as burn-in. Samples were taken every 25 iterations to thin the Markov chain.

To determine the optimal number of ancestral populations, we examined the posterior distribution of the number of populations (K). Figure 5 shows the posterior distribution. We find the posterior distribution has a single mode at $K = 4$ and non-zero probability mass for values of K up to 8. For further analyses, we use the maximum-likelihood sample from the MCMC sampling, which has 4 ancestral population components.

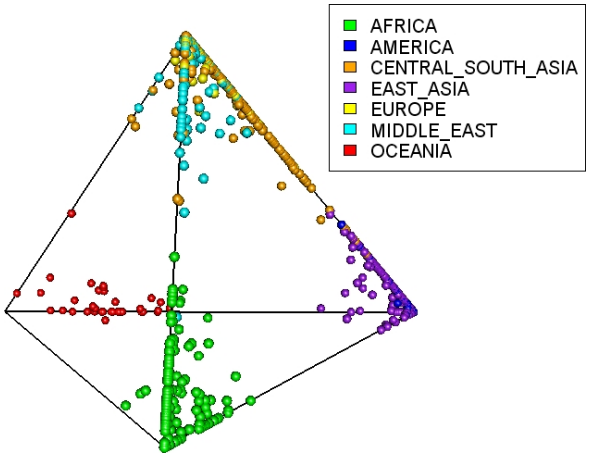

Fig. 6. The ancestry proportions for the 1048 individuals from the Human Genome Diversity Project plotted in 3-dimensional space. Each individual is represented by a small sphere and the color of the sphere depends on the continental division the individual belongs to. Different colors correspond to different continental divisions. The geographical divisions are indicated by the labels on top of the graph.

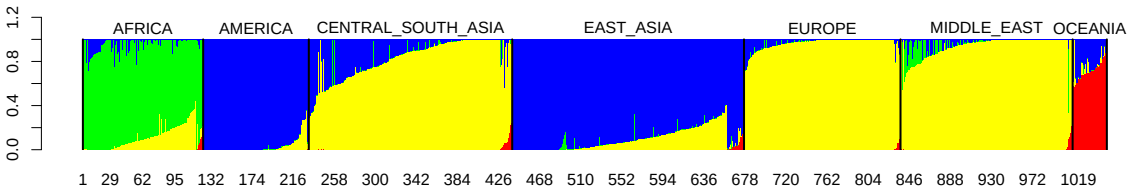

Fig. 7. The ancestry proportions for the 1048 individuals from the Human Genome Diversity Project inferred by StructHDP. Each thin line denotes the ancestry proportions for a single individual. Different colors correspond to different ancestral populations. Dark black lines separate individuals from different major geographical divisions. The geographical divisions are indicated the labels on top of the graph.

The contributions of the four ancestral populations to an individual's genome can be represented using a 4-dimensional vector whose components sum to 1. All these vectors (referred to as ancestry proportions) lie within a tetrahedron in 3-dimensional space. Each of the four vertices of the tetrahedron represents an ancestral population. To visualize the clustering, we plotted the ancestry proportions for the 1048 individuals in 3 dimensions along with the tetrahedron in which the vectors lie. In this representation, the distance of a vector from the vertices of the tetrahedron indicates the amount of admixture present in an individual's genome. The further away from a vertex the vector is (and the closer it is to the center of the tetrahedron), the more the admixture present in the individual's genome.

Figure 6 shows the resulting plot for the 1048 individuals in the HGDP dataset. In the plot, each individual is represented by a small sphere. For ease of interpretation, the individual spheres are colored based on the geographical division they belong to. In the populations we examine, the divisions are Africa, Americas, Central and South Asia, East Asia, Europe, Middle East and Oceania. These are represented by seven different colors. From the figure, we can see that individuals from a single continent cluster together in the same region of the tetrahedron. Some individual genomes are derived from a single ancestral population and lie at the vertices of the tetrahedron. Some other individuals, particularly those belonging to the Middle Eastern, Central Asian and South Asian populations, show a lot of admixture.

To analyze these results further, we plotted the ancestry proportions of the 1048 individuals as a bar graph, where every individual is represented by a thin bar with 4 components which sum to 1. Figure 7 shows the resulting bar graph.We can see that the clusters obtained correspond to the major geographical divisions of the world and the ancestral populations can be roughly described as *ancestral African* (denoted by green color), *ancestral American-East Asian* (blue), *ancestral European* (yellow) and *ancestral Oceanian* (red). From the ancestry proportions, we can see that the modern East Asian populations and American populations are similar, with the modern East Asian populations having a larger contribution from the ancestral population corresponding to Europe. Modern Asian populations also show some Oceanic ancestry (from the ancestral population denoted by red color). Modern Central and South Asian populations show an admixture of European and East Asian ancestral populations. The Middle Eastern populations show contributions from the ancestral African population and the ancestral European population. Modern Oceanic populations are an admixture of an ancestral Oceanic population with an ancestral East Asian population. All of these observations are in agreement with previous analyses of the data by Rosenberg et al. [15] and other studies of regional populations. We should note that the clusters inferred by StructHDP are not identical to the ones observed by Rosenberg et al. for $K = 4$. Rosenberg et al. observe that East Asia separates out into a separate cluster for $K = 4$ while Oceania separates from the rest of the data only for values of K larger than 4.

To analyze the similarity and differences within and between continental divisions, we computed the mean ancestry proportions for the 7 continental divisions by averaging the ancestry proportions for all individuals belonging to each continental division. We then constructed a distance matrix by computing the euclidean distance between the 4-dimensional vectors representing each continental division. Figure 8 shows the resulting distance matrix. From the figure, we can see that the distance matrix has a block structure. Modern American and East Asian populations are similar to each other and show little separation. We also see that modern European, Central-South Asian and Middle Eastern populations are close to each other. Within these 3 divisions, we see that Europeans and Middle Eastern populations group together while the Central-South Asians are further apart.

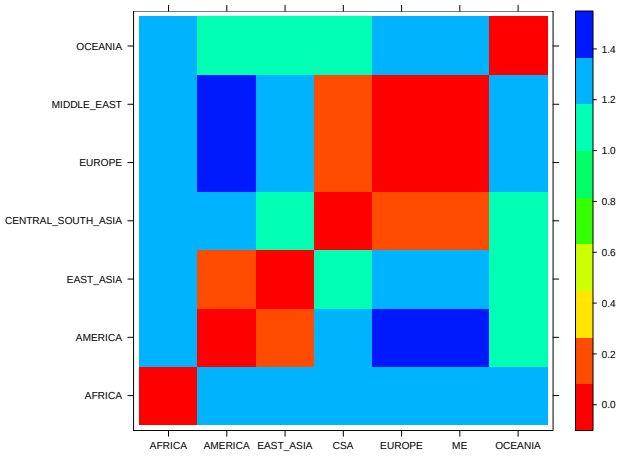

Fig. 8. A matrix representing the distances between the mean ancestry proportions of the 7 major continental divisions of the HGDP. Red color indicates less distance while blue color indicates more distance.

We hypothesized that if the inferred ancestry proportions capture the genetic variation between and across populations, then the pairwise Euclidean distance computed earlier should be correlated with genetic distance. To test this hypothesis, we computed the pairwise F_{st} distance between the 7 continental divisions of the data. To test for correlation between the pairwise Euclidean distance matrix and the pairwise F_{st} distance matrix, we used a Mantel test. A Mantel test tests the alternate hypothesis of correlation between two matrices against the null hypothesis of no correlation by permuting the rows and columns of one of the matrices and observing the distribution of the correlation statistic. The Mantel test on the Euclidean and F_{st} distance matrices shows that the correlation between the two distance matrices is 0.57 (P-value = 0.0025 with 10,000 replicates). The distribution of the observed and simulated Mantel correlation statistic is shown in Figure 9. Thus, we can see that the Euclidean distance and F_{st} distance are strongly positively correlated, which supports the inferred population structure.

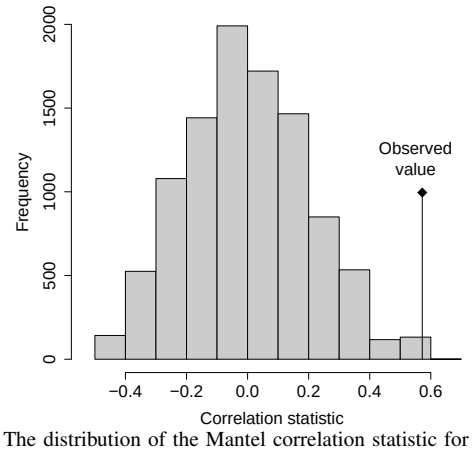

Fig. 9. The distribution of the Mantel correlation statistic for the pairwise Euclidean distance matrix and the pairwise F_{st} distance matrix. The stem indicates the observed value of the statistic. The result is significant, with the associated P-value=0.0025

To compare our results on the HGDP data with other methods, we analyzed the data using *Structurama*. However, due to computational reasons, we were unable to run *Structurama* on the full data at optimal settings. Therefore we analyzed a subset of the data that included only 100 loci per individual. We found that the posterior distribution of K inferred by *Structurama* has non-zero mass only at $K = 5$. Figure 10 shows the inferred ancestry proportions based on the mean partition from *Structurama*. We can see that *Structurama* also clusters the European, Middle Eastern and Central South Asian populations into a single cluster. However, since it does not allow partial membership, the individuals in different clusters have zero similarity. It is therefore unable to model the partial similarity between populations from different geographical divisions, e.g., the Central Asian populations and European populations.

6 DISCUSSION

We have presented StructHDP, a method for automatically inferring the number of population clusters present in a group of individuals while accounting for admixture between populations. At the same time, it also infers individual ancestry estimates under a *Structure*-like model. We demonstrated the effectiveness of our method on data simulated from an island model. We also analyzed the Taita thrush dataset and demonstrated that StructHDP chooses the number of clusters that best explain the data. Our analysis of the HGDP dataset shows that our method is able to cluster populations even when the individuals in the dataset are admixed. The ancestry proportions inferred for populations can be used to compute a distance measure between populations. We found that the Euclidean distance between populations has a strong positive correlation with the F_{st} distance between populations. The ancestry proportions therefore provide a useful low-dimensional representation of populations.

Our method uses a Hierarchical Dirichlet process to model the admixture of an unknown number of ancestral populations present in individual genomes in a given dataset. We use an MCMC sampling algorithm, Gibbs sampling, to estimate the model parameters. The number of ancestral populations that best explain the data is one of the parameters of our model. The collapsed Gibbs sampler we implemented according to [20] marginalizes the uncertainty in the population allele frequencies, thus eliminating a possible source of error in the inference. Our experiments suggest that the HDP is not sensitive to the priors on the parameters α_0 and γ since we sample them in the algorithm. The results are more sensitive to the choice of λ for the base distributions. A large value of λ tends to produce populations with uniform (high-entropy) allele frequency distributions while a small value of λ produces populations with allele frequency distributions highly skewed in favor of very few alleles (low-entropy). A complete sensitivity analysis, however, is beyond the scope of this paper. It would also be instructive to perform simulations with more realistic and complex demographic models to understand the limitations of StructHDP.

The model as described here can handle both SNP and microsatellite markers. However, one of the limitations of our method is the computational time required for the Gibbs sampling. This means that while our method can handle datasets of a few thousand markers and individuals, it cannot be efficiently used on large datasets of hundreds of thousands of markers. However, as our simulations show, even with few loci, the method performs well at recovering the number of populations required to explain the data best. Teh et al. [21] have described a way of implementing collapsed variational inference for HDPs. Applying the variational inference algorithm to StructHDP would improve its speed significantly.

In this work, we have shown how the basic admixture model can be extended to allow automatic inference of the number of populations. Just as extensions to the *Structure* model that account for recombination [5] and mutation [17] have been developed, we can also extend StructHDP to model other evolutionary processes.

Genetic datasets are often accompanied by geographical information about the genotyped individuals. In some cases, there is a single geographical label associated with each individual, while in others, there are labels at different resolutions (for example, region, nation, continent). It has been shown that geographical distance correlates well with genetic distance between populations [3, 10,

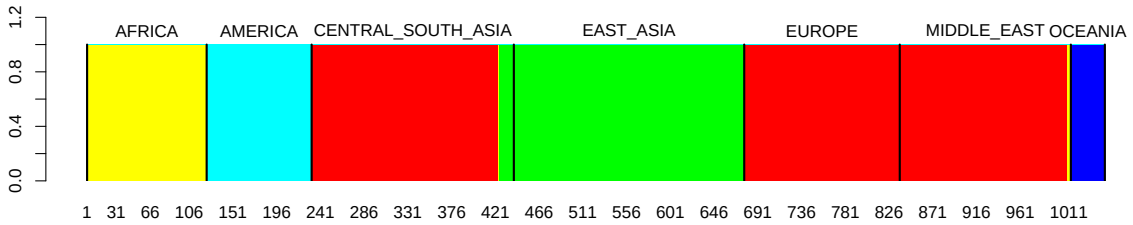

Fig. 10. The ancestry proportions for the 1048 individuals from the Human Genome Diversity Project inferred by *Structurama*. Each thin line denotes the ancestry proportions for a single individual. Different colors correspond to different ancestral populations. Dark black lines separate individuals from different major geographical divisions. The geographical divisions are indicated the labels on top of the graph.

14]. Therefore the amount of sharing of ancestral population components between modern population groups is likely to depend on their geographical distance.

In its current form, StructHDP does not make use of geographical information in the inference process. Teh et al. [20] describe how a HDP can be extended to include multiple levels of hierarchy and be generalized to a tree-like hierarchy. Use of the hierarchical geographical labels could allow us to impose a tree structure on the dataset that respects the geographical labels and enforces a level of population sharing among individuals that is consistent with their geographical labels and distances.

ACKNOWLEDGEMENTS

This research was made possible by grants NSF DBI-0546594, IIS-0713379, NIH 1R01GM087694, DARPA NBCH1080007, and an Alfred P. Sloan Fellowship to E.P.X. S.S was also supported by a Richard King Mellon Foundation Fellowship.

REFERENCES

[1]H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2]D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

[3]L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The history and geography of human genes*. Princeton Univ Pr, 1994.

[4]M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577, June 1995.

[5]D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

[6]T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.

[7]P. Galbusera, L. Lens, T. Schenck, E. Waiyaki, and E. Matthysen. Genetic variability and gene flow in the globally, critically-endangered Taita thrush. *Conservation Genetics*, 1:45–55, 2000.

[8]R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[9]J. Huelsenbeck and P. Andolfatto. Inference of population structure under a Dirichlet process prior. *Genetics*, 175(April):1787–1802, 2007.

[10]J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

[11]N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 2006.

[12]J. Pella and M. Masuda. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*, 63(3):576–596, 2006.

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

[14]S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–7, 2005.

[15]N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic Structure of Human Populations. *Science*, 298(5602):2381–2385, 2002.

[16]G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[17]S. Shringarpure and E. P. Xing. mStruct: inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, 182(2):575–593, 2009.

[18]D. Stanton and D. White. *Constructive combinatorics*. Undergraduate texts in mathematics. Springer-Verlag, 1986.

[19]H. Tang, J. Peng, P. Wang, and N. J. Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, 2005.

[20]Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[21]Y. W. Teh, K. Kurihara, and M. Welling. Collapsed Variational Inference for HDP. *Advances in Neural Information Processing Systems 20*, 20:1481–1488, 2008.

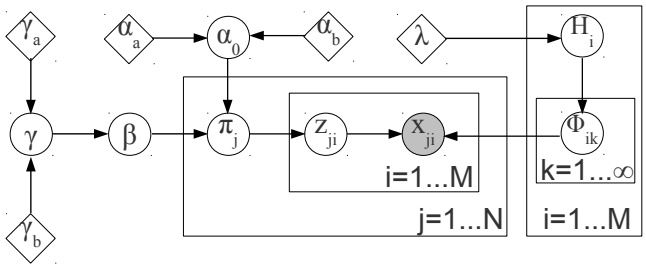

Fig. 11. Graphical model representation of StructHDP with all priors represented. The shaded circle indicates the observed alleles. The dataset has N individuals, each genotyped at M loci. For ease of representation, we do not show the ploidy of the individual in the graphical model. The diamonds indicate parameters that are supplied by the user.

APPENDIX - INFERENCE AND ESTIMATION

The population allele frequencies at locus i are assumed to be $\{\phi_{i1}, \cdots, \phi_{iK}\}$ where K can be infinity and only a finite number of the populations are used in the dataset. The prior over the allele frequencies ϕ_{ik} is H_i. In the restaurant analogy, we use t_{ji} to denote the table for customer x_{ji}, n_{jtk} to denote the number of customers in restaurant j at table t eating dish k, while m_{jk} denotes the number of tables in restaurant j serving dish k. Marginal counts are represented with dots. So $n_{jt.}$ denotes the number of customers in restaurant j at table t, and $m_{..}$ represents the total number of tables in the franchise.

Let $\mathbf{x} = (x_{ji} : \text{all } j, i), \mathbf{x}_{jt} = (x_{ji} : \text{all } i \text{ with } t_{ji} = t), \mathbf{t} = (t_{ji} : \text{all } j, i), \mathbf{z} = (z_{ji} : \text{all } j, i), \mathbf{m} = (m_{jk} : \text{all } j, k)$. When a superscript is used with a set of variables, e.g., x^{-ji} or $n_{jt.}^{-ji}$, this means that the variable corresponding to the index is removed from the set. In the example, $x^{-ji} = \mathbf{x}/x_{ji}$ and $n_{jt.}^{-ji}$ is the number of observations in group j associated with table t leaving out observation x_{ji}.

An important quantity we will use often in sampling is the conditional density of x_{ji} under mixture component k given all data except x_{ji}. This can be computed as

$$f_k^{-x_{ji}}(x_{ji}) = \frac{\int f(x_{ji}|\phi_{ik}) \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'|\phi_{ik}}) h(\phi_{ik}) d\phi_{ik}}{\int \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'|\phi_{ik}}) h(\phi_{ik}) d\phi_{ik}} \tag{12}$$

Here, we are marginalizing out the effects of the uncertainty in the allele frequencies ϕ_{ik}. For our purposes, $f(.|\theta)$ is a multinomial distribution and $h_i(.)$ is a symmetric Dirichlet distribution with parameters λ, on the simplex of dimension P if we observe P different alleles at locus i. Therefore the numerator and denominator are the normalization constants of the posterior Dirichlet distributions.

At locus i, we can represent the observed alleles as $\{a_1, \cdots, a_P\}$. Then we have that

$$f(x_{ji}|\phi_{ik}) = \prod_p \phi_{ik,p}^{\mathcal{I}[x_{ji}=a_p]} \tag{13}$$

Using this in Equation 12 gives us,

$$f_k^{-x_{ji}}(x_{ji}) = \frac{B(h_1 + \sum_{j'i', z_{j'i'}=k} \mathcal{I}[x_{j'i'} = a_1], \cdots)}{B(h_1 + \sum_{j'i' \neq ji, z_{j'i'}=k} \mathcal{I}[x_{j'i'} = a_1], \cdots)} \tag{14}$$

where $B(.)$ is the multinomial beta function, which can be written in terms of the Gamma function:

$$B(\alpha_1, \cdots, \alpha_P) = \frac{\prod_{p=1}^P \Gamma(\alpha_p)}{\Gamma(\sum_{p=1}^P \alpha_p)}$$

Sampling for the population indicator variables z is given by

$$\begin{aligned} p(z_{ji} = k|\mathbf{z}^{-ji}, \mathbf{m}, \beta) &= (n_{j.k}^{ji} + \alpha_0\beta_k) f_k^{-x_{ji}}(x_{ji}) \\ &\quad , \text{ if k is previously used} \\ &= \alpha_0\beta_u f_{k^{new}}^{-x_{ji}}(x_{ji}), \text{ if k is new} \end{aligned}$$

To sample m, we use a result derived in [20],

$$p(m_{jk} = m|\mathbf{z}, \mathbf{m}^{-jk}, \beta) = \frac{\Gamma(\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k + n_{j.k})} s(n_{j.k}, m)(\alpha_0\beta_k)^m$$

where $s(n, m)$ are unsigned Sterling numbers of the first kind.

Sampling for β is given by

$$(\beta_1, \cdots, \beta_k, \beta_u)|\mathbf{m}, \mathbf{k} \sim \text{Dir}(m_{.1}, \cdots, m_{.K}, \gamma)$$

Concentration parameter updates

For updating the concentration parameter α_0, we use the method described by [20], using a sampling scheme of auxiliary variables. For N individuals, define auxiliary variables, $\mathbf{w} = (w_j)_{j=1}^N$ and $\mathbf{s} = (s_j)_{j=1}^N$, where each $w_j \in [0, 1]$ and each s_j is a binary variable in $\{0, 1\}$. Then we have the following sampling scheme

$$\begin{aligned} q(\alpha_0|\mathbf{w}, \mathbf{s}) &\sim \text{Gamma}\left(a + m_{..} + 1 - \sum_j s_j, b + 1 - \sum_j \log(w_j)\right) \\ q(w_j|\alpha_0) &\sim \text{Beta}(\alpha_0 + 1, n_{j..}) \\ q(s_j|\alpha_0) &\sim \text{Binomial}(1, n_{j..}/\alpha_0/(1 + n_{j..}/\alpha_0)) \end{aligned}$$

To update α_0, we iterate these three steps until the value of α_0 converges. Convergence is usually quick and takes about 20-30 iterations.

For updating γ we use the method described in [4], using an auxiliary variable η. Assume γ has a gamma prior Gamma(a, b).

We have

$$\begin{aligned} q(\gamma|\eta, K) &\sim \pi_\eta \text{Gamma}(a + k, 1/(b - \log(\eta))) \\ &\quad + (1 - \pi_\eta)\text{Gamma}(a + K - 1, 1/(b - \log(\eta))) \end{aligned}$$

where the mixture weights are given by

$$\frac{\pi_\eta}{1 - \pi_\eta} = \frac{a + k - 1}{m_{..}(b - \log(\eta))}$$

Secondly, we have

$$q(\eta|\gamma, K) \sim \text{Beta}(\gamma + 1, m_{..})$$

Alternating these updates until the value of γ converges provides a method for updating γ.